



Mixed-Degradation Image Restoration using Vision-Language Guidance and Multi-Label Semantic Modulation

By

Mintesnot kir Abate

Submitted to the School of Information Technology and Engineering
in partial fulfillment of the requirements for the degree of
Master of Science in Artificial Intelligence

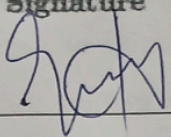
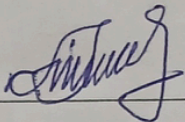
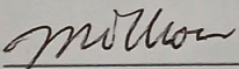
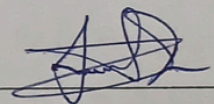
Advisor: Beakal Gizachew(PhD)

College of Technology and Built Environment
Addis Ababa University
Addis Ababa, Ethiopia

February 2026

APPROVAL

This is to certify that this thesis titled "**Mixed-Degradation Image Restoration using Vision-Language Guidance and Multi-Label Semantic Modulation**" is prepared by Mintesnot Fikir Abate and submitted in partial fulfillment of the thesis-option requirements for the Degree of Master of Science in Artificial Intelligence at the School of Information Technology and Engineering, College of Technology and Built Environment.

Name	Signature	Date
<u>Dr. Sileshi Demesie</u> (Chairman)		<u>25/05/2026</u>
<u>Dr. Beakal Gizachew</u> (Advisor)		<u>May - 21 - 2026</u>
<u>Dr. Million Meshesha</u> (External Examiner)		<u>May 21/2026</u>
<u>Dr. Adane Letta</u> (Internal Examiner)		<u>May 22, 2026</u>

ABSTRACT

Real-world images are rarely degraded by a single factor; instead, multiple corruptions such as low-light, haze, rain, and blur frequently co-exist and interact across various spatial scales. This mixed-degradation setting exposes critical limitations in current all-in-one restoration methods, which often rely on single-label representations and spatially uniform conditioning. Such approaches fail to model non-exclusive mixtures and capture what degradations are present without identifying where they occur, while providing limited diagnostic transparency for model decisions in practical applications. To address these gaps, we propose the Composite Degradation Semantic Modulation Network (CDSM-Net), a novel framework that decouples degradation reasoning from reconstruction. Our architecture integrates three core components: a wavelet-structured backbone (WAVEViM) for frequency separation, a vision-language semantic controller (MDSM) for multi-label prediction, and a self-gated spatial modulation (SGSM) interface for localized feature conditioning. By leveraging a prototype memory bank and routing weights, the MDSM provides an inspectable intermediate degradation signal and generates a global conditioning code that the SGSM translates into pixel-wise modulation without external masks. Experimental results on composite benchmarks and real-world datasets demonstrate the model’s effectiveness. On the CDD-11 benchmark, CDSM-Net achieves an average PSNR of 29.47 dB and SSIM of 0.9883, improving PSNR by +0.74 dB over OneRestore and attaining the highest average structural similarity among compared state-of-the-art methods. Additional blur-aware stress analysis on CDD-13 further supports the model’s robustness under mixed-frequency composite corruption. Evaluations on the real-world LOL-v2 dataset yield an NIQE of 5.1309, outperforming representative low-light baselines and indicating improved generalization under domain shift. Overall, these results establish CDSM-Net as a robust framework with inspectable degradation routing for resolving non-exclusive mixed degradations in diverse imaging environments.

Keywords: All-in-one image restoration, mixed/composite degradation, vision-language guidance, spatial feature modulation, wavelet decomposition, state-space models.

Acknowledgements

First and foremost, I would like to thank God. He has given me strength and encouragement throughout all the challenging moments of completing this thesis. I am truly grateful for His unconditional and endless love, mercy, and grace.

I would like also to express my heartfelt gratitude to Dr. Beakal Gizachew, my MSc advisor. His expertise, guidance, and unwavering support have been instrumental in shaping my academic journey. I am truly grateful for his invaluable insights, constructive feedback, and constant motivation. Working under his supervision has been an enriching experience that has contributed significantly to my personal and professional growth. I extend my sincerest appreciation to Dr. Beakala Gizachew for his mentorship and dedication throughout my MSc program.

Dedication

To my family.

. . .

To my academic advisor.

. . .

To my best friends.

. . .

Contents

Abstract	ii
Acknowledgements	iii
Dedication	iv
Abbreviations	viii
List of Figures	xi
List of Tables	xii
1 Introduction	1
1.1 Motivation	2
1.2 Statement of the Problem	3
1.3 Research Questions	5
1.4 Objective	5
1.4.1 General objective	5
1.4.2 Specific objectives	5
1.5 Significance	6
1.6 Scope	7
1.7 Contributions of the Study	7
1.8 Thesis structure	9
2 Literature review	10
2.1 Foundations	10
2.1.1 Image Formation	10
2.1.2 Image Degradation Overview	11
2.1.3 Image Restoration Overview	13
2.2 Image Restoration Techniques	14
2.2.1 Classical and Early Learning-Based Approaches	16
2.2.2 Deep Learning and Hybrid Architectures	17
2.2.3 Degradation Coverage and Generalization in Image Restoration	25
2.3 All-in-One Image Restoration (AiOIR)	28
2.3.1 Backbones and Architectural Patterns in AiOIR	29
2.3.2 Degradation Representation and Conditioning in AiOIR	34
2.4 Related Work	40

3	Methodology	47
3.1	Research Methodology	47
3.2	Data Acquisition	48
3.2.1	Datasets and category definitions	49
3.2.2	Dataset synthesis / generation (CDD-11 $\rightarrow$ CDD-13)	50
3.3	Benchmark Datasets and Evaluation Protocol	52
3.3.1	Paired Synthetic Composite Benchmarks (Primary)	52
3.3.2	Real-World Benchmarks (No-Reference Setting, NIQE-only)	53
3.3.3	Evaluation Metrics and Reporting Protocol	53
3.4	Data Preprocessing	54
3.4.1	Letterbox resizing with replicate padding	56
3.4.2	Validity mask for content-only supervision	56
3.4.3	Negative sample (training only)	56
3.5	Design and Development	56
3.5.1	Proposed Architecture Overview: CDSM-Net	57
3.5.2	Multi-label Degradation Semantic Modulator (MDSM)	58
3.5.3	Self-Gated Spatial Modulation (SGSM)	63
3.5.4	Wavelet-Attention Vision Mamba Backbone (WAVeViM)	66
3.5.5	Learning Objective and Optimization	69
3.5.6	Training Process	72
3.5.7	Inference Process	75
4	Experimentation and Results	77
4.1	Experimental Setup	77
4.2	Training Stability, Convergence, and Optimization	77
4.2.1	Training Stability and Convergence	78
4.2.2	Adaptive Loss Balancing via Uncertainty-Weighted Multi-Task Learning	79
4.3	Quantitative Results	81
4.3.1	Overall Results	82
4.3.2	Single-Degradation Benchmarks	84
4.3.3	All-in-One and Composite Degradation Benchmarks	85
4.3.4	Real-World Mixed-Degradation Images (No-Reference)	88
4.3.5	Efficiency and Complexity	89
4.4	Ablation Studies and Component Analysis	91
4.4.1	Impact of input data processing	91
4.4.2	Semantic Controller Ablation (MDSM & OpenCLIP Guidance)	93
4.4.3	Impact of SGSM (Spatial Selective Modulation)	94
4.5	Discussion	96
4.5.1	Key Findings	99
4.5.2	Limitations	101

5	Conclusion and recommendation	102
5.1	Conclusion	102
5.2	Recommendations and Future Work	103
	References	105

Abbreviations

Symbol	Meaning
AI	Artificial Intelligence
AiOIR	All-in-One Image Restoration
AMP	Automatic Mixed Precision
ANN	Artificial Neural Network
Avg	Average
AWGN	Additive White Gaussian Noise
BIQA	Blind Image Quality Assessment
CDD	Composite Degradation Dataset
CDSM-Net	Composite Degradation Semantic Modulation Network
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
DCP	Dark Channel Prior
DSRP	Design Science Research Process
DWT	Discrete Wavelet Transform
EMA	Exponential Moving Average
FFT	Fast Fourier Transform
FiLM	Feature-wise Linear Modulation
FR	Full-Reference
GAN	Generative Adversarial Network
GFLOPs	Giga Floating Point Operations per Second
GT	Ground Truth
HF	High-Frequency
IDWT	Inverse Discrete Wavelet Transform
IQA	Image Quality Assessment
IR	Image Restoration
ISP	Image Signal Processing
LF	Low-Frequency
LPIPS	Learned Perceptual Image Patch Similarity
MACs	Multiply-Accumulate Operations
MAP	Maximum A Posteriori
MDIR	Multi-Degradation Image Restoration
MDSM	Multi-label Degradation Semantic Modulator
MLP	Multi-Layer Perceptron
MoE	Mixture-of-Experts
MSE	Mean Squared Error
NIQE	Natural Image Quality Evaluator
NR	No-Reference

NR-IQA	No-Reference Image Quality Assessment
NSS	Natural Scene Statistics
PSF	Point Spread Function
PSNR	Peak Signal-to-Noise Ratio
RDN	Residual Dense Network
SDIR	Single-Degradation Image Restoration
SGSM	Self-Gated Spatial Modulation
SSIM	Structural Similarity Index
SSM	Structured State Space Model
TV	Total Variation
ViT	Vision Transformer
VLM	Vision-Language Model
WAVeViM	Wavelet-Attention Vision Mamba

LIST OF FIGURES

2.1	A hierarchical taxonomy of Image Restoration techniques.	15
2.2	Conceptual evolution of deep and hybrid image restoration backbones . .	24
2.3	architectural design patterns in AiOIR	29
3.1	Research Design Workflow[1].	47
3.2	Data preprocessing pipeline.	55
3.3	Overview of the proposed architecture network: CDSM-Net	57
3.4	MDSM block diagram	59
3.5	Self-Gated Spatial Modulation (SGSM)	63
4.1	Baseline training dynamics (fixed loss weights)	79
4.2	Uncertainty-weighted training dynamics	81
4.3	Learned task uncertainties	82
4.4	CDD-11 per-category performance profile (PSNR/SSIM radar) across 11 degradation categories, comparing representative all-in-one and composite- aware baselines with CDSM-Net.	87
4.5	No-reference NIQE comparison on three real-world benchmarks (lower is better).	88
4.6	Effect of input processing on restoration quality.	92

LIST OF TABLES

2.1	Qualitative comparison of deep backbone families for image restoration. .	25
2.2	Summary of AiOIR backbone families	34
2.3	Comparative analysis on Related Work	45
3.1	Datasets and metrics used for training and evaluation.	53
4.1	Implementation and optimization settings used in the experiments. . . .	77
4.2	CDD-11 overall results (average PSNR/SSIM over 11 categories). Best/second/third are highlighted in red/green/blue.	83
4.3	CDD-11 single-degradation results (Low/Haze/Rain/Snow). Best/second/third are highlighted in red/green/blue among the methods listed.	85
4.4	CDD-11 composite-category results. Composite-average highlighting: best/second/third in red/green/blue.	86
4.5	Efficiency of CDSM-Net at 256×256 on a Tesla P100-PCIE-16GB (Kaggle). *MACs/GFLOPs are THOP estimates and may under-count unsupported/custom operations.	90
4.6	Trainable parameter breakdown of CDSM-Net grouped by our methodology: restoration backbone vs. semantic/spatial control modules.	90
4.7	Ablation of input processing on CDD-11 (average over 11 categories) . .	92
4.8	CDD-11 average results (mean over 11 categories) for semantic-controller ablations.	94
4.9	SGSM ablation on CDD-11. (Higher is better).	96

Chapter 1

Introduction

Image restoration is a core problem in low-level vision because image degradations directly affect both human visual perception and the reliability of downstream computer vision systems. In practical imaging pipelines, degradations arise from sensor noise, low illumination, camera motion, defocus, atmospheric scattering, and adverse weather, often under conditions where degradation type and severity are not known in advance [2, 3, 4, 5]. Classical restoration methods based on hand-crafted priors and model-based optimization established important foundations for denoising and inverse problems, but their assumptions are often restrictive when degradations are complex or jointly present in real scenes [6, 7, 8, 9].

