國立臺灣大學電機資訊學院資訊網路與多媒體研究所

碩士論文

Graduate Institute of Networking and Multimedia

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

合成多重曝光融合資料集：從靜態資料生成動態訓練數據的方法

Synthesis Multi-Exposure Fusion Dataset: Generating Dynamic Training Sets from Static Datasets

蘇浚笙

CHUN-SHENG SU

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/NTU202504663

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

合成多重曝光融合資料集：從靜態資料生成動態訓練數據的方法

Synthesis Multi-Exposure Fusion Dataset: Generating Dynamic Training Sets from Static Datasets

本論文係 蘇浚笙 (學號 R12944032)在國立臺灣大學資訊網路與多媒體研究所完成之碩士學位論文，於民國 114 年 10 月 2 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 / Graduate Institute of Networking and Multimedia on 2 October 2025 have examined a Master's Thesis entitled above presented by CHUN-SHENG SU (student ID: R12944032) 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（所、學位學程）主管 Director: 鄭卜壬

doi:10.6342/NTU202504663

摘要

本研究針對動態多重曝光融合（Multi-Exposure Fusion, MEF）資料集的稀缺問題，提出一種方法將現有的靜態資料集轉換為可供模型訓練的動態版本，以提升模型的去鬼影能力。我們設計一套演算法，能夠分析靜態場景的多曝光影像組，並自動判定可用於合成動態組件的區域。我們進行了多層面的實驗比較，涵蓋不同演算法設定以及訓練策略，藉此找出最能優化模型效能的配置。此外，我們亦進行跨資料集的實驗，以驗證方法在不同數據來源下的泛用性。實驗結果顯示，使用經本方法轉換之靜態資料集進行訓練的模型，在多項評估指標上皆能達到與真實動態資料集訓練結果相當甚至更優的表現，證明了本研究方法的實用性與穩健性。

關鍵字：多曝光影像融合、資料集、鬼影

doi:10.6342/NTU202504663

doi:10.6342/NTU202504663

Abstract

This work addresses the scarcity of dynamic multi-exposure fusion (MEF) datasets and proposes a method to transform existing static datasets into dynamic datasets suitable for training models capable of de-ghosting. We develop an algorithm that analyzes static multi-exposure image sequences to identify candidate regions for synthesizing dynamic components. We conduct extensive experiments under various conditions, including comparisons with different algorithmic settings and training strategies. These comparisons enable us to determine the optimal configuration that consistently improves model performance. Furthermore, we perform cross-dataset experiments to demonstrate the generalizability of the proposed method. Experimental results show that models trained on the transformed static datasets achieve comparable or even better performance to those trained on real dynamic datasets, confirming the practicality and robustness of our approach.

Keywords: Multi-Exposure image fusion, dataset, ghosting

doi:10.6342/NTU202504663

doi:10.6342/NTU202504663

Contents

Page

Verification Letter from the Oral Examination Committee **i**

摘要 **iii**

Abstract **v**

Contents **vii**

List of Figures **ix**

List of Tables **xi**

Chapter 1 Introduction **1**

Chapter 2 Related Works **5**

2.1 Construct Dynamic MEF Dataset 5

2.2 Few-Shot and Self-Supervised Learning 6

2.3 Adapt or Manipulate Static Datasets 7

Chapter 3 Approach **9**

3.1 Overview 9

3.2 Synthetic Dataset Construction 10

3.2.1 Generate Masks from Over-Exposed and Labeled Images 10

3.2.2 Locate Bounding Boxes Based on Set Conditions 11

3.2.3 Synthesizing Under-Exposed Images Using Bounding Boxes 12

doi:10.6342/NTU202504663

3.3 Training Strategy 13

Chapter 4 Experiments **15**

4.1 Experiments Setting 15

4.1.1 Implementation Details 15

4.1.2 Evaluation Metrics 15

4.2 Within-dataset Comparison 17

4.2.1 Static vs. Dynamic Datasets 18

4.2.2 Bounding Box Threshold 19

4.2.3 Training Strategy 21

4.2.4 Overall Comparison 25

4.3 Cross-dataset Comparison 27

Chapter 5 Conclusion **31**

References **33**

doi:10.6342/NTU202504663

List of Figures

3.1 Generated masks with the corresponding inputs under different conditions. 11

3.2 The pipeline of the synthesis process. 13

4.1 Comparison of metric responses on two image (imgA,imgB) with the same reference. DISTS demonstrates greater alignment with perceptual quality than other traditional metrics. 17

4.2 Visual comparison of fusion results for models trained on static and dynamic datasets. 19

4.3 Visual comparison of fusion results obtained with different $[\tau_{\min}, \tau_{\max}]$ settings. 21

4.4 Illustration of ghosting artifacts in the fused result, caused by motion occurring in over-exposed areas. 22

4.5 Line plot illustrating the distribution of DISTS scores across all images in the dynamic test set. 23

4.6 Visual comparison of fusion results with and without the offset-based training strategy. 24

4.7 Visual comparison of models trained with the offset strategy on the static training set and the synthetic training set. 25

4.8 Visual comparison of models trained on static,dynamic and synthetic dataset. 27

4.9 Visual comparison between models trained on static, dynamic, and synthetic datasets, focusing on the overexposed desk lamp region to evaluate fusion quality. 27

4.10 Overview of the pipeline for generating SICE-style labels from the DDMEF dataset. 28

doi:10.6342/NTU202504663

4.11 Cross-dataset visual comparison of models trained on static and synthetic datasets. 29

doi:10.6342/NTU202504663

List of Tables

4.1 Quantitative comparison on Dynamic and Static testing set. 18

4.2 Quantitative comparison on Dynamic and Static testing sets under different $[\tau_{\min}, \tau_{\max}]$ settings. 20

4.3 Quantitative comparison between models trained with and without spatial offset. 23

4.4 Quantitative comparison of models trained on static, dynamic, and synthetic datasets. 26

4.5 Quantitative comparison of cross dataset. 29

doi:10.6342/NTU202504663

doi:10.6342/NTU202504663

Chapter 1 Introduction

Multi-Exposure Fusion (MEF) is a fundamental technique in computational photography and image processing, aiming to generate a single image with enhanced details and balanced exposure by integrating multiple shots of the same scene captured under different exposure settings.

Originally, this concept stems from High Dynamic Range (HDR) imaging, which reconstructs a radiance map from multiple exposures to overcome the limited dynamic range of standard sensors. However, HDR imaging requires radiometric calibration and tone mapping to produce a displayable image, often introducing complexity and visual artifacts. Moreover, HDR reconstruction relies on additional metadata, such as exposure times and the camera response function (CRF), which are not always accessible or reliable in practice. In contrast, MEF directly fuses the differently exposed Low Dynamic Range (LDR) images in the spatial domain, eliminating the need for HDR reconstruction while achieving visually pleasing results.

Conventional MEF algorithms focused on rule-based strategies, such as Laplacian pyramid decomposition, exposure weighting, and gradient-domain fusion, to combine complementary information from differently exposed inputs [1][2] [3]. Despite their simplicity, these approaches are limited in handling complex textures and varying illumina-

doi:10.6342/NTU202504663

tion.

With the rapid advancement of deep learning, data-driven MEF frameworks have been developed to automatically learn hierarchical representations and optimal fusion strategies. The pioneering work DeepFuse by Prabhakar et al. [4] first demonstrated the effectiveness of convolutional neural networks for exposure fusion, laying the foundation for subsequent deep learning-based approaches. Building on this foundation, recent studies such as TransMEF [5] have further extended the fusion paradigm beyond traditional convolutional architectures by incorporating Transformer-based designs. Using global context modeling and long-range dependency learning, these models achieve superior structural consistency and perceptual quality in fused images.

