COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

THESE DE DOCTORAT DE .

L'INSTITUT NATIONAL DES SCIENCES APPLIQUEES RENNES

ECOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique , Signal, Systèmes, Electronique
Spécialité : *Signal, Image, Vision*

Par
Zhengyu ZHANG

Quality Assessment of Light Field Images

Thèse présentée et soutenue à Rennes, le 11/07/2024
Unité de recherche : IETR
Thèse N° : 24ISAR 13 / D24 - 13

Rapporteurs avant soutenance :

Chaker LARABI	Professeur des Universités, Université de Poitiers
Giuseppe VALENZISE	Chargé de Recherche, HDR, CentraleSupélec, Université Paris-Saclay

Composition du Jury :

Président :	Patrick LE CALLET	Professeur des Universités, Polytech Nantes, Université de Nantes
Examinateurs :	Chaker LARABI	Professeur des Universités, Université de Poitiers
	Giuseppe VALENZISE	Chargé de Recherche, HDR, CentraleSupélec, Université Paris-Saclay
	Meriem OUTTAS	Enseignante-Chercheuse, INSA Rennes
Dir. de thèse :	Luce MORIN	Professeur des Universités, INSA Rennes
Co-dir. de thèse :	Lu ZHANG	Maître de conférences, HDR, INSA Rennes

Acknowledgement

I am sincerely grateful for the opportunity to spend three years of my Ph.D. studies in the VAADER team and IETR laboratory at INSA Rennes, where I have thrived in the company of excellent research environment, knowledgeable supervisors, and supportive teammates. Without these, this thesis would be difficult to finish. All these experiences mean a lot to me and will undoubtedly remain an unforgettable memory throughout my lifetime.

I would like to express my deepest gratitude to my Ph.D. supervisors, Prof. Luce MORIN and Maître de conférences Lu ZHANG, for providing me with the doctoral opportunity, academic guidance, and life assistance. I can't imagine how lucky I am to meet these two lovely and kind supervisors while studying in a new country. Sincerely, it is my honor to be your student! :)

I would also like to thank Prof. Wenbin ZOU and Prof. Shishun TIAN, who built a bridge between my master's and doctoral studies and played an important guiding role in my academic research and life planning.

I want to express my gratitude to Prof. Chaker LARABI, Chargé de Recherche Giuseppe VALENZISE, Prof. Patrick LE CALLET, and Enseignante-Chercheuse Meriem OUTTAS for your time and effort in my thesis and defense. I really appreciate all your comments and suggestions to make my thesis better.

I am profoundly thankful to my teammates and friends, especially Dr. Jianjun XIANG, who generously and tirelessly discussed academic issues with me and greatly benefited me.

I am also incredibly grateful to the China Scholarship Council (CSC) for recognizing my academic potential and providing me with financial support during my Ph.D. studies.

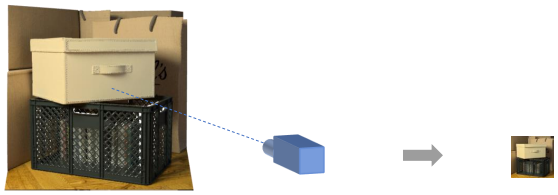
Last but not least, countless thanks are given to my parents for their selfless love and unconditional support. Finally, I would like to express my special thanks to my BaoBao Miss GU. Without her long-term understanding and encouragement, there would be no Dr. Zhengyu ZHANG!

Résumé en français

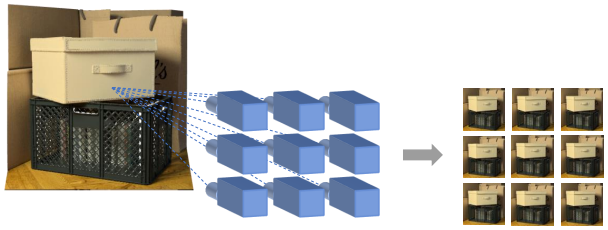
Contexte

Avec la croissance exponentielle des technologies de l'information multimédia et les progrès révolutionnaires de la technologie de communication mobile de cinquième génération, nous sommes désormais immergés dans un monde proposant de plus en plus de contenus multimédia fascinants. Le domaine des médias visuels domine et attire l'attention d'une grande majorité de la population. En tant que support courant d'informations visuelles, les images 2D traditionnelles ne peuvent pas satisfaire le désir d'une une expérience immersive. En effet, elles ne capturent que la distribution planaire de l'intensité des rayons lumineux, tout en ignorant l'impact de son chemin de propagation et de son anisotropie. Heureusement, en collectant l'éclat de tous les rayons lumineux dans l'espace libre et en démultiplexant les informations angulaires perdues dans la photographie 2D traditionnelle, la technologie de champs de lumière (Light Field) offre une expérience visuelle plus réaliste et captivante aux utilisateurs finaux. Cette capacité unique a fait des Light Field Images (LFI) un média émergent très en vogue et recherché. Des illustrations d'une image 2D traditionnelle et d'un LFI avec des vues angulaires simplifiées 3×3 sont fournies dans la Figure 1.

Comme indiqué dans une nouvelle étude de marché complète publiée par MarketsandMarkets, « Marché de champs de lumière par technologie (matériel (solutions d'imagerie, écrans de champs de lumière), logiciels), vertical (médias et divertissement, soins de santé, architecture, industrie, défense) et région (Amérique du Nord, APAC, Europe, et RoW) - Global Forecast to 2026 », le marché des LFI s'est étendu à plus de 30 pays à travers le monde. De nombreuses entreprises notables se sont déjà engagées dans le développement de produits liés aux LFI, comme NVIDIA aux États-Unis, Japan Display au Japon et Holografika en Hongrie. En outre, ce rapport souligne que, sous l'influence de divers facteurs clés, notamment la popularité croissante des stratégies de marketing personnalisées, la demande croissante de solutions d'imagerie médicale avancées et les progrès continus de la technologie des effets visuels utilisée dans les films et les jeux, la taille du marché des LFI devait passer de 76 millions de dollars en 2021 à 154 millions de dollars en 2026,



(a) Image 2D traditionnelle



(b) Image de champs de lumière (LFI)

Figure 1 – Illustrations de (a) Image 2D traditionnelle et (b) Image de champs de lumière (LFI) avec des vues angulaires simplifiées 3×3. Le contenu de l'image provient de [1].

avec un taux de croissance annuel composé (GAGR) de 15.1%.

Actuellement, la technologie de champs de lumière est appliquée dans diverses industries, révolutionnant la façon dont nous interagissons avec le contenu multimedia. Par exemple, en matière de réalité augmentée (AR), la technologie de champs de lumière offre une expérience plus réaliste et immersive en reconstruisant la profondeur, la forme et la texture de la scène, permettant ainsi aux objets virtuels de s'intégrer de manière transparente à l'environnement réel. En outre, la technologie de champs de lumière peut également améliorer la perception spatiale dans les écrans AR, rendant la position et la taille des objets virtuels plus précises et améliorant ainsi le sentiment d'immersion de l'utilisateur. Outre la réalité augmentée, la technologie de champs de lumière présente également un grand potentiel en matière de réalité virtuelle (VR). La technologie VR traditionnelle utilise généralement des images stéréoscopiques ou des vidéos à 360 degrés pour simuler des environnements virtuels, mais ces méthodes ne parviennent souvent pas à fournir une véritable impression de profondeur. Au lieu de cela, la technologie de champs de lumière capture et reproduit les informations relatives aux rayons lumineux dans des environnements virtuels pour percevoir la distance et la position des objets dans la scène virtuelle.

En fait, les scénarios d'application de la technologie de champs de lumière vont bien au-delà de l'amélioration de l'expérience visuelle en AR et VR. Il peut être utilisé dans un large éventail de domaines, notamment les soins de santé professionnels, la conception

architecturale et la formation pédagogique. Dans le domaine des soins de santé professionnels, la technologie de champs de lumière peut améliore le diagnostic des maladies en fournissant des capacités de visualisation qui permettent aux médecins d'analyser et de comprendre avec précision l'état du patient, conduisant ainsi à des plans de traitement plus efficaces. Dans la conception architecturale, la technologie de champs de lumière peut fournit aux concepteurs des modèles virtuels plus réalistes pour mieux prévisualiser et ajuster les méthodes de conception architecturale. Dans la formation pédagogique, la technologie de champs de lumière peut briser les barrières du temps et de l'espace, offrir aux étudiants une expérience d'apprentissage immersive et améliorer encore l'absorption et la compréhension des connaissances.

Certes, les riches informations angulaires enregistrées dans les LFI offrent davantage de possibilités pour leurs applications pratiques dans divers domaines. Cependant, le volume important de données LFI augmente la difficulté des processus d'acquisition, de représentation, de codage, de rendu et d'affichage, et des distorsions peuvent être introduites à chaque étape. L'utilisation de LFI déformés dans des applications pratiques affecte considérablement l'expérience visuelle des utilisateurs, ce qui contredit l'intention initiale de la technologie de champs de lumière. Par conséquent, il est nécessaire d'évaluer avec précision la qualité des LFI déformés afin de fournir un retour d'information, des conseils et une optimisation pour l'ensemble de la chaîne de traitement des LFI. Malheureusement, les mesures d'évaluation objective de la qualité 2D existantes se concentrent généralement sur la dégradation de la qualité des informations spatiales, mais ignorent l'effet de l'incohérence angulaire dans les LFI, conduisant finalement à des performances insatisfaisantes dans l'évaluation de la qualité des LFI déformées. Ces faits constituent la principale motivation de cette thèse.

Cette thèse analyse principalement l'impact des distorsions LFI en fonction des caractéristiques des LFI, dans le but d'effectuer une évaluation de la qualité des LFI (LFIQA: Light Field Image Quality Assessment) de manière objective et subjective. Bien que des efforts remarquables aient été consacrés à la recherche sur l'évaluation de la qualité des LFI, la précision prédictive est encore loin d'être idéale, par exemple, la plupart des métriques objectives existantes ne peuvent atteindre que des performances SRCC de 0.7 sur la base de données Win5-LID [2]. À cette fin, cette thèse apporte plusieurs contributions. La première contribution est la proposition de deux métriques objectives basées sur des caractéristiques ad-hoc extraites par traitement d'image classique pour l'évaluation de la qualité des LFI. Deuxièmement, cette thèse combine les techniques d'apprentissage

profond avec les caractéristiques spécifiques des LFI, aboutissant au développement de quatre mesures objectives d'évaluation de la qualité basées sur l'apprentissage profond. Enfin, constatant le manque de bases de données de référence impliquant de multiples distorsions, une nouvelle base de données d'évaluation de la qualité des LFI est donc proposée pour combler cette lacune.

Contributions de la thèse

Nous résumons brièvement les contributions de la thèse, chapitre par chapitre:

Chapitre 1 - Fondamentaux des LFI. Ce chapitre présente les principes fondamentaux des LFI. Tout d'abord, le principe théorique et le développement de champs de lumière sont présentés en détail. Ensuite, nous résumons la chaîne de traitement des LFI en quatre étapes, comprenant l'acquisition, la représentation, le codage et le rendu/affichage, et introduisons les fonctions et les technologies clés de chaque étape.

Chapitre 2 - Etat de l'art de l'évaluation de la qualité des LFI. Ce chapitre passe en revue les méthodes LFIQA de l'état de l'art sous plusieurs perspectives. Nous analysons d'abord les principaux effets des LFI lorsqu'ils subissent des distorsions, puis détaillons les protocoles expérimentaux subjectifs, les bases de données LFIQA subjectives existantes et les métriques LFIQA objectives existantes. Enfin, la stratégie d'entrainement et de validation et les critères d'évaluation des performances sont brièvement présentés.

Chapitre 3 - Métriques proposées basées sur des caractéristiques artisanales pour l'évaluation de la qualité des LFI. Compte tenu des performances insatisfaisantes et de la complexité temporelle des métriques LFIQA objectives existantes, ce chapitre présente deux nouvelles métriques LFIQA sans référence (NR) basées sur des caractéristiques conçues manuellement (hand-crafted features, appelées ici caractéristiques artisanales). La première métrique SATV-BLiF exploite les variations de texture spatio-angulaires et calcule des caractéristiques statistiques basées sur les histograms de caractéristiques de texture pour l'évaluation de la qualité LFI. La deuxième métrique MPT-MWI-BLiF offre des performances d'évaluation de qualité supérieures en extrayant des informations de texture plus complètes et en prenant en compte les caractéristiques géométriques des sous-bandes et les caractéristiques d'intensité des coefficients dans le domaine 4D-DHWT. Les résultats expérimentaux démontrent l'efficacité supérieure et les performances compétitives de la métrique SATV-BLiF par rapport à la plupart des métriques NR LFIQA existantes. De plus, la métrique MPT-MWI-BLiF atteint systéma-

tiquement de meilleures performances d'évaluation de qualité que la métrique SATV-BLiF et la plupart des métriques de l'état de l'art.

Chapitre 4 - Métriques proposées basées sur l'apprentissage profond pour l'évaluation de la qualité des LFI. Ce chapitre propose une exploration approfondie de la combinaison de la technologie d'apprentissage profond et de l'évaluation de la qualité des LFI. En conséquence, nous proposons quatre métriques LFIQA basées sur l'apprentissage profond, dont trois métriques sans référence (NR) (DeeBLiF, PVBLiF et ASEM-BLiF) et une métrique avec référence (FR) (EDDMF). La métrique DeeBLiF entraîne des modèles profonds et évalue la qualité des LFI déformés via des patchs spatio-angulaires 2D. La métrique PVBLiF étend l'idée basée sur les patchs 2D au domaine 3D et adopte une représentation 3D plus intuitive pour l'évaluation de la qualité LFI. La troisième métrique ASEM-BliF effectue l'évaluation de la qualité en prenant en compte déséquilibre d'amplitude entre les grandes variations spatiales et les faibles variations angulaires des LFI. Enfin, nous étendons la métrique DeeBLiF à la métrique EDDMF, qui fournit une description plus complète des distorsions LFI basée sur une combinaison de caractéristiques profondes à dominante angulaire et de caractéristiques artisanales à dominante spatiale. Les résultats expérimentaux montrent que notre métrique initiale DeeBLiF a de meilleures performances d'évaluation de la qualité que la plupart des métriques NR LFIQA existantes. De plus, avec les versions mises à jour de DeeBLiF, des améliorations significatives des performances peuvent être observées dans PVBLiF, ASEM-BLiF et EDDMF. Toutes les métriques proposées surpassent largement les autres métriques LFIQA de l'éat de l'art.

Chapitre 5 - Base de données proposée pour l'évaluation de la qualité des LFI. Ce chapitre se concentre sur la réalisation d'expériences subjectives pour évaluer la qualité des LFI déformés et propose donc une nouvelle base de données LFIQA standard. Tout d'abord, l'analyse des limites des bases de données existantes et la contribution de la base de données proposée sont discutées, suivies d'une présentation du contenu de la base de données proposée. Ensuite, nous présentons la mise en œuvre détaillée de nos expériences subjectives. Enfin, nous fournissons un benchmark de deux nouvelles métriques objectives NR et de nombreuses métriques objectives LFIQA sur la base de données proposée.

Table of Contents

Introduction **15**
Context . . . 15
Structure of the Thesis . . . 17

1 Fundamentals of Light Field Images **20**
1.1 Light Field Theoretical Principle . . . 20
1.2 Light Field Image Processing Chain . . . 21
1.3 Light Field Image Acquisition . . . 22
1.3.1 Multi-Camera Array . . . 23
1.3.2 Time Sequential Capture System . . . 24
1.3.3 Plenoptic Camera . . . 25
1.3.4 Computer Graphics Software . . . 26
1.4 Light Field Image Representation . . . 27
1.4.1 Sub-Aperture Image Array . . . 28
1.4.2 Pseudo Video Sequence . . . 28
1.4.3 MicroLens Image . . . 29
1.4.4 Other Representations . . . 29
1.5 Light Field Image Coding . . . 30
1.5.1 Traditional Coding Tools . . . 30
1.5.2 Light Field Image Coding Methods . . . 31
1.6 Light Field Image Rendering and Display . . . 32
1.6.1 Viewpoint-Shifting . . . 32
1.6.2 Refocusing . . . 33
1.6.3 Other Methods . . . 34
1.7 Conclusion . . . 35

2 State-of-the-art Light Field Image Quality Assessment Methods **36**
2.1 Light Field Image Distortion Analysis . . . 36
2.1.1 Super-Resolution-Related Distortions . . . 37

2.1.2 Compression-Related Distortions . . . 39
2.1.3 Other Distortions . . . 40
2.2 Subjective Experimental Protocols . . . 42
2.3 Existing Subjective Databases . . . 43
2.4 Existing Objective Metrics . . . 47
2.5 Train-Test Strategy . . . 52
2.6 Performance Evaluation Criteria . . . 53
2.7 Conclusion . . . 54

3 Proposed Handcrafted Feature-based Metrics for Light Field Images 55
3.1 Proposed SATV-BLiF Metric . . . 56
3.1.1 Spatio-Angular Textural Variations . . . 56
3.1.2 Statistical Textural Histogram and Quality Prediction . . . 57
3.2 Proposed MPT-MWI-BLiF Metric . . . 58
3.2.1 Multiplane Textural Variation Measurement . . . 59
3.2.2 Multilevel Wavelet Information Measurement . . . 62
3.2.3 Dimension Reduction and Quality Prediction . . . 65
3.3 Performance Evaluation and Discussion . . . 66
3.3.1 Hyperparameter Dependency . . . 66
3.3.2 Overall Performance Comparison . . . 67
3.3.3 Statistical Significance Test . . . 72
3.3.4 Performance on Cross-Database Validation . . . 72
3.3.5 Performance on Individual Distortion Type . . . 73
3.3.6 Ablation Studies . . . 74
3.4 Conclusion . . . 78

4 Proposed Deep Learning-based Metrics for Light Field Images 80
4.1 Proposed DeeBLiF Metric . . . 80
4.1.1 Spatio-Angular Patch Generation . . . 81
4.1.2 Two-Stream CNN Model . . . 82
4.2 Proposed PVBLiF Metric . . . 83
4.2.1 Pseudo Video Block Sequence Generation . . . 84
4.2.2 Pseudo Video Block Sequence Network . . . 85
4.2.3 Saliency- and Variance-Guided Pooling . . . 89
4.3 Proposed ASEM-BLiF Metric . . . 91

4.3.1 Angular Effect Modeling . . . 93
4.3.2 Angular-Spatial Quality Learning . . . 93
4.3.3 Discriminative Region Selection-Based Auxiliary Learning . . . 95
4.3.4 Dynamic Weighting Loss . . . 97
4.4 Proposed EDDMF Metric . . . 98
4.4.1 Patch Generation Module . . . 99
4.4.2 Hierarchical Discrepancy Network . . . 102
4.4.3 Local Discrepancy Extraction Module . . . 104
4.5 Performance Evaluation and Discussion . . . 106
4.5.1 Hyperparameter Dependency . . . 106
4.5.2 Overall Performance Comparison . . . 108
4.5.3 Statistical Significance Test . . . 112
4.5.4 Performance on Cross-Database Validation . . . 112
4.5.5 Performance on Individual Distortion Type . . . 114
4.5.6 Ablation Studies . . . 119
4.6 Conclusion . . . 122

5 Proposed Light Field Image Quality Assessment Database 124
5.1 Limitation Analysis and Contribution . . . 124
5.1.1 Limitation Analysis of Existing Databases . . . 125
5.1.2 Contribution of the Proposed Database . . . 126
5.2 Database Content . . . 126
5.2.1 Source Sequences . . . 126
5.2.2 Hypothetical Reference Circuits . . . 129
5.3 Subjective Experiment . . . 130
5.3.1 Methodology of Subjective Test . . . 131
5.3.2 Processing of Subjective Results . . . 132
5.3.3 Analysis of Subjective Scores . . . 134
5.4 Objective Measurement . . . 138
5.4.1 Proposed SAB Metric . . . 138
5.4.2 Benchmark of Objective Metrics . . . 143
5.5 Conclusion . . . 148

Conclusions and Perspectives 149
Brief Summary . . . 149

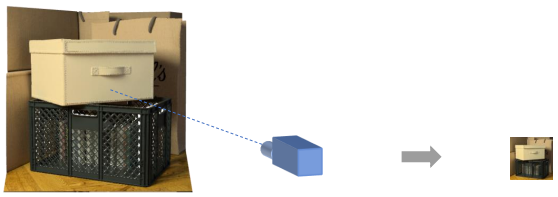
Contributions and List of Publications 149
Perspectives 152

Bibliography **153**

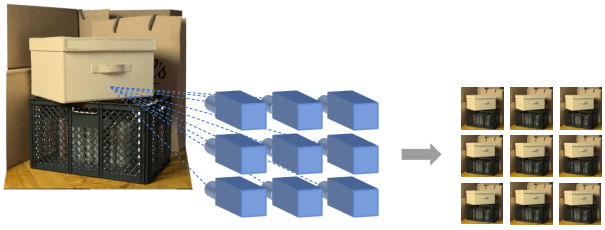
INTRODUCTION

Context

With the exponential growth of multimedia information technology and the revolutionary progress of fifth-generation mobile communication technology, people are now immersed in an ever-expanding world of fascinating media content. The realm of visual media content dominates and attracts the vast majority of people's attention. As a common carrier for conveying visual information, traditional 2D images can no longer satisfy people's yearning for immersive experience. This deficiency arises from their ability only to capture the planar intensity distribution of light rays, while ignoring the effect of its propagation path and anisotropy. Fortunately, by collecting the radiance of all light rays in free space and demultiplexing the angular information lost in traditional 2D photography, light field technology provides a more realistic and captivating visual experience for end-users. This unique capability has made Light Field Image (LFI) a highly hot and sought-after emerging medium. Illustrations of a traditional 2D image and a LFI with simplified 3×3 angular views are provided in Figure 2.



(a) Traditional 2D image



(b) Light field image

Figure 2 – Illustrations of (a) Traditional 2D image and (b) Light field image with simplified 3×3 angular views. The image content is from [1].

As reported in "Light Field Market by Technology (Hardware (Imaging Solutions, Light Field Displays), Software), Vertical (Media & Entertainment, Healthcare, Architecture, Industrial, Defense), and Region (North America, APAC, Europe, and RoW) - Global Forecast to 2026", a new and comprehensive market research published by MarketsandMarkets, the market of light field has spread to over 30 countries worldwide. Numerous notable companies have already engaged in the development of light field-related products, such as NVIDIA from the United States, Japan Display from Japan, and Holografika from Hungary. In addition, this report points out that influenced by various key factors, including the rising popularity of customized marketing strategies, increasing demand for advanced medical imaging solutions, and continuous advancements in visual effects technology utilized in movies and games, the size of light field market is expected to grow from USD 76 million in 2021 to USD 154 million in 2026, with a Compound Annual Growth Rate (GAGR) of 15.1%.

Currently, light field technology is being applied in various industries, revolutionizing how we interact with media content. For example, when it comes to Augmented Reality (AR), light field technology provides a more realistic and immersive experience by reconstructing the scene's depth, shape, and texture, enabling virtual objects to seamlessly integrate with the real environment. In addition, light field technology can also enhance spatial perception in AR displays, making the position and size of virtual objects more accurate and thus enhancing the user's sense of immersion. In addition to AR, light field technology also has great potential in Virtual Reality (VR). Traditional VR technology usually uses stereoscopic images or 360-degree videos to simulate virtual environments, but these methods often fail to provide a real sense of depth. Instead, light field technology captures and reproduces light ray information in virtual environments to perceive the distance and position of objects within the virtual scene.

In fact, the application scenarios of light field technology go far beyond improving the visual experience in AR and VR. It may be utilized in a wide range of fields, including but not limited to professional healthcare, architectural design, and education training. In professional healthcare, light field technology can enhance disease diagnosis by providing visualization capabilities, which enables doctors to accurately analyze and understand the patient's condition, thereby leading to more effective treatment plans. In architectural design, light field technology can provide designers with more realistic virtual models to better preview and adjust architectural design methods. In education training, light field technology can break the barriers of time and space, provide students with an immersive

learning experience, and further enhance knowledge absorption and understanding.

Admittedly, the rich angular information recorded in LFIs brings more possibilities for their practical applications in various fields. However, the substantial volume of LFI data increases the difficulty in the processes of acquisition, representation, coding, rendering, and display, and distortion may be introduced in each stage. The utilization of distorted LFIs in practical applications significantly affects people's visual experience, which contradicts the original intention of light field technology. Therefore, it is necessary to accurately evaluate the quality of distorted LFIs to provide feedback, guidance, and optimization for the entire LFI processing chain. Unfortunately, existing 2D objective quality assessment metrics usually focus on the quality degradation of spatial information, but ignore the effect of angular inconsistency on spatial information in LFIs, ultimately leading to unsatisfactory performance in assessing the quality of distorted LFIs. These facts are the main motivation of this thesis.

This thesis primarily analyzes the impact of LFI distortions based on the characteristics of LFIs, aiming to perform Light Field Image Quality Assessment (LFIQA) objectively and subjectively. Although remarkable efforts have been dedicated to the research of LFI quality evaluation, the predictive accuracy is still far from ideal, *e.g.*, most existing objective metrics can only achieve 0.7 SRCC performance on the Win5-LID database [2]. To this end, this thesis makes several contributions. The first contribution is the proposition of two handcrafted feature-based objective metrics for quality assessment. Secondly, this thesis combines deep learning technologies with the unique characteristics of LFIs, resulting in the development of four deep learning-based objective quality assessment metrics. Finally, considering the lack of benchmark databases that involve multiple distortions, a new quality assessment database for LFIs is thus proposed to fill this gap.

Structure of the Thesis

This thesis is organized as follows:

Chapter 1 - Fundamentals of Light Field Images. This chapter introduces the fundamentals of LFIs. Firstly, the theoretical principle and development of light fields are presented in detail. Then, we summarize the processing chain of LFIs into four stages, including acquisition, representation, coding, and rendering/display, and introduce the functions and key technologies of each stage.

Chapter 2 - State-of-the-art Light Field Image Quality Assessment Meth-

ods. This chapter reviews state-of-the-art LFIQA methods from multiple perspectives. We first analyze the main effects of LFIs when they suffer from distortions, and then detail the subjective experimental protocols, existing subjective LFIQA databases, and existing objective LFIQA metrics. Finally, the train-test strategy and performance evaluation criteria are briefly investigated.

Chapter 3 - Proposed Handcrafted Feature-based Metrics for Light Field Images. Considering the unsatisfactory quality evaluation performance and time complexity of existing objective LFIQA metrics, this chapter presents two novel No-Reference (NR) LFIQA metrics based on handcrafted features. The first metric SATV-BLiF exploits the spatio-angular textural variations and calculates the statistical textural histogram features for LFI quality evaluation. The second metric MPT-MWI-BLiF delivers superior quality assessment performance by extracting more comprehensive texture information and considering the subband geometric features and coefficient intensity features in the 4D-DHWT domain. Experimental results demonstrate the superior efficiency and competitive performance of the SATV-BLiF metric compared to most existing NR LFIQA metrics. Moreover, the MPT-MWI-BLiF metric consistently achieves better quality evaluation performance than the SATV-BLiF metric and most state-of-the-art metrics.

Chapter 4 - Proposed Deep Learning-based Metrics for Light Field Images. This chapter provides an in-depth exploration of the combination of deep learning technology and LFI quality assessment. As a result, we propose four deep learning-based LFIQA metrics, including three NR metrics (DeeBLiF, PVBLiF, and ASEM-BLiF) and one Full-Reference (FR) metric (EDDMF). The DeeBLiF metric trains deep models and evaluates the quality of distorted LFIs via 2D spatio-angular patches. The PVBLiF metric extends the 2D patch-based idea to the 3D domain and adopts a more intuitive 3D representation for LFI quality evaluation. Further, the ASEM-BLiF metric performs LFIQA by addressing the imbalance problem of large spatial variations and subtle angular variations in LFIs. Finally, we extend the DeeBLiF metric into the EDDMF metric, which provides more comprehensive description of LFI distortions based on a combination of angular-dominant deep features and spatial-dominant handcrafted features. Experimental results exhibit that our initial metric DeeBLiF has better quality evaluation performance than most existing NR LFIQA metrics. In addition, as the updated versions of DeeBLiF, significant performance improvements in various aspects can be observed in PVBLiF, ASEM-BLiF, and EDDMF. All the proposed metrics outperform other state-of-the-art LFIQA metrics by a large margin.

Chapter 5 - Proposed Light Field Image Quality Assessment Database. This chapter focuses on conducting subjective experiments to evaluate the quality of distorted LFIs and therefore proposes a novel standard LFIQA database. Firstly, the limitation analysis of existing databases and the contribution of the proposed database are discussed, followed by a presentation of the content of the proposed database. Then, we present the detailed implementation of our subjective experiments. Finally, we provide a benchmark of two novel objective NR metrics and numerous objective LFIQA metrics on the proposed database.

Chapter 1

Fundamentals of Light Field Images

Light rays in our physical world contain abundant information. By capturing and harnessing these information, light fields give rise to a wide range of practical applications with substantial real-world significance. In this chapter, we first describe the theoretical principle of light fields and provide a general processing chain for LFIs. Then, we introduce each part of the LFI processing chain, including acquisition, representation, coding, and rendering/display. Finally, the last section concludes this chapter.

1.1 Light Field Theoretical Principle

To describe all the information of light rays, Gershun [3] pioneeringly proposed the term "light field" in 1936, and assumed that light rays propagate along straight lines in a homogeneous medium, which are the carriers of radiant energy in free space. With the increasing popularity of digital imaging and geometric graphics, Adelson and Bergen [4] further refined the light field concept into a 7-parameter plenoptic function in 1991, to comprehensively describe light rays at any location (x, y, z), to any direction (θ, ϕ), in any wavelength (λ), and at any time (t), which can be expressed as $\mathcal{L}(x, y, z, \theta, \phi, \lambda, t)$.

However, the acquisition and utilization of such high-dimensional data pose significant challenges. To facilitate practical use, two assumptions were made to simplify the plenoptic function from 7 dimensions to 5 dimensions: Firstly, by assuming that the target object is static, the variable t can be excluded. Secondly, by recording each light ray independently in different color channels, *i.e.*, assuming each light ray is monochromatic, the variable λ can be removed. Moreover, in 1996, Levoy and Hanrahan [5] claimed that this 5D function can be further simplified into 4D by assuming the radiance of light rays remain constant along a straight line and thereby introducing a two-plane parameterization to represent light fields, which is illustrated in Figure 1.1.

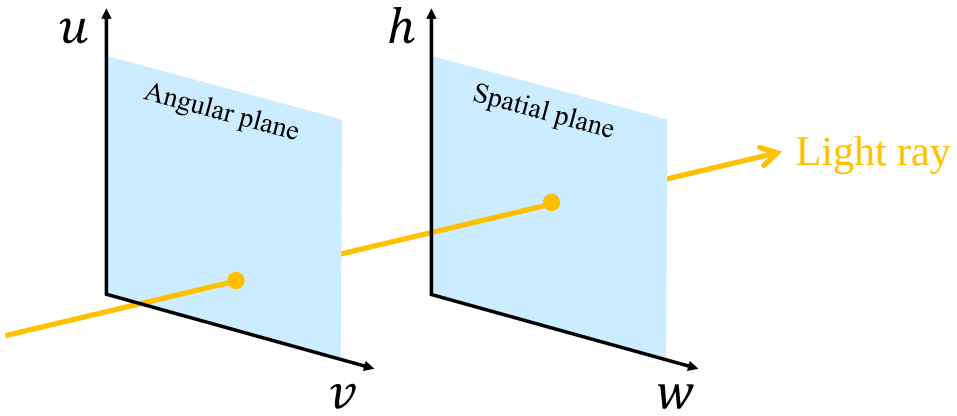


Figure 1.1 – Two-plane parameterization.

In this model, a light ray is assumed to pass through the angular plane and the spatial plane successively. By recording the coordinates of two points on two parallel planes (*i.e.*, (u,v) and (h,w)), the propagation direction of this light ray can be determined. Therefore, a light field is represented as $\mathcal{L}(u, v, h, w)$. Note that this 4D function is commonly used to describe the LFI, but it can also be utilized for light field videos by adding a temporal dimension. Owing to its simplicity and comprehensibility, most subsequent works related to light fields are based on this foundation.

1.2 Light Field Image Processing Chain

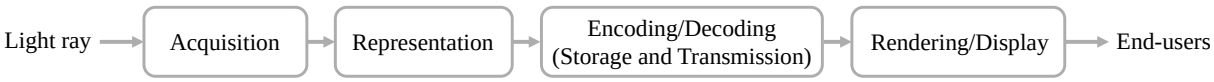


Figure 1.2 – Processing chain of LFIs.

Like any other form of media, light fields undergo numerous stages in their journey from initial acquisition to final presentation to the end-users. This thesis focuses on LFIs and divides the processing chain of LFIs into four key stages [6], [7], including acquisition, representation, coding, and rendering/display, as illustrated in Figure 1.2.

— **Acquisition**. The acquisition stage is to capture/generate LFIs that incorporate both angular and spatial information. This is generally accomplished by employing specialized optical systems or leveraging advanced computational approaches.

Section 1.3 introduces several mainstream LFI acquisition methods in detail.

— **Representation**. The raw collected LFIs typically need to be transformed into various representation formats, mainly determined by the requirements of subsequent stages. The choice of LFI representations has a profound influence on the efficiency and outcomes of the entire processing chain. Section 1.4 presents a brief overview of commonly-used LFI representations.

— **Coding**. As a typical high-dimensional data, LFI's storage and transmission impose stringent demands on data encoding/decoding due to resource limitations. The main objective at this stage is to compress LFI data for efficient storage and transmission. Section 1.5 provides a detailed description of two main coding methodologies.

— **Rendering/Display**. Since the current application scope of light field displays is quite limited, rendering technology is essential for displaying LFIs on traditional 2D/3D displays. The choice of LFI rendering methods is closely related to various user-side factors, such as intended usage scenarios, user interactivity, and display devices. Section 1.6 offers an in-depth exposition on LFI rendering and display.

In summary, the journey of LFIs from acquisition to display is a complex and multifaceted process. Each stage is crucial and interdependent, and the visual experience delivered to the end-users is the result of collective efforts across all stages. Therefore, ensuring the perceptual quality of LFIs at each stage elevates the final visual effect, paving the way for a more immersive and engaging experience.

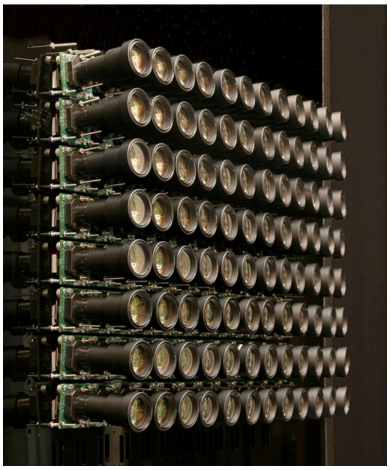
1.3 Light Field Image Acquisition

To capture a scene, a conventional camera can only record the intensity of light rays, while neglecting the direction information. In contrast, LFIs provide a more comprehensive representation of the visual world, which cannot be obtained using a single conventional camera at a single viewing perspective. According to the definition of two-plane parameterization, the two angular dimensions in LFIs inherently differentiate LFIs from 2D images, which has prompted researchers to consider how to capture additional angular information, as a crucial step towards obtaining LFIs. At present, the acquisition of LFIs is primarily achieved through four distinct approaches: multi-camera array, time sequential capture system, plenoptic camera, and computer graphics software.

1.3.1 Multi-Camera Array

The multi-camera array is a direct implementation of the two-plane parameterization model, which has been widely used for capturing LFIs in the early days. As the name suggests, this acquisition method captures various perspectives of the same scene simultaneously through multiple conventional cameras arranged on a planar or spherical surface, thereby recording the directionality of light rays. As a result, the camera number determines the angular dimensions U and V, while the sensor number within each camera determines the spatial dimensions H and W. For each shot, a LFI in the form of a 2D image array is obtained. Figure 1.3 shows two examples of multi-camera arrays for LFI acquisition, in which (a) is a planar multi-camera array from Stanford [8] and (b) is a spherical multi-camera array from Google [9].

Multi-camera arrays offer two advantages in capturing light field content: 1) Cameras can be added or removed at any time according to specific needs to obtain LFIs with different angular resolutions. 2) LFIs with high spatial resolution can be achieved using a single high-resolution camera. Correspondingly, this acquisition method also faces some challenges in practical use: 1) The camera array is generally bulky and cumbersome and may not be practical for some applications. 2) The angular resolution is directly determined by the camera number, which makes obtaining LFIs with high angular resolution very expensive. 3) Ensuring perfect synchronization of shutter timing and data acquisition for all cameras in the array is challenging in practice.



(a) Planar multi-camera array



(b) Spherical multi-camera array

Figure 1.3 – Multi-camera arrays for LFI acquisition. (a) A planar multi-camera array from Stanford [8]; (b) A spherical multi-camera array from Google [9].

1.3.2 Time Sequential Capture System

To alleviate the cost issue of multi-camera arrays, the time sequential capture system is proposed to capture a series of 2D images at consecutive times using a conventional camera. Typically, this is achieved by placing a camera on a computer-controlled moving robotic arm or gantry. After shooting continuously at each location, the angular dimensions U and V of LFIs can be rearranged according to the temporal order of 2D images. Figure 1.4 exhibits two example devices of time sequential capture systems for LFI acquisition, in which (a) is a 3D time sequential capture system proposed by Kim *et al.* [10] and (b) is a 4D time sequential capture system from Stanford [11].

However, despite their portability and cost-effectiveness, time sequential capture systems present certain limitations: 1) This method is only suitable for static scenes, as capturing dynamic scenes will introduce motion artifacts. 2) The quality of the captured LFIs may be affected by some environmental factors, such as lighting changes during camera movement.



(a) 3D time sequential capture system

(b) 4D time sequential capture system

Figure 1.4 – Time sequential capture systems for LFI acquisition. (a) A 3D time sequential capture system proposed by Kim *et al.* [10]; (b) A 4D time sequential capture system from Stanford [11].

1.3.3 Plenoptic Camera

Generally, the aforementioned two devices based on conventional cameras can only be found in companies and research institutions for advanced imaging projects. Fortunately, the emergence of plenoptic cameras [12] (also known as light field cameras) has successfully commercialized light field technology and made it accessible to individual users due to their compact design and affordability. Theoretically, a plenoptic camera consists of a main lens, a microlens array, and a series of photosensors. Based on the position of the microlens array, currently there are two types of plenoptic cameras available on the market: unfocused and focused.

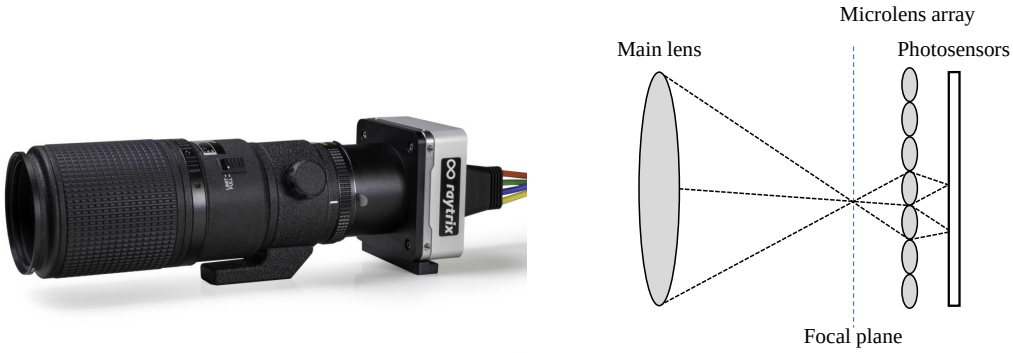


(a) Lytro camera (b) Lytro Illum camera (c) Design schematic

Figure 1.5 – Unfocused plenoptic cameras [13]. (a) Lytro camera; (b) Lytro Illum camera; (c) Design schematic.

Unfocused plenoptic cameras. In 2012, Lytro, Inc. debuted the first commercial plenoptic camera, called the Lytro camera. Then in 2014, they launched the second generation product, known as the Lytro Illum camera. The appearances of these two cameras are shown in Figure 1.5 (a) and Figure 1.5 (b), respectively. Both cameras use the unfocused imaging method, whose design schematic is illustrated in Figure 1.5 (c). In this optical model, since the microlens array is placed on the focal plane of the main lens, light rays passing through the main lens will create a sharp image at the microlens array. Then each microlens in the array defocuses the converged light and delivers it to the photosensors. The number of light rays received by each microlens in the array determines the angular dimensions U and V of LFIs, while the number of microlens in the array determines the spatial dimensions H and W of LFIs.

Focused plenoptic cameras. The resolution of unfocused plenoptic cameras is limited by hardware and cannot be changed. This inherent limitation prevents it from adapt-



(a) Raytrix camera (b) Design schematic

Figure 1.6 – Focused plenoptic cameras [14]. (a) Raytrix camera; (b) Design schematic.

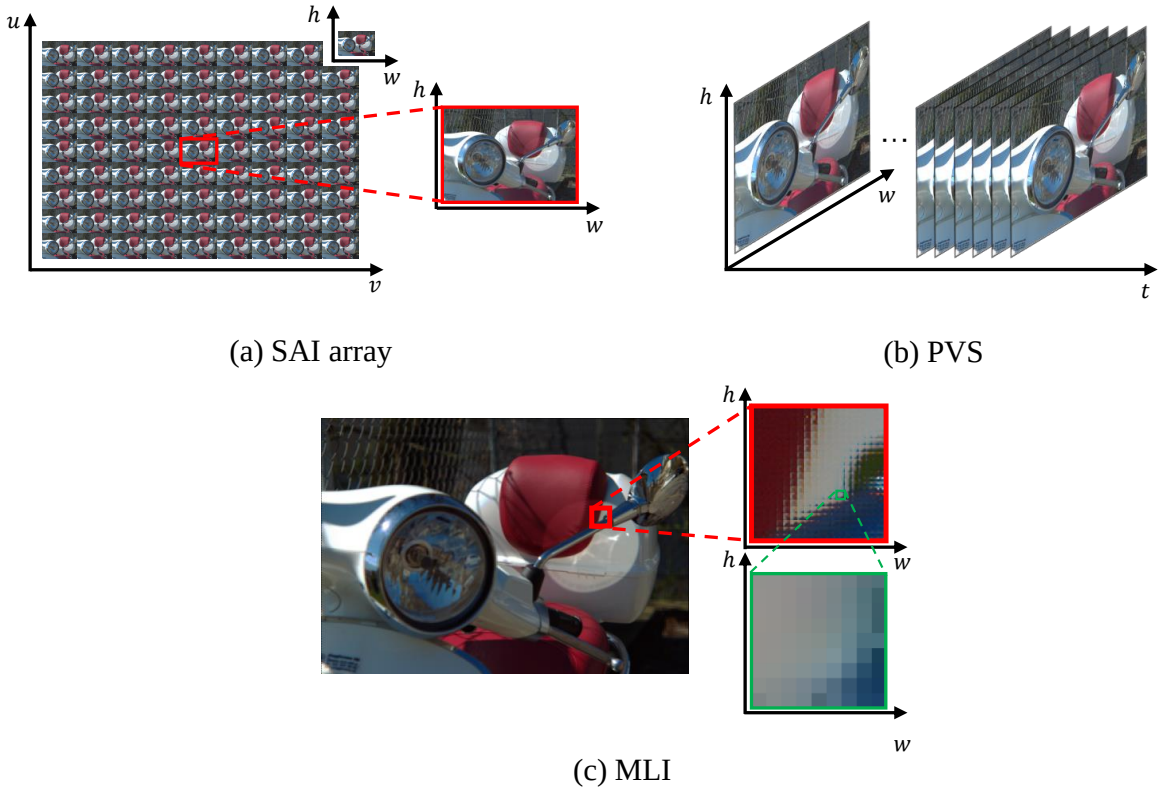
ing to ever-changing application requirements. In 2013, Raytrix GmbH [14] introduced the Raytrix camera by repositioning the microlens array in the design of unfocused plenoptic cameras. Figure 1.6 shows its appearance and design schematic. Compared to the unfocused imaging model, the microlens array in the focused imaging model is placed farther away from the main lens, which allows the microlens array to focus on the image plane. With this new design, the microlens array can capture multiple spatial points and therefore increase the spatial resolution. However, this comes with a trade-off, where an increase in spatial resolution leads to a decrease in angular resolution. Through pre-estimated scene depth and sophisticated processing algorithms, Raytrix cameras enable flexible conversion between spatial and angular resolutions.

1.3.4 Computer Graphics Software

Another popular way for generating LFIs is to use computer graphics software, such as Blender [15] and Unreal engine [16]. These software use computer graphics rendering technology to create a small real-world scene and simulate the process of capturing light rays with a plenoptic camera. Although this method cannot perfectly replicate real-world lighting, the generated LFIs are lifelike and convincing enough in most cases. Moreover, this method effectively circumvents various possible problems when capturing with a real plenoptic camera, such as hardware degradation, camera shake, and insufficient lighting conditions. As the performance of rendering and light-tracing technology improves, LFIs generated by computer graphics software are expected to get increasingly close to those from photography.

1.4 Light Field Image Representation

As mentioned in Section 1.1, the LFI can be simplified to a biplane model with four parameters. Nonetheless, visualizing this 4D format remains a challenge. Fortunately, by observing the LFI data with a subset of dimensions, the LFI can be transformed into various representations for better visualization and understanding [17]. Since all representations are derived from the same LFI data, some representations can be converted into one another, but some conversions are irreversible. Throughout the processing chain, LFIs may undergo multiple transformations, primarily based on the requirements of each stage. Figure 1.7 shows the three main representations of an example LFI, including Sub-Aperture Image array (SAI array), Pseudo Video Sequence (PVS), and MicroLens Image (MLI), which will be detailed in the following subsections. Finally, an overview of other representations will be briefly presented in the last subsection.



(a) SAI array

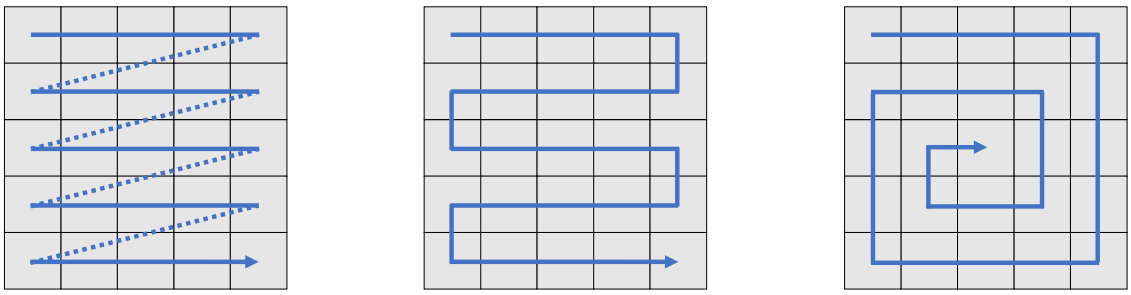
(b) PVS

(c) MLI

Figure 1.7 – Three main LFI representations. The example LFI used is from [2]. (a) SAI array; (b) PVS; (c) MLI.

1.4.1 Sub-Aperture Image Array

In the two-plane parameterization $\mathcal{L}(u, v, h, w)$, we can take the uv plane as a set of camera views and the hw plane as their image plane. By observing the image plane from a fixed view on the camera plane, a 2D pinhole image can be obtained, which is also known as a Sub-Aperture Image (SAI). After combining the SAIs of all camera views, a LFI can be represented in a SAI array, as shown in Figure 1.7 (a). In this representation, each SAI has the same resolution and records the spatial information, while the difference between individual SAIs records the angular information. The SAI array is one of the most widely-used representations in the LFI processing chain due to its clear physical meaning and distinction between angular and spatial information. For example, in the acquisition stage, LFIs collected using a multi-camera array will be represented in the form of SAI array. In the coding stage, compression algorithms are particularly effective for the SAI array representation because there is a large amount of redundant information between SAIs.



(a) Raster (b) Serpentine (c) Spiral

Figure 1.8 – Three common orders for generating PVS. (a) Raster; (b) Serpentine; (c) Spiral.

1.4.2 Pseudo Video Sequence

Pseudo Video Sequence (PVS) is a 3D LFI representation, created by arranging each SAI as a frame and displaying all SAIs in a certain order, as shown in Figure 1.7 (b). It could be observed that the two angular dimensions of LFI are reorganized into the temporal dimension of PVS. In this process, the choice of the generation order will affect the temporal information in PVS, which is directly related to the performance of subsequent encoding algorithms. Figure 1.8 presents three common orders for generating

PVS, including raster, serpentine, and spiral. Although the LFI angular information is disrupted in the PVS representation, this disruption can be easily restored in later stages. In return, there are several significant advantages to representing LFIs in the form of PVS. From an acquisition perspective, LFIs captured by a time sequential capture system will be directly represented in the PVS form. From a coding perspective, this representation enables many off-the-shelf solutions designed for videos to be available for LFIs. From a display perspective, compared to using expensive light field displays, it is very economical and accessible for end-users to use a 2D/3D display to experience light field content in the PVS representation.

1.4.3 MicroLens Image

By collecting all light rays from different viewpoints of the uv plane approaching the hw plane, a LFI can be represented as a 2D image with very high spatial resolution. This representation is called a MicroLens Image (MLI), also known as a lenslet image. As shown in Figure 1.7 (c), a MLI consists of a grid of microlens, and each microlens records the angular information corresponding to a spatial pixel. The MLI representation is the raw data acquired from plenoptic cameras, but the characteristics of each microlens in the MLI vary depending on the imaging model. In the unfocused imaging model, the microlens array focuses at the main lens and thus each microlens records an image from the back of the main lens. In contrast, in the focused imaging model, the microlens array focuses at the image plane of the main lens, so each microlens captures a portion of the image plane. Compared to the SAI array and PVS, the MLI representation offers a more intuitive presentation of angular information, thereby further providing depth information and disparity of a scene.

