UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Low-level Image Generation: Deep Learning for Makeup Transfer and Unified Visual Restoration

by Jinliang Liu

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Yi Yang

University of Technology Sydney
Faculty of Engineering and Information Technology

September 2024

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Jinliang Liu*, declare that this thesis is submitted in fulfilment of the requirements for the award of *Doctor of Philosophy, in the Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note:
Signature removed prior to publication.

DATE: 14rd September, 2024

ACKNOWLEDGMENTS

First and foremost, I wish to express my sincere gratitude to my supervisor, Professor Yi Yang. At the outset of my research, Professor Yang provides clear guidance on the direction of my study and offers patient support when I face challenges. Each step of my progress is a testament to his invaluable mentorship and encouragement.

I extend my appreciation to Prof. Xiaojun Chang for his valuable insights and occasional guidance throughout my research journey, which contribute positively to the refinement of my work.

I sincerely thank my labmates, Ma Jian and Chen Mu. During times of difficulty, they consistently provided comfort and helped me generate suitable research ideas. When I encountered challenges in writing my academic papers, they patiently assisted in revising and refining them, significantly improving the quality of my work.

I am deeply grateful to my parents for their unwavering support throughout my academic journey. They never placed any pressure on me, allowing me to pursue my studies with freedom and confidence.

I express my heartfelt thanks to my girlfriend, Xiaomeng Tian, for standing by me through everything. During times when I faced challenges or setbacks in my research, she was always there to comfort and encourage me, helping me navigate through the difficulties.

Lastly, I am grateful to the University of Technology Sydney (UTS) for providing a supportive academic environment and the necessary resources that have been instrumental in my research.

ABSTRACT

With the continuous advancement of artificial intelligence, deep learning-based computer vision technology is making significant progress. As a result, more applications are being integrated into everyday scenarios. In this thesis, we focus on two specific tasks within the field of image generation: makeup transfer and image reconstruction. Both of them have strong practical value, but there are still limitations in the implementation. In makeup transfer, the difficulty of precisely capturing facial contours often leads to generated faces appearing overly smooth and lacking in realism. To address this, we incorporate 3D facial information to accurately preserve geometric features, thereby significantly enhancing the fidelity of the makeup transfer process. In the reconstruct task, models with high accuracy often struggle to maintain real-time inference speed, which limits its application scenarios. To tackle this issue, we select video deraining as a representative task and design a Transformer-based approach. Furthermore, we incorporate a memory bank as auxiliary information, enabling precise video deraining while maintaining high-speed inference and efficient reconstruction without increasing computational overhead. Moreover, most existing reconstruction strategies are designed to address only single degradation conditions, which often results in suboptimal performance when dealing with complex degradation scenarios in the real world. To solve this issue, we design a variety of solutions. First, we introduce diffusion models, which enhance generalization across diverse degradation scenarios through the progressive generation process. Second, we develop meta batch normalization inspired by meta-learning, using precision training for domain-specific parameters to enhance generalization. Additionally, we implement test-time adaptation to improve robustness under unknown weather conditions.

List of Publications

Related to the Thesis:

1. High Fidelity Makeup via 2D and 3D Identity Preservation Net (ACM Transactions on Multimedia Computing Communications and Applications, TOMM 2024)
2. Real-time Video Deraining Network with Hierarchical Memory Bank (Ready for submission)
3. Criss-cross Diffusion Models For All-in-One Blind Image Restoration (Under review)
4. Unified Adverse Weather Removal via Meta-learning and Domain-aware Normalization (Under review)

TABLE OF CONTENTS

List of Figures

Figure **Page**

List of Tables

CHAPTER 1

INTRODUCTION

Deep learning based style transfer technology continues to make significant advancements, with diverse applications in real-world scenarios [35, 43, 56, 64, 65, 67, 71]. Makeup transfer, as a representative study of style transfer enhances the shopping experience by allowing consumers to make more efficient purchasing decisions [7, 14, 26, 44, 61, 82, 162]. Moreover, the process of removing adverse weather conditions can be regarded as removing a "style" from an image, aiming to restore the "content" of the image, i.e., the clear scene. This approach mitigates the impact of adverse weather on autonomous driving systems, enhancing the safety of these systems and the accuracy of weather monitoring, thereby significantly reducing the probability of accidents [131, 141, 182]. These application scenarios illustrate the broad impact of image generation technology across different fields, further underscoring the importance and urgency of research in this area.

While advances in deep learning make it possible to achieve the functions mentioned above [20, 70, 80, 113, 146, 148, 184, 200], these methods still face several limitations in real-world scenarios. During makeup transfer, the inability to separate facial depth information tends to result in overly smooth and unrealistic images. For the reconstruction task, current methods struggle to balance accuracy and efficiency in image reconstruction, as these methods typically focus on a single type of degradation. Furthermore, due to the domain gap between synthetic data and real-world scenarios, these methods often exhibit unsatisfactory performance under complex weather conditions.

To enhance the practicality of deep learning driven style transfer technology in real-world scenarios, our study delves into two key applications. First, we propose an improved makeup

transfer method to address the issue of facial feature information loss in existing technologies, thereby providing consumers with a more realistic shopping experience. Second, our study focuses on the image reconstruction task under severe weather conditions, aiming to overcome the limitations of current methods in complex weather scenarios by enhancing reconstruction capabilities and achieving real-time inference speed without compromising accuracy. This improvement strengthens the model generalization and enhances the safety of autonomous driving systems in diverse climates.

To address the above limitations, our research focuses on the following four aspects:

- Makeup transfer involves digitally applying the makeup style from a reference face to a target face, while ensuring that the target's identity and facial features remain unchanged.
- Real-time adverse weather reconstruction: The process of restoring clear images from those degraded by severe weather conditions, ensuring both high accuracy and real-time performance.
- All-in-One reconstruction task: A comprehensive approach to restore images affected by multiple types of degradation within a single framework, optimizing for both accuracy and efficiency.
- Real-world multi-weather reconstruction model: A model designed to restore images under various weather conditions encountered in real-world scenarios, ensuring robustness and adaptability across different climates.

Based on generative models, our system incrementally integrates four key technologies to enhance the application of image reconstruction and style transfer in real-world scenarios. These technologies progress from individual subjects to environmental contexts, and from specific tasks to broader applications. First, the makeup transfer technology utilizes generative models to achieve facial style transfer, accurately generating personalized makeup effects while preserving the target person's facial features. Second, the real-time severe weather reconstruction technology effectively addresses complex and dynamic environmental conditions by generating clear images. Subsequently, a unified all-in-one reconstruction framework handles multiple types of degradation within a single generative framework, optimizing the accuracy and efficiency of image reconstruction. Finally, the real-world multi-weather reconstruction model further enhances generalization performance in changing environments by improving the system's adaptability and robustness under different climate conditions. This series of technologies demonstrates a progressive development from

individual style transformation to environmental adaptation, culminating in the unified reconstruction of multiple scenarios, ultimately achieving comprehensive adaptation to complex real-world conditions. We perform a comprehensive set of experiments to demonstrate the effectiveness and advantages of our proposed methods. The results consistently demonstrate improvements across a variety of challenging real-world scenarios. Compared to existing approaches, our system achieves higher performance in makeup transfer and adverse weather reconstruction tasks. These experiments confirm that our generative model-based techniques offer more robust and reliable solutions.

The methods developed in this research provide significant contributions to solving real-world challenges, particularly in enhancing personalized user experiences and improving environmental adaptability. Our makeup transfer technology offers a more realistic and interactive platform for consumers, which greatly benefits industries such as virtual retail and cosmetics. Additionally, our real-time weather reconstruction models enhance the performance and safety of autonomous systems by enabling them to operate effectively under diverse and complex weather conditions. These advancements not only demonstrate the practical applicability of our techniques but also pave the way for more intelligent, adaptive systems that seamlessly integrate into everyday life.

1.1 Makeup Style Transfer

Makeup transfer refers to the process of applying makeup from a reference image onto a source image, ensuring that the original identity and facial features of the person are preserved. This involves transferring makeup styles like lipstick, eyeshadow, and blush from one face to another[51, 93, 107, 108, 199]. However, existing methods face several challenges in real-world applications. For instance, when applying eye shadow from a reference image to a target face with dramatically different eye shapes (e.g., monolids vs. double eyelids), current methods often generate unrealistic results by directly copying the makeup pattern without adapting to the target's unique eye structure. Similarly, when transferring contour makeup between faces with different bone structures, the lack of depth perception in existing methods leads to inappropriate shading that doesn't align with the target's natural facial contours. Moreover, these methods often inadvertently modify background elements, such as altering hair color or clothing texture during the transfer process, which compromises the overall realism of the results.

In this study, we introduce IP23-Net (Chapter 3), a novel deep learning framework that enhances makeup transfer by preserving facial geometric information and distinguishing

between facial foreground and background. We achieve this through a 3D Shape Identity Encoder that incorporates depth and shadow information, ensuring the preservation of individual facial features and creating a realistic three-dimensional effect. For example, our method can accurately adapt eye makeup patterns to different eye shapes by considering the depth variations around the eye area, and properly apply contouring makeup based on the target's unique facial structure. Additionally, we introduce a Background Correction Decoder to prevent alterations to the background, resulting in more natural outcomes.

To address the lack of diverse makeup datasets, we introduce the High-Resolution Synthetic Makeup (HRSM) dataset, which comprises 335,230 diverse facial images, enabling comprehensive evaluation of model performance. These images are synthesized using advanced image generation technologies, ensuring the dataset's diversity while avoiding the use of real facial data, thus not infringing on any individual's privacy. Furthermore, the use of synthetic data allows us to expand and adjust the scale and complexity of the dataset without privacy restrictions, which is essential for advancing the development and testing of deep learning algorithms in the present study. Our method surpasses existing approaches in generating realistic and identity-preserving makeup transfer results.

1.2 Reconstruction Task Analysis

1.2.1 Video Deraining

Video deraining is a representative task in the field of reconstruction, which aims to remove rain streaks and artifacts from video content to improve visual clarity [92, 125, 129, 158, 195]. Video deraining faces a major challenge: although existing methods significantly improve reconstruction accuracy, the high complexity of the models makes it difficult to achieve real-time inference speeds. For example, current state-of-the-art methods like MPRNet achieve high-quality deraining results but require approximately 0.3 seconds to process a single frame at 1080p resolution, far from the real-time requirement of 30 frames per second (0.033 seconds per frame). Some methods attempt to improve speed by using separate modules for spatial and temporal feature extraction, but this approach leads to redundant computations. For instance, when processing consecutive frames with similar rain patterns, these methods still perform full spatial feature extraction for each frame independently, despite the high temporal correlation between frames. This limitation significantly reduces their effectiveness in practical applications, especially in scenarios like autonomous driving where real-time processing is crucial for safety.

To address these issues, we design a Real-time Video Deraining Network (RVDNet) with a spatio-temporal transformer architecture that integrates spatial and temporal information in a single model, eliminating the need for different components (Chapter 4). For example, when processing a video sequence of a car driving through rain, our unified architecture can simultaneously capture both the spatial distribution of rain streaks in the current frame and their temporal evolution across frames, reducing computational redundancy. Additionally, we introduce a Long Short-Term Memory Bank (LSMB) that aims to restore features from past frames as supporting information. A concrete example of LSMB's effectiveness is in scenes with periodic rain patterns: instead of processing each frame independently, our model can retrieve similar rain pattern features from previously processed frames, significantly reducing computational cost.

Furthermore, by incorporating learnable parameters, the model gains the flexibility to select features from either nearby or distant frames, allowing it to extract the most beneficial information for reconstructing the current frame. For instance, in a scenario where rain intensity suddenly changes, our model can adaptively weight the importance of temporal information: giving more weight to recent frames when rain patterns change rapidly, while utilizing information from more distant frames when rain patterns remain stable. This design avoids repeated feature extraction and ensures the accuracy of reconstruction without increasing the model inference burden. In our experiments, RVDNet achieves processing speeds of up to 30 frames per second on 1080p videos while maintaining competitive deraining quality, thus delivering a balanced approach that meets the demands of real-time applications while maintaining high reconstruction quality.

1.2.2 All-in-One Image Restoration

The All-in-One Image Restoration task aims to effectively repair various unknown types and degrees of image quality degradation [70, 73, 119]. The importance of this task stems from the fact that images in the real world are often affected by many different factors, such as noise, blur, haze, and raindrops, which can affect the clarity and quality of the image. For instance, in outdoor surveillance scenarios, a single image might simultaneously suffer from motion blur due to camera shake, noise from low-light conditions, and rain streaks from adverse weather. Traditional image restoration methods are typically optimized for specific types of degradation and struggle to address such complex scenarios. When applying these methods sequentially (e.g., denoising followed by deblurring), each step may introduce new artifacts or amplify existing ones. For example, applying a denoising model followed by a deblurring model might remove important texture details in the first step that cannot be

recovered in the second step, resulting in over-smoothed images. Therefore, it is crucial to develop a unified model that can handle multiple types of degradation simultaneously.

Currently, the primary challenge of All-in-One image restoration tasks is how to process and restore high-frequency information without losing image details. For example, when restoring an image containing both Gaussian noise and motion blur, existing methods often struggle to distinguish between high-frequency noise that should be removed and high-frequency texture details that should be preserved. This is particularly evident in areas with fine textures such as grass, fabric patterns, or hair, where these methods tend to either over-smooth the image (losing important texture details) or under-smooth it (leaving residual noise and artifacts).

Although existing diffusion models perform well in many tasks, they often struggle with processing high-frequency texture information, leading to the loss of image edge details and resulting in blurred and unrealistic effects. This limitation is particularly noticeable in scenarios involving complex textures, such as restoring images of buildings where architectural details become smudged, or in natural scenes where fine foliage details are lost. To address this issue, we propose the Criss-cross Diffusion model (CrDiff), which introduces a static wavelet transform operation to extract high-frequency information from the degraded image (Chapter 5). For instance, when processing an image of a brick wall with both noise and motion blur, our wavelet transform can effectively separate the high-frequency patterns of the brick texture from the degradation artifacts, allowing for more precise restoration.

This high-frequency information is then used to guide the diffusion model in reconstructing these textures within the latent space through a novel high-frequency encoder. In addition, to ensure that high-frequency information is accurately captured during training, we also add a high-frequency decoder to the model to effectively distinguish between the high-frequency noise introduced by degradation and the intrinsic high-frequency details of the image. For example, when restoring a noisy and blurry photograph of a fabric with intricate patterns, our model can simultaneously remove the noise while preserving the fine details of the fabric texture. This is achieved by our high-frequency encoder-decoder architecture, which learns to differentiate between the random high-frequency patterns of noise and the structured high-frequency patterns of the actual texture.

Finally, our model effectively overcomes high-frequency information loss and significantly enhances overall image restoration quality. In our experiments, CrDiff demonstrates superior performance across various challenging scenarios, such as restoring images with mixed degradation of motion blur and rain streaks while preserving fine details like text on signs or subtle facial features. The CrDiff model performs exceptionally in various degra-

dation tasks, achieving state-of-the-art results with improvements of up to 2.3dB in PSNR compared to existing methods.

1.2.3 Unified Real-world Adverse Weather Removal

CrDiff performs well on synthetic datasets but struggles when applied to real-world data. This challenge arises because the training data consists of synthetic paired data, which exhibits a significant domain gap from real-world scenarios, particularly under complex weather conditions. For instance, when processing real-world rainy scenes, synthetic-trained models often fail to handle the complex interplay between rain streaks and environmental lighting: while synthetic rain is typically modeled as semi-transparent streaks with uniform properties, real-world rain varies dramatically in size, shape, and transparency depending on factors like wind speed, lighting conditions, and camera parameters. Similarly, in foggy conditions, synthetic training data usually assumes uniform fog density, but real-world fog exhibits complex spatial variations influenced by terrain, temperature, and humidity. These domain gaps can lead to severe performance degradation - our experiments show that models trained on synthetic data can experience up to a 40% drop in PSNR when applied to real-world scenarios.

Therefore, the research on Unified Real-world Adverse Weather Removal aims to improve the reliability and generalization ability of outdoor vision systems under various complex weather conditions [20, 80, 114, 146, 172, 212]. For example, in autonomous driving applications, a vision system needs to maintain consistent performance whether encountering light drizzle, heavy rain with wind, or patchy fog - conditions that are difficult to accurately simulate in synthetic datasets. To address this problem, we propose an innovative dual-branch network structure that integrates self-supervised learning (SSL) with a meta-learning-based bi-level optimization approach, incorporating both inner and outer loops. The SSL branch learns from unpaired real-world weather images, extracting weather-specific features without requiring corresponding clear-weather images. For instance, when processing a rainy street scene, the SSL branch can learn the characteristic patterns of real rain streaks and their interactions with street lights, while the main branch focuses on the general image restoration task.

Additionally, by selectively updating the affine parameters of the Batch Normalization layer, statistical deviations are minimized, enhancing model stability and enabling effective weather adaptation (TT-WA) during inference (Chapter 6). This approach is particularly effective in handling rapid weather changes: for example, when a vehicle drives through alternating patches of heavy and light rain, or when transitioning from clear to foggy conditions,

our TT-WA mechanism can quickly adjust the model's behavior to maintain optimal performance. In our experiments, this adaptive approach reduces the performance gap between synthetic and real-world scenarios by up to 65% compared to non-adaptive methods.

Ultimately, the developed model not only performs exceptionally well under known weather conditions but also demonstrates robust generalization to unseen scenarios, thereby significantly improving video restoration quality. Our extensive testing in real-world environments shows consistent performance across diverse weather conditions - from light rain (improving visibility by up to 85%) to heavy fog (maintaining object detection accuracy above 90% at distances up to 50 meters), demonstrating the practical utility of our approach for safety-critical applications like autonomous driving and surveillance systems.

1.3 Thesis Organization

This thesis begins with the broad field of style transfer research, initially focusing on the makeup transfer task for facial images, and then gradually extending to the analysis of environment-related image reconstruction tasks. In subsequent research, the thesis starts with the design of a single adverse weather reconstruction network, gradually expands to an "All-in-One" reconstruction model with multiple degradation types, and finally explores the design of a network with adaptive capabilities to cope with unknown weather conditions. Through comprehensive investigation, this thesis aims to design a network model aligned with real-world applications, reflecting the advancements in deep learning and demonstrating practical value in real-world scenarios. This thesis is organized as follows:

Chapter 1: Introduction

This chapter delves into the research background, motivation, significance, and objectives of this study. It thoroughly presents the primary research questions, core contributions, and provides a clear overview of the thesis structure.

Chapter 2: Literature Survey

A comprehensive review of research encompassing makeup transfer, style transfer, 3D-aware image synthesis, single image deraining, video deraining, vision transformer, single degradation restoration, and multi-degradation restoration is presented. This review lays the theoretical and technical groundwork for the methodologies and experimental designs that will be examined in later chapters, and considers the integration and application of these technologies within the framework of this study.

Chapter 3: High Fidelity Makeup via 2D and 3D Identity Preservation Net

This chapter introduces IP23-Net, a novel framework designed to enhance makeup

transfer by preserving both facial geometry and background consistency. By leveraging a 3D Shape Identity Encoder and incorporating a 3D face reconstruction model, our method ensures realistic makeup application while maintaining the natural depth and appearance of facial features. Extensive experiments, including those on a newly introduced large-scale High Resolution Synthetic Makeup Dataset, demonstrate the high fidelity and generalization capability of our approach.

Chapter 4: Real-time Video Deraining Network with Hierarchical Memory Bank

In this chapter, we introduce the Real-time Video Deraining Network (RVDNet), a novel approach that employs a spatial-temporal transformer to seamlessly integrate spatial and temporal deraining processes. Unlike traditional CNN-based methods, RVDNet utilizes a Long Short-Term Memory Bank (LSMB) to effectively merge features from both immediate and historical frames, enhancing rain layer recognition and enabling faster inference.

Chapter 5: Criss-cross Diffusion Models For All-in-One Blind Image Restoration

In this chapter, we propose the Criss-cross Diffusion model (CrDiff) for the All-in-One image restoration task, addressing the challenges of reconstructing high-frequency textures in degraded images. By leveraging static wavelet transform operations, CrDiff extracts and preserves high-frequency information, guiding the diffusion model to accurately restore image details through a novel high-frequency encoder.

Chapter 6: Unified Adverse Weather Removal via Meta-learning and Domain-aware Normalization

In this chapter, we propose a dual-branch network with an innovative self-supervised learning (SSL) branch for unified weather removal in video processing. Our approach utilizes a meta-learning-based bi-level optimization strategy to enhance the alignment between auxiliary and reconstruction objectives, improving performance and enabling Test-time Weather Adaptation (TT-WA).

Chapter 7: Conclusion and Future Work

In the last chapter, we outline future research directions, aiming to develop a unified model capable of handling tasks like image reconstruction, makeup transfer, and image generation. We also plan to integrate text information to create a multi-modal system, allowing users to guide image generation with text.

CHAPTER 2

LITERATURE SURVEY

This survey explores the techniques, development trajectories and uniqueness of makeup transfer, style transfer, 3D-aware image synthesis, image deraining, video deraining, vision transformer and image restoration. The integration of these techniques provides an important foundation for creating a unified model that is capable of excelling in a variety of generation tasks, thus meeting a wide range of applications in different scenarios, especially in image reconstruction and multi-modal generation.

2.1 Makeup and 3D Vision

2.1.1 Makeup Transfer

Makeup transfer aims to transfer a specific makeup style from a reference image to a source image. As one of the pioneering works, CycleGAN [208] investigates image style translation between unpaired images, which can be directly applied to the field of makeup. However, CycleGAN is designed for transferring global style between two pre-defined domains; thus, users are unable to customize the makeup style. Similar to eyeglass removal [52], PairedCycleGAN [7] further refines the task of makeup. In this method, two translators are trained separately: one for facial makeup and another for facial makeup removal. Notably, the two translator structures are asymmetric. BeautyGAN [82] trains a semantic segmentation network for extracting masks from different facial regions, such as skin, eyes, and mouth, to calculate the makeup loss. Consequently, the color information of the makeup area can

be restored in the source image. Chen *et al.* [14] apply the Glow model to invert the latent vectors to the RGB source image with makeup. Local Adversarial Disentangling Network [44] leverages multiple overlapping discriminators based on facial feature points for dramatic makeup transfer. PSGAN [61] proposes the Attention Mechanism Module (AMM) to explicitly address the spatial misalignment problem. The AMM utilizes point-wise spatial attention to establish makeup-to-face correspondences, enabling precise transfer of makeup features from the reference face to corresponding regions on the target face. Moreover, PSGAN introduces an innovative makeup distillation strategy that separates makeup features into different components (e.g., eye shadow, lipstick, foundation), allowing for more granular control over the makeup transfer process. SCGAN [26] is inspired by StyleGAN [67] and maps makeup styles into an intermediate style space, rather than relying on linearly projected vectors. The research by Lyu *et al.* [102] details a new technique, 3DAM-GAN, that effectively shields facial images from unauthorized detection by recognition systems. This innovation not only obscures personal identity but also proves to be highly effective in evading diverse facial recognition algorithms. It also yielded satisfactory results in the task of makeup transfer. Although the aforementioned methods can achieve makeup transfer, they fail to restore the face depth information and background of the generated images. In contrast, we propose the IP23-Net model in Chapter 3 that can generate highly stereoscopic makeup images and maintain the non-makeup area.

2.1.2 Style Transfer

In recent years, style transfer emerges as a prominent research area. The underlying concept of image style transfer involves extracting content and style features separately from source and target images, and subsequently recombining these features to reconstruct the source image with the style of the target image. Certain methods [134, 149, 154, 205], aim to transfer a source image to various styles, such as plain and cartoon styles. Nonetheless, these methods are limited to transferring the source image to a detailed style.

2.1.3 3D-aware Image Synthesis

In prior research, numerous studies focus on multi-view and 3D perceptual networks to enhance the fidelity in the image generation task [31, 48, 55, 110, 115, 132, 209], which is closely related to the temporal consistency in video generation [140] and multi-round editing [183]. HoloGAN [109] learns 3D features from rigid-body transformations to control the pose of generated objects. Schwarz *et al.* introduce GRAF [130], a method for high-resolution 3D-aware image synthesis that trains the model solely from unposed 2D images.

PIGAN [6] achieves view-consistent, camera angle-controlled, high-quality image generation by employing neural representations with periodic activation functions and neural radiance field rendering. GIRAFFE [111] incorporates a compositional 3D scene representation into the generative model, improving the controllability of generated images, while Zhang *et al.* [197, 198] further consider the pose consistency.

2.2 Deraining with Vision Transformers

2.2.1 Single Image Deraining

Over the past few years, numerous deraining techniques have been developed to recover clean images by analyzing the statistical relationship between paired rainy and clean images [22, 53, 66, 85, 99, 138, 210]. Kang et al. [66] enhance the deraining effect by using a bilateral filter to separate the image into its high- and low-frequency components, which helps in removing the rain layer while retaining the original image details. Recognizing that rain streaks typically follow a narrow range of directions, Zhu et al. [210] proposed a joint optimization framework that incorporates three different priors: rain direction prior, sparse representation prior, and rain patch prior. Chen et al. [12] developed a low-rank model that captures the spatio-temporal correlations of rain streaks in video sequences. By representing rain streaks as a low-rank matrix, they successfully separated the rain components from clean video frames. Unlike other approaches, this method does not rely on pixel-level rain detection or dictionary learning. Fu et al. [38] employed a CNN to directly map the rain layer to the clean image's detail layer, using a moderately sized network to improve rain degradation. This method further includes a predicted detail layer that is integrated with a low-pass filtered base layer to produce a rain-free image. Building on this, Yang et al. [168] introduced a multi-task deep learning framework that combines rain streak detection and removal using features from a dilated contextual network. Similarly, Qian et al. [120] applied an attention mechanism within a deep neural network for rain detection, coupled with a GAN-based architecture to reconstruct realistic, clean images. While single-image deraining has seen considerable advancements, the absence of temporal information processing constrains the performance of these models when applied to video deraining tasks. As a result, achieving comparable performance for video deraining remains a significant challenge.

2.2.2 Video Deraining

Video deraining aims to recover clean background video sequences from rainy ones. Most deep learning based video deraining methods [4, 41, 128, 129, 176] adopt the strategy of temporal corrections among video sequence frames and several methods achieve satisfactory results for removing rain. The early methods employ a photometric-based approach for modeling rain streaks [92, 125, 129, 158, 195], and utilize learn-based models to address video deraining challenges. For example, Zhang *et al.* [196] propose that the combination of time attribute and chromaticity attribute can boost the rain removal effect of the network. Wei *et al.* [159] suggest encoding rain streaks randomly using Gaussian patch mixing, which enables the proposed model to better adapt to a wider range of rain variations.

Recently, many deep neural networks are proposed and bring obviously increase to the video deraining [17, 79, 90, 91, 165]. Liu *et al.* [90] propose a recurrent neural network for pixel-level rain classification, rain removal, and background detail reconstruction. Moreover, several researchers [166, 186] build a two-stage network to firstly capture spatial information and then obtain temporal information between frames to remove rain layer. Wang *et al.* [150] propose a rain streak motions concept to enforce a consistency of rain layers between video frames. Although the above deep learning approaches for video deraining with satisfactory results, the majority of them emphasize performance above computational time. In Chapter 4, we present a revolutionary end-to-end single video deraining model that can enhance performance with super high speed.

2.2.3 Vision Transformer

The Vision Transformer (ViT)[147] integrates principles from both Computer Vision (CV) and Natural Language Processing (NLP), achieving cutting-edge performance across various tasks [5, 29, 144, 152]. Drawing inspiration from the success of transformers in the NLP domain [30], ViT takes the transformer architecture and directly applies it to images with minimal adjustments to the image classification process, leading to significant improvements over CNN-based methods. This breakthrough has encouraged many researchers to explore the use of transformers in various areas, including segmentation [16, 137], object detection [98, 211], and depth estimation [193, 201], where transformers have demonstrated impressive results. Given the demands of real-time video deraining, the Swin Transformer [98] stands out as a strong candidate. With its "sliding window" mechanism, it effectively captures both local and global image features. This enables it to rapidly detect local elements like raindrops, while its hierarchical design ensures the overall clarity and structure of the video

are preserved. These characteristics make the Swin Transformer particularly well-suited for tasks requiring real-time performance.

2.3 Image Restoration

2.3.1 Single Degradation Restoration

Image restoration plays a crucial role in computer vision, aiming to restore degraded images back to their original high-quality form. This domain involves multiple techniques such as deraining, denoising, dehazing, and enhancing images taken in low-light conditions. Recently, with the increasing availability of paired image datasets, approaches utilizing Deep Neural Networks (DNNs) have made remarkable advancements in addressing various sub-tasks related to image restoration [1, 21, 142, 206]. These studies emphasize the development of network architectures and the crafting of loss functions, aiming to restore images by capturing the underlying relationship between corrupted images and their pristine counterparts. While several models [18, 87, 180] have demonstrated excellent performance through the use of specialized modules tailored for specific degradation issues, they often require training on datasets designed for specific types of degradation, which restricts their ability to generalize across different degradation scenarios.

2.3.2 Multi-degradation Restoration

There is a growing interest in the development of unified models that, once trained, can process a variety of degraded images within a single network framework. For example, Transweather[146] innovates within the transformer architecture by integrating a decoder equipped with learnable embeddings, designed to address multiple degradation types. Furthermore, the concept of AirNet[73] introduces the All-in-One image restoration task, which is adept at effectively restoring images from a diverse spectrum of unknown degradation types. Nevertheless, AirNet adopts a contrastive learning paradigm to accomplish this unified restoration task, requiring the training of an auxiliary encoder to differentiate among varied image degradation categories. In the latest research, PromptIR [119] employs a transformer-based foundation augmented with prompt blocks. These blocks initially generate modifiable prompt parameters, subsequently utilizing these prompts to navigate the model throughout the restoration process. Specifically, the prompt blocks are designed to learn task-specific degradation patterns through a set of learnable tokens, which are then used to guide the restoration process at different scales. The framework incorporates hierarchical prompt learning, where prompts at different levels capture both local and global

degradation features. However, the prompt block integration does not significantly improve the performance of the underlying network, as the learned prompts often fail to effectively capture the complex relationships between different types of image degradation. Therefore, our work in Chapter 5 focuses on utilising diffusion models to address their inherent limitations while exploiting their powerful generative capabilities.

2.3.3 Diffusion Models for Image Restoration

Diffusion models are gaining increasing attention as a class of generative models due to their ability to model complex data distributions through a progressive process of noise addition and removal. Among these models, the Denoising Diffusion Probabilistic Model (DDPM) [50] represents a key approach. DDPM leverages a Markovian forward process, where Gaussian noise is gradually added to the data, and a reverse process, which learns to remove this noise step by step. This iterative denoising mechanism allows DDPM to accurately reconstruct data distributions from noisy inputs, making it particularly effective for tasks such as image generation and restoration.

