REFLAP: EFFICIENT AND SCALE-ROBUST REFLECTION REMOVAL VIA DUAL-DOMAIN FEATURE FUSION

REFLAP: EFFICIENT AND SCALE-ROBUST REFLECTION REMOVAL VIA DUAL-DOMAIN FEATURE FUSION

By YUNZHE LI, BEng

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Applied Science

McMaster University

MASTER OF APPLIED SCIENCE (2025)

Hamilton, Ontario, Canada (Electrical and Computer Engineering)

TITLE:	Reflap: Efficient and Scale-Robust Reflection Removal via Dual-Domain Feature Fusion
AUTHOR:	Yunzhe Li BEng (Electrical Engineering), McMaster University, Hamilton, Canada
SUPERVISOR:	Jun Chen Professor, Department of Electrical and Computer Engineering, McMaster University, ON, Canada
NUMBER OF PAGES:	xii, 43

To my dear parents, supervisor, and co-workers.

Abstract

Single image reflection removal (SIRR) remains a challenging problem due to the intricacies involved in separating layers with varying textures and intensities. While many recent methods have focused on maximizing perceptual quality or pushing performance benchmarks, their complexity and computational cost often hinder practical deployment. In this work, we propose a dual-branch reflection removal network within a Deep Laplacian Pyramid Network framework, which balances performance and efficiency through a structurally meaningful design. The frequency-domain branch, DWT-FFC, exploits Discrete Wavelet Transform and Fast Fourier Convolution inside a U-Net architecture to capture multi-scale frequency cues and suppress reflection patterns. While the spatial-domain branch, UHDM, uses pixel unshuffling, Residual Dense Blocks (RDB), and Scale Attention Modules (SAM) to improve the structural consistency of image restoration and restore fine details. For cross-domain integration to be robust, a hierarchical fusion strategy is proposed that adaptively transfers multi-scale residuals from the Laplacian-based DWT-FFC branch to guide the UHDM decoder through cross-scale attention. Various experimental results show that our method can eliminate reflections efficiently while holding onto sharp textures. Although our method does not outperform the latest state-of-the-art solutions in terms of quantitative metrics, we demonstrate that its structural simplicity, favorable model

size, fast inference speed, and lower FLOPs make it a practical and efficient choice for lightweight reflection removal in real-world applications.

Acknowledgements

I would like to sincerely thank my supervisor, Prof. Jun Chen, for his continuous guidance, encouragement, and expertise throughout this research. I am also deeply grateful to Wei Dong for his insightful suggestions and helpful feedback, which greatly contributed to the development of this work. Special thanks go to Liangyan Li for her patient support and mentorship in deep learning and image processing, especially during the early stages of my research. Finally, I thank my parents for their unwavering love, understanding, and constant encouragement, which have been a continuous source of strength.

Table of Contents

List of Figures

List of Tables

Abbreviations

Abbreviations

AI	Artificial intelligence
SIRR	Single image reflection removal
DWT-FFC	Discrete Wavelet Transform and Fast Fourier Convolution
RDB	Residual Dense Blocks
SAM	Scale Attention Modules
PSNR	Peak Signal-to-Noise Ratio
CNN	Convolutional Neural Network
SSIM	Structural Similarity Index Measure
UHDM	Ultra-high-definition Demoir´eing
Adam	Adam optimizer
MS-SSIM	Multi-Scale Structural Similarity

VGG	Visual Geometry Group
ReLU	Rectified Linear Unit
TTA	Test-Time Augmentation

Chapter 1

Introduction

Due to reflections from glass surfaces and transparent barriers, image quality decreases, which, in turn, diminishes the performance of downstream computer vision tasks such as object detection or segmentation. With consumer photography, autonomous vehicles, and augmented reality applications becoming increasingly dominant trends, it has become urgent to solve the single image reflection removal (SIRR) challenge, which remains largely difficult in computer vision at the low level. In essence, the basic formulation of SIRR involves recovering a clean image that is free of reflection from a single image input containing a mixture of two layers: background and reflections. Due to this underdetermined nature of the problem, initial solutions [44, 29, 59, 14, 2, 28] enforced strong priors or introduced more inputs such as multiple frames or polarizers. However, recent improvements in deep learning have resulted in considerable advancement concerning the single-image setting.

Existing SIRR methodologies can be clustered into three main groups: classical prior-based, deep learning spatial, and frequency domain approaches. Earlier prior-based models[33, 47, 32, 48] depended on assumptions such as gradient sparsity,

smoothness of reflections, or foreground edge prediction. Though simple in construction, these models collapse in real-world applications. Modern methods based on deep learning gave rise to more robust models like CEILNet[12], ERRNet[57], and IBCLN[30], with multiple-stage refinement, edge-preserving modules, and dual-branch reasoning. Still, these methods provide the best performance at the expense of high computational costs, numerous parameters, and poor resource-constraint adaptability. Recently, frequency-based methods[7, 3, 53, 36, 25] such as FFCR-Net[36] have emerged, showing the advantage of spectral-domain learning, although these are prone to over-smoothing and spatial detail loss. Deep Laplacian pyramid networks[25], however, first introduced for super-resolution, are promising candidates for the efficient, interpretable, and progressive image generation.

To address the limitations of existing methods that often struggle with either recovering fine details or fully suppressing strong reflections, we propose a novel two-branch neural architecture within a **Deep Laplacian Pyramid Network** framework that integrates both spatial- and frequency-domain features for effective reflection removal. Our model consists of two parallel branches: a **Laplacian decomposition branch (DWT-FFC)** [75]and an **image reconstruction branch (UHDM)**[46]. The DWT-FFC branch leverages Discrete Wavelet Transforms (DWT) and Fast Fourier Convolutions (FFC) within a U-Net[45] structure to capture multi-scale frequency information, facilitating the suppression of reflection patterns across varying scales. In parallel, the UHDM branch operates in the spatial domain and applies pixel unshuffling for resolution-aware feature extraction, followed by Residual Dense Blocks (RDB)[70] and Scale Attention Modules (SAM)[11] to enhance spatial constancy and reconstruct fine details of the background image.

Embedded within a Laplacian pyramid, our model progressively reconstructs the reflection-free image from coarse to fine resolution. This hierarchical structure not only allows early-stage inference for low-resource settings by truncating the pyramid but also ensures effective detail restoration at higher resolutions. To enhance collaboration between the frequency and spatial branches, we introduce a **hierarchical cross-scale attention fusion mechanism**, wherein multi-scale residuals from the DWT-FFC branch dynamically guide the UHDM decoder.

Our design leverages several additional advantages of the Deep Laplacian Pyramid architecture:

- **High accuracy:** Achieved through progressive residual learning and dual-domain feature extraction, resulting in sharper and more natural restoration without over-smoothing.
- **Fast and efficient inference:** Maintained by operating primarily in low-resolution spaces and avoiding early upsampling, making the model well-suited for real-time or mobile deployment.
- **Resource-aware flexibility:** Allows for intermediate-resolution outputs by truncating the pyramid, which is beneficial in constrained environments.
- **Scalable depth control:** Provides a balance between accuracy and computational efficiency, enabling the model to adapt to different application requirements.

