國立臺灣大學電機資訊學院資訊工程學研究所

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

兩階段漸進式多重曝光融合：基於中間曝光生成

Two-Phase Progressive Multiple Exposure Fusion via Intermediate Exposure Generation

黃郁夫

Yu-Fu Huang

指導教授: 莊永裕 博士

Advisor: Yung-Yu Chuang Ph.D.

中華民國 114 年 10 月

October, 2025

doi:10.6342/NTU202504549

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

兩階段漸進式多重曝光融合：基於中間曝光生成

Two-Phase Progressive Multiple Exposure Fusion via Intermediate Exposure Generation

本論文係黃郁夫君（學號 R12922152）在國立臺灣大學資訊工程學系完成之碩士學位論文，於民國 114 年 10 月 02 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Computer Science and Information Engineering on 2 October 2025 have examined a Master's thesis entitled above presented by HUANG, YU-FU (student ID: R12922152) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

莊永裕	吳賦哲	葉正聖
（指導教授 Advisor）		

系主任/所長 Director: 陳祝嵩

doi:10.6342/NTU202504549

Acknowledgements

As someone who transitioned into computer science from a different field, this academic journey has been filled with both challenges and rewards. First and foremost, I would like to express my deepest gratitude to my advisor for giving a non-CS background student this opportunity. I am fully aware that mentoring a student with relatively weak foundations requires significant trust and risk-taking. For this, I am both deeply grateful and apologetic.

I would like to thank the senior students and colleagues in our laboratory. Whether facing academic challenges or everyday problems, they were always willing to provide assistance. Their help during my early days, when I was still unfamiliar with everything, saved me considerable time and unnecessary detours.

I am grateful to the department's administrative team for efficiently handling administrative matters. Finally, I would like to thank National Taiwan University's Department of Computer Science and Information Engineering for its resources and cultivation.

doi:10.6342/NTU202504549

doi:10.6342/NTU202504549

摘要

多重曝光影像融合（MEF）面臨一項根本性的挑戰：現有使用 UNet 或注意力機制的方法必須直接處理亮度差距巨大的極端曝光影像對，這限制了融合品質，特別是在局部區域。本文提出一項新的觀察：當比較的曝光影像之間的相對差異較小時，注意力機制能計算出更可靠的權重。

我們的雙階段框架採用「由全域到局部」的漸進式策略。第一階段（UNet）在網路瓶頸處進行全域特徵融合，將極端曝光的深層特徵合併，以生成亮度均衡的中間曝光影像。此中間影像具有雙重功能：(1) 減少曝光差距以改善注意力權重計算，(2) 作為距離兩個極端曝光等距的參考錨點。第二階段（AMNet）利用該中間影像進行基於注意力的局部細化，在較小的曝光差異下，能更精確地進行特徵比對與選擇性細節恢復。

此架構的分工源於對 MEF 誤差的分析：大多數偽影來自局部曝光失敗，而非全域亮度不平衡。因此，UNet 負責全域協調，而注意力模組則專注於局部增強。我們的解耦式訓練確保每個階段都能針對其特定目標進行最佳化。實驗驗證了我們的理論：中間曝光顯著提升了注意力權重的準確性，在維持全域一致性的同時，於困難區域中達成更優秀的細節保留。

關鍵字：電腦視覺、多重曝光合成、高動態範圍影像

doi:10.6342/NTU202504549

doi:10.6342/NTU202504549

Abstract

Multi-exposure image fusion (MEF) faces a fundamental challenge: existing methods using UNet or attention mechanisms must directly process extreme exposure pairs with substantial luminance gaps, limiting fusion quality particularly in local regions. We present a novel insight that attention mechanisms compute more reliable weights when comparing exposures with reduced relative differences. Our two-phase framework implements a global-to-local progressive strategy. Phase one (UNet) performs global feature fusion at the network bottleneck, merging deep representations from extreme exposures to generate an intermediate exposure with balanced luminance. This intermediate serves dual purposes: (1) reducing exposure gaps for improved attention computation, and (2) providing a reference anchor equidistant from both extremes. Phase two (AMNet) exploits this intermediate for attention-based local refinement, where smaller exposure differences enable precise feature comparison and selective detail recovery. The architecture division follows MEF error analysis: most artifacts appear as local exposure failures rather than

doi:10.6342/NTU202504549

global imbalance. Thus, UNet handles global harmonization while attention targets local enhancement. Our decoupled training ensures each phase optimizes its specific objective. Experiments validate our theory: the intermediate exposure significantly improves attention weight accuracy, leading to superior detail preservation in challenging regions while maintaining global consistency.

Keywords: Computer Vision, Multiple Exposure Fusion, HDR

doi:10.6342/NTU202504549

Contents

doi:10.6342/NTU202504549

doi:10.6342/NTU202504549

doi:10.6342/NTU202504549

doi:10.6342/NTU202504549

List of Figures

doi:10.6342/NTU202504549

doi:10.6342/NTU202504549

List of Tables

doi:10.6342/NTU202504549

doi:10.6342/NTU202504549

Denotation

HDR	High Dynamic Range
LDR	Low Dynamic Range
MEF	Multi-Exposure Fusion
TMO	Tone Mapping Operator
CNN	Convolutional Neural Network
UNet	U-Net Architecture
AMNet	Attention Mechanism Network
HDRDB	Hybrid Dilated Residual Dense Block
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index Measure
MS-SSIM	Multi-Scale Structural Similarity Index Measure
VIF	Visual Information Fidelity

doi:10.6342/NTU202504549

SICE	Single Image Contrast Enhancement Dataset
GPU	Graphics Processing Unit
GT	Ground Truth
C	Channel Dimension
H	Image Height
W	Image Width
N	Total Number of Pixels

doi:10.6342/NTU202504549

Chapter 1 Introduction

1.1 HDR and LDR Fundamentals

Natural scenes exhibit an enormous range of luminance levels, from starlight at 10^{-3} cd/m^2 to direct sunlight exceeding 10^5 cd/m^2. However, standard digital cameras can only capture a limited dynamic range, typically 8-14 bits, resulting in Low Dynamic Range (LDR) images. This fundamental limitation causes inevitable loss of details in regions with extreme brightness or darkness—bright areas become overexposed white regions while dark areas collapse into underexposed black regions. High Dynamic Range (HDR) imaging aims to overcome this physical constraint by capturing and representing the full luminance range of real-world scenes, preserving visual details across all illumination levels that closely match human visual perception.