However, such remarkable performance is generally observed only on static scenes, where the input multi-exposure images are perfectly aligned. In real-world scenarios, camera motion or object movement during exposure bracketing often leads to misalignment among inputs, resulting in ghosting artifacts manifested as unnatural duplications or blurred regions in the fused results. To address these challenges, dynamic MEF methods have been proposed to handle motion misalignment and suppress ghosting artifacts, enabling robust fusion under real-world conditions with moving objects and handheld capture.

In the field of deep learning, datasets play an indispensable role in model training and benchmarking, providing a standardized basis for evaluating model performance. The same principle applies to learning-based MEF methods. For static MEF, large-scale datasets such as SICE [6] are widely used, offering various exposure combinations and carefully curated reference labels. In contrast, dynamic MEF datasets are relatively scarce

doi:10.6342/NTU202504663

due to the high difficulty and complexity involved in their construction. Although abundant static MEF datasets exist, they provide limited benefits for training models to handle motion-induced ghosting. The lack of dynamic MEF datasets has thus become a major obstacle hindering further progress in this field.

This research aims to address the shortage of dynamic MEF datasets by developing an algorithm that transforms static MEF datasets into dynamic counterparts suitable for training de-ghosting models. Our primary focus is on handling local misalignment caused by object motion within the scene, rather than global misalignment introduced by camera movement. This choice is motivated by two considerations. First, global misalignment can be easily simulated by applying simple geometric transformations to the input images. Second, numerous existing approaches are capable of performing global alignment before MEF processing. Therefore, we restrict our scope to local misalignment, which presents a more challenging problem and directly contributes to ghosting artifacts in dynamic MEF scenarios.

The main contributions of this work are summarized as follows:

1. We propose an algorithm that automatically converts static MEF datasets into dynamic ones, enabling effective model training without relying on costly real-world dynamic datasets.

2. We conduct extensive comparisons across algorithm settings and training strategy to thoroughly assess the impact of the proposed method.

3. We demonstrate that the proposed method generalizes well across different datasets, proving its robustness and applicability.

doi:10.6342/NTU202504663

4. Experimental results show that models trained on transformed datasets achieve comparable or even superior ghosting suppression performance compared to models trained on real dynamic datasets.

The thesis is organized as follows. Chapter 2 reviews related work on addressing the scarcity of dynamic datasets. Chapter 3 introduces the proposed dataset transformation algorithm in detail. Chapter 4 presents the experimental setup and results, including comprehensive comparisons between different configurations. Finally, Chapter 5 concludes the thesis.

doi:10.6342/NTU202504663

Chapter 2 Related Works

Several studies have sought to address the scarcity of dynamic MEF datasets, approaching the problem from various perspectives. These efforts can be broadly grouped into three categories: (1) fundamental solutions through the construction of dynamic datasets, (2) self-supervised or few-shot learning approaches, and (3) methods that adapt or manipulate existing static datasets to simulate dynamic scenarios.

2.1 Construct Dynamic MEF Dataset

For the construction of dynamic MEF datasets, a major challenge lies in obtaining appropriate ground-truth labels when motion exists between under-exposed and over-exposed input images. One proposed solution [7] involves capturing multiple sets of static multi-exposure images of the same scene, where each set is internally motion-free but differs in motion across sets. For each static set, fused images are generated using several state-of-the-art MEF algorithms, and the best result is selected via a voting strategy to serve as the reference label. A dynamic MEF input can then be synthesized by pairing a low-exposure image from one set with a high-exposure image from another, while the corresponding label is taken from the static set that contributed the selected exposure image.

doi:10.6342/NTU202504663

Although this strategy effectively addresses the problem of label acquisition, it is important to note that constructing dynamic datasets is fundamentally similar to building static ones, but under much stricter requirements. Specifically, both sets of images from the same scene must meet quality standards to form a valid dynamic pair, which significantly increases the difficulty and complexity of dataset creation. While capturing multiple exposure sets for the same scene can help alleviate this issue—since additional sets allow the creation of more dynamic pairs (e.g., three sets yield six combinations, and five sets yield twenty)—this approach still has inherent limitations. Dataset construction typically aims to maximize diversity, and when all samples are derived from the same scene with only motion variations, their contribution to improving model generalization remains limited.

2.2 Few-Shot and Self-Supervised Learning

Few-shot and self-supervised learning approaches have also been proposed to mitigate the scarcity of dynamic datasets. While these studies are primarily developed for the HDR domain, they are highly relevant to MEF due to the substantial methodological overlap between the two fields.

Zhang et al. [8] proposed a self-supervised two-stage training framework, where the first stage trains a structure-focused network on unlabeled dynamic data to learn structural priors. In the second stage, the pre-trained structure-focused network is employed to assist the main network in learning robust fusion representations. Similarly, Prabhakar et al. [9] adopted a two-phase training scheme that integrates both few-shot and self-supervised paradigms. In the first stage, a small set of labeled data is used to train an initial model,

doi:10.6342/NTU202504663

which then generates pseudo labels for unlabeled samples. These pseudo-labeled samples are incorporated in the second stage to expand the training set and further enhance the model's performance.

Despite their effectiveness, both few-shot and self-supervised approaches exhibit certain limitations. Few-shot methods often suffer from limited generalization capability, as the small number of labeled dynamic samples may not adequately capture the diversity of real-world motion patterns. Self-supervised methods, on the other hand, rely on proxy objectives or pseudo-labels, which can introduce noise and degrade performance. More importantly, both approaches fail to fully exploit the large-scale, high-quality static multi-exposure datasets that are already available, leaving a substantial amount of valuable data underutilized.

2.3 Adapt or Manipulate Static Datasets

Datasets such as SICE provide a large number of high-quality static multi-exposure image groups, making them a popular choice for training dynamic MEF models. However, because SICE consists solely of static images, some studies employ data augmentation techniques to simulate motion or misalignment. Common approaches include simple geometric transformations such as scaling, translation, and rotation, which help mimic motion and increase the diversity of input patterns [10][11].

Building upon these basic augmentations, MERF [12] introduces more sophisticated transformations. Specifically, it generates multiple deformation fields through varying degrees of affine and elastic transformations, which are then applied to the over-exposed images. When paired with the corresponding under-exposed images, these distorted im-

doi:10.6342/NTU202504663

ages form pseudo-dynamic multi-exposure sets suitable for training.

Although such approaches can mitigate global misalignment and provide some de-ghosting capability, they still struggle with significant motion variations in the scene (e.g., large changes in human body movements), often resulting in noticeable artifacts in the fused outputs. Methods that adapt static datasets primarily rely on geometric modifications, which are generally insufficient to faithfully capture the complexity of real-world dynamic scenes.

doi:10.6342/NTU202504663

Chapter 3 Approach

3.1 Overview

In this work, we focus on the setting where only two exposure images are used as inputs. Each complete image set is denoted as $\{I_u, I_o, I_{gt}\}$, where I_u represents the under-exposed image, I_o the over-exposed image, and I_{gt} the corresponding reference label for the set. Following the common practice in previous works, we designate the over-exposed image as the reference, ensuring that the ground-truth label is aligned with the motion present in the over-exposed input. This design choice is motivated by the fact that, in most cases, over-exposed images retain richer scene information compared to their under-exposed counterparts. In our method, only I_u (non-reference) is modified, while I_o and I_{gt} remain unchanged to preserve information integrity.