Our research objectives are as follows:

- To develop a hybrid Laplacian pyramid-based architecture that integrates frequency and spatial information for enhanced reflection separation and detail

preservation.

- To design a robust cross-branch fusion mechanism that facilitates adaptive multi-scale collaboration between domains.
- To establish an effective training strategy utilizing multi-level supervision and tailored loss functions for both global reconstruction and fine-grained refinement.

Our main contributions are:

- A novel reflection removal framework based on a Deep Laplacian Pyramid Network with dual-domain branches (DWT-FFC and UHDM), effectively capturing both global spectral cues and local spatial structures.
- A cross-scale attention fusion mechanism that enables seamless interaction between the two branches, improving reflection suppression and background recovery.
- Extensive experimental results demonstrate that our method achieves fast inference performance on standard benchmarks, offering a compelling balance between visual quality, efficiency, and flexibility.

Chapter 2

Related Works

2.1 Reflection Removal

Single image reflection removal (SIRR) has been a long-standing challenge in low-level vision because of the inherent ambiguity in separating two superimposed image layers from a single observation. Early approaches [51, 47, 33, 29] in this area employed classical image priors to guide the decomposition. For instance, gradient sparsity, edge prediction, and hand-crafted reflection priors assumed reflections could be smoother, with less texture, or exhibit gradient orientations that were distinct from the foreground. The two broad classes of these prior-based systems are guided filter and low-rank systems and reflection systems assuming manual annotation of reflection regions using a foreground mask. Theoretically simple, the approaches would fail to work in a large variety of real-world scenes due to hard assumptions.

With the rise of deep learning, data-driven methods have taken center stage. Early CNN-based methods such as **CEILNet** introduced edge prediction modules to preserve structure [12], while **Zhang et al.** proposed multi-stage networks that

iteratively refine reflection separation [69]. More advanced models, like **ERRNet** [58] and **IBCLN** [35], utilized dual-branch networks and recurrent refinement modules to improve perceptual quality and enhance layer disentanglement. These models significantly improved visual precision and benchmark performance.

However, most of these modern methods share common drawbacks: they are often large and deeply stacked, requiring high GPU memory and long training or inference times. Models like **IBCLN** use complex iterative reasoning and residual loops that boost PSNR [35, 56, 74] but come with heavy computational cost. In practice, the trade-off between performance and efficiency is rarely addressed — most methods optimize solely for PSNR or SSIM [18], ignoring the demands of real-time or mobile deployment [39, 19].

In contrast, our method adopts a structurally meaningful and computationally efficient architecture. While our performance in PSNR may not surpass the most recent SOTA, our design — built on dual-domain collaboration (DWT-FFC + UHDM) and a Laplacian pyramid backbone — offers clear benefits in model simplicity, inference speed, and computational cost. Our model has significantly fewer parameters, lower FLOPs [39], and faster inference time than large transformer-based or deeply recursive architectures, making it well-suited for real-world reflection removal scenarios where efficiency and interpretability matter more than marginal gains in pixel-level metrics.

2.2 Frequency-Based Restoration

Frequency-domain information plays a critical role in many image restoration tasks, including denoising, deblurring, and reflection suppression. Early frequency-based

image processing leveraged classical tools such as the Fourier Transform and Wavelet Decomposition [16, 38] to isolate image components at different scales. However, their limited learning capacity and hand-crafted nature restricted their ability to adapt to complex image degradation.

Recent work in deep learning has revived interest in frequency-based methods, particularly due to the ability to learn global structures and repetitive patterns that spatial convolutions often miss. Fast Fourier Convolution (FFC) [9] is a notable example—it introduces a learnable frequency convolution block that operates in both spatial and spectral domains, allowing models to capture long-range dependencies and global textures. Zhang et al. [68] extended this idea in FFCR-Net, a reflection removal network that uses FFC inside a dual-encoder-decoder structure to suppress reflections by attending to global spectral features. While FFCR-Net achieves notable improvements in some cases, its heavy reliance on frequency branches can lead to over-smoothing and loss of spatial details, especially when reflections and background textures overlap in frequency.

In our approach, we combine the advantages of frequency-based feature extraction with spatial-domain reconstruction. Our DWT-FFC branch incorporates Discrete Wavelet Transform (DWT) to separate multi-scale frequency components and combines it with Fast Fourier Convolution to enhance global feature awareness. Compared to FFCR-Net, our frequency branch is lighter, more interpretable, and integrates seamlessly into the Laplacian pyramid framework. More importantly, we complement it with a spatial UHDM branch that recovers local textures and fine details that the frequency branch may overlook. This dual-branch design ensures that our model avoids over-smoothing.

Despite not pushing the frontier in absolute PSNR, our approach introduces a better balance between accuracy and practical deployment metrics, including reduced FLOPs [40], smaller model size, and faster inference speed, which are qualities often overlooked by frequency-dominant designs. In lightweight applications or edge computing environments, this balance is far more valuable than minor accuracy gains.

2.3 Deep Laplacian Pyramid Networks

Laplacian pyramids were first proposed by Burt and Adelson [6] as a classical signal processing tool for image multi-scale representation. An image is decomposed into a hierarchy of band-pass filtered images (known as residuals) and low-frequency approximations. This allows for analysis, compression, and reconstruction with respect to multiple resolutions, thus making it an advantageous method used in conventional image-processing pipelines.

Inspired by its interpretability and efficiency, deep learning researchers began adapting the Laplacian pyramid into neural architectures. One of the most influential works in this direction is LapSRN [24], which extended the classical Laplacian pyramid into a trainable deep neural network for single image super-resolution. LapSRN learns to predict residuals at each scale level and progressively refines the reconstruction from coarse to fine. This coarse-to-fine learning strategy aligns well with the human visual perception system and brings several practical advantages. First, it enables effective multi-level supervision during training [27], allowing the network to focus on learning finer details incrementally. Second, by concentrating most computations in the lower-resolution stages, the model achieves better computational efficiency [46]. Third, its structural flexibility allows the network to be truncated or

expanded at different pyramid levels, providing a natural way to trade off between speed and accuracy [26]. Finally, the ability to generate multi-resolution outputs makes this architecture particularly suitable for deployment on resource-constrained platforms such as mobile and embedded systems [15, 21].

These properties make deep Laplacian pyramid networks not only effective but also highly adaptable for tasks that benefit from structural decomposition and progressive reconstruction.