1.4.4 Other Representations

Some other representations interpret LFIs from different perspectives [18]. These alternative representations are tailored to address the unique challenges and requirements of specific areas related to LFIs:

Epipolar Plane Image (EPI). By vertically stacking the rows h from SAIs in a fixed angular row u, one can obtain a 2D horizontal EPI. A 2D vertical EPI can be obtained in a similar manner. Although an EPI only contains pixel vectors from a single row or column angular dimension, the slope of lines within the EPI effectively conveys

depth information. Therefore, the EPI representation is widely employed in the field of LFI depth estimation.

Refocused Image (RI). A 2D refocused image containing focused and defocused regions is generated by superimposing multiviews with a specific slope related to the disparity. Leveraging the multiview properties of LFIs to refocus objects at different depths within an image is a fascinating application, which enables dynamic and immersive visual experience.

Geometry-related image. The geometry of LFIs can be a depth map, disparity map, saliency map, or other synthetic images. These representations reflect the LFI information to some extent and enable some LFI applications, such as view synthesis and saliency detection.

1.5 Light Field Image Coding

The traditional coding stage typically converts data into a specific format or representation for efficient storage and transmission [6]. Considering that LFIs contain a large amount of redundant information, the coding stage of LFIs focuses on compressing high-dimensional LFI data while minimizing the sacrifice of image quality. In this section, traditional coding tools as well as coding methods specifically designed for LFIs will be introduced.

1.5.1 Traditional Coding Tools

As mentioned before, based on the assumption of two-plane parameterization, a 4D LFI can be described in different 2D/3D representations. Therefore, LFIs could be processed using off-the-shelf coding tools during the coding stage.

In the earliest light field coding schemes, JPEG [19] and JPEG2000 [20] standards were commonly used as the coding tools for each viewpoint within the SAI array representation of LFIs. However, treating each viewpoint as independent data inevitably ignores the angular information of LFIs. To address this problem, some LFI coding schemes encode the MLI representation of LFIs using the same JPEG and JPEG2000 standards. Despite this, the coding efficiency remains unsatisfactory because the MLI representation contains mixed spatial and angular information that these standards cannot handle well. In addition to using 2D image codecs, many off-the-shelf 2D video codecs, such as stat-of-the-art

H.264 (AVC) [21] and H.265 (HEVC) [22], and the future video coding standard Versatile Video Coding (VVC) [23], are also applicable since LFIs can be represented in a video-like PVS form. Afterwards, some coding schemes further take inter-view information between adjacent frames into video coding, such as MultiView-HEVC (MV-HEVC) [24]. The disadvantage of these PVS-based coding methods is that they can only focus on the temporal information in the PVS rather than the angular information in the LFI.

1.5.2 Light Field Image Coding Methods

Both international standardization groups JPEG Pleno [25] and MPEG-I Visual [26], which work on the LFI representation and compression standard frameworks respectively, have called for research on novel methods in LFI coding. As a result, significant efforts have been invested in the field of LFI coding, and thus a large number of LFI coding methods have been proposed. Currently, advanced LFI coding methods can be generally divided into two types [6]: 4D transform-based and super-resolution-based.

The 4D transform-based LFI coding methods are generally extended from the naive 2D image/video coding tools by additionally considering the two angular dimensions of LFIs. For example, 2D transform-based coding methods employed for 2D images, such as 2D Discrete Cosine Transform (DCT) [27] in JPEG and 2D Discrete Wavelet Transform (DWT) [28] in JPEG2000, are extended to 3D/4D space for high-dimensional LFI coding. The 4D transform-based LFI coding methods can decompose LFIs into multiple subbands of different frequencies along four dimensions. This decomposition allows for separate encoding and even removal of some subbands, which achieves more fine-grained compression strategies. However, this method aims to disentangle the relationships between dimensions in high-dimensional LFI data, and thus performs better for LFIs containing high redundancy, such as LFIs with narrow disparity captured by plenoptic cameras. In contrast, its performance drops dramatically when applied to sparse LFI data.

The aforementioned LFI coding methods are all based on or extend from standard codecs. Another class of methods for LFI coding relies on super-resolution techniques. Note that such a super-resolution-based coding method can be applied to the angular domain as well as the spatial domain of LFIs. Initially, super-resolution methods were used to reduce the cost of acquiring high-resolution LFIs. Subsequently, they were used as a coding method in the coding stage, especially angular super-resolution (also known as angular reconstruction), which is used to synthetic missing viewpoints of angularly downsampled LFIs. Specifically, the LFI data is downsampled in the encoding phase,

followed by restoring it to the original resolution with angular reconstruction methods in the decoding phase. The feasibility of these coding methods based on angular reconstruction mainly stems from the fact that LFIs generally have narrow disparities between angular viewpoints, leading to significant redundancy between perspectives. This characteristic also enables some simple generative methods, such as linear interpolation and nearest neighbor interpolation, to be used for LFI coding. However, although traditional interpolation methods can effectively reduce the amount of LFI data, the quality of the restored LFIs is often far from ideal because they rely on nearby viewpoints to generate new perspectives and do not fully utilize the characteristics of LFIs. More recently, several state-of-the-art LFI coding methods based on angular reconstruction were proposed, which exploit global geometric information to further improve the quality of the generated novel viewpoints.

1.6 Light Field Image Rendering and Display

Although LFIs were originally designed for light field displays, which can directly display LFI data and deliver an immersive visual experience, the current coverage of light field displays is very limited. Therefore, before presenting LFIs to the end-users, a LFI rendering stage is required to convert LFI content into an appropriate format for visualization on conventional 2D/3D displays [29]. Thanks to the rich LFI information, several rendering methods open new degrees of freedom for visualizing LFI content on conventional displays, simulating the visual experience on light field displays. In this section, two widely-used visualization technologies, viewpoint-shifting and refocusing, will be described, followed by a brief overview of other rendering methods.

1.6.1 Viewpoint-Shifting

LFIs provide an immersive experience for end-users by observing the same scene from multiple viewpoints [6]. Thus, to visualize LFIs on a 2D display, viewpoint-shifting technology is an ideal solution, which allows the end-users to interactively explore different viewpoints of a scene by adjusting their viewing position. The subtle parallax created by viewpoint switching can provide a sense of depth that is impossible to achieve with traditional 2D images. Figure 1.9 presents an illustration of the viewpoint-shifting technology for LFI visualization. Note that this visualization can switch not only to the inherent

viewpoint, but also to a new viewpoint synthesized by other new methods such as Nerf and Gaussian splatting.

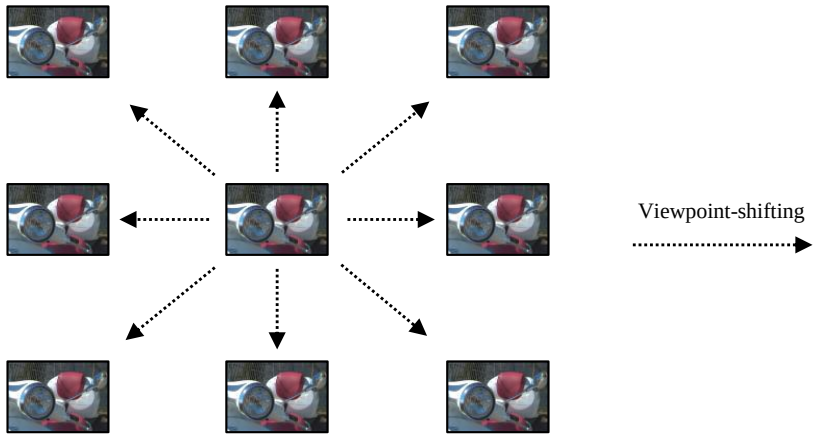


Figure 1.9 – Viewpoint-shifting technology for LFI visualization. The example LFI used is from [2].

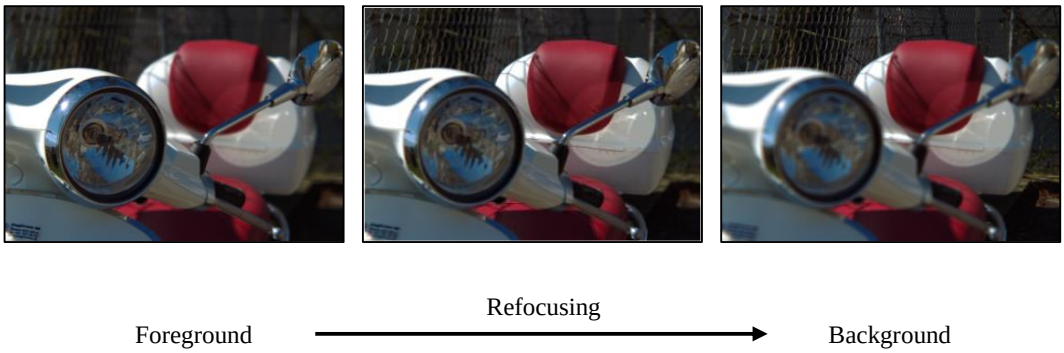


Figure 1.10 – LFI refocusing from foreground to background. The example LFI used is from [2].

1.6.2 Refocusing

Refocusing is an important technology in LFI rendering [30]. Different from traditional images where the focus is fixed during the acquisition process, LFIs can change the position of focus to achieve refocusing in post-processing. Given a target focus plane, a RI is typically generated by moving all viewpoints toward the target viewpoint in equal proportions based on the disparity and averaging the pixel-level results. This process is

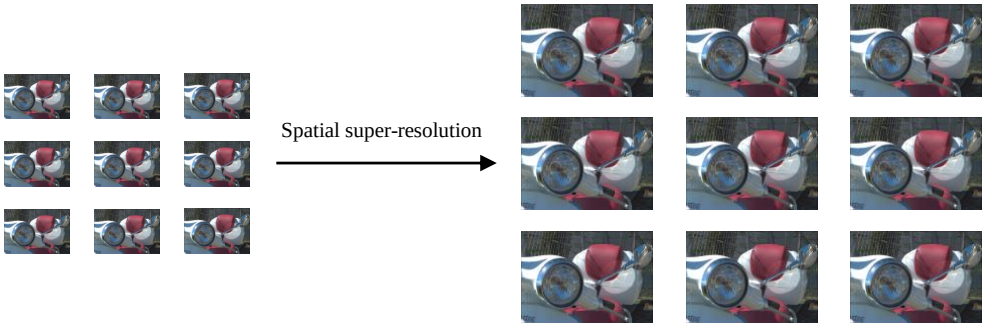


Figure 1.11 – LFI spatial super-resolution. The angular resolution of the example LFI is set to 3×3 for better visualization. The example LFI used is from [2].

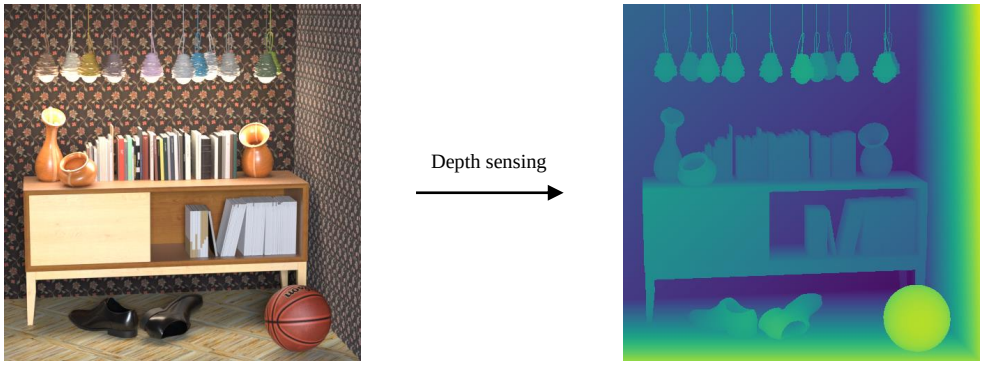


Figure 1.12 – LFI depth sensing. Only the central view of the example LFI is shown for better visualization. The example LFI used is from [1].

known as the shift-and-sum algorithm. An illustration of LFI refocusing from foreground to background is shown in Figure 1.10. It can be found that images from left to right represent the focus from the foreground (electric vehicle headlight) to the background (electric vehicle trunk). With this rendering technology, end-users are given great freedom to view different objects focused on different focal planes on a 2D display. Additionally, it is also possible to use multiple focus planes to generate an all-in-focus image.

1.6.3 Other Methods

In addition to the aforementioned technologies for LFI rendering and display, there are still many other technologies worth mentioning. These technologies can be employed at the same time to enhance the overall performance and experience.

Spatial super-resolution. Spatial super-resolution technology [31], [32] can be used in LFI rendering to restore and enhance details in each viewpoint, especially edges and

texture-rich regions. This allows users to zoom in and out to get a closer observation of image details without compromising resolution. An illustration of LFI spatial super-resolution is shown in Figure 1.11, where the angular resolution of the example LFI is set to 3×3 for better visualization.

Depth sensing. With the rich disparity information in LFIs, it is feasible to obtain the absolute distance from the shooting position to the scene using depth sensing technology [33], [34]. The depth information of a scene can further improve visual reality and experience for users. An illustration of depth sensing is shown in Figure 1.12, where only the central view of the example LFI is shown for better visualization.

Foreground de-occlusion. The multiview information in LFIs enables foreground de-occlusion by observing the occluded objects from different viewpoints [35], [36]. By combining depth-of-field control with foreground de-occlusion, users are provided with stronger interactivity, enabling them to experience image content in depth.

1.7 Conclusion

In this chapter, we present an overview of the fundamentals of LFIs, starting with the theoretical principle and development of light fields. We highlight that the two-plane parameterization for describing LFIs lays the foundation for subsequent studies and applications. Afterwards, we roughly divide the processing chain of LFIs into four stages, including acquisition, representation, coding, and rendering/display, and introduce the functions and key technologies of each stage. Specifically, for the acquisition stage, we introduce the advantages and shortcomings of four methods to obtain LFIs. Moving on to the representation stage, we provide a relatively comprehensive introduction to LFI representations and analyze their main characteristics and usage scenarios. Then, we introduce the coding stage for high-dimensional LFI data, where traditional coding tools and advanced LFI coding methods are described in detail. Finally, several widely-used technologies for visualizing LFIs on conventional displays are presented.

Chapter 2

State-of-the-art Light Field Image Quality Assessment Methods

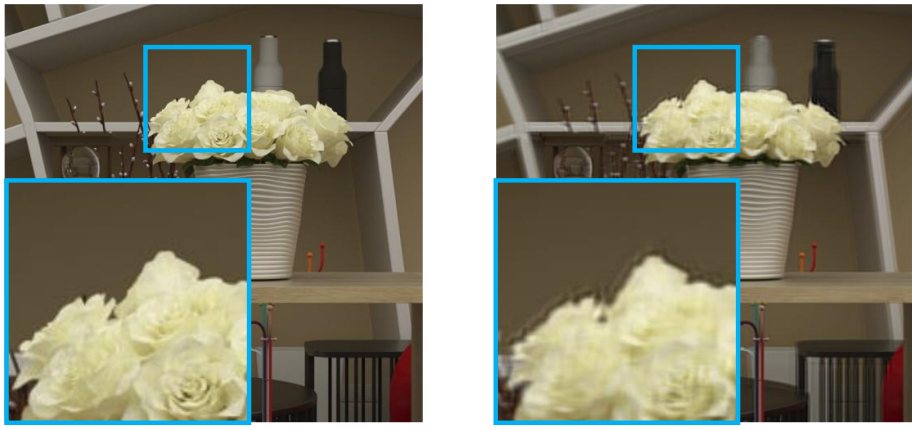
As described in the previous chapter, LFIs undergo a series of processing before being displayed to observers. In these processes, various distortions will unavoidably be introduced, which result in LFI quality degradation and ultimately affect the end-user's visual experience. Fortunately, a well-performing LFI quality monitoring indicator, known as LFIQA, is able to ensure a pleasing visual experience for observers. Besides, LFIQA is also crucial to provide reliable guidance for generative methods such as angular reconstruction and spatial super-resolution. In this chapter, we provide a comprehensive review of state-of-the-art LFIQA methods. To begin with, we provide an in-depth analysis of LFI distortions that may be suffered throughout the processing chain. Then, several common protocols for conducting subjective LFIQA experiments are described, followed by a detailed introduction of existing subjective LFIQA databases and objective LFIQA metrics. Subsequently, we present the train-test strategy and performance evaluation criteria, which are used to verify the effectiveness of objective metrics in this thesis. Finally, the conclusion of this chapter is provided.

2.1 Light Field Image Distortion Analysis

The visual effect of distorted LFIs is quite complicated and hard to evaluate due to the rich information in LFIs. However, a proper summary and analysis of these distortions can help understand how to accurately evaluate the visual perceptual quality of distorted LFIs. In this section, we roughly classify LFI distortions into three categories according to their causes and properties: super-resolution-related distortions, compression-related distortions, and other distortions.

2.1.1 Super-Resolution-Related Distortions

Super-resolution technology, including angular reconstruction and spatial super-resolution, is not only of great significance in LFI acquisition, but also crucial in LFI coding. On the one hand, this technology can effectively improve the angular or spatial resolution and obtain finer-grained information. On the other hand, its adoption may also lead to visual quality degradation in certain areas, potentially affecting the viewing experience.



(a) Reference (b) Ghosting effect

Figure 2.1 – Example of synthetic viewpoint ghosting effects in the narrow-disparity LFI. There exists a ghosting effect around the “flowers”.



(a) Reference (b) Ghosting effect

Figure 2.2 – Example of synthetic viewpoint ghosting effects in the large-disparity LFI. The ghosting effect of “clock” and “flowers” is quite serious.

Angular reconstruction is utilized in sparse LFIs to generate novel viewpoints, thereby transforming them into dense LFIs. Thus, only the synthetic viewpoints suffer from distor-



(a) Reference (b) Blurring effect

Figure 2.3 – Example of blurring effects in a single viewpoint. The boundaries of "person" are blurred. The image content is from [37].



(a) Reference (b) Detail loss

Figure 2.4 – Example of loss of detail in a single viewpoint. Many details in the "grass" are lost.

tions. In addition to the spatial distortion of synthetic viewpoints, angular inconsistency will further affect the visual experience when switching between multiple viewpoints. These distortions are summarized below.

— The ghosting effect is the main distortion caused by angular reconstruction, arising from the angular disparity between different viewpoints. The disparity leads to inaccurate object positions and ghost-like artifacts in the synthetic viewpoint. The severity of ghosting effects is influenced by the magnitude of the angular disparity. Examples of synthetic viewpoint ghosting effects in narrow-disparity and large-disparity LFIs are provided in Figure 2.1 and Figure 2.2, respectively. It can be easily seen that the synthetic viewpoint ghosting effect in the large-disparity LFI is much more obvious than that in the narrow-disparity LFI.

— The geometric distortion may be introduced by numerical rounding operations of pixel positions and inaccuracies in the depth map used for angular reconstruction. Although slight geometric distortion may be inconspicuous, it can also affect the

visual experience if this distortion accumulates over a large number of objects.

— The flickering effect may occur when switching between undistorted and distorted angular viewpoints, leading to visual discontinuity and discomfort. It should be noted that this phenomenon is particularly pronounced when the ghosting effect is present in synthetic viewpoints.

Spatial super-resolution aims to enlarge the spatial resolution of each viewpoint in LFIs from a lower resolution to a higher resolution. In this case, since all viewpoints are subject to similar distortions, the visual perception quality degradation is mainly determined by spatial distortions, as summarized below.

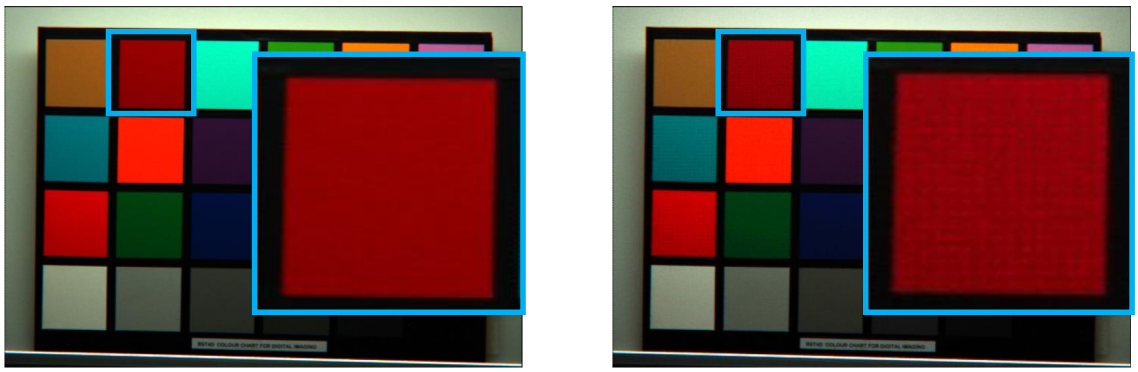
— The blurring effect is commonly observed in viewpoints after spatial super-resolution. Due to the lack of original details, spatial super-resolution algorithms fail to recover high-frequency details from low-resolution images, leading to blurring or lack of sharpness when zooming in. Figure 2.3 provides an example of blurring effects in a single viewpoint.

— The loss of detail may appear when spatial super-resolution algorithms adopt excessive smoothing operations to suppress noise and artifacts. Therefore, it is important to find a balance between noise reduction and preserving high-frequency details in the image. Figure 2.4 shows an example of loss of detail in a single viewpoint.

— The noise in the low-resolution image may be amplified during the spatial super-resolution process. The amplified noise in super-resolved viewpoints may attract the attention of observers and affect the visual experience. Note that noise amplification is more likely to occur in dark areas, where there is not enough information to be exploited.

2.1.2 Compression-Related Distortions

The lossy compression algorithm in the coding stage significantly reduces the amount of LFI data, which plays a crucial role in the transmission and storage stages. Nonetheless, this process comes with a trade-off as it necessarily compromises the data fidelity, which is typically measured by the compression ratio. The quality degradation caused by compressing LFIs at a low compression ratio may not be noticeable to human eyes, but some distortions may be observed as the compression ratio becomes higher. In addition, the type of distortion may vary depending on the compression algorithm. While all distortion types observed in 2D images can also be observed in LFIs, three representative distortions are described as follows.



(a) Reference (b) Blocking effect

Figure 2.5 – Example of blocking effects in a single viewpoint. The blocking effect can be observed in uniform areas. The image content is from [38].

— The blocking effect is typically caused by transform-based compression algorithms such as JPEG, AVC, and HEVC. These algorithms divide each viewpoint (for 2D image codecs) or video-like sequence (for 2D video codecs) of LFIs into blocks for encoding and quantization, the blocking effect may occur because some high-frequency details are discarded. Figure 2.5 provides an example of blocking effects in a single viewpoint. Moreover, the additional impact of angular discontinuity on visual perception becomes more significant at a high compression rate.

— The motion blur is often induced in compression algorithms that focus on reducing the temporal redundancy of LFIs in the PVS representation. In this case, differences between adjacent viewpoints may be compressed, resulting in unnatural transition effects as the viewpoint moves.

— The color distortion may be observed after multiple compression processes. The main reason is that compression algorithms usually compress different color channels of the image separately without considering their interdependence, resulting in a loss of color fidelity. For example, JPEG tends to turn images green as it converts colors into luminance and chrominance, while the latter is typically encoded at a lower rate.

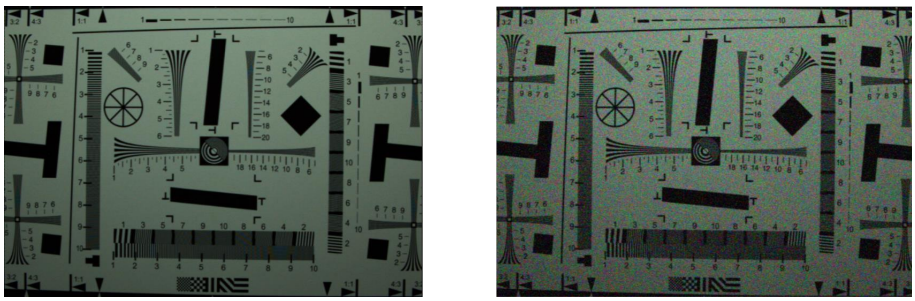
2.1.3 Other Distortions

Common distortions involved in the traditional image processing chain may also occur in the LFI processing chain. These distortions are caused by a variety of factors throughout the whole processing chain, such as acquisition equipment aging, improper interme-



(a) Reference (b) Gaussian noise

Figure 2.6 – Example of Gaussian noise in a single viewpoint. The Gaussian noise is evenly distributed in the image. The image content is from [39].



(a) Reference (b) White noise

Figure 2.7 – Example of white noise in a single viewpoint. There are some uniform bright spots in the image. The image content is from [38].

diate processing, and display hardware malfunctions. For traditional images, the current technology is mature enough and rarely produces such distortions that are perceptible to human eyes. However, for LFIs with additional angular information, these distortions may be magnified, especially when repeatedly switching to view different angular viewpoints, which may ultimately lead to visual discomfort. Here, two representative distortions are introduced below.

— The Gaussian noise is a common distortion that may be introduced in the display stage, as shown in Figure 2.6. As it contains a wide range of random noise, smoothing, and structural distortions, the Gaussian noise can be easily perceived by human eyes and lead to visual quality degradation.

— The white noise can be found during the acquisition process, which usually results

in uniform bright spots in the image. These bright spots will make people feel dazzling and difficult to ignore. Figure 2.7 gives an example of white noise in a single viewpoint.

2.2 Subjective Experimental Protocols

Currently, the research of LFIQA can be classified into two categories: subjective experiments and objective metrics. Among them, subjective experiments directly collect the human ratings for each distorted LFI and are therefore considered the most reliable way to obtain the quantitative perceptual quality. However, to ensure the accuracy of subjective experiments, it is essential to provide observers with a reasonable experimental protocol. In this section, we introduce the protocols in subjective LFIQA experiments from three aspects: visualization method, interaction mode, and test method.

Considering that LFIs contain both spatial and angular information, there are various ways to visualize LFIs, and these visualization methods are always relevant to the interaction mode. Considering that the visualization method and interaction mode can be combined to use together, we summarize them as follows.

— The PVS representation consists of a series of viewpoints arranged in a specific order, where individual viewpoints contain the spatial distortion and viewpoint differences reflect the angular distortion. Therefore, it is widely served as the passive mode of LFI visualization in subjective experiments.

— Viewpoint-shifting is a typical interactive mode to visualize LFIs. Similar to the PVS visualization, spatial distortion is intuitively presented in each viewpoint. Besides, observers can perceive the angular distortion by spontaneously changing their viewing perspective.

— Refocusing is a unique rendering method in LFI post-processing, which can also be used as a visualization method in subjective experiments. Through refocusing, spatial and angular distortions are incorporated into a series of RIs with different focus planes. Both interactive and passive modes are available in this visualization method.

— All-in-focus image is considered to merge the focused regions of all RIs with different focus planes. This visualization method is a highly condensed version of all LFI information and can only be viewed in a passive manner. This approach is rarely used because all-in-focus images are perceptually similar to traditional 2D images

and lack an immersive experience to some extent.

In addition to the presentation of LFI content, the test methods used for scoring the LFI quality are also important. The ITU-R Rec. BT.500 [40] gives a comprehensive description of the available test methods. Here, we only introduce methods that have been used in subjective LFIQA experiments, as described below.

— The Double Stimulus Continuous Quality Scale (DSCQS) method provides a pair of LFIs side-by-side, including a reference LFI and one of its distorted versions, and then requires observers to give a score according to a continuous rating scale. The observers are usually blinded to the location of the reference LFI. This test method aims to quantitatively evaluate the visual performance of a distorted LFI relative to its reference version.

— The Double Stimulus Impairment Scale (DSIS) method is roughly the same as the DSCQS method, except that: 1) The reference LFI and its distorted version are presented separately; 2) The observers are informed of the location of the reference LFI; 3) DSIS adopts a discrete rating scale while DSCQS uses a continuous rating scale. In practical use, the DSCQS method is more suitable for comparing subtle quality differences, while the DSIS method tends to determine a range of quality distortion scores.

— In the Pairwise Comparison (PC) method, a pair of LFIs are presented horizontally side by side on the same screen to observers, who are then required to make a two-alternative-forced-choice about the one they prefer. This test method simplifies decision-making by comparing stimuli in pairs and giving relative preferences rather than absolute ratings.

2.3 Existing Subjective Databases

By collecting opinion scores in subjective experiments, several publicly-available benchmark databases for LFIQA have been developed at present. These databases differ in numerous aspects, such as image content, distortion types, visualization forms, and test methods, as summarized in TABLE 2.1.

Paudyal *et al*. [41] established the first quality assessment database for LFIs, namely the SMART database. The database includes 16 reference LFIs captured by the Lytro Illum light field camera from real-world scenes. All reference LFIs are subject to 4 types of distortions related to image compression: JPEG, JPEG2k, HEVC, and Sparse Set

Table 2.1 – Summary of existing LFIQA databases. Distortions induced by deep learning-based methods are marked with *.

Name	Size	Distortion type	Distortion category	Mul. Dis.	No. SRCs	Ang. Res.	Spa. Res.	No. Obs.
SMART [41]	256	JPEG, JPEG2k, HEVC, SSDC	Compression	No	16 (real-world)	15×15	434×625	19
MPI-LFA [42]	336	HEVC	Compression	No	14 (5 real-world and 9 dense synthesis)	1×101	960×720	40
		Gaussian blur	Acquisition or Display					
		DQ, OPT, LN, NN	Angular reconstruction					
VALID [43]	100	HEVC, VP9	Compression	No	5 (real-world)	15×15	434×625	28
		[44], [45], [46]	LFI compression					
Win5-LID [2]	220	HEVC, JPEG2k	Compression	No	10 (6 real-world and 4 dense synthesis)	9×9	434×625 or 512×512	23
		LN, NN	Angular reconstruction					
		EPICNN*, LF-SYN*	LFI angular reconstruction					
NBU-LF1.0 [37]	210	NN, BI	Angular reconstruction	No	14 (8 real-world and 6 dense synthesis)	9×9	434×625 or 512×512	22
		VDSR*	Spatial super-resolution					
		EPICNN*, Zhang	LFI angular reconstruction					
SHU [38]	240	JPEG, JPEG2k	Compression	No	8 (real-world)	15×15	434×625	20
		Gaussian blur, White noise, Motion blur	Acquisition or Display					
LFDD [39]	480	JPEG, JPEG2k, BPG, VP9, AV1, AVC, HEVC	Compression	No	8 (dense synthesis)	9×9	512×512	16
		Gaussian noise, Impulse noise, Barrel, Pincushion, Unsharp mask	Acquisition or Display					
Bakir's [47]	160	HEVC, JEM	Compression	No	10 (real-world)	15×15	434×625	18
		LF-SYN*, [44]	LFI compression					

Disparity Coding (SSDC), and each distortion type has 4 distortion levels. Thus, there are 256 distorted LFIs in total. In the subjective experiments, LFIs are visualized with all-in-focus views, and the Pairwise Comparison (PC) method is adopted for quality evaluation. This database uses Bradley-Terry (BT) values as subjective quality scores. The main disadvantage of the SMART database is that only compression-related distortions are evaluated.

Adhikarla *et al.* [42] proposed a LFIQA database (called the MPI-LFA database), which consists of 14 pristine LFIs and 336 distorted LFIs spanning 6 distortion types with 6 distortion levels. The 6 distortions of this database involve 3 categories: HEVC for compression, Gaussian blur for acquisition or display, and Quantized Depth maps (DQ), OPTical flow estimation (OPT), LINEAR interpolation (LINEAR), Nearest Neighbor interpolation (NN) for angular reconstruction. The subjective test is conducted by visualizing LFIs through viewpoint-shifting and scoring using the PC method. Just-Objectionable-Difference (JOD) is used as the subjective quality score in this database. However, since the MPI-LFA database captures reference content using the time sequential capture system, all included LFIs are 3D with only one angular dimension instead of 4D with two angular dimensions, which limits a comprehensive study on the human visual perception of LFIs.

Viola and Ebrahimi [43] presented the VALID database to subjectively evaluate the quality of distorted LFIs. The database contains 5 real-world reference LFIs [48], based on which 100 distorted LFIs are generated by 5 distortion types with 4 distortion levels. All distortions are compression-related, including 2 off-the-shelf compression solutions (HEVC and VP9) and 3 relatively state-of-the-art LFI compression algorithms ([44], [45], and [46]). The database converts LFIs into PVS and RI representations for visualization and adopts the DSIS method in the subjective evaluation. Subjective quality scores ranging from 1 to 5 are provided. Insufficient content coverage is the main shortcoming of the VALID database.

The Win5-LID database proposed by Shi *et al.* [2] collects subjective quality scores when delivering a windowed 5 degree of freedom experience to observers. There are 10 reference LFIs, of which 6 are real-world [48] and the rest are dense synthetic [1]. The Win5-LID database consists of 4 distortion types (HEVC, JPEG2k, LN, and NN) with 5 distortion levels, and 2 deep learning-based distortion types (EPICNN [49] and LF-SYN [50]) with only 1 distortion level. As a result, the database has 220 distorted LFIs. The DSCQS method and refocusing technology are adopted in the subjective evaluation. The

Mean Opinion Score (MOS) label ranging from 1 to 5 is provided. In this database, the distortions caused by super-resolution methods are not taken into account.

Huang *et al*. [37] proposed a reconstruction distortion oriented LFIQA database (NBU-LF1.0 database), in which 5 types of distortions associated with reconstruction are considered: NN and BI for angular reconstruction, VDSR [51] for spatial super-resolution, EPICNN [49] and Zhang [52] for LFI angular reconstruction. The NBU-LF1.0 database has 210 distorted LFIs, which are generated by 8 real-world reference LFIs [48] and 6 dense synthetic reference LFIs [1]. The NBU-LF1.0 database employs the same visualization method as the VALID database and adopts the DSCQS method for quality rating. The MOS label on a 5-point discrete scale is presented. In the NBU-LF1.0 database, the distortions obtained by deep learning-based algorithms are considered somewhat "old-fashioned".

Shan *et al*. [38] proposed the SHU database for the evaluation of LFI quality, consisting of 240 distorted LFIs and 8 real-world pristine LFIs [48]. This database considers 5 widely-used traditional distortion types, of which JPEG and JPEG2k are compression-related, while Gaussian blur, White noise, and Motion blur are related to acquisition or display. Each distortion type has 6 distortion levels. The PVS representation is used for LFI visualization and the DSCQS method is adopted for subjective evaluation. The MOS label ranging from 0 to 5 is presented. For the SHU database, the insufficient diversity of image types and the lack of distortions caused by more advanced algorithms restrict its broad applicability.

The LFDD database is a dense synthesis LFIQA database presented by Zizien and Fliegel [39], which contains 480 distorted LFIs disturbed from 8 reference LFIs [1]. Similar to the SHU database, the LFDD database focuses on distortion types that are caused by compression techniques (JPEG, JPEG2k, BPG, VP9, AV1, AVC, and HEVC) and acquisition or display processes (Gaussian noise, Impulse noise, Barrel, Pincushion, and Unsharp mask). Each distortion type contains 5 distortion levels. In this database, LFIs are visualized in the PVS representation and the DSIS method is used for quality rating in the subjective experiments. The database provides the MOS label ranging from 1 to 5. Despite including more distorted LFIs, the LFDD database still faces the same limitations as the SHU database, *i.e.*, image types are limited, and deep learning-based methods for obtaining distortions are not considered.

Bakir *et al*. [47] proposed a LFIQA database to evaluate the quality of LFIs affected by compression artifacts. The database involves 10 real-world reference LFIs. In this

Table 2.2 – Summary of existing objective LFIQA metrics.

Metric type	Published by	Metric name	Year	Feature type
FR	Tian *et al.* [53]	MDFM	2018	Handcrafted
	Fang *et al.* [54]	-	2018	Handcrafted
	Tian *et al.* [55]	SDFM	2020	Handcrafted
	Min *et al.* [56]	-	2020	Handcrafted
	Meng *et al.* [57]	-	2020	Handcrafted
	Huang *et al.* [58]	SGFM	2022	Handcrafted
	Meng *et al.* [59]	KRIQE	2022	Handcrafted
RR	Paudyal *et al.* [60]	LF-IQM	2019	Handcrafted
	Xiang *et al.* [61]	RRLFIQA-4DDWT	2023	Handcrafted
NR	Shi *et al.* [62]	BELIF	2019	Handcrafted
	Shi *et al.* [63]	NR-LFQA	2019	Handcrafted
	Luo *et al.* [64]	LF-QMLI	2019	Handcrafted
	Ak *et al.* [65]	-	2020	Handcrafted
	Zhou *et al.* [66]	Tensor-NLFQ	2020	Handcrafted
	Xiang *et al.* [67]	PVRI	2021	Handcrafted
	Xiang *et al.* [68]	DSA	2021	Handcrafted
	Xiang *et al.* [69]	4D-DCT-LFIQA	2021	Handcrafted
	Pan *et al.* [70]	TSSV-LFIQA	2021	Handcrafted
	Chai *et al.* [71]	NR-LF-QAE	2023	Handcrafted
	Guo *et al.* [72]	-	2021	Deep-learning
	Zhao *et al.* [73]	DeLFIQE	2021	Deep-learning
	Qu *et al.* [74]	ALAS-DADS	2021	Deep-learning
	Alamgeer *et al.* [75]	DNNF-LFIQA	2022	Deep-learning

database, 160 distorted LFIs subject to 4 distortion types and 4 distortion levels are considered, including 2 compression solutions (HEVC and JEM) and 2 LFI compression algorithms (LF-SYN [50] and [44]). The subjective experimental protocol of this database is the same as that of the LFDD database, and the MOS label ranging from 1 to 5 is provided. The extensive utilization of this database is restricted since it only considers compression artifacts.

2.4 Existing Objective Metrics

Based on the established benchmark databases, a substantial quantity of objective LFIQA metrics have been designed to approximate the subjective quality scores. Currently, according to the availability of reference information, state-of-the-art objective LFIQA metrics can be roughly divided into three categories: Full-Reference (FR), Reduced-Reference (RR), and Blind/No-Reference (NR). TABLE 2.2 provides a relatively compre-

hensive summary of existing objective LFIQA metrics. These metrics are introduced as follows.

The FR LFIQA Metrics:

— **MDFM** (Multi-order Derivative Feature-based Model)
Tian *et al.* [53] performed FR LFIQA based on multi-order derivative information and proposed the MDFM metric. Specifically, the multi-order derivative feature maps are calculated by applying the discrete derivative filter to reference and distorted LFIs, respectively. After that, the maximum value of the second-order derivative features is used as a weight map to calculate the similarity score between reference and distorted LFIs.

— **Fang's metric**
In [54], Fang *et al.* proposed a FR metric for the quality evaluation of LFIs. It calculates the local and global gradient magnitudes of reference and distorted LFIs by adopting the Scharr operator on SAIs and EPIs, respectively. Finally, the quality of distorted LFIs is obtained by combining local and global features.

— **SDFM** (Symmetry and Depth Feature-based Model)
Tian *et al.* [55] also proposed a symmetry and depth feature-based metric for FR LFIQA. First, the spatial quality is measured by applying the radial symmetry transform to the luminance components of SAIs. Then, the geometry quality is calculated through structural consistency modeling based on the depth information in EPIs. The overall LFI quality is calculated by multiplying the spatial and geometry quality with a weight parameter.

— **Min's metric**
Min *et al.* [56] measured the quality of distorted LFIs from three aspects: global, local, and angular. Based on SAIs, the global quality is measured by the view structure matching, while the spatial quality is computed by the near-edge mean square error. Then a Discrete Fourier Transform (DFT)-based angular quality estimation is used to combine the global and local quality of all SAIs into an overall LFI quality.

— **Meng's metric**
Meng *et al.* [57] assumed that human eyes have the optimal parallax range for LFI quality evaluation and used dual-fan filters to constrain the parallax range. Specifically, the spatial quality is measured by applying the difference of Gaussian operators on the central SAI, while the angular quality is obtained by averaging

the structural similarity of the focus stack. The overall LFI quality is obtained by the weighted multiplication of spatial and angular quality.

— **SGFM** (Spatial and Geometry Feature-based Model)
Huang *et al.* [58] proposed a FR LFIQA metric based on spatial and geometry features. For spatial features, the detailed content of all SAIs with multi-scale features is generated using contourlet transform. For geometry features, the viewing consistency of Multi-View Sequences (MVS) is explored by applying 3D-Gabor filters.

— **KRIQE** (Key Refocused Image Quality Evaluator)
Meng *et al.* [59] proposed to evaluate the LFI quality from two key RIs: the RI with the maximum gradient magnitude and the RI with the minimal Kullback-Leibler divergence of the phase congruency. Finally, the gradient magnitude and phase congruency of these two key RIs are used to measure the overall LFI quality.

The RR LFIQA Metrics:

— **LF-IQM** (Light Field Image Quality Measurement)
Paudyal *et al.* [60] assumed that the quality degradation of LFIs can be reflected by the depth map to some extent. Therefore, the structural similarity of the depth map between reference and distorted LFIs is exploited as the predicted quality.

— **RRLFIQA-4DDWT** (Reduced-Reference Light Field Image Quality Assessment based on 4D Discrete Wavelet Transform)
Xiang *et al.* [61] proposed a 4D DWT-based RR metric to evaluate the quality of distorted LFIs. The main characteristic of this metric is that a PLFI is created as the RR information using a view synthesis algorithm, and then 4D-DWT is used to decompose the reference and distorted PLFIs. To obtain the final quality score, a spatial-angular weighting strategy that describes the importance of each location is adopted.

The NR LFIQA Metrics:

— **BELIF** (Blind quality Evaluator of LIght Field image)
Shi *et al.* [62] proposed to obtain the first three components of the LFI cyclopean image array through Tucker decomposition. Then, zero-mean Asymmetric Generalized Gaussian Distribution (AGGD) is used to fit the Mean Subtracted and Contrast Normalized (MSCN) coefficients, in which the fitting parameters are used to measure the spatial quality of LFIs. In addition, structural similarity statistical features are exploited for angular quality evaluation.

— **NR-LFQA** (No-Reference Light Field image Quality Assessment)
In addition to employing the spatial quality measurement in BELIF, Shi *et al.* [63] additionally calculated the gradient direction distribution and weighted local binary pattern on both horizontal and vertical EPIs, in order to measure the global and local angular consistency distortions in LFIs, respectively.

— **LF-QMLI** (Light Field image Quality assessment model based on MLI)
Luo *et al.* [64] proposed to measure the global and local angular inconsistency by computing the entropy distributions and texture variations of LFIs, respectively. Also, the LFI spatial degradation is evaluated by measuring information entropy in both image and frequency domains.

— **Ak's metric**
Ak *et al.* [65] exploited overall structural statistics and local non-natural structures to assess the quality of distorted LFIs. In specific, a bag-of-words codebook based on the Histogram of Oriented Gradient (HOG) is employed to represent the overall structural statistics, while a Convolutional Sparse Coding (CSC) codebook is adopted to measure the local non-natural structures.

— **Tensor-NLFQ** (Tensor oriented No-reference Light Field image Quality evaluator)
Zhou *et al.* [66] first adopted Tucker decomposition on SAI view stacks in four directions to reduce the dimensionality of LFIs. Subsequently, the quality of distorted LFIs is characterized by the AGGD fitting parameters of MSCN coefficients, DCT-based information entropy, and structural similarity properties.

— **PVRI** (Pseudo Video and Refocused Images-based light field image quality assessment)
Xiang *et al.* [67] proposed a NR LFIQA metric based on PVS and RIs. It focuses on evaluating the angular quality from the structure, motion, and disparity information of the decomposed PVS, as well as the spatial quality from the depth and semantic information of RIs. By combining these two aspects, this metric offers a comprehensive evaluation of the quality of distorted LFIs.

— **DSA** (light field image quality assessment based on Depth, Structural and Angular information)
In [68], Xiang *et al.* first extracted the depth and structure information from the mean difference image of LFIs. The Curvelet multi-scale and multi-directional decomposition is then employed, followed by Natural Scene Statistics (NSS) and en-

ergy feature extraction. Afterwards, the angular deterioration is considered through a local maximum similarity index statistics algorithm.

— **4D-DCT-LFIQA** (4D Discrete Cosine Transform-based Light Field Image Quality Assessment)
Compared to most previous works that focus on extracting 2D NSS features, Xiang *et al*. [69] proposed to estimate the quality of distorted LFIs in the 4D-DCT domain. Firstly, the sub-aperture gradient image array of LFIs is generated by high-pass filtering between adjacent SAIs. Subsequently, 4D-DCT is employed to generate a large number of block-based 4D-DCT coefficients. Finally, the frequency, orientation, energy, and amplitude features of coefficients are extracted for LFI quality assessment.

— **TSSV-LFIQA** (Tensor Slice and Singular Value for blind Light Field Image Quality Assessment)
Pan *et al*. [70] considered the LFI as a five-order tensor containing intensity, angular, and color information and adopted the Tucker decomposition for redundancy elimination. Afterwards, the sharpness and distribution information of tensor slices and the percentage of singular values are exploited for quality assessment.

— **NR-LF-QAE** (No-Reference Light Field image Quality Assessment Evaluator)
Chai *et al*. [71] evaluated the LFI quality from three aspects: group-based representations for angular consistency, multiple plane-oriented characteristics for spatial-angular texture statistics, and 2D/3D Log-Gabor features for spatial-angular structure statistics. Three features are combined to describe the quality deterioration of LFIs.

— **Guo's metric**
Guo *et al*. [72] proposed a deep Convolutional Neural Network (CNN)-based model for the quality evaluation of LFIs. It includes a SAI fusion module to integrate different SAIs, and a global context perception module to capture global context. Further, a ranking-based pre-training method is proposed to alleviate the overfitting effect.

— **DeLFIQE** (Deep learning-based Light Field Image Quality Evaluator)
Zhao *et al*. [73] proposed a CNN model using a distortion-aware multitask learning scheme to assess the quality of distorted LFIs. The model is of low complexity as it is trained and tested only on discriminative EPI patches. The overall LFI quality is obtained based on the predicted distortion category of each discriminative EPI

patch.

— **ALAS-DADS** (Auxiliary Learning with Angular and Spatial features via Deep Anglewise and Depthwise Separable convolutions)
Qu *et al.* [74] extended the concept of depthwise separable convolution to the spatial domain and angular domain of LFIs, respectively, and thus proposed a CNN model that extracts spatial and angular information simultaneously. In addition, an auxiliary learning scheme is designed, which provides more ready-made hints to help model learning.

— **DNNF-LFIQA** (Light Field Image Quality Assessment based on a Deep Neural Network that uses Frequency domain inputs)
In [75], Alamgeer *et al.* proposed a CNN model consisting of two processing streams to extract deep features from horizontal and vertical EPIs, respectively. The main characteristic of this metric is the use of the Fourier magnitude spectrum of EPIs as input.

2.5 Train-Test Strategy

To evaluate the effectiveness of objective metrics on a subjective database, a reasonable train-test strategy is a prerequisite to ensure the reliability of the verification experiment. In this thesis, we adopt leave-two-fold-out cross-validation as the train-test strategy to conduct experiments.

Specifically, for each database, we first divide all distorted LFIs into $\mathbb{K}$ folds according to their corresponding reference LFIs, where $\mathbb{K}$ denotes the number of reference scenes. As a result, each fold contains all distorted versions of a single reference LFI. Then we randomly use $\mathbb{K}$-2 folds for training and the remaining 2 folds for testing, which ensures that the training and test sets are independent of each other. After looping through all possible train-test splits (*i.e.*, $\mathbb{K}(\mathbb{K}\text{-}1)/2$ splits), the average result is reported as the final performance. Taking the Win5-LID database as an example, since this database contains 10 reference LFIs and 220 distorted LFIs, $\mathbb{K}$ is set to 10. Afterwards, the 176 distorted LFIs in 8 folds are used as the training set and the 44 distorted LFIs in the remaining 2 folds are treated as the test set. Consequently, there are a total of 45 possible train-test splits for comprehensive quality evaluation experiments.

For a fair comparison of experimental results, on the one hand, the learning-based metrics are trained on $\mathbb{K}$-2 folds, and their quality evaluation performance is tested on

the remaining 2 folds. On the other hand, the non-learning-based metrics do not require any training process for quality prediction, so they are directly performed on the same test set.

2.6 Performance Evaluation Criteria

The performance of objective metrics is evaluated by the relationship between objective quality predictions and subjective quality scores. Generally, three standard criteria are used to evaluate the performance, including Pearson Linear Correlation Coefficient (PLCC), Spearman Rank order Correlation Coefficient (SRCC), and Root Mean Square Error (RMSE). Among them, PLCC is used to evaluate the linear relationship, SRCC focuses on the monotonicity while RMSE measures the predictive accuracy. Higher PLCC and SRCC values indicate better performance, while the opposite is true for RMSE. For two sequences $\{x_1, x_2, ..., x_n\}$ and $\{y_1, y_2, ..., y_n\}$, in which n is the length of sequences, the PLCC, SRCC, and RMSE are defined as:

$$PLCC = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{2.1}$$

$$SRCC = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2.2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum (x_i - y_i)^2} \tag{2.3}$$

where d_i denotes the difference of the ranking of two sequences.

Before the calculation of PLCC and RMSE, a five-parameter logistic mapping function, recommended by the Video Quality Expert Group (VQEG) [76], is employed:

$$f(p) = \beta_1 \{\frac{1}{2} - \frac{1}{1 + e^{\beta_2(p-\beta_3)}}\} + \beta_4 p + \beta_5 \tag{2.4}$$

where $\beta_{1...5}$ are the fitting parameters, p and $f(p)$ denote the objective prediction and its nonlinear mapping result, respectively.