Our primary objectives are presented as follows: (1) After applying the proposed dataset transformation, the model can be effectively trained on the resulting dataset to achieve superior de-ghosting performance. (2) Ensure that the proposed method has strong generalizability and can be applied across different datasets.

doi:10.6342/NTU202504663

3.2 Synthetic Dataset Construction

The proposed method can be broadly divided into three main stages, each designed to progressively construct a synthetic dataset that mimics the characteristics of dynamic MEF scenarios.

3.2.1 Generate Masks from Over-Exposed and Labeled Images

We perform a pixel-level comparison between I_o and I_{gt} to generate a mask M_{ogt}. The underlying intuition is that when the difference becomes too large, reconstruction based solely on the I_o becomes unreliable, necessitating the incorporation of information from the I_u. Specifically, we compute the L2 difference between I_o and I_{gt}; if the difference at a given pixel exceeds a predefined threshold, the pixel is marked as 1 in the mask, otherwise it is marked as 0.

However, when using the ℓ_2 distance to generate masks, certain over-exposed regions may exhibit small differences because the corresponding label is also bright. These nearly saturated white areas lose most of their information and cannot be reliably reconstructed, yet they may remain unmarked. This situation most commonly occurs when the sky region in the over-exposed image becomes severely saturated. As shown in Figure 3.1, the sky region in (a) appears completely white, whereas in (b), it retains a faint color tone. This indicates that reconstructing (b) requires information from the under-exposed image. However, since the color difference between the two images is relatively small, the corresponding sky region is not marked in the mask shown in (c).

To address this issue, we additionally convert the I_o into grayscale and mark pixels

doi:10.6342/NTU202504663

with brightness values exceeding a predefined threshold as 1 in the mask. After applying this refinement, the over-exposed sky regions are correctly identified, as illustrated in (d). The final formulation is expressed as follows:

$$M_{ogt}(x,y) = \begin{cases} 1, & \text{if } \|I_o(x,y) - I_{gt}(x,y)\|_2 > \varepsilon \text{ or } G(I_o(x,y)) > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

where (x, y) denotes the pixel location, $|\cdot|_2$ represents the ℓ_2 norm, and ε is a predefined threshold. The function $G(\cdot)$ denotes the grayscale transformation applied to the over-exposed image I_o, and τ is a predefined intensity threshold. In our experiments, we set the basic values as $\varepsilon = 0.5$ and $\tau = 245$.

(a) over-exposed (b) label (c) w/o intensity mask (d) w/ intensity mask

Figure 3.1: Generated masks with the corresponding inputs under different conditions.

3.2.2 Locate Bounding Boxes Based on Set Conditions

To identify candidate regions in the I_u can be modified without compromising the reconstruction of the ground truth, we design a bounding box selection strategy based on the mask M_{ogt} generated in the previous stage.

To ensure that the modification regions are spatially reasonable and well-distributed, the bounding box sizes are predefined within a specified range, with their width and height set roughly between one-third and one-half of the image dimensions. A sliding-window strategy is then applied to traverse all possible bounding box candidates. For each candi-

doi:10.6342/NTU202504663

date B_i, the selection process follows the constraints below:

$$\frac{|M_{\text{ogt}}(B_i) = 1|}{|B_i|} \in [\tau_{\text{min}}, \tau_{\text{max}}] \tag{3.2}$$

$$\frac{|M_{\text{ogt}}(B_i) = 1| + \sum_{j \in \mathcal{S}} |M_{\text{ogt}}(B_j) = 1|}{|M_{\text{ogt}}(I) = 1|} \leq \rho_{\text{max}} \tag{3.3}$$

$$|B_i \cap M_{\text{used}}| = 0 \tag{3.4}$$

Here, M_{ogt} represents the mask from the previous stage, with a value of 1 indicating regions that need information from the under-exposed image to be properly reconstructed. $|M_{\text{ogt}}(B_i) = 1|$ and $|M_{\text{ogt}}(I) = 1|$ represent the number of pixels within B_i and the entire image that are marked as 1 in M_{ogt} , respectively. $|B_i|$ denotes the total number of pixels in the bounding box.

The range $[\tau_{\text{min}}, \tau_{\text{max}}]$ specifies the acceptable proportion of marked pixels within each bounding box. $\mathcal{S}$ denotes the set of previously selected bounding boxes, and ρ_{max} limits the global proportion of marked pixels across all selected boxes to prevent excessive modification. In our experiments, ρ_{max} is set to 0.5. Finally, M_{used} records all pixels already covered by selected boxes, and Equation 3.4 ensures that no overlap occurs among them.

Together, these constraints ensure that the selected bounding boxes accurately cover potential motion-sensitive regions while maintaining spatial balance and realism in the synthesized dynamic scenes.

3.2.3 Synthesizing Under-Exposed Images Using Bounding Boxes

With the selected bounding boxes, we proceed to the final synthesis stage. A collection of foreground images with varied human poses is prepared to serve as dynamic

doi:10.6342/NTU202504663

content for insertion. In general, increasing the diversity of poses and clothing colors enhances the effectiveness of model training.

Before synthesis, a brightness transformation is computed for each bounding box region, mapping the ground-truth image to its corresponding under-exposed counterpart. The prepared foreground regions are then adjusted using this transformation. This process ensures that the composited images appear visually consistent and natural when blended into the under-exposed inputs, while simultaneously acting as an effective form of data augmentation.

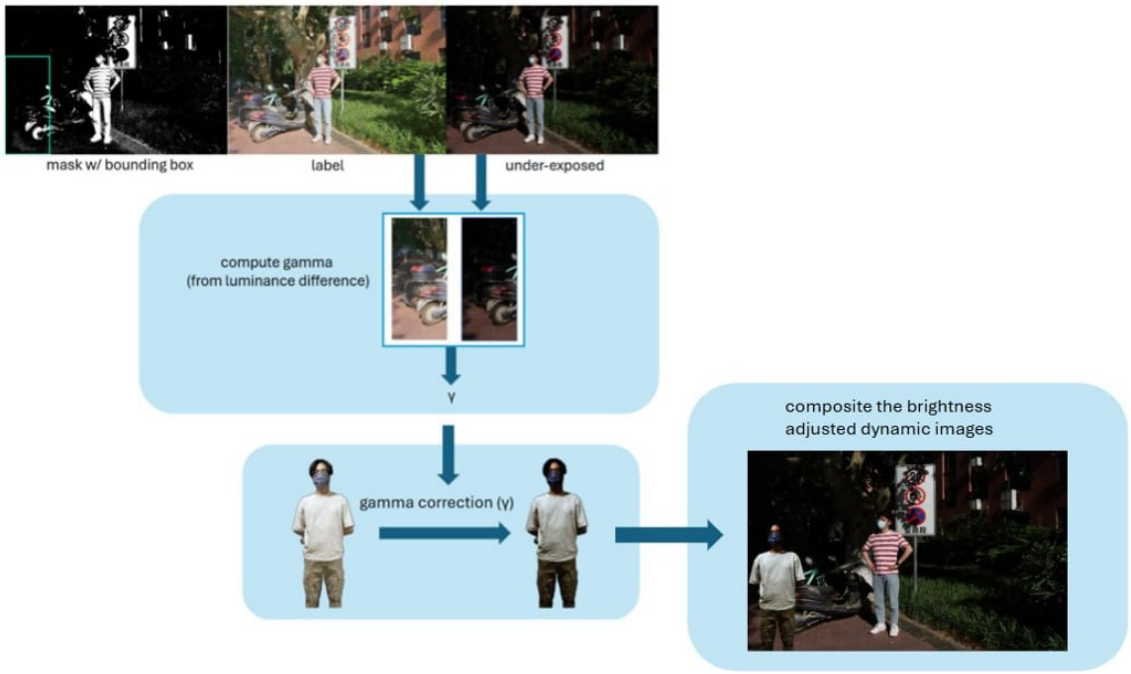


Figure 3.2: The pipeline of the synthesis process.