Data-driven restoration has significantly advanced the field by learning stronger priors from paired or synthetic training data, and modern deep models have achieved strong results across many restoration tasks. Early deep restoration models demonstrated the advantage of end-to-end learning, while later transformer- and hybrid-based models improved long-range dependency modeling and high-resolution restoration quality [10, 11, 12, 13]. Lightweight and efficient architectures further improved practicality without sacrificing accuracy in many settings [14, 15]. However, most of these successes were first established in single-degradation or task-specific settings, where the restoration objective is comparatively well-defined [14, 13, 11].

To reduce the burden of deploying separate models for different degradations, all-in-one image restoration (AiOIR) and universal restoration frameworks have become an active research direction [16, 17]. Representative methods have explored unified degradation modeling, prompt-based conditioning, ingredient-oriented learning, and instruction-guided restoration to support multiple tasks in a single framework [18, 19, 20]. More recent work has extended this direction through composite degradation modeling, diffusion-based universal restoration, adaptive frequency modulation, and degradation representation learning, reflecting the field’s shift toward broader degradation coverage and stronger

generalization [21, 22, 23, 24, 25, 26]. In parallel, efficient adaptation and pretraining strategies have also been explored to improve deployability and robustness in blind all-in-one restoration [27, 28, 29].

Another important trend is the use of vision–language priors and controllable semantic guidance for image restoration. Vision–language models such as CLIP provide transferable semantic representations, and recent restoration works have investigated how these priors can improve controllability and multi-task restoration behavior [30, 31, 32].

Recent advances in unified restoration, diffusion-based restoration, and semantic guidance have significantly expanded the scope of image restoration beyond single-task settings [21, 22, 23, 30, 31]. Nevertheless, real-world mixed degradations remain challenging because they are often non-exclusive, spatially heterogeneous, and difficult to characterize reliably under blind conditions [3, 5, 33, 34, 35].

1.1 Motivation

High-quality imagery is a prerequisite for both human-centered applications (computational photography, mobile imaging, video conferencing) and machine-centered applications (autonomous driving, surveillance, remote sensing). In these settings, restoration is not merely cosmetic: degradations can conceal fine structures, reduce contrast, and destabilize downstream perception. However, real-world capture conditions are heterogeneous and non-stationary. A deployed system must operate on unknown mixtures of degradations without access to ground-truth degradation parameters, and it must do so efficiently.

Recent advances in deep restoration have delivered impressive results on single-task benchmarks, largely by scaling expressive backbones (CNNs and Transformers) and improving training losses [36, 13, 14, 11]. Yet these successes expose a practical gap: performance often degrades when the corruption distribution shifts, when multiple degradations co-occur, or when degradations are spatially localized. This motivates restoration systems that can (a) infer an explicit degradation state, (b) condition restoration based on that state, and (c) localize corrections without requiring external spatial annotations.

Vision–language guidance is attractive because language can express degradations at a semantic level (e.g., “low-light with haze and rain”), and VLM embeddings can provide

robust global context. Nonetheless, naive conditioning with VLM features is insufficient: embeddings are entangled with content and may not produce a clean separation between “scene semantics” and “degradation semantics.” This motivates a structured controller that extracts an interpretable multi-label degradation representation and uses it to drive restoration through a principled modulation mechanism.

1.2 Statement of the Problem

All-in-one image restoration (AiOIR) has become an important direction in low-level vision because it reduces the need to maintain separate models for denoising, deblurring, deraining, dehazing, and related restoration tasks [16, 19]. Recent methods have improved unified restoration through prompts, instructions, latent priors, and diffusion-based formulations [20, 22, 23, 21]. However, real images often contain mixed and spatially varying degradations, where multiple artifacts co-exist with different intensities and overlap across the scene [3, 5, 33]. In this setting, the main challenges are: non-exclusive degradation representation, spatial specificity of conditioning, limited verifiability of degradation reasoning, and the combined challenge of multi-objective conflict and composite-degradation robustness [21, 34, 35].

A first challenge is non-exclusive degradation representation, namely identifying what degradations are present and how strongly they manifest when degradations co-occur. Several AiOIR methods use prompts, instructions, or compact latent controls to guide a unified restorer, and these strategies improve task coverage in practice [19, 20, 22]. Nevertheless, many representations remain implicit, coarse, or effectively single-choice when mixtures are complex, which can bias restoration toward a dominant degradation while weaker degradations remain under-corrected [16, 18, 21]. Recent degradation-aware priors and prompt-state modeling move toward finer degradation descriptions, but mixed/composite restoration still requires more faithful non-exclusive representation of degradation type and severity [34, 37, 38].

A second challenge is spatial specificity, because real degradations are not uniformly distributed across pixels. Some degradations are global, some are local, and others overlap with different intensities in different regions, so a global cue does not directly specify where restoration should be applied [39, 33, 21]. This is especially important in mixed

settings where localized artifacts co-exist with scene-level corruption. In conditioning-based networks, feature-wise modulation can adapt channels effectively, but channel-only conditioning is spatially uniform unless an explicit spatial mechanism is introduced [40, 41]. As a result, current AiOIR systems may still over-correct relatively clean regions or under-correct strongly degraded regions when the conditioning signal is not sufficiently spatially grounded [12, 15, 22].

A third challenge is limited diagnostic transparency of degradation reasoning. As AiOIR models increasingly rely on prompts, latent codes, diffusion guidance, or expert routing, it becomes harder to inspect whether the model correctly identified the degradation composition before restoration and why a failure occurs in real images [19, 23, 42]. This matters in practice because, under blind mixed-degradation restoration, benchmark labels are usually unavailable at inference time and visual plausibility alone is not always sufficient to verify whether the restoration process was guided by the correct degradation assumption. The issue is more pronounced for semantic guidance because general vision-language embeddings may encode image content and degradation cues jointly unless they are explicitly adapted for degradation-aware control [31, 43, 38]. Therefore, this work treats inspectable intermediate degradation reasoning primarily as a diagnostic aid for failure analysis and trustworthy deployment, rather than as an independent restoration objective [34, 21, 35].

A fourth challenge is the joint problem of multi-objective conflict and composite-degradation robustness. In mixed restoration, different degradations require competing operations, so optimizing one objective can weaken another (for example, detail amplification for deblurring may conflict with suppression behavior needed for denoising or deraining), which leads to task interference and unstable training [44, 45, 46]. Recent works explicitly recognize this through uncertainty-aware transport, adaptive prompt/state modeling, and evolutionary frequency modulation [35, 37, 47]. At the same time, robustness to composite degradations remains difficult, especially for blur-centric mixtures that suppress high-frequency details and increase failure sensitivity. Unified and diffusion and MoE-based methods improve coverage and flexibility, but they can introduce higher inference cost, opaque routing or selection behavior, or fragile robustness under hard composite settings [21, 23, 42, 48]. This work addresses these gaps by in-

roducing a unified mixed-degradation restoration framework designed to improve five practical properties simultaneously: non-exclusive degradation representation, spatially specific conditioning, diagnostically inspectable degradation reasoning, stable learning under competing objectives, and stronger robustness to composite degradations.

1.3 Research Questions

RQ1: How can vision–language priors be integrated to produce a multi-label representation of non-exclusive mixed degradations with diagnostically inspectable routing behavior?

RQ2: How can global semantic context be spatially modulated to restore localized degradations without relying on external spatial priors (e.g., masks or degradation maps)?

RQ3: To what extent does explicit frequency decomposition reduce spectral bias and improve detail recovery under blur-centric and composite degradations?

RQ4: How does dynamic multi-objective learning affect training stability and balanced performance across competing restoration objectives?

1.4 Objective

1.4.1 General objective

To design a mixed-degradation-aware image restoration framework that identifies non-exclusive degradation mixtures and restores images using spatially adaptive, frequency-aware conditioning under practical inference constraints.

1.4.2 Specific objectives

- To develop a vision–language guided multi-label semantic controller that estimates mixture weights and a compact conditioning code.
- To design a self-gated spatial modulation mechanism that injects the conditioning signal without external spatial priors.
- To develop a wavelet-structured hybrid backbone that improves blur-centric detail recovery via low/high-frequency separation.

- To formulate a stable multi-objective training strategy for competing restoration objectives.
- To evaluate performance using quantitative metrics and targeted ablations on composite benchmarks and real-world images.

1.5 Significance

Mixed-degradation restoration is important because real images are rarely corrupted by a single process. In practical capture conditions (mobile photography, surveillance, outdoor sensing, adverse weather), degradations such as low-light, haze, blur, rain, and snow often co-occur and vary across regions, so models designed for isolated degradations can become fragile in deployment.

For image restoration researchers and for practitioners building real systems, this thesis advances universal restoration by introducing a *multi-label, non-exclusive* degradation representation guided by vision–language priors and by enabling spatially selective conditioning without requiring external degradation maps or segmentation masks. Together, these choices improve controllability and provide diagnostic transparency for failure analysis while aligning with the operational reality that only a corrupted image is available at inference time.

For downstream vision applications and for reproducible benchmarking and model development, stronger mixed-degradation restoration can reduce error propagation into recognition or scene understanding pipelines, while the ablations and evaluation discipline in this thesis provide clear evidence of what works, why it works, and where it fails. In general, the study supports more interpretable, spatially adaptive, and deployable mixed-degradation restoration and highlights concrete directions for improving real-world robustness. Practically, the proposed design targets blind real-world restoration, where the degradation type and location are unknown at inference time, making the approach relevant for mobile photography, adverse-weather imaging, and pre-processing for downstream vision systems.

1.6 Scope

This thesis focuses on single-image restoration under mixed and spatially heterogeneous degradations, where multiple corruption types co-exist in the same image and their locations are unknown at inference time. The scope is defined to develop and validate a blind restoration framework that can handle non-exclusive degradation mixtures in a practical and reproducible way.