2.7 Conclusion

In this chapter, we review various aspects of state-of-the-art LFIQA methods. Considering the complex visual effects of LFI distortions, we summarize the main distortions that LFIs may involve. Next, subjective experimental protocols are introduced, followed by a comprehensive description of existing subjective LFIQA databases. The main features and shortcomings of these databases are also discussed. Afterwards, we present an exhaustive list of state-of-the-art objective LFIQA metrics based on their reliance on reference information, and then introduce their main ideas and implementations. Finally, the train-test strategy and performance evaluation criteria used in this thesis to evaluate the performance of objective metrics are provided.

Chapter 3

Proposed Handcrafted Feature-based Metrics for Light Field Images

As introduced in Chapter 2, a wealth of handcrafted feature-based quality assessment metrics have been proposed to assess the quality of distorted LFIs. However, despite the remarkable achievements, existing objective LFIQA metrics cannot well evaluate the LFI quality in various scenarios, and their time complexity is also far from ideal, making them difficult to use in practical applications. In this chapter, we design robust and efficient handcrafted features for LFI quality evaluation and propose two NR LFIQA metrics. In the first metric, we exploit the classical Local Binary Pattern (LBP) operator on LFIs and calculate spatio-angular textural variations to assess the LFI quality via the widely-used Support Vector Regression (SVR). The second metric is extended from the first metric, in which we measure the holistic textural variations of LFIs by extending the textural variation measurement to multiple planes. Besides, we capture the wavelet information of LFIs in the 4D-DHWT domain. Finally, we combine all the extracted quality-aware features and employ SVR to predict the perceptual quality of LFIs.

This chapter is presented in four sections. The first section describes the proposed Blind Light Field quality assessment metric based on Spatio-Angular Textural Variations (SATV-BLiF) in detail. The second section provides an in-depth introduction to the proposed Blind Light Field image quality assessment metric based on MultiPlane Texture and Multilevel Wavelet Information (MPT-MWI-BLiF). The experimental results and discussions are presented in the third section. Finally, the last section concludes this chapter.

3.1 Proposed SATV-BLiF Metric

A widely accepted idea is that the degradation of perceptual quality can be reflected in textural variations. The excellent ability of the LBP descriptor [77] for measuring textural variations and quality degradation has been demonstrated in many traditional IQA studies (*e.g.*, [78]). NR-LFQA [63] is the first attempt to use LBP for NR LFIQA, but where LBP is only applied on EPIs to evaluate the angular variation. We consider the potential of LBP for LFIQA has not been fully exploited. Therefore, in this section, we propose a simple yet effective metric SATV-BLiF by measuring the spatio-angular textural variations of LFI with the LBP descriptor.

Figure 3.1 exhibits the flowchart of the proposed SATV-BLiF metric, which mainly consists of the following steps. Firstly, all SAIs of distorted LFIs are converted view-by-view to grayscale. Secondly, spatial textural variations and angular textural variations are measured using the LBP operator, respectively, which are further transformed into the spatio-angular textural matrix. Thirdly, we compute statistical textural histogram features from the spatio-angular textural matrix. Finally, the connection from statistical textural histogram features to quality scores is established through SVR.

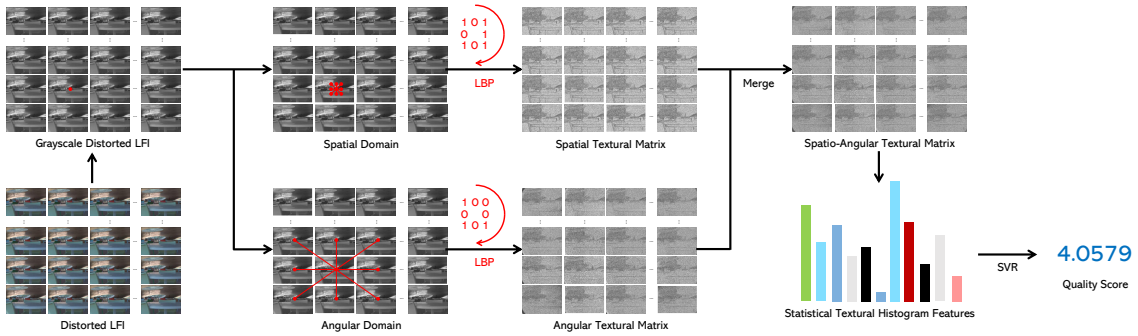


Figure 3.1 – Flowchart of the proposed SATV-BLiF metric.

3.1.1 Spatio-Angular Textural Variations

The classical LBP algorithm [77] describes the relationship between the central pixel and its surrounding neighbors. As a result, the generated LBP textural matrix $LBP_{P,R}$ is defined as:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_c,\ g_{R,p}) \cdot 2^p \tag{3.1}$$

$$s(g_c,\ g_{R,p}) = \begin{cases} 1, & g_c \leq g_{R,p} \\ 0, & g_c > g_{R,p} \end{cases} \tag{3.2}$$

where P and R determine the number and radius of the involved neighbors, respectively. Their numerical relationship is usually set as $P = 8 \cdot R$. In addition, g_c is the gray-level value of the central pixel c, while $g_{R,p}$ is the gray-level value of the circularly symmetric neighbor pixel p with the radius of R. 2^p is a binomial factor for encoding each location into a decimal number. To obtain the ability of rotational invariance with uniform patterns [77], $LBP_{P,R}$ is redefined as:

$$LBP^{riu2}_{P,R} = \begin{cases} \sum_{p=0}^{P-1} s(g_c,\ g_{R,p}), & k(LBP_{P,R}) \leq 2 \\ P+1, & otherwise \end{cases} \tag{3.3}$$

$$\begin{aligned} k(LBP_{P,R}) = &||s(g_c,\ g_{R,P-1}) - s(g_c,\ g_{R,0})|| \\ &+ \sum_{p=1}^{P-1} ||s(g_c,\ g_{R,p}) - s(g_c,\ g_{R,p-1})|| \end{aligned} \tag{3.4}$$

where the superscript ^{riu2} denotes the rotation invariant uniform patterns whose bitwise transition number k is not greater than 2.

Since LFI simultaneously contains spatial and angular information [57], we measure the textural variations of distorted LFIs from the spatial and angular domains, respectively. As a result, the spatial textural matrix $LBP^{riu2}_{P,R,S}$ and the angular textural matrix $LBP^{riu2}_{P,R,A}$ are generated, each of which would have $P+2$ patterns according to Eq. (3.3) and Eq. (3.4). In order to characterize textural variations in a unified manner, we define the spatio-angular textural matrix as:

$$LBP^{riu2}_{P,R,SA} = LBP^{riu2}_{P,R,S} + (P+2) \cdot LBP^{riu2}_{P,R,A} \tag{3.5}$$

where P and R are set to 8 and 1 by default [77].

3.1.2 Statistical Textural Histogram and Quality Prediction

After obtaining the spatio-angular textural matrix, we summarize the frequency distributions of different patterns into histogram, and then constrain the results to [0,1] using l_1-normalization. Let $\mathcal{F}_{SATV-BLiF}$ denote the generated statistical textural histogram fea-

tures of the proposed SATV-BLiF metric, its generation can be described as:

$$\mathcal{F}_{SATV-BLiF} = l_1(hist(LBP^{riu2}_{P,R,SA})) \tag{3.6}$$

where $hist(\cdot)$ is the histogram operation and $l_1(\cdot)$ is the l_1-normalization.

In order to conduct the quality assessment, a nonlinear mapping function between the statistical textural histogram features and the quality score is established. Specifically, the SVR with a radial basis function kernel is adopted as the mapping function, which is implemented using the LIBSVM package [79]. The above processing can be described as:

$$Q_{SATV-BLiF} = SVR(\mathcal{F}_{SATV-BLiF}) \tag{3.7}$$

where $Q_{SATV-BLiF}$ denotes the predicted LFI quality score, and $SVR(\cdot)$ represents the trained SVR model.

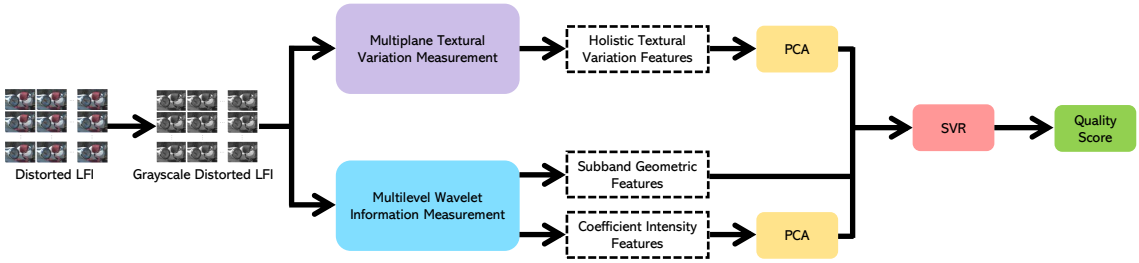


Figure 3.2 – Block diagram of the proposed MPT-MWI-BLiF metric.

3.2 Proposed MPT-MWI-BLiF Metric

Motivated by the holistic visual perception of high-dimensional LFIs and the physiological mechanism of the Human Visual System (HVS), we propose a novel blind LFIQA metric MPT-MWI-BLiF based on multiplane texture and multilevel wavelet information, which is also considered as an updated version of the SATV-BLiF metric.

Figure 3.2 presents the block diagram of the proposed MPT-MWI-BLiF metric. Firstly, given a distorted LFI represented in the SAI array form, we first convert it to grayscale in a view-by-view manner. Subsequently, two separate measurements are designed for feature extraction. With the multiplane textural variation measurement, we extract holistic textural variation features and reduce their dimensions via Principal Component Analysis (PCA). With the multilevel wavelet information measurement, we consecutively exploit

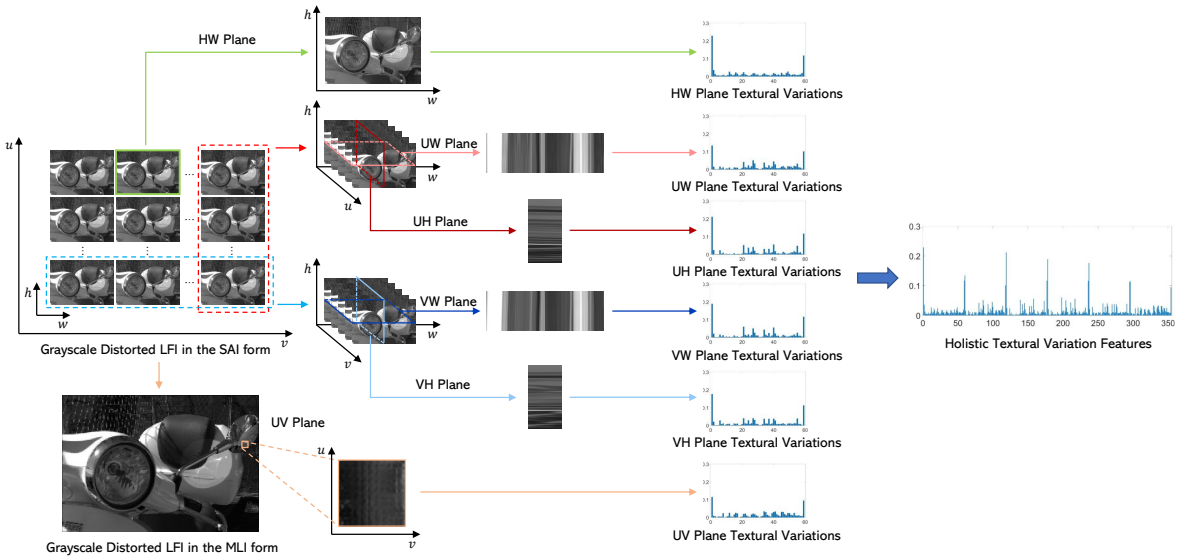


Figure 3.3 – Pipeline of Multiplane Textural Variation Measurement.

subband geometric features and the dimensionality-reduced coefficient intensity features. Finally, these three quality-aware features are combined to provide a comprehensive description of multifarious LFI distortions, followed by quantitative quality prediction using SVR.

3.2.1 Multiplane Textural Variation Measurement

According to previous HVS studies [80], HVS is a hierarchical structure that encodes visual signals layer by layer, from concrete to abstract. In addition, different layers have different sensitivities to different signals. More specifically, when the human eye is exposed to an external light source, light signals are first converted into neural signals by the photoreceptors in the retina. After basic signal processing and integration, visual signals are transmitted to the primary visual area (V1). The visual cortex in the V1 area is more sensitive to basic features like orientation, edges, color, and brightness, which help to construct a basic visual image. These visual signals are subsequently processed by the V2 area to generate more abstract features. The main focus of the V2 area is texture, location, shape, and motion, which contribute to the understanding of visual images. Among these features, texture information is well proven to be highly correlated with the distortions of images [81], thereby spawning an abundance of milestone quality assessment methods [82], [83].

Further, considering that LFI is a high-dimensional visual signal, it may be distorted differently not only due to different distortion types and levels, but also due to different representations. For example, the JPEG compression can be applied individually to each SAI of LFIs, or directly to the MLI of LFIs. In this case, the former only reduces the redundancy in spatial information, while the latter jointly compresses angular and spatial information. This implies that we should focus on the holistic visual perception brought by the interaction of various dimensions in LFIs, rather than the visual perception of a specific representation.

Motivated by the above observations, we propose a multiplane textural variation measurement, in which holistic textural variations are exploited to describe the holistic visual perception of LFIs. Our underlying hypothesis is that the interaction between different LFI sub-dimensions collectively affects its holistic visual perception. As shown in Figure 3.3, for a given four-dimensional grayscale LFI $\mathcal{L}_g \in \mathbb{R}^{U\times V\times H\times W}$, we first convert it into six individual planes by pairing any two sub-dimensions, *i.e.*, HW, UW, UH, VW, VH, and UV planes are obtained. Then, we employ LBP [77] to measure the textural variations of each plane, as many studies have demonstrated its excellent ability to reflect quality degradation [63]. Taking the UV plane for instance, for a center pixel c on the UV plane, LBP assigns binary labels to each neighboring pixel and generates decimal values representing their patterns:

$$LBP^{UV}_{P,R}(c) = \sum_{p=0}^{P-1} s(g_c,\ g_{R,p}) \cdot 2^p \tag{3.8}$$

$$s(g_c,\ g_{R,p}) = \begin{cases} 1, & g_c \leq g_{R,p} \\ 0, & g_c > g_{R,p} \end{cases} \tag{3.9}$$

where P and R are the number and radius of the involved neighbors, respectively. By default [77], they generally follow the numerical relationship of $P = 8 \cdot R$. The binomial factor 2^p is used to encode each neighboring pixel as a decimal number. Moreover, g_c and $g_{R,p}$ denote the gray-level value of the central pixel c and the circularly symmetric neighboring pixel p with R radius.

To reduce the pattern number and enhance robustness, LBP with uniform patterns is

further defined as:

$$LBP_{P,R}^{UV,u2}(c) = \begin{cases} \sum_{p=0}^{P-1} s(g_c,\ g_{R,p}) \cdot 2^p, & k(LBP_{P,R}^{UV}(c)) \leq 2 \\ -1, & otherwise \end{cases} \tag{3.10}$$

$$\begin{aligned} k(LBP_{P,R}^{UV}(c)) = &||s(g_c,\ g_{R,P-1}) - s(g_c,\ g_{R,0})|| \\ &+ \sum_{p=1}^{P-1} ||s(g_c,\ g_{R,p}) - s(g_c,\ g_{R,p-1})|| \end{aligned} \tag{3.11}$$

where $k(\cdot)$ calculates the bitwise transition number of a circle, while the superscript $u2$ denotes the uniform patterns with k not greater than 2.

After generating the pixel-wise LBP encoding, the LBP histogram is used to represent the global textural variations. The histogram can effectively capture the overall distribution characteristics without being affected by specific pixels. Further, it exhibits invariance with respect to scaling and translation. Likewise, taking the UV plane for instance, the LBP histogram recording the frequency of each pattern is calculated as:

$$H^{UV}(b) = \sum_{u=1}^{U} \sum_{v=1}^{V} \gamma(LBP_{P,R}^{UV,u2}(c) = D_b) \tag{3.12}$$

$$\gamma(A) = \begin{cases} 1, \text{ if A is true} \\ 0, \text{ if A is false} \end{cases} \tag{3.13}$$

where b and D_b denotes the b-th pattern and its corresponding decimal value. In the uniform pattern mode, there are 59 patterns (*i.e.*, $b \in [1, 59]$), including 58 uniform patterns and 1 non-uniform pattern. Accordingly, H^{UV} is a 59-dimensional histogram, which is used to measure the textural variations of the UV plane. H^{HW}, H^{UW}, H^{UH}, H^{VW}, and H^{VH} are calculated in a similar manner.

The histogram of each plane records the texture information of two sub-dimensions of the whole LFI, thereby reflecting the LFI quality degradation from a potential viewing perspective. Finally, to capture holistic textural variations and provide a comprehensive description of LFI quality, we combine the histograms of all individual planes by dimensionality concatenation and generate holistic textural variation features, denoted by $\mathcal{F}_{TV}$=$\{H^{HW}, H^{UW}, H^{UH}, H^{VW}, H^{VH}, H^{UV}\}$.

The generated $\mathcal{F}_{TV}$ contains all patterns from six LBP histograms and thus has 354

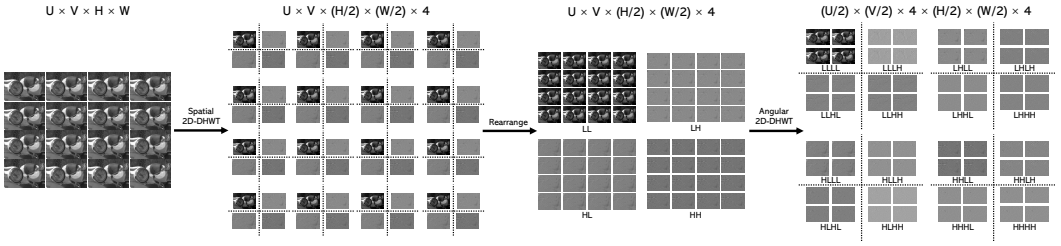


Figure 3.4 – Example of decomposing a LFI with 4D-DHWT. The angular resolution of the input LFI is set to 4×4 for better visualization.

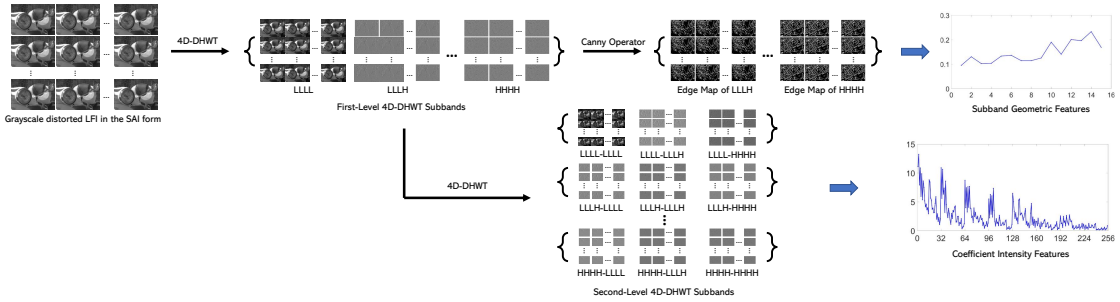


Figure 3.5 – Pipeline of Multilevel Wavelet Information Measurement.

dimensions. These features can effectively model the complex human visual perception when viewing LFIs, due to the full consideration of holistic textural variations. However, redundant features inevitably increase computational costs and affect predictive accuracy. Following [69], to obtain more differentiated and efficient features, $\mathcal{F}_{TV}$ is reduced to d_1 dimensions via PCA.

3.2.2 Multilevel Wavelet Information Measurement

As presented in neuroscience studies [84], [85], neurons in most visual cortices give specific responses to different frequencies when processing visual information. For example, neurons with visual functions in photoreceptors can be divided into two types: cone cells and rod cells [86]. Among them, cone cells are more sensitive to high frequencies that contain image details, while rod cells are more involved in perceiving the overall structure and spatial low-frequency features of the image. As a mathematical manipulation method, Discrete Wavelet Transform (DWT) can decompose a signal into multiple subbands of different frequencies, whose intensity can be analogized to the activity levels of neurons in different frequency ranges. Therefore, DWT is recognized as an effective tool for analyzing spatial and frequency domain information of 2D images [87]. Among the existing wavelets,

the Haar wavelet achieves significant success in quality evaluation due to its properties of unbiasedness and computational efficiency [88]. Mathematically, the basis Haar wavelet is defined as:

$$\begin{aligned}\frac{1}{\sqrt{2}}\phi(t) =& \frac{1}{\sqrt{2}}(\psi(t-1)-\psi(t)) \\ =& \sum_{n=-\infty}^{+\infty}(-1)^{1-n}h[1-n]\psi(t-n)\end{aligned} \tag{3.14}$$

where the mother wavelet function ϕ is defined as:

$$\phi(t) = \begin{cases} 1, & 0 \leq t < \frac{1}{2} \\ -1, & \frac{1}{2} \leq t < 1 \\ 0, & otherwise \end{cases} \tag{3.15}$$

and the scaling function ψ and its corresponding filter h are defined as:

$$\psi(t) = \begin{cases} 1, & 0 \leq t < 1 \\ 0, & otherwise \end{cases} \tag{3.16}$$

$$h[n] = \begin{cases} \frac{1}{\sqrt{2}}, & n = 0, 1 \\ 0, & otherwise \end{cases} \tag{3.17}$$

Based on the basis Haar wavelet, the 2D-DHWT of a 2D signal $\boldsymbol{F}$ is calculated by cross-multiplying different displacements and stretches, and four wavelet subbands are obtained. This process is formulated as:

$$\boldsymbol{H}\boldsymbol{F}\boldsymbol{H}^T = \begin{bmatrix} LL & LH \\ HL & HH \end{bmatrix} \tag{3.18}$$

where $\boldsymbol{H}$ denotes the Haar wavelet transform matrix. The superscript T represents the transpose operation. LL, LH, HL, and HH are four decomposed subbands containing different coefficients, with a low or high frequency in a horizontal or vertical direction.

Considering that the two additional angular dimensions distinguish LFIs from 2D images, we utilize 4D-DHWT to decompose LFIs into wavelet subbands by performing 2D-DHWT on spatial and angular information sequentially. An example of decomposing

a LFI with 4D-DHWT is illustrated in Figure 3.4. Specifically, 2D-DHWT is first applied along the spatial dimension to generate four spatial subbands for each SAI. Subsequently, the resulting subbands are rearranged to aggregate all subbands belonging to the same frequency. Finally, 2D-DHWT is employed on each set of spatial subbands along the angular dimension, generating four spatio-angular subbands. As a result, 16 spatio-angular subbands are obtained after using 4D-DHWT. These decomposed wavelet information effectively simulate the processing of visual signals in the HVS, and thus can be used to design quality-aware features and evaluate the degradation of LFI quality.

On the basis of 4D-DHWT, we propose a multilevel wavelet information measurement to exploit wavelet information, as shown in Figure 3.5. Also, given a grayscale LFI $\mathcal{L}_g \in \mathbb{R}^{U\times V\times H\times W}$, we first decompose it into 16 wavelet subbands with 4D-DHWT, which are also called first-level 4D-DHWT subbands. This process is expressed as:

$$\{\mathcal{W}_n\} = \xi(\mathcal{L}_g) \tag{3.19}$$

where $\{\mathcal{W}_n\}$ denotes all decomposed subbands of $\mathcal{L}_g$, with $n \in \{LLLL, LLLH, ..., HHHH\}$. $\xi(\cdot)$ represents the 4D-DHWT. Considering that distortions often affect the high-frequency information of an image, and object edges are typical representatives of high-frequency information, we propose to detect the edges of the geometric information in high-frequency subbands. Specifically, the Canny operator [89] is employed on each SAI of high-frequency subbands, which is denoted as:

$$C_n = Canny(\mathcal{W}_n) \tag{3.20}$$

where C_n denotes the edge map of $\mathcal{W}_n$. Note that $\mathcal{W}_{LLLL}$ is not involved here because it mainly contains low-frequency information. To measure the global geometric information, we calculate the subband geometric features $\mathcal{F}_{SG}$ as:

$$\mathcal{F}_{SG}(n) = \frac{1}{L}\sum_{l=1}^{L} C_n(l) \tag{3.21}$$

where l is the coefficient index and L represents the total number of coefficients. As a result, the length of $\mathcal{F}_{SG}$ is 15.

Moreover, in order to conduct deeper analysis and processing of information at different scales and frequency ranges, we further employ 4D-DHWT on each first-level 4D-DHWT

subband and generate second-level 4D-DHWT subbands, which is expressed as:

$$\{\hat{\mathcal{W}}_{n,m}\} = \xi(\mathcal{W}_n) \tag{3.22}$$

where $\{\hat{\mathcal{W}}_{n,m}\}$ denotes all decomposed subbands of $\mathcal{W}_n$, with $m \in \{LLLL, LLLH, ..., HHHH\}$. After decomposing all first-level 4D-DHWT subbands, 256 subbands of different frequencies are obtained. As presented in previous studies [69], the intensity of high-frequency coefficients is an important statistical feature to indicate the deterioration of image quality. In other words, if an image is distorted, the coefficient intensity distribution of different subbands will change accordingly. Thus, we calculate the coefficient intensity features $\mathcal{F}_{CI}$ as:

$$\mathcal{F}_{CI}(n,m) = \log_2\left(\frac{1}{L}\sum_{l=1}^{L}(\hat{\mathcal{W}}_{n,m}(l))^2 + 1\right) \tag{3.23}$$

Also, $\hat{\mathcal{W}}_{LLLL,LLLL}$ is not considered into $\mathcal{F}_{CI}$ because it is low-frequency information. The original length of $\mathcal{F}_{CI}$ is 255 and PCA is employed to $\mathcal{F}_{CI}$ for the sake of efficiency and effectiveness. The reduced dimension is denoted by d_2.

Consequently, subband geometric features $\mathcal{F}_{SG}$ and coefficient intensity features $\mathcal{F}_{CI}$ jointly describe the wavelet information of LFIs.

3.2.3 Dimension Reduction and Quality Prediction

After the above feature extraction, the proposed MPT-MWI-BLiF metric constructs an overall feature vector to comprehensively describe the deterioration of the perceptual quality of a LFI, *i.e.*, $\mathcal{F}_{MPT-MWI-BLiF}$=$\{\mathcal{F}_{TV}, \mathcal{F}_{SG}, \mathcal{F}_{CI}\}$. There are two hyperparameters, d_1 and d_2, in the proposed features, determining the feature length of $\mathcal{F}_{TV}$ and $\mathcal{F}_{CI}$ after PCA, respectively. In our implementation, we set d_1=d_2=35.

To conduct the quality evaluation, the SVR with a radial basis function kernel [79] is utilized to map the overall features $\mathcal{F}_{MPT-MWI-BLiF}$ to quantitative quality scores $Q_{MPT-MWI-BLiF}$, which is expressed as:

$$Q_{MPT-MWI-BLiF} = SVR(\mathcal{F}_{MPT-MWI-BLiF}) \tag{3.24}$$

where $SVR(\cdot)$ denotes the trained regression model.

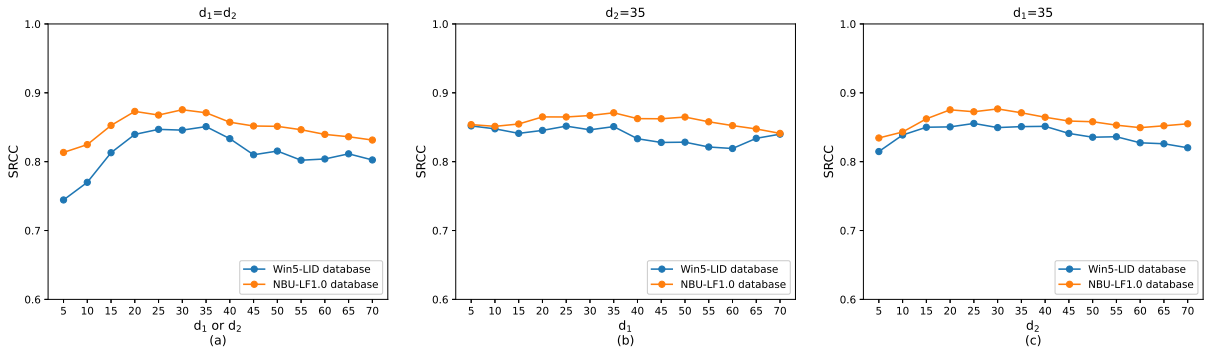


Figure 3.6 – Performance of different d_1 and d_2 values on the Win5-LID and NBU-LF1.0 databases.

3.3 Performance Evaluation and Discussion

To verify the effectiveness of our proposed SATV-BLiF metric and MPT-MWI-BLiF metric, we conduct extensive experiments on five standard LFIQA databases, including Win5-LID [2], NBU-LF1.0 [37], SHU [38], VALID10bit [43], and LFDD [39].

3.3.1 Hyperparameter Dependency

The performance of the proposed MPT-MWI-BLiF metric is affected by two hyperparameters, *i.e.*, the reduced dimension number d_1 and d_2 for $\mathcal{F}_{TV}$ and $\mathcal{F}_{CI}$, respectively. To determine the best hyperparameters, corresponding experiments are conducted on the Win5-LID and NBU-LF1.0 databases, to study the impact of different hyperparameters on the SRCC performance. Considering that there are many different combinations of d_1 and d_2 values, for simplicity, we first assume d_1=d_2, set the values from 5 to 70 with the step of 5, and report the results in Figure 3.6 (a). The abscissa denotes the values of d_1 or d_2, and the ordinate represents the SRCC performance. We can observe that a small dimension value leads to relatively poor performance for both Win5-LID and NBU-LF1.0 databases. Subsequently, the performance improves as the dimension value increases, but the opposite occurs after reaching d_1=d_2=35. Then, we further investigate the performance dependency of different d_1 values under the condition of d_2=35, as shown in Figure 3.6 (b). A similar investigation is conducted for different d_2 values, and the results are shown in Figure 3.6 (c). It can be found that a moderate d_1 or d_2 value usually leads to an acceptable performance. Therefore, we adopt d_1=d_2=35 in our MPT-MWI-BLiF metric for simplicity.

3.3.2 Overall Performance Comparison

Based on the leave-two-fold-out cross-validation strategy and performance evaluation criteria in Chapter 2, this subsection compares the proposed SATV-BLiF metric and MPT-MWI-BLiF metric with three FR 2DIQA metrics (NQM [90], IFC [91], HDR-VDR-2 [92]), two RR LFIQA metrics (LF-IQM [60], RRLFIQA-4DDWT [61]), five FR LFIQA metrics (MDFM [53], Fang's [54], Min's [56], Meng's [57], KRIQE [59]), three NR 2DIQA metrics (GWH-GLBP [93], NIQE [94], BRISQUE [95]), and nine NR LFIQA metrics (BELIF [62], NR-LFQA [63], Tensor-NLFQ [66], PVRI [67], DSA [68], 4D-DCT-LFIQA [69], TSSV-LFIQA [70], NR-LF-QAE [71], DeLFIQE [73]). Among them, only DeLFIQE is a deep learning-based LFIQA metric, while other metrics are based on handcrafted features. For 2DIQA metrics, we employ them on all SAIs and take the average result as the final performance. The reported performance of all metrics is reproduced using the released code and default parameter settings from their authors to avoid bias.

The experimental results of overall performance are shown in Table 3.1. It can be seen that due to the limited consideration of angular inconsistency, traditional 2DIQA metrics perform much worse than LFIQA metrics in quality evaluation. In addition, due to the diversity of reference scenes, distortion types and levels of the adopted LFIQA databases, most existing LFIQA metrics generally struggle to perform well on all databases. In most cases, the learning-based NR LFIQA metrics can achieve competitive or even superior performance than the FR LFIQA metrics. However, on the VALID10bit database, the FR LFIQA metrics generally obtain better results than the NR LFIQA metrics. One possible reason is that the VALID10bit database contains only 100 distorted LFIs, which may be insufficient to train a well-performing NR model. It can also be found that the proposed SATV-BLiF metric outperforms most NR metrics on the Win5-LID, NBU-LF1.0, SHU, and VALID10bit databases, but its performance on the LFDD database is not satisfactory. Even compared with the RR/FR metrics, our SATV-BLiF metric still exhibits competitive performance. While compared to SATV-BLiF, the extended version MPT-MWI-BLiF consistently improves the quality evaluation performance by a large margin on all databases. In particular, MPT-MWI-BLiF obtains more than 20% performance improvement over SATV-BLiF on the LFDD database. Also, it should be highlighted that on the two most complex databases Win5-LID and NBU-LF1.0, which contain both real-world and synthetic LFIs, the MPT-MWI-BLiF metric outperforms all existing NR/RR/FR metrics by a significant margin. All these results demonstrate the superiority of our proposed SATV-BLiF metric and MPT-MWI-BLiF metric in quality

Table 3.1 – Overall performance comparison and time complexity on five LFIQA databases. The best results of RR/FR metrics are in blue, while the best results of NR metrics are marked in **bold**.

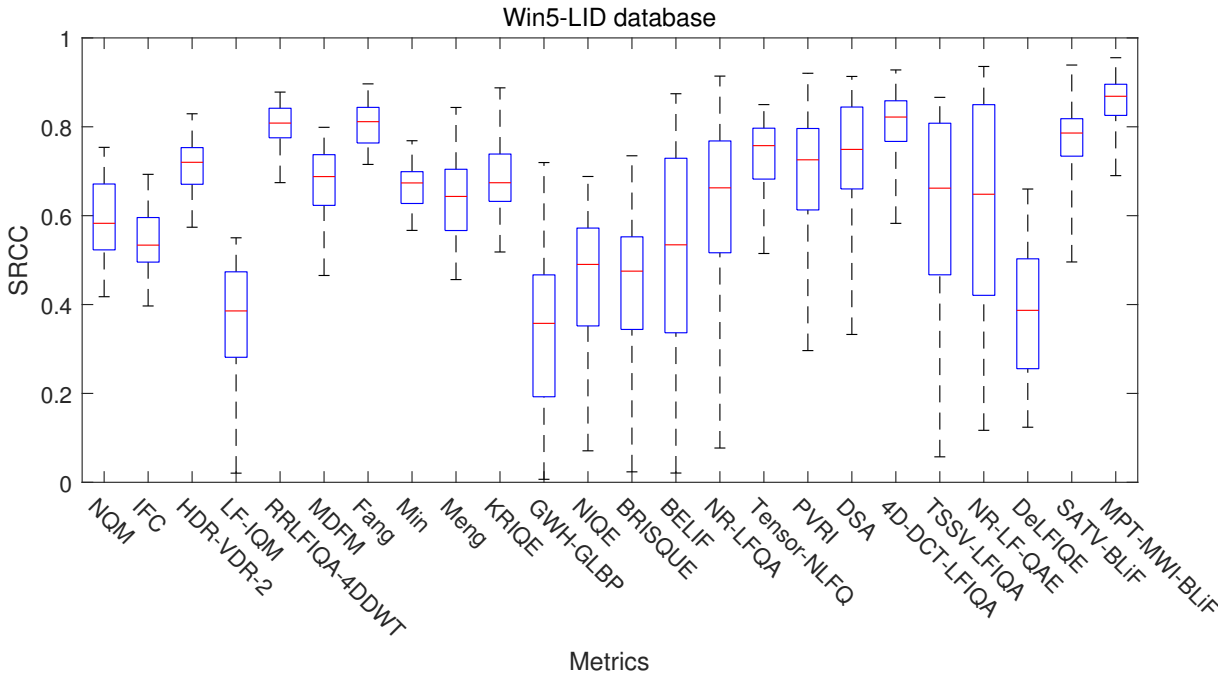
	Database	Win5-LID			NBU-LF1.0			SHU			VALID10bit			LFDD			
Type	Metric	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	Time (s)
FR 2DIQA	NQM [90]	0.6509	0.5908	0.9082	0.6479	0.6614	0.6995	0.8581	0.8618	0.5619	0.8702	0.8630	0.4586	0.4461	0.3810	0.9989	10.1941
	IFC [91]	0.4945	0.5429	1.0628	0.7440	0.6843	0.6049	0.8853	0.9068	0.4969	0.9613	0.9647	0.2260	0.3683	0.4226	1.0198	36.3739
	HDR-VDR-2 [92]	0.7315	0.7169	0.6643	0.7827	0.7924	0.6212	0.8444	0.8736	0.5861	0.9459	0.9334	0.3037	0.3986	0.5306	1.0368	89.2621
RR LFIQA	LF-IQM [60]	0.3620	0.3438	0.8930	0.4260	0.2700	0.7966	0.3601	0.2960	1.0046	0.4343	0.3097	0.8175	0.4457	0.4287	0.9950	589.7851
	RRLFIQA-4DDWT [61]	0.8305	0.8052	0.5439	0.6917	0.6741	0.6539	0.8325	0.8498	0.6072	0.9589	0.9445	0.2645	0.6457	0.6056	0.8636	128.1106
FR LFIQA	MDFM [53]	0.7303	0.6768	0.6625	0.8444	0.8138	0.4749	0.8275	0.8543	0.6149	0.9164	0.9054	0.3709	0.5725	0.5282	0.9276	0.8537
	Fang's [54]	0.8256	0.8077	0.5480	0.8589	0.8274	0.4571	0.8833	0.8856	0.5119	0.9730	0.9578	0.2135	0.5617	0.4950	0.9293	1.1574
	Min's [56]	0.7350	0.6645	0.6794	0.7112	0.6577	0.6476	0.8496	0.8470	0.5745	0.9757	0.9325	0.2051	0.5596	0.4094	0.9366	3.9845
	Meng's [57]	0.6924	0.6347	0.7001	0.8367	0.7819	0.4944	0.9282	0.9203	0.4037	0.9579	0.9383	0.2629	0.3043	0.3493	1.0593	30.4872
	KRIQE [59]	0.7380	0.6820	0.6538	0.8693	0.8377	0.4429	0.9582	0.9613	0.3091	0.9705	0.9502	0.2236	0.3500	0.3927	1.0410	115.7841
NR 2DIQA	GWH-GLBP [93]	0.4752	0.3391	0.8364	0.4947	0.3550	0.7780	0.6964	0.6102	0.7677	0.7709	0.7248	0.5622	0.5008	0.3493	0.9577	4.8472
	NIQE [94]	0.6246	0.4482	0.7584	0.4792	0.3701	0.7948	0.9187	0.8920	0.4247	0.8044	0.7053	0.5333	0.5642	0.4449	0.9405	6.0077
	BRISQUE [95]	0.6263	0.4559	0.7530	0.4969	0.3750	0.7910	0.9012	0.8747	0.4614	0.7701	0.6262	0.5674	0.5103	0.4179	0.9597	**3.1543**
NR LFIQA	BELIF [62]	0.5912	0.5119	0.7572	0.7161	0.6892	0.6291	0.8976	0.8671	0.4784	0.8099	0.7852	0.5437	0.7747	0.7072	0.7103	107.8814
	NR-LFQA [63]	0.6952	0.6275	0.6750	0.8327	0.8036	0.4895	0.9390	0.9347	0.3729	0.7687	0.7339	0.5686	0.6647	0.5864	0.8374	225.2069
	Tensor-NLFQ [66]	0.7595	0.7345	0.6327	0.7624	0.7261	0.5856	0.8649	0.8630	0.5424	0.8748	0.8553	0.4477	**0.8446**	**0.7887**	**0.6025**	697.6515
	PVRI [67]	0.7217	0.6827	0.6530	0.8017	0.7603	0.5271	0.9167	0.8771	0.4237	0.8338	0.7778	0.4831	0.6835	0.6037	0.8037	71.3578
	DSA [68]	0.7754	0.7328	0.6150	0.8482	0.8093	0.4578	0.9313	0.9195	0.3904	0.8449	0.8202	0.4882	0.7502	0.6902	0.7135	198.5443
	4D-DCT-LFIQA [69]	0.8267	0.8079	0.5512	0.8381	0.8213	0.4906	0.9400	0.9320	0.3691	0.8535	0.8308	0.4666	0.8206	0.7699	0.6411	169.2623
	TSSV-LFIQA [70]	0.6876	0.6186	0.6807	0.7876	0.7492	0.5388	0.9032	0.8856	0.4642	0.8616	0.8187	0.4655	0.7222	0.6781	0.7582	44.6696
	NR-LF-QAE [71]	0.6533	0.6239	0.6888	0.7444	0.6921	0.5808	0.9088	0.8788	0.4364	0.6013	0.5250	0.7572	0.7375	0.6788	0.7573	254.1088
	DeLFIQE [73]	0.4431	0.3860	0.8636	0.8063	0.7653	0.5220	0.6811	0.5765	0.8000	0.4907	0.4254	0.8049	0.6470	0.5809	0.8592	15.4193
	SATV-BLiF (ours)	0.7933	0.7704	0.5842	0.8515	0.8237	0.4686	0.9332	0.9284	0.3897	0.8750	0.8223	**0.4298**	0.6617	0.5543	0.8458	3.8812
	MPT-MWI-BLiF (ours)	**0.8701**	**0.8509**	**0.4714**	**0.8959**	**0.8710**	**0.3971**	**0.9403**	**0.9374**	**0.3685**	**0.8753**	**0.8616**	0.4426	0.8291	0.7667	0.6290	26.8342

evaluation.

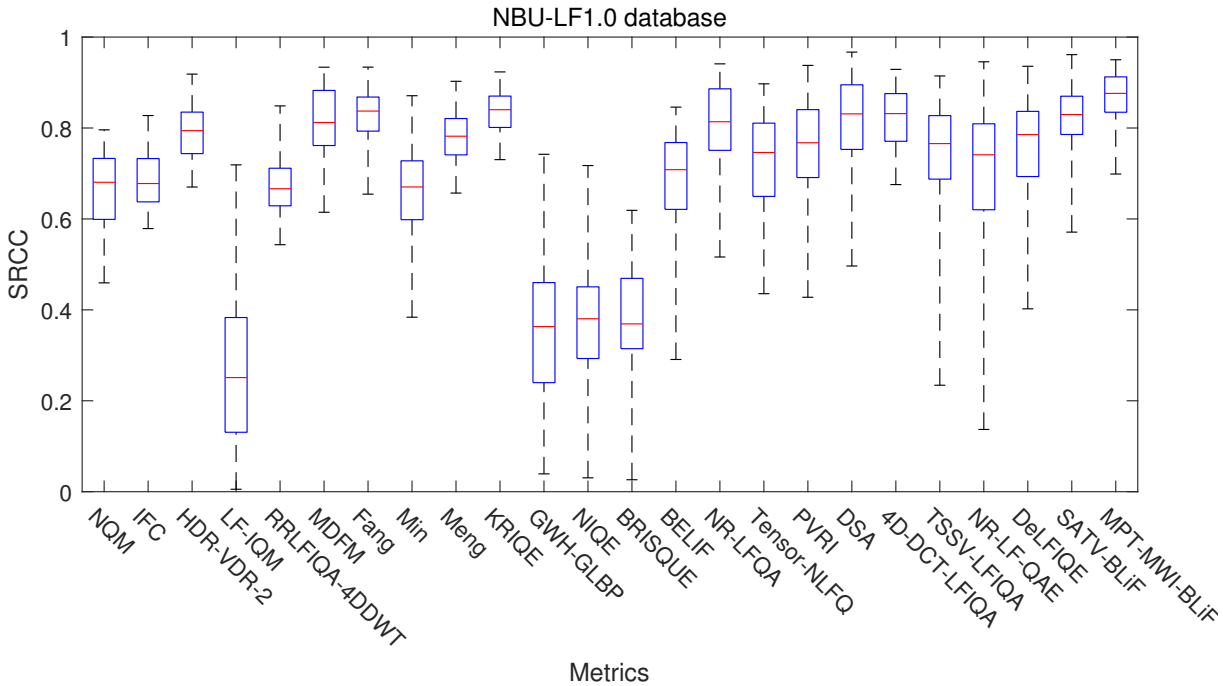
In addition to the overall performance, Table 3.1 also provides the time complexity of each metric. Here, we measure the time complexity by the runtime required to obtain the quality of a single distorted LFI. Note that we only report the runtime of each metric, excluding the time required for data loading and model initialization. For fair comparison, all metrics are tested with the same CPU-only hardware configuration. As shown in Table 3.1, the FR LFIQA metrics are generally more computational-friendly, because they allow a direct comparison of the difference between the distorted LFI and the reference LFI. In contrast, the RR/NR LFIQA metrics tend to have higher time complexity. However, the proposed SATV-BLiF metric only takes less than 4 seconds to complete the quality calculation of a LFI. Combined with its state-of-the-art quality assessment performance and ability to be independent of reference information, this metric is ideally suitable for real-time scenarios. Although MPT-MWI-BLiF is not as efficient as SATV-BLiF, its time complexity is still much lower than most existing NR LFIQA metrics.

To further investigate the stability of the proposed metric and other state-of-the-art metrics in different testing scenarios, we provide the box plots of SRCC distributions using leave-two-fold-out cross-validation in Figure 3.7. The stability experiments are carried out on the Win5-LID and NBU-LF1.0 databases, with the train-test split number of 45 and 91, respectively. In the figure, each box denotes the SRCC distribution of one metric, in which higher box position and smaller box length represent superior predictive accuracy and stability, respectively. As exhibited in Figure 3.7, the prediction performance and stability of the proposed MPT-MWI-BLiF metric are significantly better than that of state-of-the-art metrics on the Win5-LID and NBU-LF1.0 databases. While the proposed SATV-BLiF metric also performs competitively in terms of both predictive accuracy and stability. All these results fully demonstrate the excellent stability of the two proposed metrics.

Additionally, we provide the scatter plots of the objective prediction versus the subjective score for more intuitive visualization, as shown in Figure 3.8. Due to the space constraint, we only show the scatter plots of NR LFIQA metrics on the Win5-LID database. It can be found that the quality predictions of the proposed MPT-MWI-BLiF metric are more in line with the subjective score, further demonstrating its effectiveness in quality assessment.



(a)



(b)

Figure 3.7 – Box plots of SRCC distributions. (a) Win5-LID database; (b) NBU-LF1.0 database.

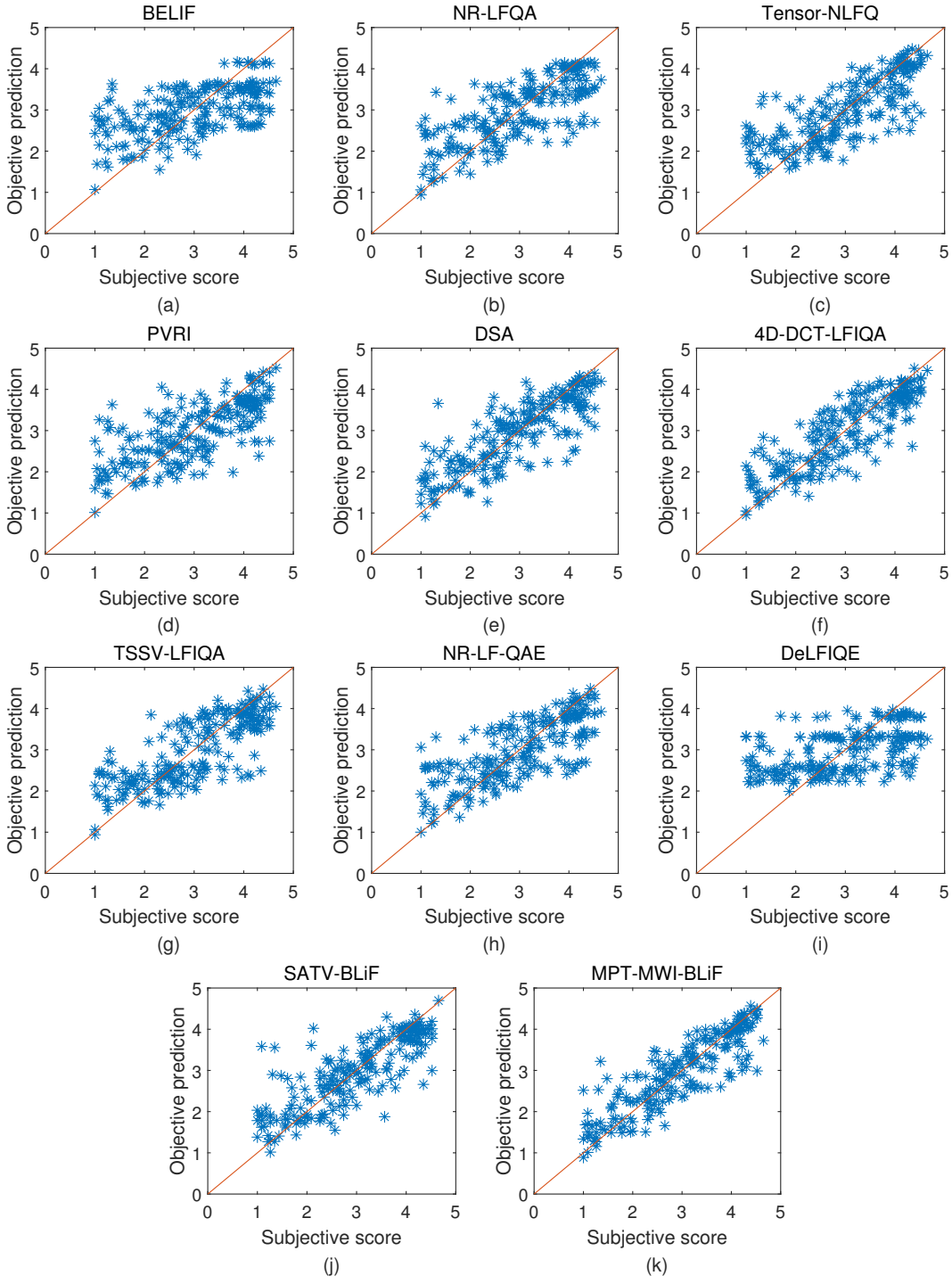


Figure 3.8 – Scatter plots of the objective prediction versus the subjective score of NR LFIQA metrics on the Win5-LID database.