3.3 Training Strategy

In addition to the proposed dataset construction strategy, we introduce a complementary training scheme to further enhance model performance. After the aforementioned transformations, during model training, a small random offset is occasionally applied to

doi:10.6342/NTU202504663

the under-exposed image whenever an input patch is cropped. This strategy encourages the model to learn stronger robustness against misalignment, thereby improving its ability to remove ghosting artifacts.

doi:10.6342/NTU202504663

Chapter 4 Experiments

4.1 Experiments Setting

4.1.1 Implementation Details

We employ the model architecture provided by DDMEF [7] for training and evaluation across various datasets. During training, image patches of size 768×768 are randomly cropped as input samples. The model is trained for 1200 epochs using the Adam optimizer with a batch size of 1 and an initial learning rate of 1×10^{-4}, which is gradually decayed after 200 epochs. All other training hyperparameters are kept consistent with the original DDMEF configuration. During testing, homography-based global alignment is first applied to the input images to correct global misalignment, and the resulting aligned images are then processed by the MEF model to produce the final fused outputs.

4.1.2 Evaluation Metrics

Since the testing datasets used in the following experiments provide corresponding ground-truth labels, we are able to employ reference-based evaluation metrics to assess the quality of the generated fused images. Specifically, we use peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), multi-scale SSIM (MS-SSIM), visual infor-

doi:10.6342/NTU202504663

mation fidelity (VIF) [13] and deep image structure and texture similarity (DISTS) [14] as objective measures of visual quality. Unlike no-reference metrics, which can only provide indirect or heuristic assessments, these reference-based measures enable a more accurate and reliable evaluation of fusion performance.

Unlike the other four metrics, DISTS has not been commonly used as an evaluation metric for MEF. Instead, it is primarily applied in the field of full-reference image quality assessment (FR-IQA), particularly in research areas such as texture retrieval, image restoration and enhancement, and image super-resolution, where it measures structural and textural similarity to better align with human visual perception. We adopt DISTS in our study due to its distinctive properties:

- **Texture-aware:** Unlike SSIM or PSNR, DISTS recognizes that replacing one texture patch with another similar patch should not result in a significant penalty if the overall texture appears consistent.
- **Geometric robustness:** DISTS is tolerant to translations, scaling, and modest geometric distortions without requiring additional augmentation.
- **Strong human correlation:** Its measurements closely align with human perceptual judgments of image quality.

From Fig. 4.1, it can be observed that imgA is visually closer to the reference image compared to imgB. However, all metrics except DISTS assign higher scores to imgB. This inconsistency arises because imgA and imgB are not perfectly aligned with the reference at the global level. Metrics such as PSNR and SSIM are highly sensitive to even slight pixel-level misalignments; although these differences are often imperceptible to the human eye, they can still lead to significant score degradation in these metrics.

doi:10.6342/NTU202504663

However, due to its texture-aware and geometric robustness properties, DISTS prioritizes overall perceptual consistency over exact pixel-wise alignment. As a result, it is less penalized by minor misalignments and provides a more reliable assessment of visual similarity in such cases.

imgA imgB reference

	PSNR ↑	SSIM ↑	MSSSIM ↑	VIF ↑	DISTS ↓
imgA	20.32	0.6760	0.7293	0.3482	**0.1038**
imgB	21.20	0.7050	0.7587	0.3545	**0.1504**

metric result

Figure 4.1: Comparison of metric responses on two image (imgA,imgB) with the same reference. DISTS demonstrates greater alignment with perceptual quality than other traditional metrics.

The under-exposed and over-exposed images are not globally aligned by default, and even after homography alignment, minor misalignments often remain. Such cases are fairly common. Our study primarily focuses on addressing local motion misalignment, and to prevent global misalignment from confounding the evaluation, we select DISTS as one of the metrics for assessing fusion quality.

4.2 Within-dataset Comparison

The within-dataset setting focuses on analyzing different synthesis algorithm and training configurations.We conduct our evaluations on the DDMEF dataset, as it is, to the best of our knowledge, the most comprehensive MEF dataset currently available. It

doi:10.6342/NTU202504663

contains both static and dynamic image sets, where the dynamic subset is essentially a reorganization of the same scenes found in the static subset. This design provides a fair and convincing basis for subsequent comparisons, since the two subsets share identical content but differ in scene dynamics.

The synthetic algorithm is applied to 100 selected groups from the static training set. The trained model is evaluated on the static and dynamic testing sets, each containing 52 image groups, to examine its performance under both static and dynamic conditions.

4.2.1 Static vs. Dynamic Datasets

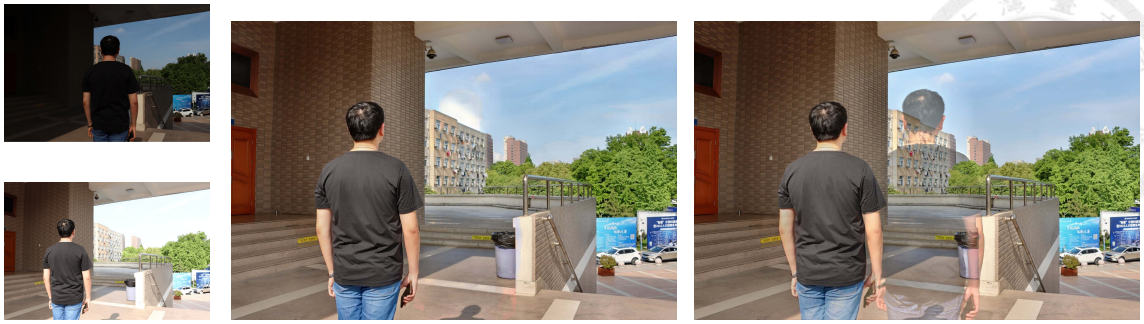
Before introducing our synthetic dataset, we first establish a baseline by examining the performance difference between static and dynamic datasets. As shown in Table 4.1, for the static testing set, models trained on static and dynamic training data perform comparably. However, a significant performance gap emerges on the dynamic testing set.

Visual results further illustrate this difference: when tested on dynamic scenes, the model trained solely on static data produces severe ghosting artifacts. These artifacts drastically degrade the image quality, leading to large score discrepancies between visually similar results. This observation highlights the importance of effective de-ghosting in dynamic multi-exposure fusion.

Table 4.1: Quantitative comparison on Dynamic and Static testing set.

Testing Set	Training set	PSNR ↑	SSIM ↑	MSSSIM ↑	VIF ↑	DISTS ↓
Dynamic	Dynamic	24.60	0.8984	0.9606	0.7163	0.0311
	Static	23.28	0.8811	0.9435	0.6705	0.0487
Static	Dynamic	24.84	0.9072	0.9667	0.7336	0.0269
	Static	25.12	0.9132	0.9707	0.7558	0.0237

doi:10.6342/NTU202504663



input dynamic training set static training set

Figure 4.2: Visual comparison of fusion results for models trained on static and dynamic datasets.

4.2.2 Bounding Box Threshold

In our synthesis algorithm, the most critical parameter is the range $[\tau_{\text{min}}, \tau_{\text{max}}]$ in Eq. (3.2), as it primarily determines whether a bounding box is valid. Other parameters, such as the bounding box sizes and the global threshold ρ_{max} in Eq. (3.3), are not highly sensitive and do not require fine-tuning; approximate values suffice. We first examine two extreme settings, [0,0.1] and [0.9,1], to analyze the effect of covering different proportions of crucial information within the bounding boxes.