While diffusion models like DDPM have shown strong performance in data generation and image reconstruction tasks, WeatherDiff [113] introduces a patch-based diffusion model that applies a denoising strategy to overlapping patches during inference, effectively managing the distortions caused by adverse weather phenomena, including rain, snow, and fog. The patch-based approach enables the model to focus on local weather patterns while maintaining global image consistency through careful patch overlap design. By processing overlapping patches, WeatherDiff can better handle various scales of weather effects while reducing boundary artifacts that commonly occur in patch-based methods. Although diffusion-based methods generally outperform GAN-based approaches in terms of visual quality [164], they still encounter difficulties in restoring high-frequency details, as noted by [171]. This limitation arises primarily from the coarse-to-fine reconstruction strategy employed by diffusion models, which focuses on recovering overall structures at the expense of finer textures and high-frequency information. To overcome these shortcomings, we propose the CrDiff model, detailed in Chapter 5, which integrates a High-Frequency Enhancement Network. Incorporating high-frequency information into the latent space enhances the model's ability to restore fine details.

2.4 Meta-learning for Weather Removal

2.4.1 Unified Adverse Weather Removal

Over the last decade, low-level vision research has primarily focused on the removal of individual weather conditions, such as dehazing [33, 97, 127], deraining [151, 170, 177], and desnowing [19, 189]. These studies encompass both image and video processing, achieving commendable results. However, existing models often struggle to adapt to different weather conditions due to their specialized architectures. This limitation has motivated researchers to explore using a single model instance to handle multiple adverse weather conditions, aiming to reduce the need for multiple models and improve generalization capabilities. At the image level, All-in-One approach [80] defines a method for Adverse Weather Removal that utilizes a single encoder and decoder to address various weather scenarios. Following this setting, TransWeather [146] proposes a transformer-based end-to-end model that uses a single encoder and decoder to restore images degraded by various weather conditions, utilizing intra-patch transformer blocks and learnable weather type embeddings to enhance performance. The intra-patch transformer blocks enable the model to capture local correlations within each patch while maintaining computational efficiency. Furthermore, the weather type embeddings allow the model to adapt its restoration strategy based on different weather conditions, making it more versatile for real-world applications. In a recent work, WeatherDiffusion [114] introduces the concept of diffusion models into the Adverse Weather Removal task, proposing a patch-based image restoration algorithm that uses denoising diffusion probabilistic models for effective and size-agnostic restoration. Additionally, research has extended to the video level. Yang *et al.* [172] propose a video adverse-weather-component suppression network (ViWS-Net) to restore videos from various adverse weather conditions, addressing the lack of temporal information and effectively handling multiple weather types. Although these methods perform well in seen weather domains, they remain unsatisfactory when facing unseen weather conditions, such as unexpected combinations of weather elements. Addressing this limitation is the primary focus of our research.

2.4.2 Meta-learning

Meta-learning, often called "learning to learn," refers to the process of training models to quickly adapt to new tasks with limited data by utilizing knowledge from previously learned tasks [122, 139]. In recent studies, researchers focus on refining learning algorithms to im-

prove model generalization to unfamiliar environments and conditions. For example, Li *et al.* [77] introduce a meta-learning approach that simulates domain shifts during training by generating virtual testing environments in each mini-batch. This method enhances the model's ability to generalize effectively to novel domains. Shu *et al.* [133] develop a Domain-Augmented Meta-Learning framework, incorporating Dirichlet mixup and distilled soft-labels to improve generalization across unseen target domains. These methods maintain domain-specific information while enabling the model to generalize across different domains, leading to improved performance on tasks it has not encountered before.

In low-level image reconstruction, meta-learning demonstrates significant potential. Soh *et al.* [136] introduce a Meta-Transfer Learning approach for Zero-Shot Super-Resolution (MZSR), which minimizes inference time by discovering a generalizable initial parameter for internal learning, enabling the model to adapt to diverse image conditions with minimal gradient updates. Park *et al.* [116] present a meta-learning-based super-resolution (SR) method that supports rapid fine-tuning during the testing phase by extracting additional details from input images to boost the performance of conventional SR networks on benchmark datasets. More recently, Chi *et al.* [23] propose a novel self-supervised meta-auxiliary learning strategy for handling dynamic scene deblurring. This method integrates both external and internal learning mechanisms, leading to enhanced performance and faster adaptation during testing. Although meta-learning's application in image reconstruction is becoming more widespread, its full potential, particularly in adverse weather conditions, remains underexplored.

CHAPTER 3

HIGH FIDELITY MAKEUP VIA 2D AND 3D IDENTITY PRESERVATION NET

This chapter delves into the task of makeup transfer, which involves transferring makeup from a reference image to a source image while preserving facial geometry and maintaining background consistency. The significance of this task lies in its potential applications in virtual try-on systems, entertainment, and digital content creation. However, existing deep learning methods often overlook the geometric structure of the source image, leading to issues such as flattening of facial features and loss of individuality, as well as difficulties in distinguishing the face from the background. To address these challenges, we propose the High Fidelity Makeup via 2D and 3D Identity Preservation Network (IP23-Net), a novel framework that leverages both 2D and 3D facial geometry information to generate more realistic and consistent results. Our method introduces a 3D Shape Identity Encoder, which captures identity and 3D shape features to preserve the three-dimensional effect of the makeup. Additionally, we incorporate a Background Correction Decoder that predicts an adaptive mask to maintain background consistency, effectively distinguishing the foreground from the background. To rigorously evaluate the effectiveness of our approach, we not only utilize popular benchmarks but also introduce a new large-scale High Resolution Synthetic Makeup Dataset, containing 335,230 diverse high-resolution face images. Experimental results demonstrate that IP23-Net achieves high-fidelity makeup transfer while preserving both facial geometry and background consistency, setting a new standard in the field.

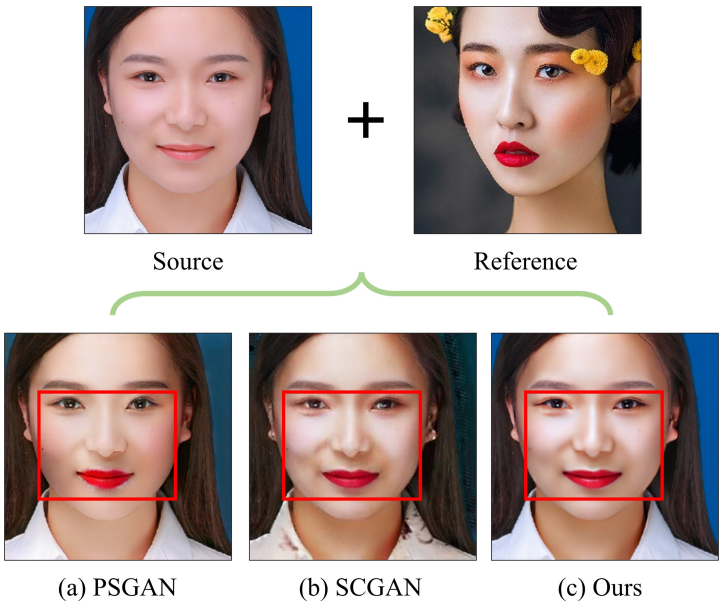


Figure 3.1: IP23-Net is capable of generating highly stereoscopic images while accurately restoring the background. We compare our output image to the ones produced by state-of-the-art existing methods, PSGAN [61] and SCGAN [26]. In addition to preserving the background, we also generate an output image featuring highly stereoscopic facial components, particularly in the mouth, nose, and eyes.

3.1 Introduction

In recent years, more people have started sharing their selfie photos on social networks like Instagram and Facebook. Good-looking selfie photos can increase attractiveness and self-confidence to people. Although cosmetics can enhance one's appearance, they also prolong the time spent on makeup application. In the last few years, deep learning techniques have gained rapid momentum and are widely applied in various domains [13, 32, 36, 36, 63, 78, 83, 84, 100, 173, 174, 204]. This study focuses on an automatic makeup model that transfers makeup from a reference image to a source image while preserving the original identity and makeup style. However, most existing methods [7, 14, 26, 44, 51, 57, 61, 62, 82, 93, 101, 108, 162, 199] face challenges in preserving facial geometric information, distinguishing between facial foreground and background, and addressing the limitations of current makeup datasets. To tackle above challenges, we introduce the 2D and 3D Identity Preservation Net (IP23-Net), an end-to-end learning framework comprising a 3D Shape Identity Encoder, a Makeup Style Encoder, and a Background Correction Decoder. As shown in Fig. 3.1, we can observe the above problems clearly.

First, previous works often lose facial geometric information during makeup transfer,

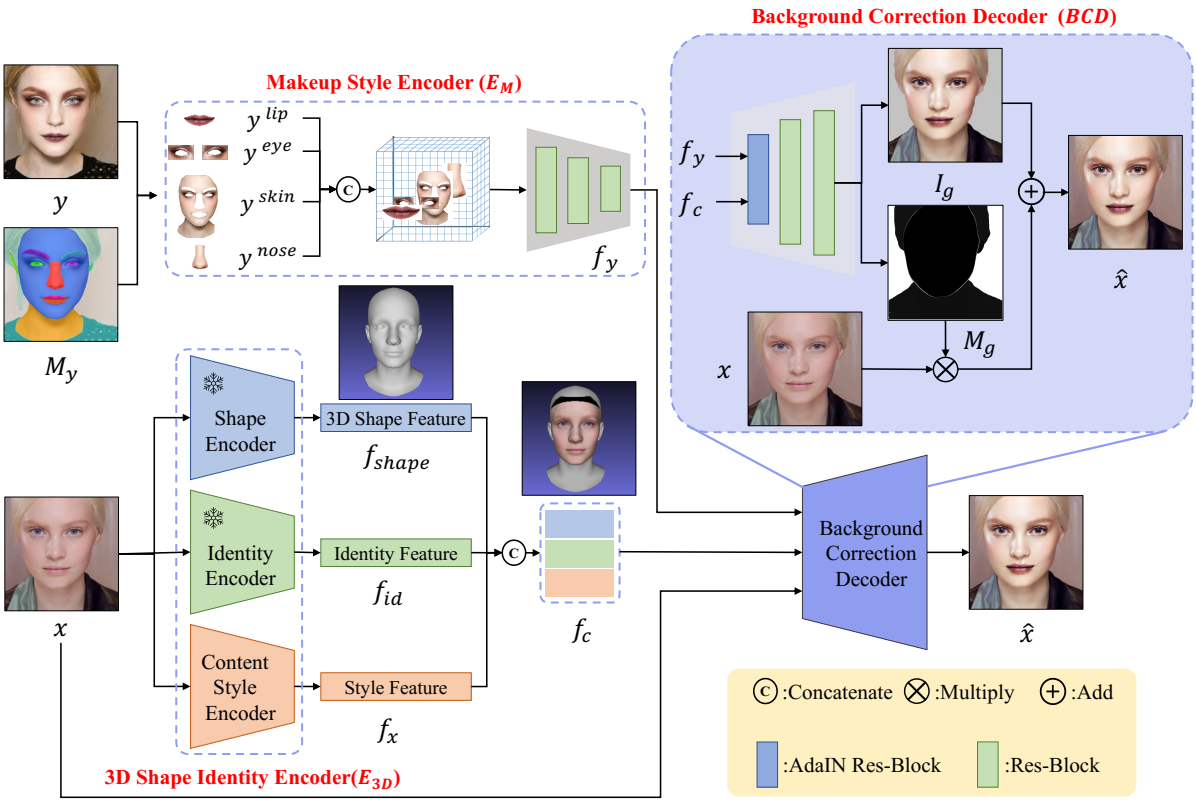


Figure 3.2: Illustration of our IP23-Net workflow. Our pipeline comprises three components: the 3D Shape Identity Encoder E_{3D}, Makeup Style Encoder E_M, and Background Correction Decoder BCD. The 3D Shape Identity Encoder is constructed from a 3D face reconstruction model [28], a face recognition network [27], and a Content Style Encoder. These elements extract features f_{shape}, f_{id}, and f_x from the source image x, forming f_c. The Makeup Style Encoder retrieves style features f_y from the reference image y. Finally, the Background Correction Decoder fuses these feature types to generate the source face with the reference makeup and restores the background.

resulting in artificial appearances and the loss of the individual's distinctive features. This issue stems from existing methods not incorporating depth information and shadow representation, both crucial for the face's three-dimensional aspect. To address this challenge, we design a 3D Shape Identity Encoder consisting of a shape encoder, identity encoder, and a content style encoder. The shape encoder, based on a 3D face reconstruction model, introduces face depth information to enhance facial contours and makeup shadow perception, achieving a visually three-dimensional effect. In the identity encoder, a face recognition network extracts facial identity information, maintaining the distinctiveness of individual facial features even after makeup application and ensuring facial uniqueness preservation. Lastly, the content style encoder extracts the face's style feature, supplementing detailed facial features. To augment the three-dimensional perception of the synthesized images, we

propose a 3D stereoscopic loss that utilizes the 3D average face parameter derived from the 3D Morphable Model (3DMM). During this process, 3D face prior features are independently extracted from both the source and generated images using the 3DMM. Following this, the pixel-wise L1 loss between the models is computed, effectively incorporating facial depth information into the generated image and achieving an enhanced stereoscopic effect. For makeup style extraction, our proposed Makeup Style Encoder obtains features from the reference image. In contrast to conventional encoders [61, 82], our approach shifts input focus from an entire image to specific facial regions based on the reference's corresponding mask. This enables IP23-Net to effectively perform makeup part selection and style blending.

Second, makeup transfer should differentiate between the facial foreground and background. However, most existing methods overlook this issue and modify background styles, leading to unrealistic artifacts. We gain inspiration from [155, 203]. Our Background Correction Decoder generates the output image by fusing features from the 3D Shape Identity Encoder and Makeup Style Encoder and predicts an adaptive face mask using the source image's 3D structure feature. This mask explicitly delineates the boundary between the face's foreground and background, maintaining end-to-end characteristics and ensuring a realistic outcome. Furthermore, the makeup transfer task lacks an effective validation dataset. The current state-of-the-art makeup MT dataset [82] comprises only 3,834 adult female facial images. As modern society evolves, the makeup user demographic has expanded from young women to various age groups and includes men. Consequently, makeup transfer techniques lack comprehensive evaluations for children, older adults, and male subjects. To address this issue, we introduce the High-Resolution Synthetic Makeup (HRSM) dataset, generated using StyleGAN2 [68], and comprising 335,230 facial images with diverse ages, poses, expressions, and backgrounds. This dataset enables a more comprehensive evaluation of model generalization and accuracy. Our IP23-Net method surpasses other SOTA techniques in generating realistic images while effectively addressing existing methods' challenges. The key contributions of this work are outlined as follows:

- We introduce a novel makeup network utilizing a 3D face reconstruction model for high-fidelity makeup transfer, facial geometric information extraction, and face mask prediction to restore the background. A 3D stereoscopic loss based on 3DMM enhances the realism of the generated makeup.

- We present the HRSM dataset, comprising 335,230 diverse face images, which serves as the largest makeup dataset available for evaluating the generalization and accuracy of automatic makeup transfer models.

- Our extensive experimental results demonstrate that IP23-Net delivers competitive results in preserving facial identity and ensuring makeup transfer quality.

3.2 Method

(1) Network Architecture

In this section, we introduce our IP23-Net, including Generator (G) and Discriminator (D). Given the source image x and the reference image y, our target is to learn a makeup transfer function $\hat{x} = G(x, y)$. Fig. 3.2 shows an overview of our network and the detail of Generator. Our pipeline consists of three parts: 3D Shape Identity Encoder (E_{3D}), Makeup Style Encoder (E_M) and Background Correction Decoder (BCD).

(2) 3D Shape Identity Encoder

The 3D Shape Identity Encoder (E_{3D}) is a vital component in our proposed makeup transfer framework, IP23-Net, designed to preserve the geometric structure and facial identity information of the source image. By incorporating 3D information, we aim to overcome the limitations of existing makeup transfer methods that often fail to maintain the fidelity of facial geometry and identity. The E_{3D} module is comprised of three key parts: a Shape Encoder, an Identity Encoder and a Content Style Encoder. To enhance the robustness of our model, we utilize pre-trained Shape Encoder and Identity Encoder with frozen parameters during network training. This approach allows us to leverage prior knowledge and optimize the network for specific tasks, without extensive training from scratch.

Shape Encoder. The Shape Encoder is designed based on a SOTA 3D Morphable Model (3DMM) [28] that extracts the facial geometry information from the source image. The facial geometry information consists of shape, texture, and depth information from the source image x. Depth information is of particular importance, as it provides rich geometric information such as contours, shapes, and local details. We define the facial geometry information extracted from the source image x as f_{shape}.

Identity Encoder. The process of 3DMM for facial modeling typically involves fitting the input face to the 3D average face model in the Basel Face Model (BFM) database [117] to predict the face model of the input image. However, since the face texture information of the predicted 3D face model is also sourced from BFM, the original details of the input image cannot be fully accurate. To address this issue and retain the identity information of the source image, we propose the use of a face recognition network [27] to extract the face identity feature f_{id} from the source image x. By using a face recognition network, we can

better capture the individual differences and details in the input image, ensuring that the generated face model is consistent with the input image.

Content Style Encoder. For the makeup transfer, capturing not only the shape and identity features but also the style information of the face such as skin color and lip color is essential for generating realistic and natural-looking images. To supplement this information and incorporate it into the generated images, we introduce the Content Style Encoder. It is designed to extract and encode the style information f_x from the source image x. The Content Style Encoder is comprised of multiple 7 res-blocks and 5 downsampling convolutional layers, which enable the capture of style information from the source image.

To generate the final makeup transferred image, we concatenate f_{shape}, f_{id}, and f_x and feed them into the decoder. However, since f_{shape} and f_{id} are obtained using pre-trained models, their feature shapes cannot match that of f_s. To ensure consistency, we adjust the feature shapes of f_{shape} and f_{id} to match that of f_x using a 1x1 convolutional layer to adjust the number of channels, and a transposed convolutional layer to adjust the spatial resolution of the features. The final output of the 3D Shape Identity Encoder is denoted as fc, and the function is defined as:

$$fc = Concat(Conv_{shape}(f_{shape}), Conv_{id}(f_{id}), f_x). \tag{3.1}$$

(3) Makeup Style Encoder

The Makeup Style Encoder is designed to extract features from the reference image using an encoder-bottleneck architecture. While most existing methods obtain style codes by simply averaging facial features in a reference image, numerous makeup styles exist for various facial components, inevitably leading to entanglements between different facial makeups. In contrast to traditional approaches [61, 82], our Makeup Style Encoder adopts the SCGAN [26] strategy.

The first processing step of the Makeup Style Encoder decomposes each reference face into four parts (lips, skin, eyes, and nose) using the face parser [72] as follows:

$$y^i = y \odot S_{y,i}. \tag{3.2}$$

We process each component of the reference face as y^i, where $i = \{lip, skin, eyes, nose\}$. $S_{y,i}$ represents the corresponding mask of the reference image, and $\odot$ denotes the Hadamard product. Then, we concatenate these codes as input $Y = [y^{lip}, y^{skin}, y^{eyes}, y^{nose}]$ for E_M. The Makeup Style Encoder consists of multiple 7 res-blocks and 5 downsampling convolutional layers, facilitating the extraction and encoding of style information from the reference image. Ultimately, we acquire the localized makeup style features, represented as f_y.

(4) Background Correction Decoder

Our Background Correction Decoder reconstructs the source image with makeup, using the content feature f_c and the style feature f_y. Specifically, our decoder comprises eight res-blocks and five upsampling convolutional layers. The Adaptive Instance Normalization (AdaIN) layer [56] achieves style transfer by altering the data distribution of the feature map. We incorporate AdaIN layers and two upsampling convolutional layers in the first five res-blocks to facilitate makeup transfer. The AdaIN layer is defined as follows:

$$AdaIN(f_c, f_y) = \sigma(f_y)\frac{f_c - \mu(f_c)}{\sigma(f_c)} + \mu(f_y), \tag{3.3}$$

where $\sigma(f_c)$ and $\mu(f_c)$ denote the mean and standard deviation of the source image features, respectively. $\sigma(f_y)$ and $\mu(f_y)$ represent the mean and standard deviation of the style image features, respectively. In general, AdaIN performs makeup transfer by transforming the mean and variance of specific channels in the feature map. We then employ another three res-blocks and upsampling convolutional layers to further fuse the feature maps and generate higher-resolution results. Note that these three res-blocks are all standard residual blocks without AdaIN layers. We incorporate the source image x as a residual to the output of the final residual blocks, which enhances the image restoration effect. Our approach is to ensure a balance between style transfer and content preservation. While AdaIN layers excel in transferring makeup style, using them excessively might compromise content information. To avoid this, we limit the number of AdaIN layers. Res-blocks without AdaIN layers play a significant role in refining and blending features after the primary makeup transfer process.

Additionally, the background information of the generated image is susceptible to damage by the reference image style during makeup transfer. To better preserve the background, we use the decoder to learn an adaptive mask and accommodate the change in face shape. Specifically, we predict a 3-channel I_g and 1-channel M_g after the upsampling. The M_g helps us identify the facial foreground and background areas. We apply the generated face I_g and blend the background of the clean source image x to create the makeup image using M_g, formulated as:

$$\hat{x} = I_g \odot M_g + (1 - M_g) \odot x, \tag{3.4}$$

where $\hat{x}$ denotes the generated makeup image. Despite changes in face shape, this decoder remains valid because the predicted mask M_g is deformable to follow the source image.

3.2.1 Optimization Objectives

(1) 3D Stereoscopic Loss

To generate highly stereoscopic images, we introduce a 3D face reconstruction model for obtaining facial depth information. Specifically, we use a pre-trained, SOTA 3D face reconstruction model (3DMM [28]) to accurately extract facial geometric information. First, we employ the regressing 3DMM coefficients model [28] to obtain the 3D face model and calculate the L1 loss:

$$L^{3D}_{attribute} = ||f^{x}_{shape} - f^{\hat{x}}_{shape}||_1 + ||f^{y}_{shape} - f^{\hat{y}}_{shape}||_1, \tag{3.5}$$

where f^{x}_{shape}, $f^{\hat{x}}_{shape}$ denote the 3D face attributes of the source image and generated makeup image, and f^{y}_{shape}, $f^{\hat{y}}_{shape}$ represent the 3D face attributes of the reference image and generated non-makeup image. Note that $\|\cdot\|_1$ is the L1-norm loss function. Second, we project the 3D facial vertices onto the 2D image to obtain landmarks. We then calculate the L1 loss between these landmarks to enhance the effect of the 3D stereoscopic loss:

$$L^{3D}_{landmark} = ||k^{x} - k^{\hat{x}}||_1 + ||k^{y} - k^{\hat{y}}||_1, \tag{3.6}$$

where k^{x}, $k^{\hat{x}}$ denote the 2D landmarks of the source image and generated makeup image, while k^{y}, $k^{\hat{y}}$ represent the 2D landmarks of the reference image and generated non-makeup image. Finally, our 3D stereoscopic loss is formulated as:

$$L_{3D} = \lambda^{3D}_{attribute} L^{3D}_{attribute} + \lambda^{3D}_{landmark} L^{3D}_{landmark}, \tag{3.7}$$

where $\lambda^{3D}_{attribute} = 2$ and $\lambda^{3D}_{landmark} = 0.2$. The weights $\lambda^{3D}_{attribute}$ and $\lambda^{3D}_{landmark}$ are determined through extensive ablation studies examining values in ranges [0.5, 4] and [0.1, 1] respectively. Through empirical evaluation, we find that $\lambda^{3D}_{attribute} = 2$ provides the optimal balance between preserving facial geometric structure and allowing flexible makeup transfer, while $\lambda^{3D}_{landmark} = 0.2$ achieves the best trade-off between landmark accuracy and generation flexibility. Lower values of these parameters lead to insufficient structural guidance, while higher values over-constrain the makeup transfer process and result in rigid transformations.

(2) Total Loss

In addition, we adopt the adversarial loss proposed by [58], denoted as L_{adv}, to enhance the realism of the generated images by ensuring they are indistinguishable from real ones through adversarial training. Additionally, we incorporate the perceptual loss introduced by [65], represented as L_{per}, to preserve the identity and high-level semantic features of the source images during the generation process. To maintain consistency in image content during bidirectional transformations, we employ the cycle consistency loss from [208], defined as L_{cyc}, which encourages accurate reconstruction of the original images after

(a) Makeup (b) Female (c) Male (d) Child (e) Elderly

Figure 3.3: Samples from the proposed High Resolution Synthetic Makeup (HRSM) dataset. The images displayed from (a) to (e), depict makeup faces, females, males, children, and elderly individuals.

translation. Finally, we utilize the makeup loss from [82], referred to as L_{HM}, to ensure that the makeup style in the generated images closely matches the reference images, particularly in terms of color distribution across key facial regions. The total loss L_{total} is:

$$L_{total} = L_{3D} + L_{adv} + L_{per} + L_{cyc} + L_{HM}. \tag{3.8}$$

3.2.2 High Resolution Synthetic Makeup Dataset

We observe that the sample number of datasets YMU [10], VMU [25], MIW [9], and MIFS [11] is less than 1,000. (see Table 3.1). The MT dataset [82] expands its size to 3,834 images, including 1,115 non-makeup images and 2,719 makeup images. Most existing approaches train and test models on the MT datasets. However, it only includes adult female samples with a relatively low resolution (361 × 361). As a result, the MT dataset lacks diverse samples (*e.g.*, males and kids) to validate the generalizability of the model. To overcome the data limitation, we propose a High Resolution Synthetic Makeup dataset. First, we deploy StyleGAN2 [68] model to generate 500,000 face images with various races, poses, expressions, and backgrounds, and the resolution is 1024 × 1024. Second, we align the generated images

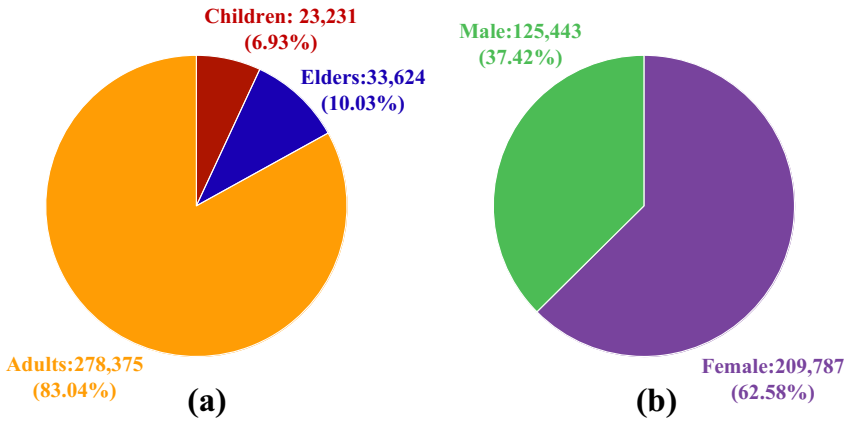


Figure 3.4: Attribute distribution of the proposed HRSM dataset. (a) Percentage of children, adults, and elders. (b) Percentage of male and female.

with 68 landmarks. Last, we obtain the gender and age labels through OpenCV. We utilize the MT dataset [82] to train a binary classifier to distinguish makeup images and non-makeup images. In this process, we use different pre-trained models such as VGG-19 [135], ResNet-50 [47] and Swin Transformer [98]. Then, we ensemble the classification results of the three models together to increase the accuracy of the labels. Most of the images are classified correctly. To further improve the label accuracy in the HRSM dataset, we set a confidence interval scope to remove images with low confidence. As a result, we obtained 225,831 makeup images and 109,399 non-makeup images for a total of 335,230 images. Fig. 3.4 demonstrates the proportion and number of each label in our dataset.

Dataset	ID number	Total Number	Age Label	Gender Label
YMU [10]	151	604		
VMU [25]	51	204		
MIW [9]	125	154		
MIFS [11]	214	642		
MT [82]	3,000+	3,834		
Ours	**330,000+**	**335,230**	✓	✓

Table 3.1: Comparison with other public makeup datasets.

The HRSM dataset is the first makeup transfer dataset with 1024 × 1024 resolution and contains 335,230 samples. Before integration into our framework, we pre-process these high-resolution images by first applying face alignment based on detected landmarks to ensure consistent facial orientation. We then normalize the pixel values to [-1, 1] range and convert all images to RGB format. The advantages of our collection method are scalability and

security. Scalability means our dataset can be expanded conveniently. Specifically, we can harness StyleGAN2 [68] model to generate more face images, and then use our classifier to distinguish the makeup images and non-makeup images. For the advantage of security, the face images we generated by StyleGAN2 [68] are not real-world images. Therefore, people using the HRSM dataset are unable to violate the privacy of other people. The detailed comparison between makeup datasets is in Table 3.1 and the examples of the HRSM dataset are illustrated in Fig. 3.3.

3.3 Experiment

3.3.1 Qualitative Comparisons

We compare our IP23-Net with other image-to-image makeup translation methods, including DIA [89], CycleGAN [208], PairedCycleGAN [7], BeautyGlow [14], LADN [44], BeautyGAN [82], PSGAN [61] and SCGAN [26] as well as two recent makeup transfer methods CPM [107] and EleGAN [163]. Fig. 3.5 demonstrates the qualitative comparison of IP23-Net with other methods on the generated makeup images. The results produced by DIA [89] show the abnormal color of the hair, and the color of the lipstick is not transferred well. In addition, the background information is affected by the reference image. Although CycleGAN [208] can be leveraged to transfer makeup, the results are incomplete. BeautyGAN [82] performs well on the makeup transfer. Nevertheless, we can find unnatural color changes in the hair part between the generated and source images. Besides, BeautyGlow [14] transfers the makeup to the source images with an uneven color distribution in the lipstick part. We use the released model from LADN [44] to obtain the generated makeup images. The background of the generated images is blurred, and some uneven color blocks appear on the face. PSGAN [61] produces fewer artifacts in the generated images. However, there is a problem of inaccurate color restoration. Specifically, the color of the lip between the generated and reference images is inconsistent. SCGAN [26] can address the spatial misalignment problem. For the CPM [107], we note that although it is able to achieve more satisfactory makeup transfer in specific scenarios, it may be slightly limited when dealing with diverse makeup style transitions. As for EleGAN [163], it performs well in localised makeup editing, but there is still room for further improvement in the naturalness and detailing of full-face makeup. The above limitations can be summarised in three points. First, the shadow information is lost in the generated image, resulting in an unrealistic nose. Second, the texture information is not preserved well. For example, the size of the lip in the generated image is smaller than

the source image. Third, the background and the hair color of the generated images are changed. Compared to these methods above, our IP23-Net can generate a high-fidelity face and preserve the background accurately.