1.2 HDR vs. MEF

Traditional HDR imaging involves capturing multiple LDR images with different exposures and combining them to create an HDR radiance map. However, for most practical applications, the ultimate goal is not the HDR image itself, but rather a high-quality photograph that can be displayed on standard LDR monitors. HDR images cannot be di-

doi:10.6342/NTU202504549

rectly displayed and require tone mapping operators (TMOs) to compress the dynamic range back to displayable range. This additional tone mapping step introduces several challenges: it often produces artifacts such as halos and gradient reversals, and the results highly depend on parameter tuning which significantly affects the final visual quality. Furthermore, with the prevalence of handheld photography in consumer and mobile devices, the HDR workflow becomes increasingly impractical as capturing stable image sequences and processing them through tone mapping adds complexity to the imaging pipeline.

Multi-exposure image fusion (MEF) has emerged as a more direct approach that fuses multiple LDR images into a single display-ready image without the intermediate HDR reconstruction or tone mapping steps. Since our objective is to obtain a visually pleasing image displayable on standard LDR monitors rather than a physically accurate HDR radiance map, MEF directly addresses this goal by producing LDR output that is immediately viewable. This streamlined approach avoids the complexities and potential artifacts introduced by tone mapping, making it a practical solution for dynamic range enhancement in real-world applications.

1.3 Evolution of MEF Methods

Early MEF methods relied on hand-crafted features and fusion rules, such as the seminal work by Mertens et al. [10] using quality measures for weight map computation. While these traditional approaches are interpretable and do not require training data, they struggle with complex scenes and extreme exposure differences.

The advent of deep learning has revolutionized MEF by enabling end-to-end learning of optimal fusion strategies. DeepFuse [11] pioneered CNN-based MEF using unsuper-

doi:10.6342/NTU202504549

vised learning. Subsequent works have explored various architectures, with two dominant paradigms emerging: encoder-decoder structures with skip connections for multi-scale feature fusion, and attention-based methods for selective feature combination.

Motion between captures remains a fundamental challenge in multi-exposure techniques, as misalignment causes ghosting artifacts that increase with the number of input images. Recent research by HoLoCo [7] demonstrated that when limited to two inputs, extreme exposure pairs achieve optimal fusion by maximizing complementary information. However, these extreme pairs present substantial luminance gaps that compromise attention mechanism performance, as features from vastly different exposures have minimal correspondence for meaningful comparison.

1.4 Contributions

MEF artifacts primarily appear as local exposure failures rather than global imbalance, suggesting the need for separate treatments: global harmonization for overall balance and local refinement for detail recovery. Our key insight is that *attention mechanisms compute more reliable weights when comparing exposures with reduced relative differences*.

Based on this insight, we propose a two-phase progressive MEF framework:

- **Phase One (UNet):** Performs global feature fusion at the network bottleneck to generate a balanced intermediate exposure.
- **Phase Two (AMNet):** Uses the intermediate exposure for attention-based local refinement with improved weight estimation.

doi:10.6342/NTU202504549

Our main contributions include: (1) demonstrating that reduced exposure differences improve attention mechanism performance in MEF, (2) proposing an architecture that explicitly separates global and local fusion, (3) introducing intermediate exposure generation to bridge extreme inputs, and (4) developing a decoupled training strategy for robust performance.

doi:10.6342/NTU202504549

Chapter 2 Related Work

2.1 Traditional Methods

Early MEF methods relied on hand-crafted features and fusion rules to combine multiple exposures. Mertens et al. [10] proposed a seminal approach using Laplacian pyramids with quality measures including contrast, saturation, and well-exposedness to compute multi-scale weight maps. Li and Kang [6] introduced guided filtering to preserve edges while suppressing artifacts through a two-scale decomposition framework that has been widely adopted. Ma et al. [9] developed a structural patch decomposition approach that separates images into average intensity, signal intensity, and structure components, achieving robustness to ghosting while maintaining vivid colors. While these traditional approaches are interpretable and do not require training data, they struggle with complex scenes and extreme exposure differences.

2.1.1 Deep Learning Methods

The advent of deep learning has revolutionized MEF by enabling end-to-end learning of optimal fusion strategies. DeepFuse [11] pioneered CNN-based MEF using unsupervised learning with perceptual losses, establishing the foundation for learning-based

doi:10.6342/NTU202504549

approaches.

Subsequent works have explored various architectural designs and learning paradigms. DPE-MEF [3] employs two cascaded UNets operating in the YCbCr color space, where the first UNet generates the luminance channel (Y) and the second produces the chrominance channels (CbCr), before converting back to RGB. This separation allows specialized processing of brightness and color information. HoLoCo [7] introduces holistic and local contrastive learning through a two-module design, where attention maps perform preliminary fusion followed by Retinex-based color correction. These methods have shown significant improvements over traditional approaches in handling complex scenes.

2.1.2 Automated Architecture and Loss Design

Recent years have witnessed growing interest in automated architecture search and optimization strategies. HSDS-MEF [16] introduces a dual-search approach that simultaneously optimizes both loss parameters and network architecture—using contrastive learning concepts to automatically optimize loss weights while leveraging validation loss to search for optimal architectures. Similarly, CRMEF [8] searches for compact architectures that maintain robustness across diverse scenes. This trend toward automation aims to reduce human bias in design choices and discover novel architectures that might not be intuitive to human designers.

2.1.3 Our Approach

In contrast to automated approaches, our work explores a different direction by incorporating domain knowledge into architectural design. We propose a two-phase frame-

doi:10.6342/NTU202504549

work with decoupled training, where Phase One generates intermediate exposures and Phase Two performs attention-based refinement. This design is motivated by the observation that extreme exposure differences pose challenges for direct feature comparison. The decoupled training strategy allows each phase to specialize in its designated task—global harmonization versus local refinement—potentially offering a complementary perspective to both end-to-end and automated search methods.

doi:10.6342/NTU202504549

doi:10.6342/NTU202504549

Chapter 3 Methodology

3.1 Overview

Our proposed two-phase progressive MEF framework addresses the fundamental challenge of extreme exposure differences in attention-based fusion. The key insight is that attention mechanisms require reduced exposure gaps for reliable weight computation. Figure 3.1 illustrates our complete architecture, consisting of Phase One (UNet) for intermediate exposure generation and Phase Two (AMNet) for attention-based refinement.