As shown in Table 4.2, the configuration of $[\tau_{\text{min}}, \tau_{\text{max}}] = [0.9, 1]$ achieves better overall performance on the dynamic testing set compared to the $[0, 0.1]$ setting, while the opposite trend is observed on the static testing set. From Fig. 4.3(a), it can be observed that the setting $[0.9, 1]$ introduces noticeable artifacts in the sky regions, whereas such artifacts are absent under $[0, 0.1]$. Conversely, in Fig. 4.3(b), strong ghosting artifacts are visible in the $[0, 0.1]$ case, while the $[0.9, 1]$ configuration, despite showing minor sky artifacts, effectively suppresses ghosting.

When $[\tau_{\text{min}}, \tau_{\text{max}}] = [0, 0.1]$, the covered regions mainly correspond to unimportant areas with respect to the ground-truth reconstruction. Consequently, this setting has lit-

doi:10.6342/NTU202504663

tle impact on the model's ability to learn fusion and still provides limited de-ghosting capability. However, it struggles to handle ghosting artifacts caused by motion within over-exposed regions, leading to inferior performance on dynamic scenes. In contrast, when $[\tau_{\min}, \tau_{\max}] = [0.9, 1]$, the covered regions include important information, which hampers the model's fusion performance. Nevertheless, this setting implicitly encourages the model to learn inpainting-like behavior, enabling it to better suppress ghosting in over-exposed areas.

Table 4.2: Quantitative comparison on Dynamic and Static testing sets under different $[\tau_{\min}, \tau_{\max}]$ settings.

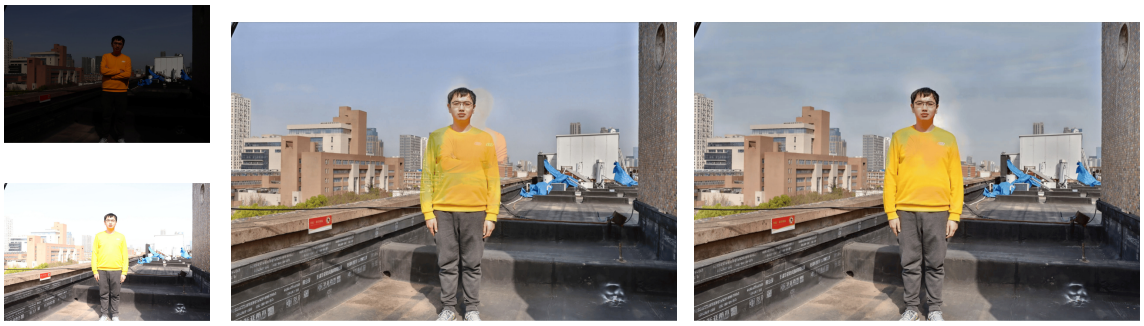
Testing set	$[\tau_{\min}, \tau_{\max}]$	PSNR ↑	SSIM ↑	MSSSIM ↑	VIF ↑	DISTS ↓
Dynamic	[0 , 0.1]	24.46	0.8919	0.9575	0.7078	0.0322
	[0.9 , 1]	24.49	0.8940	0.9584	0.7081	0.0324
	[0.1 , 0.3]	24.64	0.8930	0.9579	0.7110	0.0316
	[0.4 , 0.6]	24.65	0.8939	0.9583	0.7099	0.0319
Static	[0 , 0.1]	25.15	0.9119	0.9704	0.7542	0.0245
	[0.9 , 1]	25.08	0.9121	0.9698	0.7485	0.0254
	[0.1 , 0.3]	25.29	0.9124	0.9706	0.7567	0.0241
	[0.4 , 0.6]	25.23	0.9127	0.9700	0.7517	0.0256

Based on the above results, we can conclude that different ratios of marked pixels within the bounding boxes have distinct effects on model performance. Therefore, the optimal strategy is to adopt a balanced range that preserves the advantages of both extremes while avoiding their drawbacks. However, another important factor in $[\tau_{\min}, \tau_{\max}]$ lies in the feasibility of selecting valid bounding boxes. When the threshold is set too high, the likelihood of finding bounding boxes that meet the condition decreases, as such regions (pixels marked as 1 in the mask) typically occupy only a small portion of the image. In fact, under the [0.9, 1] condition, many images fail to produce any valid bounding boxes. Consequently, in subsequent experiments, we limit our exploration to a maximum ratio of 0.6, since higher values often result in an insufficient number of eligible bounding boxes.

doi:10.6342/NTU202504663

input $[\tau_{\min}, \tau_{\max}]$=[0,0.1] $[\tau_{\min}, \tau_{\max}]$=[0.9,1]

(a)



input $[\tau_{\min}, \tau_{\max}]$=[0,0.1] $[\tau_{\min}, \tau_{\max}]$=[0.9,1]

(b)

Figure 4.3: Visual comparison of fusion results obtained with different $[\tau_{\min}, \tau_{\max}]$ settings.

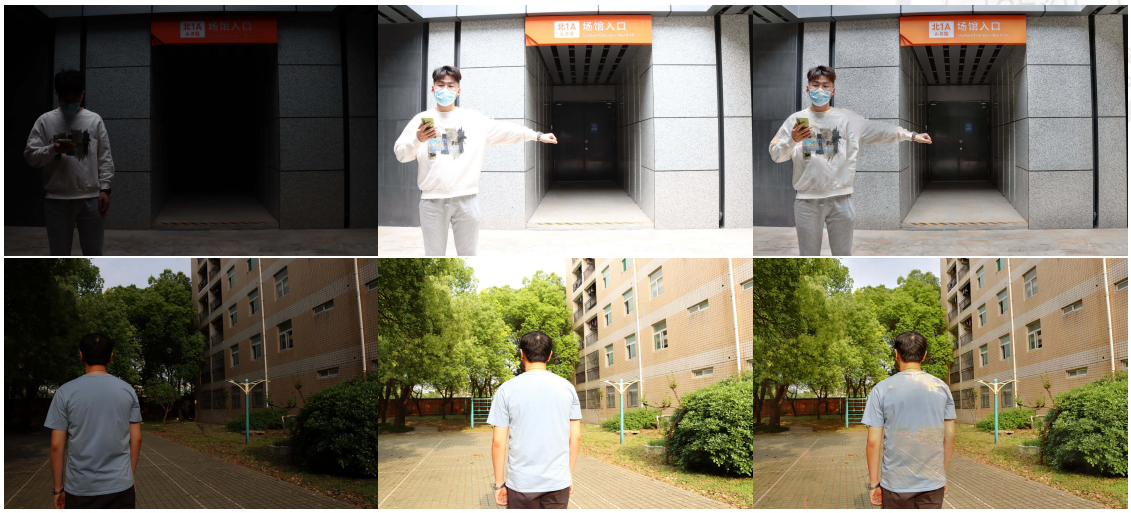
Finally, as shown in Table 4.2, the configuration $[\tau_{\min}, \tau_{\max}] = [0.1, 0.3]$ achieves the better trade-off between fusion quality and de-ghosting performance across both static and dynamic testing sets. Although its performance is comparable to the $[0.4, 0.6]$ setting, the $[0.1, 0.3]$ range allows for easier selection of valid bounding boxes. Therefore, we adopt this configuration for all subsequent experiments.