In our work, we draw inspiration from this deep Laplacian architecture and extend it to the task of reflection removal, which—like super-resolution—benefits significantly from multi-scale processing. Rather than using the Laplacian pyramid for upscaling, we repurpose its hierarchical structure to guide the design of a two-branch model that performs reflection suppression and image reconstruction at progressively finer scales. Specifically, we propose a dual-branch framework aligned with the Laplacian philosophy. Our design demonstrates how principled architectural choices can yield models that strike a practical balance between performance, clarity, and real-world usability.

Chapter 3

Method

In this section, we present our proposed reflection removal method in detail. As illustrated in Figure 3.1, our framework is constructed based on a Deep Laplacian Pyramid architecture [24] that enables progressive multi-scale restoration. The overall model comprises two complementary branches: a frequency-domain decomposition branch based on Discrete Wavelet Transform and Fast Fourier Convolution (DWT-FFC) [38, 9], and a spatial-domain reconstruction branch (UHDM) featuring pixel unshuffling [46], Residual Dense Blocks (RDB) [70], and Scale Attention Modules (SAM) [60]. At each hierarchical level of the pyramid, these two branches work in parallel, and the extracted frequency-domain residuals are fused into the UHDM decoder via cross-scale attention to guide detail reconstruction. In Section 3.1, we first introduce the overall network architecture and its Laplacian-inspired design. Section 3.2 3.3 details the construction of the DWT-FFC and UHDM branches, along with the hierarchical fusion mechanism between them. Afterwords, section 3.4 depicts the cascaded design of the framework. Finally, Section 3.5 outlines the loss functions employed to supervise the reconstruction at multiple scales.

3.1 Overall Architecture

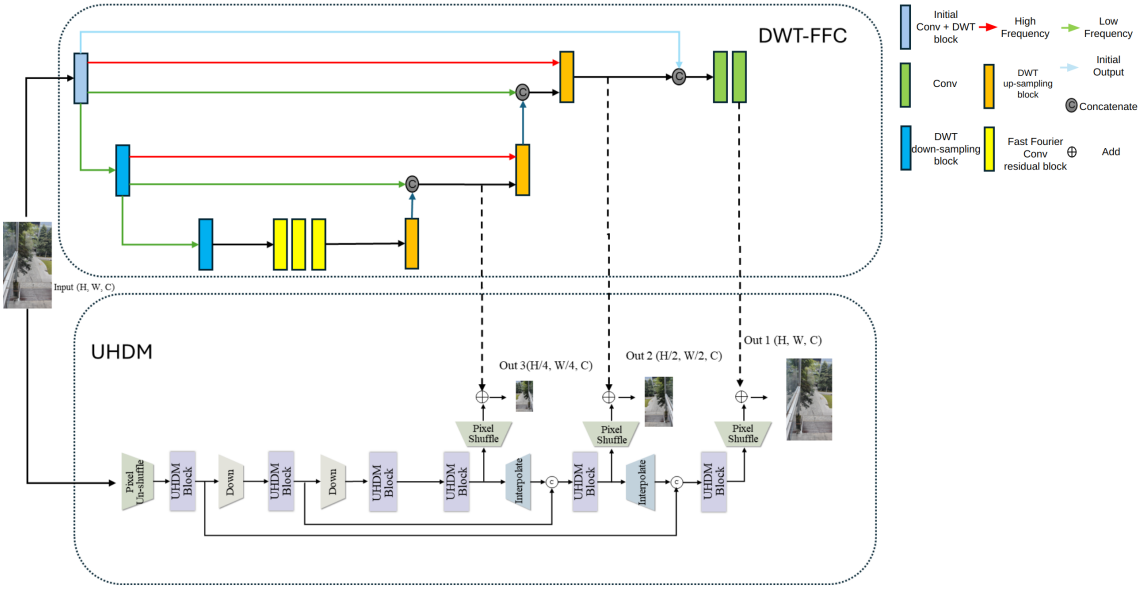

Figure 3.1: Overall Framework.

Our proposed reflection removal network, RefLap, is a dual-branch architecture that integrates classical insights from Laplacian pyramids into a modern deep learning framework. Inspired by the Deep Laplacian Pyramid Networks [24], which demonstrated the power of progressive residual learning and multi-scale decomposition for super-resolution tasks, we adopt a hierarchical encoder-decoder design. Our model leverages this framework not to super-resolve images, but to perform structural decomposition and feature fusion for the targeted task of reflection removal.

The architecture is composed of two coordinated branches: (1) a DWT-FFC branch for extracting frequency-domain features [38, 9], and (2) a UHDM branch for spatial reconstruction. As illustrated in Figure 3.1, the input image is simultaneously processed by both branches across multiple scales. The DWT-FFC branch

decomposes the input into hierarchical frequency components, capturing reflection cues and high-frequency artifacts. These features are then transferred to the UHDM branch, which reconstructs the background scene using adaptive attention-based fusion [60]. This modular design promotes interpretability, computational efficiency, and effective multi-scale interaction.

3.2 DWT-FFC Branch

Inspired by [75], we construct our DWT-FFC frequency branch as an encoder-decoder network to learn the feature mapping between reflection-contaminated and reflection-free images, leveraging dense skip connections at each feature scale. In addition to conventional convolutional operations, we introduce Discrete Wavelet Transform (DWT) for hierarchical feature decomposition. DWT enables the separation of low-frequency and high-frequency components, facilitating multi-scale feature learning. Low-frequency features are concatenated with convolutional outputs, while high-frequency components are forwarded to the upsampling module to enhance detail recovery. To further enrich the frequency-based representation, we embed Fast Fourier Convolution (FFC) residual blocks, as illustrated in Figure 3.1, which effectively fuse spatial and spectral cues to guide reflection suppression. However, using only the DWT-FFC frequency branch leads to suboptimal performance in challenging, non-homogeneous settings, largely due to data scarcity. Hence, we complement this with a second spatial branch (Section 3.3) that leverages pretrained modules for stronger prior learning.

3.2.1 Discrete Wavelet Transform (DWT)

2D DWT decomposes an input feature map using one low-pass filter (f_{LL}) and three high-pass filters (f_{LH}, f_{HL}, f_{HH}), all with fixed kernel parameters and stride 2 [38, 17, 31]. In our model, we adopt Haar wavelets [17], defined as:

$$f_{LL} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad f_{LH} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \quad f_{HL} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \quad f_{HH} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Applying these filters to the input feature map yields four sub-bands: x_{LL}, x_{LH}, x_{HL}, and x_{HH}. For instance, x_{LL} is computed as:

$$\begin{aligned} x_{LL}(i,j) = x(2i-1, 2j-1) + x(2i-1, 2j) \\ + x(2i, 2j-1) + x(2i, 2j) \end{aligned} \tag{3.2.1}$$

At each resolution level, we concatenate x_{LL} with the outputs from standard convolutional layers to jointly encode spatial and frequency-domain features [31, 10].