3.3.3 Statistical Significance Test

F-test is an effective tool to study the statistical significance relationship between any two quality assessment metrics. Following [96], we first calculate the residuals between objective predictions and subjective scores for each metric, and then employ the *vartest*2 function in MATLAB to generate the significance relationship of any two metrics, with the default confidence level value of 95%. As shown in Figure 3.9, we report the F-test results among the LFIQA metrics on the Win5-LID and NBU-LF1.0 databases. Here, "-1", "0", and "1" indicate that the row metric is statistically worse, competitive, and better than the column metric, respectively. It can be seen that the proposed MPT-MWI-BLiF metric significantly outperforms all state-of-the-art LFIQA metrics on the Win5-LID and NBU-LF1.0 databases. Also, the proposed SATV-BLiF metric performs statistically better than most RR/NR LFIQA metrics, while achieving competitive statistical performance compared to FR LFIQA metrics.

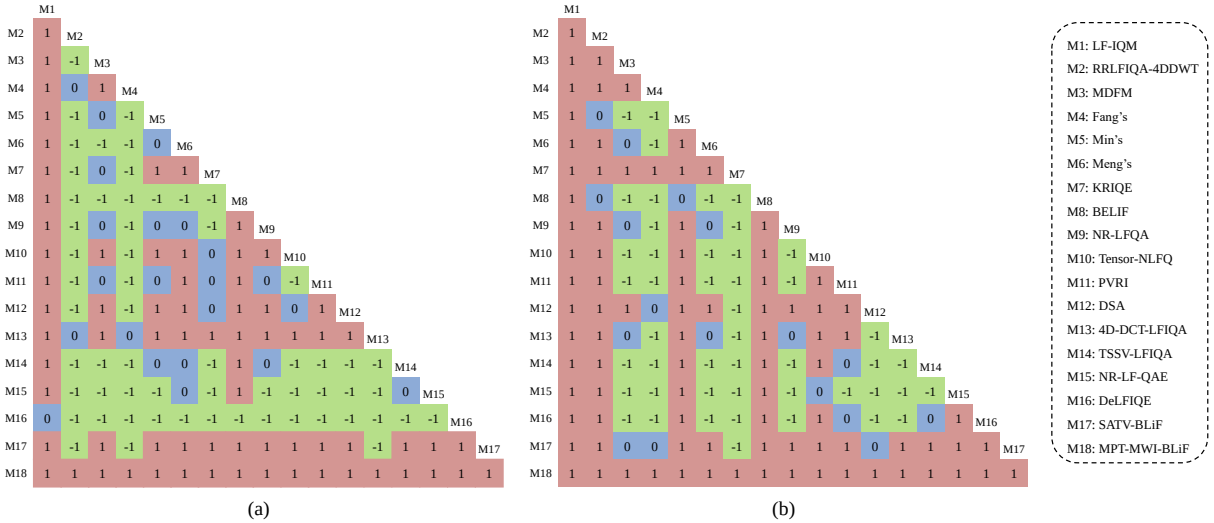


Figure 3.9 – Statistical significance between any two LFIQA metrics. (a) Win5-LID database; (b) NBU-LF1.0 database. "-1", "0", and "1" indicate that the row metric is statistically worse, competitive, and better than the column metric, respectively.

3.3.4 Performance on Cross-Database Validation

In order to verify the generalization performance of the proposed SATV-BLiF metric and MPT-MWI-BLiF metric, cross-database experiments are performed in this subsection. Specifically, considering that the Win5-LID database contains rich image categories

Table 3.2 – Performance of training on the Win5-LID database, and testing on the NBU-LF1.0, VALID10bit, and SHU databases.

Metric	Test Database	PLCC	SRCC	RMSE
SATV-BLiF	NBU-LF1.0 (NN)	0.8939	0.8682	0.3631
	VALID10bit (HEVC)	0.9271	0.8992	0.4008
	SHU (JPEG2k)	0.7675	0.7740	0.2902
MPT-MWI-BLiF	NBU-LF1.0 (NN)	0.7075	0.6799	0.5724
	VALID10bit (HEVC)	0.8242	0.8299	0.6057
	SHU (JPEG2k)	0.6255	0.6380	0.3532

and distortion types, all distorted LFIs in this database are used to train the model. Then, the LFIs with the same distortion types as Win5-LID on the NBU-LF1.0 (NN), VALID10bit (HEVC), and SHU (JPEG2k) databases are used to report the performance. Table 3.2 provides the experimental results of cross-database validation.

It can be found that the SATV-BLiF metric achieves excellent performance even when trained on another database, which fully verifies its generality and robustness. However, the cross-database performance of the MPT-MWI-BLiF metric is somewhat unsatisfactory compared to the SATV-BLiF metric. A possible reason is that LFIs are captured and processed differently for each database, the additional wavelet information considered in the MPT-MWI-BLiF metric may vary significantly across databases, which may in turn affect the cross-database performance.

3.3.5 Performance on Individual Distortion Type

Since LFI may suffer from different types of distortions during the image processing chain, we aim to investigate the robustness against distortion types in this subsection. The experimental results of PLCC, SRCC, and RMSE for individual distortion type on the Win5-LID and NBU-LF1.0 databases are provided in Table 3.3, Table 3.4, and Table 3.5, respectively. Here, "hit-count" tallies the number of times each metric obtaining a top-two result. Two CNN-based distortions in the Win5-LID database are not considered here since they have only one distortion level.

From the results, it can be observed that the FR LFIQA metrics typically perform better than the NR LFIQA metrics on a single distortion. A possible explanation is that the availability of reference information can greatly help to distinguish the same distortion type but different distortion levels. In addition, we can also observe that the perfor-

mance of most NR LFIQA metrics exhibits significant variations as the type of distortion changes. However, the proposed MPT-MWI-BLiF metric robustly achieves satisfactory performance in most distortion types. Further, in certain types of distortions, our MPT-MWI-BLiF metric significantly outperforms state-of-the-art NR LFIQA metrics. In some cases, it achieves competitive or even better performance than all existing FR metrics, *e.g.*, the NN distortion in the NBU-LF1.0 database. All three Tables corresponding to three different criteria show similar results. The above analyses fully demonstrate the robustness of our MPT-MWI-BLiF metric.

Further, the proposed SATV-BLiF metric also performs well on most distortion types, except for the "Zhang" distortion in the NBU-LF1.0 database. This may due to the fact that the "Zhang" distortion is a depth estimation-based angular distortion, in which different regions in a scene have different angular reconstruction quality according to the estimated depth. However, our SATV-BLiF metric only considers the texture information which may not be sufficient to measure such a distortion. Overall, from the perspective of hit-count numbers, our SATV-BLiF metric remains very effective for most distortion types.

3.3.6 Ablation Studies

The proposed MPT-MWI-BLiF metric consists of three features, *i.e.*, $\mathcal{F}_{TV}$, $\mathcal{F}_{SG}$, and $\mathcal{F}_{CI}$, and their effectiveness on the performance is investigated in this subsection. Table 3.6 provides the PLCC, SRCC, and RMSE performance of each feature type on the Win5-LID and NBU-LF1.0 databases. From the Table, we can see that $\mathcal{F}_{SG}$ contributes the least, while both $\mathcal{F}_{TV}$ and $\mathcal{F}_{CI}$ are very effective for quality evaluation. One possibility for this is related to the feature length. $\mathcal{F}_{SG}$ is a 15-dimensional feature vector, while both $\mathcal{F}_{TV}$ and $\mathcal{F}_{CI}$ have 35 dimensions. Intuitively, fewer features lack in providing a comprehensive description of the complex distortions in LFIs. Another possible explanation is that $\mathcal{F}_{SG}$ is extracted from the average of the edge maps of 15 first-level 4D-DHWT subbands, which only reflects the basic distortion information of LFIs. In comparison, $\mathcal{F}_{TV}$ captures holistic textural variations from all two-dimensional planes of LFIs, and $\mathcal{F}_{CI}$ contains rich frequency information from 255 second-level 4D-DHWT subbands. Therefore, $\mathcal{F}_{TV}$ and $\mathcal{F}_{CI}$ can better simulate the quality perception of human eyes. Overall, the combination of these three feature types culminates in enhanced effectiveness.

Moreover, considering that $\mathcal{F}_{TV}$ is composed of textural variations from six planes of LFIs, we conduct corresponding experiments to exploit the effectiveness of using dif-

Table 3.3 – PLCC performance of different distortion types on the Win5-LID and NBU-LF1.0 databases. The best two results of RR/FR metrics are in blue, while the best two results of NR metrics are marked in **bold**.

	Database	Win5-LID				NBU-LF1.0					
Type	Metric	HEVC	JPEG2k	LN	NN	NN	BI	EPICNN	Zhang	VDSR	Hit-count
FR 2DIQA	NQM [90]	0.9320	0.8701	0.7721	0.5927	0.8331	0.9300	0.8868	0.8810	0.7350	0
	IFC [91]	0.9331	0.8547	0.7719	0.8714	0.9195	0.9348	0.8811	0.7824	0.9326	0
	HDR-VDR-2 [92]	0.9367	0.8766	0.8236	0.8973	0.9397	0.9502	0.8832	0.8762	0.8957	1
RR LFIQA	LF-IQM [60]	0.6172	0.7263	0.8284	0.7158	0.5026	0.6764	0.5976	0.5296	0.5517	0
	RRLFIQA-4DDWT [61]	0.9741	0.9254	0.8804	0.9320	0.8904	0.9545	0.9752	0.9198	0.9762	4
FR LFIQA	MDFM [53]	0.9446	0.8878	0.8819	0.9004	0.9327	0.9463	0.9483	0.9436	0.9226	2
	Fang's [54]	0.9781	0.9365	0.9509	0.9252	0.8861	0.9649	0.9716	0.9064	0.9730	6
	Min's [56]	0.9777	0.9465	0.8798	0.7775	0.8033	0.8796	0.7890	0.8466	0.9664	2
	Meng's [57]	0.9129	0.7442	0.9486	0.8915	0.8989	0.9550	0.9734	0.7308	0.9652	2
	KRIQE [59]	0.9591	0.9185	0.9176	0.9033	0.9021	0.9616	0.9629	0.8399	0.9647	1
NR 2DIQA	GWH-GLBP [93]	0.7104	0.6002	0.6442	0.4606	0.5709	0.5950	0.7109	0.5233	0.6913	0
	NIQE [94]	0.8361	0.7317	0.6669	0.4331	0.4600	0.5225	0.7018	0.5357	0.9171	0
	BRISQUE [95]	0.8280	0.7629	0.6731	0.5055	0.4981	0.5634	0.7097	0.6310	0.8969	0
NR LFIQA	BELIF [62]	0.8062	0.7275	0.7172	0.7219	0.9244	0.8732	0.6707	0.6886	0.8749	0
	NR-LFQA [63]	0.7641	0.8098	0.7731	0.7920	0.9544	**0.9519**	0.9157	0.7108	0.8850	1
	Tensor-NLFQ [66]	0.8909	0.8340	0.8543	0.8446	0.8517	0.9199	0.8395	0.7135	0.9223	0
	PVRI [67]	0.8352	0.8702	0.7677	0.7890	0.8396	0.8967	0.8551	0.8268	0.9149	0
	DSA [68]	0.8673	0.8647	0.8338	0.8542	0.9472	0.9243	0.8670	0.8010	**0.9342**	1
	4D-DCT-LFIQA [69]	**0.9001**	**0.9365**	**0.8803**	0.8534	0.9386	0.9389	0.8183	**0.9048**	**0.9317**	**5**
	TSSV-LFIQA [70]	0.8170	0.7761	0.7396	0.7828	0.8924	0.9147	0.8148	0.8108	0.8790	0
	NR-LF-QAE [71]	0.8273	0.7178	0.7500	0.6902	0.8427	0.8503	0.8001	0.6743	0.8027	0
	DeLFIQE [73]	0.6417	0.5808	0.5809	0.5012	0.9554	0.9420	0.9069	0.7671	0.8216	0
	SATV-BLiF (ours)	0.8518	**0.9226**	0.8713	**0.8993**	**0.9726**	0.9477	**0.9336**	0.7538	0.9037	4
	MPT-MWI-BLiF (ours)	**0.9429**	0.9078	**0.9188**	**0.9379**	**0.9725**	**0.9693**	**0.9379**	**0.9065**	0.9274	**7**

Table 3.4 – SRCC performance of different distortion types on the Win5-LID and NBU-LF1.0 databases. The best two results of RR/FR metrics are in blue, while the best two results of NR metrics are marked in **bold**.

	Database	Win5-LID				NBU-LF1.0					
Type	Metric	HEVC	JPEG2k	LN	NN	NN	BI	EPICNN	Zhang	VDSR	Hit-count
FR 2DIQA	NQM [90]	0.9344	0.8540	0.7467	0.6601	0.8037	0.9228	0.7587	0.8276	0.7529	1
	IFC [91]	0.9182	0.8111	0.7818	0.8697	0.9200	0.9200	0.8054	0.6136	0.9143	1
	HDR-VDR-2 [92]	0.9568	0.8736	0.8578	0.9062	0.9213	0.9275	0.7831	0.8384	0.8986	4
RR LFIQA	LF-IQM [60]	0.2438	0.5331	0.6977	0.5938	0.3293	0.5164	0.4264	0.4007	0.3382	0
	RRLFIQA-4DDWT [61]	0.9648	0.8999	0.8211	0.8912	0.8276	0.9205	0.8358	0.8182	0.9432	3
FR LFIQA	MDFM [53]	0.9342	0.8378	0.8181	0.8491	0.8681	0.8942	0.7421	0.8432	0.8796	1
	Fang's [54]	0.9679	0.9121	0.8844	0.8869	0.8198	0.9149	0.7908	0.8001	0.9288	3
	Min's [56]	0.9657	0.9188	0.7289	0.4798	0.6402	0.7790	0.5657	0.7786	0.9344	3
	Meng's [57]	0.8878	0.6939	0.8459	0.8001	0.7568	0.9018	0.8034	0.5783	0.9225	0
	KRIQE [59]	0.9494	0.8689	0.8602	0.8584	0.8430	0.9190	0.8185	0.7242	0.9275	2
NR 2DIQA	GWH-GLBP [93]	0.6573	0.4173	0.4514	0.3392	0.3284	0.4753	0.5975	0.4559	0.6007	0
	NIQE [94]	0.8099	0.6929	0.5195	0.2821	0.3028	0.3485	0.5630	0.3990	0.8577	0
	BRISQUE [95]	0.5931	0.6131	0.4437	0.3259	0.3407	0.3568	0.5769	0.4503	0.8262	0
NR LFIQA	BELIF [62]	0.7760	0.5595	0.5552	0.5447	0.8769	0.8014	0.4816	0.4767	0.8085	0
	NR-LFQA [63]	0.7231	0.7356	0.6371	0.6893	0.9141	0.9027	**0.7585**	0.6194	0.8457	1
	Tensor-NLFQ [66]	0.8325	0.7740	0.7831	0.7606	0.8041	0.8747	0.6655	0.5861	0.8738	0
	PVRI [67]	0.7968	0.8323	0.6943	0.6714	0.7656	0.7992	0.7354	0.7339	0.8686	0
	DSA [68]	0.8449	0.8241	0.7429	0.7908	0.8990	0.8707	0.7215	0.6688	**0.8869**	1
	4D-DCT-LFIQA [69]	**0.8698**	**0.8946**	**0.8127**	0.8235	0.9027	0.9084	0.6645	**0.8446**	**0.8929**	**5**
	TSSV-LFIQA [70]	0.7471	0.6355	0.6437	0.7124	0.8414	0.8468	0.6630	0.6550	0.8023	0
	NR-LF-QAE [71]	0.7982	0.5769	0.6072	0.6120	0.7761	0.7708	0.6103	0.5824	0.7491	0
	DeLFIQE [73]	0.4628	0.4162	0.4337	0.3507	0.9024	0.8916	0.7302	0.6643	0.7192	0
	SATV-BLiF (ours)	0.7988	0.8766	0.7485	**0.8540**	**0.9282**	**0.9190**	**0.7722**	0.6498	0.8617	4
	MPT-MWI-BLiF (ours)	**0.9142**	**0.8813**	**0.8429**	**0.8867**	**0.9326**	**0.9304**	0.7347	**0.8285**	0.8851	**7**

Table 3.5 – RMSE performance of different distortion types on the Win5-LID and NBU-LF1.0 databases. The best two results of RR/FR metrics are in blue, while the best two results of NR metrics are marked in **bold**.

	Database	Win5-LID				NBU-LF1.0					
Type	Metric	HEVC	JPEG2k	LN	NN	NN	BI	EPICNN	Zhang	VDSR	Hit-count
FR 2DIQA	NQM [90]	0.3608	0.3976	0.5261	0.5750	0.4607	0.3475	0.3603	0.3261	0.6579	0
	IFC [91]	0.3552	0.3874	0.5186	0.3525	0.2978	0.3347	0.3608	0.3317	0.3158	0
	HDR-VDR-2 [92]	0.3680	0.4966	0.4652	0.3162	0.2515	0.2954	0.3638	0.2820	0.3934	1
RR LFIQA	LF-IQM [60]	0.8104	0.5730	0.3844	0.4757	0.6264	0.6881	0.5833	0.4547	0.7641	0
	RRLFIQA-4DDWT [61]	0.2129	0.3230	0.3707	0.2387	0.2915	0.2450	0.1536	0.1860	0.1784	4
FR LFIQA	MDFM [53]	0.2916	0.3789	0.3387	0.2728	0.2462	0.2793	0.2123	0.1559	0.3141	2
	Fang's [54]	0.2084	0.2749	0.2423	0.2597	0.3023	0.2268	0.1509	0.2136	0.1885	7
	Min's [56]	0.2122	0.2686	0.3691	0.4419	0.4011	0.4461	0.4240	0.2689	0.2130	2
	Meng's [57]	0.4013	0.5008	0.2470	0.3270	0.3186	0.2580	0.1610	0.3755	0.2180	1
	KRIQE [59]	0.2762	0.3179	0.2921	0.2879	0.2849	0.2301	0.1775	0.2615	0.2175	1
NR 2DIQA	GWH-GLBP [93]	0.7236	0.6117	0.5614	0.6187	0.5483	0.7549	0.5253	0.4851	0.6279	0
	NIQE []	0.5101	0.5285	0.5570	0.6362	0.6531	0.7923	0.5535	0.4930	0.3426	0
	BRISQUE [95]	0.5462	0.5023	0.5419	0.5994	0.6217	0.7618	0.5335	0.4307	0.3733	0
NR LFIQA	BELIF [62]	0.5713	0.5714	0.5111	0.4511	0.2719	0.4341	0.5565	0.4033	0.3935	0
	NR-LFQA [63]	0.6431	0.4904	0.4860	0.4028	0.2044	**0.2608**	0.2681	0.3857	0.4020	1
	Tensor-NLFQ [66]	0.4565	0.4853	0.3637	0.3737	0.3402	0.3402	0.3739	0.3727	0.3250	0
	PVRI [67]	0.5171	0.4045	0.4444	0.4026	0.3682	0.4042	0.3845	0.3078	0.3346	0
	DSA [68]	0.4853	0.4101	0.4105	0.3139	0.2129	0.3310	0.3330	0.2987	0.3184	0
	4D-DCT-LFIQA [69]	**0.4277**	**0.2783**	**0.3568**	0.3415	0.2366	0.3185	0.4184	**0.2206**	**0.3142**	**5**
	TSSV-LFIQA [70]	0.5768	0.4910	0.4647	0.3985	0.3142	0.3571	0.3788	0.3408	0.4254	0
	NR-LF-QAE [71]	0.5510	0.5849	0.4388	0.4679	0.3680	0.4572	0.4198	0.4203	0.5171	0
	DeLFIQE [73]	0.7984	0.6995	0.6223	0.6090	0.2023	0.2900	0.3155	0.3343	0.5052	0
	SATV-BLiF (ours)	0.5326	**0.3363**	0.3616	**0.2929**	**0.1618**	0.2859	**0.2544**	0.3738	0.3590	4
	MPT-MWI-BLiF (ours)	**0.3340**	0.3490	**0.3002**	**0.2345**	**0.1652**	**0.2173**	**0.2418**	**0.2258**	**0.3114**	**8**

Table 3.6 – Ablation studies of different feature combinations on the Win5-LID and NBU-LF1.0 databases. The best results are in **bold**.

Features	Win5-LID			NBU-LF1.0		
	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
$\mathcal{F}_{TV}$	0.8382	0.8195	0.5286	0.8592	0.8210	0.4578
$\mathcal{F}_{SG}$	0.7605	0.7492	0.6420	0.6702	0.6257	0.6862
$\mathcal{F}_{CI}$	0.8185	0.7893	0.5323	0.8881	0.8572	0.4068
$\mathcal{F}_{MPT-MWI-BLiF}$	**0.8701**	**0.8509**	**0.4714**	**0.8959**	**0.8710**	**0.3971**

Table 3.7 – Performance of different plane combinations for holistic textural variation features on the Win5-LID and NBU-LF1.0 databases. The best results are in **bold**.

Plane	Win5-LID			NBU-LF1.0		
	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
HW	0.5436	0.4504	0.8017	0.4082	0.3066	0.8263
UH	0.7717	0.7416	0.6264	0.8248	0.7858	0.5063
VH	0.6898	0.6693	0.6912	0.7353	0.6814	0.6119
UW	0.7647	0.7318	0.6133	0.7196	0.6751	0.6314
VW	0.7759	0.7545	0.6093	0.8267	0.7848	0.5155
UV	0.7691	0.7346	0.6222	0.6911	0.6210	0.6468
UHW	0.7882	0.7644	0.5826	0.8520	**0.8210**	0.4654
VHW	0.8235	0.8091	0.5475	0.8389	0.8028	0.4953
UVH	0.8148	0.7998	0.5544	0.8420	0.8132	0.4798
UVW	0.8129	0.7896	0.5585	0.8115	0.7810	0.5329
$UVHW$	**0.8382**	**0.8195**	**0.5286**	**0.8592**	**0.8210**	**0.4578**

ferent plane combinations. The experiments are performed on the Win5-LID and NBU-LF1.0 databases, as shown in Table 3.7. In the "Plane" column of the Table, the two-dimensional plane denotes the individual plane, *e.g.*, HW denotes the HW plane; the three-dimensional plane denotes the combination of three involved individual planes, *e.g.*, UHW denotes the combination of HW, UW, and UH planes; the four-dimensional plane $UVHW$ denotes the proposed metric with six individual planes. Table 3.7 shows that generating $\mathcal{F}_{TV}$ with all six planes consistently yields the best performance compared to any other plane combination. These results demonstrate the effectiveness of capturing the holistic textural variations for quality assessment.

3.4 Conclusion

In this chapter, we propose two NR LFIQA metrics SATV-BLiF and MPT-MWI-BLiF based on handcrafted features. Specifically, the SATV-BLiF metric adopts the LBP

descriptor to capture spatio-angular textural variations and then convert them into statistical textural histogram features for LFI quality assessment. The MPT-MWI-BLiF metric is considered an updated version of the SATV-BLiF metric, in which we first convert LFIs into six individual planes by pairing any two sub-dimensions of LFIs, and further exploit the textural variations of each plane by the LBP descriptor. As a result, the histogram of textural variations for all planes is used to capture the holistic texture information of LFIs. In addition, we decompose LFIs into multilevel wavelet subbands using 4D-DHWT, with the aim of simulating the visual perception process. After that, subband geometric features and coefficient intensity features from the first-level and second-level 4D-DHWT subbands are extracted to capture the wavelet information of LFIs. To verify the effectiveness of the two proposed metrics, we conduct quantitative experiments on five publicly-available LFIQA databases and fully compare our two metrics with a variety of state-of-the-art metrics. Extensive experimental results show that the SATV-BLiF metric has much lower time complexity than most existing NR LFIQA metrics, while achieving competitive quality evaluation performance in most scenarios. The MPT-MWI-BLiF metric comprehensively improves the quality evaluation performance of the SATV-BLiF metric, except for the cross-database performance. In terms of time complexity, the acceptable time complexity of the MPT-MWI-BLiF metric further reveals the great possibility of its applications in real-world scenarios. In the next chapter, we propose four deep learning-based LFIQA metrics that achieve much better performance in quality assessment.

Chapter 4

Proposed Deep Learning-based Metrics for Light Field Images

With the rapid development of deep learning technology, people have witnessed the tremendous potential of quality assessment metrics based on deep features far exceeding those based on handcrafted features. For high-dimensional LFIs with complex visual perception, the powerful discriminative ability of deep learning methods will undoubtedly serve as a powerful tool to promote the development of LFIQA research. For this goal, in this chapter, we explore the possibility of combining deep learning technology with LFI quality evaluation. As a result, we propose three deep learning-based NR LFIQA metrics (DeeBLiF, PVBLiF, and ASEM-BLiF) and one deep learning-based FR LFIQA metric (EDDMF). These metrics design deep models from different characteristics of LFIs to accurately evaluate the LFI quality, which will be detailed in the following sections.

This chapter is organized into six sections: First of all, Section 4.1 presents the design concept and implementation of the proposed DeeBLiF metric; then, as the extension of the DeeBLiF metric, two proposed NR metrics PVBLiF and ASEM-BLiF are detailed in Section 4.2 and Section 4.3, respectively; afterwards, the DeeBLiF metric is further extended to a FR version EDDMF, which is described in Section 4.4; finally, the experimental results and discussions are provided in Section 4.5, followed by the conclusion of this chapter in Section 4.6.

4.1 Proposed DeeBLiF Metric

There are two critical problems faced when using Convolutional Neural Network (CNN) models in the LFIQA metric: 1) Insufficient LFI data for training a CNN model. A well-performing CNN model needs enough data for training, but most existing LFIQA databases contain only a few hundred LFIs, which are insufficient for training a CNN model that can handle complex distortions. 2) No CNN model specifically designed for

LFIQA. Although a generically designed CNN model has powerful feature extraction capabilities and can mine local associations in images, using such a generic CNN model for the quality evaluation of LFIs cannot fully exploit its potential.

To address these two problems simultaneously, we propose a patch-based NR LFIQA metric based on deep learning, called Deep Blind Light Field image quality assessment metric (DeeBLiF), as illustrated in Figure 4.1. First, DeeBLiF generates a large number of spatio-angular patches from each LFI, which contain both angular and spatial information. The first problem can be effectively addressed by training a CNN model with a significant number of patches. To deal with the second problem, we propose a two-stream CNN model specifically designed for LFIQA, which is capable of extracting angular and spatial information separately. Finally, the quality scores of all patches are predicted and further pooled into an overall LFI quality score.

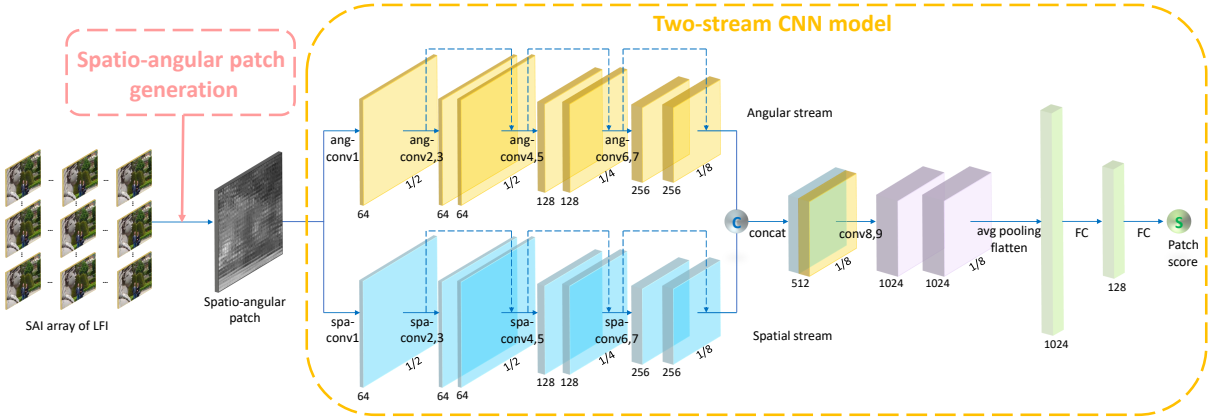


Figure 4.1 – Overall framework of the proposed DeeBLiF.

4.1.1 Spatio-Angular Patch Generation

Training a CNN model with patches is an effective approach to address the training data shortage in LFIQA, but the extracted patches need to highly reflect the characteristics of LFI, *i.e.*, containing both angular and spatial information. The diagram of the spatio-angular patch generation in DeeBLiF is shown in Figure 4.2. Given an input LFI with 9×9 SAIs, we first extract its central 5×5 SAIs to reduce computational complexity. Since the human eye is more sensitive to the luminance component Y in YCbCr color space, only the Y component is used for patch generation. After that, a 32×32 sliding window is exploited to slide evenly over each SAI with a stride of 64, and the same sam-

pling location of the central 5×5 SAIs are aggregated to form a sub-aperture patch. For example, all top-left red boxes of the central 5×5 SAIs in Figure 4.2 are aggregated to obtain a red sub-aperture patch. Finally, we convert sub-aperture patches into spatio-angular patches by clustering pixels at the same location in different SAIs, as shown at the bottom of Figure 4.2. As a result, by combining all 32×32 spatial locations with 5×5 angular pixels, each spatio-angular patch is of size 160×160.

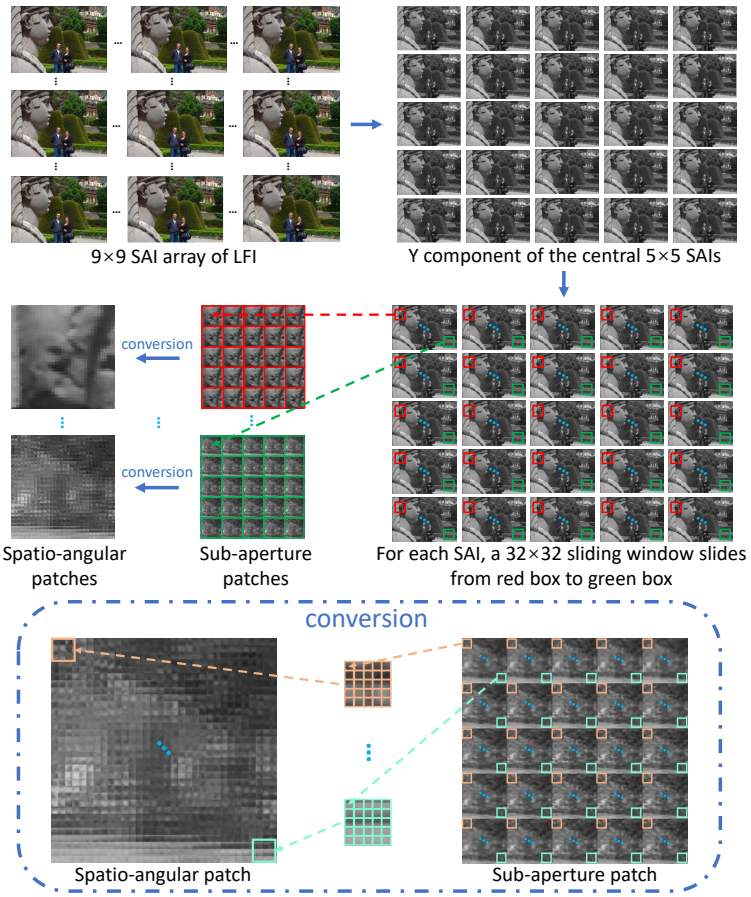


Figure 4.2 – Diagram of the spatio-angular patch generation.

4.1.2 Two-Stream CNN Model

To fully extract the information from spatio-angular patches, a two-stream CNN model is presented in DeeBLiF, as shown in Figure 4.1. The upper and lower streams, named angular and spatial streams, are designed to extract angular and spatial information, respectively. Each stream contains 7 convolutional layers and each convolutional layer is

followed by a batch normalization layer and a ReLU activation function. In addition, we add a shortcut before conv2, conv4, and conv6 for better training according to [97]. We only halve the feature map size in conv4 and conv6 to avoid losing too much information. Since adjacent pixels in spatio-angular patches are in the same position of adjacent SAIs, the classic convolutional layer using 3×3 kernels with 1 dilation is suitable for extracting the angular information from a spatio-angular patch. Therefore, all convolutional layers in the angular stream employ 3×3 kernels with 1 dilation. In addition, since spatio-angular patches are generated from the central 5×5 SAIs, subsampling the spatio-angular patch with a stride of 5 provides adjacent pixels in the same SAI. Based on this characteristic, the first convolutional layer (*i.e.*, spa-conv1) in the spatial stream exploits specifically designed 3×3 kernels with 5 dilations, which effectively extracts the spatial information from spatio-angular patches. Other than that, other convolutional configurations in the spatial stream are the same as those in the angular stream. After extracting angular and spatial information, the results of two streams are combined in the channel dimension, followed by two additional classic convolutional layers using 3×3 kernels with 1 dilation. Finally, an average pooling layer and two Fully Connected (FC) layers are utilized for regression to get the patch score.

4.2 Proposed PVBLiF Metric

This section provides the details of the proposed Pseudo Video-based Blind quality assessment metric for Light Field images (PVBLiF). As an extended version of DeeBLiF, PVBLiF aims to address two potential issues in DeeBLiF: 1) It is still unclear whether the LFI quality is well inherited by the widely-used 2D representations. Thus, we use the Pseudo Video Block Sequence (PVBS), a more intuitive 3D representation for quality evaluation. Based on the characteristics of PVBS, a CNN-based PVBSNet is proposed to fully extract the spatio-angular features and further evaluate the PVBS quality. 2) The average pooling used in the patch/block-based metric is not an optimal solution. Therefore, we propose a Saliency- and Variance-guided Pooling (SVPooling) method to reweight the importance of different PVBSs to obtain an overall LFI quality.

The overall framework of the proposed PVBLiF metric is illustrated in Figure 4.3. It mainly consists of three components: PVBS generation, PVBSNet, and SVPooling. Given a LFI represented in the SAI array form, we first transform it into PVS and then generate a large number of PVBSs. After that, the spatio-angular features of each PVBS

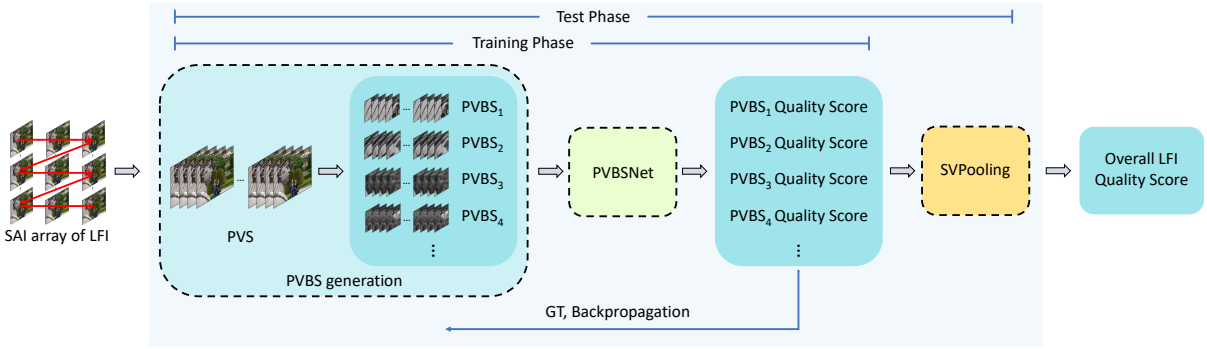


Figure 4.3 – Overall framework of the proposed PVBLiF metric.

are exploited and further quantified into a quality score. Finally, the SVPooling method is applied to integrate all the PVBS quality into an overall LFI quality. Note that PVBSNet is trained with PVBSs in an end-to-end manner, while the SVPooling method is used for post-processing in the test phase and no trainable parameters are involved.

4.2.1 Pseudo Video Block Sequence Generation

The 2D representations used in existing LFIQA metrics more or less ignore the primitive characteristics of LFIs. Instead, PVS is a 3D representation, which highly preserves the spatio-angular information of original LFIs. Derived from the idea of PVS, we exploit a 3D representation, named PVBS, to evaluate the LFI quality. The generation of PVBS and the motivation behind it are detailed below.

Let $\mathcal{L} \in \mathbb{R}^{U\times V\times H\times W\times C}$ denote the input LFI in the form of SAI array, where $U\times V$ and $H\times W$ are the angular and spatial resolutions, respectively, and C denotes the channel number. First, to reduce the computational complexity, the central $A\times A$ SAIs are extracted and further transformed into PVS (denoted as $\mathcal{P}_C \in \mathbb{R}^{(A*A)\times H\times W\times C}$), by arranging each SAI as a frame and displaying all SAIs in a certain order. Generally, there are several orders for creating a PVS [6]. In our implementation, we create PVS using raster order [6], *i.e.*, horizontally scanning the SAI array from left to right and row by row, starting from the SAI on the left superior corner. This generation order facilitates the subsequent extraction of spatio-angular features, which will be described in detail in the following subsection. Then, we convert $\mathcal{P}_C$ from RGB color space to YCbCr color space. After that, only the luminance component Y is used for subsequent feature extraction and quality prediction. Let $\mathcal{P}_{CY} \in \mathbb{R}^{(A*A)\times H\times W}$ denote the Y component of PVS, the above

operations are described as:

$$\mathcal{P}_C = PVS(Central(\mathcal{L})) \tag{4.1}$$

$$\mathcal{P}_{CY} = \{YCbCr(\mathcal{P}_C)\}_Y \tag{4.2}$$

where $Central(\cdot)$, $PVS(\cdot)$, and $YCbCr(\cdot)$ denote the selection of central SAIs, PVS transformation, and YCbCr color space conversion, respectively.

The generated PVS is a series of SAIs arranged in a specific order, where individual SAIs contain the spatial information and the changes between different SAIs reflect the angular information. However, it is difficult to evaluate the LFI quality directly using PVS, because the spatial variation is much greater than the angular variation, and the slight angular variation will be submerged. In addition, taking each PVS as an individual sample for training may suffer from the over-fitting problem since such a large amount of data requires deeper networks [72]. Therefore, we split PVS into a large number of non-overlapping blocks in the spatial domain. Each generated block is named PVBS and denoted as $\mathcal{PB} \in \mathbb{R}^{(A*A)\times S\times S}$, where $A*A$ and $S\times S$ are the length and the spatial resolution, respectively. The generation of PVBS is described as:

$$\mathcal{P}_{CY} = \{\mathcal{PB}_1, \mathcal{PB}_2, ..., \mathcal{PB}_K\} \tag{4.3}$$

where $\mathcal{PB}_k$ denotes the k-th PVBS and K is the total number of PVBSs.

In PVBLiF, A and S are set to 5 and 32 to achieve the best trade-off between precision and efficiency. Although each PVBS contains only a small amount of information derived from the original LFI, it is still a 3D representation with reduced spatial information of PVS while preserving the structural information of PVS.

4.2.2 Pseudo Video Block Sequence Network

Previous studies on LFIQA [63], [67] have demonstrated that the potential interaction between spatial and angular information, called spatio-angular information, is more closely related to the LFI quality than the spatial information. Although the 3D structure of PVBSs has great potential in simulating how LFIs are presented to human eyes, two challenges are encountered when extracting spatio-angular features for quality assessment. First, during the PVBS generation process, the original 2D angular information of LFIs is scrambled as it is transformed into the 1D temporal information of PVBSs. Second,

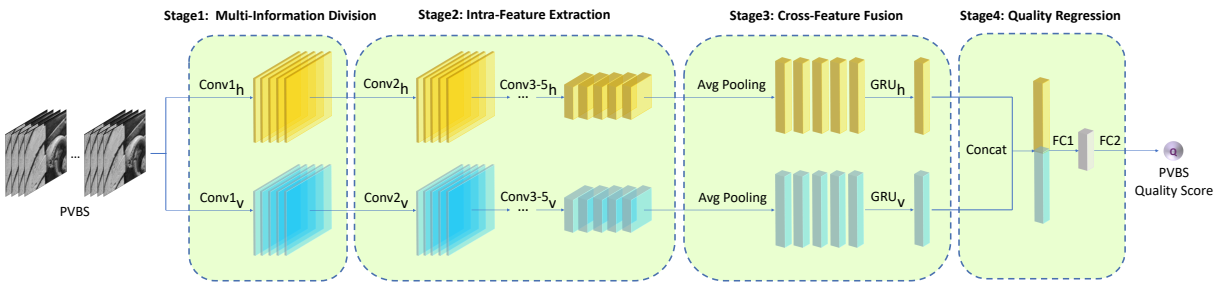


Figure 4.4 – Framework of the proposed PVBSNet.

extracting spatio-angular features from a 3D representation to evaluate quality in an end-to-end manner is non-trivial.

To address the above challenges, we propose a CNN-based network for PVBSs, named PVBSNet, to extract the spatio-angular features and further evaluate the PVBS quality. Figure 4.4 illustrates the framework of the proposed PVBSNet, where the top and bottom streams are used for extracting the horizontal and vertical angular features, respectively. In addition, the proposed network is systematically divided into four stages according to different functions: multi-information division, intra-feature extraction, cross-feature fusion, and quality regression. Note that the network is trained and tested in an end-to-end manner, and its configurations are detailed in Table 4.1. The four stages will be introduced below.

Stage 1: Multi-information division. From the original LFI to the generated PVS in raster order, the relationship between horizontal adjacent views is stored while the relationship between vertical adjacent views is destroyed. For this reason, video quality assessment methods are not suitable for PVS, nor for PVBS. The multi-information division stage is proposed to handle this challenge, which aims to divide the scrambled temporal information of PVBSs into horizontal and vertical information. To achieve this goal, we notice that in the PVBS, the horizontal angular information is recorded in the A adjacent views, while the vertical angular information is recorded in every A views, where $A\times A$ denotes the original LFI angular resolution. Thus, we propose to adopt special 3D CNN kernels to separate the regular temporal information into horizontal and vertical angular information, as shown in Figure 4.5. Specifically, based on the raster order used in generating PVBS, a 3D convolutional kernel with $A\times1\times1$ size and $A\times1\times1$ stride is used for extracting horizontal angular features, while a 3D convolutional kernel with $A\times1\times1$ size and $A\times1\times1$ dilation is used for extracting vertical angular features. As a result, horizontal and vertical angular features are extracted from the scrambled

Table 4.1 – Network configuration of the proposed PVBSNet.

Stage	Layer Name	Output Size	Kernel Size	Dilation	Stride	Padding
Stage 1	Conv1_{h}	$64\times A\times S\times S$	$A\times1\times1$	$1\times1\times1$	$A\times1\times1$	$0\times0\times0$
			Batch Normalization, Leaky ReLU			
	Conv1_{v}	$64\times A\times S\times S$	$A\times1\times1$	$A\times1\times1$	$1\times1\times1$	$0\times0\times0$
			Batch Normalization, Leaky ReLU			
Stage 2	$\text{Conv2}_{\text{h/v}}$	$64\times A\times S\times S$	$1\times3\times3$	$1\times1\times1$	$1\times1\times1$	$0\times1\times1$
			Batch Normalization, Leaky ReLU			
	$\text{Conv3}_{\text{h/v}}$	$128\times A\times S\times S$	$1\times3\times3$	$1\times1\times1$	$1\times1\times1$	$0\times1\times1$
			Batch Normalization, Leaky ReLU			
	$\text{Conv4}_{\text{h/v}}$	$256\times A\times \frac{S}{2}\times \frac{S}{2}$	$1\times3\times3$	$1\times1\times1$	$1\times2\times2$	$0\times1\times1$
			Batch Normalization, Leaky ReLU			
	$\text{Conv5}_{\text{h/v}}$	$256\times A\times \frac{S}{2}\times \frac{S}{2}$	$1\times3\times3$	$1\times1\times1$	$1\times1\times1$	$0\times1\times1$
			Batch Normalization, Leaky ReLU			
Stage 3	Avg Pooling	$256\times A$	$1\times \frac{S}{2}\times \frac{S}{2}$	$1\times1\times1$	$1\times1\times1$	$0\times0\times0$
	$\text{GRU}_{\text{h/v}}$	256×1	Gated Recurrent Unit (GRU)			
Stage 4	Concat	512×1	Concatenation layer			
	FC1	128×1	Fully Connected (FC) layer			
	FC2	1×1	Fully Connected (FC) layer			

temporal information of PVBSs. Given a PVBS $\mathcal{PB}$, let $\mathcal{F}^1$ represent the features output from stage 1, the above process is described as:

$$\mathcal{F}^1_m = Conv1_m(\mathcal{PB}) \tag{4.4}$$

where $m \in \{h, v\}$, h and v denote the horizontal and vertical angular streams, respectively. $Conv1_m(\cdot)$ denotes the 3D convolutional layer used in stage 1.

Stage 2: Intra-feature extraction. After separating horizontal and vertical angular information in stage 1, we obtain horizontal angular features $\mathcal{F}^1_h$ and vertical angular features $\mathcal{F}^1_v$. The resulting horizontal (or vertical) angular features include A slices, where the slice is defined as the spatio-horizontal (or spatio-vertical) features of all SAIs from a certain row (or column). As shown on the right side of Figure 4.5, feature maps in different colors are considered to be different slices. Considering that 2D CNNs aim to extract the spatial information and integrate it into the channel information, therefore, in the second stage, we propose to extract the spatio-angular features for each slice with a sequence of repeated 2D CNNs. Specifically, a 4-layer network with $1\times3\times3$ kernel size and $1\times1\times1$ dilation is used. To avoid losing too much spatial information, we only set the stride to $1\times2\times2$ to

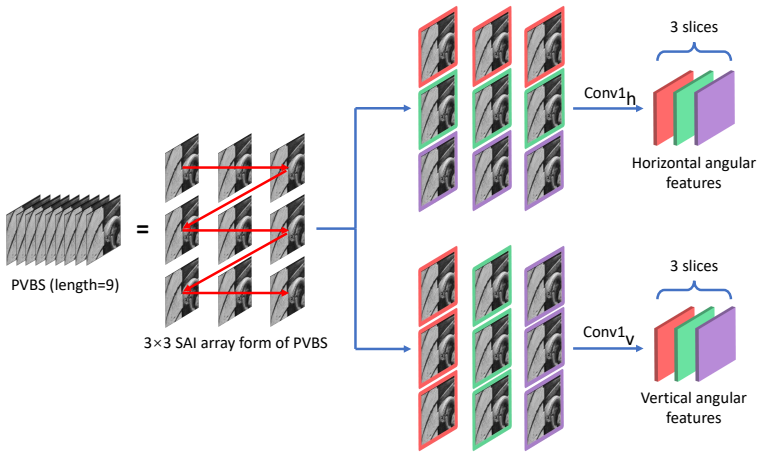


Figure 4.5 – Stage 1: Multi-information division. For better visualization, A is set to 3 for all illustrations with $A \times A$ angular resolution.

halve the spatial size in $Conv4_{h/v}$. The operations adopted in stage 2 are described as:

$$\mathcal{F}_m^N = ConvN_m(\mathcal{F}_m^{N-1}) \tag{4.5}$$

where $\mathcal{F}_m^N$ denotes the spatio-angular features output from the N-th 2D convolutional layer $ConvN_m(\cdot)$, $N \in \{2, 3, 4, 5\}$. Specifically, $\mathcal{F}_m^5$ represents the output of stage 2.

Stage 3: Cross-feature fusion. The third stage aims to establish the relationship between the spatio-angular features of each slice and fuse them into unified features. Due to the inherently narrow baseline of different views in LFIs, there are only subtle differences between different slices. In addition, all slices are arranged in order, *i.e.*, from the first row (or column) to the last row (or column). Therefore, inspired by [98], we consider it a temporal-memory issue and propose to model the long-term dependencies for fusing. Here, we perform the Gated Recurrent Unit (GRU) [99], a recurrent neural network model with gate control, to integrate the features of different slices into unified features. Before adopting the GRU model, we convert each spatio-angular feature map into a feature vector using an average pooling layer. Let $\mathcal{F}^G$ represent the feature vector output from stage 3, the process of stage 3 is described as:

$$\mathcal{F}_m^G = GRU_m(AP(\mathcal{F}_m^5)) \tag{4.6}$$

where $GRU_m(\cdot)$ denotes the GRU model, $AP(\cdot)$ denotes the average pooling layer.

Stage 4: Quality regression. In the final stage, we quantify the extracted spatio-angular

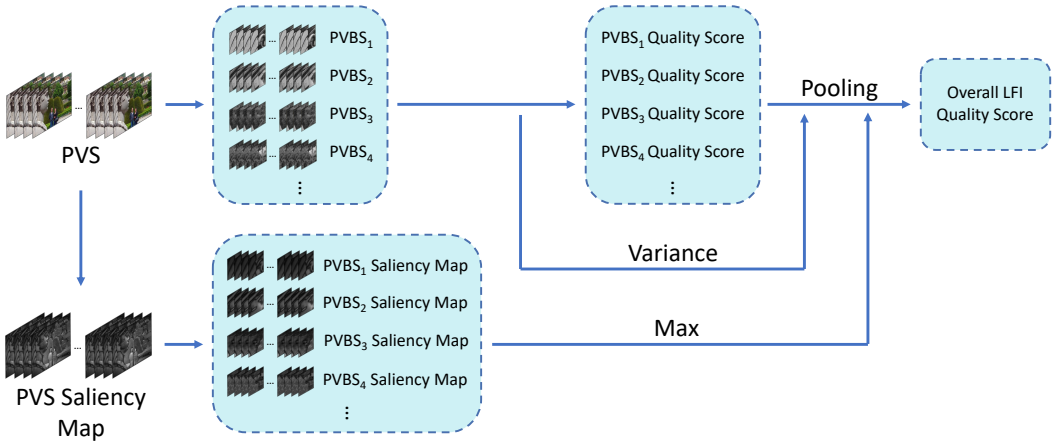


Figure 4.6 – Pipeline of the proposed SVPooling method.

features into the quality score. Specifically, the concatenation operation is applied for $\mathcal{F}_h^G$ and $\mathcal{F}_v^G$, followed by two FC layers to predict the quantified quality score. The generation of PVBS quality score Q_{PVBS} is described as:

$$Q_{PVBS} = FC2(FC1(Concat(\mathcal{F}_h^G, \mathcal{F}_v^G))) \tag{4.7}$$

where $Concat(\cdot)$, $FC1(\cdot)$, and $FC2(\cdot)$ represent the concatenation, first FC, and second FC layers, respectively.