Within scope, we study blind restoration under mixed degradations including haze, rain, snow, low-light (noise/exposure), and blur-centric composites. Experiments are conducted on the CDD family of synthetic composite-degradation datasets, and we further extend the dataset with additional blur-centric composites to analyze robustness under motion/blur interactions. The proposed framework is constrained to a frozen vision-language encoder for guidance, a learned multi-label semantic modulator, a spatial modulation interface, and a wavelet-based hybrid backbone, without assuming access to ground-truth degradation parameters at inference time. Performance is measured using standard full-reference restoration metrics computed on valid (unpadded) image regions, supported by ablation studies that isolate the contribution of each proposed component.

Outside scope, we do not target video restoration or temporal consistency, and we do not model full RAW imaging pipelines or camera-specific ISP stages. We also do not treat large-scale unpaired domain adaptation as a primary objective; instead, real-world distribution shift is used as a stress test to expose generalization gaps and motivate future work.

1.7 Contributions of the Study

To address the aforementioned limitations in mixed-degradation modeling and restoration, this thesis proposes the Composite Degradation Semantic Modulation Network (CDSM-Net). The core philosophy of this work is to move beyond the implicit, single-label paradigms of current All-in-One frameworks toward an explicit, semantically grounded, and spatially adaptive restoration architecture. The primary technical and theoretical contributions of this research are summarized as follows:

- **Multi-label Degradation Semantic Modulator (MDSM) with inspectable**

routing: We introduce a specialized semantic controller that overcomes the limitations of softmax-based classifiers, which force dominant degradation representations to suppress minor co-occurring corruptions. Unlike black-box prompt-learning methods, the MDSM projects CLIP-extracted features onto a learnable degradation prototype memory bank and produces differentiable routing weights (α) that quantify the detected degradation types and their relative intensities in a non-competing, sigmoid-based manner. This provides an inspectable intermediate degradation signal that supports practical failure diagnosis when restoration quality degrades.

- **Channel-Distinct Self-Gated Spatial Modulation (SGSM):** We propose a spatially adaptive conditioning interface designed to resolve heterogeneous mixtures without the need for external segmentation masks or expensive iterative sampling. While standard spatial modulation techniques (e.g., FiLM) apply uniform gating maps across all feature channels, our SGSM generates unique spatial gate maps for each channel independently, derived directly from the feature maps. This allows the global semantic code to be translated into precisely localized, spatially variant restoration actions, effectively handling mixtures where degradations vary across pixel coordinates.
- **Wavelet-inspired Mamba Backbone (WAVEViM):** To resolve the computational bottleneck inherent in diffusion-based AiOIR models, we integrate a linear-complexity State Space Model (SSM) backbone. The WAVEViM architecture utilizes wavelet-inspired separation to explicitly decouple low-frequency base representations (e.g., haze and low-light) from high-frequency structural details (e.g., blur and noise). This prevents the frequency entanglement artifacts commonly observed in models that process all spectral components jointly.
- **Uncertainty-Weighted Multi-Objective Optimization:** We address the theoretical challenge of task-conflicting gradients—where optimization for high-frequency restoration conflicts with low-frequency tasks—by adapting a learnable, likelihood-uncertainty-based loss function. This balancing mechanism ensures that dominant degradations do not suppress the detection and restoration of subtle, co-occurring artifacts, providing a stable optimization trajectory for complex composite training.

1.8 Thesis structure

The remainder of this thesis is organized as follows. Chapter 2 first introduces key fundamentals, then reviews related work, progressing from single-degradation restoration to multi-degradation and all-in-one restoration, and then analyzing degradation representation and conditioning mechanisms. Chapter 3 presents the proposed methodology, including the CDSM-Net architecture, multi-label semantic modulation, spatial gating, the WAVEViM backbone, and the training objective. Chapter 4 describes the experimental setup and implementation details, and reports quantitative results, real-world no-reference evaluation, ablations, and comparisons with baselines. Finally, Chapter 5 summarizes the key findings, concludes the thesis, and outlines limitations and future research directions.

Chapter 2

Literature review

2.1 Foundations

This section discusses the key foundations required for the proposed solution.

2.1.1 Image Formation

To understand image degradation and restoration, we first recall standard image-formation models from digital imaging and computer vision. In the physical world, visual information exists as a continuous distribution of light energy. We denote scene radiance by $L(x, y, \lambda, t)$, representing light intensity at spatial coordinates (x, y) , wavelength λ , and time t [2]. The imaging process converts this continuous radiance into a discrete digital image $I[u, v]$ through optical integration, sampling, quantization, and subsequent processing [49]. The equations in this subsection are therefore used as standard background formulations and notation, not as new derivations of this thesis.

Continuous to Discrete Transformation: Following standard imaging treatments [2, 49, 50], a camera sensor integrates incoming radiance over exposure time and wavelength (weighted by the sensor spectral sensitivity), producing an irradiance signal that is then discretized. To obtain a digital image, we apply **sampling** (spatial discretization into pixels indexed by (u, v)) and **quantization** (mapping sensor voltages to finite bit-depth).

1. **Sampling:** The sensor array discretizes the spatial domain into pixels with pitch $(\Delta x, \Delta y)$, so $I_{discrete}[u, v]$ represents the sampled irradiance at the pixel grid.
2. **Quantization:** The analog sensor readout is converted into a digital integer $I[u, v] \in [0, 2^B - 1]$ (bit depth B) through a quantization function with sensor gain and offset terms, following standard digital imaging formulations [2].

The Camera Image Signal Processing (ISP) Pipeline: Real-world images typi-

cally undergo a non-linear ISP pipeline before being saved, altering noise statistics and the appearance of degradations [49]. In practice, this pipeline includes demosaicing, white balance, color correction, tone mapping, and gamma correction applied to raw sensor data before conversion to the final sRGB image.

Sensor Noise and Dynamic Range: A fundamental constraint in image formation is **noise**. In practice, sensor noise is heteroscedastic: it includes a signal-independent read noise term and a signal-dependent shot noise term. Following the practical Poisson–Gaussian sensor-noise modeling literature [3], we use the common approximation:

$$y \sim \mathcal{N}(x, \sigma_{read}^2 + \sigma_{shot}^2(x)), \quad [3] \quad (2.1)$$

where x denotes the latent clean sensor signal, y the observed noisy measurement, σ_{read}^2 the read noise variance, and $\sigma_{shot}^2(x)$ the intensity-dependent shot noise variance [3].

Furthermore, sensors have finite **Dynamic Range (DR)**. When irradiance falls outside the measurable range, values clip, permanently losing information in shadows/highlights. This is a major cause of failure in low-light or saturated regions and motivates restoration methods that can reconstruct structure despite missing measurements [4].

2.1.2 Image Degradation Overview

The physical acquisition pipeline described in Section 2.1.1 establishes that digital images are inevitably approximations of the continuous scene radiance. While ideal imaging assumes a lossless mapping, real acquisition imposes a transformation $\mathbf{y} = \Phi(\mathbf{x})$, where $\mathbf{x}$ denotes the latent clean image, $\mathbf{y}$ the observed degraded image, and $\Phi(\cdot)$ the (often unknown) degradation operator. Historically, literature treated Φ as a single deterministic operator; however, modern applications increasingly exhibit stochastic, spatially variant, and composite degradations that occur together.

2.1.2.1 Image Degradation Models

The evolution of restoration literature is fundamentally a progression from linear, additive assumptions toward non-linear, composite formulations.

The Linear Space-Invariant Assumption. Following the classical linear degradation model in image restoration, the observed image is often written as convolution with a blur kernel $\mathbf{k}$ plus additive noise $\mathbf{n}$ [51]:

$$\mathbf{y} = \mathbf{k} * \mathbf{x} + \mathbf{n}, \quad [51] \quad (2.2)$$

where $*$ denotes convolution, $\mathbf{x}$ is the clean image, and $\mathbf{y}$ is the observed image. While this enabled foundational filters (e.g., Wiener), it cannot model signal-dependent noise, clipping, and many non-linear ISP effects, limiting realism in HDR and low-light settings.

The Atmospheric Scattering Model. To capture weather-specific corruptions, physics-based formulations were introduced. Following the standard atmospheric scattering model used in dehazing literature [52, 33], haze/fog is described as:

$$\mathbf{y}(x) = \mathbf{J}(x)t(x) + \mathbf{A}(1 - t(x)), \quad [52, 33] \quad (2.3)$$

where $\mathbf{J}(x)$ is the clean scene radiance at location x , $\mathbf{A}$ is global atmospheric light, and $t(x)$ is the transmission map (often depth-dependent). Although effective for dehazing, it typically assumes simplified conditions and may fail when haze co-occurs with heterogeneous degradations such as rain streaks or local blur.

A critical synthesis of recent literature reveals a gap in modeling *mixed degradations*. In real scenes, degradations may form a nested chain where the order of operations matters (non-commutative), and different composite processes can yield similar observations. This ambiguity makes inversion severely ill-posed and motivates restoration frameworks that leverage stronger priors (including semantic cues) to disambiguate mixed corruptions.

2.1.2.2 Types of Image Degradation

We classify common degradations into categories to highlight why mixed settings are challenging: they differ in spatial extent, frequency characteristics, and whether they corrupt or occlude information.

Noise Artifacts. Early work often assumed signal-independent AWGN, but real sensors exhibit signal-dependent noise, especially in low-light where shot noise dominates [3]. Consequently, assuming a single global noise level can over-smooth dark textures while under-denoising brighter regions.

Blur and Optical Aberrations. Blur suppresses high-frequency content and may arise from defocus or camera/scene motion. In dynamic scenes, blur can be spatially variant (different motion across the image), so restoration may require estimating a kernel field or relying on strong learned structural priors [53].

Environmental and Weather Corruptions. Weather degradations often occlude information. Rain introduces streak layers and partial occlusions, and low-light conditions amplify noise and can introduce non-linear color shifts via ISP processing [54, 55]. The complexity of these spatially variant and composite degradations makes purely analytical priors insufficient, motivating learning-based restoration methods discussed next.

2.1.3 Image Restoration Overview

The degradation models above define a forward process $\mathbf{y} = \Phi(\mathbf{x}) + \mathbf{n}$ that destroys information. Image Restoration (IR) therefore seeks an estimate $\hat{\mathbf{x}}$ from $\mathbf{y}$ such that $\hat{\mathbf{x}} \approx \Phi^{-1}(\mathbf{y})$, an ill-posed problem that requires priors.

2.1.3.1 Classical Image Restoration as an Inverse Problem

A classical restoration view formulates the problem as regularized inverse estimation, where a data-fidelity term is balanced against a prior or regularizer on the latent image [56, 9]. While effective for mild degradations, hand-crafted priors struggle to reconstruct complex textures and to model non-linear composite degradations.