3.2.2 Fast Fourier Convolution (FFC)

Fast Fourier Convolution (FFC) enhances global context understanding by decomposing input features into local and global branches [9, 72]. The local branch employs two standard convolutional layers, while the global branch uses spectral transforms based on channel-wise 2D FFT operations. The spectral processing follows three key steps:

(a) Forward FFT and Conversion to Real Domain:

$$\text{Real FFT2D: } \mathbb{R}^{H\times W\times C} \to \mathbb{C}^{H\times \frac{W}{2}\times C}, \quad \text{ComplexToReal: } \mathbb{C}^{H\times \frac{W}{2}\times C} \to \mathbb{R}^{H\times \frac{W}{2}\times 2C}$$

(b) Feature Transformation in Frequency Domain:

$$\text{ReLU} \cup \text{BN} \cup \text{Conv}_{1\times 1} : \mathbb{R}^{H\times \frac{W}{2}\times 2C} \to \mathbb{R}^{H\times \frac{W}{2}\times 2C}$$

(c) Inverse FFT and Fusion:

$$\text{RealToComplex} : \mathbb{R}^{H\times \frac{W}{2}\times 2C} \to \mathbb{C}^{H\times \frac{W}{2}\times C}, \quad \text{Inverse Real FFT2D: } \mathbb{C}^{H\times \frac{W}{2}\times C} \to \mathbb{R}^{H\times W\times C}$$

The outputs from both branches are fused via concatenation and further processed by a 1×1 convolution. Two FFC units are stacked to form a residual FFC block, and we include three such residual blocks in our model to leverage global semantics and spatial precision for effective reflection removal.

3.3 UHDM Branch

To enhance spatial feature representation and address the limitations of frequency-domain methods alone, we integrate a strong spatial branch inspired by the Efficient and Scale-Robust Demoiréing Network (ESDNet) [65]. This branch, referred to as UHDM, consists of an encoder-decoder architecture with skip connections, equipped with dilated residual dense blocks (RDBs) and semantic-aligned scale-aware modules (SAMs) for multi-scale feature refinement. This design enables our model to effectively capture spatial patterns and adaptively suppress reflection artifacts of various

scales and complexities.

3.3.1 Residual Dense Blocks (RDBs)

Each level $i \in \{1, 2, 3\}$ in the encoder and decoder consists of a convolutional unit followed by a dilated residual dense block to refine feature representations. The RDB is built upon dense connectivity [70] and integrates dilated convolutions to expand the receptive field without increasing the number of parameters. Given input feature F_i^0 at level i, the l-th feature inside the block is computed as:

$$F_i^l = C^l([F_i^0, F_i^1, \ldots, F_i^{l-1}]), \quad l = 1, 2, \ldots, L, \tag{3.3.1}$$

where C^l denotes a 3×3 convolution with dilation rate d_l, followed by ReLU activation, and $[\cdot]$ is the concatenation operator. A 1×1 convolution is then used to match the original channel dimensions, and a residual connection yields the refined output:

$$F_i^r = F_i^0 + \mathrm{Conv}_{1\times1}(F_i^L). \tag{3.3.2}$$

3.3.2 Semantic-Aligned Scale-Aware Module (SAM)

To enhance scale adaptability, the Semantic-Aligned Module (SAM) extracts and dynamically fuses multi-scale features at the same semantic level. Given an input feature map $F^r \in \mathbb{R}^{H\times W\times C}$, we generate pyramid features via bilinear downsampling:

$$F_\downarrow^r = \mathrm{Down}(F^r), \quad F_{\downarrow\downarrow}^r = \mathrm{Down}(F_\downarrow^r). \tag{3.3.3}$$

Each scale is processed through shared convolutional blocks E_0, E_1, and E_2, producing Y_0, Y_1, and Y_2 respectively:

$$Y_0 = E_0(F^r), \quad Y_1 = \mathrm{Up}(E_1(F^r_{\downarrow})), \quad Y_2 = \mathrm{Up}(E_2(F^r_{\downarrow\downarrow})). \tag{3.3.4}$$

Here, Up denotes bilinear upsampling to match the spatial size of F^r.

To perform cross-scale dynamic fusion, we first compute global descriptors via global average pooling:

$$v_i = \frac{1}{H \times W} \sum_{s=1}^{H} \sum_{t=1}^{W} Y_i(s,t), \quad i = 0, 1, 2. \tag{3.3.5}$$

Concatenating v_0, v_1, and v_2, we estimate fusion weights through an MLP:

$$[w_0, w_1, w_2] = \mathrm{MLP}([v_0, v_1, v_2]). \tag{3.3.6}$$

The final output of SAM is computed by adaptive weighted fusion:

$$F^{\mathrm{out}} = F^r + w_0 \odot Y_0 + w_1 \odot Y_1 + w_2 \odot Y_2, \tag{3.3.7}$$

where $\odot$ denotes channel-wise multiplication. The result F^{out} is passed to the next decoder level for progressive image reconstruction.

In comparison to other multi-scale fusion designs [66, 71], SAM operates within the same semantic stage, ensuring semantic alignment across scales. This approach improves both efficiency and accuracy, acting as an implicit classifier without requiring manually defined scale attributes.

3.4 Cascaded Design with Cross-Branch Supervision

To enhance the connection between the two branches of our network, we adopt a cascaded design inspired by the Deep Laplacian Pyramid Networks [24] in which the frequency-domain DWT-FFC branch explicitly guides the spatial-domain UHDM branch. Rather than treating each sub-network as an independent processing stream, we connect them hierarchically, allowing information to flow from one to the other across each scale level. Specifically, at each resolution scale, the output features from the DWT-FFC branch are forwarded to the UHDM decoder branch via channel-wise concatenation. These features include both low-frequency and high-frequency residuals extracted through discrete wavelet transforms [38] and refined using Fast Fourier Convolution blocks [9]. When concatenated with the corresponding decoder features in UHDM, they serve as priors that help suppress residual reflections and enhance structural fidelity.

This cascaded setup introduces a form of deep supervision across branches: the DWT-FFC branch is not only optimized through its own reconstruction loss but also indirectly supervised via its impact on the UHDM branch's output. Each decoder stage in UHDM is responsible for progressively reconstructing the background image using its own local information, enriched with global spectral guidance from DWT-FFC. The fused features are processed through residual dense blocks [70] and semantic-aligned scale-aware modules [65], which selectively refine the signal using cross-scale attention.

Hierarchical alignment guarantees that the decoder in each stage is given frequency

priors at the right spatial resolution, remaining consistent with the decomposition of the Laplacian pyramid. This ensures efficient feature reuse for scale-aware restoration and promotes generalization to complex reflection patterns. Hence, the cascading of branches not only lends interpretability to the model but in fact improves performance while keeping the architecture largely simple.

3.5 Loss Function

To enhance the optimization process, we adopt a deep supervision strategy, which has been shown effective in prior work [67]. Our network produces hierarchical outputs $\hat{I}_1$, $\hat{I}_2$, and $\hat{I}_3$ at different decoder levels. Each of these predictions is directly supervised by the corresponding ground truth image, promoting convergence and improving gradient flow during training.