Figure 3.5: Qualitative comparison with existing models. Both source and reference images are selected from the MT dataset for a fair comparison. IP23-Net effectively transfers the makeup style from the reference image to the source image. In addition, our method preserves the background as well as the original identity.

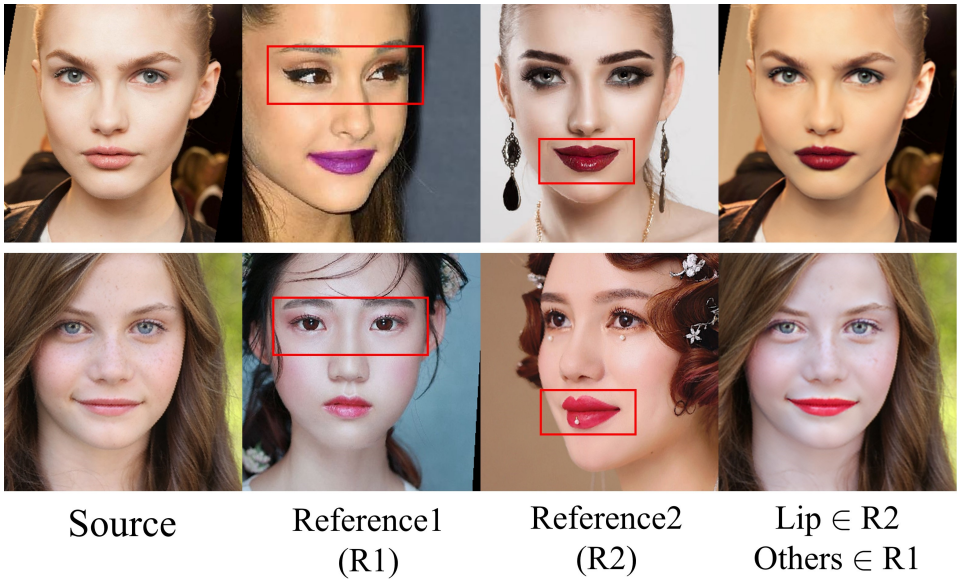


Figure 3.6: Results of partial makeup transfer. The lipstick style for the generated image is from Reference2, and the other makeup styles are from Reference1. The source and reference images are from the MT dataset.

We evaluate the realism of images generated by IP23-Net using the Fréchet Inception Distance (FID) [49], comparing it against BeautyGAN [82], PSGAN [61], and SCGAN [26] under uniform conditions. We select makeup and non-makeup images from the MT dataset [82], transforming non-makeup images into a consistent makeup style. Different reference images are used to calculate the average FID. IP23-Net achieves the lowest FID score of 45.59,

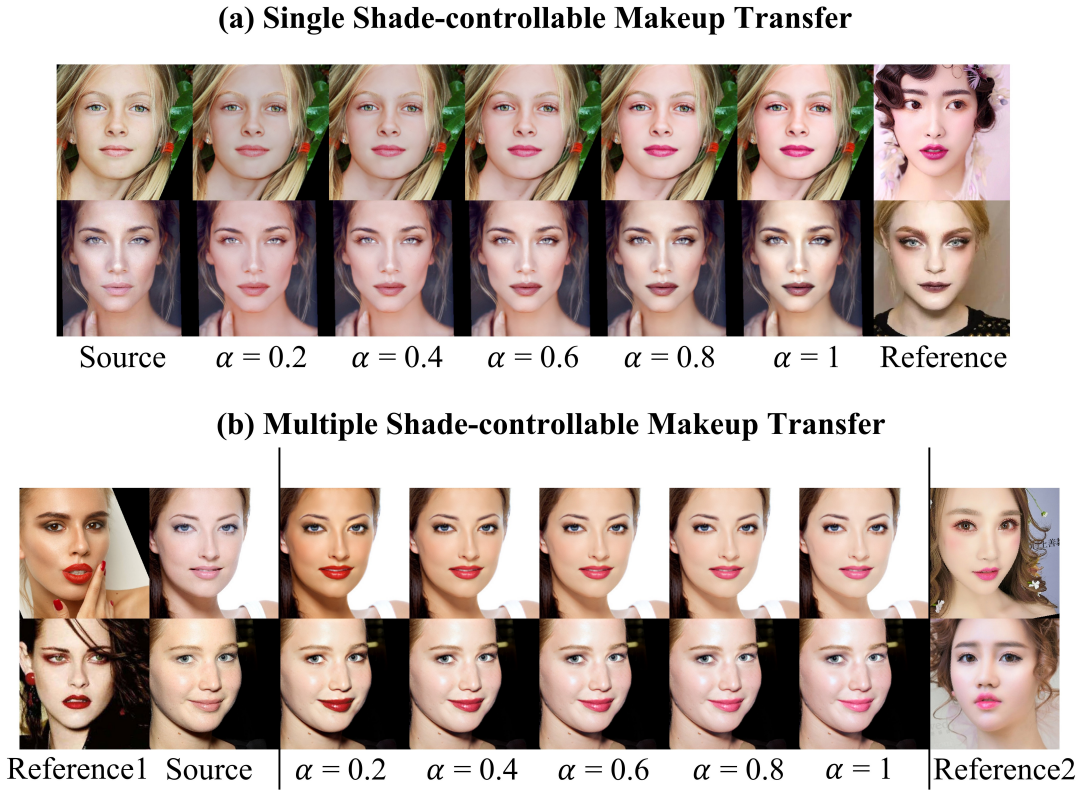


Figure 3.7: (a) Shade-controllable makeup transfer results. The results are sorted from left to right according to the degree of makeup from light to heavy. (b) The makeup style of the generated images gradually changes from Reference1 to Reference2 from left to right. Note that the displayed source and reference images are from the MT dataset.

indicating superior image quality. To ensure that identity is preserved during makeup transfer, we employ ArcFace [27] to measure the facial similarity before and after the makeup transformation. Comparison tests with other methods indicate that IP23-Net obtained the highest score of 0.9743, demonstrating better feature maintenance. Additionally, we incorporate the Natural Image Quality Evaluator (NIQE) to gauge the visual quality of the images. NIQE is a non-referenced measure that assesses the naturalness of an image based on the statistical properties of a natural, undistorted image. Our IP23-Net scores 5.6377 in NIQE, outperforming others in visual naturalness and reducing perceptual distortion, thus confirming the effectiveness of our approach in maintaining image quality and fidelity. The results are shown in the Table 3.2.

To assess the preservation of expression and pose, we employed a pre-trained MTCNN model to detect facial landmarks in source images. We then calculated the cosine similarity (CosSim $\in [0,1]$) of these landmarks to evaluate our method's efficacy in maintaining

Method	IP23-Net	SCGAN[26]	PSGAN[61]	BeautyGAN[82]
FID ↓	**45.59**	47.78	51.78	57.43
Arcface ↑	**0.9743**	0.9611	0.9721	0.9691
NIQE ↓	**5.6377**	6.1762	5.7311	5.8729

Table 3.2: Quantitative comparison between IP23-Net and other competitive methods in terms of FID, NIQE (lower is better) and Arcface (higher is better).

Method	IP23-Net	SCGAN[26]	PSGAN[61]	BeautyGAN[82]
CosSim ↑	**0.9994**	0.9971	0.9992	0.9992

Table 3.3: Quantitative comparison of expressions and poses with other competitors.

Method	IP23-Net	SCGAN[26]	PSGAN[61]	BeautyGAN[82]
Time (s/img)	0.0241	0.1272	1.1586	**0.0156**

Table 3.4: Comparison of time efficiency of our method with other competitors.

facial expressions and poses. As Table 3.3 shows, IP23-Net and other baseline methods demonstrate effective preservation of expressions and poses.

3.3.2 Computational Complexity Analysis.

For the tests regarding the computational complexity of our IP23-Net, we also compared it to previous approaches, and for a fair comparison, we used a single RTX6000 GPU uniformly in our tests, and all models are evaluated using the test section of the MT dataset to ensure fairness. Table 3.4 illustrates that IP23-Net significantly outstrips competitors with an average inference speed of only 0.0241 seconds per image. This compares favorably to the times recorded by PSGAN and SCGAN. Such efficiency is maintained even with the inclusion of complex pre-trained models like the 3D Morphable Model and Arcface in our encoder framework, which are highly optimized for feature extraction and impose minimal computational burden. IP23-Net thus offers an optimal blend of rapid inference and sturdy performance, essential for real-time processing scenarios.

3.3.3 Generalization to Unseen Images

In this study, we introduce the novel High Fidelity Makeup Transfer Network via 2D and 3D Identity Preservation, termed as IP23-Net. To test the generalizability of our IP23-Net. We train the model on the publicly available MT dataset [82] and test on the unseen images from High Resolution Synthetic Makeup (HRSM) Dataset. Fig. 3.8 demonstrates the efficacy of our

method. Moreover, to comprehensively assess the robustness of our model, we select unseen male images from the High Resolution Synthetic Makeup (HRSM) dataset for additional testing. Fig. 3.9 illustrates the results of our makeup application on male subjects. To further validate the performance and generalizability of our proposed IP23-Net method, we have also conducted visualizations of makeup transfer effects on the Flickr-Faces-HQ dataset [67] and EDFace-Celeb-1M dataset [187] (shown in Fig. 3.10). In addition, in order to more fully demonstrate the performance of our method in the transfer of different makeup styles, we purposely add Fig. 3.11 to visualise a randomly selected case. As you can see, the left part shows the results of a heavier makeup transfer, while the right part shows the results of a more traditional Asian makeup style transfer. By comparing the results, we can see the flexibility and efficiency of IP23-Net in handling different makeup styles. From all the above evidence, it is evident that our model not only delivers high-quality results across various makeup styles but also exhibits an impressive balance between transformation and preservation. Notably, the backgrounds of the characters remain intact after the makeup transfer, maintaining the overall integrity of the image.

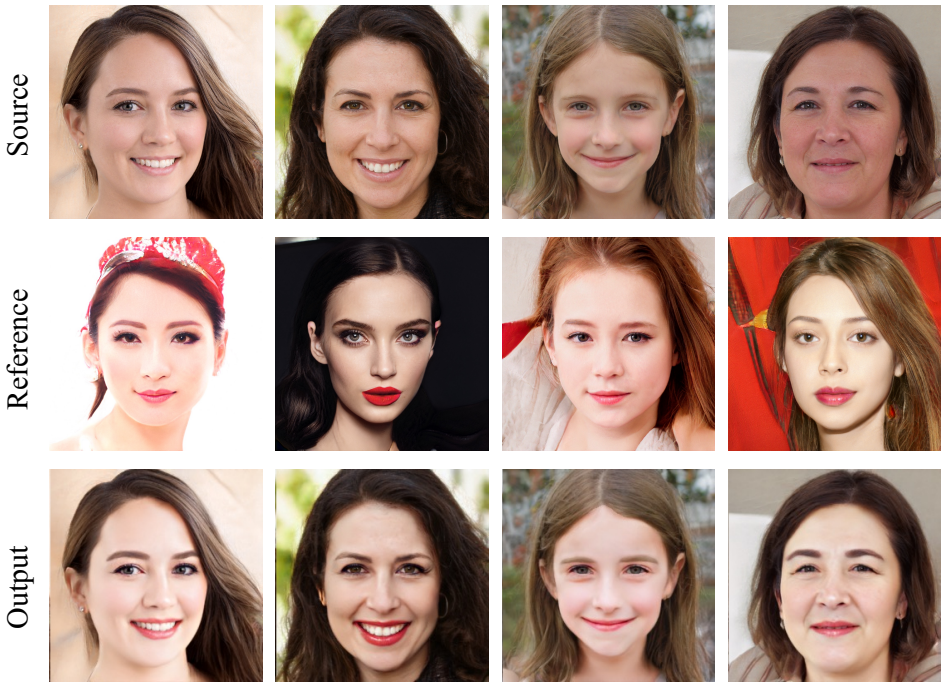


Figure 3.8: Results of makeup transfer. The source and reference images are from the HRSM (ours) dataset.

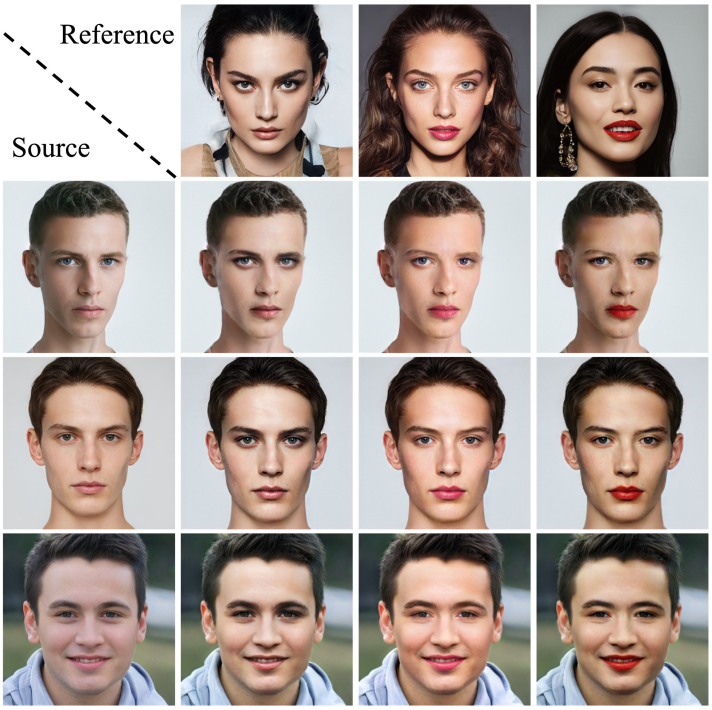


Figure 3.9: Male makeup results. The source and reference images are from the HRSM (ours) dataset.

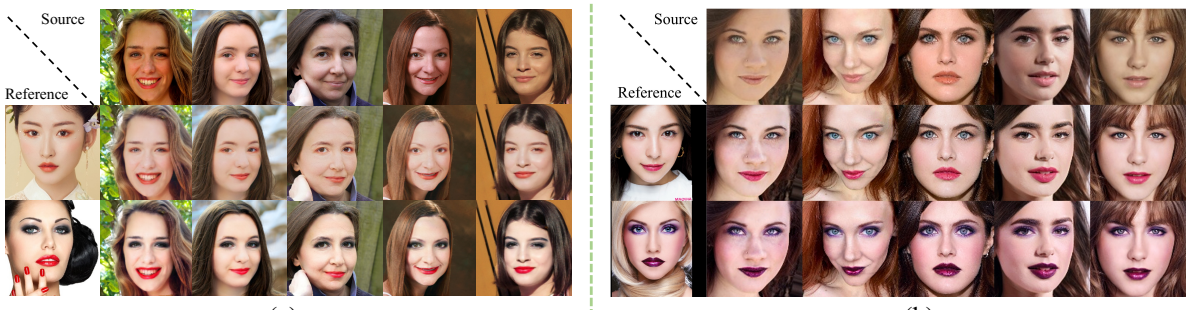


Figure 3.10: Results of makeup transfer by IP23-Net. (a) The source and reference images are from the Flickr-Faces-HQ (FFHQ) dataset [67]. (b) The source and reference images are from the EDFace-Celeb-1M dataset [187].

3.3.4 Comparison of Facial Details

Fig. 3.12 left part presents a detailed comparison of our IP23-Net with state-of-the-art methods PSGAN [61] and SCGAN [26]. The makeup images generated by our network exhibit notable superiority in retaining more character details and shadows, especially within the highlighted regions. This is a significant improvement over the prior methods, enhancing

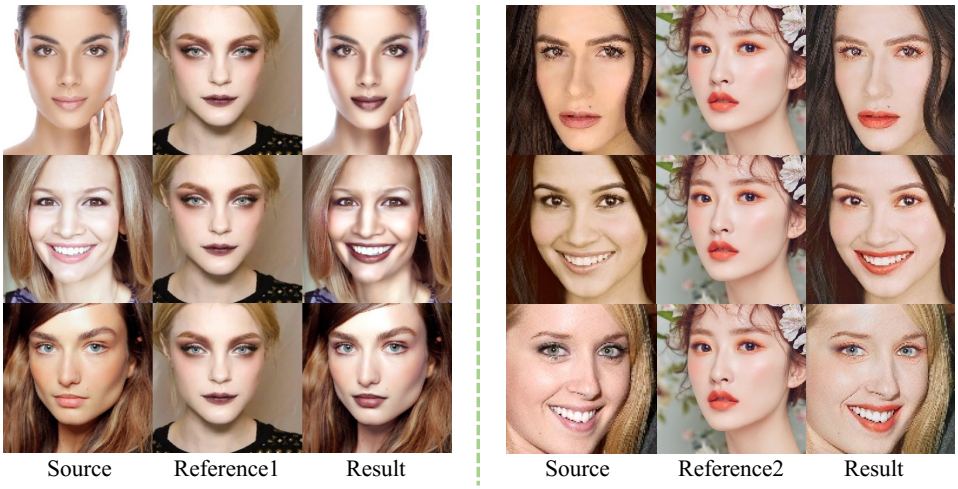


Figure 3.11: Results of makeup transfer by IP23-Net. The different source images are selected randomly, two makeup styles with widely differing styles are selected as reference images.

the three-dimensionality of the characters and making the makeup appear more realistic. A key feature of our approach is its ability to enhance these aspects without compromising the unique personality of the characters, thereby achieving a balance between transformation and authenticity that previous methods have struggled to attain. This balance results in a more holistic and appealing representation of makeup transfer, thereby pushing the boundaries of what's achievable in this field. Furthermore, our approach not only excels in executing high-quality makeup transfer but also demonstrates a strong commitment to maintaining the original image's integrity. This effectiveness is achieved by restoring the background of the characters and non-makeup areas, as illustrated in an ID photo presented in Fig. 3.12 right part. This figure offers a detailed comparison with state-of-the-art methods PSGAN [61] and SCGAN [26]. In contrast to these previous methods, IP23-Net demonstrates high accuracy in reconstructing the character background, further preserving the authenticity of the original image. The content within the black box accentuates the superior capacity of IP23-Net for character detail preservation, enhancement of image three-dimensionality, and maintenance of original image integrity. The effectiveness of our approach in background restoration, combined with its proficiency in makeup transfer, marks a significant advancement in the field. It establishes a balance between aesthetic enhancement and preservation of the original image, ensuring a realistic and holistic representation in makeup transfer tasks.

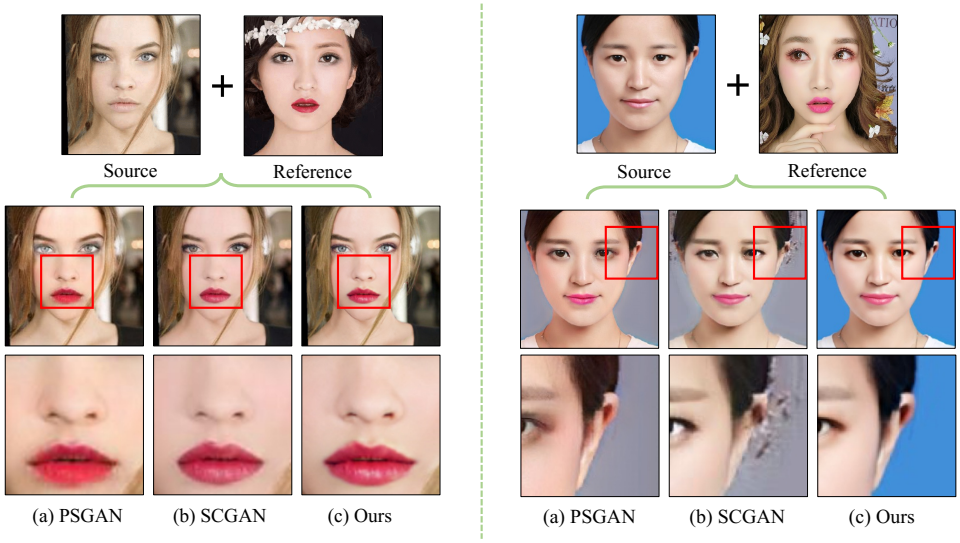


Figure 3.12: Local detailed comparison. Both source and reference images are selected from the MT dataset [82]. From the left side of the figure, it is evident that IP23-Net is able to generate highly stereoscopic makeup images with accurate background restoration, By comparing the content in the black box, we find that our generated lipstick is plump and glossy. In addition, according to the right part of the figure, we can find out that our method is also satisfactory for background reconstruction.

3.3.5 Partial Makeup Transfer

IP23-Net supports partial makeup transfer. People are able to choose the area of makeup by themselves from different reference images. Given a source image x and two reference images y_1 and y_2. We can use the lip part from y_2 and other parts from y_1 to obtain the new mixed makeup area $Y = [{y_1}^{skin}, {y_1}^{eyes}, {y_1}^{nose}, {y_2}^{lip}]$, where $[\cdot,\cdot]$ denotes concatenation operation. The input to Facial Component Style Encoder is the facial components, and we can select specific parts from any reference image. The partial makeup transfer results are presented in Fig. 3.6. We can find that the style of the lip in the generated image is from Reference2, and the others (eyes, nose, and skin) are from Reference1.

3.3.6 Shade-controllable Makeup Transfer

Our model can control the degree of makeup easily. We use a style-code f_{all} to represent the input feature of Background Correction Decoder, consisting of content feature f_c and makeup style feature f_y. We set a coefficient $\alpha \in [0,1]$ to control the weight of f_c and f_y, which is defined as $f_{combined} = \alpha f_y + (1-\alpha) f_c$. Fig. 3.7 (a) shows the makeup degree from light to heavy. We can see that the makeup degree of the source image gets higher when the coefficient α increases. In addition, IP23-Net supports the makeup style fuse from multiple

Method	IP23-Net	SCGAN[26]	PSGAN[61]	BeautyGAN[82]
RSB ↑	**53.68**	16.55	24.72	5.05

Table 3.5: Evaluation results of different makeup transfer methods in user studies, represented by the Ratio Selected as Best (RSB) in percentage.

reference images. We can also use the coefficient α to change the makeup style of the source image from Reference1 to Reference2. The style features f_{y_1} and f_{y_2} are from Reference1 and Reference2, respectively. The style-code f_{all} is calculated by $f_{all} = (1-\alpha) f_{y_1} + \alpha f_{y_2}$. Fig. 3.7 (b) shows the makeup style changes from Reference1 to Reference2 when the coefficient α increases. In other words, the makeup style is closer to the reference image that contributes more style feature to f_{all}.

3.3.7 User Study

We conducted comprehensive user studies to quantitatively evaluate the robustness and visual quality of IP23-Net with three makeup transfer methods, BeautyGAN, PSGAN and SCGAN. A total of 20 participants took part in these user studies. In the studies, we randomly selected 15 pairs of makeup and non-makeup images. The aim was to investigate the completeness of the makeup migration and whether the contours of the face were well preserved, in addition to whether the non-makeup areas of the characters were affected. In all user studies, participants were asked to select the result with the best visual quality and the most accurate transfer. Table 3.5 illustrates the results of the user studies, where our IP23-Net outperforms all state-of-the-art methods. We have reason to believe that this is due to the fact that we have introduced 3D face information reasonably well into the training process of the model, thus preserving the depth information of the face well.

3.3.8 Ablation Study

(1) Impact of 3D Stereoscopic Loss on Image Quality

To verify the effect of different losses, we perform ablation studies on the MT dataset. Specifically, we first add losses one by one to train different IP23-Nets under the same experimental setting. We then calculate the FID [49] between the reference image and the generated makeup image. Table 3.6 demonstrates the impact of each loss function in terms of the performance. The baseline model, which employed only adversarial loss, has achieved an FID of 80.49. By incorporating cycle consistency loss, the FID significantly decreased to 60.99, showing the importance of preserving the subject's identity during makeup transfer.

Loss	Performance				
Adversarial Loss	✓	✓	✓	✓	✓
Cycle Consistency Loss		✓	✓	✓	✓
Perceptual Loss			✓	✓	✓
Makeup Loss				✓	✓
3D Stereoscopic Loss					✓
FID ↓	80.49	60.99	59.06	48.06	**45.59**

Table 3.6: Ablation study for different losses.

The addition of perceptual loss led to an FID reduction to 59.06, implying that high-level semantic information from pre-trained models contributes to the preservation of facial details and makeup style. With the introduction of makeup loss, performance improved further, and the FID dropped to 48.06, emphasizing its effectiveness in preserving and transferring makeup attributes. Lastly, the inclusion of 3D stereoscopic loss resulted in the lowest FID of 45.59, signifying the value of 3D facial structure information for enhanced makeup transfer and more natural results.

The ablation studies clearly demonstrate that the combination of adversarial loss, cycle consistency loss, perceptual loss, makeup loss, and 3D stereoscopic loss is crucial for achieving optimal performance in our makeup transfer network.

(2) Fixed-Parameter Encoder Performance Evaluation

In our IP23-Net architecture, the design of the 3D Shape Identity Encoder incorporates two distinct encoder components: the Shape Encoder and the Identity Encoder. The Shape Encoder uses a 3D Morphable Model (3DMM) [28], which exists in a parametric form and is primarily utilized for reconstructing 3D facial structures from 2D images. This model reconstructs the 3D facial structure from 2D images by comparing the input facial data with the average facial model in the Basel Face Model (BFM) database. These parameters are pre-calculated based on extensive detailed 3D facial scan data and are not obtained through training. Therefore, in the Shape Encoder, we use a fixed 3DMM that accurately captures facial geometry, which is crucial for high-fidelity makeup transfer. On the other hand, the Identity Encoder employs the Arcface model [27], an advanced facial recognition network used to extract facial identity features. We added experiments to compare the differences between using the fixed-parameter Arcface model and an Arcface model with trainable parameters in the Identity Encoder. We compare two key performance metrics on the test set of the MT dataset: Arcface Similarity (higher is better) and Fréchet Inception Distance (FID) (lower is better).

Configuration	FID ↓	Arcface ↑
IP23-Net (Identity Encoder)	47.23	0.9631
IP23-Net (* Identity Encoder)	**45.59**	**0.9743**

Table 3.7: Performance comparison of IP23-Net with fixed (*) and trainable Identity Encoder configurations in terms of FID (lower is better) and Arcface Similarity (higher is better).

In IP23-Net, the fixed-parameter(*) Identity Encoder demonstrates superior performance over the trainable parameter configuration for identity consistency and image authenticity in makeup transfer tasks. As shown in Table 3.7, the reduced Fréchet Inception Distance (FID) from 47.23 to 45.59, indicating enhanced efficiency, and the increase in similarity from 0.9743 to 0.9631, highlighting its effectiveness. The fixed-parameter approach, pre-trained on a diverse dataset, effectively captures facial features, ensuring high-fidelity in makeup transfer. Its constant parameters during training contribute to stable outputs across various makeup styles. This stability is crucial for accurately replicating reference makeup styles while preserving the source identity.

(3) Effects of Shape Encoder.

We conduct additional experiments, selecting the seminal work in makeup transfer, BeautyGAN, as the foundation for our study. On this basis, we integrate our innovative feature, the Shape Encoder, and employ 3D stereoscopic loss as the supervisory mechanism. In line with the method we describe in this chapter, we incorporate the features generated by the Shape Encoder (f_{shape}) into the network through concatenation. We then apply a 1x1 convolutional layer to adjust the number of channels, ensuring effective integration and functioning of the features. To illustrate the enhancement's impact more clearly, we provide a visual comparison that contrasts with the original BeautyGAN method in Fig. 3.13. By observation, we can see that although the overall difference between the two generated makeup images is not very large, if we look closely at the lip part in the black box, we can still feel that the resulting image with the addition of f_{shape} has a more defined facial contour.

Finally, we believe that despite the integration of the Shape Encoder in BeautyGAN, there are some limitations in handling the interaction of f_{shape} features as the framework does not include our specially designed Background Correction Decoder. Due to insufficient feature fusion, the potential of the Shape Encoder could not be fully utilised, resulting in a less than obvious enhancement of the visualisation. Based on these observations, we conclude that Shape Encoder can be useful in different approaches, but its value can only be maximised in the IP23-Net framework.

(4) Evaluating the Impact of HRSM Dataset.

Figure 3.13: Comparative Visualization of Makeup Transfer with and without the Shape Encoder.

Integrating the High-Resolution Synthetic Makeup (HRSM) dataset with the Makeup Transfer (MT) dataset is expected to improve our model's robustness by diversifying its training data. The HRSM dataset, with its high-resolution and varied synthetic facial images, complements the MT dataset's style and complexity. This integration aims to broaden the model's understanding of diverse facial features and makeup styles, potentially improving its adaptability to complex real-world scenarios and its accuracy in makeup style transfer.

To test this hypothesis, we conducted a visual experiment, comparing models trained on the MT dataset alone with those trained on both MT and HRSM datasets, focusing on real-world image processing. The comparative results are illustrated in Fig. 3.14. Crucially, all test images in our experiment originate from real-world scenarios and are not part of the model's training set, ensuring unbiased performance evaluation. The comparative analysis revealed that models trained with the HRSM dataset more accurately replicated the makeup styles from reference images while preserving the identity of the source images. This finding supports the notion that integrating the HRSM dataset into training enhances the model's performance in real-world applications.

(5) Impact of face shape information on Image Quality

In our study, we extract facial depth information (f_{shape}) from the source images using a shape encoder. This is intended to enhance the three-dimensionality of the facial contours while preserving the original identity features of the subjects. An ablation study in order to validate the effectiveness of depth information in enhancing the three-dimensionality of makeup and preserving identity features is conducted. The results, as shown in Fig. 3.15, demonstrate that models utilizing f_{shape} perform better in maintaining facial contours and identity features. This proves the significant role of integrating depth information in improving the overall performance and output quality of the makeup transfer process. This discovery provides intuitive insights into the critical components of our method.

Discussion. The importance of depth information cannot be ignored in the process of

Figure 3.14: Comparative results of makeup transfer models trained without (w.o. HRSM) and with (w. HRSM) the HRSM dataset, using real-world source and reference images.

makeup transfer. It allows makeup to be applied in a way that respects the natural profile and height of the face. This means that makeup is not just adding layers of color to the face, but takes into account the three-dimensional structure of the face, such as the height of the bridge of the nose, the hollows of the eye sockets, and the curves of the cheeks. Such a makeup transfer not only looks more natural, but also better adapts to the unique facial features of different individuals.