2.1.3.2 Sparse Coding and Dictionary Learning

To better capture local structure, sparse representation assumes a patch $\mathbf{p} \in \mathbb{R}^d$ can be approximated using a few atoms from an over-complete dictionary $\mathbf{D} \in \mathbb{R}^{d \times K}$ ($K \gg d$) [8]. Restoration then solves for sparse coefficients $\boldsymbol{\alpha}$ such that $\mathbf{p} \approx \mathbf{D}\boldsymbol{\alpha}$. Although sparse coding improves texture modeling compared to simple smoothness priors, it requires

costly iterative inference and remains primarily local, limiting performance on large-scale degradations such as haze and heavy rain [9].

2.1.3.3 Deep Learning Foundations for Image Restoration

Deep learning replaces iterative optimization with an end-to-end learned inverse mapping. **CNNs** efficiently learn hierarchical local features and commonly use residual learning to predict artifact maps instead of clean pixels directly [57, 58]. However, locality can limit robustness when restoration requires long-range context. **Vision Transformers** address this by enabling global interactions via self-attention and have shown strong results in restoration tasks requiring non-local reasoning (e.g., dehazing), though their cost grows quickly with resolution [11].

More recently, Structured State Space Models (Mamba/SSMs) provide a global receptive field with linear complexity by propagating information through a learned state representation, making them attractive for high-resolution restoration under memory constraints [59]. Nevertheless, these backbones are generic: they do not explicitly represent what degradation is present. This motivates conditioning and guidance mechanisms that identify and modulate restoration behavior under mixed and non-exclusive degradations, which we review in subsequent sections.

2.2 Image Restoration Techniques

Having established the physical models of mixed degradation and the quantitative benchmarks for evaluation, we now critically synthesize the algorithmic evolution of Image Restoration (IR). The progression of this field can be viewed as a continuous shift in the locus of modeling: moving from hand-crafted analytical priors describing image statistics [56] to data-driven parametric models approximating the inverse degradation function Φ^{-1} directly.

The historical trajectory begins with variational methods and sparse coding, which offered mathematical guarantees but struggled with the non-linear complexity of real-world corruptions. This limitation catalyzed the era of Convolutional Neural Networks (CNNs), which excelled at local pattern matching but lacked long-range reasoning. Current state-of-the-art research has thus converged toward **Hybrid Architectures** integrating Trans-

formers, State Space Models (Mamba), and Wavelet representations—and explicitly incorporating **Semantic Guidance** to navigate the ill-posed nature of composite degradations.

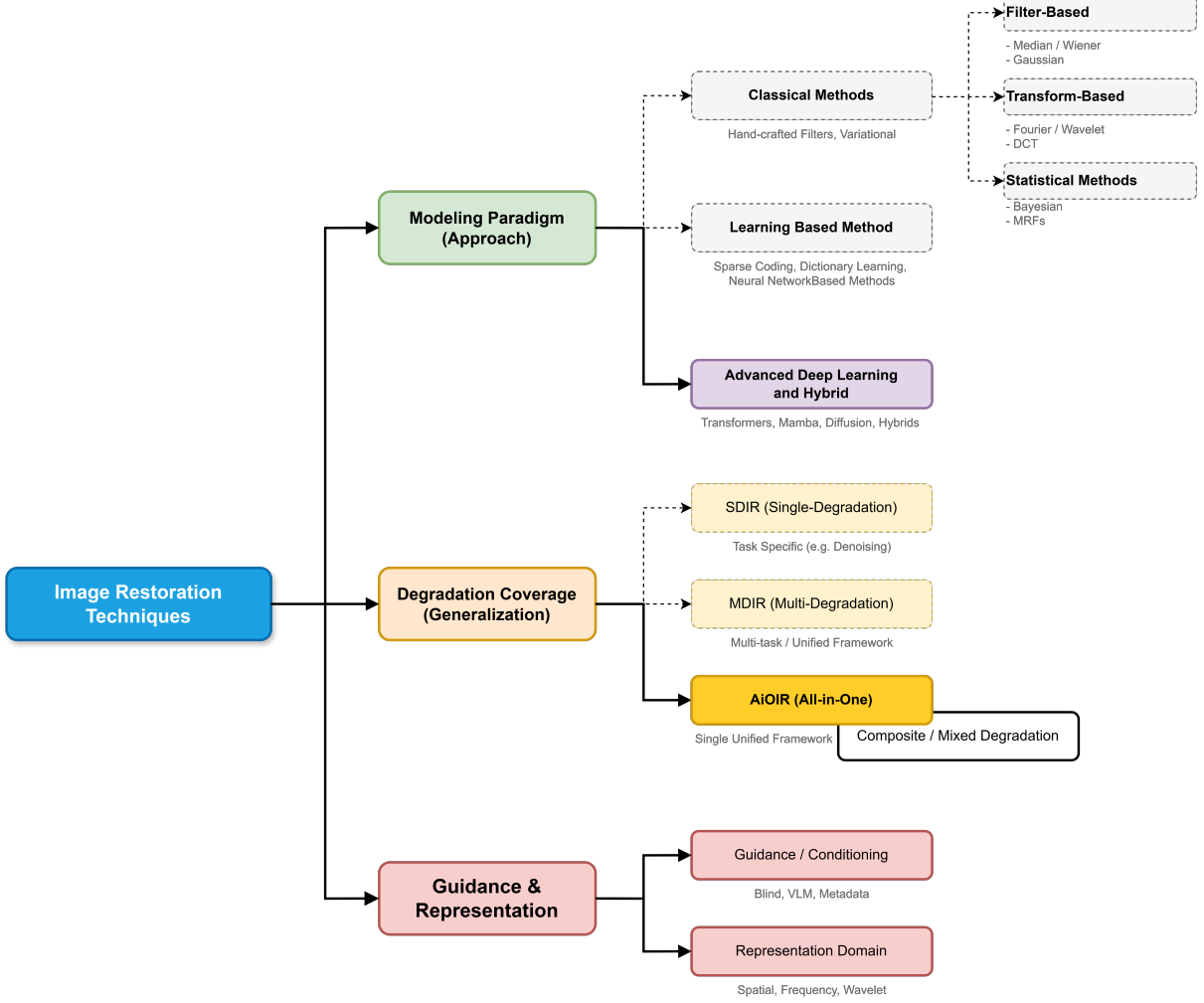


Figure 2.1: A hierarchical taxonomy of Image Restoration techniques.

As illustrated in taxonomy Figure 2.1, We organizes image restoration methods in three primary axes. In the taxonomy figure above, this thesis literature review only considers the solid lines and broken-line branches will not be explored in this literature review. First, the **Modeling Paradigm** distinguishes between classical optimization and end-to-end deep learning. Second, **Degradation Coverage** categorizes models by their generalization capability: from Single-Degradation (SDIR) to Multi-Degradation (MDIR) and finally All-in-One (AiOIR) frameworks capable of handling mixed artifacts. Third, **Guidance and Representation** differentiates models based on their use of external priors (e.g., text, metadata) and their computational domain (spatial vs. frequency). In

alignment with the research objectives of this thesis, our analysis focuses on the solid-line trajectories specifically **AiOIR** architectures leveraging **Advanced Hybrid Backbones** and **Semantic Guidance** while treating classical methods as necessary theoretical precursors.

2.2.1 Classical and Early Learning-Based Approaches

Before the advent of deep learning, restoration was predominantly framed as an optimization task leveraging analytic priors. **Classical Methods** typically assumed a linear degradation model $\mathbf{y} = \mathbf{k} * \mathbf{x} + \mathbf{n}$ and solved a regularized inverse problem using priors such as Total Variation and Tikhonov regularization [56, 9]. While these methods provided closed-form or convex solutions in simplified settings, they fundamentally lack semantic awareness.

The introduction of **Sparse Coding** and Dictionary Learning marked an intermediate evolution. By learning an over-complete dictionary $\mathbf{D}$ such that patches $\mathbf{x}_i \approx \mathbf{D}\alpha_i$ with sparse α_i [8], these methods adapted to local image statistics better than fixed analytical priors. However, they remained bound by two critical limitations:

1. **Computational Inefficiency:** Inference required iterative optimization (e.g., OMP or ISTA) for every image, precluding real-time application.
2. **Local-Global Disconnect:** Operating strictly on small patches (e.g., 8×8), sparse coding failed to model global degradation phenomena such as illumination gradients or large-scale blur trajectories.

These limitations highlight a fundamental gap: classical and sparse models are effective only when the degradation Φ is simple, known, and mathematically well-behaved. They falter entirely under the *composite, non-linear, and semantically dependent* corruptions typical of modern autonomous driving datasets (e.g., CDD-11). This necessitates the shift toward **Deep Learning and Hybrid Architectures**, which replace explicit optimization with learned parametric approximations capable of capturing the full complexity of the inverse mapping.

2.2.2 Deep Learning and Hybrid Architectures

In the previous section’s taxonomy we distinguished between **what** to restore (the degradation representation and guidance signal) and **how** to restore (the backbone that maps degraded images to clean estimates). Deep neural networks sit squarely in the latter category: they implement a parameterized inverse operator that attempts to undo an unknown degradation by exploiting statistical regularities in natural images. In this section, we therefore concentrate on deep and hybrid restoration backbones—from early CNN and U-Net models to Transformer, Mamba/state-space, diffusion-based, and hybrid designs—and analyze how they trade off local detail fidelity, global context modeling, efficiency, and robustness to mixed degradations. Issues of explicit degradation representation and conditioning will be treated in Sections 2.3.2 and 2.3.3; here we deliberately focus on the architectural side of the inverse problem.

Evaluating every individual paper by quantitative benchmarks is infeasible in a single thesis section. Instead, we organize our synthesis around eight qualitative properties that repeatedly emerge in the literature on all-in-one and mixed-degradation restoration:

- **C1: Local fidelity** (edge/texture preservation and avoidance of over-smoothing).
- **C2: Global context** (ability to model large-scale structures, haze fields, long-range correlations).
- **C3: Efficiency** (computational and memory cost at high resolution, suitability for deployment).
- **C4: Mixed-degradation robustness** (behavior when multiple degradations co-exist or compositions deviate from training distributions).
- **C5: Controllability and interpretability** (how explicitly the backbone exposes control over “what is restored”).
- **C6: Degradation disentanglement** (ability to separate corruption features from semantic content or between different degradation types).
- **C7: Spatial adaptivity** (ability to adapt restoration strength and type across spatially non-uniform degradations).

- **C8: Deployment complexity** (number of separate models, training pipelines, and ancillary components required).

The subsections below review the main backbone families through this lens. We are not aiming for an exhaustive catalogue; instead, we seek to identify common patterns, limitations, and gaps that later motivate explicit degradation representation and guidance mechanisms.

2.2.2.1 CNN/U-Net family

Convolutional neural networks (CNNs) and U-Net-like encoder–decoders were the first dominant deep architectures for low-level restoration. DnCNN [58] learns a residual mapping $\mathcal{R}_\theta(\mathbf{y}) \approx \mathbf{n}$ and recovers $\hat{\mathbf{x}} = \mathbf{y} - \mathcal{R}_\theta(\mathbf{y})$ for Gaussian denoising. IRCNN [60] embeds a CNN denoiser as a learned prior within an optimization framework, bridging classical MAP estimation with deep priors. Residual dense networks (RDN) [61] introduce dense connectivity inside residual blocks to stabilize deep architectures for super-resolution, while MIRNet [62] and its successors exploit multi-scale residual aggregation and attention for real-image denoising and enhancement. More recent designs such as NAFNet [14] demonstrate that carefully simplified residual blocks can match state-of-the-art performance while removing many nonlinearities and normalizations.