Given that moiré patterns often corrupt image structure by introducing unnatural stripe-like textures, we incorporate a feature-based perceptual loss [22] in addition to the standard pixel-wise loss. Specifically, our final loss function is composed of three terms: an ℓ_1 loss to encourage pixel-level fidelity, a perceptual loss $\mathcal{L}_p$ to enforce feature similarity in deep VGG space, and a structural similarity (SSIM) loss [56] to preserve overall image structure. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{3} \mathcal{L}_1(I_i, \hat{I}_i) + 0.001 \cdot \mathcal{L}_p(I_i, \hat{I}_i) + 0.4 \cdot \mathcal{L}_{ssim}(I_i, \hat{I}_i) \tag{3.5.1}$$

For the perceptual loss $\mathcal{L}_p$, we extract features from the `conv3_3` layer (after ReLU activation) of a pre-trained VGG16 network and compute the ℓ_1 distance between the predicted and ground truth features. We empirically set the perceptual loss weight

to 0.001 and the SSIM loss weight to 0.4 to balance its influence during training.

This hybrid loss formulation allows our model to better distinguish and suppress moiré artifacts while preserving both fine details and high-level structural cues across scales.

Chapter 4

Experiments

This project is part of the NTIRE2025 Single Image Reflection Removal in the Wild Challenge, in which the goal is to develop a method that will robustly and efficiently remove reflections in real-world images. Herein, we present a thorough evaluation of our proposed method for reflection removal. Section 4.1 introduces the NTIRE2025 Single Image Reflection Removal in the Wild Challenge, describing the evaluation protocol and real-world constraints of the benchmark. The train and test datasets used in all experiments are described in Section 4.2. Section 4.3 details the implementation settings, including optimizers, training schedules, and inference settings. Section 4.4 gives quantitative and qualitative evidence on the performance of our method. Section 4.5 describes ablation studies to show the effects of the individual components of the network.

4.1 Challenge Details

The Single Image Reflection Removal (SIRR) in the Wild Challenge is one of the official competitions associated with NTIRE 2025, held in conjunction with CVPR [63]. This challenge aims to advance the development of reflection removal methods that generalize well to real-world conditions by providing a benchmark composed of real-world images and ground-truth references, evaluated using both objective metrics (e.g., PSNR, SSIM [56]) and human-based subjective scores [5]. Participants are tasked with designing algorithms that take a single reflection-contaminated image as input and output a reflection-free version. To ensure fairness, all submitted methods must be reproducible, and each team is limited to one final submission. The goal is to promote robust, generalizable approaches and bridge the gap between academic research and industrial deployment in practical reflection removal applications.

4.2 Datasets

In this challenge, the dataset used for training and evaluation is the OpenRR-1k, a novel real-world benchmark specifically curated for the task of single image reflection removal [63, 8]. Unlike previous datasets that rely on artificial setups—such as removing glass or using black cloth to capture reflection-free images—OpenRR-1k employs an AI-assisted and human-refined data collection protocol to ensure high-quality and naturally aligned transmission-reflection pairs. Initially, reflection-free transmission images are generated using AI-based reflection removal tools built into OPPO smartphones [8]. These results are then refined using professional image editing software

like Photoshop [1] and MeituPic [61] to eliminate residual artifacts, producing visually clean ground-truths. The final dataset consists of 1,000 real-world image pairs, split into 800 training samples, 100 validation samples (OpenRR-1k_val), and 100 test samples (OpenRR-1k_test). The dataset captures diverse real-world conditions, with a rich distribution of image subjects—such as landscapes, animals, and transportation—and varying lighting conditions, including daytime, nighttime, and indoor scenarios. Compared to synthetic or constrained real-world datasets, OpenRR-1k offers a more practical and challenging benchmark for evaluating reflection removal methods in authentic environments.

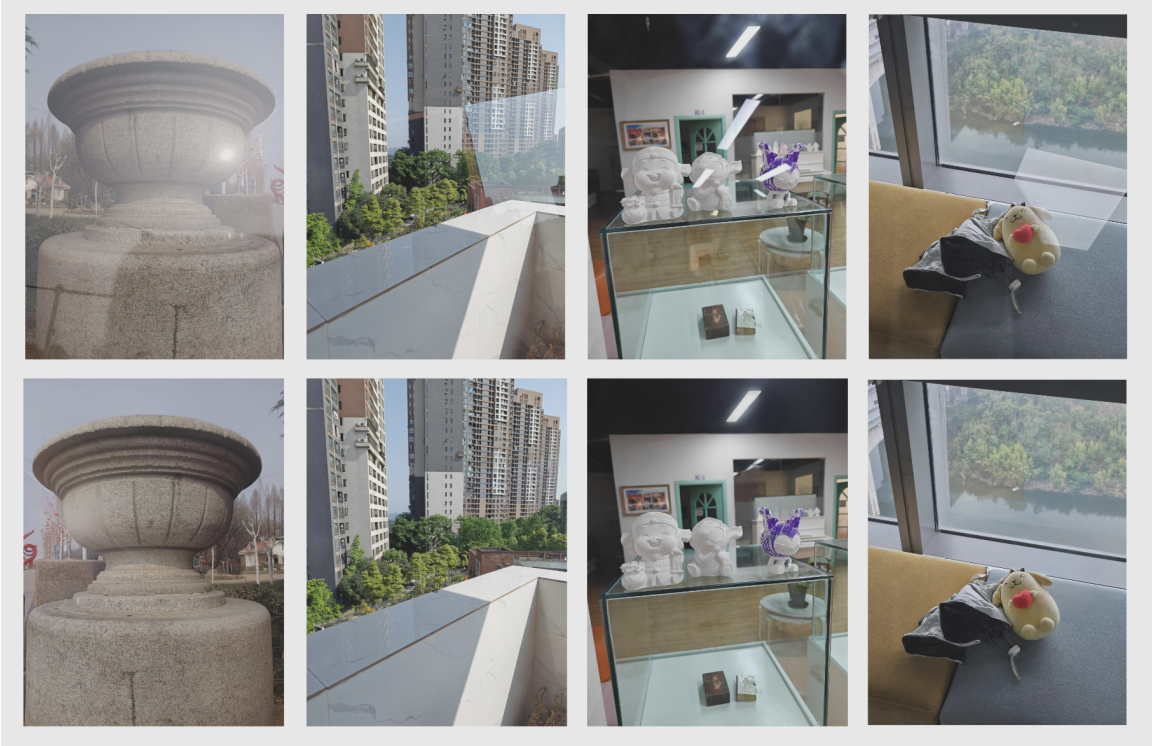

Figure 4.1: Example training pairs from the subset of the NTIRE 2025 SIRR challenge dataset.