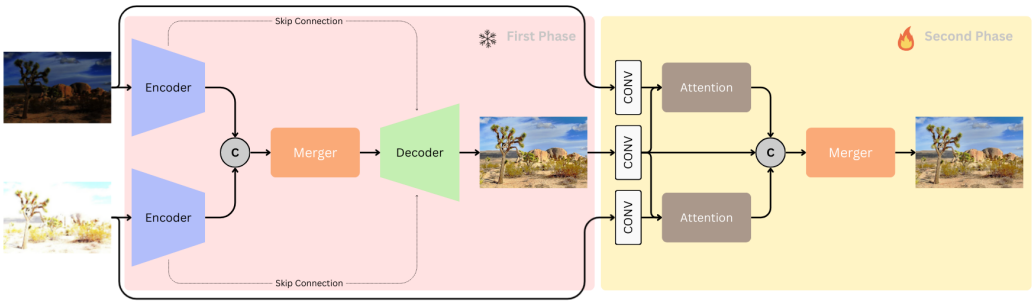


Figure 3.1: Overview of our two-phase progressive MEF framework. Phase One generates an intermediate exposure from extreme inputs, which Phase Two uses for attention-based refinement.

doi:10.6342/NTU202504549

3.2 Phase One: UNet for Intermediate Exposure Generation

3.2.1 Architecture Design

Phase One employs a dual-encoder UNet architecture inspired by DDMEF [13], which demonstrated the effectiveness of separate encoders for extreme exposure processing. However, unlike DDMEF which performs fusion throughout all skip connections, we specifically design our architecture to concentrate fusion solely at the bottleneck layer. Furthermore, our bottleneck fusion module adopts a more sophisticated design inspired by AHDRNet [17], utilizing our proposed HDRDB for comprehensive multi-scale context aggregation.

As shown in Figure 3.2, the network processes under-exposed and over-exposed images through separate encoders, allowing specialized feature extraction for each exposure level.

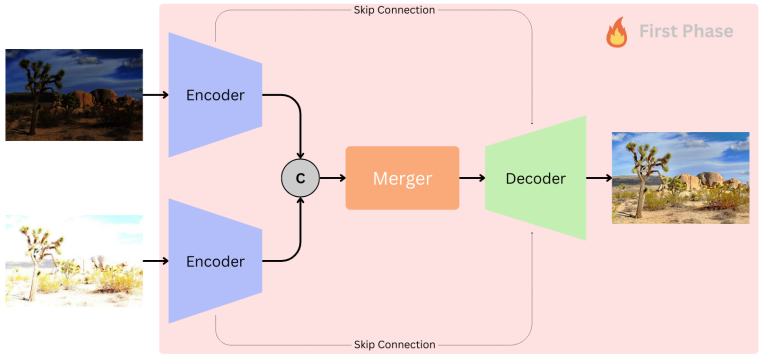


Figure 3.2: Phase One UNet architecture with dual encoders for processing extreme exposures.

doi:10.6342/NTU202504549

3.2.2 Network Components

Figure 3.3 details the architectural components used in our framework.

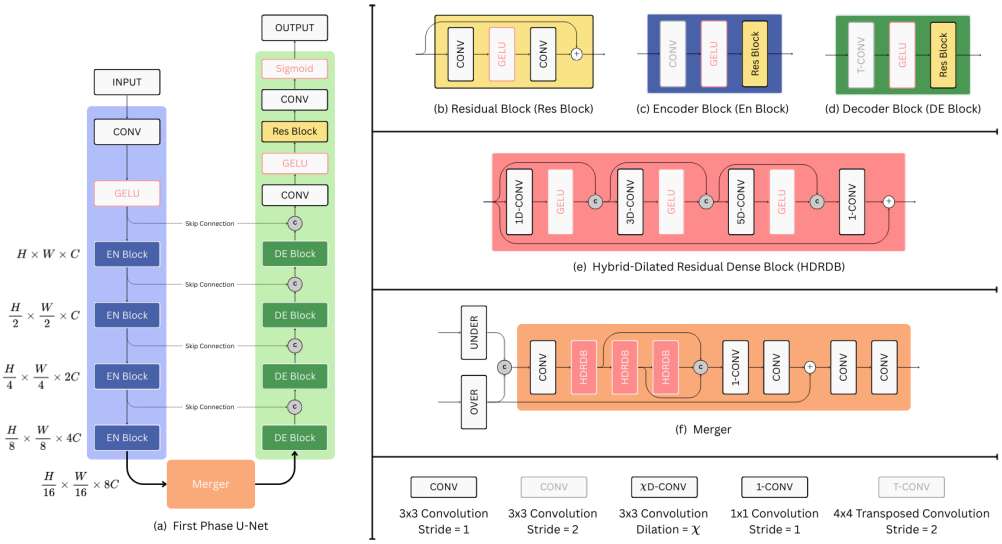


Figure 3.3: Detailed architecture of building blocks: (a) First Phase U-Net structure, (b) Residual Block, (c) Encoder Block, (d) Decoder Block, (e) Hybrid Dilated Residual Dense Block (HDRDB), and (f) Merger module.

Dual Encoders: Each encoder independently processes its input through an initial projection followed by four encoder blocks:

$$F^0_{under} = \text{GELU}(\text{Conv}_{3\times3}(I_{under})), \quad F^0_{over} = \text{GELU}(\text{Conv}_{3\times3}(I_{over})) \tag{3.1}$$

$$F^i_{under} = \text{EnBlock}^i_{under}(F^{i-1}_{under}), \quad F^i_{over} = \text{EnBlock}^i_{over}(F^{i-1}_{over}), \quad i \in \{1,2,3,4\} \tag{3.2}$$

Each encoder block progressively downsamples spatial resolution while increasing channel dimensions from C to $8C$.

Bottleneck Fusion: At the deepest level (spatial resolution $H/16 \times W/16$), the

doi:10.6342/NTU202504549

Merger module combines features from both encoders:

$$F_{merged} = \text{Merger}_{UNet}(F^4_{under}, F^4_{over}) \tag{3.3}$$

The Merger module consists of our proposed Hybrid Dilated Residual Dense Block (HDRDB), which processes the concatenated features. HDRDB is an enhanced version of AHDR-Net's DRDB that employs varying dilation rates of $[1, 2, 3, 2, 1]$ to avoid the gridding artifacts [2] associated with uniform dilation rates, ensuring continuous receptive field coverage.