A key reason we chose to use 3DMM is its ability to efficiently apply this 3D depth information in a 2D image. Despite the fact that the final output is on a 2D plane, the depth information obtained through 3DMM can enhance the three-dimensionality and dynamics of facial features. This approach overcomes the limitations in traditional 2D image processing, making makeup transfer not just a simple migration of colors, but a more comprehensive artistic creation that takes into account the 3D structure of the face. Therefore, 3DMM plays a crucial role in our makeup transfer process. It not only provides precise information about the depth of the face, but also ensures a natural and personalized makeup effect. We could observe the illumination on nose are significantly different.

3.4 Conclusion

In this chapter, we propose a new makeup network called 2D and 3D Identity Preservation Net in an attempt to overcome several limitations of the current makeup transfer framework. IP23-Net first distills the facial geometric information and identity feature from the source image by 3D Shape Identity Encoder. Then, we leverage Makeup Style Encoder to extract

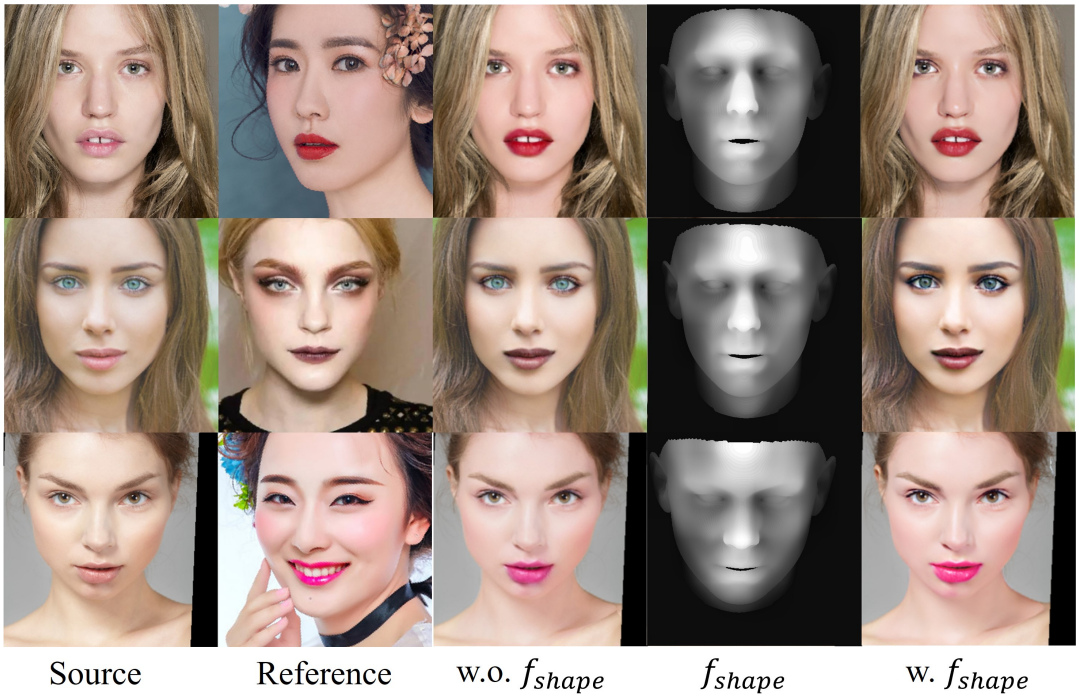


Figure 3.15: This figure shows makeup transfer with (w.) and without (w.o.) facial shape feature f_{shape}. The source and reference images are from the MT dataset. We can observe that faces with f_{shape} exhibit a stronger sense of three-dimensionality in the nose and lip areas. Without f_{shape} (third column), the facial features tend to lose their natural depth variation. In contrast, with f_{shape} (rightmost column), the model preserves important depth cues through more pronounced shadowing and highlights, particularly in defining the nasal bridge contours and lip volume. The grayscale f_{shape} visualization (fourth column) directly encodes the learned shape representation, where intensity values correspond to the degree of geometric prominence in facial features.

the local makeup style from the reference image. Besides, a 3D Stereoscopic loss is proposed to provide structure supervision to achieve high-fidelity makeup transfer. Background Correction Decoder utilizes makeup and identity features for makeup transfer while restoring the background by distinguishing face and background using the face shape extracted through 3DMM. Moreover, we introduce a High Resolution Synthetic Makeup dataset by StyleGAN2 [68], which is the largest makeup dataset. Extensive experiments on the MT dataset demonstrate that our approach can achieve competitive makeup transfer results and preserve the background accurately.

CHAPTER 4

REAL-TIME VIDEO DERAINING NETWORK WITH HIERARCHICAL MEMORY BANK

In this chapter, we study the video deraining task, which is a crucial aspect of reconstruction tasks and better aligns with real-world applications like vision-based autonomous driving. Video deraining aims to eliminate rain streaks and artifacts from video content. Existing methods using Convolutional Neural Network (CNN) deliver clear results but face challenges with slower inference speeds because of the complex architectures. To address these challenges, we introduce the Real-time Video Deraining Network (RVDNet), based on a spatial-temporal transformer, which integrates spatial and temporal deraining processes within a unified model. Contrary to conventional CNN-based video deraining techniques, our model integrates both temporal variations and spatial rain distortions without the need for distinct components. Furthermore, we employ a Long Short-Term Memory Bank (LSMB) to store features, sourced from encoders of the rainy input frames. LSMB adeptly merges immediate and historical frame attributes for clear rain layer pattern discernment and fortifies frame-to-frame communication, bolstering rain layer recognition and ensuring prompt inference. Comprehensive evaluations across three public datasets confirm that our method surpasses state-of-the-art benchmarks in terms of accuracy and operational speed. We will release the code to the public. In summary, this chapter presents our approach to advancing video deraining through the development of RVDNet, which significantly improves both performance and speed over existing methods.

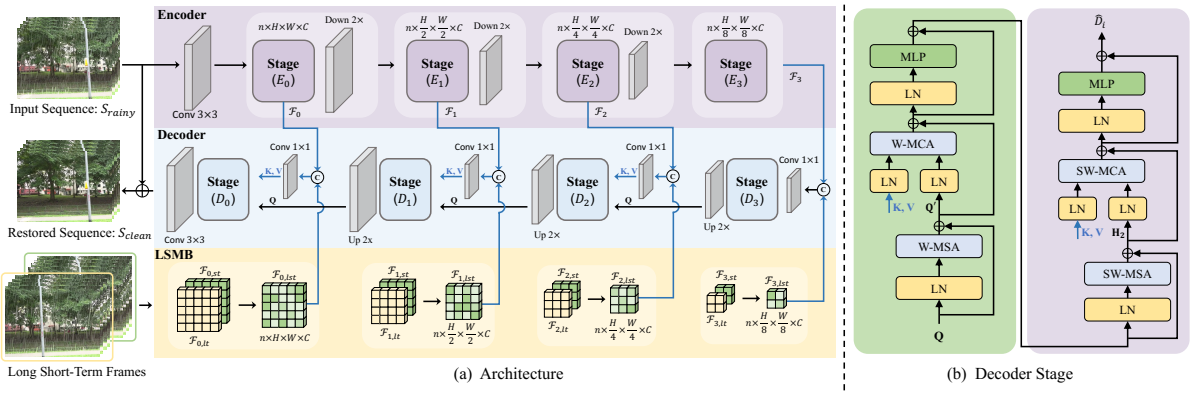


Figure 4.1: (a) The RVDNet architecture utilizes a UNet-style approach, integrating transformer blocks in both the Encoder and Decoder sections, and highlights the features of the Long Short-term Memory Bank (LSMB). (b) The basic transformer block used in Decoder stage D_i, $i \in 0, 1, 2$ of RVDNet.

4.1 Introduction

In outdoor settings, adverse weather conditions often cause cameras to record compromised visuals, with rain layers notably diminishing visibility and impacting vision-based system efficacy. Rain-induced disturbances create significant challenges for computer vision tasks, including autonomous driving, augmented reality, and video editing. These disturbances can obscure crucial visual elements or specific object attributes. Thus, deraining has emerged as a pivotal area in computer vision research, emphasizing the optimization of intelligent systems under rain-affected conditions. In outdoor environments, researchers continue to make progress in developing robust deraining methods to eliminate rain from images and videos. Deraining techniques can be divided into two primary categories: image deraining and video deraining. Image deraining [21, 169] harnesses spatial information to remove rain from images. Meanwhile, video deraining methods [79, 91, 126, 166, 170] further utilize temporal information across multiple video frames to recover content affected by rain. Several contemporary video-based methods [167, 186] employ a two-stage network for this purpose. For instance, ESTINet [186] captures spatial information before processing temporal information between frames. On the other hand, RDDNet [150] integrates the concept of rain streak motions, ensuring consistency of rain layers across video frames by referencing rain-layer annotations.

While current video deraining efforts showcase considerable progress, two primary challenges linger in the field. (1) *Architecture Inefficiency.* Several of the advanced video deraining models [167, 170, 188] differentiate between spatial and temporal correlations, employing complex cascaded networks. These dual-structured models face latency issues

due to their intricate parameter setups, which restrict their use in real-time applications. (2) *Limited Temporal Information Utilization.* A notable limitation in contemporary techniques is that, although processing sequences with multiple consecutive frames enhances intra-sequence interactions, the network's capability to restore the scene experiences only minor improvements. This issue largely stems from overlapping content in successive frames, leading to a lack of informational diversity. Incorporating a broader range of frames might tackle this issue, but it markedly elevates the computational demands, amplifying the concerns underscored in point (1).

To address the above limitations, we propose an end-to-end Real-time Video Deraining Network (RVDNet) based on the transformer to achieve a real-time inference speed with satisfactory performance. Specifically, inspired by the recent success of vision transformers [147] in video understanding, we design RVDNet in a cascaded UNet-style architecture to effectively incorporate all the spatial and temporal information for background reconstruction. Incorporating a hierarchical structure, we introduce a Long Short-Term Memory Bank (LSMB) to retain features from preceding frames. Features stored within the LSMB are classified based on temporal variations into Long-Term and Short-Term features. By integrating a more enriched background context at dynamically learned ratios into the attention computation of the transformer block, we aim to enhance the network's restoration efficacy for current input frames. The LSMB draws its feature storage from multi-scale features provided at each stage by the encoder. This approach not only prevents redundant feature extraction but also ensures the network optimally accomplishes real-time inference tasks. The contents of the LSMB, once stored, become accessible in the decoder phase, serving as a direct basis for reconstructing derained frames. Fundamentally, the LSMB serves to strengthen the connection between successive video frames, diminish computational burdens through the utilization of conserved features, and maintain an efficient inference pace while enhancing the network's deraining capabilities.

- We make the following contributions. First, we propose an end-to-end Real-time Video Deraining Network (RVDNet) to extract better spatial-temporal information for video deraining in a single model
- We develop a lightweight LSMB to enhance the reconstruction of the network. To the best of our knowledge, RVDNet is the first real-time transformer-based framework for video-deraining.
- We achieve new state-of-the-art performance on widely-used video deraining benchmarks, including NTURain, RainSynLight25 and RainSynHeavy25.

4.2 Method

4.2.1 Network Architecture

For a rainy sequence S_{rainy} comprising n input rainy frames $\{f_{t-n}, ..., f_{t-1}, f_t\}$ of dimensions $n \times H \times W \times 3$, our objective is to devise a deraining model that can eliminate the rain layer and produce the clean sequence S_{clean}. The network structure is shown in Fig. 4.1(a).

Encoder. A singular convolutional layer with a kernel size of 3×3 efficiently extracts C features from rainy frames. The encoder has four stages, represented as E_i, with $i \in \{0,1,2,3\}$. Within E_i, Swin Transformer blocks utilize shifting non-overlapping windows to balance computational efficiency with learning long-range dependencies. Given window size $M \times M$, it partitions input video frames into non-overlapping windows. Post Layer Normalization (LN), Window-based Multi-head Self-Attention (W-MSA) [98] processes local attention. Subsequently, a Multi-Layer Perception (MLP) coupled with an LN layer undergoes further transformation. Another Swin Transformer block introduces Shifted Window-based Multi-head Self-Attention (SW-MSA) [98], integrating cross-window connections. In this block, the only difference is a shift in input features by $\lfloor \frac{M}{2} \rfloor \times \lfloor \frac{M}{2} \rfloor$ before partitioning, enabling it to capture dependencies both spatially and temporally. SW-MSA outputs are then downsampled via a convolutional layer, except for E_3. Incidentally, $\mathscr{F}_i$, are stored in the Long Short-Term Memory Bank (LSMB) to optimize decoder performance.

Decoder. Same as the encoder part, the decoder incorporates four stages, termed as D_i, with $i \in \{0,1,2,3\}$. Excluding the final stage D_0, each subsequent stage incorporates a convolutional layer for upsampling. The cascaded UNet-style structure ensures each decoder stage D_i has two inputs, excluding D_3. These consist of output features from the preceding decoder stage as ($\mathbf{Q}$) and features from LSMB as ($\mathbf{K}, \mathbf{V}$). Besides employing W-MSA and SW-MSA, our approach integrates Window-based Multi-head Cross Attention (W-MCA) and Shifted Window-based Cross Attention (SW-MCA) between $\mathbf{Q}$ and $\mathbf{K}, \mathbf{V}$ (as shown in Fig. 4.1(b)). The specific operation sequence is as follows:

$$
\begin{aligned}
\mathrm{Q}' &= \text{W-MSA}(\text{LN}(\mathrm{Q})) + \mathrm{Q} \\
\mathrm{H}_1 &= \text{W-MCA}(\text{LN}(\mathrm{Q}'), \text{LN}(\mathrm{K}), \text{LN}(\mathrm{V})) + \mathrm{Q}' \\
\mathrm{H}_1' &= \text{MLP}(\text{LN}(\mathrm{H}_1)) + \mathrm{H}_1 \\
\mathrm{H}_2 &= \text{SW-MSA}(\text{LN}(\mathrm{H}_1')) + \mathrm{H}_1' \\
\mathrm{H}_2' &= \text{SW-MCA}(\text{LN}(\mathrm{H}_2), \text{LN}(\mathrm{K}), \text{LN}(\mathrm{V})) + \mathrm{H}_2 \\
\hat{D}_i &= \text{MLP}(\text{LN}(\mathrm{H}_2')) + \mathrm{H}_2'.
\end{aligned}
\tag{4.1}
$$

$\hat{D}_i$ represents the output from each individual stage of the Decoder. Finally, instead of

employing a tail reconstruction module with numerous residual-blocks, we utilize a singular convolutional layer with a 3×3 kernel size to translate features to the RGB level. These modifications aim to maintain the real-time inference capability of our RVDNet.

4.2.2 Long Short-Term Memory Bank

Traditional methods often focus on information derived from adjacent input frames. Recognizing the constraints of depending only on short-term local changes to capture full rain layer data, we propose a hierarchical feature matching and propagation approach using multiple sequences. Memory networks demonstrate stability, particularly evident in video object segmentation [112]. Inspired by this, we introduce the Long Short-Term Memory Bank (LSMB) for video deraining.

Within the LSMB framework, multi-scale features originate from the current input of n consecutive frames through the encoder, and from an additional $2n$ previous frames. Such an approach reduces computational demands and enhances network inference speed. These multi-scale features from different frames are segmented into three distinct categories: the features corresponding to the initial n frames are considered long-term and denoted as $\mathscr{F}_{i,lt}$. The features from the following $f_{n+1,\ldots,2n}$ frames are termed short-term, symbolized as $\mathscr{F}_{i,st}$. Lastly, the features stemming from the latest rainy frames, specifically $f_{2n+1,\ldots,3n}$, are designated $\mathscr{F}_i$, where i indicates the encoder's stage and falls within $\{0,1,2\}$.

Short-Term Features. $\mathscr{F}_{i,st}$ captures the current data associated with $\mathscr{F}_i$. Considering the typically smooth transitions between sequential video frames, $\mathscr{F}_{i,st}$ offers an in-depth scene context, enhancing background reconstruction. However, due to minimal background variations, the information provided by $\mathscr{F}_{i,st}$ can contain redundant portions.

Long-Term Features. The temporal gap between $\mathscr{F}_{i,lt}$ and $\mathscr{F}_i$ presents difficulties, particularly during pronounced scene changes or video irregularities such as camera reframing. Sequences from earlier frames might provide a more robust reference for reconstruction. In situations characterized by these irregularities, $\mathscr{F}_{i,lt}$ serves as an essential long-term reference, ensuring a stable foundation for scene reconstruction. This methodology leverages the inherent stability of long-term sequences, amplifying reconstruction accuracy. In crafting our Real-time Video Deraining Network (RVDNet), we merge the insights from both long and short-term features to enhance deraining. Both $\mathscr{F}_{i,lt}$ and $\mathscr{F}_{i,st}$ are integral during the reconstruction stage. We've integrated a flexible parameter $\alpha \in [0,1]$ that adjusts adaptively throughout the training phase. Considering RVDNet's real-time priority, maintaining computational efficiency is paramount. We concatenate $\mathscr{F}_{i,lt}$, $\mathscr{F}_{i,st}$, and $\mathscr{F}_i$ and then utilize a 1x1 convolutional layer to fine-tune these combined features, ensuring alignment with the

Method		NTURain		RainSynLight25		RainSynHeavy25		FPS	Parameters
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	↑	(Millions) ↓
FastDerain [60]	TIP'18	30.32	0.9262	29.42	0.8683	19.25	0.5385	2.52	-
SpacCNN [17]	CVPR'18	33.11	0.9474	32.78	0.9239	21.21	0.5854	0.11	-
FCDN [166]	CVPR'19	36.05	0.9676	35.80	0.9622	27.72	0.8239	1.11	-
SLDNet [170]	CVPR'20	34.89	0.9540	34.28	0.9586	26.51	0.7966	-	-
S2VD [177]	CVPR'21	37.37	0.9683	34.66	0.9403	27.03	0.8255	8.07	**2.05**
ESTINet [188]	TPAMI'22	37.48	0.9700	36.12	0.9631	28.48	0.8242	1.49	91.96
RDDNet [150]	ECCV'22	**37.71**	**0.9720**	**38.61**	**0.9766**	**32.39**	**0.9318**	0.54	30.64
RVDNet-S	Ours	37.31	0.9718	38.31	0.9734	30.83	0.9293	**25.64**	**4.02**
RVDNet-L	Ours	**38.85**	**0.9771**	**39.22**	**0.9793**	**33.28**	**0.9374**	**19.61**	5.58

Table 4.1: Quantitative comparison of our network and compared methods on three public datasets NTURain [17], RainSynLight25 and RainSynHeavy25 [42]. FPS is computed on Nvidia Quadro RTX 6000 machine and on NTURain dataset. Best results are denoted in red and the second best results are denoted in blue.

original $\mathscr{F}_i$ dimensions. The formula is defined as:

$$\begin{aligned} \mathscr{F}_{i,lst} &= \alpha \cdot \mathscr{F}_{i,lt} + (1-\alpha) \cdot \mathscr{F}_{i,st} \\ \mathscr{F}_{i,lst} &= \text{Conv}_{1\times1}(\text{Concat}(\mathscr{F}_i, \mathscr{F}_{i,lst})). \end{aligned} \tag{4.2}$$

4.3 Experiment

4.3.1 Implementation Details

Datasets. The NTURain dataset, curated by Chen *et al.* [17], includes images from cameras at varied motion speeds, with 24 training sequences of rainy scenes and 8 testing sequence pairs. It also contains seven authentic rainy videos. RainSynLight25 provides 190 RGB training sequence pairs and 27 testing pairs, each with rainy and clean versions. These clean images derive from CIF, HDTV, and HEVC standard sequences, enhanced by rain streaks from a probabilistic model [39]. RainSynHeavy25, similar to RainSynLight25, features more distinct rain streaks characterized by clear lines and sparkle noises.

Training. We propose RVDNet in two versions, denoted as small (S) and large (L) architectures, which correspond to the number of Swin Transformer blocks employed in each stage of the encoder E_i and decoder D_i. Specifically, for the small and large architectures, we set the number of Swin Transformer blocks to 4, and 6, respectively. The network is optimized with an $\mathbf{L_1}$ loss and we utilize PyTorch to implement our RVDNet and train it using the Adam

optimizer on two NVIDIA RTX 6000 GPUs. Our network takes in four frames as input to perform video deraining.

Inference. During the initial n frames, predictions rely on the RVDNet baseline. With the second batch of inputs, LSMB initiates, using the first batch as $\mathscr{F}_{i,st}$. As it progresses, $\mathscr{F}_{i,st}$ and $\mathscr{F}_{i,lt}$ activate, improving inference using enhanced features.

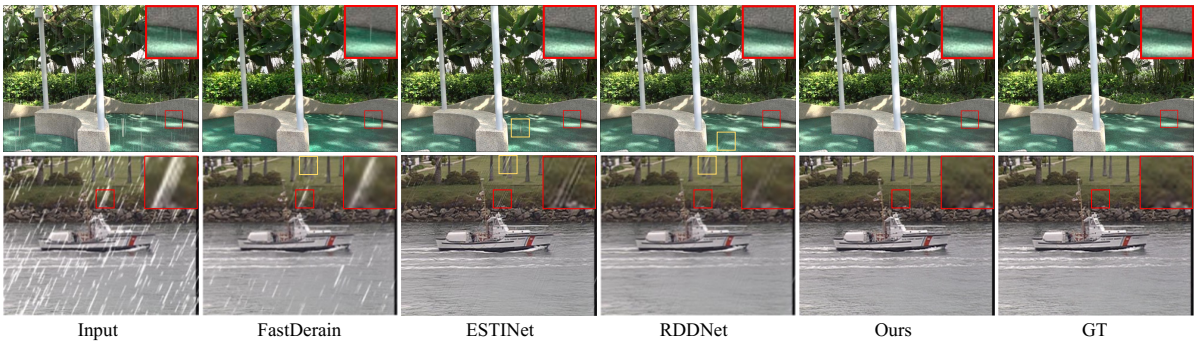


Figure 4.2: Visual comparison of different deraining methods on NTURain dataset (Upper Part) and SynLight25 dataset (Lower Part). The yellow box indicates the comparison of rain streak removal. The red box indicates the comparison of detail retention.

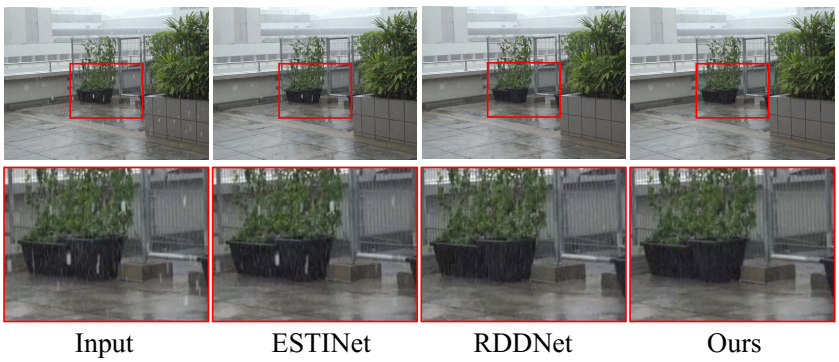


Figure 4.3: Visual comparison of different deraining methods on real-world rainy frame from NTURain dataset [17]. The red box indicates the comparison of detail retention.

RVDT-L	**Metric 1**	**Metric 2**	**Metric 3**
Basic	✓	✓	✓
Short-Term		✓	✓
Long Short-Term			✓
PSNR (↑)	37.41	38.32	**38.85**
SSIM (↑)	0.9716	0.9759	**0.9771**

Table 4.2: Ablation Study on RVDNet-L Performance using the NTURain dataset.

Input Frames	PSNR (↑)	SSIM (↑)
2 frames	37.41	0.9716
4 frames	38.85	0.9771
6 frames	38.92	0.9775
8 frames	**38.94**	**0.9776**

Table 4.3: Ablation study on the impact of input frames in RVDNet-L.

4.3.2 Quantitative Evaluation

Table 4.1 provides a thorough comparison between our proposed models, RVDNet-S and RVDNet-L, and other advanced video deraining methods, such as FastDerain [60], SpacCNN [17], FCDN [166], SLDNet [170], S2VD [177], ESTINet [188], and RDDNet [150] across three public datasets. RVDNet-S, designed for real-time requirements beyond 24 FPS, shows prominent performance with an impressive 25.64 FPS, emphasizing its efficiency in real-time situations. Meanwhile, RVDNet-L, developed with a focus on detailed restoration, consistently achieves top PSNR and SSIM scores across all datasets, indicating its prowess in preserving intricate image details. The unique attributes of our models address deraining challenges, from fast processing to detail preservation.

4.3.3 Qualitative Evaluation

We evaluated our RVDNet-L model against recent advanced techniques such as FastDerain [60], ESTINet [188], and RDDNet [150] on the SynLight25 and NTURain datasets (refer to Fig.4.2 and Fig.4.3). Our approach effectively retains pristine backgrounds post rain removal, evidenced by enhanced PSNR/SSIM values. Fig.4.2 demonstrates that our model achieves thorough rain layer removal. While the visualization from RDDNet shares similarities with our output, our model showcases superior inference speed during detailed background reconstruction. The test frame in Fig.4.3 is from the NTURain real-world dataset, with close-up views highlighting our approach's exceptional background restoration capabilities.

4.3.4 Ablation Study

The ablation study conducted on RVDNet-L, as presented in Table 4.2, underscores the significance of temporal elements in video deraining. Starting with a PSNR of 37.41 and SSIM of 0.9716 for the basic model, the introduction of the short-term component brought about a marginal improvement. However, when the long short-term component was added, the metrics increased notably to a PSNR of 38.85 and an SSIM of 0.9771. This highlights the

importance of leveraging both immediate and extended temporal associations for proficient rain pattern detection and video enhancement.

Additionally, we conduct experiments to investigate the impact of input frame numbers, as shown in Table 4.3. The results reveal that while using only 2 frames leads to suboptimal performance (PSNR: 37.41, SSIM: 0.9716), increasing to 4 frames significantly improves the deraining quality (PSNR: 38.85, SSIM: 0.9771). Further increasing the input frames to 6 or 8 yields minimal improvements (less than 0.1dB PSNR), suggesting that 4 frames provide sufficient temporal information for effective rain removal while maintaining computational efficiency.

4.4 Conclusion

In conclusion, the RVDNet approach addresses video deraining challenges effectively. Merging spatial-temporal data and utilizing an LSMB, it enhances inter-frame connections and reduces computational demands. Its exemplary results on leading benchmarks highlight the model's significant potential in handling rain-affected visuals.

CHAPTER 5

CRISS-CROSS DIFFUSION MODELS FOR ALL-IN-ONE BLIND IMAGE RESTORATION

After optimizing the balance between speed and performance in video deraining, we expand our research into the field of All-in-One image restoration, utilizing a unified restoration model across multiple degradation scenarios rather than employing separate models for individual cases. In this chapter, we first introduce the diffusion model to address the All-in-One image restoration task, instead of relying on conventional Transformer and CNN-based methods. The diffusion model is trained via a probabilistic process of incremental denoising, enabling it to effectively capture diverse image patterns and demonstrate strong generalization capabilities, making it particularly well-suited for addressing a wide range of unknown image degradation challenges. However, after observing the reconstruction process over multiple iterations of the diffusion model, it becomes evident that the reconstruction of high-frequency texture information is often random and inaccurate due to the suppression of edge information as noise.

To address this issue, we propose the Criss-cross Diffusion Model (CrDiff) for All-in-One image restoration. This model leverages static wavelet transform operations to extract high-frequency information from degraded images and guides the diffusion model to reconstruct high-frequency textures through a novel high-frequency encoder in the latent space. Furthermore, to ensure that this encoder captures accurate high-frequency information during training, we include a matching high-frequency decoder in the training process. Extensive experiments on popular benchmarks for multiple degradation tasks show that CrDiff achieves excellent performance.

5.1 Introduction

Image degradation is a common problem that affects the quality and clarity of images captured by cameras. It can be caused by various factors, such as noise, blur, haze, and rain, interfering with the capture process. Image restoration is a task that aims to recover a high-quality clean image from a degraded one. Traditional approaches focus on the exploration of the image prior, such as sparse [99, 104], low-rank [45, 161], self-similarity [24] *etc.* Recently, deep learning-based methods [1, 21, 142, 206] have achieved remarkable results in image restoration. However, these methods are usually designed and trained for specific degradation types and levels, such as denoising [2, 8, 178], deraining [81, 202], and dehazing [34, 124]. Therefore, they cannot handle complex and diverse degradation scenarios that may occur in real-world applications. To expand their applicability, some methods have proposed to develop *All-in-one* models that can restore images from various types and levels of degradation. For example, AirNet [73] uses an extra encoder to learn different types of image degradation by contrastive learning. PromptIR [119] leverages the prompt learning approach to enhance the degradation-specific representations.

Nevertheless, existing methods enhance the generalization of models through supplementary designs; however, these improvements still rely on conventional architectures and fail to deliver an all-in-one solution that can uniformly address all types of degradation. To address this limitation, we introduce a **Cr**iss-cross **Diff**usion model (CrDiff) based on a diffusion model that learns the data distribution by adding and removing noise from images [50]. Specifically, we observe that the diffusion model tends to put more effort into restoring color patches while ignoring high-frequency details in the image which contains the details of the original image, such as edges, textures, and patterns. Based on the above observations, we extend the vanilla diffusion model to incorporate high-frequency information from degraded images. We develop the Stationary Wavelet Transform (SWT) to filter high-frequency information from the degraded image and via a lightweight extra encoder E_h to integrate the high-frequency information in the latent space. We then design a high-frequency fusion block to use the high-frequency information as a guiding signal in the latent space, helping the model reconstruct detailed features more accurately. Moreover, we design a paired decoder D_h between E_h to reconstruct the high-frequency information. It aims to ensure the ability to extract high-frequency information of E_h. By combining the above design, our CrDiff model can restore the input degraded image with high quality in an *All-in-One* manner without any prior knowledge. Extensive experiments demonstrate the robust generalization capabilities of our Crdiff model, which consistently achieves state-of-the-art results across various image restoration tasks, including denoising, deraining, and dehazing,

all within a unified framework.