4.2.3 Training Strategy

Although we identified the configuration that best balances fusion and de-ghosting performance in the previous stage, the model trained under this condition still exhibits noticeable artifacts in overexposed regions, as shown in Fig 4.4. Since the model trained on the dynamic dataset successfully addresses this issue, we conducted a deeper investigation

doi:10.6342/NTU202504663

into the dynamic training data.



Under Exposed Over Exposed Fused Result

Figure 4.4: Illustration of ghosting artifacts in the fused result, caused by motion occurring in over-exposed areas.

Specifically, we fine-tuned the model trained on the synthetic dataset on each individual sample of the dynamic training set to determine which samples contributed most to resolving the artifacts. Interestingly, we found that only a small subset of the training samples led to improvement. A common characteristic among these samples is that the under and over-exposure input images exhibit global misalignment, with slight positional shifts between them. This observation inspired us to incorporate such characteristics into the training process of our synthetic dataset.

Building upon the synthetic dataset from the previous stage, we further introduce occasional small spatial offsets to the cropped under-exposed images during training to mimic global misalignment. As shown in Table 4.3, this modification yields a modest performance gain on the dynamic test set, while causing a slight drop in performance on the static test set. By plotting the DISTS scores of all dynamic test samples in Fig 4.5, we observe a few cases with large performance gaps, mostly corresponding to motion in overexposed regions. The offset-based strategy effectively suppresses ghosting in these

doi:10.6342/NTU202504663

cases, resulting in larger score differences for these samples; however, since such scenarios are relatively rare in the dataset, the overall improvement in quantitative metrics remains limited. As shown in Fig. 4.6, the visual results clearly highlight the significant impact of this strategy.

Table 4.3: Quantitative comparison between models trained with and without spatial offset.

Testing set	Offest Applied	PSNR ↑	SSIM ↑	MSSSIM ↑	VIF ↑	DISTS ↓
Dynamic	X	24.64	0.8930	0.9579	0.7110	0.0316
	✓	24.68	0.8976	0.9666	0.7194	0.0313
Static	X	25.29	0.9124	0.9706	0.7567	0.0241
	✓	25.12	0.9112	0.9744	0.7498	0.0258

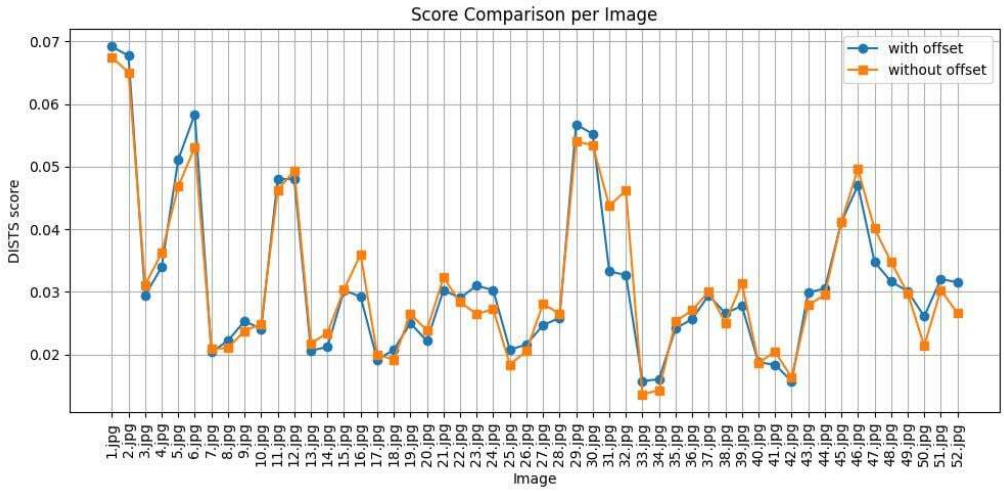
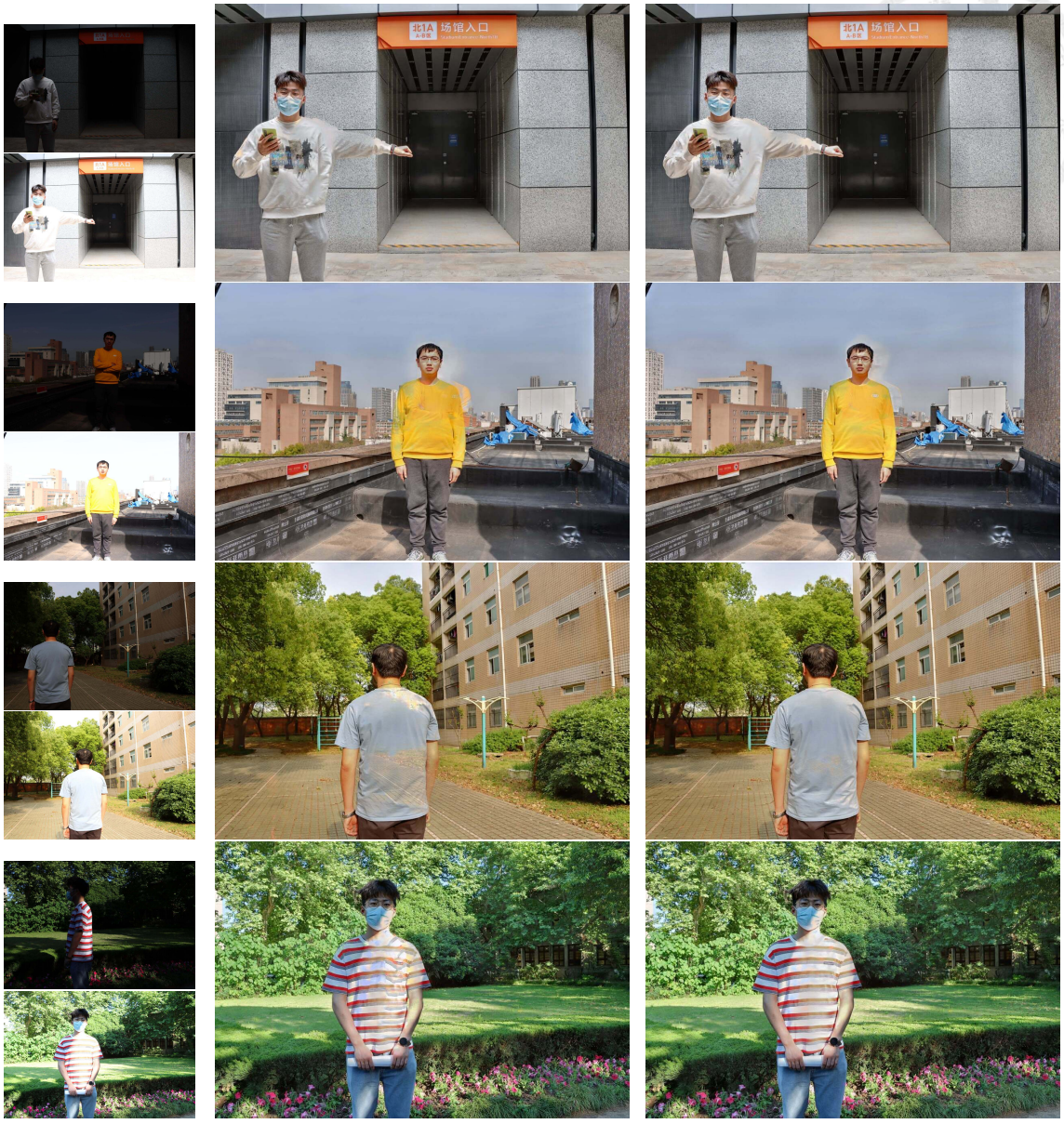


Figure 4.5: Line plot illustrating the distribution of DISTS scores across all images in the dynamic test set.

doi:10.6342/NTU202504663



input without offset with offset

Figure 4.6: Visual comparison of fusion results with and without the offset-based training strategy.