Shared Decoder: The decoder reconstructs the intermediate exposure through four decoder blocks with skip connections:

$$F^i_{dec} = \text{DeBlock}^i(F^{i+1}_{dec}, F^i_{under}, F^i_{over}), \quad i \in \{3, 2, 1, 0\} \tag{3.4}$$

where $F^4_{dec} = F_{merged}$. Skip connections from both encoders are concatenated with the upsampled features at each level. The final output is generated through:

$$I_{inter} = \text{Sigmoid}(\text{Conv}_{3\times3}(\text{ResBlock}(\text{GELU}(\text{Conv}_{3\times3}(F^0_{dec}))))) \tag{3.5}$$

3.2.3 Training Objective

Phase One is trained independently to generate optimal intermediate exposures. Given under-exposed I_u and over-exposed I_o images with ground truth I_{gt}, we minimize:

$$\mathcal{L} = \mathcal{L}_{\text{L1}}(I_{inter}, I_{gt}) + \lambda_{vgg}\mathcal{L}_{\text{VGG}}(I_{inter}, I_{gt}) + \lambda_{ssim}\mathcal{L}_{\text{SSIM}}(I_{inter}, I_{gt}) \tag{3.6}$$

where the intermediate exposure is generated by $I_{inter} = \text{UNet}(I_u, I_o)$.

doi:10.6342/NTU202504549

3.3 Phase Two: AMNet for Attention-based Refinement

3.3.1 Architecture Design

Phase Two adopts the attention mechanism from AHDRNet [17], which demonstrated effective ghost-free HDR reconstruction through attention-guided feature fusion. However, AHDRNet was designed for HDR imaging with three inputs (under, medium, and over exposures), where the medium exposure serves as a reference for computing attention weights. In contrast, MEF research typically works with only two extreme exposures as input, lacking the crucial medium reference.

Our approach addresses this limitation by using the generated intermediate exposure from Phase One to serve the role of the missing medium exposure. This adaptation enables reliable attention-based fusion in the MEF setting where only extreme exposures are available. The generated intermediate provides a balanced reference point, reducing the exposure gap for more accurate feature comparison and weight computation.

As illustrated in Figure 3.1, AMNet takes three inputs: the original under-exposure, the generated intermediate exposure, and the original over-exposure. The critical innovation is that our intermediate exposure is learned to be optimal for attention computation rather than being a captured medium exposure, allowing the network to generate reference exposures specifically tailored for the fusion task.

The attention mechanism computes weights based on feature similarity:

$$\alpha_i = \text{Attention}(F_{inter}, F_i), \quad i \in \{under, over\} \tag{3.7}$$

doi:10.6342/NTU202504549

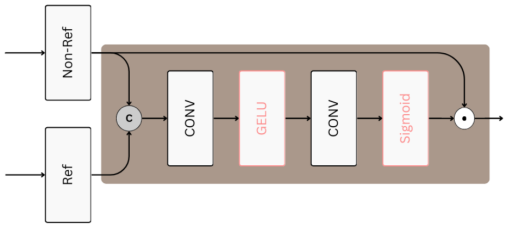


Figure 3.4: Attention mechanism in Phase Two, computing weights between intermediate and extreme exposures.

where F_{inter}, F_{under}, and F_{over} are encoded features from the intermediate, under, and over-exposed images respectively. The Attention function internally performs feature concatenation, convolution, and softmax normalization to generate the attention weights.

3.3.2 Attention-guided Fusion

The key innovation is using the intermediate exposure as a reference for attention computation. Since the exposure gap between intermediate-under and intermediate-over is smaller than under-over directly, the attention weights are more reliable:

$$I_{final} = \text{Merger}_{AMNet}(\text{Concat}[\alpha_u \odot F_{under}, F_{inter}, \alpha_o \odot F_{over}]) \tag{3.8}$$

where $\odot$ denotes element-wise multiplication. Note that the intermediate features F_{inter} are concatenated without weighting, as they already provide a balanced reference, while the extreme exposures are modulated by their respective attention weights to selectively incorporate their complementary information.

doi:10.6342/NTU202504549

3.3.3 Training Objective

Crucially, Phase Two is trained on realistic intermediate exposures generated by the frozen Phase One model:

$$I_{inter} = \text{UNet}_{frozen}(I_u, I_o) \tag{3.9}$$

The training objective for Phase Two uses the same loss formulation:

$$\mathcal{L} = \mathcal{L}_{\text{L1}}(I_{final}, I_{gt}) + \lambda_{vgg}\mathcal{L}_{\text{VGG}}(I_{final}, I_{gt}) + \lambda_{ssim}\mathcal{L}_{\text{SSIM}}(I_{final}, I_{gt}) \tag{3.10}$$

where the final output is $I_{final} = \text{AMNet}(I_u, I_{inter}, I_o)$.

3.4 Loss Functions

Our training employs a combination of reconstruction losses to ensure both pixel-level accuracy and perceptual quality.

3.4.1 L1 Loss

The L1 loss enforces pixel-wise reconstruction accuracy:

$$\mathcal{L}_{\text{L1}} = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i| \tag{3.11}$$

where Y is the ground truth image, $\hat{Y}$ is the predicted output, and N is the total number of pixels. L1 loss is preferred over L2 loss as it is less sensitive to outliers and produces sharper results.

doi:10.6342/NTU202504549

3.4.2 VGG Perceptual Loss

The VGG perceptual loss [5] captures high-level semantic and textural similarities by comparing features extracted from pre-trained VGG-16 network [12]:

$$\mathcal{L}_{\text{VGG}} = \sum_j \frac{1}{C_j H_j W_j} \|\phi_j(Y) - \phi_j(\hat{Y})\|_2^2 \tag{3.12}$$

where ϕ_j represents the feature maps at layer j of the VGG-16 network, and C_j, H_j, W_j are the number of channels, height, and width of the feature maps respectively. We extract features from layers `conv1_2`, `conv2_2`, `conv3_4`, `conv4_4`, and `conv5_4` to capture multi-scale perceptual information.