From a modeling viewpoint, these networks implement a composition of local convolutions $\mathbf{F}_{\ell+1} = \phi_\ell(\mathbf{W}_\ell * \mathbf{F}_\ell)$ stacked within an encoder–decoder topology. Their strengths are well understood: **strong local fidelity (C1)**, where convolutions with skip connections preserve edges and textures and give stable PSNR/SSIM for single degradations; and **good efficiency (C3)** combined with **relatively low deployment complexity (C8)**, as convolutional operators scale linearly with image size and hardware support is mature.

However, the same inductive bias creates limitations. **Global context (C2)** is limited because the receptive field grows only linearly with depth and kernel size, which is problematic for large-scale haze, glow, or non-local rain streaks. Furthermore, **mixed-degradation robustness (C4)** is weak when multiple artifacts co-exist; features corresponding to noise, blur, and details are easily entangled, leading to over-smoothing or residual artifacts. Finally, there is **little explicit disentanglement (C6)**, **spatial**

adaptivity (C7), or **controllability (C5)**, as the mapping f_θ is largely blind to the specific degradation composition. These limitations largely motivated the shift towards attention-based and more globally-aware backbones.

2.2.2.2 Transformers for restoration

Transformer-based restorers were introduced to address CNNs’ lack of explicit global modeling. SwinIR [11] replaces convolutions with Swin Transformer blocks using windowed self-attention in a hierarchical fashion, and achieves strong results on super-resolution, denoising, and compression artifact removal. Restormer [13] proposes multi-Dconv head transposed attention and gated feed-forward networks, offering an efficient attention backbone for high-resolution restoration. Uformer [12] embeds locally-enhanced window attention in a U-shaped architecture, while U²-Former [63] nests multiple U-shaped Transformer structures to strengthen multi-scale reasoning. MAXIM [15] shows that spatially-gated MLPs can play a similar role to attention in a carefully designed UNet-like hierarchy.

The core operator is self-attention,

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \mathbf{V}, \quad [64] \quad (2.4)$$

which allows each token to attend to all others. In restoration, this helps capture long-range correlations such as globally consistent haze thickness or repeated structural patterns. When guidance or auxiliary tokens are available, this mechanism naturally generalizes to cross-attention, where queries come from image features and keys/values from condition tokens, although most backbone papers still focus on self-attention only.

In terms of the eight properties, Transformers significantly improve global context (C2) and maintain high local fidelity (C1), but at the cost of: **higher computational and memory cost (C3)**, where global attention scales as $\mathcal{O}(N^2)$ in the number of tokens N and windowed attention remains a compromise with only approximate global context; **mixed-degradation robustness (C4)** that still relies on very diverse training data rather than explicit modeling of degradation compositions; and **limited progress on controllability, disentanglement, and spatial adaptivity (C5–C7)**, as most models are still trained as blind regressors. These practical limitations have motivated

research into architectures that retain long-range modeling with improved efficiency, such as Mamba and other state-space models.

2.2.2.3 Mamba / State-space restorers

State-space models (SSMs) and their recent selective variants (e.g., Mamba) treat images as sequences and model dependencies via learned linear dynamical systems. A generic SSM update can be written as

$$\mathbf{s}_{t+1} = \mathbf{A}\mathbf{s}_t + \mathbf{B}\mathbf{u}_t, \quad \mathbf{o}_t = \mathbf{C}\mathbf{s}_t, \quad [65] \quad (2.5)$$

where $\mathbf{u}_t$ is the input token and $\mathbf{o}_t$ the output. Mamba-style architectures implement selective scan mechanisms that achieve long-range dependency modeling with linear complexity in sequence length.

VMamba [66] demonstrates that selective state-space blocks can serve as competitive vision backbones on classification and dense prediction, with better scaling than Transformers. MambaIR [67] ports this idea to restoration, combining Mamba blocks with local enhancement modules to address pixel forgetting and channel redundancy. Variants such as EAMamba and related designs integrate SSM blocks into U-shaped encoder-decoders for efficient, all-around restoration [68].

Compared to Transformers, Mamba-style restorers exhibit a different trade-off profile: they offer strong global context (C2) with more favorable efficiency (C3) at high resolution, thanks to linear scanning; and they preserve competitive local fidelity (C1) when combined with convolutional or local enhancement modules.

At the same time, they inherit several open issues common to attention-based backbones: they are architecturally younger, so training stability and best practices for multi-scale integration are still evolving (affecting deployment complexity, C8); and they typically remain blind to degradation semantics, and thus do not directly address mixed-degradation robustness, disentanglement, or spatial adaptivity (C4–C7). Thus, while SSMs alleviate the cost of global modeling, they do not by themselves resolve the ambiguity introduced by composite degradations.

2.2.2.4 Diffusion priors for restoration

Diffusion models provide a different perspective: instead of learning a direct inverse mapping, they learn a generative prior over clean images and perform restoration through an iterative denoising process. Using the standard Gaussian noising form adopted in diffusion-based restoration literature [69, 70, 71], we write the forward process as

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \quad [72] \quad (2.6)$$

and a learned reverse process produces samples $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$ via a denoiser ϵ_θ . Palette [69] shows that a single conditional diffusion backbone can solve diverse image-to-image tasks (colorization, inpainting, JPEG artifact removal) with excellent perceptual quality. DDRM [70] uses a pre-trained diffusion model as an unsupervised prior for generic linear inverse problems, covering super-resolution, deblurring, and inpainting without task-specific retraining. Plug-and-play approaches integrate diffusion denoisers into iterative optimization for restoration [71].

Historically, GAN-based restorers such as DeblurGAN-v2 [73] and ESRGAN [74] also pursued perceptual realism using adversarial losses. While GANs offer sharp textures and low perception-distortion error, they suffer from training instability and hallucination risk, which becomes particularly problematic for safety-critical applications.

From the perspective of the eight properties, diffusion and related generative priors exhibit a characteristic profile: they offer strong global context and high perceptual fidelity (C1, C2), often outperforming purely discriminative models on perceptual metrics; they are computationally expensive (C3), with many denoising steps required, which complicates real-time deployment (C8); and unless explicitly conditioned on degradation information, they also do not automatically disentangle or spatially adapt to mixed degradations (C4–C7), acting instead as powerful but generic priors. Consequently, many recent works employ diffusion not as a full-resolution backbone, but as a latent- or low-frequency refiner on top of faster discriminative networks [22].

2.2.2.5 Hybrid Deep Architectures

Hybrid architectures explicitly combine different modeling principles to compensate for the limitations of any single backbone family. Instead of asking one network to satisfy all properties C1–C8 simultaneously, they decompose the restoration process into sub-modules that specialize along different dimensions: domain (spatial vs. frequency), scale (low vs. high resolution), semantics (structure vs. texture), or algorithmic style (feed-forward vs. iterative).

Hybrid restoration pipelines combine multiple transforms, subnetworks, and fusion stages so that different modules can specialize in complementary representations such as spatial/frequency domains or local/global reasoning. Simple instances include two-branch models separating low- and high-frequency content, while more complex hybrids combine CNN, Transformer, and state-space blocks within the same architecture, or couple a fast feed-forward backbone with a slower generative refiner.

Domain and frequency hybrids. Wavelet and frequency-aware models such as WNet [75] and ANet [76] decompose the input using a discrete wavelet transform, route low-frequency and high-frequency bands through dedicated convolutional branches, and then fuse them. This explicit separation improves local fidelity (C1) and degradation disentanglement (C6) for artifacts that are strongly frequency localized (e.g., moiré patterns, ringing, structured rain). Similar ideas appear in later deraining and deblurring methods that embed DWT or FFT constraints at intermediate stages.

Backbone hybrids (CNN/Transformer/SSM). Another line of work mixes backbones at different depths: shallow stages may use CNNs to exploit local inductive bias, while deeper stages use Transformers or Mamba blocks to capture global context. U-shaped or nested architectures [63, 15, 66, 67, 68] often alternate between local and global modules, allowing different stages to capture complementary multi-scale representations. These hybrids improve global context (C2) compared to pure CNNs while retaining reasonable efficiency (C3) compared to full-depth Transformers.

Backbone + generative prior hybrids. Hybrid systems also appear when discriminative backbones are combined with diffusion or GAN priors. For example, DeblurGAN-v2 [73] uses a feed-forward generator with an adversarial loss to improve perceptual quality, and ESRGAN [74] refines SR outputs by adding a GAN-based perceptual head. More

recent all-in-one diffusion frameworks such as AutoDIR [22] couple a discriminative backbone with a latent diffusion model: the backbone provides coarse restoration, while the diffusion prior refines visual details and corrects ambiguous regions. In this hybrid setting, the backbone and the diffusion prior can be viewed as two specialized components with different computational footprints and priors.

Across these designs, hybrid backbones show clear advantages: they improve local fidelity and global context simultaneously (C1, C2) by assigning different responsibilities to different modules (e.g., LF vs. HF branches, local vs. global blocks); they allow more flexible efficiency/performance trade-offs (C3, C8), for example by limiting expensive attention or diffusion operations to selected stages or domains; and frequency-domain or multi-branch decompositions partially improve disentanglement and spatial adaptivity (C6, C7), especially for degradations with distinct spectral signatures.

However, the literature reveals several persistent gaps that are especially problematic for mixed-degradation and real-world scenarios: **implicit degradation reasoning**, as most hybrids still treat degradation understanding as an implicit by-product of feature learning without exposing an explicit representation of mixture or layout; **limited robustness to unseen mixtures**, where networks trained on synthetic combinations often overfit and degrade under complex, non-commutative compositions like (2.8); and **weak controllability**, as internal routing decisions are usually fixed after training, preventing practitioners from specifying **what** to emphasize or suppress at inference time. These limitations indicate that backbone design alone is insufficient to address mixed-degradation and real-world restoration. In addition to powerful CNN, Transformer, Mamba, diffusion, and hybrid backbones, there is a growing need for explicit mechanisms that represent degradation states and modulate the backbone accordingly.