Since this strategy is similar to many dynamic MEF methods—training models for dynamic scenarios by applying basic geometric operations to static inputs—we further apply it to the static training set to observe its effect. As shown in Fig 4.7, the results are significantly worse than those obtained with the synthetic dataset: not only does the ghosting issue persist, but the overall fusion quality also decreases. This demonstrates that the strategy is effective only when combined with the synthetic dataset, and its standalone

doi:10.6342/NTU202504663

impact on static data is limited.

input static dataset w offset synthetic dataset w offset

Figure 4.7: Visual comparison of models trained with the offset strategy on the static training set and the synthetic training set.

4.2.4 Overall Comparison

Finally, we conduct a comprehensive comparison among models trained on the static, dynamic, and synthetic datasets to summarize the overall findings. As shown in Table 4.4, the model trained on the synthetic dataset achieves comparable quantitative performance to the one trained on the dynamic dataset, while both significantly outperform the model trained on the static dataset. As illustrated in Fig 4.8, when the input images contain large motion differences, the models trained on the synthetic and dynamic datasets exhibit strong de-ghosting capabilities, whereas the static-trained model produces noticeable ghosting artifacts.

We observe that on the static test set, the model trained on the synthetic dataset achieves performance comparable to that of the model trained on the static dataset, with

doi:10.6342/NTU202504663

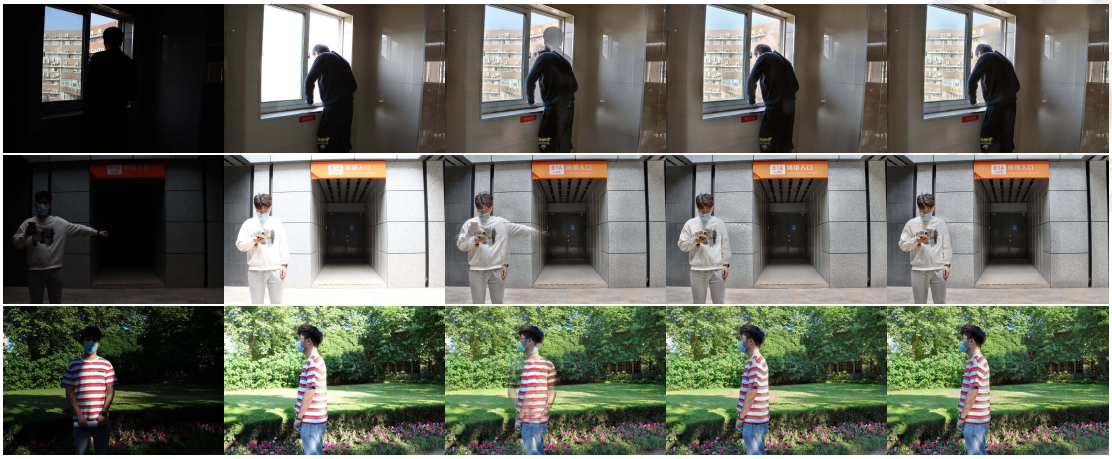
Table 4.4: Quantitative comparison of models trained on static, dynamic, and synthetic datasets.

Testing set	Training Set	PSNR ↑	SSIM ↑	MSSSIM ↑	VIF ↑	DISTS ↓
Dynamic	Static	23.28	0.8811	0.9435	0.6705	0.0487
	Dynamic	24.60	0.8984	0.9606	0.7163	0.0311
	synthetic	24.68	0.8976	0.9666	0.7194	0.0313
Static	Static	25.12	0.9132	0.9707	0.7558	0.0237
	Dynamic	24.84	0.9072	0.9667	0.7336	0.0269
	synthetic	25.12	0.9112	0.9744	0.7498	0.0258

both exhibiting a small performance gap relative to the model trained on the dynamic dataset. As illustrated in Fig 4.9, when reconstructing the lamp highlights, only the model trained on the dynamic dataset produces slightly blurred highlights, indicating less effective preservation of fine details, whereas the synthetic and static set models better retain these details.

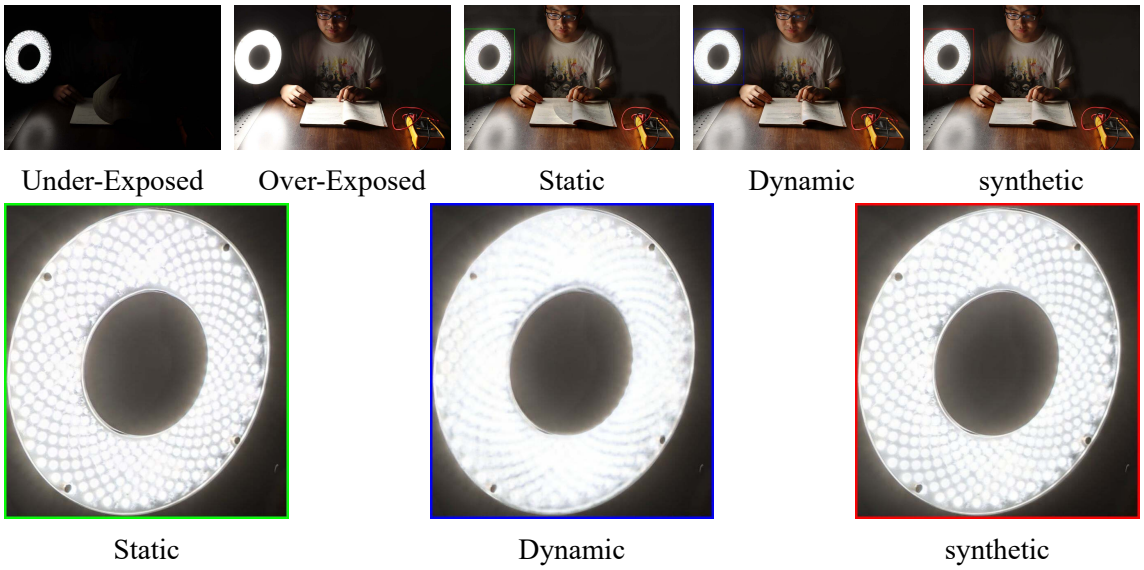
Based on the above experimental results, we can conclude that using the synthetic dataset as the training set yields performance comparable to that of the dynamic dataset, particularly in handling large motion differences, while significantly outperforming the static dataset in dynamic scenarios. On static test data, the synthetic dataset also maintains competitive performance, closely matching the model trained on the static dataset and demonstrating only a minor gap relative to the dynamic dataset. These observations indicate that the synthetic dataset effectively bridges the gap between static and dynamic datasets, providing both strong de-ghosting ability and reliable detail preservation.

doi:10.6342/NTU202504663



Under-Exposed Over-Exposed Static Dynamic synthetic

Figure 4.8: Visual comparison of models trained on static,dynamic and synthetic dataset.



Under-Exposed Over-Exposed Static Dynamic synthetic

Static Dynamic synthetic

Figure 4.9: Visual comparison between models trained on static, dynamic, and synthetic datasets, focusing on the overexposed desk lamp region to evaluate fusion quality.