The main contributions are summarized as follows:

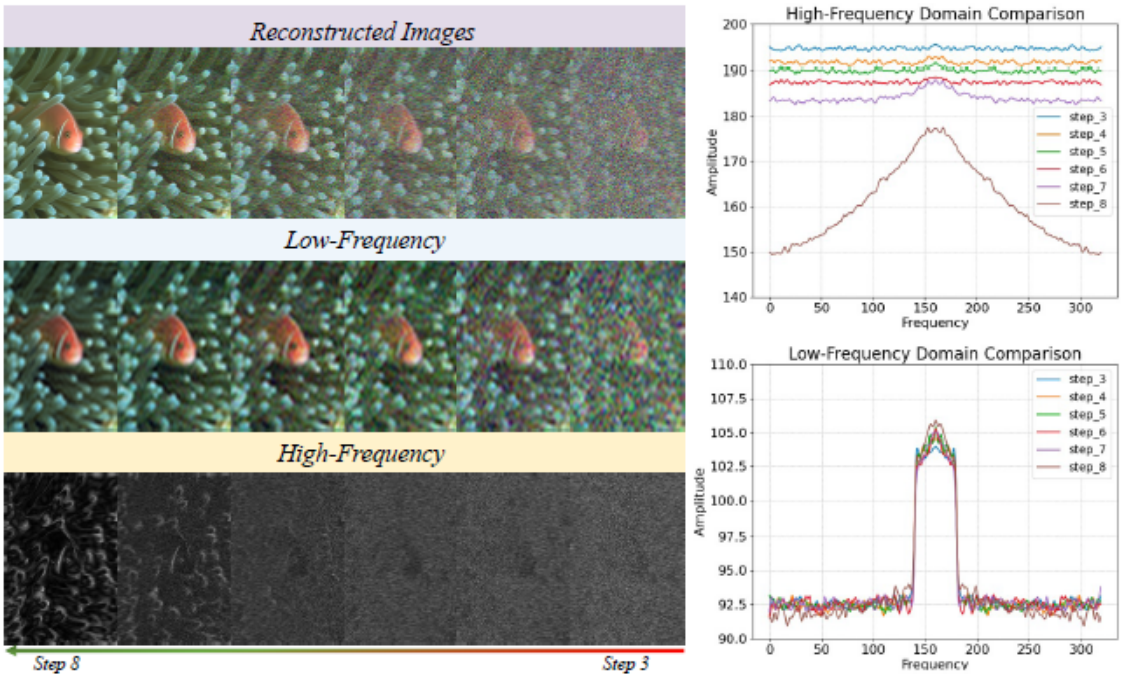
- We develop a Criss-Cross Diffusion Model (CrDiff) for *All-in-One* image restoration, which relies only on the input image for clean image recovery, without the need for prior knowledge of the image's degradation.
- We introduce a High-frequency Enhancement Network and High-frequency Fusion Block, which enables our model to adaptively utilize high-frequency features to restore the vivid details of the image.
- We conduct extensive experiments to confirm our model's superior performance across a spectrum of image restoration tasks, including image denoising, deraining, and dehazing, using a single, unified model.

5.2 Method

5.2.1 Motivation

Reason for Using Diffusion Models. The recent Diffusion-based Image Restoration methods have achieved satisfactory results in the field of image reconstruction, due to the fact that the diffusion model introduces different levels of noise during the training process to simulate various degradation processes, such as rain layer, haze, and so on. This multi-layer noise not only simulates real-world degradation, but also allows the model to learn to recover images under different degrees of distortion and damage, which largely increases the generalisation of the model. Instead of completely removing all the noise at once, the model gradually reduces the noise level during the inverse iteration process.

Limitations of Diffusion Models in Reconstruct Tasks. Through the analysis of Fig.5.1, we found that there is an obvious asynchrony phenomenon in the reconstruction process of low-frequency information and high-frequency information in the diffusion model in the image reconstruction task. Specifically, the reconstruction of low-frequency information by the model shows a trend of gradual clarity, gradually transitioning from a blurred state to clear details. However, the reconstruction of high-frequency information shows significant fluctuations, and its detail recovery is mainly concentrated in the final stage of the reconstruction process. This delayed recovery of high-frequency information indicates that the model has a weak ability to capture high-frequency details in the early stages, making



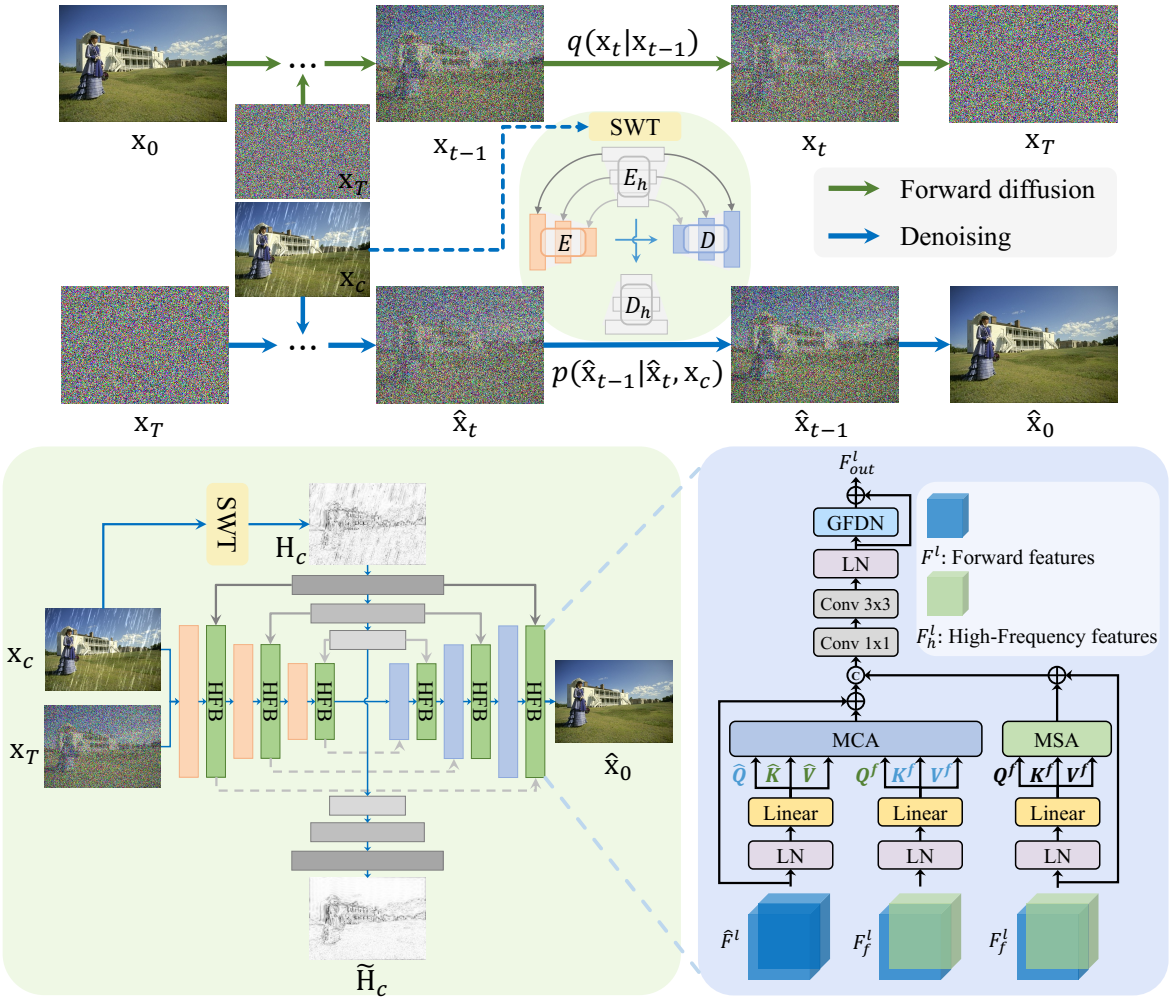
(a) (b)

Figure 5.1: (a) The denoising process of the Vanilla diffusion model, shown over 8 steps, visualizes the final 6 iterations. The top row illustrates the iterative denoising progression, while the low- and high-frequency components, separated via Stationary Wavelet Transform (SWT), are displayed in the second and third rows. The low-frequency component quickly restores the image's color and overall structure with minimal variation, while the high-frequency component exhibits more noticeable changes throughout the denoising process. (b) The spectrograms show changes in low- and high-frequency components during reconstruction. In the low-frequency plots, the waveform proximity indicates consistent recovery of the overall structure across steps, reflecting the model's stability. In contrast, the high-frequency plots display greater variability, suggesting significant differences in the model's ability to recover fine details throughout the process.

it difficult to achieve high-quality image reconstruction. Given that high-frequency information usually contains key details in the image and becomes a core factor affecting the reconstruction effect, we try to strengthen the reconstruction strategy for high-frequency information in the diffusion model, focusing on enhancing the model's early perception and processing capabilities of high-frequency information.

5.2.2 Criss-cross Diffusion

Our CrDiff is based on Denoising Diffusion Probabilistic Models and introduces an additional High-frequency Enhancement Network as a powerful auxiliary network, in addition, to ensure the perfect interaction of the information, we further design the High-frequency Fusion Block (HFB), the overall structure of CrDiff, as shown in Fig. 5.2 (a).



(a) **Cr**iss-cross **Diff**usion model (CrDiff) (b) High-frequency Fusion Block (HFB)

Figure 5.2: The upper part of the figure shows the diffusion as well as denoising process of the diffusion model. (a) shows the Criss-cross Diffusion model (CrDiff) we designed, where x_c as the input degraded image is our reconstruction target. (b) Shows the internal structure of the High-frequency Fusion Block (HFB) we designed, and the operation flow of processing and fusing the high-frequency information to enhance the details and textures of the image.

(1) Denoising Diffusion Probabilistic Models

In the field of image restoration, Denoising Diffusion Probabilistic Models (DDPMs) provide a robust framework to handle various types of degradation. DDPM simulates image degradation through a gradual forward diffusion process that progressively transforms the clean image $\mathbf{x}_0$ into a noisy image $\mathbf{x}_T$, formalized into a Markov chain:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}). \tag{5.1}$$

Here, α_t is the noise level coefficient, which monotonically decreases with step t. Each step

of noise addition follows a Gaussian distribution, where $\mathbf{x}_{t-1}$ is the state of the image at the previous time step. The ultimate goal of the forward diffusion is to reach a noise state $\mathbf{x}_T$, which follows a standard Gaussian distribution $\mathcal{N}(0,\mathbf{I})$. The image at any intermediate step $\mathbf{x}_t$ can be obtained through its relation with the initial image $\mathbf{x}_0$, namely:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}). \tag{5.2}$$

Here, $\bar{\alpha}_t$ is the product of all noise scaling coefficients α_i up to time step t. Through this method, we can simulate the process from a clear image to a completely random noisy image. In the reverse process of DDPM, the aim is to reconstruct the original lossless image $\hat{\mathbf{x}}_0$. During the reverse process, we begin with a noisy image $\mathbf{x}_T$, which is generated in the forward process by gradually adding noise. The goal of the reverse process is to gradually remove these noise elements, ultimately restoring the clean image $\hat{\mathbf{x}}_0$. At time step t, the reverse process estimates the conditional probability distribution of $\mathbf{x}_{t-1}$, given by:

$$p(\hat{\mathbf{x}}_{t-1}|\hat{\mathbf{x}}_t, \mathbf{x}_c) = \mathcal{N}(\hat{\mathbf{x}}_{t-1}; \mu_\theta(\hat{\mathbf{x}}_t, t, \mathbf{x}_c), \Sigma_\theta(\hat{\mathbf{x}}_t, t, \mathbf{x}_c)). \tag{5.3}$$

Here, μ_θ is the conditional mean, while Σ_θ is the conditional variance, both of which are computed from a parameterized neural network. They take the degraded image $\mathbf{x}_c$ as an input, in conjunction with the current reconstruction image $\hat{\mathbf{x}}_t$, and together decide the state of reconstruction in the next step.

The noise $\epsilon_\theta(\hat{\mathbf{x}}_t, t, \mathbf{x}_c)$ predicted by the neural network is used to approximate the noise component of $\mathbf{x}_{t-1}$, and the reconstructed image is updated according to the following formula:

$$\hat{\mathbf{x}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\hat{\mathbf{x}}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\hat{\mathbf{x}}_t, t, \mathbf{x}_c)). \tag{5.4}$$

In this way, the reverse process iterates, each step gradually decreasing noise and restoring clearer image features, until finally estimating a clean image $\hat{\mathbf{x}}_0$.

(2) High-frequency Enhancement Network

High-frequency Encoder. In addressing the issue of high-frequency information loss during the diffusion process, we propose the introduction of an auxiliary guiding mechanism. This mechanism is embodied in the design of the High-frequency Encoder, denoted as E_h. The key feature of E_h is that its network structure mirrors that of the backbone encoder, facilitating the effective fusion of multi-scale information.

Contrasting with the backbone encoder, the input to E_h is not the image itself but the high-frequency information extracted from the image using the Stationary Wavelet Transform (SWT). This approach underscores our commitment to preserving and processing

high-frequency details that are crucial in certain image processing tasks. The SWT is advantageous due to its excellent boundary processing, spatio-temporal localization, and multi-scale analysis capabilities. It can effectively identify and analyze image details and structural information at different scales, making it particularly suitable for tasks such as image denoising, texture analysis, and image segmentation. When processing a 2D image, it is especially important to use the general formula of the wavelet transform for each colour channel of the image. The formula for the input degraded image $x_c(i, j)$ can be applied in this case. The formula is defined as follows:

$$\mathrm{C}_c(i, j) = \sum_m \sum_n F_1(m) F_2(n) \cdot x_c(i - 2m, j - 2n). \tag{5.5}$$

Here, $\mathrm{C}_c(i, j)$ denotes the wavelet coefficients post-transformation, with $F_1(m)$ and $F_2(n)$ representing the filters applied to the image. These filters can be low-pass filters (L), which capture the smooth part of the image, or high-pass filters (H), which capture the detailed part of the image. By combining different kinds of L and H, it is possible to obtain the finest horizontal and vertical high-frequency detail information HH_c, LH_c, HL_c, and low-frequency information LL_c for x_c.

The obvious advantage of this SWT operation is the ability to clearly distinguish between the various wavelet coefficients, which is essential for accurate image analysis and processing. To focus the attention of E_h on high-frequency information, we strategically discard the LL_c part. For the remaining high-frequency information $\forall c \in \{\mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}, c \in \mathbb{R}^{H \times W \times C}$, we concatenate along the channel dimension and obtain $\mathrm{H}_c \in \mathbb{R}^{H \times W \times 3C}$ as the input to E_h. After that, we employ the extractor E_h on H_c to derive layered feature representations F_h^l, where each layer l contributes to capturing distinct aspects of the high-frequency components. In summary, we design it in such a way that E_h serves as a bridge connecting the encoder E of the diffusion model backbone and the denoising decoder D.

High-frequency Decoder. In order to ensure the correct perception of high-frequency information by E_h during the training process, we designed the corresponding High-frequency Decoder denoted as D_h. This decoder is used only in the training phase. Specifically, by reconstructing the high-frequency details, D_h enables E_h to better tune its weights and parameters to accurately capture the key subtle features in the diffusion process. This not only enhances the overall model's ability to handle high frequency and detailed information, but also allows E_h to understand and encode this critical information more deeply through D_h during the training phase. The design optimises the quality of the latent spatial features used by E_h during diffusion. Abandoning the use of D_h in the inference phase maintains the efficiency and simplicity of the model, while considering the computational cost and

speed at runtime to maintain both efficiency and model performance. Similarly, the features obtained from E_h are connected to D_h in a skip connection to get the output $\tilde{H}_c$.

(3) High-frequency Fusion Block (HFB)

In the architecture of the diffusion model, we introduce a key component - the High-frequency Fusion Block (HFB), which aims to improve the ability of the model to preserve high frequency information. The design of this module allows the model to use high-frequency information in the latent space to enhance the transmitted features prior to dimension sampling, thereby improving the restoration of image details and textures. The specific structure of HFB is shown in Fig. 5.2 (b).

We add HFBs after each block in the backbone, denoted as $\{\text{HFB}^1, \text{HFB}^2, \ldots, \text{HFB}^L\}$. These modules correspond to the layers in the network $\{F^1, F^2, \ldots, F^L\}$. Each HFB^l is designed to fuse the output F^l from the diffusion model block at its corresponding layer l with the multi-scale high-frequency features F_h^l obtained from E_h. For each level l in the network, HFB^l intervenes before the dimension sampling, ensuring that the diffusion block F^l in the backbone network can be fully integrated with the high-frequency features F_h^l. The fusion operation of HFB can be expressed as $F_{out}^l = \text{HFB}^l(F^l, F_h^l)$. Here, F_{out}^l is the fused feature representation which is fed to the next layer of the network. In addition, our F_{out}^l features in E are also transmitted to the decoder D through the skip connection.

In the HFB, we concatenate the two components along the channel dimension, resulting in F_f^l. The process begins with layer normalization (LN), followed by linear projection to compute queries Q^f, keys K^f, and values V^f from F_f^l. Subsequently, we calculate the Multi-head Self-Attention (MSA):

$$\text{MSA}(F_f^l) = \text{SoftMax}(Q^f \cdot (K^f)^T) \times V^f. \tag{5.6}$$

Furthermore, we compute the Mutual-Cross-Attention (MCA) between F^l and F_f^l. Specifically, MCA enables each feature map to focus on another, extracting relevant contextual information from the spatial regions of the other. Since the value of dimension C changes during the generation of F_f^l, for alignment, we concatenate F^l with itself to obtain $\hat{F}^l \in \mathbb{R}^{H\times W\times 2C}$. Both F_f^l and $\hat{F}^l$ are processed through layer normalization (LN). Subsequently, linear projection is used to compute the corresponding Q^f, K^f, V^f, and $\hat{Q}$, $\hat{K}$, $\hat{V}$. The formulas for MCA are defined as follows:

$$\text{MCA}_1(F_f^l, \hat{F}^l) = \text{SoftMax}(Q^f \cdot (\hat{K})^T) \cdot \hat{V} \tag{5.7}$$

$$\text{MCA}_2(\hat{F}^l, F_f^l) = \text{SoftMax}(\hat{Q} \cdot (K^f)^T) \cdot V^f, \tag{5.8}$$

To maximize the preservation of useful information and save computational cost, we concatenate MCA_1 and MCA_2 and then compress their dimensions to the input size through a

1×1 convolution layer. The function is defined as:

$$\text{MCA}(\hat{F}^l, F_f^l) = \text{Conv}_{1\times1}[\text{MCA}_1; \text{MCA}_2], \tag{5.9}$$

Here, [;] denotes the concatenation operation. Then, these improved representations are integrated through a $\text{Conv}_{1\times1}$ convolution, reducing their dimensions. Finally, a $\text{Conv}_{3\times3}$ operation refines these features in the Gated Depthwise Convolutional feed-forward Network (Gated-Dconv feed forward network, GDFN) [179], ensuring precise and controlled transformations in the restoration task.

5.2.3 Training Stage

In Section 5.2.2, the objective function for the vanilla diffusion model (θ), is concisely defined as follows:

$$\mathbf{L}_{\text{diff}}(\theta) = \mathbb{E}_{q(\mathbf{x}_{0:T}|\mathbf{x}_c)}\Big[\sum_{t=1}^{T} ||\epsilon - \epsilon_\theta(\hat{\mathbf{x}}_t, t, \mathbf{x}_c)||_1 + ||\hat{\mathbf{x}}_0 - \mathbf{x}_{gt}||_1\Big] \tag{5.10}$$

The first term $\sum_{t=1}^{T} ||\epsilon - \epsilon_\theta(\hat{\mathbf{x}}_t, t, \mathbf{x}_c)||_1$ measures the MAE loss between the predicted and actual noise. The second term $||\hat{\mathbf{x}}_0 - \mathbf{x}_{gt}||_1$ quantifies the MAE loss between the reconstructed image and the ground truth $\mathbf{x}_{gt}$.

Furthermore, we also introduce a detail reconstruction loss. Specifically, we perform the same SWT operation on $\mathbf{x}_{gt}$ and concatenate the obtained high-frequency information along the channel dimensions to obtain H_{gt}, after which we compute the MSE loss by applying the outputs $\tilde{\text{H}}_c$ obtained from the High-frequency Enhancement Network (ϕ), the formula for this loss is:

$$\mathscr{L}_{\text{detail}}(\phi) = \text{MSE}(\tilde{\text{H}}_c, \text{H}_{gt}) = \frac{1}{N}\sum_{i=1}^{N}(\tilde{\text{H}}_c^{(i)} - \text{H}_{gt}^{(i)})^2. \tag{5.11}$$

Ultimately, the total optimization objective combines the high-frequency information loss $\mathscr{L}_{\text{detail}}(\phi)$ and the diffusion model loss $\mathscr{L}_{\text{diff}}(\theta)$. The combined loss function is defined as:

$$\mathscr{L}_{\text{total}}(\theta, \phi) = \mathscr{L}_{\text{diff}}(\theta) + \gamma \cdot \mathscr{L}_{\text{detail}}(\phi). \tag{5.12}$$

γ is a trade-off coefficient that balances the effects of the two loss functions.

5.3 Experiments

5.3.1 Setup

Datasets. We prepare datasets for different restoration tasks, following closely the prior work [119]. For image denoising in the single-task setting, we use a combined set of BSD400 [3]

and WED [103] datasets for training. The BSD400 dataset contains 400 training images and the WED dataset has 4,744 images. From clean images of these datasets, we generate the noisy images by adding Gaussian noise with different noise levels $\sigma \in \{15, 25, 50\}$. Testing is performed on BSD68 [105] and Urban100 [54] datasets. For single-task image deraining, we use the Rain100L [156] dataset, which consists of 200 clean-rainy image pairs for training, and 100 pairs for testing. For image dehazing in the single-task setting, we utilize SOTS [76] dataset that contains 72,135 training images and 500 testing images. Finally, to train a unified model under the All-in-One setting, we combine all four aforementioned datasets and train a single model that is later evaluated on multiple tasks.

Schedules. In this study, we have set the training period for the model to 400 epochs , with an image training size of 224 × 224 pixels. The training is conducted on two NVIDIA A40 GPUs, processing 48 images per batch. The AdamW optimizer is adopted for optimization. The batch size is set to 12 and $\gamma = 0.5$.

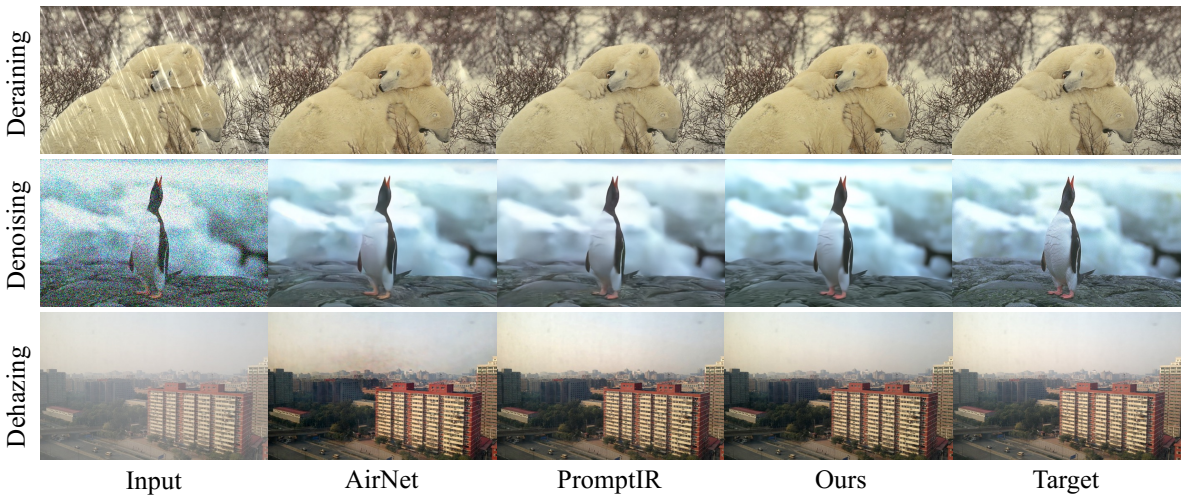


Figure 5.3: Deraining, Denoising and Dehazing results for All-in-One methods. Our method effectively reconstructs clean images of all degradation types.

5.3.2 Comparison Experiment

The robustness of our proposed CrDiff is tested on three core image restoration tasks: denoising, dehazing and deraining. Following the protocols established in [74], our evaluation framework is divided into two distinct approaches: the comprehensive All-in-One strategy, where a single model is tasked with correcting all types of image degradation, and the dedicated single-task strategy, where unique models are tailored to each specific restoration challenge. Best results are denoted in red and the second best results are denoted in blue.

Method	Dehazing on SOTS	Deraining on Rain100L	Denoising on BSD68 dataset $\sigma = 15$	$\sigma = 25$	$\sigma = 50$	Average
DL [37]	26.92 / 0.391	32.62 / 0.931	33.05 / 0.914	30.41 / 0.861	26.90 / 0.740	29.98 / 0.875
LPNet [40]	20.84 / 0.828	24.88 / 0.784	26.47 / 0.778	24.77 / 0.748	21.26 / 0.552	23.64 / 0.738
BRDNet [143]	23.23 / 0.895	27.42 / 0.895	32.26 / 0.898	29.76 / 0.836	26.34 / 0.836	27.80 / 0.843
FDGAN [34]	24.71 / 0.924	29.89 / 0.933	30.25 / 0.910	28.81 / 0.868	26.43 / 0.776	28.02 / 0.883
MPRNet [181]	25.28 / 0.954	33.57 / 0.954	33.54 / 0.927	30.89 / 0.880	27.56 / 0.779	30.17 / 0.899
AirNet [74]	27.94 / 0.962	34.90 / 0.967	33.92 / 0.933	31.26 / 0.888	28.00 / 0.797	31.20 / 0.910
PromptIR [119]	30.58 / 0.974	36.37 / 0.972	33.98 / 0.933	31.31 / 0.888	28.06 / 0.799	32.06 / 0.913
CrDiff (Ours)	31.24 / 0.981	40.37 / 0.986	34.35 / 0.941	32.12 / 0.903	29.02 / 0.815	33.42 / 0.925

Table 5.1: Comparisons under All-in-One restoration setting: single model trained on a combined set of images originating from different degradation types. Our method achieves the best results on all three representative image restoration tasks, especially on the Deraining task where the value of PSNR directly increases by 3 dB.

(1) Multiple Degradation All-in-One Results

In order to evaluate the performance of CrDiff, we carefully compare it with the latest research in the current field PrompIR [119] and the classical approach AirNet [74] as well as involving other related approaches such as BRDNet [143], LPNet [40], FDGAN [34], MPRNet [181], to ensure the comprehensiveness of the assessment. In the dehazing task, CrDiff outperforms PrompIR (30.58/0.974) and AirNet (27.94/0.962) on the SOTS dataset [76] (PSNR of 31.24, SSIM of 0.981). For deraining, CrDiff achieves a PSNR of 40.37 and an SSIM of 0.986 on the Rain100L dataset [37], significantly outperforming PrompIR (36.37/0.972) and AirNet (34.90/0.967). This significant improvement demonstrates the superiority of CrDiff in removing raindrops and recovering image details. For the denoising task on the BSD68 dataset [105], CrDiff outperforms PrompIR and AirNet at all noise levels ($\sigma = 15, 25, 50$), especially at higher noise levels ($\sigma = 50$). The PSNR of CrDiff is 29.02 and the SSIM is 0.815, which are much higher than the other methods. This result highlights CrDiff's ability to maintain image quality and detail in high noise environments, as well as its robustness and adaptability to different denoising tasks.

(2) Single Degradation One-by-One Results

Denoising. In a single-task denoising experiment on the BSD68 dataset [105], our CrDiff network particularly highlights its performance advantages in a high-noise environment. For the noise level $\sigma = 50$, CrDiff achieves a PSNR of 29.51 and an SSIM of 0.834, significantly higher than other methods such as PromptIR [119]. This significant improvement stems from CrDiff's focus on high-frequency detail recovery, which is especially critical in high-noise environments. In contrast, at low noise levels ($\sigma = 15$ and $\sigma = 25$), while CrDiff still

Method	PSNR ↑ / SSIM ↑		
	$\sigma = 15$	$\sigma = 25$	$\sigma = 50$
DnCNN [190]	33.89 / 0.930	31.23 / 0.883	27.92 / 0.789
IRCNN [191]	33.87 / 0.929	31.18 / 0.882	27.88 / 0.790
FFDNet [192]	33.87 / 0.929	31.21 / 0.882	27.96 / 0.789
BRDNet [143]	34.10 / 0.929	31.43 / 0.885	28.16 / 0.794
AirNet [74]	34.14 / 0.936	31.48 / 0.893	28.23 / 0.806
Restormer [179]	34.29 / 0.937	31.64 / 0.895	28.41 / 0.810
PromptIR [119]	34.34 / 0.938	31.71 / 0.897	28.49 / 0.813
CrDiff (Ours)	34.45 / 0.938	32.32 / 0.911	29.51 / 0.834

Table 5.2: Quantitative comparison of our method and other state-of-the-art methods on denoising dataset BSD68.

Method	PSNR ↑	SSIM ↑
DIDMDN [185]	23.79	0.773
UMR [175]	32.39	0.921
SIRR [157]	32.37	0.926
MSPFN [59]	33.50	0.948
LPNet [40]	33.61	0.958
AirNet [74]	34.90	0.977
Restormer [179]	36.74	0.978
PromptIR [119]	37.04	0.979
CrDiff (Ours)	38.64	0.988

Table 5.3: Evaluating Deraining task on Rain100L dataset. Our CrDiff achieving top results of 38.64 in PSNR and 0.988 in SSIM, highlighting its advanced performance in deraining scenarios.

leads, the advantage is smaller. This is because the need for high-frequency detail recovery is less urgent at low noise than at high noise.

Dehazing. In the dehazing task, although CrDiff marginally outperforms with a PSNR of 31.85 and an SSIM of 0.977, its advantage over PromptIR [119] (31.31/0.973) is relatively modest. This can primarily be attributed to the specific demands of dehazing tasks, which prioritize accurate restoration of color and brightness, diverging slightly from CrDiff's forte in high-frequency detail recovery.

Deraining. In the task of rain removal, CrDiff demonstrated a significant performance improvement, achieving a PSNR of 38.64 and an SSIM of 0.988, substantially surpassing

Method	PSNR ↑	SSIM ↑
AODNet [75]	20.29	0.877
EPDN [121]	22.57	0.863
FDGAN [34]	23.15	0.921
AirNet [74]	23.18	0.900
Restormer [179]	30.87	0.969
PromptIR [119]	31.31	0.973
CrDiff (Ours)	31.85	0.977

Table 5.4: Quantitative results on the SOTS dataset in terms of PSNR and SSIM.

other methods such as PromptIR [119], which achieved a PSNR of 37.04 dB and an SSIM of 0.979. This significant improvement is due to CrDiff's effective handling of high-frequency details, particularly important for capturing rain layer characteristics.

5.3.3 Ablations Studies

Impact of Stationary Wavelet Transform (SWT): We verify the effectiveness of the SWT operation on the deraining task and test the impact of the high-frequency enhancement network on different inputs. As shown in Table 5.5, allowing E_h to learn high-frequency information as targeted guidance achieves the best results.