4.3 Implementation Details

We implemented our reflection removal model using the PyTorch framework [43] and trained it exclusively on the official NTIRE25-SIRR dataset following the competition guidelines [63]. The dataset consists of 2,000 paired images for training and validation, with high-resolution inputs exhibiting diverse reflection artifacts captured in the wild. No external or synthetic datasets were introduced during training to ensure fair comparison under the challenge rules.

For training, we adopted a patch-based strategy where image patches of size 352×352 were randomly sampled and augmented with geometric transformations including horizontal flips and random rotations to improve generalization. Each training batch contained 12 image pairs. The network was trained for 150,000 iterations with a base learning rate of 2×10^{-4}, decayed following the cosine annealing schedule [37], decreasing to 1×10^{-6} over two full cycles for stable convergence. This scheduling strategy follows the optimization practice introduced in DWT-FFC [75].

Optimization was carried out using the Adam optimizer [23] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To stabilize training, we applied gradient clipping with a threshold of 0.1 [42] and enabled data prefetching to accelerate I/O during batch loading. To ensure reproducibility and adaptive checkpointing, we saved the model states at regular intervals and selected the best-performing checkpoint based on validation PSNR.

Our loss function consisted of three components: (1) the Charbonnier loss for pixel-level accuracy [76], (2) a perceptual loss computed from VGG-19 feature layers (weighted at 0.01) [22], and (3) a multi-scale structural similarity loss (MS-SSIM) weighted at 0.4 [55]. This composite loss function balances low-level fidelity and perceptual quality for effective reflection removal.

For testing, we followed the NTIRE25-SIRR challenge protocol by directly processing entire high-resolution images without cropping. We also employed a simple test-time augmentation (TTA) strategy involving horizontal and vertical flips as well as 90-degree rotations. The final output was obtained by averaging the outputs from all augmented variants.

All experiments were conducted on a single NVIDIA RTX 3090 GPU. The total model size is 12.2M parameters, and the inference time for a 352×352 image is approximately 0.1 seconds. Our implementation strikes a balance between structural design efficiency, fast inference, and perceptual quality.

Input	Training Time	Epochs	Extra data	Diffusion	Attention	Quantization	# Params. (M)	Runtime	GPU
(352, 352, 3)	25h	225	No	No	No	No	12.2	0.1s on GPU	RTX3090

Table 4.1: Training Configuration

4.4 Results

Figure 4.2: Sample Qualitative Results of the challenge dataset.

Our proposed method was evaluated on the NTIRE25 Single Image Reflection Removal (SIRR) in the Wild Challenge test set [63]. We report the PSNR and SSIM metrics—standard indicators of image restoration performance [56, 18]. Our model achieves a PSNR of 29.09 and an SSIM of 0.9493 on the benchmark dataset.

Team name	**PSNR ↑**	**SSIM↑**
X-Reflection	33.7606	0.9685
AIIA	32.4062	0.9611
Okkk	33.5411	0.9674
MVP Lab	33.3140	0.9682
KLETech-CEVI	31.7977	0.9601
ACVLab	32.6355	0.9662
i am a bug	32.4648	0.9603
Reflep(ours)	29.09	0.9493

Table 4.2: PSNR and SSIM performance of different teams

While it may not lead to absolute PSNR or SSIM values compared to some over-parameterized models, it strikes a strong balance between efficiency, structure, and visual quality. Built with only 12.2 million parameters, the model processes a 352×352 image patch in just 0.1 seconds on an RTX3090 GPU, making it highly suitable for real-time or resource-constrained deployments [40, 43].

Baseline Model	**Time (s)**	**#Params**
RDNet	0.44	243M
RAGNet	0.33	130M
DSRNet	0.36	137.6M
ERRNet	0.14	80M
Ours	**0.10**	**12.2M**

Table 4.3: Comparison of inference time and model size with other teams' baseline models

Qualitatively, our model demonstrates robust reflection suppression and perceptual sharpness across diverse input conditions. By leveraging the dual-branch DWT-FFC for frequency-domain decomposition [38, 9] and UHDM for progressive spatial reconstruction—the model effectively disentangles reflection layers while preserving natural details. Visual examples show that RefLap generates clean, visually appealing outputs even in challenging scenes with non-uniform reflections.

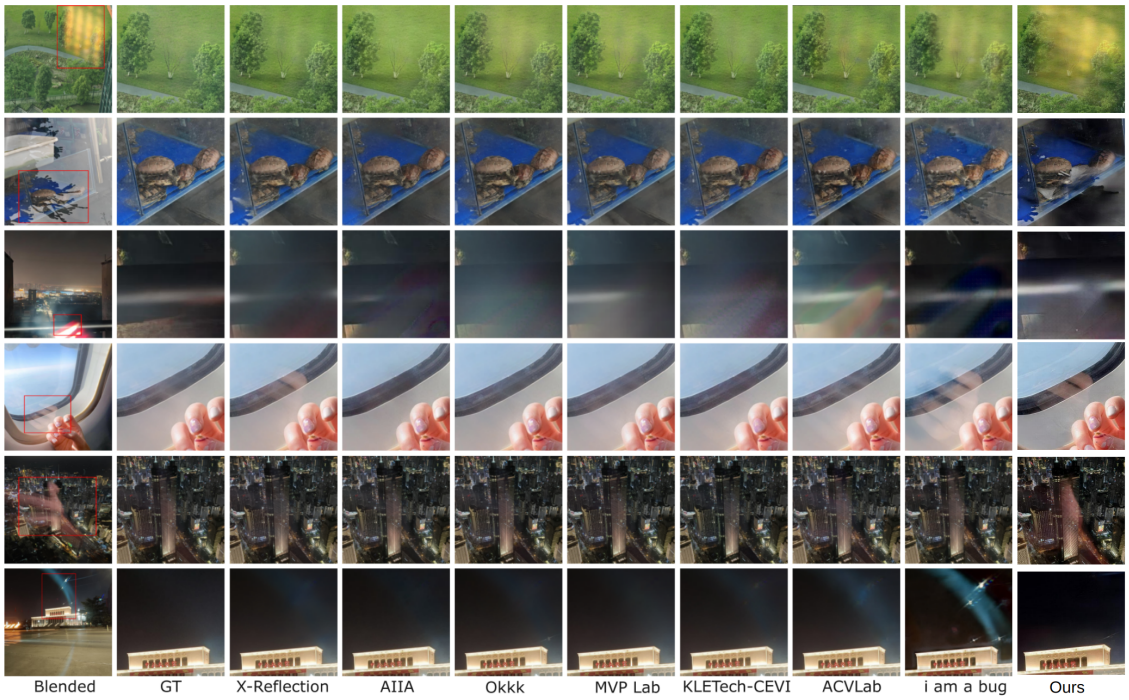

Figure 4.3: Qualitative results comparisons with other methods in the Challenge.