4.2.3 Saliency- and Variance-Guided Pooling

The PVBS quality reflects the local quality of LFIs since it derives from the partial information of LFIs. To obtain the overall LFI quality, a pooling method is required. Our previous metric DeeBLiF employs average pooling to convert all local quality scores into an overall LFI quality score. In other words, all regions of the same scene are considered to contribute equally to the quality perception of human vision. However, as concluded in many previous studies (*e.g.*, [100], [101]), human eyes naturally pay more attention to certain regions when observing a scene. In addition, we consider that some regions do not have sufficient information to reflect the LFI quality degradation. Therefore, in the PVBLiF metric, we propose a Saliency- and Variance-guided Pooling (SVPooling) method to obtain the overall LFI quality, as illustrated in Figure 4.6.

1) Saliency. First, we reweight the contribution of different PVBSs according to their saliency maps. Here, saliency maps are generated by applying a saliency detection ap-

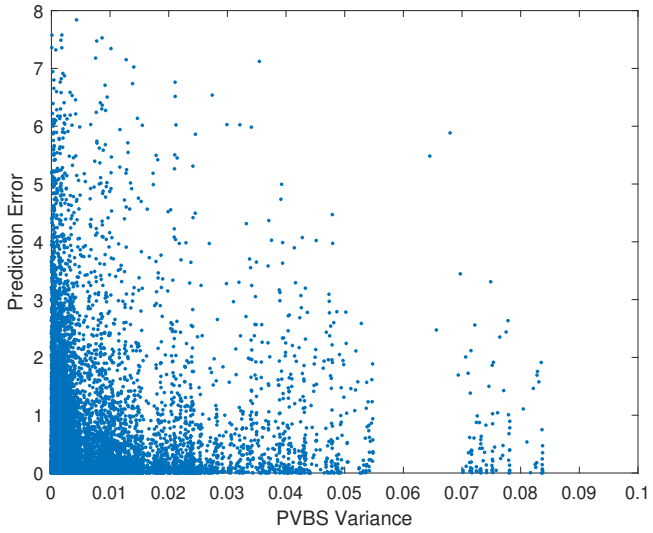


Figure 4.7 – Scatter plots of all PVBSs in terms of quality prediction error against variance on the Win5-LID database [2].

proach to all views of SAI array and converting them into the PVBS form. Specifically, we choose SDSP [102] for saliency detection. Further, human eyes pay more attention to a certain area, often because there are some noteworthy points in it. Based on this principle, we consider the maximum value of the saliency map as the PVBS weight. Let S_{PVBS} and W_{PVBS} denote the saliency map and weight of PVBS, respectively, the generation of W_{PVBS} is described as:

$$W_{PVBS} = Max\{S_{PVBS}\} \tag{4.8}$$

where $Max\{\cdot\}$ represents the maximum operation.

2) Variance. Then, we delve into the relationship between the quality prediction accuracy and the variance of PVBSs, by assuming that PVBSs with complex textures may better reflect their quality. To prove our conjecture, we conducted an experiment on the Win5-LID database [2] and visualized the experimental results as suggested in [103]. Figure 4.7 shows the scatter plots of all PVBSs in terms of quality prediction error against variance. Here, the prediction error is calculated as the square error between the quality prediction and its corresponding subjective quality score. It can be found that low-variance PVBSs are not reliable for evaluating the quality of PVBSs as their prediction errors are quite scattered. In other words, although some low-variance PVBSs have accurate predictions, it is difficult to discern the prediction quality of low-variance PVBSs. In contrast,

the error distribution of high-variance PVBSs is more concentrated in low prediction error, implying that high-variance PVBSs should be biased when pooling the overall LFI quality. For this reason, we propose to discard the low-variance PVBSs directly to avoid introducing inaccurate predictions from low-variance PVBSs. In this way, the predicted overall LFI quality will be dominated by the quality of high-variance PVBSs. In order to distinguish between low-variance and high-variance PVBSs, for simplicity, we divide all PVBSs into two equal parts sorted by variance. Let Ω denote the assemble of high-variance PVBSs, the generation of Ω is described as:

$$\mathcal{T} = Median\{V_{\mathcal{PB}_1}, V_{\mathcal{PB}_2}, ..., V_{\mathcal{PB}_K}\} \tag{4.9}$$

$$\mathcal{PB}_k \in \Omega, \quad if\ V_{\mathcal{PB}_k} > T \tag{4.10}$$

where $\mathcal{PB}_k$ and $V_{\mathcal{PB}_k}$ denote the k-th PVBS and its variance, respectively, $\mathcal{T}$ denotes the threshold, $Median\{\cdot\}$ represents the median operation.

3) Pooling. Finally, we obtain the overall LFI quality score using only high-variance PVBSs and their corresponding weights:

$$Q_{PVBLiF} = \frac{\sum_\Omega Q_{PVBS} * W_{PVBS}}{\sum_\Omega W_{PVBS}} \tag{4.11}$$

where Q_{PVBLiF} is the overall LFI quality score of the proposed PVBLiF metric. Note that no trainable parameters are involved in SVPooling, which is considered a post-processing method in our PVBLiF metric.

4.3 Proposed ASEM-BLiF Metric

Most existing LFIQA metrics, including our DeeBLiF metric, fail to establish a reasonable correlation between spatial and angular information in LFIs, especially ignoring the imbalance problem of large spatial variations and subtle angular variations. A likely result is that spatial information dominates feature extraction and quality degradation learning, further causing unsatisfactory quality assessment performance. To this end, in this section, we propose a novel Blind Light Field image quality assessment metric based on Angular-Spatial Effect Modeling, abbreviated as ASEM-BLiF.

Figure 4.8 presents the overall framework of the proposed ASEM-BLiF metric, which mainly consists of four components: Angular Effect Modeling (AEM), Angular-Spatial

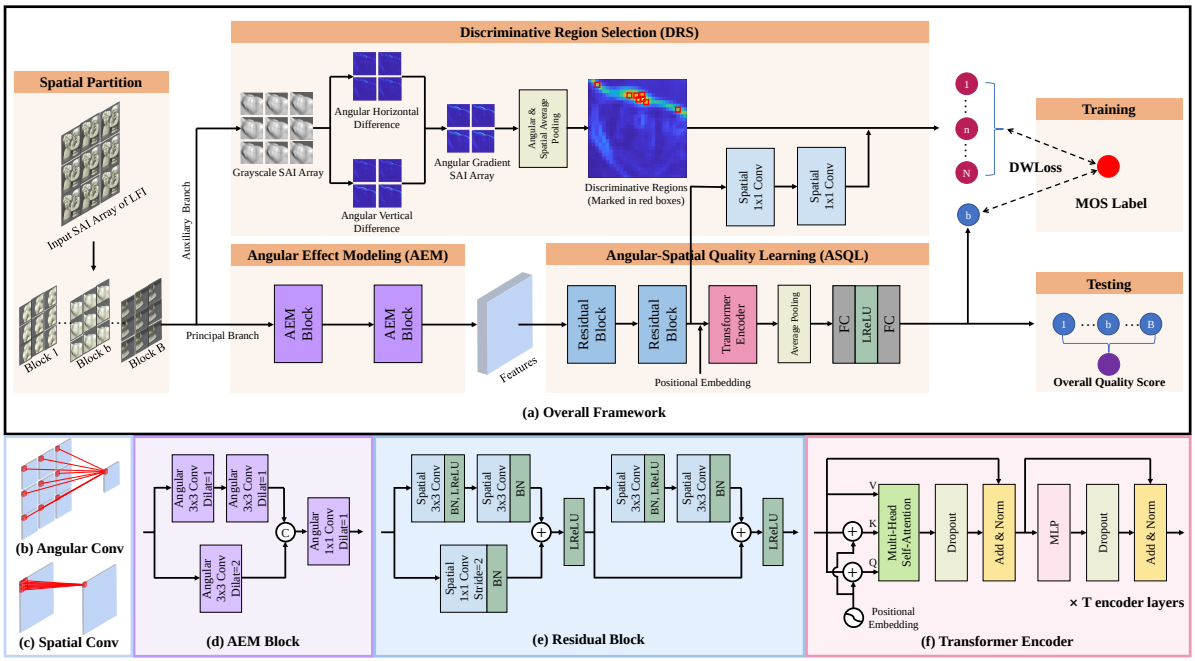


Figure 4.8 – Overall framework of the proposed ASEM-BLiF metric. For better visualization, the angular resolution of the input SAI array of LFI is set to 3×3.

Quality Learning (ASQL), Discriminative Region Selection (DRS), and Dynamic Weighting Loss (DWLoss). Given an input LFI $\mathcal{L} \in \mathbb{R}^{U\times V\times H\times W\times C}$, we first spatially partition $\mathcal{L}$ into B overlapping blocks to ensure a sufficient training set, denoted as $\mathcal{LB}_b \in \mathbb{R}^{U\times V\times S\times S\times C}$, $b = 1, 2, ..., B$. Then two branches are designed, the principal branch participates in both the training and testing stages, while the auxiliary branch is used only in the training stage. In the principal branch, the angular information of the input block is first captured by the AEM module, followed by the ASQL module for local angular-spatial effect modeling and quality prediction from a global perspective. In the auxiliary branch, the DRS module is utilized to selectively obtain discriminative regions with large angular sparsity, whose quality is subsequently predicted and exploited for auxiliary learning. During training, a DWLoss is further employed to achieve an optimal balance between the learning of principal and auxiliary branches. During testing, the overall LFI quality score is obtained by averaging the predicted quality score of all blocks.

4.3.1 Angular Effect Modeling

The subtle angular variation is one of the most distinctive characteristics of LFIs [42]. With the final goal of modeling angular-spatial effect for quality assessment, the design of the AEM module is inspired by the following two points to handle the angular information: First, to minimize the imbalance of large spatial variations and subtle angular variations, the raw angular information should be captured independent of spatial information. Second, angular hierarchical features extracted from different angular disparities are beneficial to encode the angular information with subtle variations.

As shown in Figure 4.8 (a), the proposed AEM module consists of two AEM blocks (Figure 4.8 (d)). In each AEM block, two sub-branches with angular convolutions (Figure 4.8 (b)) are designed in parallel to extract angular hierarchical features without considering any spatial information. Note that, due to the low angular resolution (typically 9×9) of LFIs, the padding of all angular convolutions is set to zero to avoid introducing irrelevant information. Specifically, as shown in Figure 4.8 (b), the top sub-branch encodes the small angular disparity by utilizing two 3×3 angular convolutions with a dilation rate of 1, while the bottom sub-branch uses a 3×3 angular convolution with a dilation rate of 2 to encode the large angular disparity. The motivation of these operations is to ensure that both sub-branches have the same output size and receptive field, which facilitates the subsequent combination. Finally, the output features of these two branches are concatenated and fused by a 1×1 angular convolution. Given an input block $\mathcal{LB} \in \mathbb{R}^{U\times V\times S\times S\times C}$, the above process is described as:

$$\mathcal{F}_A = \phi_A(\mathcal{LB}) \tag{4.12}$$

where $\mathcal{F}_A \in \mathbb{R}^{S\times S\times C}$ denotes the output features of the AEM module $\phi_A(\cdot)$.

4.3.2 Angular-Spatial Quality Learning

The main objective of the ASQL module is to model the angular-spatial effect and further predict the global quality. Firstly, two residual blocks (Figure 4.8 (e)) based on the widely-used ResNet [97] are performed to model the local angular-spatial effect. The motivation is that the angular information from multiple viewpoints is captured in the channel dimension of $\mathcal{F}_A$, so the angular-spatial effect can be modeled by conventional spatial convolutions (Figure 4.8 (c)) which simultaneously extract spatial and channel features. Note that such a feature extraction part can be easily extended by using other

mainstream backbones. Here, we use only a small number of spatial convolutions for feature extraction to prevent spatial information from dominating the angular-spatial effect. Let the above feature extraction denote as $\phi_L(\cdot)$, the generated local angular-spatial features $\mathcal{F}_L \in \mathbb{R}^{m \times m \times d}$ can be described as:

$$\mathcal{F}_L = \phi_L(\mathcal{F}_A) \tag{4.13}$$

Here, m is equal to $S/4$ since we adopt two residual blocks and each block halves the spatial resolution. d denotes the depth of feature maps.

However, using a small number of spatial convolutions with small receptive fields inevitably lacks the consideration of global content information [104]. Different local regions may have different angular-spatial effects [59], thus it is necessary to establish the relationship between the local regions and the global quality. To this end, we subsequently adopt the Transformer derived from DETR [105] to achieve this goal. Specifically, we convert $\mathcal{F}_L$ into a sequence of features with size $m^2 \times d$, where m^2 and d represent the sequence length and feature depth, respectively:

$$\mathcal{F}_L = [\mathcal{F}_1; \mathcal{F}_2, ...; \mathcal{F}_{m^2}], \quad \mathcal{F}_L \in \mathbb{R}^{m^2 \times d} \tag{4.14}$$

To mine the self-attention relationship between local regions, the Multi-head Self-Attention (MSA) [105] extended from the standard Query-Key-Value-Self-Attention (QKV-SA) architecture is adopted. First, the $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ matrices are separately obtained from the input sequence $\mathcal{F}_L$. Note that, since the position information of local regions is lost after the operation in Eq. (4.14), an extra learnable positional embedding $p_e \in \mathbb{R}^{m^2 \times d}$ is added to the sequence before generating the $\boldsymbol{Q}$ and $\boldsymbol{K}$ matrices [105]:

$$\boldsymbol{Q} = (\mathcal{F}_L + p_e)\boldsymbol{W}_q, \quad \boldsymbol{W}_q \in \mathbb{R}^{d \times d'} \tag{4.15}$$

$$\boldsymbol{K} = (\mathcal{F}_L + p_e)\boldsymbol{W}_k, \quad \boldsymbol{W}_k \in \mathbb{R}^{d \times d'} \tag{4.16}$$

$$\boldsymbol{V} = \mathcal{F}_L\boldsymbol{W}_v, \quad \boldsymbol{W}_v \in \mathbb{R}^{d \times d'} \tag{4.17}$$

where d' is set to d/Y, Y is the head number of MSA, $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v$ are the linear projection matrices of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, respectively. Then, let x_0 denote the combination

of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, we obtain $MSA(x_0)$ as:

$$SA_y(x_0) = Softmax(\boldsymbol{Q}\boldsymbol{K}^{\mathsf{T}}/\sqrt{d'})\boldsymbol{V} \tag{4.18}$$

$$MSA(x_0) = Concat(SA_1(x_0); ...; SA_Y(x_0))\boldsymbol{W}_{ln} \tag{4.19}$$

where $\boldsymbol{W}_{ln} \in \mathbb{R}^{d\times d}$ is the linear projection matrix used as the final step before obtaining the MSA output.

Based on MSA, the standard Transformer encoder consisting of T encoder layers is designed linearly. As shown in Figure 4.8 (f), each encoder layer contains MSA, Multi-Layer Perception (MLP), Dropout (Drop), Add, and Normalization (Norm) operations. As a result, we generate the self-attention features x_T, and further obtain the global features $\mathcal{F}_G$ by Average (Avg) pooling:

$$x'_t = Norm(Drop(MSA(x_{t-1})) + x_{t-1}),\ t = 1, ..., T \tag{4.20}$$

$$x_t = Norm(Drop(MLP(x'_{t-1})) + x'_{t-1}),\ t = 1, ..., T \tag{4.21}$$

$$\mathcal{F}_G = Avg(x_T), \quad \mathcal{F}_G \in \mathbb{R}^{1\times d},\ x_T \in \mathbb{R}^{m^2\times d} \tag{4.22}$$

Finally, two FC layers and one LeakyReLU (LReLU) activation layer are adopted for producing a global quality score Q_G of the input block:

$$Q_G = FC(LReLU(FC(\mathcal{F}_G))) \tag{4.23}$$

The above processes model the local angular-spatial effect and establish the global relationship between different local regions to predict the global quality, which are viewed as the principal branch of the proposed metric.

4.3.3 Discriminative Region Selection-Based Auxiliary Learning

In addition to the learning of the global quality prediction in the principal branch, we further design an auxiliary branch to exploit the local angular-spatial effect $\mathcal{F}_L$ for learning from a local perspective. As shown in Figure 4.8 (a), we first predict the local quality score for each spatial position of $\mathcal{F}_L$ with two 1×1 convolutions:

$$Q_L = Conv_{1\times1}(\mathcal{F}_L) \tag{4.24}$$

where $Q_L \in \mathbb{R}^{m \times m}$ denotes the predicted local quality scores of different spatial positions.

As widely discussed in many studies (*e.g.*, [103]), when an image suffers from distortions, the quality of different regions in the image will be affected to varying degrees. In addition, since human eyes tend to focus on the anomalies, errors, and incoherent regions of an image, these low-quality regions are more likely to negatively impact global quality. Along this vein, we argue that different regions in LFIs contain different angular-spatial effects and corresponding local quality, and only the quality of certain regions has a close correlation with the global LFI quality. Therefore, adopting all regions as supplementary information for learning will be detrimental to the training process. As a result, we propose a DRS module and utilize discriminative regions for auxiliary learning to improve learning efficiency and prediction accuracy. Since the angular-spatial effect denotes the effect of angular inconsistency on spatial quality, regions with large angular sparsity are regarded as discriminative regions, which contain strong angular-spatial effect [59], [67]. As shown in Figure 4.8 (a), for an input block $\mathcal{LB}$, the grayscale SAI array, denoted as $G \in \mathbb{R}^{U \times V \times S \times S}$, is first generated. Then the angular horizontal difference $D^{hor}_{u,v}$ and angular vertical difference $D^{ver}_{u,v}$ between adjacent SAIs are calculated as:

$$D^{hor}_{u,v} = G_{u+1,v} - G_{u,v} \tag{4.25}$$

$$D^{ver}_{u,v} = G_{u,v+1} - G_{u,v} \tag{4.26}$$

where $u = 1, ..., U\text{-}1$, $v = 1, ..., V\text{-}1$, and $G_{u,v}$ denotes the SAI of (u, v) angular viewpoint in G. Then the angular gradient SAI array $G \in \mathbb{R}^{(U-1)\times(V-1)\times S\times S}$ is calculated as:

$$G_{u,v} = \sqrt{{D^{hor}_{u,v}}^2 + {D^{ver}_{u,v}}^2} \tag{4.27}$$

where $G_{u,v}$ denotes the SAI of (u, v) angular viewpoint in G. Then we combine different angular viewpoints in G to $G_a \in \mathbb{R}^{S\times S}$ using angular average pooling $\psi_A(\cdot)$:

$$G_a = \psi_A(G) = \frac{1}{(U-1)(V-1)} \sum_{u=1}^{U-1} \sum_{v=1}^{V-1} G_{u,v} \tag{4.28}$$

The generated G_a has the same spatial resolution as the input $\mathcal{LB}$ and stores the angular sparsity by recording the gradient intensity, in which regions with larger gradient intensity are more pronounced to have stronger angular-spatial effects. In order to select discriminative regions corresponding to Q_L with spatial resolution $m \times m$, we fur-

ther reduce the spatial resolution of G_a by applying spatial average pooling $\psi_S(\cdot)$ for non-overlapping $(S/m)\times(S/m)$ regions:

$$G_s = \psi_S(G_a) \tag{4.29}$$

where $G_s \in \mathbb{R}^{m\times m}$ is the gradient intensity corresponding to Q_L.

Let $G_s^{\Omega_N}$ denote the assemble of N $(0<N<m^2)$ discriminative regions with the largest gradient intensity in G_s, the angular-spatial effect of these regions can be exploited to improve the quality-aware feature learning. Here, N is a hyperparameter and its effect on the performance of the proposed metric will be further discussed in the experiment section. Correspondingly, the predicted local quality scores of these discriminative regions are denoted as $Q_L^{\Omega_N}$.

4.3.4 Dynamic Weighting Loss

The training of the auxiliary branch is driven by minimizing the differences between the local quality scores Q_L of discriminative regions Ω_N and the MOS label Q_{MOS}, based on the Mean Square Error (MSE) function:

$$L_a = \frac{1}{N}\frac{1}{B_s}\sum_{n=1}^{N}\sum_{i=1}^{B_s}(Q_L^{\Omega_n,i} - Q_{MOS}^i)^2 \tag{4.30}$$

where L_a denotes the loss function of the auxiliary branch, B_s represents the batch size.

Similarly, the loss function of the principal branch is defined based on the MSE between global quality scores Q_G and the MOS label Q_{MOS}:

$$L_p = \frac{1}{B_s}\sum_{i=1}^{B_s}(Q_G^i - Q_{MOS}^i)^2 \tag{4.31}$$

where L_p denotes the loss function of the principal branch.

Intuitively, the easiest way to construct the final loss function is to add L_p and L_a according to fixed weights [73], [72]. However, this naive way is suboptimal in our proposed metric. This is mainly attributed to the fact that the auxiliary branch is only involved during training, while the final training objective of our metric is the principal branch. In addition, the learning of local angular-spatial effect should serve as the basis for the learning of global quality prediction. Based on such a hypothesis, the auxiliary branch should be biased at the early training stage, while the principal branch should be prominent

at the late training stage. As a result, we design a DWLoss to assign dynamic learning weights for principal and auxiliary branches:

$$DWLoss = \frac{Ep_c}{Ep_t} L_p + (1 - \frac{Ep_c}{Ep_t}) L_a \tag{4.32}$$

where Ep_c and Ep_t denote the number of current and total epochs, respectively.

At the beginning of the training process, the proposed DWLoss assigns the largest weight to the auxiliary branch, aiming to learn the angular-spatial effect from a local perspective. Subsequently, the weight of the auxiliary branch gradually decreases, while the weight of the principal branch gradually increases. Finally, the global quality prediction is dominant at the end of the training, leading to a better training result than the naive way.

4.4 Proposed EDDMF Metric

As demonstrated by the DeLFIQE metric [73] and our DeeBLiF metric, evaluating the LFI quality with patches is a promising approach to reduce computational complexity. However, these patch-based LFIQA metrics are all NR metrics. Compared to the NR metrics, the FR metrics are undoubtedly more reliable with access to reference information. To this end, we propose an Efficient Deep Discrepancy Measuring Framework (EDDMF) for FR LFIQA in this section.

First, since LFI patches need to reflect the characteristics of high-dimensional LFIs, we exploit two kinds of LFI patches, spatio-angular patches and sub-aperture patches, both of which contain angular and spatial quality degradation information of LFIs. In addition, although spatial and angular information are two different characteristics of LFIs, existing LFIQA metrics generally adopt homogeneous features for both information, *i.e.*, CNN features only or handcrafted features only. However, we argue that they should be treated differently to maximize efficiency and effectiveness. For angular information, due to the non-intuitive nature of angular information deterioration, it is often perceived in combination with spatial information. Therefore, CNN features with strong discriminative ability are more suitable for angular information. For spatial information, existing handcrafted feature-based FR metrics can handle most spatial distortions well when reference information is available. As a result, a combination of angular-dominant CNN features and spatial-dominant handcrafted features is presented as a more comprehensive description

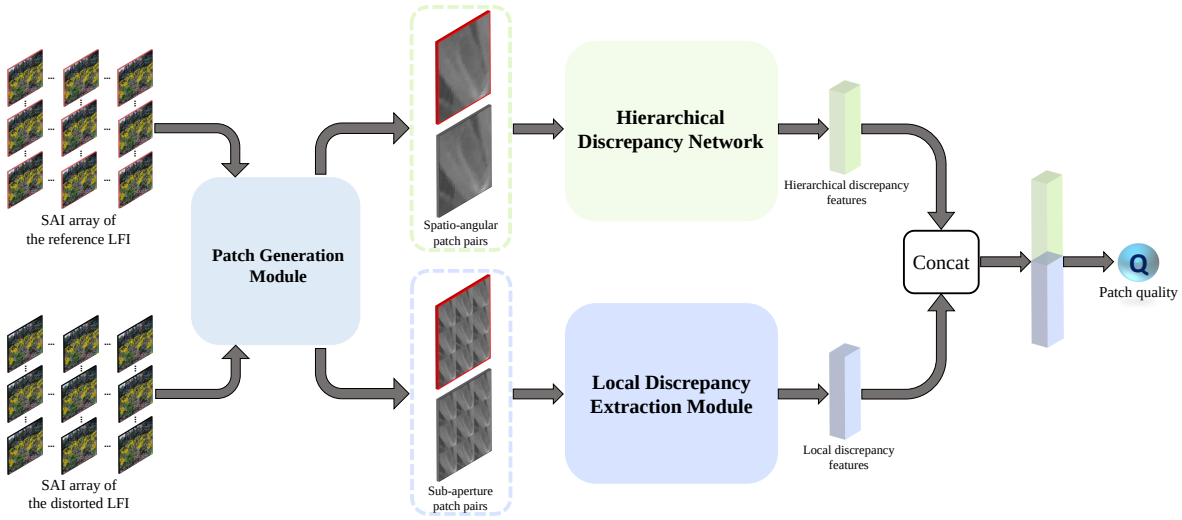


Figure 4.9 – Overview of the proposed EDDMF. Note that patches with and without red borders denote reference patches and distorted patches, respectively.

of LFI patches.

The overview of the proposed EDDMF is illustrated in Figure 4.9, which contains three main components: patch generation module, hierarchical discrepancy network, and local discrepancy extraction module. First, given a distorted LFI and its corresponding reference LFI, both spatio-angular patches and sub-aperture patches of the reference and distorted LFIs are generated by the patch generation module. Then, each spatio-angular patch pair is composed of a spatio-angular patch and its corresponding reference version, while each sub-aperture patch pair consists of a sub-aperture patch and its corresponding reference version. Note that patches with and without red borders in Figure 4.9 denote reference patches and distorted patches, respectively. Second, hierarchical discrepancy features are obtained by the hierarchical discrepancy network using spatio-angular patch pairs. Third, sub-aperture patch pairs are exploited to extract local discrepancy features through the local discrepancy extraction module. Finally, hierarchical discrepancy features and local discrepancy features are combined to predict the patch quality, and then the overall quality of the distorted LFI is obtained by averaging the quality of all patches.

4.4.1 Patch Generation Module

Based on the assumption that LFI patches can reflect the LFI quality as long as they contain enough angular information, a patch generation module is presented to extract

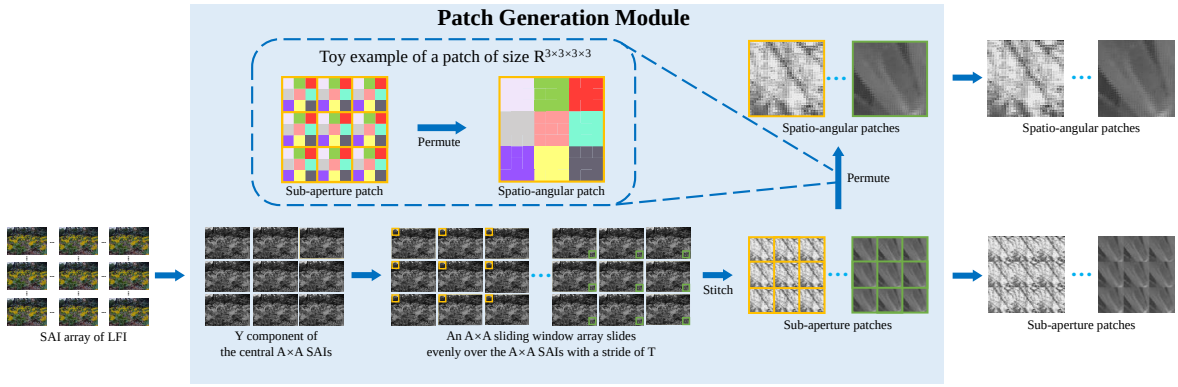


Figure 4.10 – Pipeline of the proposed patch generation module. For better visualization, A is set to 3 for all illustrations with $A \times A$ angular resolution.

patches that are capable of reflecting the LFI quality, as shown in Figure 4.10.

Following the DeeBLiF metric, for an input LFI in the form of SAI array, denoted as $\mathcal{L} \in \mathbb{R}^{U \times V \times H \times W \times C}$, we convert it into YCbCr color format and use only the luminance component Y of the central $A \times A$ SAIs for subsequent patch generation:

$$\mathcal{L}_Y = \{rgb2ycbcr(\mathcal{L})\}_Y \tag{4.33}$$

$$\mathcal{L}_{YC} = Central(\mathcal{L}_Y) \tag{4.34}$$

where $\mathcal{L}_Y \in \mathbb{R}^{U \times V \times H \times W}$ and $\mathcal{L}_{YC} \in \mathbb{R}^{A \times A \times H \times W}$ denote the luminance component Y of the $U \times V$ SAIs and the central $A \times A$ SAIs, respectively.

To extract both angular and spatial information in LFIs, an $A \times A$ sliding window array is defined. Each sliding window is defined as a square of size $S \times S$ to obtain the same amount of spatial information from the vertical and horizontal directions. Then the sliding window array is placed on the $A \times A$ SAIs for evenly cropping and sliding, with a stride of T. For each sliding position, $A \times A$ blocks of size $S \times S$ are obtained and a sub-aperture patch is generated by stitching all blocks. After sliding over $\mathcal{L}_{YC}$, a set of sub-aperture patches $\mathbb{P}_s$ is generated:

$$I = \lceil \frac{H - S}{T} \rceil, \ J = \lceil \frac{W - S}{T} \rceil \tag{4.35}$$

$$P_s^{i,j} = Stitch\{B_{1,1}, ..., B_{Ax,Ay}, ..., B_{A,A}\}^{i,j} \tag{4.36}$$

$$\mathbb{P}_s = \{P_s^{1,1}, ..., P_s^{i,j}, ..., P_s^{I,J}\} \tag{4.37}$$

where I and J represent the number of patches generated in the vertical and horizontal directions, respectively. $B_{Ax,Ay} \in \mathbb{R}^{S\times S}$ denotes the Ax-th row and the Ay-th column block cropped from $\mathcal{L}_{YC}$. $P_s^{i,j} \in \mathbb{R}^{A\times A\times S\times S}$ denotes the sub-aperture patch with the i-th vertical index and the j-th horizontal index.

Further, for each sub-aperture patch $P_s^{i,j}$, a permutation operation is applied to exchange its angular coordinates and spatial coordinates for obtaining a spatio-angular patch. A toy example of a patch of size 3×3×3×3 is illustrated in the dashed box in Figure 4.10. In this case, the permutation operation can be visualized as grouping all pixels of the same color. After applying the permutation operation to all $P_s^{i,j}$ in $\mathbb{P}_s$, a set of spatio-angular patches $\mathbb{P}_a$ is generated as:

$$P_a^{i,j} = Permute(P_s^{i,j}) \tag{4.38}$$

$$\mathbb{P}_a = \{P_a^{1,1}, ..., P_a^{i,j}, ..., P_a^{I,J}\} \tag{4.39}$$

where $P_a^{i,j} \in \mathbb{R}^{S\times S\times A\times A}$ denotes the spatio-angular patch with the i-th vertical index and the j-th horizontal index.

Finally, $\mathbb{P}_a$ and $\mathbb{P}_s$ are the outputs of the patch generation module. Specifically, $\mathbb{P}_a$ contains $I\times J$ spatio-angular patches while $\mathbb{P}_s$ includes $I\times J$ sub-aperture patches. The data sizes of the original LFI L, the generated $\mathbb{P}_a$ and $\mathbb{P}_s$, and their ratio are calculated as:

$$D_L = U\times V\times H\times W\times C \tag{4.40}$$

$$D_{\mathbb{P}_a} = D_{\mathbb{P}_s} = A\times A\times S\times S\times I\times J \tag{4.41}$$

$$R = \frac{D_{\mathbb{P}_a}}{D_L} = \frac{D_{\mathbb{P}_s}}{D_L} \tag{4.42}$$

where D_L, $D_{\mathbb{P}_a}$, and $D_{\mathbb{P}_s}$ represent the data sizes of L, $\mathbb{P}_a$, and $\mathbb{P}_s$, respectively. R is the ratio between $D_{\mathbb{P}_a}$ (or $D_{\mathbb{P}_s}$) and D_L. In our implementation, three hyperparameters related to the generation of patches, A, S, and T are set to 5, 32, and 64, respectively. Taking a LFI with 9×9 angular resolution, 434×625 spatial resolution, and 3 channel number for example, according to Eq. (4.42), we only use 3% of the original LFI data for quality assessment, which greatly reduces the computational complexity of the whole framework. In addition, since the generated patches consist of angular and spatial information from different locations of the original LFI, they are able to reflect the overall LFI quality.

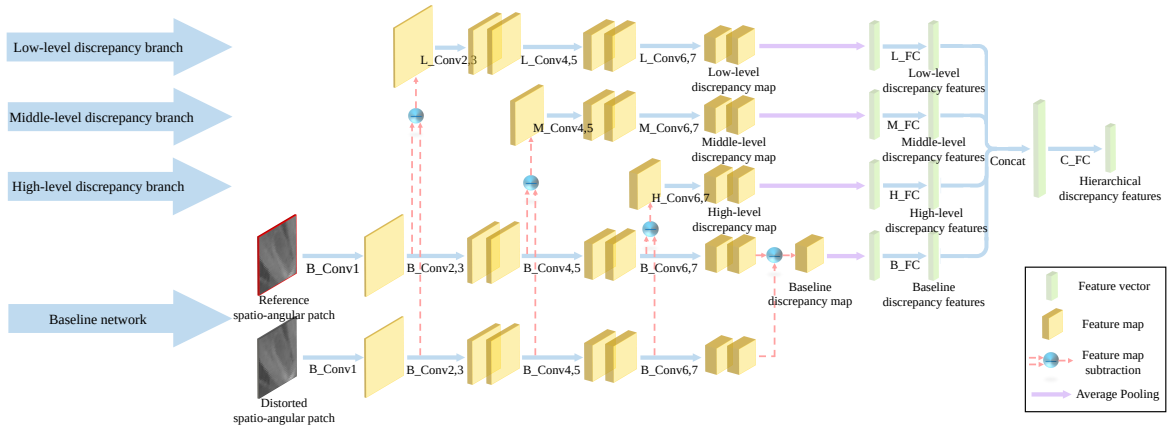


Figure 4.11 – Architecture of the proposed hierarchical discrepancy network, which consists of a baseline network and three discrepancy branches of different levels.

4.4.2 Hierarchical Discrepancy Network

The angular information in LFIs essentially distinguishes LFIs from traditional 2D images. Therefore, the key to designing a well-performing LFIQA metric lies in effectively capturing the angular effect on spatial information. Generally, LFI's angular information will be affected to varying degrees when different types and levels of distortion are introduced. For example, JPEG distortion focuses on distorting spatial information, while its effect on angular information is very limited. Conversely, the opposite is true for the interpolation distortion. Thus, we argue that the solidified angular feature extraction fails to accommodate the ever-changing distortion types and levels, thereby resulting in suboptimal performance in quality evaluation.

Given a spatio-angular patch, we notice that shallow convolutional layers capture information between nearest neighbor pixels, *i.e.*, angular features. As the network deepens and the feature map size decreases, information between distant pixels is captured accordingly, *i.e.*, spatial features, which can also be known as spatio-angular features since they are extracted based on the shallow angular features. Inspired by the above observations, we design a hierarchical discrepancy network to extract the hierarchical discrepancy features between reference and distorted spatio-angular patches, in which different branches extract spatio-angular discrepancy features with different angular encoding levels.

Figure 4.11 illustrates the architecture of our proposed hierarchical discrepancy network. The proposed network consists of a baseline network and three discrepancy branches, which generate the baseline, low-level, middle-level, and high-level discrepancy maps, re-

Table 4.2 – Network configuration of the proposed hierarchical discrepancy network.

Layer Name	**Kernel Size**	**Channel**	**Stride**	**Padding**
B_Conv1	3×3	[1, 32]	2	1
	BN, ReLu			
B/L_Conv2	3×3	[32, 64]	2	1
	BN, ReLu			
B/L_Conv3	3×3	[64, 64]	1	1
	BN, ReLu			
$B/L/M_Conv4$	3×3	[64, 128]	2	1
	BN, ReLu			
$B/L/M_Conv5$	3×3	[128, 128]	1	1
	BN, ReLu			
$B/L/M/H_Conv6$	3×3	[128, 256]	2	1
	BN, ReLu			
$B/L/M/H_Conv7$	3×3	[256, 256]	1	1
	BN, ReLu			
Average Pooling				
$B/L/M/H_FC$	-	[256, 256]	-	-
$Concat$	-	[4×256, 256]	-	-
C_FC	-	[1024, 128]	-	-

spectively. Then each discrepancy map is converted into a feature vector using an average pooling layer, followed by a FC layer (denoted as $B/L/M/H_FC$, respectively). Finally, a FC layer, denoted as C_FC, is applied after concatenating the four feature vectors to generate the final hierarchical discrepancy features. The baseline network contains two sub-networks for extracting features from reference and distorted spatio-angular patches, respectively. Each sub-network includes seven convolutional layers, *i.e.*, $B_Conv1 - B_Conv7$. Note that the two sub-networks share the same weights in all convolutional layers. At the end of the baseline network, the feature maps extracted by the two sub-networks are subtracted, *i.e.*, the baseline discrepancy map between reference and distorted spatio-angular patches is obtained. Additionally, three discrepancy branches are exploited to extract the low-level, middle-level, and high-level discrepancy maps, respectively. The three branches start from the subtraction of reference and distorted feature maps generated by B_Conv1, B_Conv3, and B_Conv5, respectively, and have the same structure as the rest of the baseline network.

As shown in Table 4.2, all convolution layers use 3×3 kernel size with 1 padding, and each convolutional layer is followed by a Batch Normalization (BN) layer and a ReLu activation function. As the number of convolutional layers increases, the output channel

increases while the output feature map size decreases. Specifically, we halve the feature map size by setting the stride to 2 in B_Conv1, B/L_Conv2, $B/L/M_Conv4$, and $B/L/M/H_Conv6$, respectively. Therefore, the final feature map size is 1/16 of the original input patch. Taking a sub-network in the baseline network for example, with an input patch of size 160×160×1, the size and channel of the feature map are changed as follows: $160\times160\times1 - 80\times80\times32 - 40\times40\times64 - 40\times40\times64 - 20\times20\times128 - 20\times20\times128 - 10\times10\times256 - 10\times10\times256$. After applying the average pooling layer, the FC structure $4\times256 - 4\times256 - 1024 - 128$ is adopted to generate hierarchical discrepancy features. Finally, the final output of our proposed hierarchical discrepancy network is 128 dimensions, consisting of spatio-angular discrepancy features with different angular encoding levels.

4.4.3 Local Discrepancy Extraction Module

As mentioned before, hierarchical discrepancy features are more angularly discriminative due to the hierarchical design based on spatio-angular patches. As a result, the spatial features used for quality assessment are insufficient to some extent. Therefore, we aim to extract local discrepancy features, which focus more on spatial quality degradation and serve as complementary features to hierarchical discrepancy features. Specifically, we propose a local discrepancy extraction module to generate the local discrepancy features between reference and distorted sub-aperture patches, which provide a more comprehensive description of the patch quality degradation.

The pipeline of the proposed local discrepancy extraction module is illustrated in Figure 4.12. Given a distorted sub-aperture patch $P_s \in \mathbb{R}^{A\times A\times S\times S}$ and its reference version $\hat{P}_s \in \mathbb{R}^{A\times A\times S\times S}$, they are both composed by $A\times A$ blocks of size $S\times S$. Since each block contains local information derived from a single SAI, we measure the local discrepancy between reference and distorted blocks to evaluate the impact of spatial quality degradation for each SAI. Specifically, for each distorted block and its reference version, we measure their local discrepancy score by adopting a handcrafted feature-based FR 2DIQA metric instead of a CNN model. The reason behind this is that existing handcrafted feature-based FR 2DIQA metrics can deal well with the degradation of spatial information in 2D images. Finally, a total of $A\times A$ local discrepancy scores are generated and further concatenated into a feature vector, named local discrepancy features. In our implementation, we use the representative Structural Similarity Index Metric (SSIM) [106] as the FR 2DIQA metric due to its low computational complexity and excellent performance. The above process

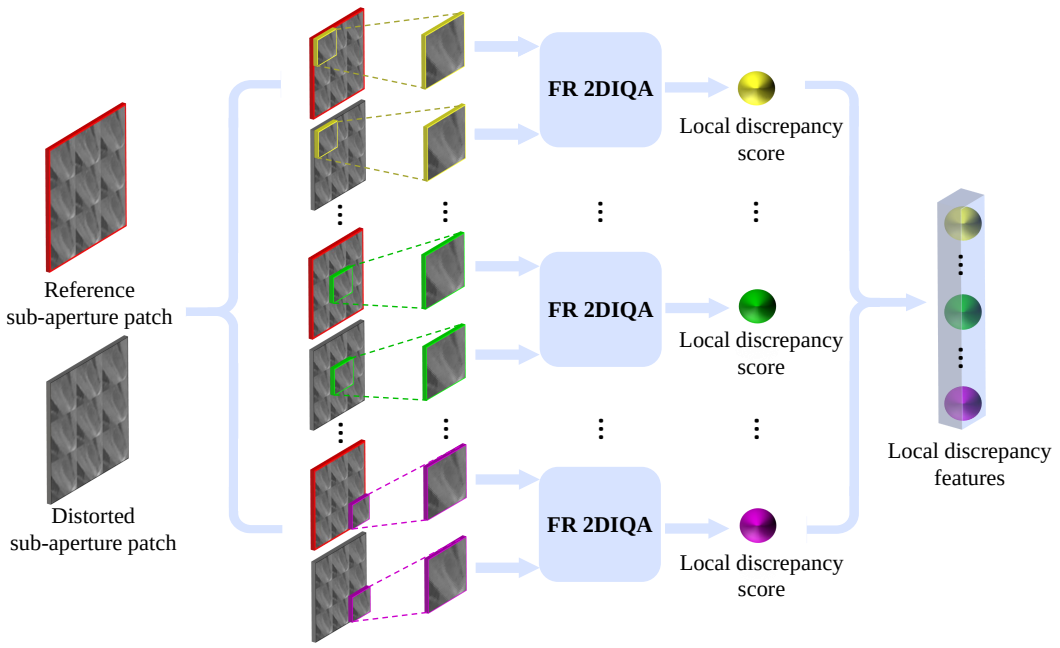


Figure 4.12 – Pipeline of the proposed local discrepancy extraction module. For better visualization, A is set to 3 for all illustrations with $A \times A$ angular resolution.

can be described as:

$$S_i = SSIM(\hat{B}_i, B_i) \tag{4.43}$$

$$\mathcal{F}_{local} = Concat(S_1, S_2, ..., S_{A \times A}) \tag{4.44}$$

where $\hat{B}_i$ and B_i denote the i-th reference and distorted blocks of the sub-aperture patch P_s, respectively, S_i is the local discrepancy score between $\hat{B}_i$ and B_i, $\mathcal{F}_{local}$ is the local discrepancy features with $A \times A$ dimensions.

Since the same distortion may have different impacts on different SAIs and each local discrepancy score can only measure the spatial degradation of a single SAI, the angular degradation is neglected. To this end, we combine all the local discrepancy scores to estimate the angular degradation of sub-aperture patches. Although angular information is considered, local discrepancy features are more spatially discriminative and are thus used as complementary features to the angular-dominant hierarchical discrepancy features.

4.5 Performance Evaluation and Discussion

As in the experiment part of Chapter 3, in this section, we also use five standard LFIQA databases (Win5-LID [2], NBU-LF1.0 [37], SHU [38], VALID10bit [43], and LFDD [39]) to report the experimental results and verify the effectiveness of our proposed metrics DeeBLiF, PVBLiF, ASEM-BLiF, and EDDMF.

4.5.1 Hyperparameter Dependency

In this subsection, we investigate the performance dependency of different hyperparameter values in different proposed metrics.

1) PVBLiF. The PVBLiF metric involves two manual hyperparameters ($A \times A$ and $S \times S$) that determine the central angular resolution of LFIs and the spatial resolution of PVBSs. Table 4.3 exhibits the experimental results on the Win5-LID database. Note that experimental results of one hyperparameter are reported when the other hyperparameter is fixed. For hyperparameter A, it is observed that setting a small value for A leads to relatively poor performance, whereas similar performance can be achieved by setting A to 5 or higher. One possible explanation is that the selection of A determines the amount of angular information, and the lack of angular information will directly affect the prediction accuracy. Besides, experimental results also show that the PVBLiF metric can be fully exploited with moderate angular information. For hyperparameter S, we can see that adopting a moderate S value achieves better performance than a small or large value. The reason for this phenomenon could be that as the value of S increases, the spatial information of a single PVBS increases while the number of training samples decreases, and vice versa. A more reasonable trade-off can be obtained with a moderate S value, resulting in better performance. Based on the above analyses, we set A and S to 5 and 32 in PVBLiF, respectively.

2) ASEM-BLiF. The performance of the ASEM-BLiF metric is affected by the number of the selected discriminative regions N. We perform hyperparameter experiments on the Win5-LID database to investigate its impact. As shown in Table 4.4, the performance is compared when N is set to 1, 5, 10, 20, 50, 100, 200, and w/o. Here, w/o represents that all regions are incorporated into the learning process. From the Table we can see that the performance is relatively poor if the value of N is too large or too small. The main reason could be that insufficient discriminative regions lead to inadequate information for auxiliary training, while excessive discriminative regions introduce too many low-quality

Table 4.3 – Hyperparameter dependency of the PVBLiF metric on the Win5-LID database. $A\times A$ denotes the central angular resolution and $S\times S$ denotes the spatial resolution. The best results are in **bold**.

Hyperparameter	Value	PLCC	SRCC	RMSE
$A\times A$	3×3	0.8559	0.8411	0.4989
	5×5	**0.8749**	**0.8580**	**0.4660**
	7×7	0.8721	0.8564	0.4705
	9×9	0.8740	0.8607	0.4669
$S\times S$	8×8	0.8567	0.8428	0.4920
	16×16	**0.8749**	**0.8580**	**0.4660**
	32×32	0.8733	0.8551	0.4727
	64×64	0.8604	0.8445	0.4904

Table 4.4 – Hyperparameter dependency of the ASEM-BLiF metric on the Win5-LID database. N denotes the number of the selected discriminative regions. The best results are in **bold**.

Hyperparameter	Value	PLCC	SRCC	RMSE
N	1	0.8731	0.8594	0.4741
	5	0.8863	0.8683	0.4477
	10	0.9015	0.8723	0.4232
	20	0.9072	**0.8949**	0.4090
	50	**0.9085**	0.8889	**0.4083**
	100	0.8967	0.8746	0.4309
	200	0.8869	0.8650	0.4512
	w/o	0.8849	0.8633	0.4552

predictions, ultimately impeding the learning process. Therefore, in our implementation, we set N to a moderate value of 20 which yields satisfactory performance on the Win5-LID database.

3) EDDMF. There are three hyperparameters for generating spatio-angular patches and sub-aperture patches in EDDMF: $A\times A$, the central angular resolution; $S\times S$, the size of each sliding window; and T, the stride of the sliding window array. Table 4.5 exhibits the performance dependency of these three hyperparameters on the Win5-LID database. Note that experimental results of one hyperparameter are reported when other hyperparameters are fixed. For hyperparameter A, similar performance can be obtained when $A\times A$ is set to 5×5 or larger, while a smaller value results in suboptimal performance,

Table 4.5 – Hyperparameter dependency of the EDDMF metric on the Win5-LID database. $A\times A$ denotes the central angular resolution, $S\times S$ denotes the size of each sliding window, and T denotes the stride of the sliding window array. The best results are in **bold**.

Hyperparameter	Value	PLCC	SRCC	RMSE
$A\times A$	3×3	0.8428	0.8155	0.5178
	5×5	**0.8654**	**0.8354**	**0.4839**
	7×7	0.8635	0.8351	0.4860
	9×9	0.8629	0.8349	0.4891
$S\times S$	8×8	0.8506	0.8287	0.5102
	16×16	0.8589	0.8305	0.4946
	32×32	**0.8654**	**0.8354**	**0.4839**
	64×64	0.8630	0.8232	0.5022
T	32	**0.8668**	**0.8401**	**0.4819**
	48	0.8625	0.8365	0.4883
	64	0.8654	0.8354	0.4839
	80	0.8517	0.8225	0.5046

it is possibly because there is not enough angular information to measure the LFI quality. For hyperparameter S, adopting a moderate value, *i.e.*, $S\times S$ is set to 32×32, achieves the best performance. One possible explanation is that as the S value increases, a single patch contains more spatial information while the total number of generated patches decreases. For hyperparameter T, the smaller the value set, the more LFI patches can be obtained. Thus, it can be found that a smaller value achieves better performance since more samples are used for training. However, to keep a better trade-off between computational complexity and quality evaluation performance, we adopt a moderate value of 64 in EDDMF.