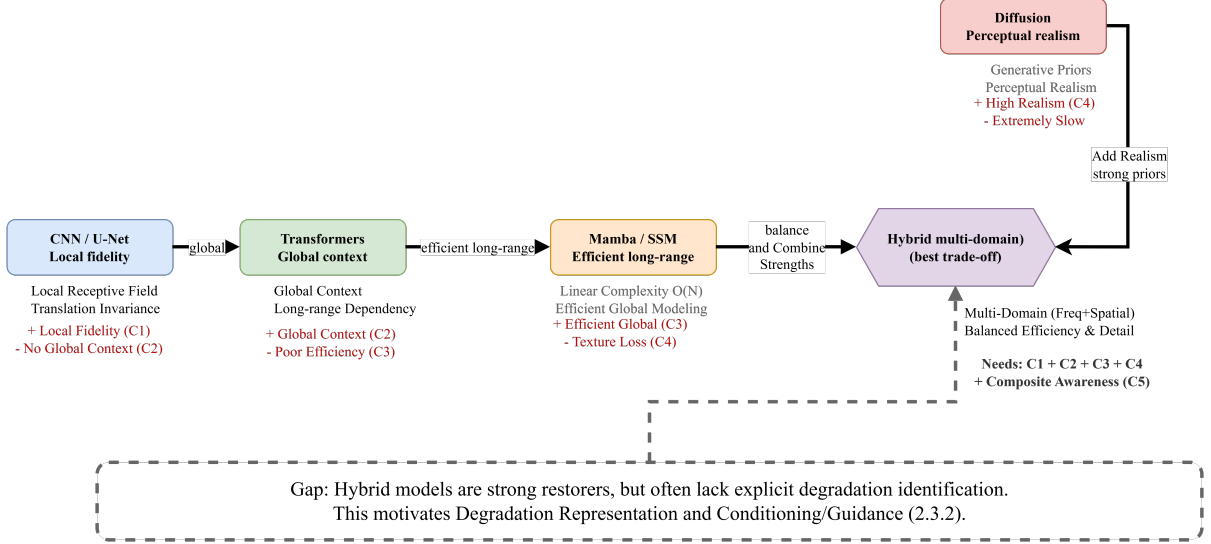


Figure 2.2: Conceptual evolution of deep and hybrid image restoration backbones

As we summarized the gap oriented Conceptual evolution of deep and hybrid image restoration backbones in Figure 2.2, CNNs emphasize local fidelity; Transformers introduce global context; state-space/Mamba models improve efficiency for long-range modeling; diffusion priors enhance perceptual realism; and hybrid architectures combine complementary strengths to approximate a better trade-off. However, despite their capability, most hybrids still lack explicit degradation identification and guidance, motivating the need for degradation representation and conditioning mechanisms discussed in later sections.

2.2.2.6 Summary of Deep and Hybrid Image Restoration Techniques

Table 2.1 summarizes the major backbone families discussed above in terms of the qualitative properties C1–C8 and lists representative papers for each family. Each cell indicates a qualitative assessment: ✓ (strong), △ (moderate), and × (weak).

Table 2.1: Qualitative comparison of deep backbone families for image restoration.

Family	C1	C2	C3	C4	C5	C6	C7	Representative papers
CNN / U-Net	✓	△	✓	△	×	×	△	DnCNN [58], RDN [61], NAFNet [14], IRCNN [60], MIRNet [62]
Transformer / MLP	✓	✓	△	△	△	×	△	SwinIR [11], Uformer [12], MAXIM [15], Restormer [13], U ² -Former [63]
Mamba / SSM	✓	✓	✓	△	×	×	△	VMamba [66], MambaIR [67], EA-Mamba [68]
Diffusion / GAN priors	✓	✓	×	△	△	△	△	Palette [69], DDRM [70], PnP diffusion [71], DeblurGAN-v2 [73], ESRGAN [74]
Hybrid / frequency / backbone+prior	✓	✓	△	△	△	△	△	WDNet [75], AWWNet [76], AutoDIR [22]

C1 local fidelity, C2 global context, C3 computational efficiency, C4 robustness under mixed degradations, C5 controllability and interpretability, C6 degradation disentanglement, C7 spatial adaptivity.

Overall, we observe that CNN/U-Net and multi-stage CNNs are strong on local fidelity and efficiency but limited in global context and explicit mixed-degradation handling; Transformers and Mamba backbones substantially improve global reasoning, with Mamba offering better scaling, but they typically remain blind to the precise degradation composition; diffusion and GAN priors provide excellent perceptual quality at significant computational cost; and wavelet or backbone+prior hybrids occupy a favorable region of the quality–efficiency space by combining complementary modules. Yet across all families, degradation understanding is largely implicit, which becomes a serious limitation under real-world mixed degradations. This motivates the next step in our review, where we move from the general image restoration techniques view to degradation coverage and generalization (Section 2.2.3), backbone design patterns (Section 2.3.1), and then to the joint discussion of degradation representation and conditioning in AiOIR (Section 2.3.2).

2.2.3 Degradation Coverage and Generalization in Image Restoration

A convenient way to formalize **coverage** and **generalization** is to treat restoration as learning an inverse map under a **degradation state** variable. Let the clean image be

$\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ and the observed image be

$$\mathbf{y} = \mathcal{D}(\mathbf{x}; \mathbf{z}) + \boldsymbol{\varepsilon}, \quad (2.7)$$

where $\mathcal{D}$ is a (possibly non-linear) corruption operator parameterized by a state $\mathbf{z}$ (type, strength, and potentially spatial layout), and $\boldsymbol{\varepsilon}$ is stochastic noise. In mixed/composite settings, $\mathbf{z}$ is **multi-factor** and the forward process is often **non-commutative**:

$$\mathbf{y} = \left(\mathcal{D}_{t_m}^{(m)} \circ \cdots \circ \mathcal{D}_{t_1}^{(1)} \right) (\mathbf{x}) + \boldsymbol{\varepsilon}, \quad \mathcal{D}_a \circ \mathcal{D}_b \neq \mathcal{D}_b \circ \mathcal{D}_a. \quad (2.8)$$

Given a restorer f_θ , its **generalization** is captured by the risk on a target degradation distribution:

$$\mathcal{R}_{\text{test}}(f_\theta) = \mathbb{E}_{(\mathbf{x}, \mathbf{z}) \sim p_{\text{test}}} [\ell(f_\theta(\mathbf{y}), \mathbf{x})], \quad \mathbf{y} \text{ from (2.7)}. \quad (2.9)$$

The evolution SDIR $\rightarrow$ MDIR $\rightarrow$ AiOIR can be read as expanding the supported set $\mathcal{Z}$ of states while controlling the **generalization gap** induced by shifts $p_{\text{train}}(\mathbf{z}) \neq p_{\text{test}}(\mathbf{z})$, under practical constraints (data, compute, and deployment); in this thesis, **multi-degradation** refers to the *coverage/capability setting* in which one shared model is trained to handle several degradation families, whereas **mixed degradation** refers to the *input condition* in which two or more degradations co-exist in the same observation, so a mixed image may indeed be produced by sequential composition of degradations, but the two terms are not synonymous because the former describes what the model is expected to cover and the latter describes what is present in a particular image [77, 18, 21].

Single-Degradation Image Restoration (SDIR). SDIR fixes the degradation family (often even its noise model), i.e., $\mathbf{z} \equiv (t, \phi)$ with a single t , and learns a specialized inverse f_{θ_t} that minimizes $\mathcal{R}_{\text{test}}$ only on that narrow support. This yields strong in-domain performance because the model can internalize task-specific inductive biases; canonical examples include discriminative denoisers such as DnCNN [58] and denoiser-prior formulations such as IRCNN [60]. However, the same specialization becomes a liability once t changes or when ϕ drifts: the learned inverse approximates $\mathcal{D}(\cdot; \mathbf{z})^{-1}$ only on a thin manifold of states, so robustness decays rapidly under out-of-family corruptions and under compositions (2.8). Practically, SDIR scales poorly in deployment because supporting $|\mathcal{T}|$

degradation types typically requires a model set $\{f_{\theta_t}\}_{t \in \mathcal{T}}$ plus a selector, inflating storage, maintenance, and validation cost roughly as $\mathcal{O}(|\mathcal{T}|)$ even before considering mixtures.

Multi-Degradation Image Restoration (MDIR). MDIR expands coverage by training a shared model on multiple degradations, often optimizing a weighted risk

$$\min_{\theta} \mathbb{E}_{t \sim \pi(t)} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\cdot|t)} [\ell(f_{\theta}(\mathbf{y}), \mathbf{x})], \quad (2.10)$$

so a single parameter set must serve several t 's. Multi-task restorers such as MPRNet [36] show that careful architecture/training can share structure across tasks and reduce the need for per-degradation retraining, while plug-and-play lines (e.g., DPIR) push this idea by reusing a single deep denoiser prior across multiple inverse problems [78]. Yet, MDIR exposes two recurring tensions: (i) **negative transfer** (features beneficial for one degradation can harm another), and (ii) **state ambiguity** (the model must implicitly infer t and ϕ from $\mathbf{y}$). Consequently, MDIR often requires broader and more balanced training data to avoid bias toward dominant degradations, and its robustness to **unseen combinations** remains limited because the training distribution typically samples a small subset of the combinatorial space induced by (2.8).

All-in-One Image Restoration (AiOIR). AiOIR pushes coverage further toward a single deployable model that must handle **unknown** degradation type/strength and, increasingly, composite mixtures. In the notation of (2.7), AiOIR aims to learn f_{θ} with large supported state set $\mathcal{Z}$ while remaining stable under shifts in $p(\mathbf{z})$. Representative directions include (a) learning an internal degradation estimator and routing/adaptation mechanism as in AirNet [77] and adaptive-filter designs [79], (b) prompt-driven or task-token paradigms that enlarge the conditioning space without training separate models, e.g., PromptIR [19], and (c) explicit degradation-representation learning to improve invariance and discrimination across states [25]. For composite degradations specifically, OneRestore emphasizes that the central difficulty is not only stronger corruption but the **interaction** between factors (order, coupling, and spatial heterogeneity) and accordingly benchmarks restoration under mixed states [21].

Across $\text{SDIR} \rightarrow \text{MDIR} \rightarrow \text{AiOIR}$, the bottleneck shifts from designing a strong in-

verse for a fixed t to managing uncertainty over $\mathbf{z}$ under distribution shift. The resulting gap is consistent across recent AiOIR literature: reliable **degradation-state representation** (especially multi-label $\mathbf{z}$ for composite inputs), robustness to **unseen mixtures** and spatially varying compositions in (2.8), and controllable adaptation with acceptable compute for deployment.

2.3 All-in-One Image Restoration (AiOIR)

Building on the SDIR→MDIR→AiOIR progression (Section 2.2.2), we focus on **all-in-one image restoration**—the setting where a **single** model must restore images under **unknown** and potentially **mixed** degradations. Unlike SDIR, AiOIR requires a **control mechanism**: the model must infer **what** corruption is present and modulate **how** to remove it within one shared framework. The fundamental challenge shifts from defining an inverse operator to defining a **control mechanism**: the model must correctly identify **what** is wrong and significantly modulate **how** to fix it within a single unified framework.