4.5 Ablation Study

To better understand the contribution of different components in our RefLap framework, we conduct ablation experiments on the NTIRE25-SIRR validation set [63]. We evaluate two key design choices: (1) the role of the multi-scale SSIM loss in guiding perceptual reconstruction [56, 73], and (2) the importance of the DWT-FFC branch in enhancing frequency-aware feature learning [38, 9]. The evaluation metrics include PSNR and SSIM, and all models are trained with the same hyperparameter settings for a fair comparison.

4.5.1 Effect of Removing SSIM Loss

To assess the contribution of the multi-scale SSIM (MS-SSIM) loss in training, we trained a variant of RefLap with only the Charbonnier and VGG perceptual loss, excluding the SSIM loss term. As shown in Table 4.2, removing the SSIM loss results in a noticeable drop in both PSNR and SSIM. Although the model still preserves global structures, it produces outputs with weaker local contrast and slightly more residual reflection, especially around textured regions.

These results suggest that the SSIM loss plays a crucial role in improving local structural consistency, especially in regions with complex reflections. Its removal leads to perceptual degradation, validating the necessity of incorporating MS-SSIM for high-quality reflection removal.

Table 4.4: Ablation study results (PSNR and SSIM).

Configuration	**PSNR ↑**	**SSIM ↑**
w/o SSIM loss	27.67	0.863
w/o DWTFFC	26.33	0.849
Full Model	**29.09**	**0.949**

4.5.2 Effect of Removing the DWT-FFC Branch

To evaluate the significance of the frequency-domain DWT-FFC branch, we train another variant of RefLap where the DWT-FFC branch is removed entirely. The remaining UHDM-only pipeline is still capable of generating reasonable outputs, but performance degrades both quantitatively and qualitatively. As summarized in table

4.2, this variant achieves lower PSNR and SSIM, particularly struggling with low-frequency ghosting artifacts and fine edge preservation.

Qualitative analysis further reveals that the absence of the DWT-FFC module limits the model's ability to distinguish reflection layers from high-frequency background details. This affirms that the fusion of frequency and spatial representations is essential for robust reflection removal, especially under challenging illumination and texture conditions.

Chapter 5

Future Improvements

While RefLap demonstrates a compelling balance between efficiency and performance, several avenues remain to further enhance its generalizability, robustness, and deployment readiness. We outline three key areas for future improvement:

5.1 Enhancing Real-World Generalization through Diverse Data

Our current training process relies entirely on the NTIRE25-SIRR dataset; this, despite being high in quality, contains mostly synthetic image pairs. Such synthetic reflections usually do not possess the complexity and diversity determined by real-world conditions, such as distortions by curved glass, uneven lighting, or layered reflections [52, 64]. For better generalization, future studies can work toward including more diverse datasets, particularly real-world reflection benchmarks [35, 68]. In addition, domain adaptation [49] or semi-supervised learning techniques [4] could

be employed to allow the model to learn more robust and invariant representations, enabling it to perform well on a greater variety of natural scenes of interest.

5.2 Layer Decomposition for Physically-Grounded Learning

Currently, RefLap reconstructs the clean image directly from the corrupted input without modeling the underlying image formation process. While this end-to-end learning has practical simplicity, it may limit the interpretability and precision in challenging cases. A future version of our framework could incorporate a reflection-background layer decomposition module — either via explicit alpha matte estimation [13] or by leveraging auxiliary supervision to disentangle reflection layers [58, 34]. Such physically-grounded modeling can help the network focus more precisely on reflection-specific regions, improving both restoration quality and visual consistency. Furthermore, coupling this with attention-based mechanisms [50] can help guide the decoder to focus more on highly reflective or semi-transparent areas.

5.3 Temporal Consistency and Video Extension

While RefLap is targeted at reflection removal for a single image, numerous real-world applications involve sequential frames with utmost importance being given to temporal coherence [20]. Accordingly, it is required to put on the video input the present method and consider frame-wise flickering and temporal consistency between outputs. This can be introduced by bringing about temporal attention modules [62], optical

flow-based feature warping [41], or recurrent architectures [54]. On the other hand, the reflection patterns vary gradually with time; modeling such temporal dynamics may also serve as a context for a steady and more precise reflection removal in video sequences.

Chapter 6

Conclusion

In this work, we presented RefLap, an efficient dual-branch reflection removal network that combines frequency-domain and spatial-domain cues within a Deep Laplacian Pyramid framework. While our method does not surpass the latest state-of-the-art approaches in terms of quantitative performance metrics, it demonstrates clear advantages in terms of model simplicity, computational efficiency, and inference speed. By integrating the DWT-FFC branch for frequency-aware feature extraction and the UHDM branch for fine-grained spatial reconstruction, our approach achieves a strong balance between performance and resource usage. These characteristics make RefLap a practical solution for real-world deployment scenarios where efficiency and responsiveness are critical. Future work will explore enhancing generalization, integrating explicit layer separation, and extending the model to handle video data with temporal consistency.

Bibliography

[1] Adobe Inc. Adobe photoshop, 2023. `https://www.adobe.com/products/photoshop.html`.

[2] Y. Aizu and A. Matsuoka. Reflection removal using multiple polarized images with different exposure times. *Applied Optics*, 61(10):2586–2593, 2022.

[3] Y. H. Ali and M. S. Mohsen. Remove reflection using wavelet transformation estimation. *Iraqi Journal of Science*, 59(1C):629–634, 2018.

[4] D. Berthelot et al. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 2019.

[5] Y. Blau and T. Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6228–6237, 2018.

[6] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

[7] J. Cai, K. Yang, L. Ouyang, L. Fu, J. Ding, H. Sun, C. M. Ho, and Z. Meng. F2t2-hit: A u-shaped fft transformer and hierarchical transformer for reflection removal. *arXiv preprint arXiv:2506.05489*, 2025.

[8] J. Cai, K. Yang, L. Ouyang, H. Sun, L. Fu, and Z. Meng. Openrr-1k: A scalable dataset for real-world reflection removal. In *arXiv preprint arXiv:2506.08299*, 2025.

[9] L. Chi, B. Jiang, and Y. Mu. Fast fourier convolution. In *NeurIPS*, 2020.

[10] e. a. Deshpande. Spectral wavelet dropout: Regularization in the wavelet domain. *arXiv preprint arXiv:2409.18951*, 2024.

[11] W. Dong, Y. Min, H. Zhou, and J. Chen. Sg-llie: Scale-aware low-light image enhancement via structure-guided transformer design. *arXiv preprint arXiv:2504.14075*, 2025.

[12] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[13] Q. Fan et al. Rethinking image matting: A simple baseline for high-accuracy matting. In *ECCV*, 2020.

[14] H. Farid and E. H. Adelson. Separating reflections and lighting using independent components analysis. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1:262–267, 1999.

[15] X. Fu, B. Liang, et al. Lightweight pyramid networks for image deraining. In *ACM International Conference on Multimedia*, 2018.