4.5.2 Overall Performance Comparison

The performance of the proposed metrics DeeBLiF, PVBLiF, ASEM-BLiF, and EDDMF is compared with that of a variety of state-of-the-art metrics (including our proposed metrics in Chapter 3): three FR 2DIQA metrics (NQM [90], IFC [91], HDR-VDR-2 [92]), two RR LFIQA metrics (LF-IQM [60], RRLFIQA-4DDWT [61]), five FR LFIQA metrics (MDFM [53], Fang's [54], Min's [56], Meng's [57], KRIQE [59]), three NR 2DIQA metrics (GWH-GLBP [93], NIQE [94], BRISQUE [95]), and eleven NR LFIQA metrics (BE-

LIF [62], NR-LFQA [63], Tensor-NLFQ [66], PVRI [67], DSA [68], 4D-DCT-LFIQA [69], TSSV-LFIQA [70], NR-LF-QAE [71], DeLFIQE [73], SATV-BLiF, MPT-MWI-BLiF). Also, we perform the 2DIQA metrics on all SAIs and report the average performance. The reported performance of all metrics is reproduced using the released code and default parameter settings from their authors to avoid bias.

Table 4.6 provides the experimental results of overall performance and time complexity. We can see that the proposed DeeBLiF metric, which is our initial attempt at using deep learning in LFIQA, outperforms most existing state-of-the-art LFIQA metrics on the Win5-LID, NBU-LF1.0, SHU, and LFDD databases. This strongly demonstrates the great potential of deep learning-based LFIQA metrics. However, the performance of the DeeBLiF metric has not been fully exploited as there is still a significant gap compared to that of our MPT-MWI-BLiF metric on the Win5-LID and NBU-LF1.0 databases. Fortunately, by comprehensively considering the characteristics of LFIs, we extend DeeBLiF to two more powerful NR LFIQA metrics PVBLiF and ASEM-BLiF. As a result, PVBLiF and ASEM-BLiF gain varying degrees of performance improvements over DeeBLiF in most cases. Overall, the performance of PVBLiF is slightly better than that of ASEM-BLiF. However, ASEM-BLiF significantly outperforms PVBLiF on the VAILD10bit database. A possible explanation is that ASEM-BLiF pays more attention to the training strategy and thus it is particularly effective when the size of available training samples is limited. The time complexity of three proposed deep learning-based NR LFIQA metrics DeeBLiF, PVBLiF, and ASEM-BLiF is relatively low compared to other existing NR LFIQA metrics. In addition, the proposed EDDMF metric, a FR extension of the DeeBLiF metric, also exhibits outstanding quality evaluation performance on all databases, especially on the LFDD database. Further, it should be highlighted that our EDDMF metric has a very low time complexity. All these results demonstrate the effectiveness and efficiency of our proposed metrics.

Then, we present the box plots of SRCC distributions using leave-two-fold-out cross-validation in Figure 4.13, to compare the stability of four proposed metrics and other existing LFIQA metrics. The stability experiments are performed on the Win5-LID and NBU-LF1.0 databases, whose train-test split numbers are 45 and 91, respectively. Boxes with smaller sizes and higher positions indicate better stability and performance of quality evaluation, respectively. Figure 4.13 shows that our three deep learning-based LFIQA metrics, PVBLiF, ASEM-BLiF, and EDDMF, have similar predictive accuracy and stability on both databases, while consistently outperforming our DeeBLiF metric and other

Table 4.6 – Overall performance comparison and time complexity on five LFIQA databases. The best results of RR/FR metrics are in blue, while the best results of NR metrics are marked in **bold**.

	Database	Win5-LID			NBU-LF1.0			SHU			VALID10bit			LFDD			
Type	Metric	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	Time (s)
FR 2DIQA	NQM [90]	0.6509	0.5908	0.9082	0.6479	0.6614	0.6995	0.8581	0.8618	0.5619	0.8702	0.8630	0.4586	0.4461	0.3810	0.9989	10.1941
	IFC [91]	0.4945	0.5429	1.0628	0.7440	0.6843	0.6049	0.8853	0.9068	0.4969	0.9613	0.9647	0.2260	0.3683	0.4226	1.0198	36.3739
	HDR-VDR-2 [92]	0.7315	0.7169	0.6643	0.7827	0.7924	0.6212	0.8444	0.8736	0.5861	0.9459	0.9334	0.3037	0.3986	0.5306	1.0368	89.2621
RR LFIQA	LF-IQM [60]	0.3620	0.3438	0.8930	0.4260	0.2700	0.7966	0.3601	0.2960	1.0046	0.4343	0.3097	0.8175	0.4457	0.4287	0.9950	589.7851
	RRLFIQA-4DDWT [61]	0.8305	0.8052	0.5439	0.6917	0.6741	0.6539	0.8325	0.8498	0.6072	0.9589	0.9445	0.2645	0.6457	0.6056	0.8636	128.1106
FR LFIQA	MDFM [53]	0.7303	0.6768	0.6625	0.8444	0.8138	0.4749	0.8275	0.8543	0.6149	0.9164	0.9054	0.3709	0.5725	0.5282	0.9276	0.8537
	Fang's [54]	0.8256	0.8077	0.5480	0.8589	0.8274	0.4571	0.8833	0.8856	0.5119	0.9730	0.9578	0.2135	0.5617	0.4950	0.9293	1.1574
	Min's [56]	0.7350	0.6645	0.6794	0.7112	0.6577	0.6476	0.8496	0.8470	0.5745	0.9757	0.9325	0.2051	0.5596	0.4094	0.9366	3.9845
	Meng's [57]	0.6924	0.6347	0.7001	0.8367	0.7819	0.4944	0.9282	0.9203	0.4037	0.9579	0.9383	0.2629	0.3043	0.3493	1.0593	30.4872
	KRIQE [59]	0.7380	0.6820	0.6538	0.8693	0.8377	0.4429	0.9582	0.9613	0.3091	0.9705	0.9502	0.2236	0.3500	0.3927	1.0410	115.7841
	EDDMF (ours)	0.8654	0.8354	0.4839	0.9024	0.8743	0.3866	0.9443	0.9206	0.3575	0.9533	0.9419	0.2461	0.9406	0.9077	0.3794	1.8775
NR 2DIQA	GWH-GLBP [93]	0.4752	0.3391	0.8364	0.4947	0.3550	0.7780	0.6964	0.6102	0.7677	0.7709	0.7248	0.5622	0.5008	0.3493	0.9577	4.8472
	NIQE [94]	0.6246	0.4482	0.7584	0.4792	0.3701	0.7948	0.9187	0.8920	0.4247	0.8044	0.7053	0.5333	0.5642	0.4449	0.9405	6.0077
	BRISQUE [95]	0.6263	0.4559	0.7530	0.4969	0.3750	0.7910	0.9012	0.8747	0.4614	0.7701	0.6262	0.5674	0.5103	0.4179	0.9597	**3.1543**
NR LFIQA	BELIF [62]	0.5912	0.5119	0.7572	0.7161	0.6892	0.6291	0.8976	0.8671	0.4784	0.8099	0.7852	0.5437	0.7747	0.7072	0.7103	107.8814
	NR-LFQA [63]	0.6952	0.6275	0.6750	0.8327	0.8036	0.4895	0.9390	0.9347	0.3729	0.7687	0.7339	0.5686	0.6647	0.5864	0.8374	225.2069
	Tensor-NLFQ [66]	0.7595	0.7345	0.6327	0.7624	0.7261	0.5856	0.8649	0.8630	0.5424	0.8748	0.8553	0.4477	0.8446	0.7887	0.6025	697.6515
	PVRI [67]	0.7217	0.6827	0.6530	0.8017	0.7603	0.5271	0.9167	0.8771	0.4237	0.8338	0.7778	0.4831	0.6835	0.6037	0.8037	71.3578
	DSA [68]	0.7754	0.7328	0.6150	0.8482	0.8093	0.4578	0.9313	0.9195	0.3904	0.8449	0.8202	0.4882	0.7502	0.6902	0.7135	198.5443
	4D-DCT-LFIQA [69]	0.8267	0.8079	0.5512	0.8381	0.8213	0.4906	0.9400	0.9320	0.3691	0.8535	0.8308	0.4666	0.8206	0.7699	0.6411	169.2623
	TSSV-LFIQA [70]	0.6876	0.6186	0.6807	0.7876	0.7492	0.5388	0.9032	0.8856	0.4642	0.8616	0.8187	0.4655	0.7222	0.6781	0.7582	44.6696
	NR-LF-QAE [71]	0.6533	0.6239	0.6888	0.7444	0.6921	0.5808	0.9088	0.8788	0.4364	0.6013	0.5250	0.7572	0.7375	0.6788	0.7573	254.1088
	DeLFIQE [73]	0.4431	0.3860	0.8636	0.8063	0.7653	0.5220	0.6811	0.5765	0.8000	0.4907	0.4254	0.8049	0.6470	0.5809	0.8592	15.4193
	SATV-BLiF (ours)	0.7933	0.7704	0.5842	0.8515	0.8237	0.4686	0.9332	0.9284	0.3897	0.8750	0.8223	0.4298	0.6617	0.5543	0.8458	3.8812
	MPT-MWI-BLiF (ours)	0.8701	0.8509	0.4714	0.8959	0.8710	0.3971	0.9403	0.9374	0.3685	0.8753	0.8616	0.4426	0.8291	0.7667	0.6290	26.8342
	DeeBLiF (ours)	0.8427	0.8186	0.5160	0.8583	0.8229	0.4588	0.9548	0.9419	0.3185	0.9283	0.8783	0.3443	0.8827	0.8086	0.5267	4.8533
	PVBLiF (ours)	**0.8749**	**0.8580**	0.4660	**0.9060**	**0.8883**	**0.3746**	**0.9554**	**0.9501**	**0.3160**	0.9231	0.8869	0.3568	**0.9214**	**0.8640**	**0.4334**	8.3795
	ASEM-BLiF (ours)	0.8707	0.8509	**0.4654**	0.9048	0.8871	0.3751	0.9422	0.9254	0.3614	**0.9399**	**0.9227**	**0.3069**	0.9082	0.8412	0.4610	5.2377

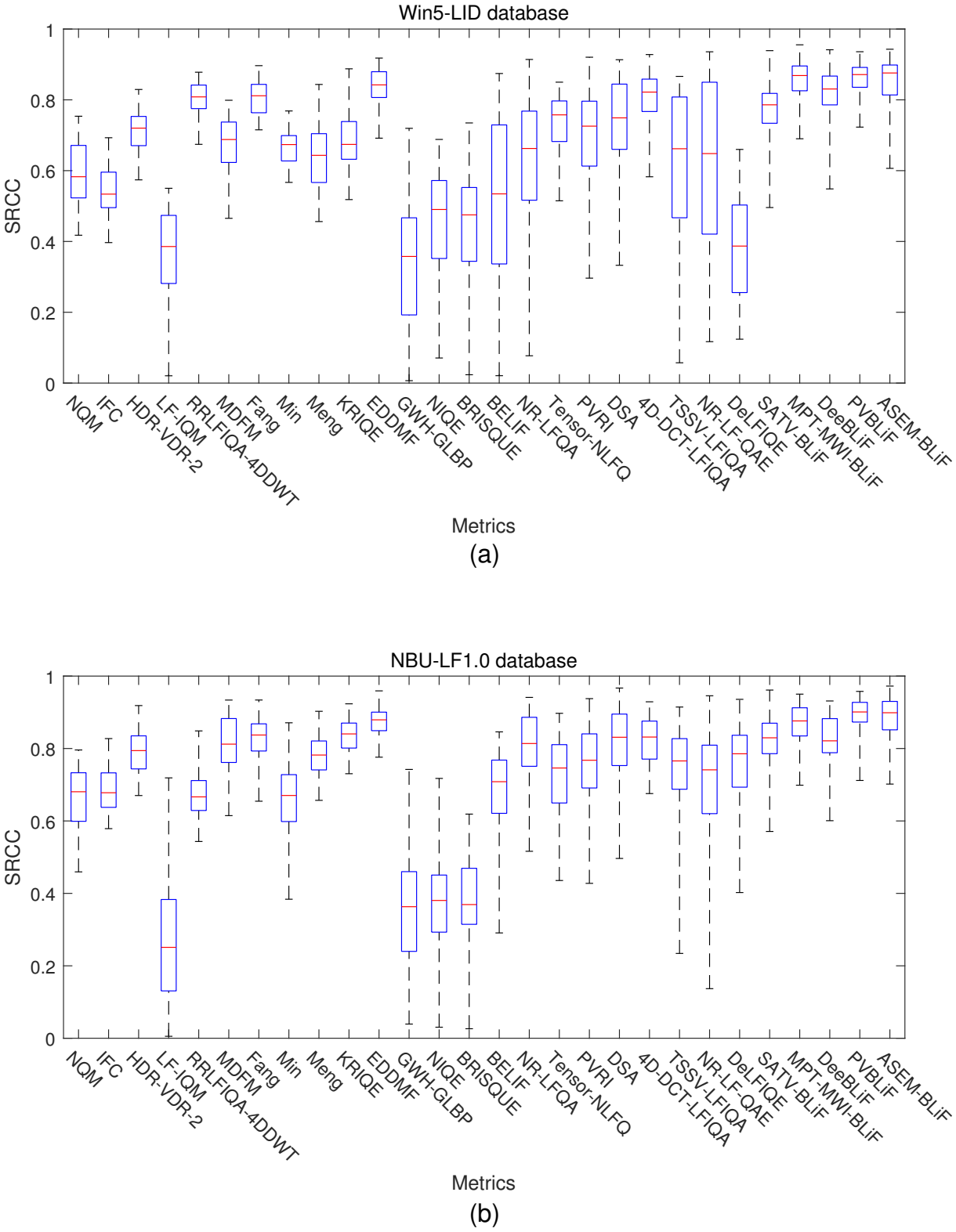


Figure 4.13 – Box plots of SRCC distributions. (a) Win5-LID database; (b) NBU-LF1.0 database.

existing metrics. In general, we are able to observe the superiority of deep learning-based methods in terms of both predictive accuracy and stability.

To further verify the effectiveness of our metrics in quality evaluation, the scatter plots of the objective prediction versus the subjective score of FR/NR LFIQA metrics are provided in Figure 4.14, where the red line represents the perfect prediction. From the scatter plots, we can find that the predictions of our four metrics (Subfigures (f), (r), (s), and (t) of Figure 4.14) are more consistent with the subjective scores than other state-of-the-arts, demonstrating their superior ability to simulate the human visual perception.

4.5.3 Statistical Significance Test

In the subsection, we perform the F-test to compare the statistical significance between any two LFIQA metrics following [96]. This experiment involves a total of 22 RR/FR/NR LFIQA metrics, aiming to provide a comprehensive analysis of statistical significance. Specifically, we first calculate the residuals between objective predictions and subjective scores for each metric. Then, the *vartest*2 function in MATLAB is used to calculate the significance relationship of any two metrics, in which the confidence level is set to 95%. Figure 4.15 presents the statistical significance results on the Win5-LID and NBU-LF1.0 databases, in which "-1", "0", and "1" indicate that the row metric is statistically worse, competitive, and better than the column metric, respectively. From the results, we can see that our proposed metrics DeeBLiF (M20), PVBLiF (M21), ASEM-BLiF (M22), and EDDMF (M8) perform statistically better than most existing LFIQA metrics on both databases. While among these four metrics, the statistical significance performance of PVBLiF, ASEM-BLiF, and EDDMF is slightly better than that of DeeBLiF. These statistical significance results demonstrate the quality evaluation capabilities of our proposed metrics.

4.5.4 Performance on Cross-Database Validation

Since generalization ability is a crucial factor for designing an effective IQA metric, we investigate the cross-database performance of our four proposed metrics following the implementation in Section 3.3.4. The results are shown in Table 4.7. We can find that all four metrics still perform well on HEVC and JPEG2k even when trained on another database, implying a relatively good generalization ability on compression distortions. However, the three NR metrics (DeeBLiF, PVBLiF, and ASEM-BLiF) show relatively

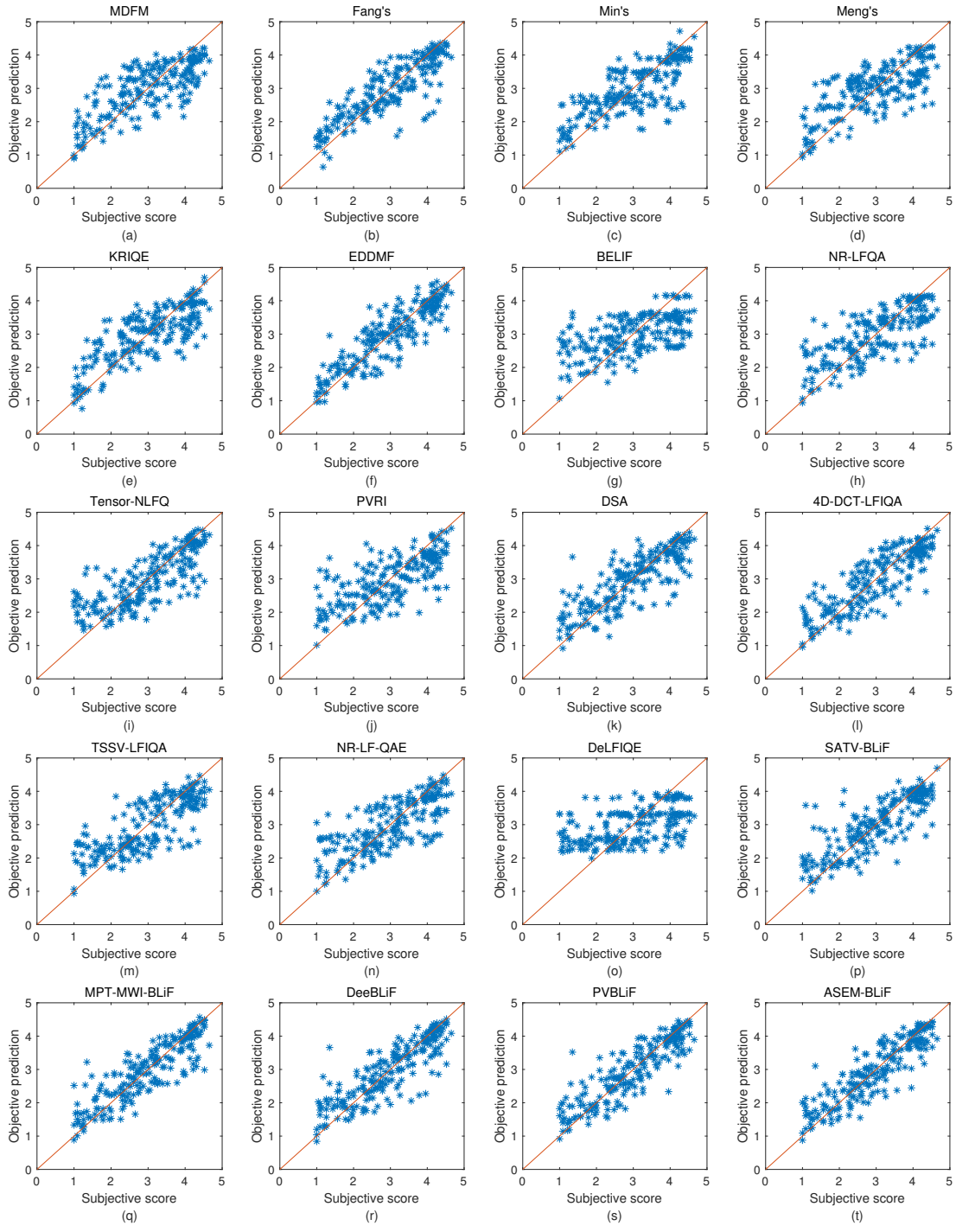
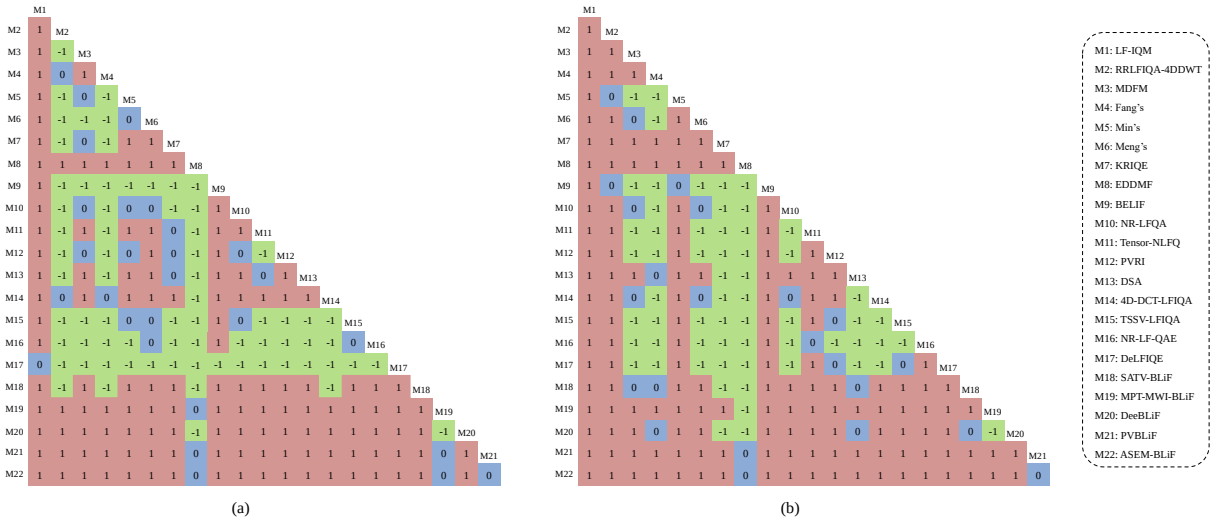


Figure 4.14 – Scatter plots of the objective prediction versus the subjective score of FR/NR LFIQA metrics on the Win5-LID database.



(a) (b)

Figure 4.15 – Statistical significance between any two LFIQA metrics. (a) Win5-LID database; (b) NBU-LF1.0 database. "-1", "0", and "1" indicate that the row metric is statistically worse, competitive, and better than the column metric, respectively.

poor performance on the NN distortion. The reason for this phenomenon has been discussed in [67], *i.e.*, the NN distortion is implemented differently in the Win5-LID and NBU-LF1.0 databases. For the Win5-LID database, the LFI containing 9×9 SAIs is reconstructed in five different distortion levels, each of which involves 50, 40, 30, 20, and 10 randomly selected SAIs to reconstruct the original LFI, respectively. For the NBU-LF1.0 database, the 9×9 SAIs are first down-sampled to 5×5, 3×3, and 2×2 fixed SAIs, respectively, and then the NN interpolation is used for reconstruction. We can easily find that the distortion levels between these two databases do not overlap, which leads to relatively poor performance on the NN distortion. Nonetheless, the FR metric EDDMF still performs well on the NN distortion. This may be because providing reference information effectively bridges the gap between the two databases.

4.5.5 Performance on Individual Distortion Type

An excellent quality assessment metric should show efficacy in various distortion types. Therefore, this subsection investigates and compares the performance of different metrics on different distortion types. The experimental results of PLCC, SRCC, and RMSE for individual distortion type on the Win5-LID and NBU-LF1.0 databases are provided in Table 4.8, Table 4.9, and Table 4.10, respectively. Here, we do not consider two CNN-based distortions in the Win5-LID database because they have only one distortion level.

Table 4.7 – Performance of training on the Win5-LID database, and testing on the NBU-LF1.0, VALID10bit, and SHU databases.

Metric	Test Database	PLCC	SRCC	RMSE
DeeBLiF	NBU-LF1.0 (NN)	0.4929	0.4698	0.9694
	VALID10bit (HEVC)	0.8217	0.7832	0.8011
	SHU (JPEG2k)	0.9133	0.8910	0.2439
PVBLiF	NBU-LF1.0 (NN)	0.7335	0.7017	0.6832
	VALID10bit (HEVC)	0.8469	0.8427	0.7764
	SHU (JPEG2k)	0.8926	0.8578	0.3011
ASEM-BLiF	NBU-LF1.0 (NN)	0.6945	0.5547	0.7513
	VALID10bit (HEVC)	0.7318	0.8585	0.7288
	SHU (JPEG2k)	0.8835	0.9061	0.2120
EDDMF	NBU-LF1.0 (NN)	0.9699	0.9144	0.1765
	VALID10bit (HEVC)	0.8647	0.7815	0.8917
	SHU (JPEG2k)	0.8318	0.8090	0.3456

In general, we can see that the PLCC, SRCC, and RMSE performance show similar rules as concluded in Section 3.3.5, *i.e.*, the FR LFIQA metrics typically perform better than the NR LFIQA metrics for a single distortion. Then, it can be found that among NR LFIQA metrics, deep learning-based metrics generally outperform handcrafted feature-based metrics, while the opposite is true in the FR LFIQA metrics. We believe this may be because traditional features are discriminative enough to evaluate the quality of LFIs with the same distortion type but different distortion levels when reference information is available. In contrast, the advantage of deep features obtained by model training can be fully exploited when reference information is inaccessible. However, the four proposed metrics DeeBLiF, PVBLiF, ASEM-BLiF, and EDDMF achieve state-of-the-art performance in most distortion types. The only poor performance is DeeBLiF's performance on the "Zhang" distortion in the NBU-LF1.0 database. Considering that three extended versions of DeeBLiF (PVBLiF, ASEM-BLiF, and EDDMF) still perform competitively on the "Zhang" distortion, we argue that DeeBLiF's poor performance on the "Zhang" distortion is due to its suboptimal design for angular feature extraction.

Table 4.8 – PLCC performance of different distortion types on the Win5-LID and NBU-LF1.0 databases. The best two results of RR/FR metrics are in blue, while the best two results of NR metrics are marked in **bold**.

	Database	Win5-LID				NBU-LF1.0					
Type	Metric	HEVC	JPEG2k	LN	NN	NN	BI	EPICNN	Zhang	VDSR	Hit-count
FR 2DIQA	NQM [90]	0.9320	0.8701	0.7721	0.5927	0.8331	0.9300	0.8868	0.8810	0.7350	0
	IFC [91]	0.9331	0.8547	0.7719	0.8714	0.9195	0.9348	0.8811	0.7824	0.9326	0
	HDR-VDR-2 [92]	0.9367	0.8766	0.8236	0.8973	0.9397	0.9502	0.8832	0.8762	0.8957	1
RR LFIQA	LF-IQM [60]	0.6172	0.7263	0.8284	0.7158	0.5026	0.6764	0.5976	0.5296	0.5517	0
	RRLFIQA-4DDWT [61]	0.9741	0.9254	0.8804	0.9320	0.8904	0.9545	0.9752	0.9198	0.9762	3
FR LFIQA	MDFM [53]	0.9446	0.8878	0.8819	0.9004	0.9327	0.9463	0.9483	0.9436	0.9226	1
	Fang's [54]	0.9781	0.9365	0.9509	0.9252	0.8861	0.9649	0.9716	0.9064	0.9730	6
	Min's [56]	0.9777	0.9465	0.8798	0.7775	0.8033	0.8796	0.7890	0.8466	0.9664	2
	Meng's [57]	0.9129	0.7442	0.9486	0.8915	0.8989	0.9550	0.9734	0.7308	0.9652	2
	KRIQE [59]	0.9591	0.9185	0.9176	0.9033	0.9021	0.9616	0.9629	0.8399	0.9647	0
	EDDMF (ours)	0.9662	0.9249	0.8747	0.9073	0.9699	0.9660	0.9544	0.9204	0.9534	3
NR 2DIQA	GWH-GLBP [93]	0.7104	0.6002	0.6442	0.4606	0.5709	0.5950	0.7109	0.5233	0.6913	0
	NIQE [94]	0.8361	0.7317	0.6669	0.4331	0.4600	0.5225	0.7018	0.5357	0.9171	0
	BRISQUE [95]	0.8280	0.7629	0.6731	0.5055	0.4981	0.5634	0.7097	0.6310	0.8969	0
NR LFIQA	BELIF [62]	0.8062	0.7275	0.7172	0.7219	0.9244	0.8732	0.6707	0.6886	0.8749	0
	NR-LFQA [63]	0.7641	0.8098	0.7731	0.7920	0.9544	0.9519	0.9157	0.7108	0.8850	0
	Tensor-NLFQ [66]	0.8909	0.8340	0.8543	0.8446	0.8517	0.9199	0.8395	0.7135	0.9223	0
	PVRI [67]	0.8352	0.8702	0.7677	0.7890	0.8396	0.8967	0.8551	0.8268	0.9149	0
	DSA [68]	0.8673	0.8647	0.8338	0.8542	0.9472	0.9243	0.8670	0.8010	0.9342	0
	4D-DCT-LFIQA [69]	0.9001	**0.9365**	0.8803	0.8534	0.9386	0.9389	0.8183	0.9048	0.9317	1
	TSSV-LFIQA [70]	0.8170	0.7761	0.7396	0.7828	0.8924	0.9147	0.8148	0.8108	0.8790	0
	NR-LF-QAE [71]	0.8273	0.7178	0.7500	0.6902	0.8427	0.8503	0.8001	0.6743	0.8027	0
	DeLFIQE [73]	0.6417	0.5808	0.5809	0.5012	0.9554	0.9420	0.9069	0.7671	0.8216	0
	SATV-BLiF (ours)	0.8518	0.9226	0.8713	0.8993	**0.9726**	0.9477	0.9336	0.7538	0.9037	1
	MPT-MWI-BLiF (ours)	0.9429	0.9078	**0.9188**	**0.9379**	**0.9725**	**0.9693**	0.9379	**0.9065**	0.9274	**5**
	DeeBLiF (ours)	0.9389	0.9254	**0.9021**	0.9207	0.9610	0.9499	0.9395	0.6659	0.9487	1
	PVBLiF (ours)	**0.9768**	**0.9388**	0.8941	**0.9286**	0.9688	0.9636	**0.9498**	0.9011	**0.9605**	**5**
	ASEM-BLiF (ours)	**0.9625**	0.9163	0.8994	0.9187	0.9679	**0.9644**	**0.9468**	**0.9138**	**0.9575**	**5**

Table 4.9 – SRCC performance of different distortion types on the Win5-LID and NBU-LF1.0 databases. The best two results of RR/FR metrics are in blue, while the best two results of NR metrics are marked in **bold**.

	Database	Win5-LID				NBU-LF1.0					
Type	Metric	HEVC	JPEG2k	LN	NN	NN	BI	EPICNN	Zhang	VDSR	Hit-count
FR 2DIQA	NQM [90]	0.9344	0.8540	0.7467	0.6601	0.8037	0.9228	0.7587	0.8276	0.7529	0
	IFC [91]	0.9182	0.8111	0.7818	0.8697	0.9200	0.9200	0.8054	0.6136	0.9143	1
	HDR-VDR-2 [92]	0.9568	0.8736	0.8578	0.9062	0.9213	0.9275	0.7831	0.8384	0.8986	4
RR LFIQA	LF-IQM [60]	0.2438	0.5331	0.6977	0.5938	0.3293	0.5164	0.4264	0.4007	0.3382	0
	RRLFIQA-4DDWT [61]	0.9648	0.8999	0.8211	0.8912	0.8276	0.9205	0.8358	0.8182	0.9432	3
FR LFIQA	MDFM [53]	0.9342	0.8378	0.8181	0.8491	0.8681	0.8942	0.7421	0.8432	0.8796	1
	Fang's [54]	0.9679	0.9121	0.8844	0.8869	0.8198	0.9149	0.7908	0.8001	0.9288	3
	Min's [56]	0.9657	0.9188	0.7289	0.4798	0.6402	0.7790	0.5657	0.7786	0.9344	3
	Meng's [57]	0.8878	0.6939	0.8459	0.8001	0.7568	0.9018	0.8034	0.5783	0.9225	0
	KRIQE [59]	0.9494	0.8689	0.8602	0.8584	0.8430	0.9190	0.8185	0.7242	0.9275	2
	EDDMF (ours)	0.9498	0.8855	0.7411	0.8300	0.9144	0.9351	0.8061	0.8188	0.9168	1
NR 2DIQA	GWH-GLBP [93]	0.6573	0.4173	0.4514	0.3392	0.3284	0.4753	0.5975	0.4559	0.6007	0
	NIQE [94]	0.8099	0.6929	0.5195	0.2821	0.3028	0.3485	0.5630	0.3990	0.8577	0
	BRISQUE [95]	0.5931	0.6131	0.4437	0.3259	0.3407	0.3568	0.5769	0.4503	0.8262	0
NR LFIQA	BELIF [62]	0.7760	0.5595	0.5552	0.5447	0.8769	0.8014	0.4816	0.4767	0.8085	0
	NR-LFQA [63]	0.7231	0.7356	0.6371	0.6893	0.9141	0.9027	0.7585	0.6194	0.8457	0
	Tensor-NLFQ [66]	0.8325	0.7740	0.7831	0.7606	0.8041	0.8747	0.6655	0.5861	0.8738	0
	PVRI [67]	0.7968	0.8323	0.6943	0.6714	0.7656	0.7992	0.7354	0.7339	0.8686	0
	DSA [68]	0.8449	0.8241	0.7429	0.7908	0.8990	0.8707	0.7215	0.6688	0.8869	0
	4D-DCT-LFIQA [69]	0.8698	**0.8946**	0.8127	0.8235	0.9027	0.9084	0.6645	**0.8446**	0.8929	2
	TSSV-LFIQA [70]	0.7471	0.6355	0.6437	0.7124	0.8414	0.8468	0.6630	0.6550	0.8023	0
	NR-LF-QAE [71]	0.7982	0.5769	0.6072	0.6120	0.7761	0.7708	0.6103	0.5824	0.7491	0
	DeLFIQE [73]	0.4628	0.4162	0.4337	0.3507	0.9024	0.8916	0.7302	0.6643	0.7192	0
	SATV-BLiF (ours)	0.7988	0.8766	0.7485	0.8540	**0.9282**	0.9190	0.7722	0.6498	0.8617	1
	MPT-MWI-BLiF (ours)	0.9142	0.8813	**0.8429**	**0.8867**	**0.9326**	**0.9304**	0.7347	0.8285	0.8851	4
	DeeBLiF (ours)	0.9103	0.8686	0.7914	0.8628	0.9168	0.8986	0.8027	0.5832	0.9062	0
	PVBLiF (ours)	**0.9768**	**0.9026**	0.8328	0.8750	0.9162	**0.9304**	**0.8347**	0.8138	**0.9335**	**5**
	ASEM-BLiF (ours)	**0.9436**	0.8564	**0.8398**	**0.8762**	0.9169	**0.9335**	**0.8080**	**0.8328**	**0.9086**	**7**

Table 4.10 – RMSE performance of different distortion types on the Win5-LID and NBU-LF1.0 databases. The best two results of RR/FR metrics are in blue, while the best two results of NR metrics are marked in **bold**.

	Database	Win5-LID				NBU-LF1.0					
Type	Metric	HEVC	JPEG2k	LN	NN	NN	BI	EPICNN	Zhang	VDSR	Hit-count
FR 2DIQA	NQM [90]	0.3608	0.3976	0.5261	0.5750	0.4607	0.3475	0.3603	0.3261	0.6579	0
	IFC [91]	0.3552	0.3874	0.5186	0.3525	0.2978	0.3347	0.3608	0.3317	0.3158	0
	HDR-VDR-2 [92]	0.3680	0.4966	0.4652	0.3162	0.2515	0.2954	0.3638	0.2820	0.3934	0
RR LFIQA	LF-IQM [60]	0.8104	0.5730	0.3844	0.4757	0.6264	0.6881	0.5833	0.4547	0.7641	0
	RRLFIQA-4DDWT [61]	0.2129	0.3230	0.3707	0.2387	0.2915	0.2450	0.1536	0.1860	0.1784	4
FR LFIQA	MDFM [53]	0.2916	0.3789	0.3387	0.2728	0.2462	0.2793	0.2123	0.1559	0.3141	2
	Fang's [54]	0.2084	0.2749	0.2423	0.2597	0.3023	0.2268	0.1509	0.2136	0.1885	7
	Min's [56]	0.2122	0.2686	0.3691	0.4419	0.4011	0.4461	0.4240	0.2689	0.2130	2
	Meng's [57]	0.4013	0.5008	0.2470	0.3270	0.3186	0.2580	0.1610	0.3755	0.2180	1
	KRIQE [59]	0.2762	0.3179	0.2921	0.2879	0.2849	0.2301	0.1775	0.2615	0.2175	1
	EDDMF (ours)	0.2619	0.3108	0.3630	0.2849	0.1765	0.2416	0.2133	0.1952	0.2677	1
NR 2DIQA	GWH-GLBP [93]	0.7236	0.6117	0.5614	0.6187	0.5483	0.7549	0.5253	0.4851	0.6279	0
	NIQE [94]	0.5101	0.5285	0.5570	0.6362	0.6531	0.7923	0.5535	0.4930	0.3426	0
	BRISQUE [95]	0.5462	0.5023	0.5419	0.5994	0.6217	0.7618	0.5335	0.4307	0.3733	0
NR LFIQA	BELIF [62]	0.5713	0.5714	0.5111	0.4511	0.2719	0.4341	0.5565	0.4033	0.3935	0
	NR-LFQA [63]	0.6431	0.4904	0.4860	0.4028	0.2044	0.2608	0.2681	0.3857	0.4020	0
	Tensor-NLFQ [66]	0.4565	0.4853	0.3637	0.3737	0.3402	0.3402	0.3739	0.3727	0.3250	0
	PVRI [67]	0.5171	0.4045	0.4444	0.4026	0.3682	0.4042	0.3845	0.3078	0.3346	0
	DSA [68]	0.4853	0.4101	0.4105	0.3139	0.2129	0.3310	0.3330	0.2987	0.3184	0
	4D-DCT-LFIQA [69]	0.4277	**0.2783**	0.3568	0.3415	0.2366	0.3185	0.4184	**0.2206**	0.3142	2
	TSSV-LFIQA [70]	0.5768	0.4910	0.4647	0.3985	0.3142	0.3571	0.3788	0.3408	0.4254	0
	NR-LF-QAE [71]	0.5510	0.5849	0.4388	0.4679	0.3680	0.4572	0.4198	0.4203	0.5171	0
	DeLFIQE [73]	0.7984	0.6995	0.6223	0.6090	0.2023	0.2900	0.3155	0.3343	0.5052	0
	SATV-BLiF (ours)	0.5326	0.3363	0.3616	0.2929	**0.1618**	0.2859	0.2544	0.3738	0.3590	1
	MPT-MWI-BLiF (ours)	0.3340	0.3490	**0.3002**	**0.2345**	**0.1652**	**0.2173**	0.2418	0.2258	0.3114	**4**
	DeeBLiF (ours)	0.3406	0.3257	**0.2964**	0.2701	0.1843	0.2736	**0.2283**	0.4365	0.2614	2
	PVBLiF (ours)	**0.2215**	**0.2887**	0.3324	**0.2526**	0.1669	0.2433	**0.2146**	0.2350	**0.2386**	**5**
	ASEM-BLiF (ours)	**0.2776**	0.3388	0.3488	0.2654	0.1727	**0.2360**	0.2381	**0.2066**	**0.2453**	**4**

Table 4.11 – Ablation studies of different stream combinations in DeeBLiF on the Win5-LID and NBU-LF1.0 databases. The best results are marked in **bold**.

Database	Stream	PLCC	SRCC	RMSE
Win5-LID	Angular	0.8355	0.8088	0.5233
	Spatial	0.8224	0.7974	0.5440
	Angular + Spatial	**0.8427**	**0.8186**	**0.5160**
NBU-LF1.0	Angular	0.8220	0.7995	0.4816
	Spatial	0.8173	0.7811	0.4908
	Angular + Spatial	**0.8583**	**0.8229**	**0.4588**

Table 4.12 – Ablation studies of different stages in PVBLiF on the Win5-LID and NBU-LF1.0 databases. The best results are marked in **bold**.

Database	Stage	PLCC	SRCC	RMSE
Win5-LID	*w/o* stage 1	0.8064	0.7772	0.5754
	w/o stage 2	0.8005	0.7821	0.5823
	w/o stage 3	0.8714	0.8529	0.4911
	PVBLiF	**0.8749**	**0.8580**	**0.4660**
NBU-LF1.0	*w/o* stage 1	0.8019	0.7811	0.4843
	w/o stage 2	0.8345	0.8149	0.4887
	w/o stage 3	0.8997	0.8842	0.3905
	PVBLiF	**0.9060**	**0.8883**	**0.3746**

4.5.6 Ablation Studies

To further explore the effectiveness of each module of our metric, we conduct ablation studies in this subsection.

1) DeeBLiF. The network design in the DeeBLiF metric involves two parallel streams: angular stream and spatial stream. We perform ablation studies of different stream combinations and report the results in Table 4.11. It can be noticed that adopting a two-stream CNN model achieves better performance than adopting an angular or spatial stream alone, which indicates that extracting angular and spatial distorted information separately helps the quality evaluation of distorted LFIs.

2) PVBLiF. In PVBLiF, the proposed PVBSNet contains four stages with different functions, and the efficacy of each stage deserves further investigation. For this goal, ablation studies of multi-information division (stage 1), intra-feature extraction (stage 2), and cross-feature fusion (stage 3) are performed on the Win5-LID and NBU-LF1.0

Table 4.13 – Ablation studies of with and without SVPooling in PVBLiF on the Win5-LID and NBU-LF1.0 databases. The best results are marked in **bold**.

Database	Saliency	Variance	PLCC	SRCC	RMSE
Win5-LID			0.8668	0.8521	0.4762
	✓		0.8694	0.8562	0.4739
		✓	0.8706	0.8540	0.4714
	✓	✓	**0.8749**	**0.8580**	**0.4660**
NBU-LF1.0			0.9018	0.8840	0.3823
	✓		0.9045	**0.8883**	0.3775
		✓	0.9040	0.8848	0.3853
	✓	✓	**0.9060**	**0.8883**	**0.3746**

databases, and the results are reported in Table 4.12. Here, we do not delve into the efficacy of stage 4, as it is deemed necessary in this metric. From the Table, we can see that without stage 1, it is difficult for CNNs to learn discriminative spatio-angular features from the scrambled temporal information in PVS, thus the quality assessment performance drops dramatically. Likewise, stage 2 also plays a crucial role in extracting deep features for quality evaluation. The incorporation of stage 3 promotes a marginal improvement in the final performance.

In addition, PVBLiF assesses the LFI quality by evaluating the quality of each PVBS, thus a post-processing method is required to pool all the PVBS quality into the overall LFI quality. Most previous works, including our DeeBLiF, adopt the widely-used average pooling to obtain the LFI quality. Therefore, to verify the effectiveness of the proposed SVPooling method and its two components (saliency and variance), we conduct ablation studies on the Win5-LID and NBU-LF1.0 databases. The experimental results are reported in Table 4.13. We can see that using saliency or variance alone for post-processing can improve performance to a certain extent. The combination of saliency and variance, *i.e.*, the proposed SVPooling, yields the best quality assessment performance. This may be because the SVPooling method is able to filter out the low-variance PVBSs with inaccurate quality predictions and reweight each PVBS quality according to its visual saliency.

3) ASEM-BLiF. Considering that ASEM-BLiF consists of several modules that contribute to the final quality evaluation performance, we conduct ablation experiments to explore their effectiveness. Table 4.14 reports the experimental results on the Win5-LID and NBU-LF1.0 databases, where w/o is the abbreviation of *without*, and Aux. Learn. denotes the auxiliary learning branch. The results demonstrate a significant decrease in

Table 4.14 – Ablation studies of different modules in ASEM-BLiF on the Win5-LID and NBU-LF1.0 databases. The best results are marked in **bold**.

Database	Module	PLCC	SROCC	RMSE
Win5-LID	*w/o* AEM	0.5465	0.5538	0.8085
	w/o DRS	0.8849	0.8633	0.4552
	w/o Aux. Learn.	0.8816	0.8643	0.4498
	w/o DWLoss	0.9003	0.8854	0.4248
	ASEM-BLiF	**0.9072**	**0.8949**	**0.4090**
NBU-LF1.0	*w/o* AEM	0.6645	0.6254	0.6531
	w/o DRS	0.9024	0.8712	0.3747
	w/o Aux. Learn.	0.9066	0.8773	0.3692
	w/o DWLoss	0.9057	0.8878	0.3681
	ASEM-BLiF	**0.9087**	**0.8952**	**0.3613**

Table 4.15 – Ablation studies of different feature combinations in EDDMF on the Win5-LID and NBU-LF1.0 databases. Here, $\mathcal{B}$ denotes baseline discrepancy features, $\mathcal{H}$ denotes hierarchical discrepancy features, and $\mathcal{L}$ denotes local discrepancy features. The best results are marked in **bold**.

Database	Feature	PLCC	SROCC	RMSE
Win5-LID	$\mathcal{B}$	0.8036	0.7693	0.5618
	$\mathcal{B} + \mathcal{H}$	0.8605	0.8288	0.4906
	$\mathcal{B} + \mathcal{L}$	0.8480	0.8230	0.5068
	$\mathcal{B} + \mathcal{H} + \mathcal{L}$	**0.8654**	**0.8354**	**0.4839**
NBU-LF1.0	$\mathcal{B}$	0.8514	0.7996	0.4706
	$\mathcal{B} + \mathcal{H}$	0.8972	0.8621	0.3934
	$\mathcal{B} + \mathcal{L}$	0.8651	0.8315	0.4574
	$\mathcal{B} + \mathcal{H} + \mathcal{L}$	**0.9024**	**0.8743**	**0.3866**

the performance of our metric on both databases when the AEM module is excluded, indicating the crucial role of capturing angular information for LFIQA. In addition, we can also find that using auxiliary learning without the DRS module (*i.e.*, *w/o* DRS) performs even worse than not using auxiliary learning (*i.e.*, *w/o* Aux. Learn.). A possible explanation is that if all regions are introduced for auxiliary learning, the regions with insufficient angular-spatial information will lead to decreased performance. This demonstrates that the selected discriminative regions are necessary for auxiliary learning. Moreover, incorporating DWLoss into the training process can further promote learning towards better performance. Finally, the combination of all proposed modules culminates in enhanced effectiveness.

4) EDDMF. Since EDDMF combines several features to describe LFI quality degradation, we conduct ablation studies to investigate the efficacy of different feature combinations. Table 4.15 exhibits the experimental results on the Win5-LID and NBU-LF1.0 databases, in which $\mathcal{B}$ denotes baseline discrepancy features, $\mathcal{H}$ denotes hierarchical discrepancy features, and $\mathcal{L}$ denotes local discrepancy features. From the Table, we can observe a significant performance improvement when incorporating $\mathcal{H}$ into $\mathcal{B}$ or $\mathcal{B}$+$\mathcal{L}$, indicating that the angular-dominant hierarchical discrepancy features are quite discriminative for quality assessment. In addition, we can see that incorporating $\mathcal{L}$ into $\mathcal{B}$, *i.e.*, the combination of baseline CNN features and handcrafted features, also significantly improves the performance. Besides, even combining $\mathcal{L}$ with $\mathcal{B}$+$\mathcal{H}$ can yield slight improvement, which demonstrates the effectiveness of the spatial-dominant local discrepancy features. Moreover, it can be found that the boost with $\mathcal{H}$ is more significant than that with $\mathcal{L}$, indicating that the angular-dominant hierarchical discrepancy features contribute more to quality evaluation than the spatial-dominant local discrepancy features. Finally, the combination of $\mathcal{B}$, $\mathcal{H}$, and $\mathcal{L}$ constitutes our proposed EDDMF and consistently achieves the best performance on both databases.

4.6 Conclusion

In this chapter, we delve into integrating deep learning technology into the task of LFI quality evaluation and propose four deep learning-based LFIQA metrics, including three NR metrics (DeeBLiF, PVBLiF, and ASEM-BLiF) and one FR metric (EDDMF). The core idea of the first metric DeeBLiF is to evaluate the LFI quality by evaluating the quality of their 2D patches. After that, PVBLiF extends this idea to 3D and adopts

a more intuitive 3D representation, followed by a more reasonable pooling method based on saliency and variance to obtain the overall LFI quality. In addition, ASEM-BLiF is further developed to address the imbalance problem of large spatial variations and subtle angular variations in LFIs. Finally, we extend DeeBLiF into a FR version EDDMF, in which a combination of angular-dominant deep features and spatial-dominant handcrafted features is presented as a more comprehensive description of LFI distortions. Following Section 3, verification experiments are performed on five standard LFIQA databases to ensure validity and reproducibility. Quantitative experimental results show that our initial metric DeeBLiF already performs better than most existing NR LFIQA metrics. Moreover, the updated versions PVBLiF, ASEM-BLiF, and EDDMF further outperform DeeBLiF in several aspects, such as overall predictive accuracy, time complexity, stability, cross-database performance, and individual distortion robustness. All these results fully demonstrate the great potential of deep learning technology in LFIQA. In the next chapter, we design a novel subjective experiment to evaluate the distorted LFI quality and present a novel perceptual quality assessment database with the associated subjective quality scores.

Chapter 5

Proposed Light Field Image Quality Assessment Database

Currently, objective metrics and subjective experiments are two mainstream methods for LFIQA. Several objective LFIQA metrics have been proposed in previous chapters. Therefore, in this chapter, we focus on conducting subjective experiments to evaluate the quality of distorted LFIs and propose a novel standard LFIQA database. Specifically, the proposed database focuses on the distortions introduced by deep learning-based LFI angular reconstruction and spatial super-resolution methods, individually and multiplely. Besides, in the case of multiple distortions, the adoption order of two distortions is taken into consideration. Further, our database presents three types of LFIs that suffer from distortions: real-world, dense synthesis, and sparse synthesis. As a result, the quality of distorted LFIs was subjectively assessed by 32 valid observers using the Pairwise Comparison (PC) protocol. Secondly, we develop two novel objective NR metrics for LFI quality evaluation, based on the features extracted from spatial gradients, angular-spatial statistics, and binocular disparity. Finally, a benchmark of the proposed metrics and numerous state-of-the-art quality assessment metrics on the proposed database is presented.