A common abstraction models an observed degraded image $\mathbf{y} \in \mathbb{R}^{H \times W \times 3}$ as

$$\mathbf{y} = \mathcal{D}(\mathbf{x}; \mathbf{z}) + \boldsymbol{\epsilon}, \quad (2.11)$$

where $\mathbf{x}$ is the clean target, $\boldsymbol{\epsilon}$ is noise, and $\mathbf{z}$ denotes a latent **degradation state**. In SDIR, $\mathbf{z}$ is typically assumed to lie in a single family (e.g., blur-only); in MDIR, $\mathbf{z}$ spans multiple families but often with limited compositional diversity; in AiOIR, $\mathbf{z}$ is **unknown at inference** and may represent a **mixture** of degradations, e.g., $\mathbf{z} = (\alpha_1, \dots, \alpha_M)$ with $\alpha_m \geq 0$ and $\sum_m \alpha_m = 1$, or even spatially-varying maps $\mathbf{z}(\mathbf{p})$ over pixels $\mathbf{p}$.

AiOIR then seeks a **single** restoration model f_θ that generalizes over a wide degradation set $\mathcal{Z}$:

$$\hat{\mathbf{x}} = f_\theta(\mathbf{y}; \mathbf{c}), \quad \mathbf{c} = r_\psi(\mathbf{y}, \mathbf{s}), \quad (2.12)$$

where $\mathbf{c}$ is a conditioning signal (possibly empty), produced by a representation module r_ψ from the input $\mathbf{y}$ and optional side information $\mathbf{s}$ (e.g., a task token, a prompt, or an instruction). Eq. (2.12) is intentionally generic: it covers degradation-agnostic AiOIR (set $\mathbf{c} = \emptyset$), label-/token-conditioned models, prompt-driven models, and interactive instruction-following systems.

This formulation necessitates a two-axis decomposition of the literature:

- **The Backbone (Restoration Engine) (Section 2.3.1):** what restoration engine is used and how the network is structured to share computation across tasks.
- **Degradation intelligence (Section 2.3.2):** (i) how degradation is represented (r_ψ) and (ii) how it is injected to control restoration (conditioning operator implicit in f_θ).

2.3.1 Backbones and Architectural Patterns in AiOIR

Early restoration systems often relied on **task-specific models** (one per degradation), which scales poorly and assumes prior knowledge of $\mathbf{z}$. Since AiOIR explicitly removes that assumption, we focus on the dominant **shared-parameter** designs that attempt to approximate a family of inverse operators within a single framework. Following common taxonomies in recent AiOIR literature [77, 18, 21, 80, 81, 82, 83, 27, 26, 84], we group backbone patterns into six families (Fig. 2.3).

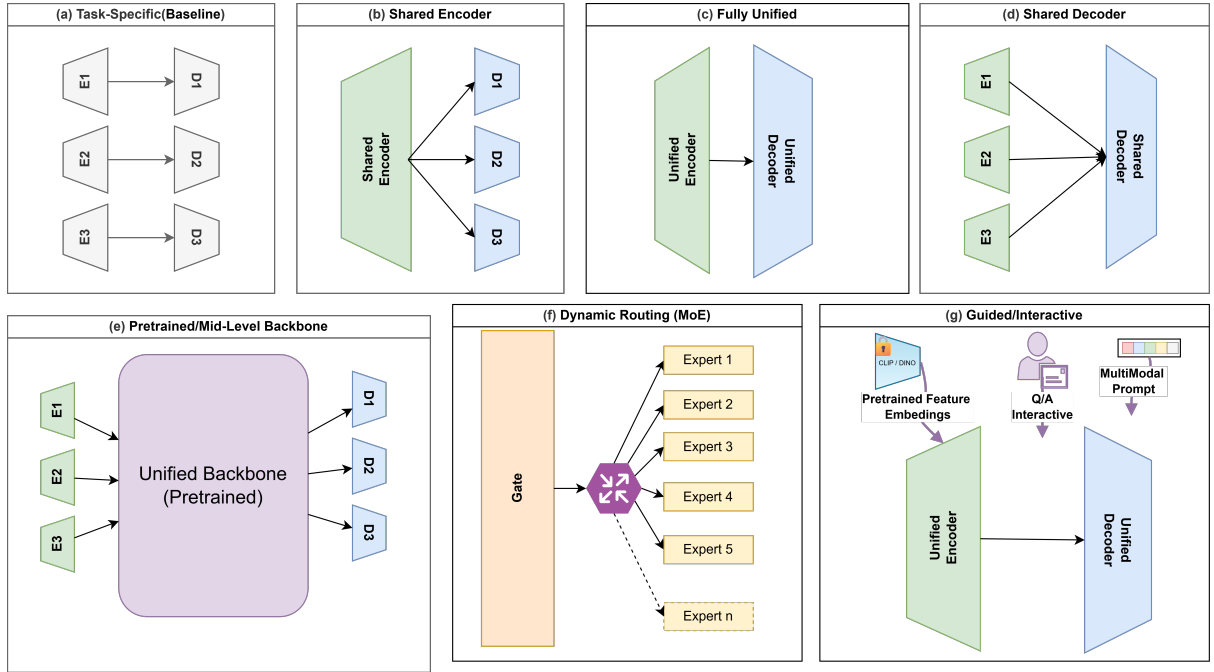


Figure 2.3: architectural design patterns in AiOIR

Evolution from shared-parameter multi-task designs to unified backbones, pretrained mid-level priors, dynamic routing (MoE), and guided/interactive systems. The key unresolved gap across families is that architectural strength does not automatically yield

explicit, reliable degradation identification under mixed/unknown degradations, motivating the joint degradation representation and conditioning discussion in Section 2.3.2.

A unifying view of these families is obtained by writing f_θ as either a single shared restorer or a mixture of specialized restorers:

$$\hat{\mathbf{x}} = \sum_{k=1}^K \pi_k(\mathbf{c}, \mathbf{y}) f_{\theta_k}(\mathbf{y}; \mathbf{c}), \quad (2.13)$$

where $K = 1$ reduces to a fully shared backbone, and $K > 1$ yields mixture-of-**experts** routing. This formulation lets us compare architectures by **where** sharing occurs (encoder, decoder, bottleneck), **how** specialization is expressed (f_{θ_k}), and **what** drives selection (π_k via $\mathbf{c}$).

(A) Shared Encoder with Multi-Head Decoders These models share an encoder $E_\phi(\mathbf{y})$ and branch into multiple heads/decoders specialized per degradation family. The key benefit is efficient reuse of low-level features while allowing head-level specialization. However, this pattern tends to **reintroduce hidden task selection**: either (i) a head must be chosen (explicitly or implicitly), or (ii) supervision must strongly disentangle head behaviors. In practice, as the number of degradations grows (and especially for mixtures), head boundaries become ambiguous, increasing training complexity and risking negative transfer across heads. Multi-task restoration backbones that inspired this strategy (shared trunk with task-specific modules) include representative multi-task designs such as MAXIM [15] and related shared-feature multi-restorers, but AiOIR settings increasingly move beyond explicit head separation due to mixed-degradation ambiguity.

The brittleness of discrete head specialization under unknown/mixed $\mathbf{z}$ motivates designs that remove explicit heads and rely on a fully unified restoration pathway.

(B) Fully Unified Encoder–Decoder Fully unified models use a single encoder–decoder without task-specific heads, aiming to learn a universal mapping $\mathbf{y} \mapsto \hat{\mathbf{x}}$ across degradations. A representative line is adverse-weather “one-model” restoration, e.g., TransWeather [85], which demonstrates that a single transformer encoder–decoder can learn to handle diverse weather corruptions without selecting a task branch. AirNet [77] further frames AiOIR as restoration under unknown corruption and uses a unified pipeline

to cover multiple degradations. Unified designs often inherit strong restoration engines (CNN/U-Net or transformer variants) from general restoration research—e.g., Restormer [13], NAFNet [14], MAXIM [15]—and then train them under multi-degradation supervision.

The unification that improves scalability also makes degradation reasoning mostly **implicit**: the same pathway must internally infer $\mathbf{z}$ and perform inversion. This often works for training-covered degradations, but mixed and long-tail combinations can produce failure modes such as (i) **under-restoration** (leaving residual haze/rain) or (ii) **over-restoration** (texture hallucination or detail wash-out), because the model lacks an explicit handle to separate what is “content” vs. “corruption”. This limitation becomes sharper as $\mathbf{z}$ becomes compositional or spatially varying.

To reduce ambiguity without multiplying full decoders, later designs retain a shared backbone but attach lightweight task heads or decoders with more structured parameterization.

(C) Shared Backbone with Lightweight Task Heads (or Multiple Decoders)

This family shares a common feature extractor/backbone and attaches multiple lightweight heads/decoders (sometimes per degradation, sometimes per cluster), aiming to balance universal features with limited specialization. Such designs can reduce negative transfer compared to a single fully unified decoder, but they inherit a key tension: increasing the number of heads helps specialization yet increases supervision needs and deployment complexity. BID is a representative example of using a shared backbone with multiple decoders in multi-degradation restoration benchmarks [86], illustrating the practical trade-off between expressiveness and scalability.

When mixed degradations occupy a continuous space, discrete head assignments become ill-posed unless the model also learns a reliable degradation representation and routing rule. Thus, this family naturally points toward either **pretraining** (to stabilize shared representations) or **dynamic routing** (to soften hard assignments).

Pretraining aims to strengthen the shared backbone so that fewer task-specific parameters are needed, shifting the burden from architectural branching to reusable mid-level representations.

(D) Pretrained Mid-Level Backbones Pretrained-backbone AiOIR reuses mid-level representations from large-scale pretraining to improve generalization and data-efficiency. IPT [10] is a representative pretraining-driven transformer for multiple low-level tasks, showing how shared representations can be adapted across restoration problems via task conditioning. Recent AiOIR systems increasingly combine such pretrained backbones (e.g., ImageNet/MAE-style initialization or large-scale pretraining) with lightweight adapters to reduce per-task parameter growth, and to improve robustness when training data cannot cover all mixtures.

While pretraining improves representation quality, it does not automatically solve **degradation disentanglement**: a strong backbone can still fail to **separate** multiple simultaneous degradations without explicit cues. Empirically, the strongest gains appear when pretraining is paired with mechanisms that estimate or encode degradation state (next family).

If explicit heads are too rigid and pure pretraining is insufficient, a natural compromise is **dynamic specialization** through routing, yielding mixture-of-experts AiOIR.

(E) Dynamic Routing / Mixture-of-Experts (MoE) MoE architectures instantiate Eq. (2.13) with $K > 1$ and a learned router $\pi_k(\mathbf{c}, \mathbf{y})$ that selects experts. The promise is to approximate a set of inverse operators while keeping inference efficient by activating only a subset of experts. Recent AiOIR works increasingly adopt this strategy: Multi-Expert Adaptive Selection / task-balancing frameworks explicitly address conflicts between restorations and expert selection to balance multi-task demands [87]; UniRestorer performs multi-granularity degradation estimation and expert selection to reduce sensitivity to degradation-estimation errors [26]; LoRA-IR injects low-rank expert adaptations with a degradation-guided router to improve efficiency and adaptability [27]; and M2Restore fuses Mamba-CNN streams with MoE gating for robust all-in-one restoration under complex mixtures [84].