3.4.3 SSIM Loss

The Structural Similarity Index Measure (SSIM) loss [14] preserves structural information:

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(Y, \hat{Y}) \tag{3.13}$$

where SSIM is computed as:

$$\text{SSIM}(Y, \hat{Y}) = \frac{(2\mu_Y \mu_{\hat{Y}} + C_1)(2\sigma_{Y\hat{Y}} + C_2)}{(\mu_Y^2 + \mu_{\hat{Y}}^2 + C_1)(\sigma_Y^2 + \sigma_{\hat{Y}}^2 + C_2)} \tag{3.14}$$

Here, μ_Y and $\mu_{\hat{Y}}$ are the mean values, σ_Y^2 and $\sigma_{\hat{Y}}^2$ are the variances, $\sigma_{Y\hat{Y}}$ is the covariance, and C_1, C_2 are constants for numerical stability. SSIM is computed locally using an 11×11 Gaussian window.

doi:10.6342/NTU202504549

3.4.4 Loss Weighting

Both Phase One and Phase Two use the same loss weights: $\lambda_{vgg} = 0.01$ and $\lambda_{ssim} = 0.1$. These weights emphasize structural preservation while maintaining a balance with perceptual quality. The weights were determined through empirical validation on the validation set.

doi:10.6342/NTU202504549

doi:10.6342/NTU202504549

Chapter 4 Experiments

4.1 Implementation Details

4.1.1 Network Configuration

For Phase One, we set the base channel number $C = 32$. The encoder progressively increases channels: $32 \rightarrow 64 \rightarrow 128 \rightarrow 256$. The HDRDB in the merger uses dilation rates of $[1, 2, 5]$ to capture multi-scale context without gridding artifacts.

4.1.2 Training Details

Our framework is implemented in PyTorch and trained on NVIDIA RTX 3090 GPUs. Images are randomly cropped to 256×256 during training with horizontal flipping for data augmentation. We use the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

Our decoupled training procedure consists of:

1. **Phase One Training:** Train UNet for 300 epochs with initial learning rate 10^{-4}, kept constant for the first 200 epochs then linearly decayed to 5×10^{-5} over the last 100 epochs.

2. **Intermediate Generation:** Generate intermediate exposures for the entire training

doi:10.6342/NTU202504549

set using the trained UNet.

3. **Phase Two Training:** Train AMNet for 300 epochs with the same learning rate schedule.

This decoupled approach ensures each phase optimizes for its specific objective without interference, leading to better specialization and overall performance.

4.2 Experimental Setup

4.2.1 Dataset

We train and evaluate our method on the SICE (Single Image Contrast Enhancement) dataset [1], which contains multi-exposure sequences captured from real-world scenes. For ground truth generation, SICE processed each sequence through 13 different fusion algorithms, then asked 13 amateur photographers and 5 volunteers without much photographing experience to select the best result as the reference image. This provides subjective quality targets that represent general user preferences among various fusion approaches.

Currently, MEF research lacks standardized training and testing splits, leading to inconsistent evaluation across different methods. However, examining recent literature reveals that most studies utilizing explicit training sets employ at least 360 image sequences for training. The original SICE dataset contains 589 sequences, but many suffer from severe misalignment due to handheld capture, while others contain nearly black or white exposures that provide minimal information. Following common practice in static MEF research, we exclude these problematic sequences.

doi:10.6342/NTU202504549

After filtering out problematic cases, we randomly select 360 high-quality aligned sequences for training and 93 sequences for testing from the remaining clean data, ensuring our evaluation focuses on the fusion task rather than alignment challenges. Each selected sequence contains extremely under-exposed and over-exposed image pairs with sufficient content variation, making them ideal for evaluating MEF methods under challenging yet realistic conditions. This split size aligns with the training data scale used in comparable recent works while maintaining a substantial test set for robust evaluation.

4.2.2 Evaluation Metrics

We employ four widely-used metrics to comprehensively evaluate fusion quality:

- **PSNR (Peak Signal-to-Noise Ratio):** Measures pixel-level reconstruction accuracy. Higher values indicate better quality.
- **SSIM (Structural Similarity Index)** [14]: Evaluates structural preservation considering luminance, contrast, and structure. Values range from 0 to 1.
- **MS-SSIM (Multi-Scale SSIM)** [15]: Extends SSIM across multiple scales for more robust perceptual evaluation.
- **VIF (Visual Information Fidelity)** [4]: Assesses information preservation from source images. Higher values indicate better information retention.

4.3 Comparison with State-of-the-Art Methods

We compare our method against five recent MEF approaches introduced in the related work section: SPD-MEF (structural patch decomposition), DPE-MEF (dual UNet

doi:10.6342/NTU202504549

architecture), HoLoCo (holistic and local contrastive learning), HSDS-MEF (dual-search optimization), and CRMEF (compact architecture search).

4.3.1 Quantitative Results

Table 4.1 presents quantitative comparisons on the SICE test set. Our method achieves the best performance across all metrics, with particularly significant improvements in PSNR (22.9934) and VIF (0.7761).

Table 4.1: Quantitative comparison on SICE dataset. Best results in red, second best in blue.

Method	PSNR↑	SSIM↑	MS-SSIM↑	VIF↑
SPD-MEF	17.5891	0.7643	0.8599	0.6294
DPE-MEF	18.4047	0.8065	0.8963	0.6716
HoLoCo	20.0829	0.8236	0.8822	0.4678
HSDS-MEF	20.0444	0.7771	0.9058	0.5661
CRMEF	20.0893	0.8273	0.8740	0.6570
Ours	22.9934	0.8768	0.9359	0.7761

The substantial improvement in PSNR (2.9 dB over the second-best) demonstrates superior pixel-level accuracy. The high VIF score indicates effective information preservation from both source images, validating our two-phase approach's ability to combine complementary information.