Impact of High-Frequency Enhancement Network: In our ablation study, detailed in Table 5.6, we demonstrate the impact of high-frequency components in image restoration tasks. The basic model serves as a baseline, while the addition of the High-Frequency Encoder (E_h) significantly enhances performance. The incorporation of the High-Frequency Decoder (D_h) further improves results, achieving top metrics in PSNR and SSIM. This highlights the importance of each component in refining our model's capability to restore images with high precision, especially in preserving high-frequency details.

CrDiff	PSNR↑/SSIM↑
w.o SWT	37.92 / 0.977
w. SWT_{full}	38.43 / 0.986
w. $\text{SWT}_{\text{high-freq}}$	**38.64 / 0.988**

Table 5.5: Performance comparison of SWT on deraining dataset Rain100L. The best results are highlighted in bold.

Basic	E_h	D_h	**Deraining** PSNR↑ / SSIM↑	**Dehazing** PSNR↑ / SSIM↑	**Denoising** PSNR↑ / SSIM↑
✓			36.53 / 0.969	29.87 / 0.970	26.58 / 0.745
✓	✓		37.98 / 0.982	31.02 / 0.971	28.89 / 0.928
✓	✓	✓	**38.64 / 0.988**	**31.85 / 0.977**	**29.51 / 0.834**

Table 5.6: Quantitative ablation results illustrating the efficacy of the High-frequency encoder (E_h) and decoder (D_h) in CrDiff. The best results are highlighted in bold.

CrDiff	PSNR↑/SSIM↑
w.o HFB	37.12 / 0.979
w. HFBs in E_h	37.58 / 0.981
w. HFBs in D_h	38.33 / 0.985
w. HFBs in E_h & D_h	**38.64 / 0.988**

Table 5.7: Performance comparison of CrDiff with(w.) / without(w.o) the High-frequency Fusion Block on deraining dataset Rain100L. The best results are highlighted in bold.

Impact of High-frequency Fusion Block (HFB): From the results shown in Table 5.7, it can be seen that adding HFBs throughout delivers the best results.

5.4 Conclusion

In summary, our Criss-Cross Diffusion (CrDiff) model demonstrates robust and exceptional capabilities in key image restoration tasks such as denoising, dehazing and deraining. The success of the model is largely due to our innovative design of the High-Frequency Enhancement Network and the High-Frequency Fusion Block, which effectively exploit high-frequency information. These components enable CrDiff to achieve state-of-the-art results, particularly in challenging conditions with significant noise and degradation.

CHAPTER 6

UNIFIED ADVERSE WEATHER REMOVAL VIA META-LEARNING AND DOMAIN-AWARE NORMALIZATION

In the previous chapter, we presented CrDiff, which demonstrated significant advancements in All-in-One reconstruction tasks. However, a significant limitation of current methods, including CrDiff, is their suboptimal performance in real-world scenarios, primarily due to the domain discrepancy between synthetic and real-world data. To address this issue, we propose a dual-branch network with an additional self-supervised learning (SSL) branch for the unified adverse weather removal task. The SSL branch employs a pair-free self-supervised approach, enabling the extraction of weather-specific features from unpaired data. We further enhance the training process using a meta-learning-based bi-level optimization method, aligning the objectives of the auxiliary SSL branch with the reconstruction branch. To mitigate knowledge interference and avoid instability caused by uniform parameter updates in meta-learning, we update only the affine parameters of the Batch Normalization (BN) layers, which capture domain-specific information. Our proposed Test-time Weather Adaptation (TT-WA) method outperforms state-of-the-art techniques in restoring videos affected by various adverse weather conditions and demonstrates strong generalization to unseen weather scenarios in both synthetic and real-world settings.

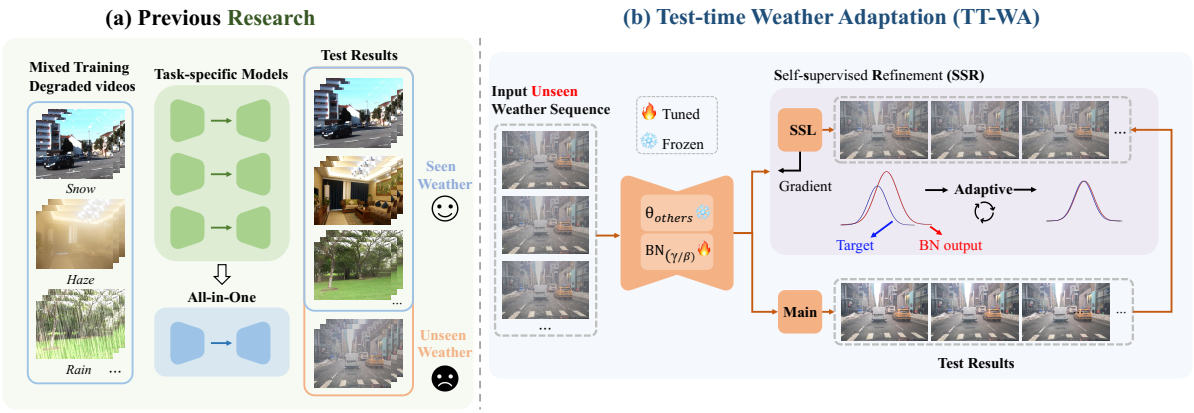


Figure 6.1: **(a) Previous Research:** Current studies on adverse weather conditions such as rain, haze, and snow mainly use task-specific models, each targeting a single type of weather degradation. While effective for known conditions, these models perform poorly under unseen scenarios. All-in-one approaches attempt to handle multiple conditions with one model but still struggle with unknown weather adaptations. **(b) Proposed Test-Time Weather Adaptation (TT-WA) Method:** Our novel TT-WA method inputs unseen weather sequences during testing and uses self-supervised loss (SSL) for online gradient updates on batch normalization layers' affine parameters (γ and β). The Iterative Self-Refinement (ISR) (***ISR***) method iteratively optimizes the input sequence, enhancing the model's adaptation to various weather degradations. This approach ensures high efficiency and stability across diverse and complex weather conditions, as shown by the dynamic adjustments in the figure.

6.1 Introduction

Adverse weather conditions, such as rain, snow, and haze, occur in outdoor videos and significantly degrade visibility [46]. Such degradation critically impairs the performance of vision applications, including object detection [69, 94], semantic segmentation [118, 145], and autonomous driving [106, 182]. A significant number of research efforts aim to mitigate the impact of adverse weather conditions; however, the majority primarily focus on addressing single-type weather conditions. Such as dehazing [33, 97, 127], deraining [151, 170, 177], and desnowing [19, 189] in images and videos. Although above methods perform well in specific domains, most typically handle only one type of weather. Thus, deploying these methods on edge platforms is challenging due to the need for multiple models. To overcome the single-weather focus, recent research shifts towards developing unified frameworks to remove various adverse weather conditions [20, 80, 114, 146, 172, 212]. Li *et al.* [80] propose an All-in-One network that removes various adverse weather conditions from images, offering the comprehensive solution. Under the expanding success of diffusion models in numerous domains, Ozdenizci *et al.* [114] apply a patch-based diffusion model for image restoration

across different weather conditions. Next, research progresses to the video domain. Yang *et al.* [172] develop ViWS-Net, a framework that restores videos affected by adverse weather using a video transformer encoder and long short-term temporal modeling.

However, despite the above methods performing well under synthetic adverse weather conditions, their reconstruction effectiveness in **real-world scenarios** is often unsatisfactory due to the limitations of synthetic data and its differences from real data. Existing datasets mainly consist of synthetic samples, which fail to adequately represent the complexity of real-world weather conditions. Moreover, these methods lack sufficient **adaptation strategies** to handle the diversity and unpredictability of real data. Consequently, when these models encounter weather conditions outside their training data distribution in practical applications, their performance is often unsatisfactory. Therefore, the issue we aim to address is ***how to enhance model adaptability to real-world adverse weather conditions***.

To tackle this challenge, we utilize a **Test-time Weather Adaptation (TT-WA)** with meta-learning, which dynamically updates the network via self-supervised learning for adaptation to unseen weather conditions. However, updating the entire network with limited unseen weather domain data can destabilize the model's reconstruction ability during adaptation. Inspired by the research in Knowledge Model Editing, which aims to update specific knowledge without affecting the overall model performance. We decide to decouple the domain adaptation ability from the reconstruction ability in our model, updating only the domain-specific knowledge. According to the results presented by Li et al. [86], domain-specific knowledge is embedded in the affine coefficients (γ, β) of the Batch Normalization (BN) layers. Therefore, we decide to only update the γ, β from BN layer and keep the rest of the parameters frozen during TT-WA. It supports the model to achieve more natural domain adaptation to unseen weather conditions. The overall framework and its comparison with previous approaches are illustrated in Fig. 6.1.

Moreover, we utilize a **Meta-BN training stage**. Specifically, we utilize a bi-level optimization, treating each weather condition as a separate task and training them sequentially. Importantly, we only update the γ, β from BN layer and keep the rest of the parameters frozen during training. This operation ensures that the model learns weather domain adaptation, while also retaining previously learned reconstruction knowledge during successive task learning. Finally, we comprehensively test our TT-WA method under synthetic single weather scenarios, such as rain, haze, and snow. Additionally, we evaluate our method under complex real world weather conditions. The experimental results indicate that our method achieves SOTA performance in both qualitative and quantitative results.

Our contributions can be summarized as:

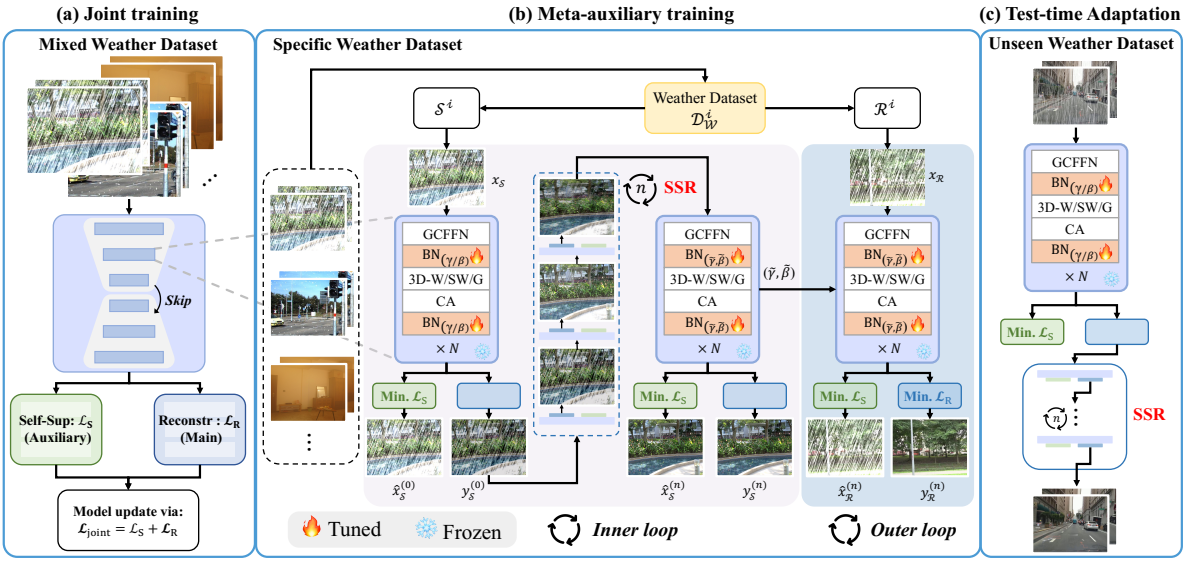


Figure 6.2: **Unified Framework for Adverse Weather Removal:** This figure showcases the primary stages of our approach: **(a) Joint Training** on a mixed weather dataset develops a generalized model, **(b) Meta-BN Training** involves rapid model adaptation through Iterative Self-Refinement (ISR) and Domain-aware Batch Normalization (BN), which fine-tunes affine parameters (γ, β) to enhance robustness to specific weather conditions, and **(c) Test-time Adaptation** dynamically adjusts to unseen weather scenarios by optimizing self-supervised loss ($\mathscr{L}_s$). Details include the dual-branch network structure, emphasizing efficient and targeted adaptation across varying weather conditions.

- We develop a Test-time Weather Adaptation (TT-WA) framework that dynamically adjusts the model during inference by updating only the BN affine parameters. This approach effectively adapts to unseen weather conditions, maintains model stability, and ensures computational efficiency.
- We introduce a Meta-BN training stage that selectively updates only the affine parameters of the Batch Normalization (BN) layers. This adjustment enhances model generalization and robustness by focusing on domain-specific knowledge while preserving the pre-trained reconstruction capabilities.
- We extensively evaluate our approach on both synthetic and real-world weather conditions, demonstrating superior performance in both qualitative and quantitative metrics, confirming its effectiveness and generalizability.

6.2 Method

6.2.1 Overview

In this paper, we apply a **Test-time Weather Adaptation (TT-WA)** method with **Meta-BN training** to obtain a robust weather model and an effective adaptation mechanism. During

the training process, we utilize **M** types of degraded weather conditions, denoted as $\{\mathscr{D}_{\boldsymbol{W}}^{\boldsymbol{i}}\}_{i=1}^{\mathbf{M}}$. Each weather condition $\mathscr{D}_{\boldsymbol{W}}^{\boldsymbol{i}}$ contains a set of ground truth data, i.e., $\mathscr{D}_{\boldsymbol{W}}^{\boldsymbol{i}} = (\mathbf{x}_W^i, \mathbf{y}_W^i)$, where $\mathbf{x}$ and $\mathbf{y}$ represent input degraded sequences and corresponding clean sequences, respectively. For the inference step TT-WA, we define the input unseen weather dataset as $\mathscr{D}_{\boldsymbol{U}}$. Our objective is to adapt the pre-trained model to this input domain using a limited amount of unseen weather data. Next, we will provide a detailed explanation of ***Model architecture***, and ***Training strategy*** in sequence. Our approach is shown in Fig. 6.2.

6.2.2 Model Architecture

To effectively extract spatial and temporal information and enhance weather adaptation, we design a **Dual-branch Network** and **Enhanced Transformer Blocks**. Next, we provide a detailed explanation.

i) Dual-branch Network: We employ a four-level symmetrical U-Net style encoder-decoder network. After the decoder, the network splits into two branches, each composed of multiple residual modules and do not share parameters. The first branch is defined as the ***Reconstruction branch***; it aims to reconstruct clean backgrounds from input degraded sequences. For the second branch, termed the ***Adaptive branch***, it employs self-supervised learning (SSL) to reconstruct the degraded input sequences. The objective is to enhance the model's understanding of the intrinsic structure of the data, thereby improving its generalization capabilities when dealing with various types of unknown weather.

ii) Enhanced Transformer Blocks: In the encoder-decoder network, each level comprises N Enhanced transformer blocks. Specifically, each block consists of 3D window-based, shifted-window, and global multi-head self-attention mechanisms (3D-W/SW/G-MSA), combined with a channel-attention (CA) module.

For the feedforward component, we utilize Gated Convolutional Feedforward Networks (GCFFN). Moreover, we utilize 3D Batch Normalization (3D-BN) instead of Layer Normalization (LN). Additionally, residual skip connections are integrated into the network to enhance information flow and gradient propagation. The process is as follows:

$$\begin{aligned} \mathbf{x}' &= \text{3D-W/SW/G-MSA}(\text{BN}(\mathbf{x})) + \mathbf{x}, \\ \mathbf{x}'' &= \text{CA}(\text{BN}(\mathbf{x}')) + \mathbf{x}', \\ \mathbf{y} &= \text{GCFFN}(\text{BN}(\mathbf{x}'')) + \mathbf{x}'', \end{aligned} \tag{6.1}$$

where $\mathbf{x}$ and $\mathbf{y}$ represent the input and output of each transformer block, respectively. Based on the above architecture, our model effectively integrates sophisticated feature extraction with adaptive mechanisms to ensure robust performance across varying weather conditions.

6.2.3 Training Strategy

Our training strategy consists of two stages: ***Joint training*** and ***Meta-BN training***. The Joint training aims to provide a robust pre-trained model for degraded weather reconstruction by training on various degraded conditions. The Meta-BN training enhances the model's adaptive capabilities, ensuring effective generalization to unseen weather conditions. Next, we will describe the details of each stage.

i) Joint training. We perform large-scale training by aggregating all data from $\{\mathscr{D}_{\boldsymbol{W}}^{\boldsymbol{i}}\}_{i=1}^{\mathbf{M}}$ and uniformly sampling mini-batches, which ensures the model's foundational reconstruction capabilities. The Reconstruction branch utilizes paired data to perform reconstruction loss $\mathscr{L}_{\boldsymbol{Rec}}$. In contrast, the adaptive branch employs the input degraded sequences for self-supervised learning to provide the SSL loss $\mathscr{L}_{\boldsymbol{SSL}}$. The above losses are combined into the joint loss $\mathscr{L}_{\boldsymbol{Joint}}$ as follows:

$$\mathscr{L}_{\boldsymbol{Joint}} = \mathscr{L}_{\boldsymbol{Rec}} + \lambda \mathscr{L}_{\boldsymbol{SSL}}. \tag{6.2}$$

Both $\mathscr{L}_{\boldsymbol{Rec}}$ and $\mathscr{L}_{\boldsymbol{SSL}}$ are L1 losses, balanced by the parameter λ.

ii) Meta-BN training. We employ a bi-level optimization approach in the Meta-BN training, which consists of an ***Inner loop*** and an ***Outer loop***. Furthermore, we apply each weather domain data $\mathscr{D}_{\boldsymbol{W}}^{\boldsymbol{i}}$ from the training set, dividing it into two subsets: support set $\mathscr{S}^{\boldsymbol{i}}$. It contains degraded videos for self-supervised training following the meta-learning strategy. In contrast, another subset of the data is designated as the reference set $\mathscr{R}^{\boldsymbol{i}}$, which is used to refine the model's reconstruction capability. For the ***1) Inner loop***, we apply mini-batches $x_{\mathscr{S}}$ from $\mathscr{S}^i$ for self-supervised learning to adapt to the current weather conditions. Importantly, we only update the affine coefficients (γ, β) from BN, while freezing the model's weight matrix to preserve the learned comprehensive feature representations and prevent disruption. The equation is as follows:

$$(\tilde{\gamma}, \tilde{\beta}) = (\gamma, \beta) - \eta_1 \nabla_{(\gamma,\beta)} \mathscr{L}_{\boldsymbol{SSL}}(\mathscr{S}^i; \theta, (\gamma, \beta)), \tag{6.3}$$

where θ denotes all the weight matrices, and η_1 is the learning rate for the inner loop. This method allows us to quickly update the model to adapt to the current weather conditions.

Within the Inner loop, we introduce an **I**terative **S**elf-**R**efinement (ISR) design, which aims to ensure that the model not only accurately reconstructs different degraded weather conditions but also effectively handles varying degrees of degradation. Specifically, we feed the initial input sequence $x_{\mathscr{S}}$ into the network, generating two outputs: the reconstructed input sequence $\hat{x}_{\mathscr{S}}^{(0)}$ and the predicted clean sequence $y_{\mathscr{S}}^{(0)}$. In this process, $\hat{x}_{\mathscr{S}}^{(0)}$ is employed for self-supervised learning, while $y_{\mathscr{S}}^{(0)}$ is fed back into the network for further refinement.

Through multiple iterations, the reconstruction quality is progressively enhanced, enabling the model to better handle varying levels of degradation. This process can be described as:

$$y_{\mathscr{S}}^{(n)} = f(y_{\mathscr{S}}^{(n-1)};\theta), \tag{6.4}$$

where $y_{\mathscr{S}}^{(0)}$ is the initial reconstructed sequence and $y_{\mathscr{S}}^{(n)}$ is the sequence after the n-th iteration. For the ***2) Outer loop***, we utilize the parameters $(\tilde{\gamma},\tilde{\beta})$ obtained from the inner loop to adapt to the dataset $\mathscr{R}^{i}$, enhancing the performance of the Reconstruction branch. Specifically, this process aims to ensure that the adapted model can generalize across the entire dataset, thereby ensuring the robustness of the primary reconstruction task:

$$(\gamma,\beta) := (\gamma,\beta) - \eta_2 \nabla_{(\gamma,\beta)} \mathscr{L}_{\boldsymbol{Joint}}(\mathscr{R}^{i};\theta,(\tilde{\gamma},\tilde{\beta})). \tag{6.5}$$

$\mathscr{L}_{\boldsymbol{Joint}}$ is as defined in the Eq. 6.2, and η_2 is the learning rate for the outer loop. Through this iterative refinement, the model progressively enhances the quality of the reconstructed clean sequences and adapts to varying weather conditions. Alternating between Inner loop and Outer loop ensures the model's robustness and generalization across diverse scenarios.

6.2.4 Test-time Weather Adaptation (TT-WA)

Based on the update design of Meta-BN training, we apply TT-WA and adopt a more efficient testing method. Specifically, we suppose the input unseen video is $\mathscr{D}_{\boldsymbol{U}}$, which is divided into T consecutive sequences. First, the t-th sequence, denoted as $x_{\mathscr{U},t}$, and the BN affine parameters (γ,β) are updated as follows:

$$(\gamma,\beta)^{(t+1)} = (\gamma,\beta)^{(t)} - \eta_3 \nabla_{(\gamma,\beta)} \mathscr{L}_{\boldsymbol{SSL}}(x_{\mathscr{U},t};\theta,(\gamma,\beta)^{(t)}), \tag{6.6}$$

where, $t = 1,2,\ldots,T-1$ and η_3 is the learning rate. It is important to note that once the $(t+1)$-th sequence is input into the network, the updated affine coefficients $(\gamma,\beta)^{(t+1)}$ are used, avoiding the need for re-inputting the t-th sequence.

During the TT-WA process, Iterative Self-Refinement (ISR) is also employed. The initial input sequence undergoes self-refinement, denoted as:

$$y_{\mathscr{U},t}^{(0)} = x_{\mathscr{U},t}, \quad t = 1,2,\ldots,T \tag{6.7}$$

$$y_{\mathscr{U},t}^{(n+1)} = f(y_{\mathscr{U},t}^{(n)};\theta), \quad n = 0,1,\ldots,N-1 \tag{6.8}$$

where $y_{\mathscr{U},t}^{(0)}$ is the initial state of the t-th sequence before refinement and $y_{\mathscr{U},t}^{(n+1)}$ is the refined sequence after n iterations. This approach effectively adapts to input videos under different weather conditions, ensuring that the model can quickly and adaptively adjust during testing, maintaining high-quality reconstruction.

6.3 Experiments

6.3.1 Datasets

Multiple video datasets containing adverse weather conditions were utilized in our experiments. Following the setting of [172], we adopt RainMotion [151], REVIDE [194] and KITTI-snow [172] as seen weather. RainMotion is the latest video deraining dataset synthesized based on NTURains [177]. This dataset includes five large rain streak masks, each following natural motion trajectories, thereby better simulating realistic rainy scenes. REVIDE is the first real-world video dehazing dataset recorded under high-fidelity real hazy conditions for indoor scenes. The hazy scenes in this dataset are highly realistic and feature high resolution. KITTI-snow is a synthesized outdoor dataset comprising 50 videos, where snowflakes have varying properties and are processed with Gaussian blur, making the video desnowing task more challenging. Furthermore, we assess the performance of our **TT-WA** on two datasets, VRDS [160] and RVSD [15], to demonstrate its robustness and generalization to various unseen weather conditions. VRDS is a synthesized video dataset of joint rain streaks and raindrops with a total of 102 videos, while RVSD is a realistic video desnowing dataset with a total of 110 videos containing both snow and fog achieved by the rendering engine. Additionally, we collect a variety of real-world videos affected by different weather conditions to demonstrate the effectiveness of our approach in practical applications.

6.3.2 Implementation

The proposed framework is trained using NVIDIA RTX 4090 GPUs and implemented on the PyTorch platform. To ensure robustness and effectiveness, we randomly crop the video frames to a resolution of 256×256 pixels for training. The batch size is set to 9, with each batch containing an average of three extreme weather scenarios, each scenario including 4 consecutive frames. This configuration allows the model to learn temporal dependencies effectively. For optimization, we employ the AdamW optimizer. For the Joint training, we combine the training sets of the three datasets to form a mixed set, which is used to learn a generic model. The learning rate is set to 1×10^{-4} and the parameter λ is set to 0.1 to balance the loss components and enhance the training process. We employ Meta-BN training to alternately train on the three datasets. The learning rate η_1 is set to 3×10^{-4} for the inner loop, while η_2 is set to 3×10^{-5} for the outer loop. For the TT-WA, η_3 is also set to 3×10^{-4}. Additionally, for the Iterative Self-Refinement, we set the number of iterations n to 3, allowing the model to progressively refine its predictions.

	Method	Type	Source	Datasets									
				Original Weather		Rain		Haze		Snow		Average	
Derain	**PReNet** [123]	Image	CVPR'19	27.06	0.9077	26.80	0.8814	17.64	0.8030	28.57	0.9401	24.34	0.8748
	SLDNet [170]	Video	CVPR'20	20.31	0.6272	21.24	0.7129	16.21	0.7561	22.01	0.8550	19.82	0.7747
	S2VD [177]	Video	CVPR'21	24.09	0.7944	28.39	0.9006	19.65	0.8607	26.23	0.9190	24.76	0.8934
	RDD-Net [151]	Video	ECCV'22	31.82	0.9423	30.34	0.9300	18.36	0.8432	30.40	0.9560	26.37	0.9097
Dehaze	**GDN** [95]	Image	ICCV'19	19.69	0.8545	29.96	0.9370	19.01	0.8805	31.02	0.9518	26.66	0.9231
	MSBDN [33]	Image	CVPR'20	22.01	0.8759	26.70	0.9146	22.24	0.9047	27.07	0.9340	25.34	0.9178
	VDHNet [127]	Video	TIP'19	16.64	0.8133	29.87	0.9272	16.85	0.8214	29.53	0.9395	25.42	0.8960
	PM-Net [97]	Video	MM'22	23.83	0.8950	25.79	0.8880	23.57	0.9143	18.71	0.7881	22.69	0.8635
Desnow	**DesnowNet** [96]	Image	TIP'18	28.30	0.9530	25.19	0.8786	16.43	0.7902	27.56	0.9181	23.06	0.8623
	DDMSNET [189]	Image	TIP'21	32.55	0.9613	29.01	0.9188	19.50	0.8615	32.43	0.9694	26.98	0.9166
	HDCW-Net [19]	Image	ICCV'21	31.77	0.9542	28.10	0.9055	17.36	0.7921	31.05	0.9482	25.50	0.8819
Desnow	**MPRNet** [180]	Image	CVPR'21	— —	— —	28.22	0.9165	20.25	0.8934	30.95	0.9482	26.47	0.9194
	EDVR [153]	Video	CVPR'19	— —	— —	31.10	0.9371	19.67	0.8724	30.27	0.9440	27.01	0.9178
	RVRT [88]	Video	NIPS'22	— —	— —	30.11	0.9132	21.16	0.8949	26.78	0.8834	26.02	0.8972
	RTA [207]	Video	CVPR'22	— —	— —	30.12	0.9186	20.75	0.8915	29.79	0.9367	26.89	0.9156
Multi-Weather	**All-in-one** [80]	Image	CVPR'20	— —	— —	26.62	0.8948	20.88	0.9010	30.09	0.9431	25.86	0.9130
	UVRNet [70]	Image	TMM'22	— —	— —	22.31	0.7678	20.82	0.8575	24.71	0.8873	22.61	0.8375
	TransWeather [146]	Image	CVPR'22	— —	— —	26.82	0.9118	22.17	0.9025	28.87	0.9313	25.95	0.9152
	TKL [20]	Image	CVPR'22	— —	— —	26.73	0.8935	22.08	0.9044	31.35	0.9515	26.72	0.9165
	WeatherDiffusion [114]	Image	TPAMI'23	— —	— —	25.86	0.9125	20.10	0.8442	26.40	0.9113	24.12	0.8893
	WGWS-Net [212]	Image	CVPR'23	— —	— —	29,64	0.9310	17.71	0.8113	31.58	0.9528	26.31	0.9265
	ViWS-Net [172]	Video	ICCV'23	— —	— —	31.52	0.9433	24.51	0.9187	31.49	0.9562	29.17	0.9394
	TT-WA (ours)	Video	— —	— —	— —	**32.85**	**0.9591**	**25.23**	**0.9212**	**32.71**	**0.9722**	**30.26**	**0.9508**

Table 6.1: **Evaluation of Quantitative Performance for Video Weather Removal under Unseen Conditions.** For the Original Weather, the methods are trained on specific weather training sets and evaluated on corresponding testing sets. For Rain, Haze, and Snow, the methods are trained on a combined training set and evaluated on individual weather-specific testing sets. The average performance metrics are computed for Rain, Haze, and Snow. PSNR and SSIM are utilized as evaluation metrics.

6.3.3 Performance Evaluation

We reference the experimental results from [172], which are widely recognized, and conduct comparisons based on their findings. As shown in Table 6.1, we compare our proposed method against five state-of-the-art methods under three known weather conditions, similar to the approach taken by [172]. These conditions include derain, dehaze, desnow, restoration, and all-in-one adverse weather removal. For all-in-one adverse weather removal, we compare our method with six representative single-image methods: All-in-one [80], UVR-Net [70], TransWeather [146], TKL [20], WeatherDiffusion [114], WGWS-Net [212], and the video-level method ViWS-Net [172]. In the comparison results, we employ a color-coding scheme where the top-performing method is highlighted in " pink ", and the second-best

method is highlighted in " blue " to facilitate clear visualization of performance rankings across different evaluation metrics.

Method	**VRDS** [160]		**RVSD** [15]	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
All-in-one [80]	19.12	0.5882	18.56	0.7128
TransWeather [146]	20.98	0.7187	20.53	0.7486
TKL [20]	19.71	0.7024	19.02	0.7212
WeatherDiffusion [114]	20.21	0.6989	17.61	0.6554
WGWS-Net [212]	21.33	0.7242	18.85	0.7439
ViWS-Net [172]	21.62	0.7131	19.43	0.7510
TT-WA (ours)	**23.11**	**0.7421**	**22.79**	**0.7823**

Table 6.2: Quantitative evaluation on **unseen weather** conditions for video adverse weather removal.

Quantitative Comparison on Seen Domain. In this experiment, we compare the performance of various image enhancement methods under different weather conditions, including deraining, dehazing, desnowing, restoration, and multi-weather methods. Our primary focus is on multi-weather methods due to their capability to handle multiple complex weather conditions. Moreover, we also evaluate single-task methods.