4.3 Cross-dataset Comparison

In this section, we apply our proposed algorithm to the SICE dataset, which is currently the largest and most widely used dataset for training MEF models. The model is then trained on the transformed dataset and evaluated on the DDMEF dynamic test set to assess its performance in handling dynamic scenes.

doi:10.6342/NTU202504663

Due to stylistic differences in the labels across datasets, we cannot directly compare the outputs of a model trained on SICE with the original labels of the DDMEF dataset. As illustrated in Fig 4.10, we first train an MEF model on the original SICE dataset and use it to generate SICE-style labels for the DDMEF static test set. Subsequently, the model trained on the synthetic SICE dataset produced via our method is evaluated on the DDMEF dynamic test set, and its fused outputs can then be compared against the generated SICE-style labels. This comparison is possible because, in the DDMEF dataset, the static and dynamic test sets are derived from the same image groups, with the only difference being the selection of input images.

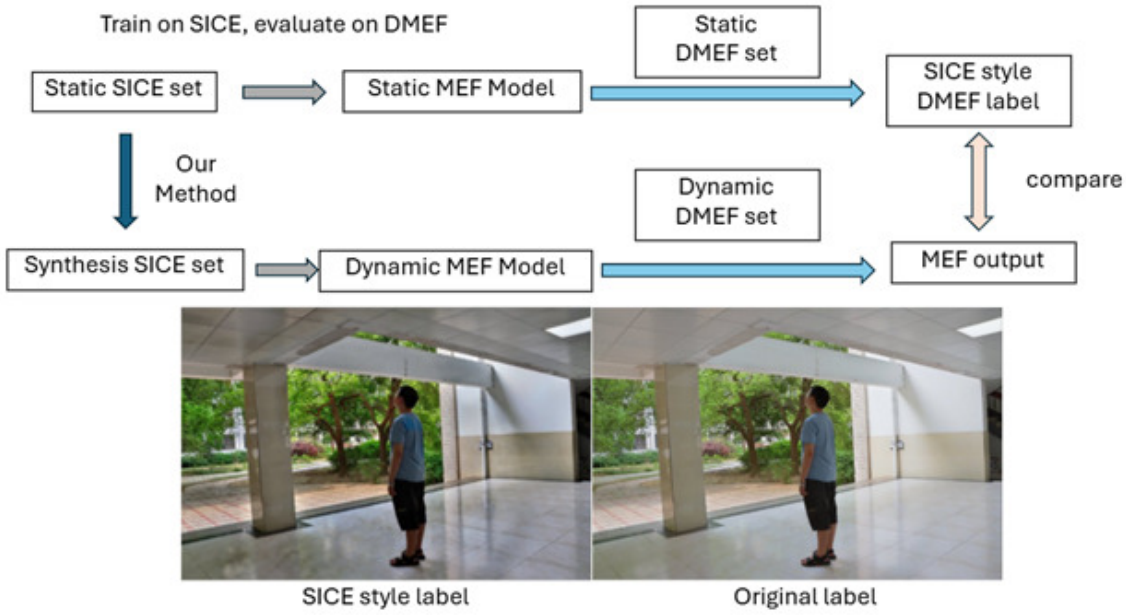


Figure 4.10: Overview of the pipeline for generating SICE-style labels from the DDMEF dataset.

Since the SICE dataset itself does not contain dynamic scenes, our comparison is limited to the synthetic dataset generated by our method and the original SICE static dataset. From Table 4.5, it can be observed that the model trained on the synthetic dataset consistently outperforms the model trained on the original static SICE dataset across nearly all metrics. As shown in Fig 4.11, ghosting artifacts that could not be resolved by the model

doi:10.6342/NTU202504663

trained on the original SICE dataset are effectively mitigated when trained on the synthetic dataset. These results demonstrate that our method is not limited to a single dataset; rather, it exhibits strong generalizability and robustness, making it applicable across different datasets.

Table 4.5: Quantitative comparison of cross dataset.

Testing set	Training Set	PSNR ↑	SSIM ↑	MSSSIM ↑	VIF ↑	DISTS ↓
Dynamic	Static	27.16	0.9375	0.9570	0.7723	0.039
	synthetic	28.93	0.9461	0.9679	0.7438	0.025

input static synthetic

Figure 4.11: Cross-dataset visual comparison of models trained on static and synthetic datasets.

doi:10.6342/NTU202504663

doi:10.6342/NTU202504663

Chapter 5 Conclusion

In this study, we successfully developed a method to transform static MEF datasets into synthetic datasets that enable models to learn effective ghosting removal. Through experiments, we demonstrate that model trained on the synthetic dataset achieves both strong ghosting removal and high-quality fusion, often surpassing model trained solely on dynamic datasets. Furthermore, cross-dataset evaluations indicate that our method exhibits strong generalizability and is not restricted to a particular dataset. This approach allows existing large-scale static MEF datasets to be leveraged for training dynamic MEF models, while dynamic MEF datasets can be more extensively reserved for testing to provide a comprehensive evaluation.

In future work, we aim to make the synthesis process more automated, such as automatically analyzing dataset properties (e.g., exposure range) to select the most suitable parameters. In addition, incorporating diffusion models into the synthesis pipeline—for instance, to generate dynamic content within the bounding boxes—could greatly improve the diversity and realism of the synthesized dataset.

doi:10.6342/NTU202504663

doi:10.6342/NTU202504663

References

[1] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In 15th Pacific Conference on Computer Graphics and Applications (PG'07), pages 382–390. IEEE, 2007.

[2] Rui Shen, Imran Cheng, Jiayi Shi, and Arijit Basu. Generalized random walks for fusion of multi-exposure images. IEEE Transactions on Image Processing, 20(12):3634–3646, Dec 2011.

[3] Sheng Li and Xin Kang. Fast multi-exposure image fusion with median filter and recursive filter. IEEE Transactions on Consumer Electronics, 58(2):626–632, May 2012.

[4] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In Proceedings of the IEEE international conference on computer vision, pages 4714–4722, 2017.

[5] Linhao Qu, Shaolei Liu, Manning Wang, and Zhijian Song. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 2126–2134, 2022.

doi:10.6342/NTU202504663

[6] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing, 27(4):2049–2062, 2018.

[7] Xiao Tan, Huaian Chen, Rui Zhang, Qihan Wang, Yan Kan, Jinjin Zheng, Yi Jin, and Enhong Chen. Deep multi-exposure image fusion for dynamic scenes. IEEE Transactions on Image Processing, 32:5310–5325, 2023.

[8] Zhilu Zhang, Haoyu Wang, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Self-supervised high dynamic range imaging with multi-exposure images in dynamic scenes. arXiv preprint arXiv:2310.01840, 2023.

[9] K Ram Prabhakar, Gowtham Senthil, Susmit Agrawal, R Venkatesh Babu, and Rama Krishna Sai S Gorthi. Labeled from unlabeled: Exploiting unlabeled data for few-shot deep hdr deghosting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4875–4885, 2021.

[10] Sheng-Yeh Chen and Yung-Yu Chuang. Deep exposure fusion with deghosting via homography estimation and attention learning. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1464–1468. IEEE, 2020.

[11] Wenhui Hong, Hao Zhang, and Jiayi Ma. Ofpf-mef: An optical flow guided dynamic multi-exposure image fusion network with progressive frequencies learning. IEEE Transactions on Multimedia, 26:8581–8595, 2024.

[12] Wenhui Hong, Hao Zhang, and Jiayi Ma. Merf: A practical hdr-like image generator via mutual-guided learning between multi-exposure registration and fusion. IEEE Transactions on Image Processing, 33:2361–2376, 2024.

doi:10.6342/NTU202504663

[13] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. IEEE Transactions on image processing, 15(2):430–444, 2006.

[14] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence, 44(5):2567–2581, 2020.

doi:10.6342/NTU202504663