[16] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, 2002.

[17] A. Haar. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 69(3):331–371, 1910.

[18] A. Horé and D. Ziou. Image quality metrics: Psnr vs. ssim. In *20th International Conference on Pattern Recognition*, pages 2366–2369, 2010.

[19] J. Hu, Y. Wang, Q. Zhao, Y. Tai, L. Xia, C.-K. Tang, and J. Yang. How to train neural networks for flare removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/TPAMI.2021.3091737.

[20] J. Jang et al. Multi-frame reflection removal with consistency constraints. In *CVPR*, 2021.

[21] C. Jin, L.-J. Deng, and T.-Z. Huang. Lppn: Efficient deep pyramid network for pansharpening. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021.

[22] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 624–632, 2017.

[25] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 624–632, 2017.

[26] W.-S. Lai, J.-B. Huang, and Y. Zhang. Ms-lapsrn: Multi-scale deep super-resolution network for image upscaling. *IJCV*, 126(2-3):232–245, 2018.

[27] C.-Y. Lee, S. Xie, W. Li, et al. Training deep convolutional networks with deep supervision. In *ICML*, 2015.

[28] Y. Lei, B. Ren, Q. Wu, C. Xu, Q. Chen, and J. Jia. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1752–1760, 2020.

[29] A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007.

[30] C. Li, Y. Yang, K. He, S. Lin, and J. E. Hopcroft. Single image reflection removal through cascaded refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3565–3574, 2020.

[31] Q. Li, L. Shen, S. Guo, and Z. Lai. Wavecnet: Wavelet integrated cnns to suppress aliasing effect for noise-robust image classification. In *arXiv preprint arXiv:2107.13335*, 2021.

[32] X. Li, K. Xing, Q. Liu, and D. Chen. Single image reflection removal based on dark channel sparsity prior. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–1, 2023.

[33] Y. Li and M. S. Brown. Single image layer separation using relative smoothness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2759, 2014.

[34] Y. Li et al. Single image reflection removal through cascaded refinement. In *CVPR*, 2018.

[35] Y. Li et al. Single image reflection removal with real-world training data. In *CVPR*, 2020.

[36] J. Liu, Z. He, Z. Su, Y. Wang, and G. Yu. Fast fourier convolution for accelerating high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12361–12370, 2021.

[37] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[38] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11 (7):674–693, 1989.

[39] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations (ICLR)*, 2017.

[40] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz. Pruning convolutional neural networks for resource efficient inference. In *ICLR*, 2017.

[41] S. Niklaus et al. Context-aware synthesis for video frame interpolation. In *CVPR*, 2018.

[42] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *ICML*, 2013.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 2019.

[44] A. Popescu, J. Broekens, and M. van Someren. GAMYGDALA: An emotion engine for games. *Affective Computing, IEEE Transactions on*, 5(1):32–44, 2014.

[45] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention (MICCAI)*, pages 234–241. Springer, 2015.

[46] W. Shi, J. Caballero, F. Evans, and et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[47] Y.-H. Shih, D. Krishnan, F. Durand, and W. T. Freeman. Reflection removal using ghosting cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3193–3201, 2015.

[48] Y.-H. Shih, C.-K. Lin, and T. Lee. Reflection removal using ghosting cues. *Computer Vision and Image Understanding*, 137:1–12, 2015.

[49] E. Tzeng et al. Adversarial discriminative domain adaptation. In *CVPR*, 2017.

[50] A. Vaswani et al. Attention is all you need. In *NeurIPS*, 2017.

[51] R. Wan, B. Shi, L.-Y. Duan, A. H. Tan, and A. C. Kot. Benchmarking single-image reflection removal algorithms. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3922–3930, 2017.

[52] R. Wan et al. Single image reflection removal with perceptual losses. *IEEE Transactions on Image Processing*, 2021.

[53] A. Wang and collaborators. Single image reflection removal through multi-scale gradient refinement. In *Some Conference*, 2021.

[54] X. Wang et al. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, 2019.

[55] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. *The 37th Asilomar Conference on Signals, Systems Computers*, 2003.

[56] 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.

[57] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8170–8179, 2019.

[58] Y. Wei et al. Single image reflection separation with perceptual losses. In *CVPR*, 2019.

[59] L. B. Wolff. Polarization-based material classification from specular reflection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(11):1059–1071, 1990.

[60] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[61] Xiamen Meitu Technology Co., Ltd. Meitupic: Professional mobile photo editing app, 2023. `https://www.meitu.com/`.

[62] N. Xu et al. Temporal modulation network for temporal consistency of style transfer and video editing. In *ECCV*, 2019.

[63] K. Yang, J. Cai, L. Ouyang, F.-A. Vasluianu, R. Timofte, J. Ding, H. Sun, L. Fu, J. Li, C. Ho, Z. Meng, et al. Ntire 2025 challenge on single image reflection removal in the wild: Datasets, methods and results. In *CVPR Workshops*, pages 1301–1311, 2025.

[64] K. Yang et al. Diverse reflection removal dataset for robust generalization. *CVPR Workshops*, 2023.

[65] N. Yu, T. Li, Q. Yan, and C. C. Loy. Towards efficient and scale-robust demoiréing via frequency-aware decomposition and aggregation. In *CVPR*, pages 4282–4292, 2023.

[66] K. Zhang, W. Wang, W. Liu, R. Hu, and Y. Zhu. Mopnet: Motion pattern network for image deblurring. In *ECCV*, pages 614–631, 2020.

[67] L. Zhang, Y. Lin, and Q. Liu. Moiré pattern removal via adaptive feature decomposition. In *CVPR*, pages 12832–12841, 2021.

[68] L. Zhang et al. Ffcr-net: Frequency-aware reflection removal via dual-domain learning. *IEEE Transactions on Image Processing*, 2023.

[69] X. Zhang, R. Ng, and Q. Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4786–4794, 2018.

[70] Y. Zhang, Y. Tian, Y. Kong, and et al. Residual dense network for image super-resolution. *arXiv preprint arXiv:1802.08797*, 2018.

[71] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Residual non-local attention networks for image restoration. In *International Conference on Learning Representations (ICLR)*, 2019.

[72] e. a. Zhao. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image enhancement. In *CVPR*, 2024.

[73] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. In *IEEE Transactions on Computational Imaging*, volume 3, pages 47–57, 2016.

[74] Y. Zhong, K. Wang, L. Chen, Y. Xu, C. Li, S. Lin, and N. Yu. Language-guided image reflection separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1733–1743, 2024.

[75] H. Zhou, W. Dong, Y. Liu, and J. Chen. Breaking through the haze: An advanced non-homogeneous dehazing method based on fast fourier convolution and convnext. In *CVPR Workshops*, 2023.

[76] F. Ähnelt and Authors. Guided frequency loss for image restoration. *arXiv preprint arXiv:2309.15563*, 2023.