The results in Table 6.1 show that multi-weather methods generally perform more stably and comprehensively. Our method achieves the best overall performance with a Peak Signal-to-Noise Ratio of 32.85 and a Structural Similarity Index of 0.9508, surpassing all comparison methods. WeatherDiffusion and WGWFS-Net also perform well but fall short of our method's results, especially under rain and snow conditions. Single-task methods perform well under specific conditions but show limitations when trained on a mixed set and tested on specific weather conditions. For instance, RDD-Net excels in deraining but performs poorly under haze and snow. GDN is effective in dehazing but unstable in other conditions. DDMSNET is strong in desnowing but average elsewhere. Restoration methods like EDVR perform adequately in dehazing and desnowing but less so in deraining. These methods, while effective in their respective tasks, do not match the versatility of multi-weather methods under mixed conditions. Notably, our method outperforms the recent ViWS-Net, demonstrating higher image quality and stability under complex weather conditions like rain, haze, and snow. This further confirms the robustness and adaptability of our approach.

Quantitative Comparison on Unseen Domain. To evaluate the generalization capability of our model, we use the VRDS and RVSD datasets. Specifically, we use only the test subsets of

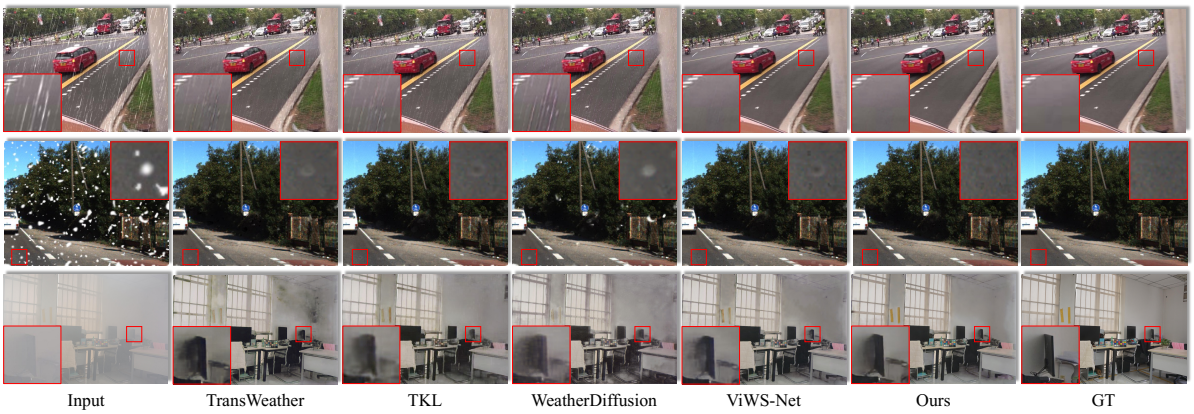


Figure 6.3: **Qualitative Comparison of seen weather conditions on synthetic data between our approach and state-of-the-art methods.** The competing algorithms are selected to demonstrate results on example frames degraded by rain, haze, and snow, respectively. The red boxes highlight detailed comparisons. Please zoom in on the images for enhanced visualization.

these datasets as unseen domains to test the pre-trained model. To assess the performance, we select PSNR and SSIM as the evaluation metrics. As shown in Table 6.2, our model performs excellently on both datasets. On the VRDS dataset, our model achieves a PSNR of 23.11 dB and an SSIM of 0.7421, significantly outperforming other methods, especially the latest ViWS-Net method (PSNR 21.62 dB, SSIM 0.7131). On the RVSD dataset, our model achieves a PSNR of 22.79 dB and an SSIM of 0.7823, also surpassing the latest TransWeather method (PSNR 20.53 dB, SSIM 0.7486). These results indicate that our model excels in handling weather variations across different datasets, demonstrating strong robustness.

Visual Comparison. To more intuitively demonstrate the effectiveness of our approach, Fig. 6.3 presents the visual comparison of our method with four state-of-the-art methods under conditions of rain, fog, and snow. The data for the rain, fog, and snow environments are sourced from the test sets of RainMotion, REVIDE, and KITTI-snow, respectively. Our method consistently exhibits excellent visual quality across various weather conditions. By closely examining the enlarged areas within the red boxes, it is evident that our method significantly reduces the interference of raindrops and snow particles, outperforming other methods. In foggy environments, our method effectively removes residual haze and maintains a clear background with impressive results. Additionally, Fig. 6.4 and Fig.6.5 demonstrate our visual comparison results on the VRDS and RVSD test sets, respectively. Since we did not incorporate the VRDS and RVSD training sets, their test sets can be considered as unseen data. Despite this, we can still observe differences. By comparing the contents within the

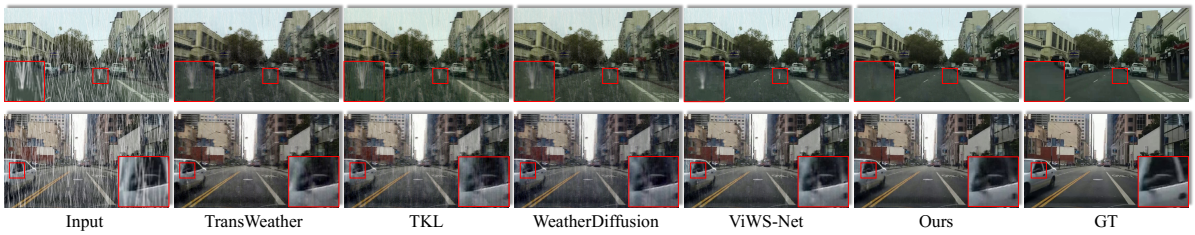


Figure 6.4: **Qualitative comparison of unseen weather conditions on synthetic data from the VRDS dataset.** We compare our approach with state-of-the-art methods in heavy rain scenarios. The red boxes highlight detailed comparisons. Please zoom in on the images for enhanced visualization.

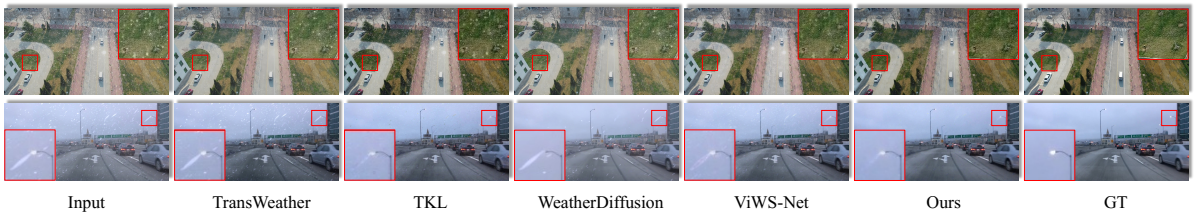


Figure 6.5: **Qualitative comparison of unseen weather conditions on synthetic data from the RVSD dataset.** We compare our approach with state-of-the-art methods in a complex snow-and-fog scenario. The red boxes highlight detailed comparisons. Please zoom in on the images for enhanced visualization.

red boxes, it is apparent that the superiority of our method remains evident.

6.3.4 Ablation Study

Batch Norm vs. Layer Norm. One of the main changes in our enhanced Transformer block design is replacing the default Layer Normalization (LN) with 3D-Batch Normalization (3D-BN) to align more effectively with the demands of reconstructing under multiple adverse weather conditions. Table 6.3 presents our experimental results, demonstrating the performance improvements achieved with 3D-BN. In the derain task, the PSNR increases to 32.85 dB and the SSIM to 0.9591; in the dehaze task, the PSNR improves to 25.23 dB and the SSIM to 0.9212; and in the desnow task, the PSNR rises to 32.71 dB and the SSIM to 0.9722. These results indicate that 3D-BN is better suited to handling video data from different domains, thereby enhancing the overall performance of the model.

Ablation studies for different components. In Table 6.4, we validate the effectiveness of our proposed method through a series of comparative experiments. First, comparing Index 1 and 2, we observe that the inclusion of the Self-Supervised Learning (SSL) branch does

Normalization	Derain		Dehaze		Desnow	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
w. LN	32.41	0.9568	24.89	0.9198	32.55	0.9713
w. 3D-BN	**32.85**	**0.9591**	**25.23**	**0.9212**	**32.71**	**0.9722**

Table 6.3: Ablation studies comparing LN and 3D-BN in enhanced transformer blocks.

Index	*SSL*	θ *Update*	*Train. Alg.*	*TT-WA*	Derain		Dehaze		Desnow	
					PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
1	✗	All	Joint	✗	28.37	0.9113	22.24	0.9041	28.97	0.9213
2	✓	All	Joint	✗	28.79	0.9154	22.69	0.9082	29.41	0.9243
3	✓	All	Joint & Meta	✗	28.91	0.9157	22.88	0.9081	29.65	0.9247
4	✓	BN	Joint & Meta	✗	30.17	0.9201	23.17	0.9140	30.04	0.9352
5	✓	$BN_{(\gamma,\beta)}$	Joint & Meta	✗	31.30	0.9438	24.04	0.9177	31.17	0.9488
6	✓	$BN_{(\gamma,\beta)}$	Joint & Meta	✓	**32.85**	**0.9591**	**25.23**	**0.9212**	**32.71**	**0.9508**

Table 6.4: Ablation studies for different components of our framework. *SSL* denotes SSL branch. *Param.* denotes which parameters are updating, including the whole network ("All"), BN layer ("BN") or only the affine parameters ("Aff"). *TS* denotes the training scheme. *Adapt* denotes whether adapting to each target domain.

not significantly impact the results under Joint training alone. This is reasonable because SSL needs to be combined with Meta-BN training to fully exert its effect. However, merely adding Meta-BN training does not significantly improve performance and even results in some metrics declining, as seen in the comparison between Index 2 and 3. This is intuitive because, during the learning process of different specific weather datasets, the model may forget previously acquired knowledge while learning new information. To further enhance the model's performance, we train only the Batch Normalization (BN) layers during Meta-BN training, as seen in the comparison between Index 3 and 4. This approach results in a significant performance improvement, indicating that focusing on BN layers can effectively mitigate the forgetting of previously learned knowledge when training on new datasets. Subsequently, we optimize further by training only the affine parameters γ and β within the BN layers, as shown in the comparison between Index 4 and 5. This further improvement verifies the crucial role of affine parameters, which is intuitive as these parameters can effectively adjust the feature distribution across different domains. Finally, during testing, we introduce the Test-time Weather Adaptation strategy, as seen in the comparison between Index 5 and 6. This strategy further enhances the model's performance across various tasks, achieving the best results and demonstrating that real-time adaptation is particularly

effective in handling dynamic weather changes.

ISR $(x;\theta,n)$	**Derain**		**Dehaze**		**Desnow**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
$n=1$	32.33	0.9584	24.19	0.9181	32.35	0.9643
$n=2$ (default)	**32.85**	0.9591	25.23	0.9212	32.71	0.9722
$n=3$	32.83	**0.9595**	**25.29**	**0.9214**	**32.74**	**0.9729**

Table 6.5: Ablation studies for different components of our framework.

Ablation studies for Iterative Self-Refinement (ISR). Table 6.5 shows the impact of different iteration numbers n on the ISR method in the tasks of deraining, dehazing, and desnowing. Overall, as the iteration number n increases, PSNR and SSIM improve, but the gains tend to level off, especially between $n=2$ and $n=3$. Specifically, in the deraining task, performance is optimal at $n=2$; in the dehazing task, there is a significant improvement at $n=2$; and in the desnowing task, the performance is similar for $n=2$ and $n=3$. Moderately increasing the iteration number significantly enhances performance, with $n=2$ achieving a good balance between performance and computational efficiency. The experimental results validate the effectiveness of self-supervised refinement in the ISR method, with an iteration number of $n=2$ performing best across all tasks.

Method	**Parameters (M)**	**Inference time (s)**
TransWeather [146]	37.68	0.49
TKL [20]	**28.71**	0.51
WeatherDiffusion [114]	82.96	342.76
ViWS-Net [172]	57.82	0.46
TT-WA (Ours)	47.76	**0.38**

Table 6.6: Quantitative comparison of computational complexity between the selected models and ViWS-Net. The best values are denoted in bold.

6.4 Conclusion

This chapter proposes a dual-branch network structure based on meta-learning and domain-aware normalization to uniformly handle video reconstruction problems under various severe weather conditions. By combining the self-supervised learning branch and the Reconstruction branch, the model effectively extracts specific weather features and improves the recovery ability under known and unknown weather conditions. The meta-learning

optimization strategy avoids knowledge interference by only updating the affine parameters of the batch normalization layer, enhancing the stability and generalization ability of the model. Experimental results show that this method outperforms existing technologies on multiple benchmark datasets, especially under complex weather conditions such as rain, fog, and snow. It also achieves strong generalization ability for unseen weather through the test-time adaptation mechanism (TT-WA), showing good practical application potential.

CHAPTER 7

CONCLUSION AND FUTURE WORK

This paper designs and proposes four core technologies around the image reconstruction task: makeup style transfer, real-time video deraining, multi-degraded image restoration, and test-time weather adaptation (TT-WA). Although these technologies are targeted at different task scenarios, their core essence is style transfer. Whether it is makeup style transfer or image reconstruction, they all involve the conversion from one visual style to another. By unifying these technologies into a generation task framework, we have developed a system that can handle complex scenes while improving the accuracy and real-time performance of image reconstruction. This research not only expands the application boundaries of generation and reconstruction techniques, but also provides effective solutions for various practical application scenarios.

The four technologies we proposed have a unified logical core in the generation task. Makeup style transfer uses the IP23-Net framework to achieve the migration from no makeup to makeup style, retaining facial geometric features while enhancing facial stereoscopic perception. This is closely related to the need for detail preservation in image reconstruction. Real-time video deraining technology uses the RVDNet network framework to convert visual information containing raindrops into clear and rain-free scenes, demonstrating the potential of style transfer in dynamic video processing. The multi-degraded image restoration technology CrDiff solves the style reconstruction problem under complex degradation conditions by combining static wavelet transform and diffusion model. The TT-WA technology improves the adaptability of the model under unknown weather conditions through self-supervised learning (SSL) and meta-learning, ensuring that the generation task can still

be carried out stably in various complex weather scenarios. These four technologies have not only made breakthroughs in their respective fields, but also formed a synergistic effect through their common generation properties.

These generation technologies have demonstrated important value in multiple practical applications. Makeup style transfer technology provides a personalized experience for virtual makeup trials, and image reconstruction technology, especially in autonomous driving, ensures public safety by improving visual clarity in bad weather. Multi-degraded image restoration in disaster monitoring, by restoring severely degraded images, ensures the accurate transmission of key visual information, and helps speed up emergency response. TT-WA technology enhances the adaptability of visual systems under complex weather conditions and is widely used in autonomous driving, drone monitoring, and smart city monitoring systems, improving the stability and efficiency of these systems.

Future Work. With the continuous development of deep learning technology, we will further optimize these generation technologies, especially in terms of real-time and diversified application capabilities. Future research will explore the fusion of multi-modal data, such as combining text with image generation to achieve more intelligent and personalized generation tasks. At the same time, we will also strive to improve the generalization ability of the model in complex real-world environments and ensure that these generation technologies play a greater role in more application scenarios.

BIBLIOGRAPHY

[1] A. ABUOLAIM AND M. S. BROWN, *Defocus deblurring using dual-pixel data*, in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, eds., vol. 12355 of Lecture Notes in Computer Science, Springer, 2020, pp. 111–126.

[2] S. ANWAR AND N. BARNES, *Real image denoising with feature attention*, in 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 3155–3164.

[3] P. ARBELÁEZ, M. MAIRE, C. C. FOWLKES, AND J. MALIK, *Contour detection and hierarchical image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell., 33 (2011), pp. 898–916.

[4] P. C. BARNUM, S. G. NARASIMHAN, AND T. KANADE, *Analysis of rain and snow in frequency space*, Int. J. Comput. Vis., 86 (2010), pp. 256–274.

[5] T. B. BROWN, B. MANN, N. RYDER, M. SUBBIAH, J. KAPLAN, P. DHARIWAL, A. NEELAKANTAN, P. SHYAM, G. SASTRY, A. ASKELL, S. AGARWAL, A. HERBERT-VOSS, G. KRUEGER, T. HENIGHAN, R. CHILD, A. RAMESH, D. M. ZIEGLER, J. WU, C. WINTER, C. HESSE, M. CHEN, E. SIGLER, M. LITWIN, S. GRAY, B. CHESS, J. CLARK, C. BERNER, S. MCCANDLISH, A. RADFORD, I. SUTSKEVER, AND D. AMODEI, *Language models are few-shot learners*, in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., 2020.

[6] E. R. CHAN, M. MONTEIRO, P. KELLNHOFER, J. WU, AND G. WETZSTEIN, *Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 5799–5809.

[7] H. CHANG, J. LU, F. YU, AND A. FINKELSTEIN, *Pairedcyclegan: Asymmetric style transfer for applying and removing makeup*, in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 40–48.

[8] M. CHANG, Q. LI, H. FENG, AND Z. XU, *Spatial-adaptive network for single image denoising*, in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, eds., vol. 12375 of Lecture Notes in Computer Science, Springer, 2020, pp. 171–187.

[9] C. CHEN, A. DANTCHEVA, AND A. ROSS, *Automatic facial makeup detection with application in face recognition*, in International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain, J. Fiérrez, A. Kumar, M. Vatsa, R. N. J. Veldhuis, and J. Ortega-Garcia, eds., IEEE, 2013, pp. 1–8.

[10] ——, *An ensemble of patch-based subspaces for makeup-robust face recognition*, Inf. Fusion, 32 (2016), pp. 80–92.

[11] C. CHEN, A. DANTCHEVA, T. SWEARINGEN, AND A. ROSS, *Spoofing faces using makeup: An investigative study*, in IEEE International Conference on Identity, Security and Behavior Analysis, ISBA 2017, New Delhi, India, February 22-24, 2017, IEEE, 2017, pp. 1–8.

[12] D. CHEN, C. CHEN, AND L. KANG, *Visual depth guided color image rain streaks removal using sparse coding*, IEEE Trans. Circuits Syst. Video Technol., 24 (2014), pp. 1430–1455.

[13] D. CHEN, Y. ZHUANG, Z. SHEN, C. YANG, G. WANG, S. TANG, AND Y. YANG, *Cross-modal data augmentation for tasks of different modalities*, IEEE Trans. Multim., 25 (2023), pp. 7814–7824.

[14] H. CHEN, K. HUI, S. WANG, L. TSAO, H. SHUAI, AND W. CHENG, *Beautyglow: On-demand makeup transfer framework with reversible generative network*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 10042–10050.

[15] H. CHEN, J. REN, J. GU, H. WU, X. LU, H. CAI, AND L. ZHU, *Snow removal in video: A new dataset and A novel method*, in IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, IEEE, 2023, pp. 13165–13176.

[16] J. CHEN, Y. LU, Q. YU, X. LUO, E. ADELI, Y. WANG, L. LU, A. L. YUILLE, AND Y. ZHOU, *Transunet: Transformers make strong encoders for medical image segmentation*, CoRR, abs/2102.04306 (2021).

[17] J. CHEN, C. TAN, J. HOU, L. CHAU, AND H. LI, *Robust video content alignment and compensation for rain removal in a CNN framework*, in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6286–6295.

[18] L. CHEN, X. LU, J. ZHANG, X. CHU, AND C. CHEN, *Hinet: Half instance normalization network for image restoration*, in IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 182–192.

[19] W. CHEN, H. FANG, C. HSIEH, C. TSAI, I. CHEN, J. DING, AND S. KUO, *ALL snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss*, in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 4176–4185.

[20] W. CHEN, Z. HUANG, C. TSAI, H. YANG, J. DING, AND S. KUO, *Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 17632–17641.

[21] X. CHEN, H. LI, M. LI, AND J. PAN, *Learning A sparse transformer network for effective image deraining*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, 2023, pp. 5896–5905.

[22] Y. CHEN AND C. HSU, *A generalized low-rank appearance model for spatio-temporally correlated rain streaks*, in IEEE International Conference on Computer Vision,

ICCV 2013, Sydney, Australia, December 1-8, 2013, IEEE Computer Society, 2013, pp. 1968–1975.

[23] Z. CHI, Y. WANG, Y. YU, AND J. TANG, *Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 9137–9146.

[24] K. DABOV, A. FOI, V. KATKOVNIK, AND K. O. EGIAZARIAN, *Image denoising by sparse 3-d transform-domain collaborative filtering*, IEEE Trans. Image Process., 16 (2007), pp. 2080–2095.

[25] A. DANTCHEVA, C. CHEN, AND A. ROSS, *Can facial cosmetics affect the matching accuracy of face recognition systems?*, in IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2012, Arlington, VA, USA, September 23-27, 2012, IEEE, 2012, pp. 391–398.

[26] H. DENG, C. HAN, H. CAI, G. HAN, AND S. HE, *Spatially-invariant style-codes controlled makeup transfer*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 6549–6557.

[27] J. DENG, J. GUO, N. XUE, AND S. ZAFEIRIOU, *Arcface: Additive angular margin loss for deep face recognition*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 4690–4699.

[28] Y. DENG, J. YANG, S. XU, D. CHEN, Y. JIA, AND X. TONG, *Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set*, in IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 285–295.

[29] J. DEVLIN, M. CHANG, K. LEE, AND K. TOUTANOVA, *BERT: pre-training of deep bidirectional transformers for language understanding*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, eds., Association for Computational Linguistics, 2019, pp. 4171–4186.

[30] ——, *BERT: pre-training of deep bidirectional transformers for language understanding*, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, eds., Association for Computational Linguistics, 2019, pp. 4171–4186.

[31] T. DEVRIES, M. Á. BAUTISTA, N. SRIVASTAVA, G. W. TAYLOR, AND J. M. SUSSKIND, *Unconstrained scene generation with locally conditioned radiance fields*, in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 14284–14293.

[32] Y. DING, H. FAN, M. XU, AND Y. YANG, *Adaptive exploration for unsupervised person re-identification*, ACM Trans. Multim. Comput. Commun. Appl., 16 (2020), pp. 3:1–3:19.

[33] H. DONG, J. PAN, L. XIANG, Z. HU, X. ZHANG, F. WANG, AND M. YANG, *Multi-scale boosted dehazing network with dense feature fusion*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 2154–2164.

[34] Y. DONG, Y. LIU, H. ZHANG, S. CHEN, AND Y. QIAO, *FD-GAN: generative adversarial networks with fusion-discriminator for single image dehazing*, in The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 10729–10736.

[35] A. M. ELGAMMAL, B. LIU, M. ELHOSEINY, AND M. MAZZONE, *CAN: creative adversarial networks, generating "art" by learning about styles and deviating from style norms*, in Proceedings of the Eighth International Conference on Computational Creativity, ICCC 2017, Atlanta, Georgia, USA, June 19-23, 2017, A. K. Goel, A. Jordanous, and A. Pease, eds., Association for Computational Creativity (ACC), 2017, pp. 96–103.

[36] H. FAN, L. ZHENG, C. YAN, AND Y. YANG, *Unsupervised person re-identification: Clustering and fine-tuning*, ACM Trans. Multim. Comput. Commun. Appl., 14 (2018), pp. 83:1–83:18.

[37] Q. FAN, D. CHEN, L. YUAN, G. HUA, N. YU, AND B. CHEN, *A general decoupled learning framework for parameterized image operators*, IEEE Trans. Pattern Anal. Mach. Intell., 43 (2021), pp. 33–47.

[38] X. FU, J. HUANG, X. DING, Y. LIAO, AND J. W. PAISLEY, *Clearing the skies: A deep network architecture for single-image rain removal*, IEEE Trans. Image Process., 26 (2017), pp. 2944–2956.

[39] Y. GAL AND Z. GHAHRAMANI, *Dropout as a bayesian approximation: Representing model uncertainty in deep learning*, in Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, M. Balcan and K. Q. Weinberger, eds., vol. 48 of JMLR Workshop and Conference Proceedings, JMLR.org, 2016, pp. 1050–1059.

[40] H. GAO, X. TAO, X. SHEN, AND J. JIA, *Dynamic scene deblurring with parameter selective sharing and nested skip connections*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 3848–3856.

[41] K. GARG AND S. K. NAYAR, *Detection and removal of rain from videos*, in 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June - 2 July 2004, Washington, DC, USA, IEEE Computer Society, 2004, pp. 528–535.

[42] ——, *Photorealistic rendering of rain streaks*, ACM Trans. Graph., 25 (2006), pp. 996–1002.

[43] L. A. GATYS, A. S. ECKER, AND M. BETHGE, *Image style transfer using convolutional neural networks*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 2414–2423.

[44] Q. GU, G. WANG, M. T. CHIU, Y. TAI, AND C. TANG, *LADN: local adversarial disentangling network for facial makeup and de-makeup*, in 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 10480–10489.

[45] S. GU, L. ZHANG, W. ZUO, AND X. FENG, *Weighted nuclear norm minimization with application to image denoising*, in 2014 IEEE Conference on Computer Vision

and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, IEEE Computer Society, 2014, pp. 2862–2869.

[46] H. GUPTA, O. KOTLYAR, H. ANDREASSON, AND A. J. LILIENTHAL, *Robust object detection in challenging weather conditions*, in IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024, IEEE, 2024, pp. 7508–7517.

[47] K. HE, X. ZHANG, S. REN, AND J. SUN, *Deep residual learning for image recognition*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 770–778.

[48] P. HENDERSON, V. TSIMINAKI, AND C. H. LAMPERT, *Leveraging 2d data to learn textured 3d mesh generation*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 7495–7504.

[49] M. HEUSEL, H. RAMSAUER, T. UNTERTHINER, B. NESSLER, AND S. HOCHREITER, *Gans trained by a two time-scale update rule converge to a local nash equilibrium*, in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds., 2017, pp. 6626–6637.

[50] J. HO, A. JAIN, AND P. ABBEEL, *Denoising diffusion probabilistic models*, in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., 2020.

[51] D. HORITA AND K. AIZAWA, *SLGAN: style- and latent-guided generative adversarial network for desirable makeup transfer and removal*, in Proceedings of the 4th ACM International Conference on Multimedia in Asia, MMAsia 2022, Tokyo, Japan, December 13-16, 2022, S. Jiang, K. Aizawa, P. Chen, and K. Yanai, eds., ACM, 2022, pp. 29:1–29:5.

[52] B. HU, Z. ZHENG, P. LIU, W. YANG, AND M. REN, *Unsupervised eyeglasses removal in the wild*, IEEE Trans. Cybern., 51 (2021), pp. 4373–4385.

[53] D. HUANG, L. KANG, Y. F. WANG, AND C. LIN, *Self-learning based image decomposition with applications to single image denoising*, IEEE Trans. Multim., 16 (2014), pp. 83–93.

[54] J. HUANG, A. SINGH, AND N. AHUJA, *Single image super-resolution from transformed self-exemplars*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, IEEE Computer Society, 2015, pp. 5197–5206.

[55] S. HUANG, Z. YANG, L. LI, Y. YANG, AND J. JIA, *Avatarfusion: Zero-shot generation of clothing-decoupled 3d avatars using 2d diffusion*, in Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023, A. El-Saddik, T. Mei, R. Cucchiara, M. Bertini, D. P. T. Vallejo, P. K. Atrey, and M. S. Hossain, eds., ACM, 2023, pp. 5734–5745.

[56] X. HUANG AND S. J. BELONGIE, *Arbitrary style transfer in real-time with adaptive instance normalization*, in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 1510–1519.

[57] Z. HUANG, Z. ZHENG, C. YAN, H. XIE, Y. SUN, J. WANG, AND J. ZHANG, *Real-world automatic makeup via identity preservation makeup net*, in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, C. Bessiere, ed., ijcai.org, 2020, pp. 652–658.

[58] P. ISOLA, J. ZHU, T. ZHOU, AND A. A. EFROS, *Image-to-image translation with conditional adversarial networks*, in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 5967–5976.

[59] K. JIANG, Z. WANG, P. YI, C. CHEN, B. HUANG, Y. LUO, J. MA, AND J. JIANG, *Multi-scale progressive fusion network for single image deraining*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 8343–8352.

[60] T. JIANG, T. HUANG, X. ZHAO, L. DENG, AND Y. WANG, *Fastderain: A novel video rain streak removal method using directional gradient priors*, IEEE Trans. Image Process., 28 (2019), pp. 2089–2102.

[61] W. JIANG, S. LIU, C. GAO, J. CAO, R. HE, J. FENG, AND S. YAN, *PSGAN: pose and expression robust spatial-aware GAN for customizable makeup transfer*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 5193–5201.

[62] W. JIANG, S. LIU, C. GAO, R. HE, B. LI, AND S. YAN, *Beautify as you like*, in MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020, C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, and R. Zimmermann, eds., ACM, 2020, pp. 4542–4544.

[63] Y. JIN, W. JIANG, Y. YANG, AND Y. MU, *Zero-shot video event detection with high-order semantic concept discovery and matching*, IEEE Trans. Multim., 24 (2022), pp. 1896–1908.

[64] Y. JING, Y. YANG, Z. FENG, J. YE, Y. YU, AND M. SONG, *Neural style transfer: A review*, IEEE Trans. Vis. Comput. Graph., 26 (2020), pp. 3365–3385.

[65] J. JOHNSON, A. ALAHI, AND L. FEI-FEI, *Perceptual losses for real-time style transfer and super-resolution*, in Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II, B. Leibe, J. Matas, N. Sebe, and M. Welling, eds., vol. 9906 of Lecture Notes in Computer Science, Springer, 2016, pp. 694–711.

[66] L. KANG, C. LIN, AND Y. FU, *Automatic single-image-based rain streaks removal via image decomposition*, IEEE Trans. Image Process., 21 (2012), pp. 1742–1755.

[67] T. KARRAS, S. LAINE, AND T. AILA, *A style-based generator architecture for generative adversarial networks*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 4401–4410.

[68] T. KARRAS, S. LAINE, M. AITTALA, J. HELLSTEN, J. LEHTINEN, AND T. AILA, *Analyzing and improving the image quality of stylegan*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 8107–8116.

[69] M. KRISTO, M. IVASIC-KOS, AND M. POBAR, *Thermal object detection in difficult weather conditions using YOLO*, IEEE Access, 8 (2020), pp. 125459–125476.

[70] A. KULKARNI, P. W. PATIL, S. MURALA, AND S. GUPTA, *Unified multi-weather visibility restoration*, IEEE Trans. Multim., 25 (2023), pp. 7686–7698.

[71] Y. LECUN, Y. BENGIO, AND G. E. HINTON, *Deep learning*, Nat., 521 (2015), pp. 436–444.

[72] C. LEE, Z. LIU, L. WU, AND P. LUO, *Maskgan: Towards diverse and interactive facial image manipulation*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 5548–5557.