This chapter is structured as follows: The limitation analysis of existing databases and the contribution of the proposed database are firstly provided in Section 5.1. Then, Section 5.2 introduces the content of our database. After that, Section 5.3 states the implementations of our subjective experiment. A benchmark of two novel objective NR metrics and numerous objective LFIQA metrics on the proposed database is presented in Section 5.4. In the end, conclusions are drawn in Section 5.5.

5.1 Limitation Analysis and Contribution

Section 2.3 provides detailed descriptions of existing databases, each with different motivations and characteristics. On this basis, this section first introduces the limita-

tion analysis of existing databases and then provides the contributions of our proposed database.

5.1.1 Limitation Analysis of Existing Databases

As mentioned in Section 1.3, obtaining LFIs with high angular resolution by multi-camera arrays or time sequential capture systems is expensive and laborious. Although the emergence of light field cameras has greatly reduced the acquisition cost, the captured LFIs must contend with a trade-off between angular resolution and spatial resolution due to the limited number of imaging sensors. Instead of directly capturing high-resolution LFIs, angular reconstruction and spatial super-resolution are two feasible methods to increase the LFI resolution in the post-processing stage [17]. However, these methods inevitably lead to some unusual distortions in LFIs that differ from those common in 2D images, ultimately affecting the visual experience of human eyes.

At present, significant efforts have been invested in subjective experiments to study the distortion influence caused by angular reconstruction in LFIs [2], [42]. Besides, some subjective experiments investigated the impact of LFI spatial super-resolution on human visual perception [37]. However, existing subjective experiments and their resulting databases have not thoroughly explored the relationship between human visual perception and LFIs distorted by these two types of generative methods. The limitations are summarized as follows:

1) With the widespread adoption of deep learning technologies, many deep learning-based LFI angular reconstruction and spatial super-resolution methods are proposed and achieve automatic feature learning through backpropagation. As a consequence, the associated distortions are also more intricate and difficult to assess. Nevertheless, existing databases do not adequately cover this aspect.

2) Current studies only investigate the impact of angular reconstruction or spatial super-resolution in LFIs individually. However, in practice, these two methods may be implemented in combination and induce multiple distortions. Further, the adoption order of angular reconstruction and spatial super-resolution may also affect the degradation degree of LFI quality. These all deserve further investigation.

3) Studying how different types of LFIs are affected by different types of distortions can help to better understand the HVS. In other words, the limited consideration of the diversity of LFI types restricts the comprehension of HVS to some extent. However, most existing databases only consider a single type of LFIs that suffers from different

distortions.

4) The absence of exploration in the above three aspects results in a lack of standard databases and corresponding objective metrics. In addition, there is a pressing need to establish a benchmark of existing objective metrics to facilitate a comprehensive effectiveness analysis for practical applications.

5.1.2 Contribution of the Proposed Database

To fill the aforementioned blank in the subjective LFIQA database, we conduct subjective experiments by employing the PC protocol for quality scoring, and present a novel LFIQA database with the associated subjective quality scores. Different from previous databases, our database focuses on LFIs that are only disturbed by deep learning-based LFI angular reconstruction and spatial super-resolution methods. A total of 12 types of distortions are considered, encompassing both individual and multiple distortions. Note that when multiple distortions are present, the adoption order of two distortions is also taken into account. In addition, we select three types of source LFIs, including real-world, dense synthesis, and sparse synthesis, and evaluate their perceptual quality after suffering different distortions. For better comparison, we provide a summary of existing LFIQA databases and our proposed database in Table 5.1.

5.2 Database Content

In this section, we provide an in-depth introduction to the content in our proposed database, including Source Sequences (SRCs) and Hypothetical Reference Circuits (HRCs).

5.2.1 Source Sequences

The selection of SRCs plays a crucial role in establishing a high-quality database. In order to cover a variety of practical application scenarios and enhance the accuracy and generalization of subjective test results, it is necessary to ensure the diversity of SRCs. In this work, we consider the diversity of SRCs from two aspects: type diversity and content diversity. For the type diversity, three types of SRCs are involved, including real-world, dense synthesis, and sparse synthesis. For content diversity, several attributes are considered in the selection of SRCs, such as Spatial Information (SI), Colorfulness (CF), contrast, and brightness.

Table 5.1 – Summary of existing LFIQA databases. Distortions induced by deep learning-based methods are marked with *.

Name	Size	Distortion type	Distortion category	Mul. Dis.	No. SRCs	Ang. Res.	Spa. Res.	No. Obs.
SMART [41]	256	JPEG, JPEG2k, HEVC, SSDC	Compression	No	16 (real-world)	15×15	434×625	19
MPI-LFA [42]	336	HEVC	Compression	No	14 (5 real-world and 9 dense synthesis)	1×101	960×720	40
		Gaussian blur	Acquisition or Display					
		DQ, OPT, LN, NN	Angular reconstruction					
VALID [43]	100	HEVC, VP9	Compression	No	5 (real-world)	15×15	434×625	28
		[44], [45], [46]	LFI compression					
Win5-LID [2]	220	HEVC, JPEG2k	Compression	No	10 (6 real-world and 4 dense synthesis)	9×9	434×625 or 512×512	23
		LN, NN	Angular reconstruction					
		EPICNN*, LF-SYN*	LFI angular reconstruction					
NBU-LF1.0 [37]	210	NN, BI	Angular reconstruction	No	14 (8 real-world and 6 dense synthesis)	9×9	434×625 or 512×512	22
		VDSR*	Spatial super-resolution					
		EPICNN*, Zhang	LFI angular reconstruction					
SHU [38]	240	JPEG, JPEG2k	Compression	No	8 (real-world)	15×15	434×625	20
		Gaussian blur, White noise, Motion blur	Acquisition or Display					
LFDD [39]	480	JPEG, JPEG2k, BPG, VP9, AV1, AVC, HEVC	Compression	No	8 (dense synthesis)	9×9	512×512	16
		Gaussian noise, Impulse noise, Barrel, Pincushion, Unsharp mask	Acquisition or Display					
Bakir's [47]	160	HEVC, JEM	Compression	No	10 (real-world)	15×15	434×625	18
		LF-SYN*, [44]	LFI compression					
Ours	120	HDDRNet*, LFASR*	LFI angular reconstruction	Yes	10 (4 real-world, 3 dense synthesis and 3 sparse synthesis)	9×9	434×625 or 512×512	32
		LF-IINet*, DistgSSR*	LFI spatial super-resolution					
		HDDRNet* → LF-IINet*, HDDRNet* → DistgSSR*, LFASR* → LF-IINet*, LFASR* → DistgSSR*	LFI angular reconstruction → LFI spatial super-resolution					
		LF-IINet* → HDDRNet*, LF-IINet* → LFASR*, DistgSSR* → HDDRNet*, DistgSSR* → LFASR*	LFI spatial super-resolution → LFI angular reconstruction					



SRC01 (BouquetFlower1) SRC02 (Duck) SRC03 (Fruits) SRC04 (Symmetric)

(a) Real-world



SRC05 (Bottles) SRC06 (Roses_table) SRC07 (Toy_friends)

(b) Dense synthesis



SRC08 (Chair) SRC09 (Flowers_clock) SRC10 (Flying_furniture)

(c) Sparse synthesis

Figure 5.1 – Illustration of the central view of SRCs in the proposed database.

Specifically, the SRCs of the proposed database consist of 10 reference LFIs, and we provide an illustration of the central view of the selected SRCs in Figure 5.1. Among them, $BouquetFlower1$, $Duck$, $Fruits$, and $Symmetric$ are captured in real-world scenarios [107], $Bottles$, $Roses_table$, and $Toy_friends$ are dense synthesis [108], while $Chair$, $Flowers_clock$, and $Flying_furniture$ are sparse synthesis [108]. The source content has a rich range of coverage, including but not limited to vibrant/dull colors, small/large objects, distant/close-up views, and indoor/outdoor scenes. Further, following [2], [37], we provide the SI and CF values of our selected SRCs to demonstrate the content diversity, as shown in Figure 5.2. Note that the SI and CF values are calculated according to the ITU-T Rec. P.910 [109] and Hasler's method [110], respectively. Although each real-world

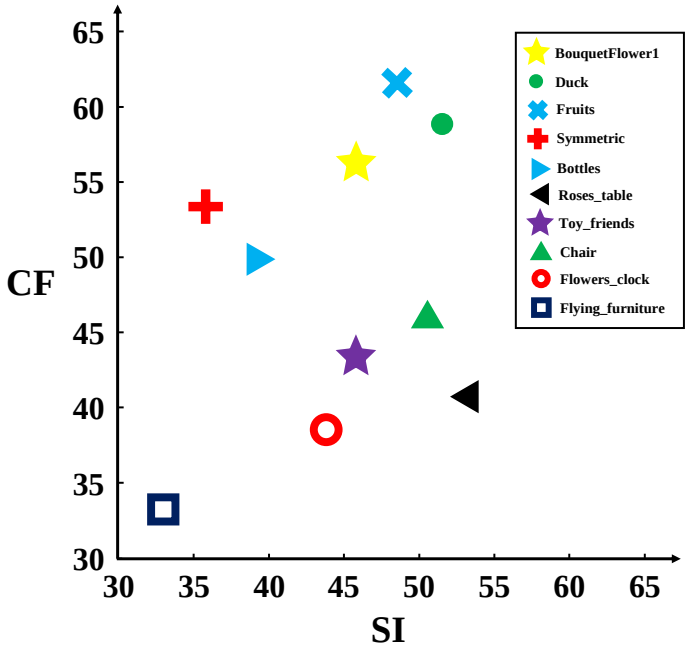


Figure 5.2 – Distribution of the SI and CF values of the selected SRCs.

reference LFI in [48] provides 15×15 angular views, we only keep the central 9×9 angular views to avoid introducing edge views with inherent noise. As a result, the resolution of real-world reference LFIs and other synthesis reference LFIs are 9×9×434×625 and 9×9×512×512, respectively.

5.2.2 Hypothetical Reference Circuits

To investigate the impact of distortions introduced by deep learning-based LFI angular reconstruction and spatial super-resolution methods individually and multiplely, we define 12 Hypothetical Reference Circuits (HRCs) in the proposed database. Given a source LFI, we introduce distortions by halving its angular/spatial resolution and then restoring its original resolution using angular reconstruction/spatial super-resolution methods. Specifically, two angular reconstruction methods (HDDRNet [111] and LFASR [112]) and two spatial super-resolution methods (LF-IINet [113] and DistgSSR [114]) are used. These methods have different technical focuses based on deep learning, and their descriptions are as follows:

HDDRNet proposed by Meng *et al.* [111] is a two-stage scheme using a deep convolutional network with dense connections for LFI angular reconstruction. By progressively performing spatial-angular restoration and high-frequency texture refinement, the model

utilizes the geometry information encoded in angular views to reconstruct LFIs.

LFASR proposed by Jin *et al*. [112] is a totally end-to-end LFI angular reconstruction approach based on CNNs. This approach employs a depth estimation module to capture intrinsic scene geometry information, and utilizes warping and blending modules to produce novel angular views.

LF-IINet proposed by Liu *et al*. [113] develops an intra-inter view interaction network to realize LFI spatial super-resolution. The two parallel branches of the model extract global inter-view information and exploit correlations among intra-view information, respectively, and further interact with each other for better representation learning.

DistgSSR proposed by Wang *et al*. [114] employs a generic mechanism to disentangle LFIs from different dimensions. It is achieved through the utilization of a set of domain-specific convolutions. Subsequently, these disentangled features are combined and applied to perform spatial super-resolution for LFIs.

In our database, we employ these four methods on source LFIs individually and multiplely, resulting in 12 HRCs that involve 4 individual distortions (HDDRNet, LFASR, LF-IINet, and DistgSSR) and 8 multiple distortions (HDDRNet → LF-IINet, HDDRNet → DistgSSR, LFASR → LF-IINet, LFASR → DistgSSR, LF-IINet → HDDRNet, LF-IINet → LFASR, DistgSSR → HDDRNet, and DistgSSR → LFASR), as summarized in TABLE 5.1. Here, the methods on the left and right of the arrow "→" represent the methods used successively. Further, using the two methods in different orders during the generation of multiple distortions are considered as different distortion types. Unlike most existing LFIQA databases, the proposed database no longer considers the severity levels of each distortion type, but focuses more on the relationship between different distortion types. This is mainly because we have observed that most objective metrics are able to distinguish different severity levels within the same distortion type [59], [63], [67]. As a result, our database consists of 120 distorted LFIs.

5.3 Subjective Experiment

Based on the collected LFI content, we conduct a novel subjective experiment to obtain quality scores for all distorted LFIs. This section first presents the methodology of our subjective test, followed by the processing of subjective results. Finally, the analysis of subjective scores is provided.

5.3.1 Methodology of Subjective Test

According to the ITU-T Rec. P.910 [109], there are some recommended testing protocols for subjective quality scoring, mainly including Absolute Category Rating (ACR), Degradation Category Rating (DCR), and Pairwise Comparison (PC). In the ACR and DCR protocols, the observers are required to rate the images directly based on the given scoring table. However, as concluded in [42], these protocols are considered to be insensitive to subtle but noticeable degradation of LFI quality, which always result in observer confusion regarding the selection of the desired rating. Instead, the PC protocol simplifies decision-making by comparing stimuli in pairs and giving relative preferences rather than absolute ratings. Therefore, we adopt the PC protocol in our subjective test.

In each trial, a pair of LFIs is presented horizontally side by side on the same screen to observers, who are then required to make a two-alternative-forced-choice about the one they prefer. Since it is difficult to display all LFI information on the screen at once, the Graphical User Interface (GUI) only displays one view of each LFI. Observers can change their viewing perspective by moving the mouse [2], which simulates the process of eye movement when observing a LFI. Such an interactive method provides observers with greater freedom to perceive the angular and spatial differences between two LFIs and thereby make accurate choices. Note that there is no time limit for observer interactions, but once a selection is made it cannot be changed. Before the formal test, we offer 10 pre-training pairs from other scenarios to familiarize observers with the usage of the interactive GUI.

The proposed database contains 120 distorted LFIs derived from 10 reference LFIs, *i.e.*, each LFI has 12 distorted versions. In the PC protocol, the reference version is treated as one stimuli and hidden in all pairs, thus there are 156 testing pairs for each reference LFI. Since the fixed display order may lead to a decrease in the observer's accuracy in judging latter pairs, we shuffle the display order for each observer. However, according to the three different types of LFIs in our database, we divide all testing pairs into three parts and display them separately. This is based on the consideration that repeatedly switching viewing different types of LFIs may affect the observer's perception. To minimize the negative impact of visual fatigue, observers can take breaks at any time during the test. Nonetheless, observers still need to take a mandatory 5-minute break after completing each part. Generally, the total duration required for each observer to complete all tests will not exceed 90 minutes.

The viewing conditions in our experiments are set based on the ITU-T Rec. P.910

[109] to ensure that the results are reliable. The display device is the LENOVO T2424PA monitor with a resolution of 1920×1080. The region outside the testing pairs is filled with a constant gray color.

Finally, a total of 32 valid observers participated in our subjective experiments, including 21 males and 11 females, ranging in age from 21 to 29. None of the observers had any prior knowledge in the field of image quality evaluation. The normal (or corrected-to-normal) visual acuity and normal color vision of all observers were examined using the Snellen and Ishihara charts, respectively.

5.3.2 Processing of Subjective Results

After completing the PC test, the ranking results of all observers are typically arranged in a count matrix, and the final step is to convert this matrix into a scalar quality score for each distorted LFI. We perform this conversion according to [115]. Specifically, let $\boldsymbol{C}$ denote a D-by-D count matrix, which can be described as:

$$\boldsymbol{C} = \begin{bmatrix} 0 & c_{12} & ... & c_{1D} \\ c_{21} & 0 & ... & c_{2D} \\ ... & ... & ... & ... \\ c_{D1} & c_{D2} & ... & 0 \end{bmatrix} \tag{5.1}$$

where D=13 in our database, representing the sum of one reference and all distorted versions in a single scene. The element c_{ij} denotes the count of cases that the quality of distorted LFI i was selected as better than that of distorted LFI j. Then the probability matrix $\boldsymbol{P}$ can be empirically estimated as:

$$\boldsymbol{P} = \begin{bmatrix} 0 & \hat{p}_{12} & ... & \hat{p}_{1D} \\ \hat{p}_{21} & 0 & ... & \hat{p}_{2D} \\ ... & ... & ... & ... \\ \hat{p}_{D1} & \hat{p}_{D2} & ... & 0 \end{bmatrix} \tag{5.2}$$

where the element $\hat{p}_{ij}$ denotes the probability that the quality of distorted LFI i was selected as better than that of distorted LFI j, which can be described as:

$$\hat{p}_{ij} = \frac{c_{ij}}{c_{ij} + c_{ji}}, \quad i \neq j \tag{5.3}$$

Then, we adopt the Thurstone Case V model [116] as the observer model, in which the perceptual quality of distorted LFI i is assumed as a normal distribution g_i:

$$g_i \sim N(q_i, \sigma_i) \tag{5.4}$$

where the true quality score q_i is assumed to be the distribution mean, and the standard deviation σ_i denotes the combination of inter-observer and intra-observer variances. Thus, the distance between two distorted versions i and j is also considered as a normal distribution g_{ij}:

$$g_{ij} \sim N(q_{ij}, \sigma_{ij}) \tag{5.5}$$

where $g_{ij} = g_i - g_j$, $q_{ij} = q_i - q_j$, and $\sigma_{ij} = \sigma_i^2 + \sigma_j^2$.

In addition, the probability of choosing the quality of distorted LFI i over that of distorted LFI j can be calculated using the cumulative normal distribution based on g_{ij}:

$$\hat{p}_{ij} \approx \boldsymbol{P}(g_i > g_j) = \boldsymbol{P}(g_{ij} > 0) = \Phi(q_{ij}, \sigma_{ij}) \tag{5.6}$$

where $\hat{p}_{ij}$ is approximated as $\boldsymbol{P}(g_i > g_j)$. As shown in Eq. (5.6), given a $\hat{p}_{ij}$ and a σ_{ij}, we can obtain q_{ij}, the distance of true quality scores between distorted LFIs i and j, using the inverse of the cumulative normal distribution function $\Phi(\cdot)$. According to the assumption of the Thurstone Case V model [116], σ_{ij} is a constant, whose selection generally satisfies the case that a probability of 0.75 corresponds to a distance score of 1 Just-Objectionable-Difference (JOD) unit. As a result, σ_{ij} is set to 1.4826 [115].

Based on the above assumptions, the probability that the quality of distorted LFI i was selected over that of distorted LFI j in c_{ij} trials of the total trial number of $n_{ij} = c_{ij} + c_{ji}$ can be described by the binomial distribution:

$$\boldsymbol{P}(\hat{q}_{ij}|c_{ij}, n_{ij}) = \binom{n_{ij}}{c_{ij}} \Phi(\hat{q}_{ij}, \sigma_{ij})^{c_{ij}} (1 - \Phi(\hat{q}_{ij}, \sigma_{ij}))^{n_{ij} - c_{ij}} \tag{5.7}$$

and the estimated quality score $\hat{q}_2$,...,$\hat{q}_n$ can be calculated using the maximum likelihood estimation:

$$\arg\max_{\hat{q}_2,\ldots,\hat{q}_n} \prod_{i,j} \ln(\boldsymbol{P}(\hat{q}_{ij}|c_{ij}, n_{ij})) \tag{5.8}$$

Here, due to the use of the natural logarithm function, most of the obtained quality scores are negative, ranging from -7.7319 to 0.3751. To avoid any ambiguity and confusion caused by negative scores, we directly add an offset value of 8 to all scores to ensure they are

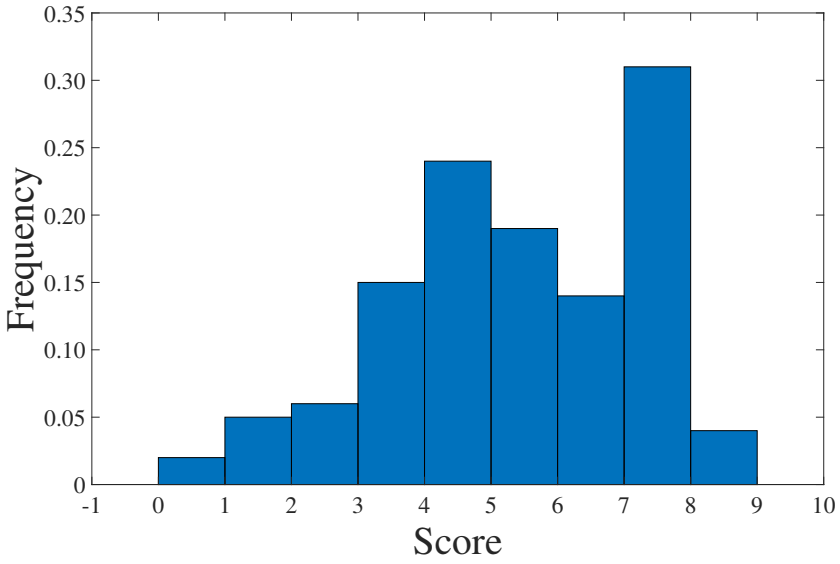


Figure 5.3 – Distribution of all quality scores in our database.

positive. Thus, the final quality scores of our proposed database range from 0.2681 to 8.3751.

5.3.3 Analysis of Subjective Scores

In this subsection, we conduct statistical analyses on the obtained subjective quality scores. To ensure the clarity of expression and simplicity of diagrams, in the following, we will refer to the two involved LFI angular reconstruction methods, HDDRNet and LFASR, as A1 and A2, respectively. Similarly, we will refer to the two involved LFI spatial super-resolution methods, LF-IINet and DistgSSR, as S1 and S2, respectively.

1) Distribution of all quality scores. Figure 5.3 presents the quality score distribution of distorted LFIs in our database. It can be observed that the quality scores of distorted LFIs have a wide and reasonable distribution, further demonstrating the rationality of the distortion type selection and subjective experiment design in our database.

2) Quality score summary of different scenes. Figure 5.4 provides the quality score summary of 12 involved distortion types for different scenes, where (a)-(d) are real-world scenes, (e)-(g) and (h)-(j) are dense and sparse synthetic scenes, respectively. The figure shows that the same distortion has different effects on LFIs in different scenarios, resulting in varying degrees of visual quality degradation. These results are consistent with the conclusion obtained by Adhikarla *et al.* in [42], *i.e.*, the LFI quality is scene-dependent.

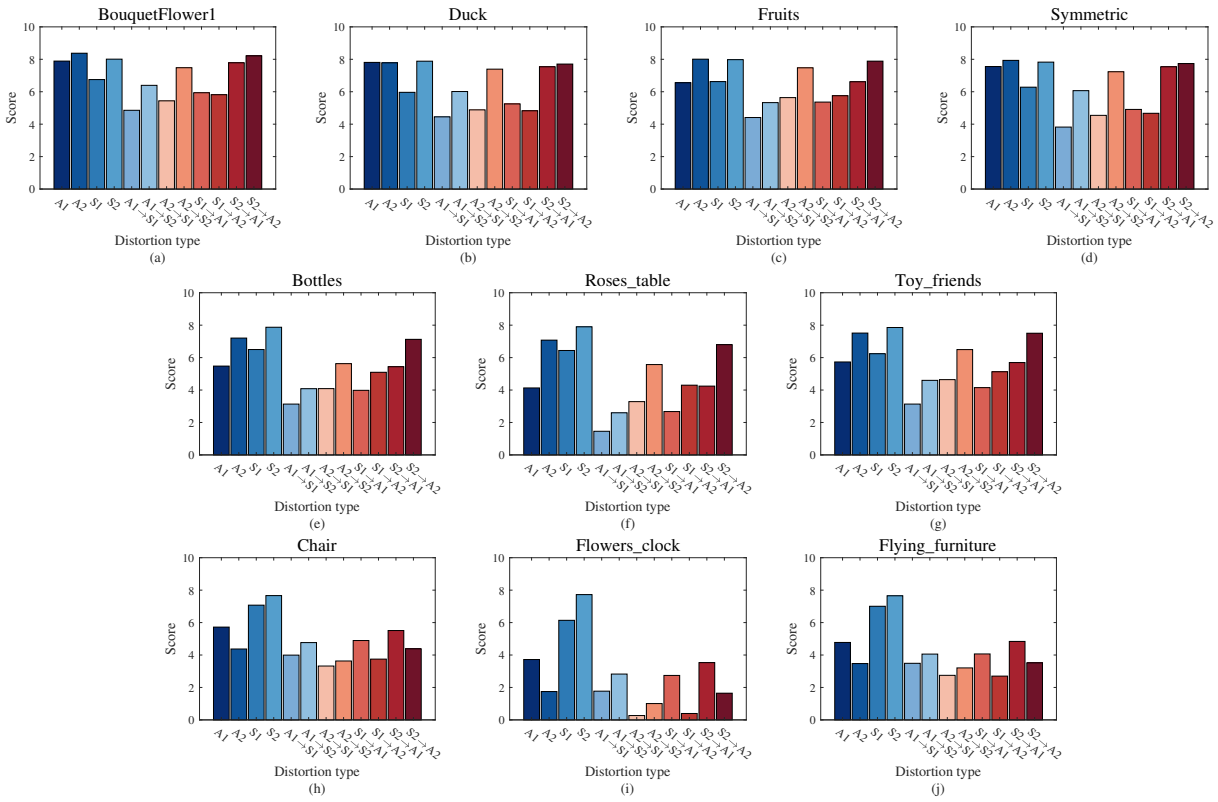


Figure 5.4 – Quality score summary of 12 involved distortions for different scenes. (a) BouquetFlower1; (b) Duck; (c) Fruits; (d) Symmetric; (e) Bottles; (f) Roses_table; (g) Toy_friends; (h) Chair; (i) Flowers_clock; (j) Flying_furniture.

Nonetheless, we can find that when suffering from distortions, real-world scenes tend to have better visual perception than computer-synthesized scenes. A possible explanation is that real-world scenes contain more complex textures and content, which present favorable conditions for generative tasks. Further, we can also see that the visual effect of sparse synthetic scenes is generally poorer than that of dense synthetic scenes. This phenomenon may be attributed to the fact that existing LFI angular reconstruction methods (including HDDRNet and LFASR used in our database) mainly focus on LFIs with narrow disparities, and have limited reconstruction performance for LFIs with large disparities.

3) Relationship between the impact of individual and multiple distortions. In Figure 5.5, we summarize the quality score distribution with respect to different generative methods used in our database, where (a)-(d) are related to HDDRNet (A1), LFASR (A2), LF-IINet (S1), and DistgSSR (S2) methods, respectively. For the distortions induced by the HDDRNet method (Figure 5.5 (a)), the individual distortion A1 has similar visual

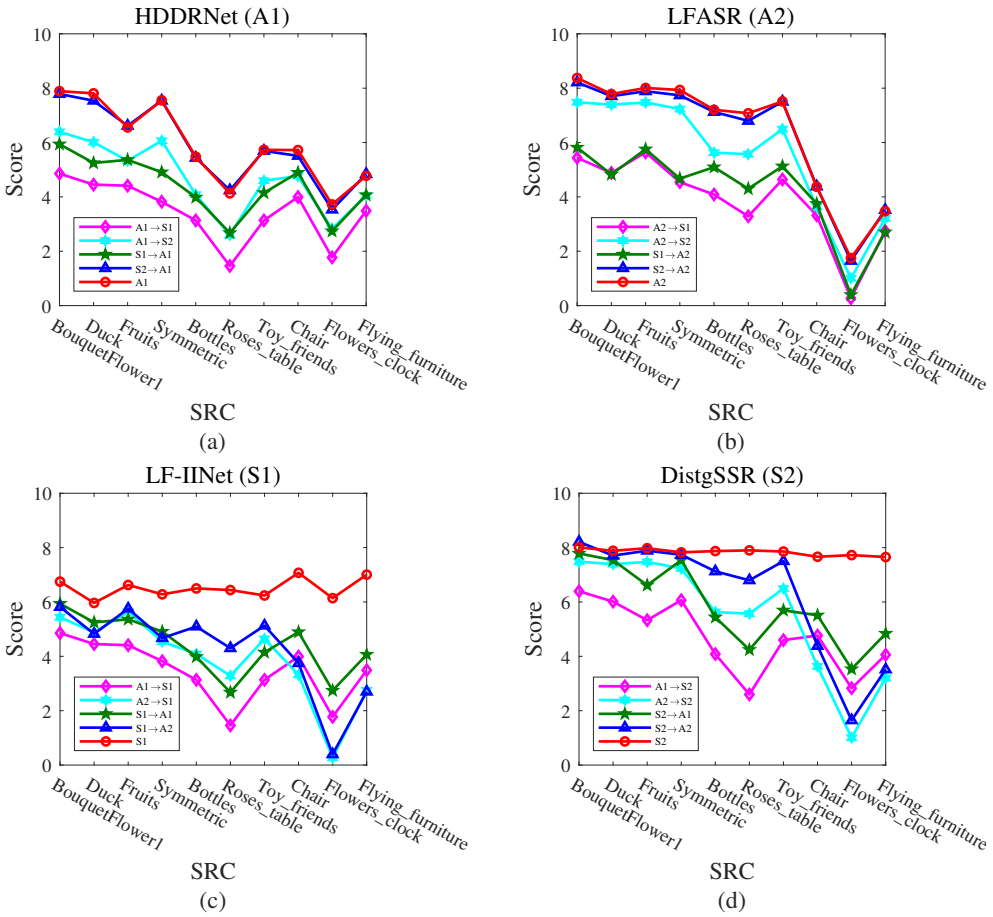


Figure 5.5 – Quality score distribution with respect to different generative methods used in our database. (a) HDDRNet; (b) LFASR; (c) LF-IINet; (d) DistgSSR.

quality scores as the multiple distortion S2→A1, while causing significantly less damage to visual perception compared to other three multiple distortions (A1→S1, A1→S2, and S1→A1). Similar conclusions can be drawn in the case of LFASR (Figure 5.5 (b)). However, when the two LFI spatial super-resolution methods LF-IINet (Figure 5.5 (c)) and DistgSSR (Figure 5.5 (d)) are used, the individual distortion consistently exhibits higher visual quality scores than all multiple distortions. Overall, these score distributions are consistent with our expectation, *i.e.*, whenever LFIs suffer an additional distortion, their visual quality should decrease accordingly.

4) Impact of different adoption orders of two distortions in multiple distortions. The combined use of LFI angular reconstruction and spatial super-resolution methods introduces multiple distortions in LFIs, in which the impact of different adoption orders of two distortions is investigated in this work. As shown in Figure 5.6, (a)-(d) present the

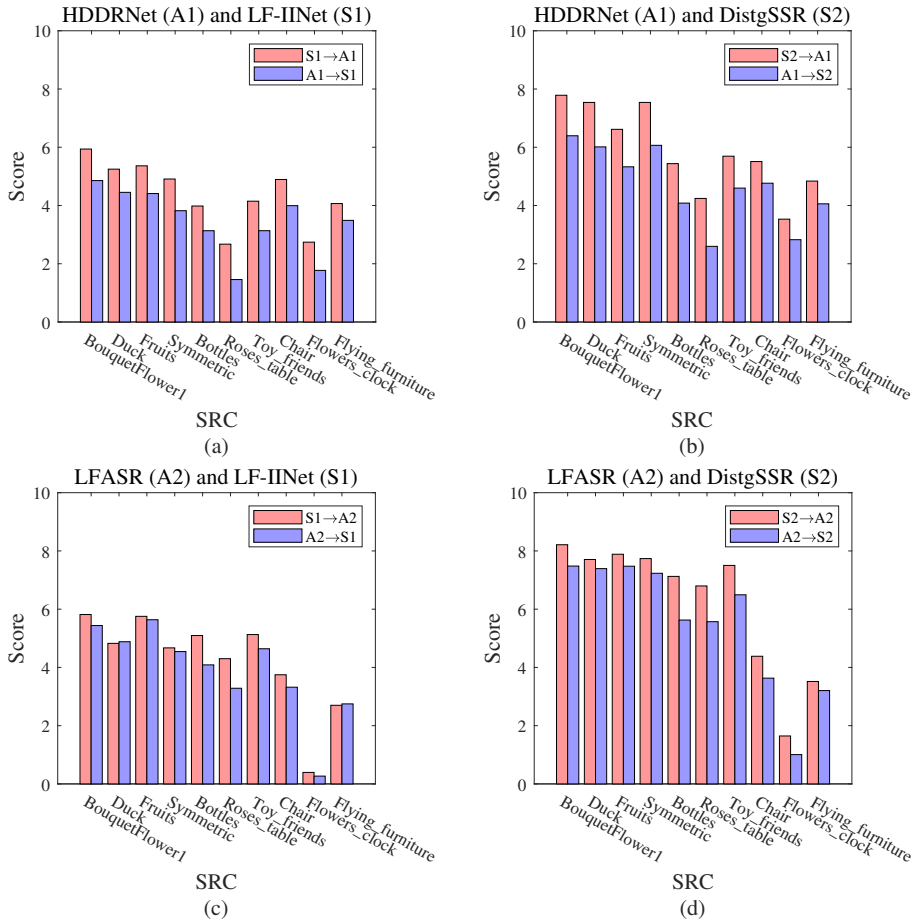


Figure 5.6 – Quality score distribution with respect to multiple distortions induced by two different methods. (a) HDDRNet and LF-IINet; (b) HDDRNet and DistgSSR; (c) LFASR and LF-IINet; (d) LFASR and DistgSSR.

quality score distribution with respect to multiple distortions induced by two different methods respectively: (a) HDDRNet and LF-IINet; (b) HDDRNet and DistgSSR; (c) LFASR and LF-IINet; (d) LFASR and DistgSSR. We can clearly see that adopting a LFI spatial super-resolution method followed by a LFI angular reconstruction method generally obtains a better visual effect than adopting the reverse order. We believe it is mainly because LFIs contain richer spatial information than angular information (or have higher spatial resolution than angular resolution). This inherent attribute enables the LFI spatial super-resolution task to utilize more spatial information to expand spatial resolution and achieve better visual perception, which benefits the subsequent LFI angular reconstruction task. On the contrary, if the LFI angular reconstruction method is

adopted first, it would lead to limited reconstruction performance and poor visual effects due to inadequate information. The above findings provide certain guidance for practical applications involving multiple distortions.

5.4 Objective Measurement

The high-dimensional characteristics of LFIs lead to intricate visual perception when LFIs are subject to distortions. Therefore, we need to comprehensively evaluate the visual effect of distorted LFIs from multiple aspects. In this section, we propose a novel objective NR LFIQA metric (SAB) by exploring Spatial gradients, Angular-spatial statistics, and Binocular disparity. To meet the real-time needs of some practical purposes, a lightweight version of the SAB metric is further presented. Finally, a benchmark of the two proposed metrics and numerous existing LFIQA metrics on our proposed database is established.

5.4.1 Proposed SAB Metric

Figure 5.7 describes the block diagram of the proposed SAB metric, which exploits the quality-aware features of distorted LFIs from three aspects: spatial gradients, angular-spatial statistics, and binocular disparity. Finally, all extracted features are combined and converted into quantified quality scores using SVR.

1) Spatial Gradient Feature Extraction. The spatial information in LFIs has similar characteristics to that in traditional 2D images, and it tends to suffer from blur-like distortions caused by spatial-related tasks, such as spatial super-resolution. These kinds of distortions often modify the local anisotropy and smooth the edge sharpness of images. Considering the well-documented effectiveness of image gradients in assessing blur-like distortions [117], we utilize the Relative Gradient Orientation (RGO) and Relative Gradient Magnitude (RGM) to measure the local anisotropy and edge sharpness, respectively, in order to evaluate the LFI quality.

Specifically, for a given LFI $\boldsymbol{L} = L(u, v, h, w)$, we first convert it into a stack of grayscale Sub-Aperture Images (SAIs) $\boldsymbol{S} = \{S_{u,v}(h, w)\}$ focusing on spatial information. For simplicity, we denote each single SAI as $S(h, w)$ without u and v in this subsection. Firstly, the Gradient Orientation (GO) of $S(h, w)$ is calculated as:

$$\angle \nabla S(h, w) = \arctan(\frac{S_y(h, w)}{S_x(h, w)}) \tag{5.9}$$

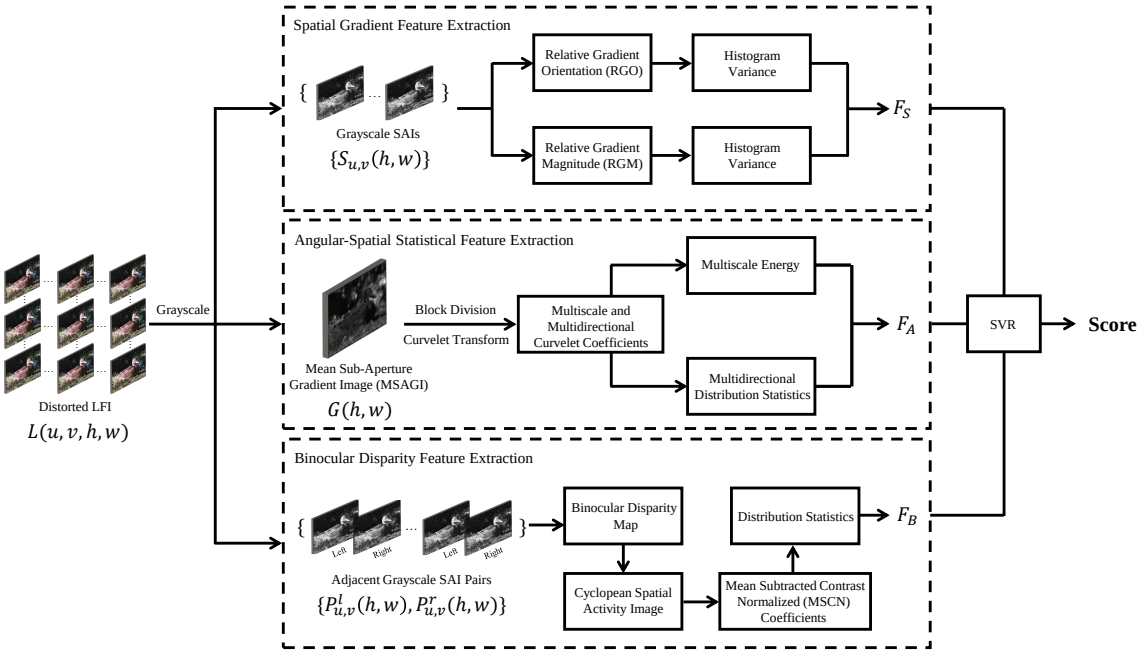


Figure 5.7 – Block diagram of the proposed SAB metric.

where $S_x(h,w)$ and $S_y(h,w)$ represent the approximate values of the directional derivatives in the horizontal and vertical directions, respectively. Similarly, the local average GO of $S(h,w)$ is calculated as:

$$\angle\nabla\bar{S}(h,w) = \arctan\left(\frac{\bar{S}_y(h,w)}{\bar{S}_x(h,w)}\right) \tag{5.10}$$

where the horizontal and vertical average directional derivatives, $\bar{S}_x(h,w)$ and $\bar{S}_y(h,w)$, are calculated on the 3×3 neighboring pixels (*i.e.*, M=N=3), with a set of relative coordinate offsets Ω={(-1,-1), (-1,0), (-1,1), (0,-1), (0,0), (0,1), (1,-1), (1,0), (1,1)}. Accordingly, $\bar{S}_x(h,w)$ and $\bar{S}_y(h,w)$ are described as:

$$\bar{S}_x(h,w) = \frac{1}{M \times N} \sum_{(m,n)\in\Omega} S_x(h-m, w-n) \tag{5.11}$$

$$\bar{S}_y(h,w) = \frac{1}{M \times N} \sum_{(m,n)\in\Omega} S_y(h-m, w-n) \tag{5.12}$$

The RGO and RGM of $S(h,w)$ are then calculated as:

$$\angle\nabla S(h,w)_{RGO} = \angle\nabla S(h,w) - \angle\nabla\bar{S}(h,w) \tag{5.13}$$

$$|\nabla S(h,w)|_{RGM} = \sqrt{\begin{array}{l}(S_x(h,w)-\bar{S}_x(h,w))^2 \\ \quad +(S_y(h,w)-\bar{S}_y(h,w))^2\end{array}} \tag{5.14}$$

Subsequently, inspired by Ruderman's work [118], we utilize the histogram variance of RGO and RGM as the quality-aware features to measure the spatial degradation of $S(h,w)$:

$$V_{RGO} = var(hist(\angle\nabla S(h,w)_{RGO})) \tag{5.15}$$

$$V_{RGM} = var(hist(|\nabla S(h,w)|_{RGM})) \tag{5.16}$$

In addition to obtain the V_{RGO} and V_{RGM} of $S(h,w)$ with the original size, we also consider these two variances from the down-sampled version of $S(h,w)$, denoted as V'_{RGO} and V'_{RGM}, respectively. Finally, we combine all these variances to obtain the spatial gradient features $\mathcal{F}_S$:

$$\mathcal{F}_S = [V_{RGO}, V_{RGM}, V'_{RGO}, V'_{RGM}] \tag{5.17}$$

Note that the spatial gradient features of the whole LFI are obtained by averaging the spatial gradient features of all SAIs.

2) Angular-Spatial Statistical Feature Extraction. The angular distortions induced by LFI angular reconstruction methods affect visual perception in an implicit manner. Generally, angular distortions are perceived in conjunction with spatial information, as they are revealed only when refocusing or changing viewpoints. Therefore, we advocate prioritizing angular-spatial distortions rather than angular distortions. Following [69], given a grayscale LFI, we first calculate the horizontal and vertical angular differences $D^{hor}_{u,v}(h,w)$ and $D^{ver}_{u,v}(h,w)$ to mitigate the impact of the extensive information redundancy among different SAIs:

$$D^{hor}_{u,v}(h,w) = S_{u+1,v}(h,w) - S_{u,v}(h,w) \tag{5.18}$$

$$D^{ver}_{u,v}(h,w) = S_{u,v+1}(h,w) - S_{u,v}(h,w) \tag{5.19}$$

The Mean Sub-Aperture Gradient Image (MSAGI), denoted as $G(h,w)$, is then calculated as:

$$G_{u,v}(h,w) = \sqrt{D^{hor}_{u,v}(h,w)^2 + D^{ver}_{u,v}(h,w)^2} \tag{5.20}$$

$$G(h,w) = \underset{u,v}{avg}(G_{u,v}(h,w)) \tag{5.21}$$

where $\underset{u,v}{avg}(\cdot)$ is the average operation in the angular domain.

Curvelet transform offers multiscale and multidirectional representations with its efficient sparse expression capabilities, making it a powerful analysis tool in the field of image processing. Also, Curvelet transform is a commonly-used decomposition method in quality assessment research [68]. Thus, we convert the MSAGI into multiscale and multidirectional Curvelet coefficients $C_{s,d}$ based on a set of image blocks with size $B \times B$:

$$C_{s,d} = \sum_{0 \le h,w < B} G(h,w)\overline{\Phi_{s,d}(h,w)} \tag{5.22}$$

where s and d are the parameters related to scale and direction. $\overline{\Phi_{s,d}(h,w)}$ denotes the conjugate function of the basis Curvelet transform. $C_{s,d}$ contains 5 scales and each scale has different directional information. $C_{1,d}$ and $C_{5,d}$ record low-frequency information and high-frequency information, respectively, with only one direction. $C_{4,d}$ contains the richest detailed information with 64 directions. In our implementation, B is set to 256.

Then, we extract statistical features based on $C_{s,d}$ for quality assessment. First, considering that distortions typically disrupt the energy distribution at different scales, we calculate the multiscale energy as quality-aware features $\mathcal{F}^s_{eng}$:

$$\mathcal{F}^s_{eng} = \log_{10} |C_{s,d}| \tag{5.23}$$

where $s \in \{1,2,3,4,5\}$. Thus, the length of $\mathcal{F}^s_{eng}$ is 5.

Besides, the distribution changes of multidirectional information also reflect the LFI quality to a certain extent. To this end, we compute the multidirectional distribution statistics on $C_{4,d}$, as it consists of the richest 64 directional information. Specifically, we divide the 64 directions into four equal parts in sequential order, and calculate the kurtosis and skewness of the first three discriminative parts, resulting in $\mathcal{F}_{kur}$ and $\mathcal{F}_{ske}$:

$$\mathcal{F}_{kur} = [kur(C_{4,1-16}), kur(C_{4,17-32}), kur(C_{4,33-48})] \tag{5.24}$$

$$\mathcal{F}_{ske} = [ske(C_{4,1-16}), ske(C_{4,17-32}), ske(C_{4,33-48})] \tag{5.25}$$

Consequently, the angular-spatial statistical features $\mathcal{F}_A$ are generated by combining $\mathcal{F}^s_{eng}$, $\mathcal{F}_{kur}$, and $\mathcal{F}_{ske}$:

$$\mathcal{F}_A = [\mathcal{F}^s_{eng}, \mathcal{F}_{kur}, \mathcal{F}_{ske}] \tag{5.26}$$

Note that, $\mathcal{F}_A$ is an 11-dimensional feature vector.

3) Binocular Disparity Feature Extraction. When distortion is introduced, LFIs with

different disparities will produce different visual effects. Therefore, we propose to consider the binocular visual effect and extract the binocular disparity features, which provide a complementary description for the degradation of angular-spatial information in LFIs. With a given grayscale LFI, we first combine each pair of horizontally adjacent SAIs and generate a large number of pseudo stereoscopic image pairs, denoted as $\{P^l_{u,v}(t), P^r_{u,v}(t)\}$, where $t{=}(h, w)$ are spatial coordinates. For each stereopair, we generate the binocular disparity map $d(t)$ by maximizing the structural similarity score [119]. Then, following [120], we define $O^l(t)$ as a $A \times A$ (A is set to 17 as default [120]) neighborhood window of the left image $P^l(t)$, and $\varepsilon(O^l(t))$ is the spatial activity of $O^l(t)$:

$$\varepsilon(O^l(t)) = \log_2(\sigma^2(O^l(t)) + 1) + \delta \tag{5.27}$$

where δ is set to 0.01 to ensure the stability of the subsequent calculations. $O^r(t)$ and $\varepsilon(O^r(t))$ of the right image $P^r(t)$ can be defined in a similar manner. After that, we calculate the cyclopean spatial activity image $I(t)$ as:

$$I(t) = \frac{\varepsilon(O^l(t)) \cdot P^l(t) + \varepsilon(O^r(t + d(t))) \cdot P^r(t + d(t))}{\varepsilon(O^l(t)) + \varepsilon(O^r(t + d(t)))} \tag{5.28}$$

The generated cyclopean spatial activity image is capable of capturing and preserving the differences between left and right views, especially the visual discontinuity regions between the two views. Then we calculate the saliency-weighted Mean Subtracted Contrast Normalized (MSCN) coefficients $\hat{I}(t)$ of the cyclopean spatial activity image $I(t)$ as:

$$\hat{I}(t) = \frac{1}{1 + |\nabla d(t)|} \cdot \frac{I(t) - \mu(I(t))}{\sigma(I(t)) + 1} \tag{5.29}$$

where $|\nabla d(t)|$ denotes the gradient magnitude of $d(t)$. $\mu(\cdot)$ and $\sigma(\cdot)$ represent the calculations of local mean and standard deviation, respectively. Also, based on the consideration that the changes of the statistical distributions of MSCN coefficients are associated with the quality deterioration, we compute the kurtosis and skewness of $\hat{I}(t)$ as the binocular disparity features $\mathcal{F}_B$:

$$\mathcal{F}_B = [kur(\hat{I}(t)), ske(\hat{I}(t))] \tag{5.30}$$

The average binocular disparity features of all stereopairs are calculated as the binocular disparity features of the whole LFI.

4) Quality Score Prediction. After the aforementioned feature extractions, we obtain

the overall quality-aware features $\mathcal{F}_{SAB} = [\mathcal{F}_S, \mathcal{F}_A, \mathcal{F}_B]$ for the proposed SAB metric, and $\mathcal{F}_{SAB-light} = [\mathcal{F}_S, \mathcal{F}_A]$ for its lightweight version. Finally, the quality score is predicted by the SVR with a radial basis function kernel:

$$Q_{SAB} = SVR(\mathcal{F}_{SAB}) \tag{5.31}$$

$$Q_{SAB-light} = SVR(\mathcal{F}_{SAB-light}) \tag{5.32}$$

where $SVR(\cdot)$ denotes the trained SVR model.

5.4.2 Benchmark of Objective Metrics

In this subsection, we establish a relatively comprehensive benchmark on the proposed database following the experimental settings in Chapter 3 and Chapter 4. The objective metrics involved in this benchmark include the two proposed metrics (SAB and SAB-light) and a total of 22 objective LFIQA metrics covering multiple types: two RR LFIQA metrics (LF-IQM [60] and RRLFIQA-4DDWT [61]), six FR LFIQA metrics (MDFM [53], Fang's [54], Min's [56], Meng's [57], KRIQE [59], and EDDMF), and fourteen NR LFIQA metrics (BELIF [62], NR-LFQA [63], Tensor-NLFQ [66], PVRI [67], DSA [68], 4D-DCT-LFIQA [69], TSSV-LFIQA [70], NR-LF-QAE [71], DeLFIQE [73], SATV-BLiF, MPT-MWI-BLiF, DeeBLiF, PVBLiF, and ASEM-BLiF). Among them, EDDMF, DeLFIQE, DeeBLiF, PVBLiF, and ASEM-BLiF are deep learning-based, while other metrics are based on handcrafted features.