4.3.2 Qualitative Results

Visual comparisons across diverse scenes are presented in Figure 4.1. Each comparison includes the extreme input pairs, results from competing methods, and normalized pixel value plots along marked columns to analyze exposure distribution.

doi:10.6342/NTU202504549

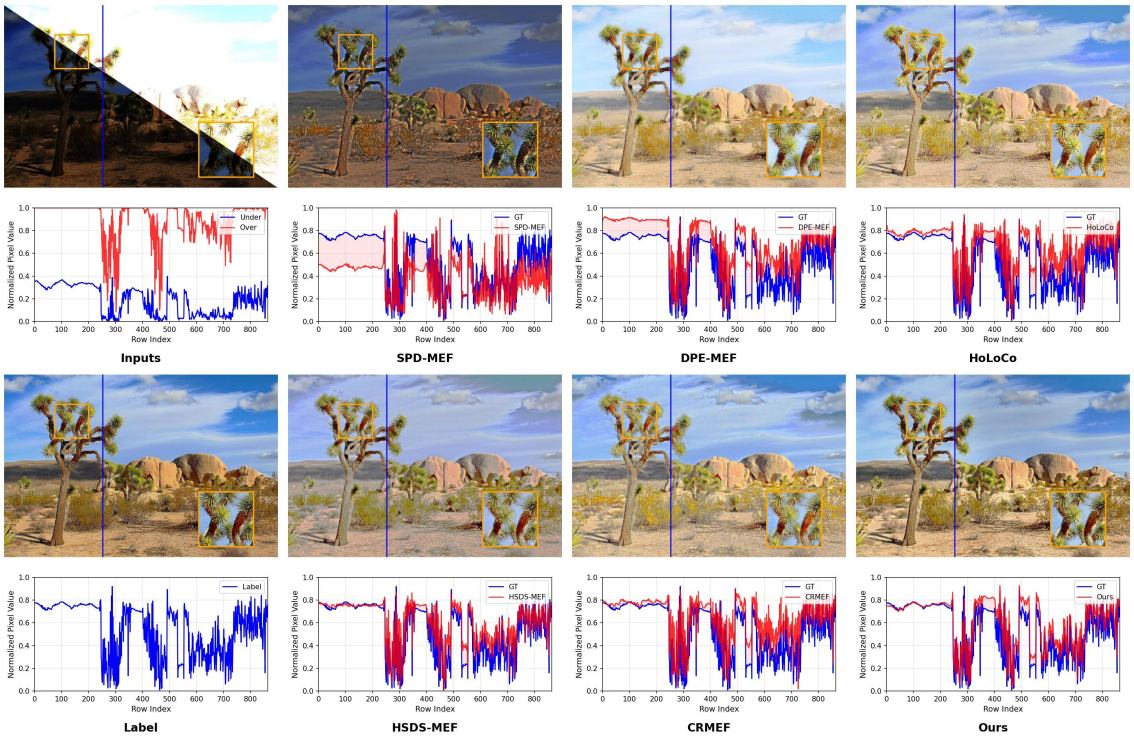
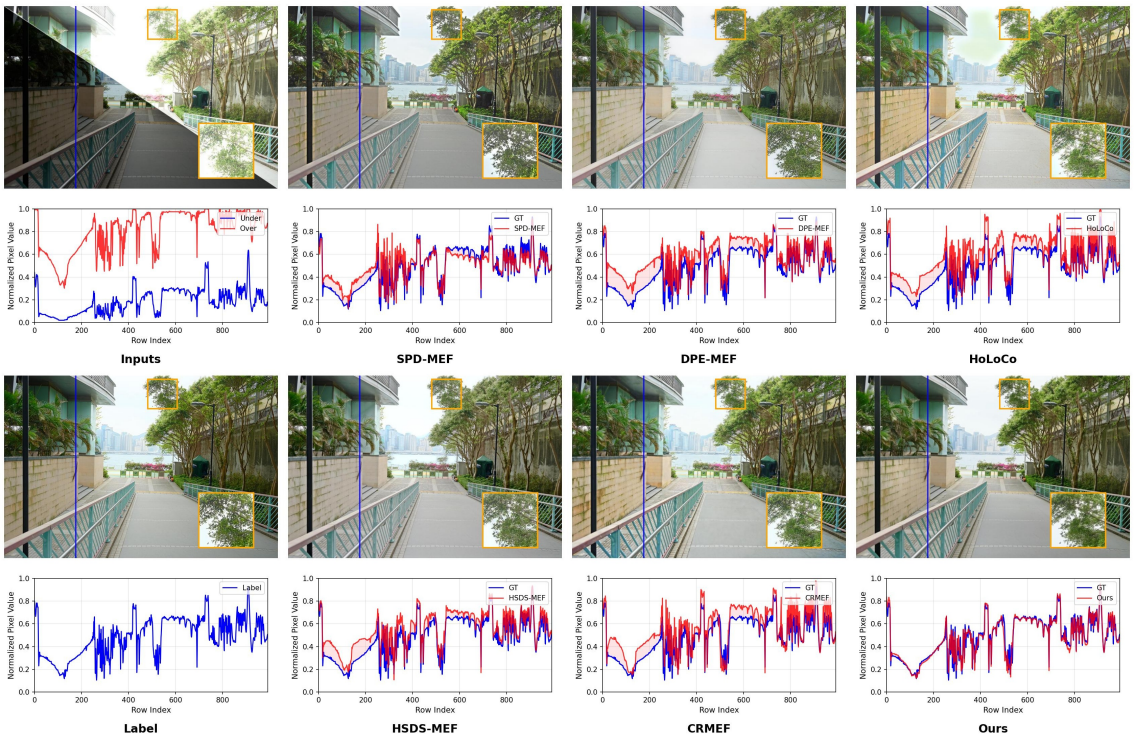
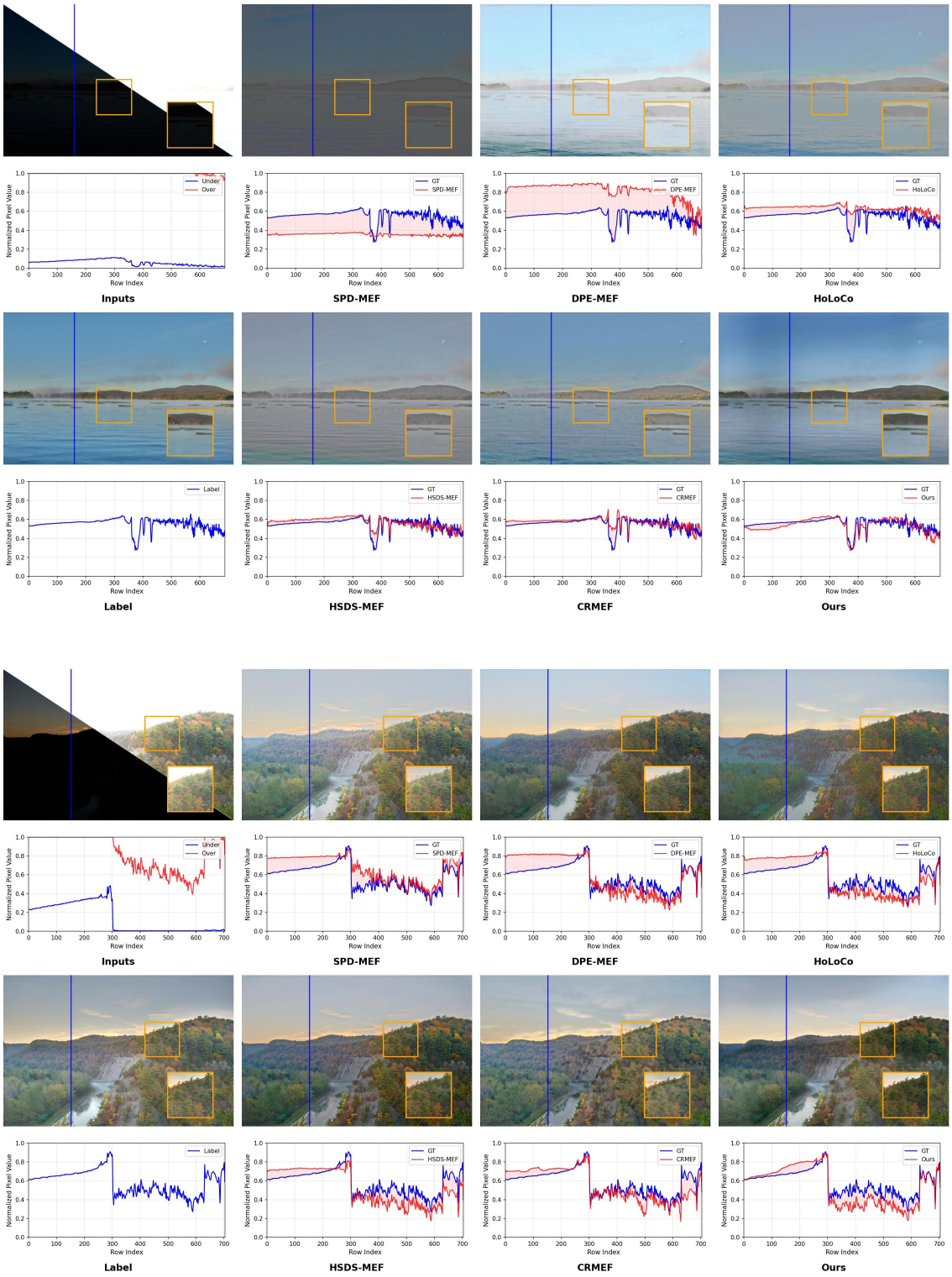