[73] B. LI, X. LIU, P. HU, Z. WU, J. LV, AND X. PENG, *All-in-one image restoration for unknown corruption*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 17431–17441.

[74] ——, *All-in-one image restoration for unknown corruption*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 17431–17441.

[75] B. LI, X. PENG, Z. WANG, J. XU, AND D. FENG, *Aod-net: All-in-one dehazing network*, in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 4780–4788.

[76] B. LI, W. REN, D. FU, D. TAO, D. FENG, W. ZENG, AND Z. WANG, *Benchmarking single-image dehazing and beyond*, IEEE Trans. Image Process., 28 (2019), pp. 492–505.

[77] D. LI, Y. YANG, Y. SONG, AND T. M. HOSPEDALES, *Learning to generalize: Meta-learning for domain generalization*, in Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, S. A. McIlraith and K. Q. Weinberger, eds., AAAI Press, 2018, pp. 3490–3497.

[78] K. LI, Z. YANG, L. CHEN, Y. YANG, AND J. XIAO, *CATR: combinatorial-dependence audio-queried transformer for audio-visual video segmentation*, in Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023, A. El-Saddik, T. Mei, R. Cucchiara, M. Bertini, D. P. T. Vallejo, P. K. Atrey, and M. S. Hossain, eds., ACM, 2023, pp. 1485–1494.

[79] M. LI, Q. XIE, Q. ZHAO, W. WEI, S. GU, J. TAO, AND D. MENG, *Video rain streak removal by multiscale convolutional sparse coding*, in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6644–6653.

[80] R. LI, R. T. TAN, AND L. CHEONG, *All in one bad weather removal using architectural search*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 3172–3182.

[81] S. LI, I. B. ARAUJO, W. REN, Z. WANG, E. K. TOKUDA, R. H. JUNIOR, R. M. C. JUNIOR, J. ZHANG, X. GUO, AND X. CAO, *Single image deraining: A comprehensive benchmark analysis*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 3838–3847.

[82] T. LI, R. QIAN, C. DONG, S. LIU, Q. YAN, W. ZHU, AND L. LIN, *Beautygan: Instance-level facial makeup transfer with deep generative adversarial network*, in 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018, S. Boll, K. M. Lee, J. Luo, W. Zhu, H. Byun, C. W. Chen, R. Lienhart, and T. Mei, eds., ACM, 2018, pp. 645–653.

[83] X. LI, H. DING, W. ZHANG, H. YUAN, J. PANG, G. CHENG, K. CHEN, Z. LIU, AND C. C. LOY, *Transformer-based visual segmentation: A survey*, CoRR, abs/2304.09854 (2023).

[84] X. LI, H. YUAN, W. LI, H. DING, S. WU, W. ZHANG, Y. LI, K. CHEN, AND C. C. LOY, *Omg-seg: Is one model good enough for all segmentation?*, CoRR, abs/2401.10229 (2024).

[85] Y. LI, R. T. TAN, X. GUO, J. LU, AND M. S. BROWN, *Rain streak removal using layer priors*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 2736–2744.

[86] Y. LI, N. WANG, J. SHI, J. LIU, AND X. HOU, *Revisiting batch normalization for practical domain adaptation*, in 5th International Conference on Learning Representations,

ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings, OpenReview.net, 2017.

[87] J. LIANG, J. CAO, G. SUN, K. ZHANG, L. V. GOOL, AND R. TIMOFTE, *Swinir: Image restoration using swin transformer*, in IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021, IEEE, 2021, pp. 1833–1844.

[88] J. LIANG, Y. FAN, X. XIANG, R. RANJAN, E. ILG, S. GREEN, J. CAO, K. ZHANG, R. TIMOFTE, AND L. V. GOOL, *Recurrent video restoration transformer with guided deformable attention*, in Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds., 2022.

[89] J. LIAO, Y. YAO, L. YUAN, G. HUA, AND S. B. KANG, *Visual attribute transfer through deep image analogy*, ACM Trans. Graph., 36 (2017), pp. 120:1–120:15.

[90] J. LIU, W. YANG, S. YANG, AND Z. GUO, *Erase or fill? deep joint recurrent rain removal and reconstruction in videos*, in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 3233–3242.

[91] ——, *D3r-net: Dynamic routing residue recurrent network for video rain removal*, IEEE Trans. Image Process., 28 (2019), pp. 699–712.

[92] P. LIU, J. XU, J. LIU, AND X. TANG, *Pixel based temporal analysis using chromatic property for removing rain from videos*, Comput. Inf. Sci., 2 (2009), pp. 53–60.

[93] S. LIU, W. JIANG, C. GAO, R. HE, J. FENG, B. LI, AND S. YAN, *PSGAN++: robust detail-preserving makeup transfer and removal*, IEEE Trans. Pattern Anal. Mach. Intell., 44 (2022), pp. 8538–8551.

[94] W. LIU, G. REN, R. YU, S. GUO, J. ZHU, AND L. ZHANG, *Image-adaptive YOLO for object detection in adverse weather conditions*, in Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, AAAI Press, 2022, pp. 1792–1800.

[95] X. LIU, Y. MA, Z. SHI, AND J. CHEN, *Griddehazenet: Attention-based multi-scale network for image dehazing*, in 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 7313–7322.

[96] Y. LIU, D. JAW, S. HUANG, AND J. HWANG, *Desnownet: Context-aware deep network for snow removal*, IEEE Trans. Image Process., 27 (2018), pp. 3064–3073.

[97] Y. LIU, L. WAN, H. FU, J. QIN, AND L. ZHU, *Phase-based memory network for video dehazing*, in MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022, J. Magalhães, A. D. Bimbo, S. Satoh, N. Sebe, X. Alameda-Pineda, Q. Jin, V. Oria, and L. Toni, eds., ACM, 2022, pp. 5427–5435.

[98] Z. LIU, Y. LIN, Y. CAO, H. HU, Y. WEI, Z. ZHANG, S. LIN, AND B. GUO, *Swin transformer: Hierarchical vision transformer using shifted windows*, in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 9992–10002.

[99] Y. LUO, Y. XU, AND H. JI, *Removing rain from a single image via discriminative sparse coding*, in 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, IEEE Computer Society, 2015, pp. 3397–3405.

[100] Y. LUO AND Y. YANG, *Large language model and domain-specific model collaboration for smart education*, Frontiers Inf. Technol. Electron. Eng., 25 (2024), pp. 333–341.

[101] Y. LYU, J. DONG, B. PENG, W. WANG, AND T. TAN, *SOGAN: 3d-aware shadow and occlusion robust GAN for makeup transfer*, in MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021, H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. César, F. Metze, and B. Prabhakaran, eds., ACM, 2021, pp. 3601–3609.

[102] Y. LYU, Y. JIANG, Z. HE, B. PENG, Y. LIU, AND J. DONG, *3d-aware adversarial makeup generation for facial privacy protection*, IEEE Trans. Pattern Anal. Mach. Intell., 45 (2023), pp. 13438–13453.

[103] K. MA, Z. DUANMU, Q. WU, Z. WANG, H. YONG, H. LI, AND L. ZHANG, *Waterloo exploration database: New challenges for image quality assessment models*, IEEE Trans. Image Process., 26 (2017), pp. 1004–1016.

[104] J. MAIRAL, M. ELAD, AND G. SAPIRO, *Sparse representation for color image restoration*, IEEE Trans. Image Process., 17 (2008), pp. 53–69.

[105] D. R. MARTIN, C. C. FOWLKES, D. TAL, AND J. MALIK, *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics*, in Proceedings of the Eighth International Conference On Computer Vision (ICCV-01), Vancouver, British Columbia, Canada, July 7-14, 2001 - Volume 2, IEEE Computer Society, 2001, pp. 416–425.

[106] V. MUSAT, I. FURSA, P. NEWMAN, F. CUZZOLIN, AND A. BRADLEY, *Multi-weather city: Adverse weather stacking for autonomous driving*, in IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021, IEEE, 2021, pp. 2906–2915.

[107] T. NGUYEN, A. T. TRAN, AND M. HOAI, *Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 13305–13314.

[108] T. V. NGUYEN AND L. LIU, *Smart mirror: Intelligent makeup recommendation and synthesis*, in Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017, Q. Liu, R. Lienhart, H. Wang, S. K. Chen, S. Boll, Y. P. Chen, G. Friedland, J. Li, and S. Yan, eds., ACM, 2017, pp. 1253–1254.

[109] T. NGUYEN-PHUOC, C. LI, L. THEIS, C. RICHARDT, AND Y. YANG, *Hologan: Unsupervised learning of 3d representations from natural images*, in 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 7587–7596.

[110] T. NGUYEN-PHUOC, C. RICHARDT, L. MAI, Y. YANG, AND N. J. MITRA, *Blockgan: Learning 3d object-aware scene representations from unlabelled images*, in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., 2020.

[111] M. NIEMEYER AND A. GEIGER, *GIRAFFE: representing scenes as compositional generative neural feature fields*, in IEEE Conference on Computer Vision and Pattern

Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 11453–11464.

[112] S. W. OH, J. LEE, N. XU, AND S. J. KIM, *Video object segmentation using space-time memory networks*, in 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 9225–9234.

[113] O. ÖZDENIZCI AND R. LEGENSTEIN, *Restoring vision in adverse weather conditions with patch-based denoising diffusion models*, IEEE Trans. Pattern Anal. Mach. Intell., 45 (2023), pp. 10346–10357.

[114] ——, *Restoring vision in adverse weather conditions with patch-based denoising diffusion models*, IEEE Trans. Pattern Anal. Mach. Intell., 45 (2023), pp. 10346–10357.

[115] P. PALA AND S. BERRETTI, *Reconstructing 3d face models by incremental aggregation and refinement of depth frames*, ACM Trans. Multim. Comput. Commun. Appl., 15 (2019), pp. 23:1–23:24.

[116] S. PARK, J. YOO, D. CHO, J. KIM, AND T. H. KIM, *Fast adaptation to super-resolution networks via meta-learning*, in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVII, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, eds., vol. 12372 of Lecture Notes in Computer Science, Springer, 2020, pp. 754–769.

[117] P. PAYSAN, R. KNOTHE, B. AMBERG, S. ROMDHANI, AND T. VETTER, *A 3d face model for pose and illumination invariant face recognition*, in Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009, 2-4 September 2009, Genova, Italy, S. Tubaro and J. Dugelay, eds., IEEE Computer Society, 2009, pp. 296–301.

[118] A. PFEUFFER AND K. DIETMAYER, *Robust semantic segmentation in adverse weather conditions by means of sensor data fusion*, in 22th International Conference on Information Fusion, FUSION 2019, Ottawa, ON, Canada, July 2-5, 2019, IEEE, 2019, pp. 1–8.

[119] V. POTLAPALLI, S. W. ZAMIR, S. H. KHAN, AND F. S. KHAN, *Promptir: Prompting for all-in-one image restoration*, in Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023,

NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds., 2023.

[120] R. QIAN, R. T. TAN, W. YANG, J. SU, AND J. LIU, *Attentive generative adversarial network for raindrop removal from a single image*, in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 2482–2491.

[121] Y. QU, Y. CHEN, J. HUANG, AND Y. XIE, *Enhanced pix2pix dehazing network*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 8160–8168.

[122] S. RAVI AND H. LAROCHELLE, *Optimization as a model for few-shot learning*, in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017.

[123] D. REN, W. ZUO, Q. HU, P. ZHU, AND D. MENG, *Progressive image deraining networks: A better and simpler baseline*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 3937–3946.

[124] W. REN, S. LIU, H. ZHANG, J. PAN, X. CAO, AND M. YANG, *Single image dehazing via multi-scale convolutional neural networks*, in Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II, B. Leibe, J. Matas, N. Sebe, and M. Welling, eds., vol. 9906 of Lecture Notes in Computer Science, Springer, 2016, pp. 154–169.

[125] W. REN, J. TIAN, Z. HAN, A. B. CHAN, AND Y. TANG, *Video desnowing and deraining based on matrix decomposition*, in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 2838–2847.

[126] ——, *Video desnowing and deraining based on matrix decomposition*, in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 2838–2847.

[127] W. REN, J. ZHANG, X. XU, L. MA, X. CAO, G. MENG, AND W. LIU, *Deep video dehazing with semantic segmentation*, IEEE Trans. Image Process., 28 (2019), pp. 1895–1908.

[128] V. SANTHASEELAN AND V. K. ASARI, *A phase space approach for detection and removal of rain in video*, in Intelligent Robots and Computer Vision XXIX: Algorithms and Techniques, Burlingame, California, USA, January 22-26, 2012, J. Röning and D. P. Casasent, eds., vol. 8301 of SPIE Proceedings, SPIE, 2012, p. 830114.

[129] ——, *Utilizing local phase information to remove rain from video*, Int. J. Comput. Vis., 112 (2015), pp. 71–89.

[130] K. SCHWARZ, Y. LIAO, M. NIEMEYER, AND A. GEIGER, *GRAF: generative radiance fields for 3d-aware image synthesis*, in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., 2020.

[131] F. SEZGIN, D. VRIESMAN, D. STEINHAUSER, R. LUGNER, AND T. BRANDMEIER, *Safe autonomous driving in adverse weather: Sensor evaluation and performance monitoring*, in IEEE Intelligent Vehicles Symposium, IV 2023, Anchorage, AK, USA, June 4-7, 2023, IEEE, 2023, pp. 1–6.

[132] X. SHEN, J. MA, C. ZHOU, AND Z. YANG, *Controllable 3d face generation with conditional style code diffusion*, in Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, M. J. Wooldridge, J. G. Dy, and S. Natarajan, eds., AAAI Press, 2024, pp. 4811–4819.

[133] Y. SHU, Z. CAO, C. WANG, J. WANG, AND M. LONG, *Open domain generalization with domain-augmented meta-learning*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 9624–9633.

[134] A. SIAROHIN, G. ZEN, C. MAJTANOVIC, X. ALAMEDA-PINEDA, E. RICCI, AND N. SEBE, *Increasing image memorability with neural style transfer*, ACM Trans. Multim. Comput. Commun. Appl., 15 (2019), pp. 42:1–42:22.

[135] K. SIMONYAN AND A. ZISSERMAN, *Very deep convolutional networks for large-scale image recognition*, in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, eds., 2015.

[136] J. W. SOH, S. CHO, AND N. I. CHO, *Meta-transfer learning for zero-shot super-resolution*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 3513–3522.

[137] R. STRUDEL, R. G. PINEL, I. LAPTEV, AND C. SCHMID, *Segmenter: Transformer for semantic segmentation*, in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 7242–7252.

[138] S. SUN, S. FAN, AND Y. F. WANG, *Exploiting image structural similarity for single image rain removal*, in 2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014, IEEE, 2014, pp. 4482–4486.

[139] F. SUNG, Y. YANG, L. ZHANG, T. XIANG, P. H. S. TORR, AND T. M. HOSPEDALES, *Learning to compare: Relation network for few-shot learning*, in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 1199–1208.

[140] Y. SUO, Z. ZHENG, X. WANG, B. ZHANG, AND Y. YANG, *Jointly harnessing prior structures and temporal consistency for sign language video generation*, ACM Trans. Multim. Comput. Commun. Appl., 20 (2024), pp. 185:1–185:18.

[141] N. U. A. TAHIR, Z. ZHANG, M. ASIM, J. CHEN, AND M. A. EL-AFFENDI, *Object detection in autonomous vehicles under adverse weather: A review of traditional and deep learning approaches*, Algorithms, 17 (2024), p. 103.

[142] X. TANG, X. ZHAO, J. LIU, J. WANG, Y. MIAO, AND T. ZENG, *Uncertainty-aware unsupervised image deblurring with deep residual prior*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, 2023, pp. 9883–9892.

[143] C. TIAN, Y. XU, AND W. ZUO, *Image denoising using deep CNN with batch renormalization*, Neural Networks, 121 (2020), pp. 461–473.

[144] H. TOUVRON, M. CORD, M. DOUZE, F. MASSA, A. SABLAYROLLES, AND H. JÉGOU, *Training data-efficient image transformers & distillation through attention*, in Proceedings of the 38th International Conference on Machine Learning, ICML 2021,

18-24 July 2021, Virtual Event, M. Meila and T. Zhang, eds., vol. 139 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 10347–10357.

[145] A. VALADA, J. VERTENS, A. DHALL, AND W. BURGARD, *Adapnet: Adaptive semantic segmentation in adverse environmental conditions*, in 2017 IEEE International Conference on Robotics and Automation, ICRA 2017, Singapore, Singapore, May 29 - June 3, 2017, IEEE, 2017, pp. 4644–4651.

[146] J. M. J. VALANARASU, R. YASARLA, AND V. M. PATEL, *Transweather: Transformer-based restoration of images degraded by adverse weather conditions*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 2343–2353.

[147] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER, AND I. POLOSUKHIN, *Attention is all you need*, in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds., 2017, pp. 5998–6008.

[148] G. WANG, J. C. YE, AND B. D. MAN, *Deep learning for tomographic image reconstruction*, Nat. Mach. Intell., 2 (2020), pp. 737–748.

[149] Q. WANG, S. LI, X. ZHANG, AND G. FENG, *Multi-granularity brushstrokes network for universal style transfer*, ACM Trans. Multim. Comput. Commun. Appl., 18 (2022), pp. 107:1–107:17.

[150] S. WANG, L. ZHU, H. FU, J. QIN, C. SCHÖNLIEB, W. FENG, AND S. WANG, *Rethinking video rain streak removal: A new synthesis model and a deraining network with video rain prior*, in Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIX, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds., vol. 13679 of Lecture Notes in Computer Science, Springer, 2022, pp. 565–582.

[151] ——, *Rethinking video rain streak removal: A new synthesis model and a deraining network with video rain prior*, in Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIX, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds., vol. 13679 of Lecture Notes in Computer Science, Springer, 2022, pp. 565–582.

[152] W. WANG, E. XIE, X. LI, D. FAN, K. SONG, D. LIANG, T. LU, P. LUO, AND L. SHAO, *Pyramid vision transformer: A versatile backbone for dense prediction without convolutions*, in 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 548–558.

[153] X. WANG, K. C. K. CHAN, K. YU, C. DONG, AND C. C. LOY, *EDVR: video restoration with enhanced deformable convolutional networks*, in IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 1954–1963.

[154] X. WANG AND J. YU, *Learning to cartoonize using white-box cartoon representations*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 8087–8096.

[155] X. WANG, L. ZHU, Y. WU, AND Y. YANG, *Symbiotic attention for egocentric action recognition with object-centric alignment*, IEEE Trans. Pattern Anal. Mach. Intell., 45 (2023), pp. 6605–6617.

[156] Y. WANG, S. LIU, C. CHEN, AND B. ZENG, *A hierarchical approach for rain or snow removing in a single color image*, IEEE Trans. Image Process., 26 (2017), pp. 3936–3950.

[157] W. WEI, D. MENG, Q. ZHAO, Z. XU, AND Y. WU, *Semi-supervised transfer learning for image rain removal*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 3877–3886.

[158] W. WEI, L. YI, Q. XIE, Q. ZHAO, D. MENG, AND Z. XU, *Should we encode rain streaks in video as deterministic or stochastic?*, in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 2535–2544.

[159] ——, *Should we encode rain streaks in video as deterministic or stochastic?*, in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 2535–2544.

[160] H. WU, Y. YANG, H. CHEN, J. REN, AND L. ZHU, *Mask-guided progressive network for joint raindrop and rain streak removal in videos*, in Proceedings of the 31st

ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023, A. El-Saddik, T. Mei, R. Cucchiara, M. Bertini, D. P. T. Vallejo, P. K. Atrey, and M. S. Hossain, eds., ACM, 2023, pp. 7216–7225.

[161] J. XU, L. ZHANG, D. ZHANG, AND X. FENG, *Multi-channel weighted nuclear norm minimization for real color image denoising*, in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 1105–1113.

[162] L. XU, Y. DU, AND Y. ZHANG, *An automatic framework for example-based virtual makeup*, in IEEE International Conference on Image Processing, ICIP 2013, Melbourne, Australia, September 15-18, 2013, IEEE, 2013, pp. 3206–3210.

[163] C. YANG, W. HE, Y. XU, AND Y. GAO, *Elegant: Exquisite and locally editable GAN for makeup transfer*, in Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVI, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds., vol. 13676 of Lecture Notes in Computer Science, Springer, 2022, pp. 737–754.

[164] T. YANG, P. REN, X. XIE, AND L. ZHANG, *GAN prior embedded network for blind face restoration in the wild*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 672–681.

[165] W. YANG, J. LIU, AND J. FENG, *Frame-consistent recurrent video deraining with dual-level flow*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 1661–1670.

[166] ——, *Frame-consistent recurrent video deraining with dual-level flow*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 1661–1670.

[167] W. YANG, R. T. TAN, J. FENG, Z. GUO, S. YAN, AND J. LIU, *Joint rain detection and removal from a single image with contextualized deep networks*, IEEE Trans. Pattern Anal. Mach. Intell., 42 (2020), pp. 1377–1393.

[168] W. YANG, R. T. TAN, J. FENG, J. LIU, Z. GUO, AND S. YAN, *Deep joint rain detection and removal from a single image*, in 2017 IEEE Conference on Computer Vision

and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 1685–1694.

[169] W. YANG, R. T. TAN, S. WANG, Y. FANG, AND J. LIU, *Single image deraining: From model-based to data-driven and beyond*, IEEE Trans. Pattern Anal. Mach. Intell., 43 (2021), pp. 4059–4077.

[170] W. YANG, R. T. TAN, S. WANG, AND J. LIU, *Self-learning video rain streak removal: When cyclic consistency meets temporal correspondence*, in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 1717–1726.

[171] X. YANG, D. ZHOU, J. FENG, AND X. WANG, *Diffusion probabilistic model made slim*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, 2023, pp. 22552–22562.

[172] Y. YANG, A. I. AVILÉS-RIVERO, H. FU, Y. LIU, W. WANG, AND L. ZHU, *Video adverse-weather-component suppression network via weather messenger and adversarial backpropagation*, in IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, IEEE, 2023, pp. 13154–13164.

[173] Y. YANG, Y. ZHUANG, AND Y. PAN, *Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies*, Frontiers Inf. Technol. Electron. Eng., 22 (2021), pp. 1551–1558.

[174] Z. YANG, Y. WEI, AND Y. YANG, *Collaborative video object segmentation by multi-scale foreground-background integration*, IEEE Trans. Pattern Anal. Mach. Intell., 44 (2022), pp. 4701–4712.

[175] R. YASARLA AND V. M. PATEL, *Uncertainty guided multi-scale residual learning-using a cycle spinning CNN for single image de-raining*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 8405–8414.

[176] S. YOU, R. T. TAN, R. KAWAKAMI, Y. MUKAIGAWA, AND K. IKEUCHI, *Adherent raindrop modeling, detectionand removal in video*, IEEE Trans. Pattern Anal. Mach. Intell., 38 (2016), pp. 1721–1733.

[177] Z. YUE, J. XIE, Q. ZHAO, AND D. MENG, *Semi-supervised video deraining with dynamical rain generator*, in IEEE Conference on Computer Vision and Pattern

Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 642–652.

[178] Z. YUE, H. YONG, Q. ZHAO, D. MENG, AND L. ZHANG, *Variational denoising network: Toward blind noise modeling and removal*, in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, eds., 2019, pp. 1688–1699.

[179] S. W. ZAMIR, A. ARORA, S. KHAN, M. HAYAT, F. S. KHAN, AND M. YANG, *Restormer: Efficient transformer for high-resolution image restoration*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 5718–5729.

[180] S. W. ZAMIR, A. ARORA, S. H. KHAN, M. HAYAT, F. S. KHAN, M. YANG, AND L. SHAO, *Multi-stage progressive image restoration*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 14821–14831.

[181] ——, *Multi-stage progressive image restoration*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 14821–14831.

[182] S. ZANG, M. DING, D. SMITH, P. TYLER, T. RAKOTOARIVELO, AND M. A. KÂAFAR, *The impact of adverse weather conditions on autonomous vehicles: How rain, snow, fog, and hail affect the performance of a self-driving car*, IEEE Veh. Technol. Mag., 14 (2019), pp. 103–111.

[183] L. ZENG, Z. ZHENG, Y. WEI, AND T. CHUA, *Instilling multi-round thinking to text-guided image generation*, CoRR, abs/2401.08472 (2024).

[184] L. ZHAI, Y. WANG, S. CUI, AND Y. ZHOU, *A comprehensive review of deep learning-based real-world image restoration*, IEEE Access, 11 (2023), pp. 21049–21067.

[185] H. ZHANG AND V. M. PATEL, *Density-aware single image de-raining using a multi-stream dense network*, in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 695–704.

[186] K. ZHANG, D. LI, W. LUO, W. LIN, F. ZHAO, W. REN, W. LIU, AND H. LI, *Enhanced spatio-temporal interaction learning for video deraining: A faster and better framework*, CoRR, abs/2103.12318 (2021).

[187] K. ZHANG, D. LI, W. LUO, J. LIU, J. DENG, W. LIU, AND S. ZAFEIRIOU, *Edface-celeb-1 M: benchmarking face hallucination with a million-scale dataset*, IEEE Trans. Pattern Anal. Mach. Intell., 45 (2023), pp. 3968–3978.

[188] K. ZHANG, D. LI, W. LUO, W. REN, AND W. LIU, *Enhanced spatio-temporal interaction learning for video deraining: A faster and better framework*, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), (2022).

[189] K. ZHANG, R. LI, Y. YU, W. LUO, AND C. LI, *Deep dense multi-scale network for snow removal using semantic and depth priors*, IEEE Trans. Image Process., 30 (2021), pp. 7419–7431.

[190] K. ZHANG, W. ZUO, Y. CHEN, D. MENG, AND L. ZHANG, *Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising*, IEEE Trans. Image Process., 26 (2017), pp. 3142–3155.

[191] K. ZHANG, W. ZUO, S. GU, AND L. ZHANG, *Learning deep CNN denoiser prior for image restoration*, in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 2808–2817.

[192] K. ZHANG, W. ZUO, AND L. ZHANG, *Ffdnet: Toward a fast and flexible solution for cnn-based image denoising*, IEEE Trans. Image Process., 27 (2018), pp. 4608–4622.

[193] N. ZHANG, F. NEX, G. VOSSELMAN, AND N. KERLE, *Lite-mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, 2023, pp. 18537–18546.

[194] X. ZHANG, H. DONG, J. PAN, C. ZHU, Y. TAI, C. WANG, J. LI, F. HUANG, AND F. WANG, *Learning to restore hazy video: A new real-world dataset and a new method*, in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 9239–9248.

[195] X. ZHANG, H. LI, Y. QI, W. K. LEOW, AND T. K. NG, *Rain removal in video by combining temporal and chromatic properties*, in Proceedings of the 2006 IEEE International

Conference on Multimedia and Expo, ICME 2006, July 9-12 2006, Toronto, Ontario, Canada, IEEE Computer Society, 2006, pp. 461–464.

[196] ——, *Rain removal in video by combining temporal and chromatic properties*, in Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME 2006, July 9-12 2006, Toronto, Ontario, Canada, IEEE Computer Society, 2006, pp. 461–464.

[197] X. ZHANG, Z. ZHENG, D. GAO, B. ZHANG, P. PAN, AND Y. YANG, *Multi-view consistent generative adversarial networks for 3d-aware image synthesis*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 18429–18438.

[198] X. ZHANG, Z. ZHENG, D. GAO, B. ZHANG, Y. YANG, AND T. CHUA, *Multi-view consistent generative adversarial networks for compositional 3d-aware image synthesis*, Int. J. Comput. Vis., 131 (2023), pp. 2219–2242.

[199] Y. ZHANG, L. QU, L. HE, W. LU, AND X. GAO, *Beauty aware network: An unsupervised method for makeup product retrieval*, in Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019, L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, eds., ACM, 2019, pp. 2558–2562.

[200] Y. ZHANG, L. WEI, Q. ZHANG, Y. SONG, J. LIU, H. LI, X. TANG, Y. HU, AND H. ZHAO, *Stable-makeup: When real-world makeup transfer meets diffusion model*, CoRR, abs/2403.07764 (2024).

[201] C. ZHAO, Y. ZHANG, M. POGGI, F. TOSI, X. GUO, Z. ZHU, G. HUANG, Y. TANG, AND S. MATTOCCIA, *Monovit: Self-supervised monocular depth estimation with a vision transformer*, in International Conference on 3D Vision, 3DV 2022, Prague, Czech Republic, September 12-16, 2022, IEEE, 2022, pp. 668–678.

[202] Y. ZHENG, X. YU, M. LIU, AND S. ZHANG, *Residual multiscale based single image deraining*, in 30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019, BMVA Press, 2019, p. 147.

[203] Z. ZHENG, X. WANG, N. ZHENG, AND Y. YANG, *Parameter-efficient person re-identification in the 3d space*, IEEE Trans. Neural Networks Learn. Syst., 35 (2024), pp. 7534–7547.

[204] Z. ZHENG, L. ZHENG, AND Y. YANG, *A discriminatively learned CNN embedding for person reidentification*, ACM Trans. Multim. Comput. Commun. Appl., 14 (2018), pp. 13:1–13:20.

[205] Z. ZHONG, L. ZHENG, Z. ZHENG, S. LI, AND Y. YANG, *Camstyle: A novel data augmentation method for person re-identification*, IEEE Trans. Image Process., 28 (2019), pp. 1176–1190.

[206] D. ZHOU, Z. YANG, AND Y. YANG, *Pyramid diffusion models for low-light image enhancement*, in Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China, ijcai.org, 2023, pp. 1795–1803.

[207] K. ZHOU, W. LI, L. LU, X. HAN, AND J. LU, *Revisiting temporal alignment for video restoration*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, IEEE, 2022, pp. 6043–6052.

[208] J. ZHU, T. PARK, P. ISOLA, AND A. A. EFROS, *Unpaired image-to-image translation using cycle-consistent adversarial networks*, in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 2242–2251.

[209] J. ZHU, Z. ZHANG, C. ZHANG, J. WU, A. TORRALBA, J. TENENBAUM, AND B. FREEMAN, *Visual object networks: Image generation with disentangled 3d representations*, in Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., 2018, pp. 118–129.

[210] L. ZHU, C. FU, D. LISCHINSKI, AND P. HENG, *Joint bi-layer optimization for single-image rain streak removal*, in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 2545–2553.

[211] X. ZHU, W. SU, L. LU, B. LI, X. WANG, AND J. DAI, *Deformable DETR: deformable transformers for end-to-end object detection*, in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021.

[212] Y. ZHU, T. WANG, X. FU, X. YANG, X. GUO, J. DAI, Y. QIAO, AND X. HU, *Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions*, in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, 2023, pp. 21747–21758.