TABLE 5.2 exhibits the benchmark of objective metrics in terms of quality assessment performance and running time on the proposed database. Although most LFIQA metrics share a common emphasis on the degradation of angular-spatial information, they show significant differences in quality assessment capabilities. These findings indicate that different quality assessment metrics exhibit different sensitivities to multiple distortions, and also highlight the guiding significance of our database in the selection of objective LFIQA metrics. Typically, due to the availability of reference information, RR/FR metrics show superior quality evaluation performance than NR metrics. In addition, the NR metrics based on deep learning generally outperform those based on handcrafted features. However, despite the use of handcrafted features, the proposed SAB metric still performs better than most FR/RR metrics or deep learning-based metrics, without incorporating any reference information. This may be because our SAB metric considers the potential

Table 5.2 – Overall performance comparison and time complexity on the proposed database. The best results of RR/FR metrics are in blue, while the best results of NR metrics are marked in **bold**.

Type	Metric	PLCC	SRCC	RMSE	Time (s)
RR LFIQA	LF-IQM [60]	0.4506	0.3409	1.4498	589.7851
	RRLFIQA-4DDWT [61]	0.8500	0.8251	0.8223	128.1106
FR LFIQA	MDFM [53]	0.6367	0.4864	1.1822	0.8537
	Fang's [54]	0.8935	0.8712	0.7100	1.1574
	Min's [56]	0.8987	0.8617	0.7204	3.9845
	Meng's [57]	0.7684	0.6677	1.0075	30.4872
	KRIQE [59]	0.8815	0.8557	0.7523	115.7841
	EDDMF (ours)	0.9469	0.9234	0.5171	1.8775
NR LFIQA	BELIF [62]	0.7915	0.7335	0.9887	107.8814
	NR-LFQA [63]	0.8043	0.7624	0.9065	225.2069
	Tensor-NLFQ [66]	0.7705	0.6787	1.0375	697.6515
	PVRI [67]	0.8520	0.8148	0.8366	71.3578
	DSA [68]	0.8739	0.8386	0.7992	198.5443
	4D-DCT-LFIQA [69]	0.8080	0.7583	0.9644	169.2623
	TSSV-LFIQA [70]	0.8054	0.7494	0.9661	44.6696
	NR-LF-QAE [71]	0.7876	0.7427	0.9533	254.1088
	DeLFIQE [73]	0.7349	0.6795	1.0807	15.4193
	SATV-BLiF (ours)	0.8494	0.8064	0.8552	3.8812
	MPT-MWI-BLiF (ours)	0.8576	0.8210	0.8587	26.8342
	DeeBLiF (ours)	0.8813	0.8434	0.7653	4.8533
	PVBLiF (ours)	0.8948	0.8743	0.7091	8.3795
	ASEM-BLiF (ours)	0.9023	0.8490	0.7094	5.2377
	SAB-light (ours)	0.8838	0.8661	0.7485	**3.6741**
	SAB (ours)	**0.9144**	**0.8934**	**0.6693**	133.3048

impact of LFI distortions from multiple aspects. Further, both our SAB metric and its lightweight version achieve superior performance compared to the metrics that also utilize handcrafted features. In addition to the quality evaluation performance, the running time of each metric is also provided in TABLE 5.2. We can see that the handcrafted feature-based NR LFIQA metrics tend to be more time-consuming compared to the deep learning-based ones, but this comparison is somewhat unfair because deep learning-based methods require significant extra time and more powerful parallel computing platforms to train a model. Among the NR metrics based on handcrafted features, our SAB metric has no clear advantage in running time although it has the best quality evaluation performance. However, the SAB-light metric achieves an optimal trade-off by effectively reducing running time without significantly compromising predictive accuracy.

Then, we provide the box plots of SRCC distributions for all metrics on the proposed

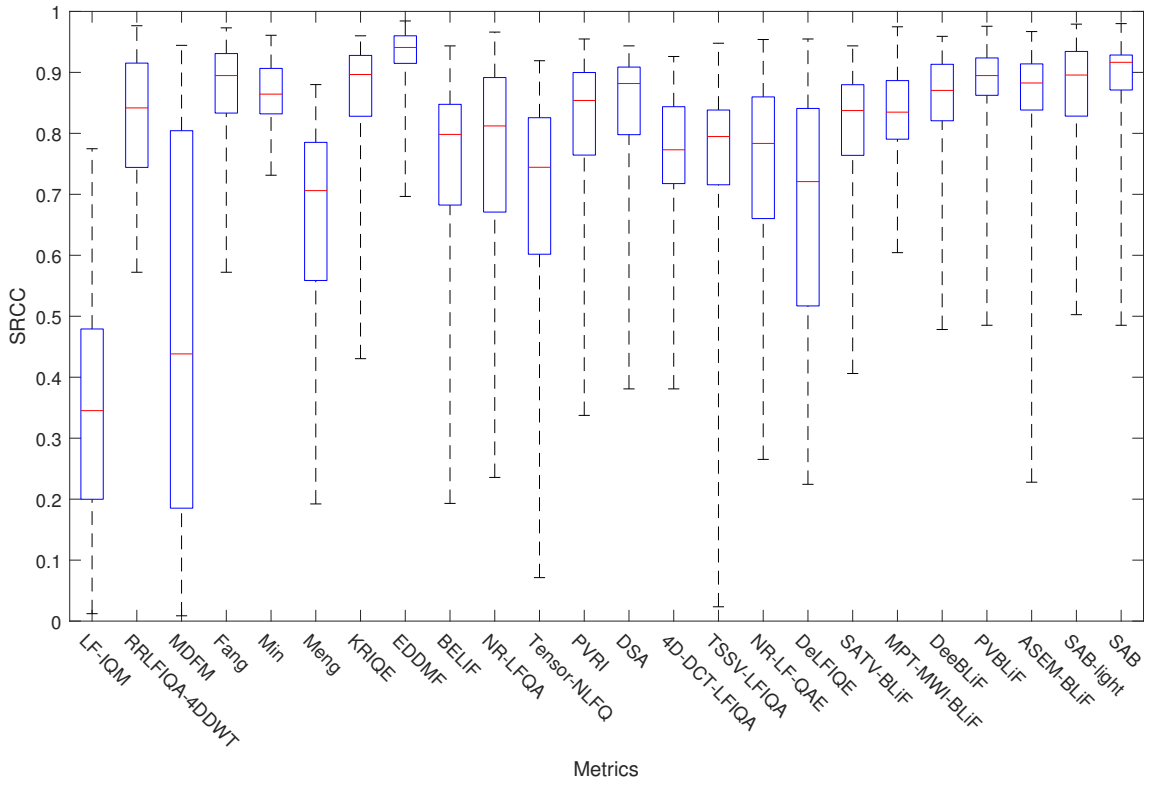


Figure 5.8 – Box plots of SRCC distributions on the proposed database.

database in Figure 5.8. For a specific box, shorter length and higher vertical location represent better stability and predictive accuracy, respectively. Accordingly, we can see that the proposed SAB metric and its lightweight version have remarkable stability and quality evaluation ability, outperforming most existing handcrafted feature-based metrics. Compared to our proposed deep learning-based metrics, SAB and SAB-light perform competitively with our NR metric DeeBLiF, PVBLiF, and ASEM-BLiF, and only perform worse than our FR metric EDDMF.

Additionally, a F-test is performed to study the statistical significance relationship between any two LFIQA metrics. The statistical significance results are provided in Figure 5.9, where "-1", "0", and "1" indicate that the row metric is statistically worse, competitive, and better than the column metric, respectively. We can see that our SAB metric statistically outperforms all existing NR LFIQA metrics. Besides, the SAB-light metric performs significantly better than all handcrafted feature-based NR LFIQA metrics, while obtaining competitive results compared to most deep learning-based NR LFIQA metrics.

To further demonstrate the superiority of the proposed metric in quality evaluation,

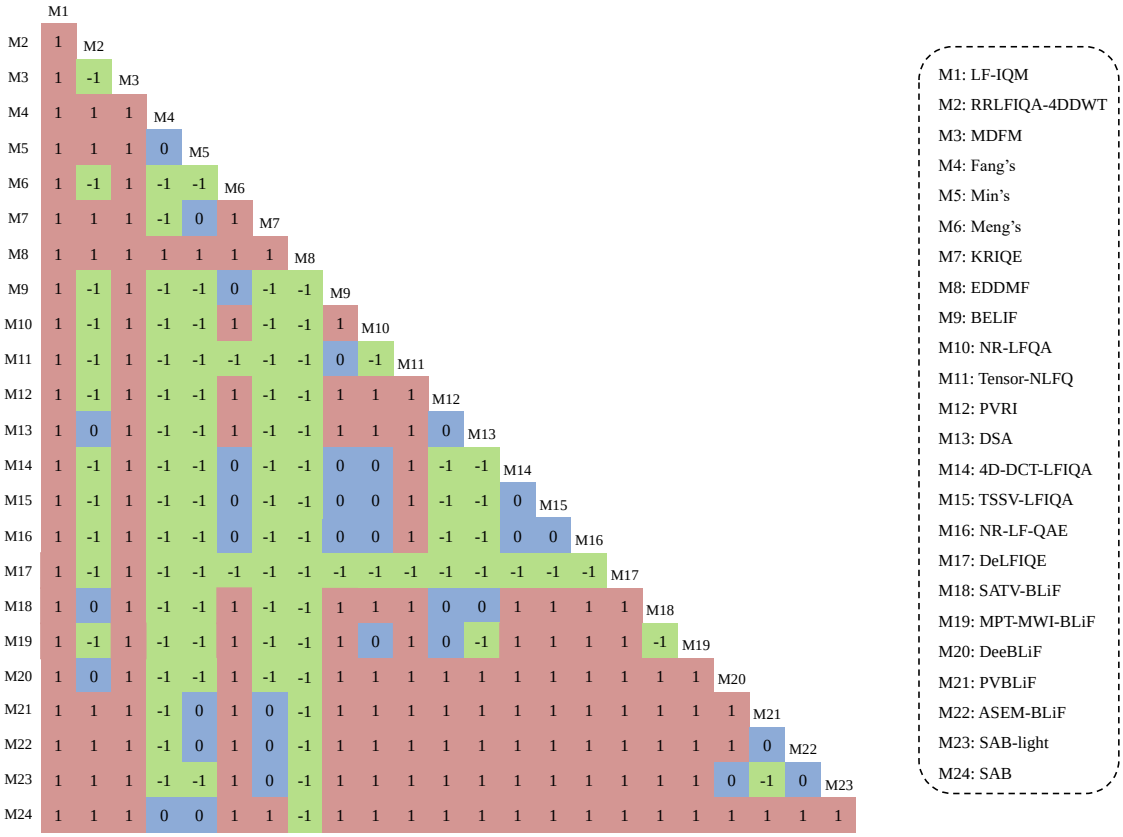


Figure 5.9 – Statistical significance between any two LFIQA metrics on the proposed database. "-1", "0", and "1" indicate that the row metric is statistically worse, competitive, and better than the column metric, respectively.

Figure 5.10 presents the scatter plots of predicted scores versus subjective scores of LFIQA metrics. We can see that the scatter plots of our SAB metric show very close proximity to the red line (*i.e.*, perfect predictions) with only fewer outliers. These results provide strong evidence of its effectiveness.

Finally, to investigate the effectiveness of different features in our SAB metric, ablation studies are conducted on the proposed database, as shown in TABLE 5.3. We can see that both $\mathcal{F}_S$ and $\mathcal{F}_A$ show significant efficacy in quality assessment. The combination of these two features constitutes the SAB-light metric and achieves remarkable quality evaluation performance. Further, it could also be observed that relying solely on $\mathcal{F}_B$ for quality assessment does not yield satisfactory results, as $\mathcal{F}_B$ only includes 2 quality-aware values. Intuitively, too few features may fail to provide a comprehensive description of LFI distortions. However, the incorporation of $\mathcal{F}_B$ with any other features consistently leads to

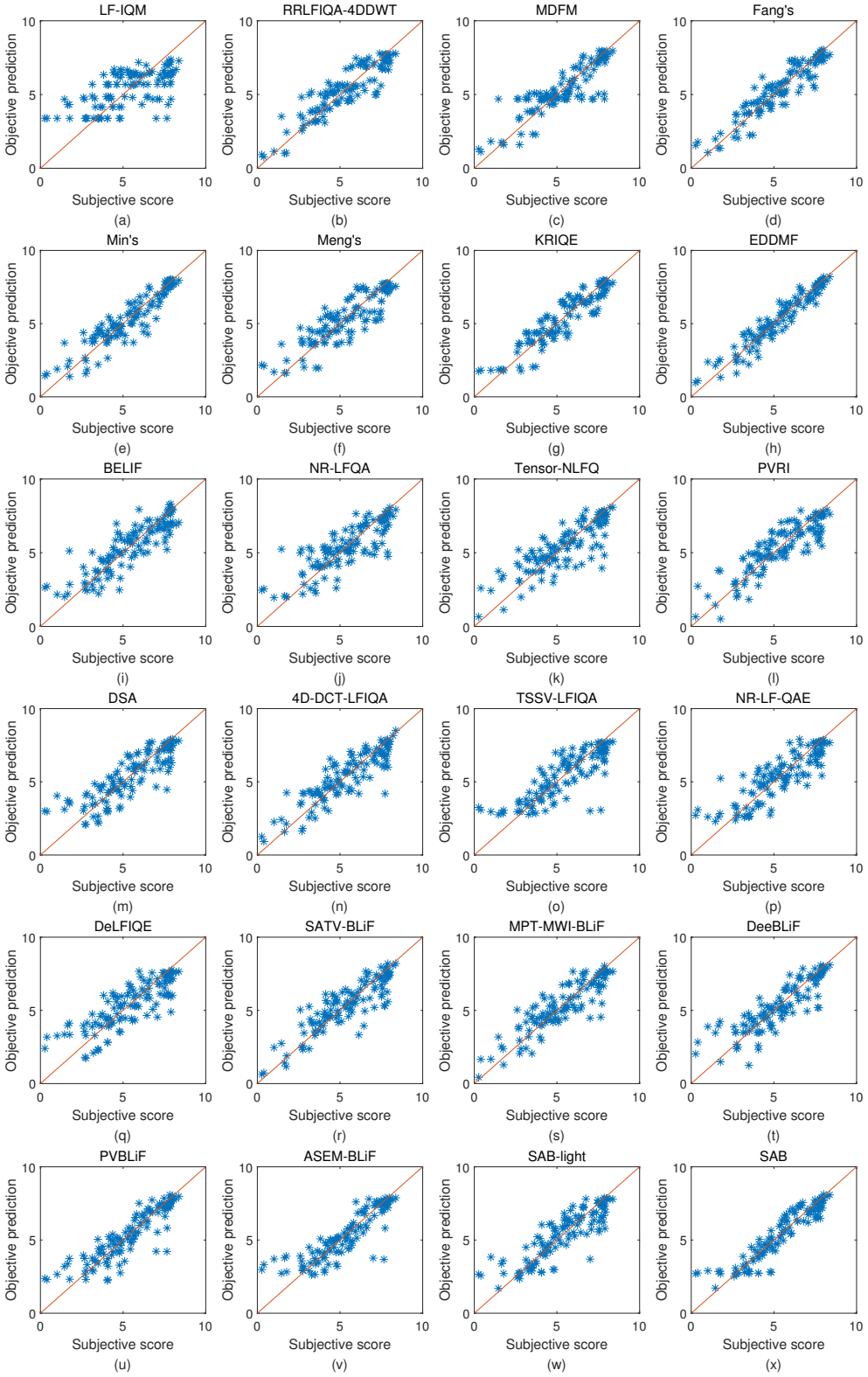


Figure 5.10 – Scatter plots of the objective prediction versus the subjective score of LFIQA metrics on the proposed database.

Table 5.3 – Ablation studies of different feature combinations on the proposed database. The best results are marked in **bold**.

$\mathcal{F}_S$	$\mathcal{F}_A$	$\mathcal{F}_B$	PLCC	SRCC	RMSE
✓			0.8114	0.7661	1.0026
	✓		0.7961	0.7445	0.9531
		✓	0.4882	0.3797	1.4500
✓	✓		0.8838	0.8661	0.7485
✓		✓	0.8509	0.7987	0.8798
	✓	✓	0.8328	0.7947	0.8681
✓	✓	✓	**0.9144**	**0.8934**	**0.6693**

noticeable performance improvements, indicating its capability to serve as discriminative supplementary information. Finally, By combining $\mathcal{F}_S$, $\mathcal{F}_A$, and $\mathcal{F}_B$, the SAB metric is formed, resulting in the best quality evaluation performance.

5.5 Conclusion

In this section, we fill the gap in subjective experimental studies of multiple distortions in LFIs, and therefore propose a new perceptual quality assessment database for LFIs. Specifically, the proposed database focuses on the distortions induced by deep learning-based LFI angular reconstruction and spatial super-resolution methods, individually and multiplely. The impact of different adoption orders of the two distortions in multiple distortions is also studied in our database. Further, our database collects three different types of LFIs to present a comprehensive study on the human visual perception of LFIs. As a result, 10 reference scenes and 120 distorted LFIs subject to various distortion types are included in our database. Through subjective testing with the participation of 32 observers, we obtain subjective quality scores and conduct a thorough statistical analysis on these scores. In addition, by exploring quality-aware features from spatial gradients, angular-spatial statistics, and binocular disparity, we propose a novel NR LFIQA metric named SAB and also develop its lightweight version with the consideration of efficiency. Finally, a benchmark of the two proposed metrics and numerous state-of-the-art LFIQA metrics on the proposed database is established. Experimental results demonstrate the superiority of our proposed metrics.

Conclusions and Perspectives

Brief Summary

As a novel imaging technology, LFI has received extensive attention from both academia and industry, further offering the possibility for a wide range of applications. However, throughout the LFI processing chain, LFIs inevitably suffer from some unusual distortions that are different from those commonly seen in 2D images. To this end, LFIQA plays a crucial role in providing dependable feedback to control the effect of distortions in LFIs.

This thesis is dedicated to evaluating the perceptual quality of LFIs through objective metrics and subjective tests. For objective metrics, we propose two handcrafted feature-based LFIQA metrics and four deep learning-based metrics for LFI quality evaluation. For subjective tests, we conduct subjective experiments to assess the LFI quality and propose a novel LFIQA database. Finally, a benchmark of numerous state-of-the-art objective metrics on the proposed database is presented.

Contributions and List of Publications

The main contributions and novelties of this thesis are summarized below.

1. A simple yet effective Blind Light Field image quality assessment metric by measuring Spatio-Angular Textural Variations (SATV-BLiF) is proposed (Chapter 3). This metric exploits the LBP descriptor to capture spatio-angular textural variations, and calculates statistical textural histogram features for LFI quality evaluation. Experimental results show that the SATV-BLiF metric has much lower time complexity than most existing NR LFIQA metrics, while achieving competitive quality evaluation performance in most scenarios.
2. A Blind Light Field image quality assessment metric by exploring MultiPlane Texture and Multilevel Wavelet Information (MPT-MWI-BLiF) is developed (Chapter 3). As an extended version of the SATV-BLiF metric, the proposed MPT-MWI-BLiF metric provides more comprehensive texture information from various planes and further considers wavelet information to assess the LFI quality. Experimental

results show that the MPT-MWI-BLiF metric consistently outperforms its initial version SATV-BLiF and most existing NR LFIQA metrics by a large margin.

3. A patch-based Deep Blind Light Field image quality assessment metric (DeeBLiF) is proposed (Chapter 4). It generates 2D spatio-angular patches to address the problem of insufficient LFI training data, and introduces a two-stream CNN model to fully excavate the spatio-angular information in these patches. Experimental results show that the DeeBLiF metric outperforms most existing handcrafted feature-based LFIQA metrics, demonstrating the effectiveness of introducing deep learning technology into LFIQA.

4. A Pseudo Video-based Blind quality assessment metric for Light Field images (PVBLiF) is designed (Chapter 4). Instead of using 2D representations, this metric explores a more intuitive 3D representation for LFI quality evaluation. Further, a HVS-inspired pooling method based on saliency and variance is proposed to generate the overall LFI quality. Experimental results exhibit that compared to all existing LFIQA metrics, the PVBLiF metric comprehensively obtains superior quality assessment performance with relatively low time complexity.

5. A Blind Light Field image quality assessment metric based on Angular-Spatial Effect Modeling (ASEM-BLiF) is proposed (Chapter 4). To improve the performance of LFI quality evaluation, the ASEM-BLiF metric focuses on alleviating the imbalance problem of large spatial variations and subtle angular variations in LFIs. Experimental results show that the ASEM-BLiF metric performs better than most existing LFIQA metrics, and its low time complexity makes it possible to implement in real scenarios.

6. An Efficient Deep Discrepancy Measuring Framework (EDDMF) for full-reference light field image quality assessment is proposed (Chapter 4). This metric is also extended from the DeeBLiF metric, in which a combination of angular-dominant deep features and spatial-dominant handcrafted features is exploited to provide a more comprehensive description of LFI distortions. Experimental results demonstrate the superiority of the EDDMF metric in terms of overall predictive accuracy, time complexity, stability, cross-database performance, and individual distortion robustness.

7. A subjective experiment is conducted and a new LFIQA database is provided (Chapter 5). The proposed database focuses on the distortions introduced by deep

learning-based LFI angular reconstruction and spatial super-resolution methods, individually and multiplely. Also, the adoption order of two distortions is considered in multiple distortions. Further, our database provides three types of LFIs that suffer from distortions: real-world, dense synthesis, and sparse synthesis. The quality of distorted LFIs was subjectively assessed by 32 valid observers using the PC protocol, which resulted in a new LFIQA database.

8. Based on the quality-aware features extracted from spatial gradients, angular-spatial statistics, and binocular disparity, an objective NR LFIQA metric (SAB) and its lightweight version are designed (Chapter 5). In addition, we present a relatively comprehensive benchmark of state-of-the-art objective LFIQA metrics (including FR, RR, and NR) on the proposed database, which provides certain guidance and inspiration for future LFIQA research.

So far, our work resulted in seven publications (including three published journal papers, two published conference papers, and two journal papers under review), which are described as follows.

Three published journal papers:

— Z. Zhang, S. Tian, W. Zou, L. Morin, and L. Zhang, «EDDMF: An efficient deep discrepancy measuring framework for full-reference light field image quality assessment», IEEE Trans. on Image Process., vol. 32, pp. 6426-6440, 2023.

— Z. Zhang, S. Tian, W. Zou, L. Morin, and L. Zhang, «PVBLiF: A pseudo video-based blind quality assessment metric for light field image», IEEE J. Sel. Topics Signal Process., vol. 17, no. 6, pp. 1193-1207, 2023.

— Z. Zhang, S. Tian, Y. Zhang, W. Zou, L. Morin, and L. Zhang, «Blind perceptual quality assessment of LFI based on angular-spatial effect modeling», IEEE Trans. Broadcast., vol. 70, no. 1, pp. 290-304, 2024.

Two published conference papers:

— Z. Zhang, S. Tian, W. Zou, L. Morin, and L. Zhang, «DeeBLiF: Deep blind light field image quality assessment by extracting angular and spatial information», in Proc. IEEE Int. Conf. Image Process., 2022, pp. 2266-2270.

— Z. Zhang, S. Tian, W. Zou, Y. Zhang, L. Morin, and L. Zhang, «Blind quality assessment of light field image based on spatio-angular textural variation», in Proc. IEEE Int. Conf. Image Process., 2023, pp. 2385-2389.

Two journal papers under review:

— Z. Zhang, S. Tian, J. Zhou, L. Morin, and L. Zhang, «A new benchmark database

and objective metric for light field image quality evaluation», IEEE Trans. on Circuits Syst. Video Technol., 2024.
— Z. Zhang, S. Tian, J. Xiang, W. Zou, L. Morin, and L. Zhang, «Blind light field image quality assessment by exploring multiplane texture and multilevel wavelet information», Pattern Recognit., 2024.

Perspectives

Although certain efforts have been made in this thesis, the research LFIQA is still far from ideal due to various factors. Therefore, we put forward some perspectives on future research directions as follows.

1. The subjective perceptual quality of LFIs affected by multiple distortions is investigated in this work. However, research on objective evaluation methods of LFI quality is still blank. Considering the prevalence of multiple distortions in LFIs in practice, it would be interesting to develop objective LFIQA metrics focusing on multiple distortions.
2. All deep learning-based objective LFIQA metrics proposed in this work are based on patches/blocks, with the core goal of enlarging the training set to obtain sufficient training samples. However, this method inevitably weakens the connection between image regions. To address this shortcoming, objective metrics based on end-to-end manners are the future research trend.
3. With the emergence of handheld light field cameras, acquiring LFIs has become convenient and effortless. However, it is time-consuming and laborious to conduct subjective experiments and establish LFIQA databases with associated quality scores. Therefore, designing unsupervised objective LFIQA metrics using massive unlabeled LFI data is very attractive for practical use.
4. The subjective scores obtained from traditional 2D/3D displays may not be able to fully reflect the characteristics of human eyes, since LFIs are not presented in their original 4D form. In the future, it would be necessary to collect subjective opinions of observers using a light field display.

Sources primary

[1] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, « A dataset and evaluation methodology for depth estimation on 4d light fields », *in Proc. Asian Conf. Comput. Vis.*, 2016, pp. 19–34.

[2] L. Shi, S. Zhao, W. Zhou, and Z. Chen, « Perceptual evaluation of light field image », *in Proc. of IEEE Int. Conf. Image Process.*, 2018, pp. 41–45.

[3] G. Arun, « The light field », *J. Math. Phys.*, vol. 18, pp. 51–151, 1936.

[4] E. H. Adelson and J. R. Bergen, « The plenoptic function and the elements of early vision », *in Proc. Comput. Models Vis. Process.*, 1991, pp. 3–20.

[5] M. Levoy and P. Hanrahan, « Light field rendering », *in Proc. Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 31–42.

[6] C. Conti, L. D. Soares, and P. Nunes, « Dense light field coding: a survey », *IEEE Access*, vol. 8, pp. 49 244–49 284, 2020.

[7] I. Ihrke, J. Restrepo, and L. Mignard-Debise, « Principles of light field imaging: briefly revisiting 25 years of research », *IEEE Signal Process. Mag.*, vol. 33, *5*, pp. 59–69, 2016.

[8] B. Wilburn, N. Joshi, V. Vaish, *et al.*, « High performance imaging using large camera arrays », *ACM Trans. Graph.*, vol. 24, *3*, pp. 765–776, 2005.

[9] M. Broxton, J. Flynn, R. Overbeck, *et al.*, « Immersive light field video with a layered mesh representation », *ACM Trans. Graph.*, vol. 39, *4*, p. 86, 2020.

[10] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, « Scene reconstruction from high spatio-angular resolution light fields », *ACM Trans. Graph.*, vol. 32, *4*, p. 73, 2013.

[11] S. G. Laboratory, « The new stanford light field archive », 2020. [Online]. Available: http://lightfield.stanford.edu/index.html.

[12] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, « Light field photography with a hand-held plenoptic camera », *Comput. Sci. Tech. Rep.*, vol. 2, *11*, pp. 1–11, 2005.

[13] Project CLIM. [Online]. Available: http://clim.inria.fr/DataSoftware.html.

[14] Raytrix. [Online]. Available: https://raytrix.de/.

[15] Blender. [Online]. Available: https://www.blender.org/.

[16] Unreal Engine. [Online]. Available: https://www.unrealengine.com/.

[17] G. Wu, B. Masia, A. Jarabo, *et al.*, « Light field image processing: an overview », *IEEE J. Sel. Topics Signal Process.*, vol. 11, *7*, pp. 926–954, 2017.

[18] S. Alamgeer and M. C. Q. Farias, « A survey on visual quality assessment methods for light fields », *Signal Process., Image Commun.*, vol. 110, p. 116 873, 2023.

[19] G. K. Wallace, « The jpeg still picture compression standard », *Commun. ACM*, vol. 34, *4*, pp. 30–44, 1991.

[20] C. Christopoulos, A. Skodras, and T. Ebrahimi, « The jpeg2000 still image coding system: an overview », *IEEE Trans. Consum. Electron.*, vol. 46, *4*, pp. 1103–1127, 2000.

[21] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, « Overview of the h.264/avc video coding standard », *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, *7*, pp. 560–576, 2003.

[22] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, « Overview of the high efficiency video coding (hevc) standard », *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, *12*, pp. 1649–1668, 2012.

[23] B. Bross, Y.-K. Wang, Y. Ye, *et al.*, « Overview of the versatile video coding (vvc) standard and its applications », *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, *10*, pp. 3736–3764, 2021.

[24] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, « Overview of the multiview and 3d extensions of high efficiency video coding », *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, *1*, pp. 35–49, 2016.

[25] T. Ebrahimi, S. Foessel, F. Pereira, and P. Schelkens, « Jpeg pleno: toward an efficient representation of visual reality », *IEEE Multimedia*, vol. 23, *4*, pp. 14–20, 2016.

[26] M. Domanski, O. Stankiewicz, K. Wegner, and T. Grajek, « Immersive visual media mpeg-i: 360 video, virtual navigation and beyond », *in Proc. Int. Conf. Syst., Signals Image Process.*, 2017, pp. 1–9.

[27] N. Ahmed, T. Natarajan, and K. R. Rao, « Discrete cosine transform », *IEEE Trans. Comput.*, vol. 100, *1*, pp. 90–93, 1974.

[28] M. J. Shensa, « The discrete wavelet transform: wedding the a trous and mallat algorithms », *IEEE Trans. Signal Process.*, vol. 40, *10*, pp. 2464–2482, 1992.

[29] P. Paudyal, F. Battisti, P. L. Callet, J. Gutiérrez, and M. Carli, « Perceptual quality of light field images and impact of visualization techniques », *IEEE Trans. Broadcast.*, vol. 67, *2*, pp. 395–408, 2021.

[30] H. Huang, H. Zeng, Y. Tian, J. Chen, J. Zhu, and K.-K. Ma, « Light field image quality assessment: an overview », *in Proc. IEEE Conf. Multimedia Inf. Process. Retr.*, 2020, pp. 348–353.

[31] M. Rossi and P. Frossard, « Geometry-consistent light field super-resolution via graph-based regularization », *IEEE Trans. Image Process.*, vol. 27, *9*, pp. 4207–4218, 2018.

[32] Z. Cheng, Z. Xiong, and D. Liu, « Light field super-resolution by jointly exploiting internal and external similarities », *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, *8*, pp. 2604–2616, 2020.

[33] W. Zhou, E. Zhou, G. Liu, L. Lin, and A. Lumsdaine, « Unsupervised monocular depth estimation from light field image », *IEEE Trans. Image Process.*, vol. 29, pp. 1606–1617, 2020.

[34] A. Chuchvara, A. Barsi, and A. Gotchev, « Fast and accurate depth estimation from sparse light fields », *IEEE Trans. Image Process.*, vol. 29, pp. 2492–2506, 2020.

[35] M. L. Pendu, X. Jiang, and C. Guillemot, « Light field inpainting propagation via low rank matrix completion », *IEEE Trans. Image Process.*, vol. 27, *4*, pp. 1981–1993, 2018.

[36] Y. Wang, T. Wu, J. Yang, L. Wang, W. An, and Y. Guo, « Deoccnet: learning to see through foreground occlusions in light fields », *in Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 118–128.

[37] Z. Huang, M. Yu, G. Jiang, K. Chen, Z. Peng, and F. Chen, « Reconstruction distortion oriented light field image dataset for visual communication », *in Proc. Int. Symp. Net. Comp. Commun.*, 2019, pp. 1–5.

[38] L. Shan, P. An, C. Meng, X. Huang, C. Yang, and L. Shen, « A no-reference image quality assessment metric by multiple characteristics of light field images », *IEEE Access*, vol. 7, pp. 127 217–127 229, 2019.

[39] A. Zizien and K. Fliegel, « Lfdd: light field image dataset for performance evaluation of objective quality metrics », *in Proc. Appl. Digit. Image Process. XLII*, vol. 11510, 2020, Art. no. 115102U.

[40] ITU-R BT.500-13. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication, 2012.

[41] P. Paudyal, F. Battisti, M. Sjöström, R. Olsson, and M. Carli, « Toward the perceptual quality evaluation of compressed light field images », *IEEE Trans. Broadcast.*, vol. 63, *3*, pp. 507–522, 2017.

[42] V. K. Adhikarla, M. Vinkler, D. Sumin, *et al.*, « Towards a quality metric for dense light fields », *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3720–3729.

[43] I. Viola and T. Ebrahimi, « Valid: visual quality assessment for light field images dataset », *in Proc. Int. Conf. Quality Multimedia Exper.*, 2018, pp. 1–3.

[44] S. Zhao and Z. Chen, « Light field image coding via linear approximation prior », *in Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 4562–4566.

[45] W. Ahmad, R. Olsson, and M. Sjostrom, « Interpreting plenoptic images as multiview sequences for improved compression », *in Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 4557–4561.

[46] I. Tabus, P. Helin, and P. Astola, « Lossy compression of lenslet images from plenoptic cameras combining sparse predictive coding and jpeg2000 », *in Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 4567–4571.

[47] N. Bakir, S. A. Fezza, W. Hamidouche, K. Samrouth, and O. Déforges, « Subjective evaluation of light field image compression methods based on view synthesis », *in Proc. Eur. Signal Process. Conf.*, 2019, p. 5.

[48] M. Rerabek and T. Ebrahimi, « New light field image dataset », *in Proc. Int. Conf. Quality Multimedia Exper.*, 2016, pp. 1–2.

[49] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, « Light field reconstruction using deep convolutional network on epi », *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6319–6327.

[50] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, « Learning-based view synthesis for light field cameras », *ACM Trans. Graph.*, vol. 35, *6*, pp. 1–10, 2016.

[51] J. Kim, J. K. Lee, and K. M. Lee, « Accurate image super-resolution using very deep convolutional networks », *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[52] S. Zhang, H. Sheng, D. Yang, J. Zhang, and Z. Xiong, « Micro-lens-based matching for scene recovery in lenslet cameras », *IEEE Trans. Image Process.*, vol. 27, *3*, pp. 1060–1075, 2018.

[53] Y. Tian, H. Zeng, L. Xing, J. Chen, J. Zhu, and K.-K. Ma, « A multi-order derivative feature-based quality assessment model for light field image », *J. Vis. Commun. Image Represent.*, vol. 57, pp. 212–217, 2018.

[54] Y. Fang, K. Wei, J. Hou, W. Wen, and N. Imamoglu, « Light filed image quality assessment by local and global features of epipolar plane image », *in Proc. IEEE Int. Conf. Multimedia Big Data*, 2018, pp. 1–6.

[55] Y. Tian, H. Zeng, J. Hou, J. Chen, J. Zhu, and K.-K. Ma, « A light field image quality assessment model based on symmetry and depth features », *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, *5*, pp. 2046–2050, 2020.

[56] X. Min, J. Zhou, G. Zhai, P. L. Callet, X. Yang, and X. Guan, « A metric for light field reconstruction, compression, and display quality evaluation », *IEEE Trans. Image Process.*, vol. 29, pp. 3790–3804, 2020.

[57] C. Meng, P. An, X. Huang, C. Yang, and D. Liu, « Full reference light field image quality evaluation based on angular-spatial characteristic », *IEEE Signal Process. Lett.*, vol. 27, pp. 525–529, 2020.

[58] H. Huang, H. Zeng, J. Hou, J. Chen, J. Zhu, and K.-K. Ma, « A spatial and geometry feature-based quality assessment model for the light field images », *IEEE Trans. Image Process.*, vol. 31, pp. 3765–3779, 2022.

[59] C. Meng, P. An, X. Huang, C. Yang, L. Shen, and B. Wang, « Objective quality assessment of lenslet light field image based on focus stack », *IEEE Trans. Multimedia*, vol. 24, pp. 3193–3207, 2021.

[60] P. Paudyal, F. Battisti, and M. Carli, « Reduced reference quality assessment of light field images », *IEEE Trans. Broadcast.*, vol. 65, *1*, pp. 152–165, 2019.

[61] J. Xiang, P. Chen, Y. Dang, R. Liang, and G. Jiang, « Pseudo light field image and 4d wavelet-transform-based reduced-reference light field image quality assessment », *IEEE Trans. Multimedia*, vol. 26, pp. 929–943, 2023.

[62] L. Shi, S. Zhao, and Z. Chen, « Belif: blind quality evaluator of light field image with tensor structure variation index », *in Proc. of IEEE Int. Conf. Image Process.*, 2019, pp. 3781–3785.

[63] L. Shi, W. Zhou, Z. Chen, and J. Zhang, « No-reference light field image quality assessment based on spatial-angular measurement », *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, *11*, pp. 4114–4128, 2019.

[64] Z. Luo, W. Zhou, L. Shi, and Z. Chen, « No-reference light field image quality assessment based on micro-lens image », *in Proc. IEEE Picture Coding Symp.*, 2019, pp. 1–5.

[65] A. Ak, S. Ling, and P. L. Callet, « No-reference quality evaluation of light field content based on structural representation of the epipolar plane image », *in Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2020, pp. 1–6.

[66] W. Zhou, L. Shi, Z. Chen, and J. Zhang, « Tensor oriented no-reference light field image quality assessment », *IEEE Trans. Image Process.*, vol. 29, pp. 4070–4084, 2020.

[67] J. Xiang, M. Yu, G. Jiang, H. Xu, Y. Song, and Y.-S. Ho, « Pseudo video and refocused images-based blind light field image quality assessment », *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, *7*, pp. 2575–2590, 2021.

[68] J. Xiang, G. Jiang, M. Yu, Y. Bai, and Z. Zhu, « No-reference light field image quality assessment based on depth, structural and angular information », *Signal Process.*, vol. 184, p. 108 063, 2021.

[69] J. Xiang, G. Jiang, M. Yu, Z. Jiang, and Y.-S. Ho, « No-reference light field image quality assessment using four-dimensional sparse transform », *IEEE Trans. Multimedia*, vol. 25, pp. 457–472, 2023.

[70] Z. Pan, M. Yu, G. Jiang, H. Xu, and Y.-S. Ho, « Combining tensor slice and singular value for blind light field image quality assessment », *IEEE J. Sel. Topics Signal Process.*, vol. 15, *3*, pp. 672–687, 2021.

[71] X. Chai, F. Shao, Q. Jiang, X. Wang, L. Xu, and Y.-S. Ho, « Blind quality evaluator of light field images by group-based representations and multiple plane-oriented perceptual characteristics », *IEEE Trans. Multimedia*, vol. 26, pp. 607–622, 2023.

[72] Z. Guo, W. Gao, H. Wang, J. Wang, and S. Fan, « No-reference deep quality assessment of compressed light field images », *in Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[73] P. Zhao, X. Chen, V. Chung, and H. Li, « Delfiqe—a low-complexity deep learning-based light field image quality evaluator », *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[74] Q. Qu, X. Chen, V. Chung, and Z. Chen, « Light field image quality assessment with auxiliary learning based on depthwise and anglewise separable convolutions », *IEEE Trans. Broadcast.*, vol. 67, *4*, pp. 837–850, 2021.

[75] S. Alamgeer and M. C. Q. Farias, « Deep learning-based light field image quality assessment using frequency domain inputs », *in Proc. IEEE Int. Conf. Qual. Multimedia Exper.*, 2022, pp. 1–6.

[76] Video Quality Experts Group (VQEG). [Online]. Available: http://www.its.bldrdoc.gov/vqeg/vqeg-home.

[77] T. Ojala, M. Pietikainen, and T. Maenpaa, « Multiresolution gray-scale and rotation invariant texture classification with local binary patterns », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, *7*, pp. 971–987, 2002.

[78] W. Zhou and L. Yu, « Binocular responses for no-reference 3d image quality assessment », *IEEE Trans. Multimedia*, vol. 18, *6*, pp. 1077–1084, 2016.

[79] C.-C. Chang and C.-J. Lin, « Libsvm: a library for support vector machines », *ACM Trans. Intell. Symp. Technol.*, vol. 2, *3*, pp. 1–27, 2011.

[80] N. Kruger, P. Janssen, S. Kalkan, *et al.*, « Deep hierarchies in the primate visual cortex: what can we learn for computer vision? », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, *8*, pp. 1847–1871, 2013.

[81] S. Kastner, P. D. Weerd, and L. G. Ungerleider, « Texture segregation in the human visual cortex: a functional mri study », *J. Neurophysiol.*, vol. 83, pp. 2453–2457, 2000.

[82] Y. Fang, J. Yan, L. Li, J. Wu, and W. Lin, « No reference quality assessment for screen content images with both local and global feature representation », *IEEE Trans. Image Process.*, vol. 27, *4*, pp. 1600–1610, 2018.

[83] D. D. Sandic-Stankovic, D. D. Kukolj, and P. L. Callet, « Quality assessment of dibr-synthesized views based on sparsity of difference of closings and difference of gaussians », *IEEE Trans. Image Process.*, vol. 31, pp. 1161–1175, 2022.

[84] F. Heitger, L. Rosenthaler, R. V. D. Heydt, E. Peterhans, and O. Kübler, « Simulation of neural contour mechanisms: from simple to end-stopped cells », *Vis. Res.*, vol. 32, *5*, pp. 963–981, 1992.

[85] S. A. Engel, G. H. Glover, and B. A. Wandell, « Retinotopic organization in human visual cortex and the spatial precision of functional mri », *Cerebral cortex*, vol. 7, *2*, pp. 181–192, 1997.

[86] D. Mustafi, A. H. Engel, and K. Palczewski, « Structure of cone photoreceptors », *Progress in retinal and eye research*, vol. 28, *4*, pp. 289–302, 2009.

[87] S. G. Mallat, « A theory for multiresolution signal decomposition: the wavelet representation », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, *7*, pp. 674–693, 1989.

[88] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, « A haar wavelet-based perceptual similarity index for image quality assessment », *Signal Process., Image Commun.*, vol. 61, pp. 33–43, 2018.

[89] J. Canny, « A computational approach to edge detection », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, *6*, pp. 679–698, 1986.

[90] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, « Image quality assessment based on a degradation model », *IEEE Trans. Image Process.*, vol. 9, *4*, pp. 636–650, 2000.

[91] H. R. Sheikh, A. C. Bovik, and G. de Veciana, « An information fidelity criterion for image quality assessment using natural scene statistics », *IEEE Trans. Image Process.*, vol. 14, *12*, pp. 2117–2128, 2005.

[92] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, « Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions », *ACM Trans. Graph.*, vol. 30, *4*, pp. 1–14, 2011.

[93] Q. Li, W. Lin, and Y. Fang, « No-reference quality assessment for multiply-distorted images in gradient domain », *IEEE Signal Process. Lett.*, vol. 23, *4*, pp. 541–545, 2016.

[94] A. Mittal, R. Soundararajan, and A. C. Bovik, « Making a 'completely blind' image quality analyzer », *IEEE Signal Process. Lett.*, vol. 20, *3*, pp. 209–212, 2013.

[95] A. Mittal, A. K. Moorthy, and A. C. Bovik, « No-reference image quality assessment in the spatial domain », *IEEE Trans. Image Process.*, vol. 21, *12*, pp. 4695–4708, 2012.

[96] S. Tian, L. Zhang, L. Morin, and O. Déforges, « Niqsv+: a no-reference synthesized view quality assessment metric », *IEEE Trans. Image Process.*, vol. 27, *4*, pp. 1652–1664, 2018.

[97] K. He, X. Zhang, S. Ren, and J. Sun, « Deep residual learning for image recognition », *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[98] D. Li, T. Jiang, and M. Jiang, « Quality assessment of in-the-wild videos », *in Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2351–2359.

[99] K. Cho, B. V. Merrienboer, C. G. ang F. Bougares, H. Schwenk, and Y. Bengio, « Learning phrase representations using rnn encoder-decoder for statistical machine translation », *in Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2014, pp. 1724–1734.

[100] D. M. Chandler and S. S. Hemami, « Vsnr: a wavelet-based visual signal-to-noise ratio for natural images », *IEEE Trans. Image Process.*, vol. 16, *9*, pp. 2284–2298, 2007.

[101] L. Zhang, Y. Shen, and H. Li, « Vsi: a visual saliency-induced index for perceptual image quality assessment », *IEEE Trans. Image Process.*, vol. 23, *10*, pp. 4270–4281, 2014.

[102] L. Zhang, Z. Gu, and H. Li, « Sdsp: a novel saliency detection method by combining simple priors », *in Proc. IEEE Int. Conf. Image Process.*, 2013, pp. 171–175.

[103] L.-M. Po, M. Liu, W. Y. F. Yuen, *et al.*, « A novel patch variance biased convolutional neural network for no-reference image quality assessment », *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, *4*, pp. 1223–1229, 2019.

[104] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, « Attention is all you need », *in Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[105] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, « End-to-end object detection with transformers », *in Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[106] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, « Image quality assessment: from error visibility to structural similarity », *IEEE Trans. Image Process.*, vol. 13, *4*, pp. 600–612, 2004.

[107] INRIA Lytro image dataset. [Online]. Available: https://www.irisa.fr/temics/demos/lightField/LowRank2/datasets/datasets.html.

[108] J. Shi, X. Jiang, and C. Guillemot, « A framework for learning depth from a flexible subset of dense and sparse light field views », *IEEE Trans. Image Process.*, vol. 28, *12*, pp. 5867–5880, 2019.

[109] Subjective Video Quality Assessment Methods for Multimedia Applications, document ITU-T Rec. P.910, International Telecommunication Union, 2023.

[110] D. Hasler and S. E. Suesstrunk, « Measuring colorfulness in natural images », *Proc. SPIE*, vol. 5007, pp. 87–96, 2003.

[111] N. Meng, H. K.-H. So, X. Sun, and E. Lam, « High-dimensional dense residual convolutional neural network for light field reconstruction », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, *3*, pp. 873–886, 2021.

[112] J. Jin, J. Hou, H. Yuan, and S. Kwong, « Learning light field angular super-resolution via a geometry-aware network », *in Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11 141–11 148.

[113] G. Liu, H. Yue, J.Wu, and J. Yang, « Intra-inter view interaction network for light field image super-resolution », *IEEE Trans. Multimedia*, vol. 25, pp. 256–266, 2021.

[114] Y. Wang, L. Wang, G. Wu, *et al.*, « Disentangling light fields for super-resolution and disparity estimation », *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, *1*, pp. 425–443, 2023.

[115] M. Pérez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk, « From pairwise comparisons and rating to a unified quality scale », *IEEE Trans. Image Process.*, vol. 29, *1*, pp. 1139–1151, 2019.

[116] L. L. Thurstone, « A law of comparative judgement », *Psychol. Rev.*, vol. 34, *4*, pp. 273–286, 1927.

[117] L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, « Blind image quality assessment by relative gradient statistics and adaboosting neural network », *Signal Process., Image Commun.*, vol. 40, pp. 1–15, 2016.

[118] D. L. Ruderman, « The statistics of natural images », *Netw.: Comput. Neural Syst.*, vol. 5, *4*, pp. 517–548, 1994.

[119] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, « Full-reference quality assessment of stereopairs accounting for rivalry », *Signal Process., Image Commun.*, vol. 28, *9*, pp. 1143–1155, 2013.

[120] L. Liu, B. Liu, C.-C. Su, H. Huang, and A. C. Bovik, « Binocular spatial activity and reverse saliency driven no-reference stereopair quality assessment », *Signal Process., Image Commun.*, vol. 58, pp. 287–299, 2017.

COLLEGE DOCTORAL BRETAGNE
MATHS, TELECOMS
INFORMATIQUE, SIGNAL
SYSTEMES, ELECTRONIQUE

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

Titre : Évaluation de la qualité des images de champs de lumière

Mot clés : Évaluation de la qualité, Images de champs de lumière, Métrique objective, Expérience subjective, Système visuel humain

Résumé : Les images de champs de lumière (LFI) suscitent un intérêt et une fascination remarquables en raison de leur importance croissante dans les applications immersives. Étant donné que les LFI peuvent être déformés à différentes étapes, de l'acquisition à la visualisation, l'évaluation de la qualité des images de champs de lumière (LFIQA) est d'une importance vitale pour surveiller les dégradations potentielles de la qualité des LFI.

La première contribution (Chapitre 3) de ce travail se concentre sur le développement de deux métriques LFIQA sans référence (NR) fondées sur des caractéristiques ad-hoc, dans lesquelles les informations de texture et les coefficients ondelettes sont exploités pour l'évaluation de la qualité.

Puis dans la deuxième partie (Chapitre 4), nous explorons le potentiel de la technologie de l'apprentissage profond (deep learning) pour l'évaluation de la qualité des LFI, et nous proposons quatre métriques LFIQA basées sur l'apprentissage profond, dont trois métriques sans référence (NR) et une métrique Full-Reference (FR).

Dans la dernière partie (Chapitre 5), nous menons des évaluations subjectives et proposons une nouvelle base de données normalisée pour la LFIQA. De plus, nous fournissons une étude comparative (benchmark) de nombreuses métriques objectives LFIQA de l'état de l'art, sur la base de données proposée.

Title: Quality Assessment of Light Field Images

Keywords: Quality assessment, Light field images, Objective metric, Subjective experiment, Human visual system

Abstract: Light Field Image (LFI) has garnered remarkable interest and fascination due to its burgeoning significance in immersive applications. Since LFIs may be distorted at various stages from acquisition to visualization, Light Field Image Quality Assessment (LFIQA) is vitally important to monitor the potential impairments of LFI quality.

The first contribution (Chapter 3) of this work focuses on developing two handcrafted feature-based No-Reference (NR) LFIQA metrics, in which texture information and wavelet information are exploited for quality evaluation.

Then in the second part (Chapter 4), we explore the potential of combining deep learning technology with the quality assessment of LFIs, and propose four deep learning-based LFIQA metrics according to different LFI characteristics, including three NR metrics and one Full-Reference (FR) metric.

In the last part (Chapter 5), we conduct subjective experiments and propose a novel standard LFIQA database. Moreover, a benchmark of numerous state-of-the-art objective LFIQA metrics on the proposed database is provided.