Figure 4.1: Visual comparison of different MEF methods across multiple test scenes. For each scene: first row shows Inputs, SPD-MEF, DPE-MEF, and HoLoCo; second row shows Label (ground truth), HSDS-MEF, CRMEF, and our method. Blue lines indicate analyzed columns for normalized pixel value plots, yellow boxes highlight regions of interest for detail comparison.



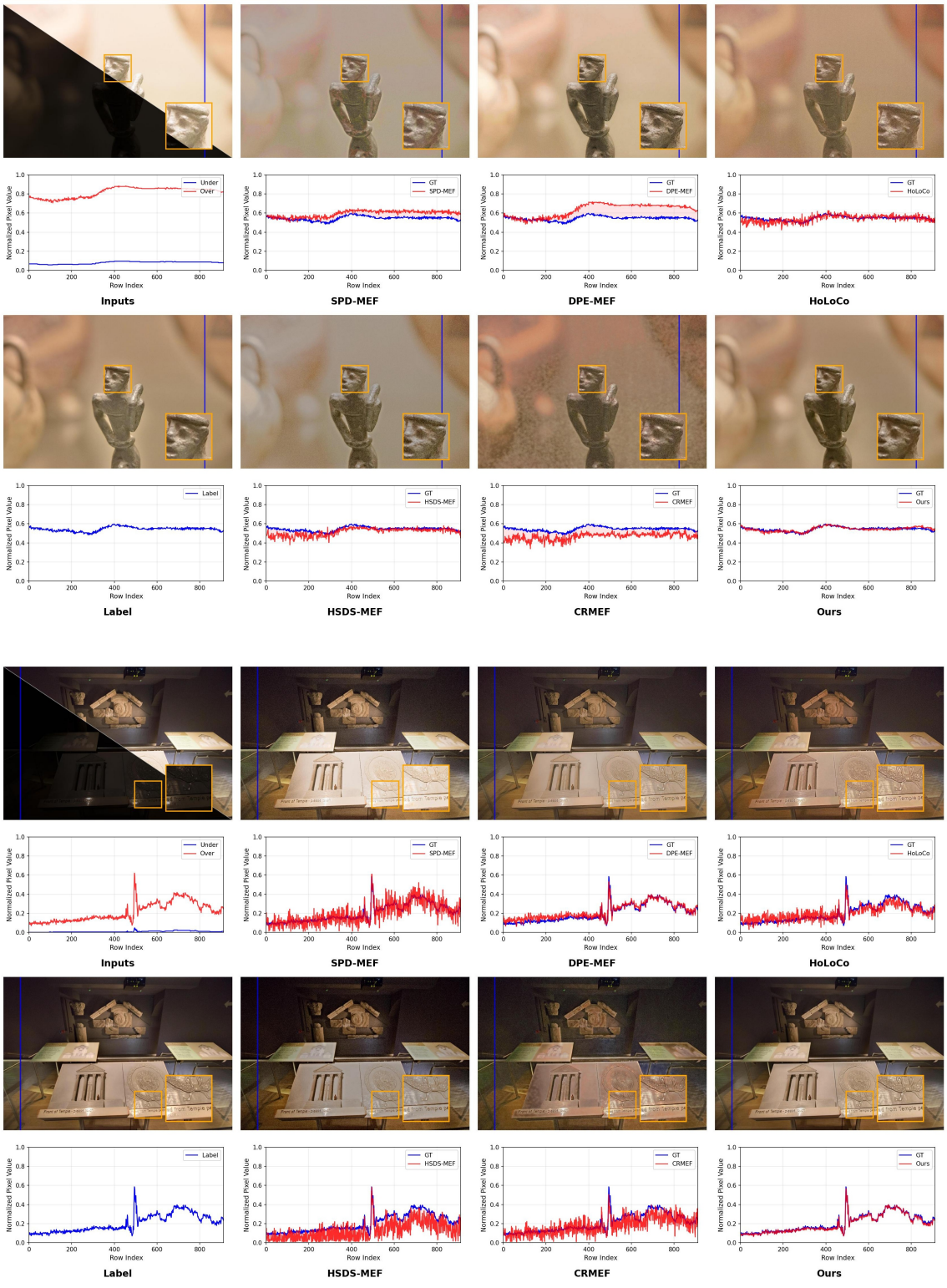
doi:10.6342/NTU202504549



4.3.3 Ablation Study

To validate the effectiveness of our two-phase design, we compare the performance of Phase One alone versus the complete two-phase framework:

doi:10.6342/NTU202504549



While the numerical improvements appear modest (PSNR +0.36 dB, SSIM +1.59%, MS-SSIM +0.20%), these metrics do not fully capture the significant local enhancements achieved by Phase Two. Global metrics like PSNR and SSIM are dominated by well-exposed regions that constitute the majority of pixels, thereby underrepresenting improve-

doi:10.6342/NTU202504549

Table 4.2: Comparison between Phase One and complete framework on SICE test set.

Configuration	PSNR↑	SSIM↑	MS-SSIM↑	VIF↑
First-Phase	22.6324	0.8631	0.9340	0.8022
Ours (Second-Phase)	22.9934	0.8768	0.9359	0.7761

ments in challenging over-exposed and under-exposed areas.

Figure 4.2 reveals the true impact of our two-phase design. The zoomed regions demonstrate that Phase Two dramatically improves detail recovery in over-exposed areas.

Figure 4.2: Visual comparison of Phase One and Phase Two outputs. Despite modest numerical improvements, Phase Two achieves significant enhancement in over-exposed regions (bottom row), recovering text details that were lost in Phase One.

These findings confirm that:

1. Phase One successfully generates a globally balanced intermediate exposure
2. Phase Two's attention-based refinement specifically targets and improves challenging local regions
3. The modest numerical gains mask substantial perceptual improvements in over/under-exposed areas
4. Our two-phase design effectively addresses both global and local fusion challenges

The disparity between numerical metrics and visual quality underscores a limitation of current evaluation metrics for MEF—they inadequately weight improvements in extreme exposure regions, which are often the most critical for practical applications.

doi:10.6342/NTU202504549

Chapter 5 Conclusion

This work presents a two-phase progressive MEF framework that offers a new perspective on handling extreme exposure differences. Our experimental results demonstrate competitive performance on the SICE dataset. The key insights from our research include: (1) separating global and local processing allows specialized optimization, with UNet handling global harmonization while attention focuses on local refinement; (2) intermediate exposure generation bridges the gap for attention mechanisms, making weights more discriminative when processing extreme pairs; and (3) decoupled training enables better phase specialization compared to joint optimization.

While our method shows promising results for static scenes, real-world photography presents additional challenges. Modern handheld devices inevitably introduce motion between captures, creating fundamental ambiguity—both exposure changes and motion manifest as pixel variations. Distinguishing valuable exposure information from motion artifacts remains an open problem.

We believe multi-phase progressive architectures could address this challenge. An initial phase could handle motion detection and compensation, followed by exposure harmonization and detail refinement in subsequent phases. This decomposition would allow each phase to specialize in its task, potentially outperforming current joint optimization

doi:10.6342/NTU202504549

approaches. Future work should embrace real-world capture complexity, modeling motion and exposure as interconnected challenges rather than independent problems, moving toward more practical solutions for computational photography.

doi:10.6342/NTU202504549

References

[1] J. Cai, S. Gu, and L. Zhang. Learning a deep single image contrast enhancer from multi-exposure images. In IEEE Transactions on Image Processing, volume 27, pages 2049–2062. IEEE, 2018.

[2] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Understanding convolution for semantic segmentation. arXiv preprint arXiv:1706.05587, 2017.

[3] D. han, L. Li, X. Guo, and J. Ma. Multi-exposure image fusion via deep perceptual enhancement. Information Fusion, 79:248–262, 2022.

[4] Y. Han, Y. Cai, Y. Cao, and X. Xu. A new image fusion performance metric based on visual information fidelity. Information Fusion, 14(2):127–135, 2013.

[5] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision, pages 694–711. Springer, 2016.

[6] S. Li and X. Kang. Image fusion with guided filtering. IEEE Transactions on Image Processing, 22(7):2864–2875, 2013.

[7] H. Liu, J. Ma, H. Xu, J. Liu, and X. Zhang. HoLoCo: Holistic and local contrastive

doi:10.6342/NTU202504549

learning network for multi-exposure image fusion. Information Fusion, 95:237–249, 2023.

[8] Z. Liu, J. Liu, G. Wu, Z. Chen, X. Fan, and R. Liu. Searching a compact architecture for robust multi-exposure image fusion. IEEE Transactions on Circuits and Systems for Video Technology, 34(7):6224–6237, 2024.

[9] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang. Robust multi-exposure image fusion: A structural patch decomposition approach. IEEE Transactions on Image Processing, 26(5):2519–2532, 2017.

[10] T. Mertens, J. Kautz, and F. Van Reeth. Exposure fusion. In Pacific Conference on Computer Graphics and Applications, pages 382–390. IEEE, 2007.

[11] K. R. Prabhakar, V. S. Srikar, and R. V. Babu. DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In IEEE International Conference on Computer Vision (ICCV), pages 4714–4722, 2017.

[12] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[13] X. Tan, H. Chen, R. Zhang, Q. Wang, Y. Kan, J. Zheng, Y. Jin, and E. Chen. Deep multi-exposure image fusion for dynamic scenes. IEEE Transactions on Image Processing, 32:5310–5325, 2023.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, 2004.

[15] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image

doi:10.6342/NTU202504549

quality assessment. Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2:1398–1402, 2003.

[16] G. Wu, H. Fu, J. Liu, L. Ma, X. Fan, and R. Liu. Hybrid-supervised dual-search: Leveraging automatic learning for loss-free multi-exposure image fusion. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 5985–5993, 2024.

[17] Q. Yan, D. Gong, Q. Shi, A. van den Hengel, C. Shen, I. Reid, and Y. Zhang. Attention-guided network for ghost-free high dynamic range imaging. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1751–1760, 2019.

doi:10.6342/NTU202504549