The MoE gain depends critically on the **router signal $\mathbf{c}$** : if $\mathbf{c}$ cannot reliably represent multi-degradation composition (or if it collapses to single-label decisions), routing becomes unstable and may fail on novel mixtures. Moreover, many routers are global (image-level), while mixed degradations can be spatially localized, suggesting the need for richer degradation representations and spatially-aware conditioning (Sections 2.3.2–

2.3.3).

Routing improves specialization, but the strongest recent trend is to treat degradation understanding as a **semantic** problem—using prompts, instructions, and pretrained priors to generate $\mathbf{c}$ and modulate restoration.

(F) Guided and Interactive Architectures (prompts, priors, instructions, Q&A)

Guided AiOIR expands Eq. (2.12) by making $\mathbf{s}$ a human- or model-provided control signal and letting r_ψ leverage pretrained priors. A major direction uses **vision–language priors**: CLIP-based control and related VLM guidance provide semantically meaningful embeddings to steer restoration under unknown degradations [30]. DA-CLIP-style approaches explicitly train degradation-aware CLIP representations to guide restoration and improve robustness to out-of-distribution degradations [31]. Other works formulate restoration as **prompting**: PromptIR injects learnable prompts to adapt a unified restoration backbone across degradations [19], and ingredient-oriented reformulations like IDR encourage decomposition of degradations into compositional “ingredients” to better model multi-degradation mixtures [18].

A complementary interactive direction uses **instructions and question answering**. InstructIR treats restoration as instruction following, allowing natural language control over what to improve in the image [82]. PromptGIP unifies image processing under a visual prompting Q&A paradigm, enabling task selection and control through a structured question representation [81]. UniProcessor [83] similarly uses text-induced control features injected via cross-attention to support unified low-level processing with multimodal control and initial mixed-degradation handling capabilities. Diffusion-based universal restoration also appears in this guided space: AutoDIR frames AiOIR with latent diffusion and interactive prompts in a question-answering style [22]; “Restore Anything with Masks” (RAM) uses mask-based modeling to focus on content rather than explicit degradation classification [28]; and recent diffusion refinement variants further explore controllable restoration and realism under diverse degradations [88].

A comparative overview of these architectural paradigms is provided in Table 2.2. The literature suggests that backbone progress improves scalability and capacity allocation, but mixed-degradation robustness is still limited when degradation cues remain implicit or weakly grounded. Consequently, recent AiOIR research increasingly shifts

Table 2.2: Summary of AiOIR backbone families

Family	Feature(Task) Specialization	Scalability	Data Efficiency	Semantic Control	Rep. Works
Shared encoder + multi-head decoders	✓	×	×	×	MAXIM [15]
Fully unified encoder-decoder	×	✓	≈	×	TransWeather [85], AirNet [77], Restormer [13], NAFNet [14]
Shared backbone + lightweight heads	≈	≈	×	×	BIDeN [86]
Pretrained mid-level backbones	×	≈	✓	×	IPT [10]
Dynamic routing / MoE	✓	✓	≈	≈	MoE [87], UniRestorer [26], LoRA-IR [27], M2Restore [84]
Guided + interactive (Prompts / Priors)	≈	≈	≈	✓	PromptIR [19], IDR [18], InstructIR [82], PromptGIP [81], UniProcessor [83], DA-CLIP [31], AutoDIR [22]

Legend: ✓ = supported; ≈ = limited support; × = no/very low support.

attention from which backbone to use toward the joint problem of representing degradation states under compositional and spatially varying corruptions and injecting these cues into restoration through conditioning and guidance mechanisms such as modulation, cross-attention, routing, prompts, and frequency-aware control. This transition is especially important for multi-label $\mathbf{z}$, where mis-representation or mis-conditioning is a common source of artifacts under real mixed degradations.

2.3.2 Degradation Representation and Conditioning in AiOIR

Following the SDIR→MDIR→AiOIR evolution, the central bottleneck is no longer only whether a unified network can support multiple restoration tasks, but how it represents the degradation state (“what is degraded”) and how that representation is injected to steer restoration (“how and where to restore”) when multiple degradations co-exist in a single image [77, 85, 18, 21, 31]. This distinction is important because image restoration is already ill-posed for a single corruption, and the ambiguity increases substantially

under mixed degradations unless the degradation state is parameterized, inferred, or semantically grounded.

A compact mixed-degradation view. Let $\mathbf{z}$ denote the degradation state (or its proxy) and $\mathbf{c}$ the conditioning signal derived from it (e.g., labels, latent embeddings, prompts, CLIP/VLM tokens, or statistics). In real scenes, degradations are frequently spatially non-uniform, so the restoration problem is better viewed as a spatially varying mixture of local degradation operators rather than as a single global corruption state [21, 22, 31]. This implies two coupled requirements for AiOIR: (i) the degradation representation should support composite and non-exclusive corruption states, and (ii) the conditioning mechanism should preserve spatial faithfulness rather than applying a single global correction to the entire image.

For a concise comparison across representation and conditioning choices, we use five desirable properties: mixed-degradation expressiveness, spatial specificity, interpretability/identifiability, controllability, and practicality/open-set robustness. The literature shows that no single family simultaneously maximizes all five properties; instead, each family makes a different trade-off [21, 31, 22, 26].

2.3.2.1 Degradation representation

Single-label and discrete task-ID representations. A common starting point is to represent degradation using a discrete task identity (e.g., a one-hot code or class label), which is compact, easy to supervise, and effective when the corruption class is known or can be reliably classified [85, 77]. However, this representation is structurally mismatched to mixed degradations because it assumes mutual exclusivity. In practice, single-label classifiers typically use a softmax-based objective, which allocates probability mass competitively across classes; this encourages the dominant degradation cue to suppress weaker but still important degradations (e.g., mild blur under stronger haze or low-light), yielding mis-conditioning in composite settings, and is generally inconsistent with non-exclusive multi-label targets [89]. As a result, discrete labels are efficient for closed-set task routing but weak for representing non-exclusive mixtures unless the label space is expanded combinatorially, which quickly becomes impractical [77, 18].

Continuous physical and model-based representations. Physical parameterizations encode degradations through interpretable variables such as blur kernels, atmospheric scattering parameters, or noise levels. Their main advantage is interpretability and direct severity control, and they connect naturally to restoration priors and classical estimation frameworks [90, 7, 6]. Recent universal restoration systems also revisit physically meaningful descriptors to improve control under composite corruptions [21]. The limitation is coverage: real-world degradations often deviate from idealized analytic models, and multiple degradations can be confounded (for example, low contrast caused by haze can resemble blur-induced smoothing in some regions). Thus, physical representations are valuable when the degradation family is well characterized, but they are often incomplete for open-world mixed degradations.

Learned latent degradation embeddings. A dominant modern direction is to infer a latent degradation embedding from the degraded image and use it to guide restoration [77, 18, 80]. This avoids manual parameter design and can capture composite cues implicitly. It is also flexible enough to be integrated with CNN, Transformer, or hybrid backbones. The main limitation is identifiability: latent embeddings can be entangled, meaning that different latent codes may produce similar restorations, while similar degradations can map to different embeddings under domain shift. This weakens interpretability and can reduce stable compositional control unless additional structure (e.g., ingredient decomposition, semantic alignment, or calibration) is imposed [18, 80, 24].

Statistics-based and no-reference quality descriptors. Another representation family uses measurable statistics or no-reference image quality descriptors (e.g., BRISQUE, NIQE, IL-NIQE, or BIQA-style cues) as degradation proxies [91, 92, 93, 22]. These descriptors are attractive because they require little or no annotation and can provide a degradation “handle” even when the exact corruption type is unknown. They are also useful for estimating severity trends (e.g., blur-sensitive edge/texture statistics). However, they are typically non-specific for mixed degradations: several corruptions may produce similar statistics, and many descriptors are global or patch-level summaries rather than faithful spatial degradation maps. Consequently, they are useful auxiliary cues but are rarely sufficient as the sole representation for spatially varying composite restoration.

Semantic prompts, CLIP/VLM cues, and open-set representation. Prompt-based and CLIP/VLM-guided methods provide a strong semantic prior for blind and open-set restoration because they can represent degradations in a richer token/embedding space than closed-set labels [19, 94, 95, 30, 31, 22]. This is a major strength in real-world AiOIR, where degradations are diverse and not fully captured by a fixed label set. However, semantic representations introduce a critical separation problem: the embedding may encode scene semantics (content) more strongly than degradation semantics, especially under heavy corruption or domain shift. In other words, CLIP/VLM features are highly informative for blind restoration, but restoration-oriented use requires mechanisms that separate degradation-related cues from scene-content semantics (or adapt semantic encoders to corrupted inputs) to reduce content leakage and stabilize control [31, 21, 22]. This separation issue becomes more pronounced when restoration decisions must be attributed to multiple co-existing degradations rather than a single dominant cue.

Multi-label degradation states for non-exclusive mixtures. For mixed degradations, a more suitable representation is a multi-label state that allows multiple active factors:

$$\mathbf{z} \in [0, 1]^K, \quad z_k = \sigma(\mathbf{w}_k^\top \mathbf{h}), \quad k = 1, \dots, K, \quad [89] \quad (2.14)$$

where each dimension can represent the presence or severity of a degradation component. Compared with single-label softmax formulations, this removes forced mutual exclusivity and better reflects real composite degradations [18, 96, 26, 21]. Nevertheless, multi-label representation alone is not sufficient: degradation correlations can still cause confusion, calibration is nontrivial, and a global multi-label vector does not by itself indicate where each degradation occurs. Therefore, representation design must be coupled with an appropriate conditioning mechanism.

2.3.2.2 Conditioning and guidance mechanisms

Once a degradation representation (or proxy) is available, the next question is how to inject it into the restorer so that restoration is controllable, stable, and spatially faithful.

Existing AiOIR methods use several conditioning families, each with different trade-offs in expressiveness, compute, and robustness [21, 31, 19, 24, 26].

Early fusion and global feature modulation. Early fusion concatenates condition signals at the input or shallow feature stages. It is computationally simple and compatible with labels, continuous descriptors, or multi-label vectors, but the influence of the condition may dissipate through depth, especially in deep unified restorers handling multiple degradations [77, 21]. Feature-wise modulation (e.g., FiLM/AdaIN-style conditioning) improves controllability by explicitly scaling and shifting intermediate features according to the condition [40, 97, 24, 80]. This is often effective for task steering, but global modulation remains limited for spatially varying mixtures because the same condition is broadcast across the image. As a result, global modulation can restore the dominant degradation reasonably well while under-correcting localized minor degradations.

Spatially aware conditioning. To address local mixtures, conditioning must become spatially adaptive rather than purely global. A generic spatial modulation form is:

$$\mathbf{F}'_\ell(\mathbf{p}) = \gamma_\ell(\mathbf{c}, \mathbf{p}) \odot \mathbf{F}_\ell(\mathbf{p}) + \beta_\ell(\mathbf{c}, \mathbf{p}), \quad [40, 41] \quad (2.15)$$

which is closely related to spatially adaptive normalization mechanisms such as SPADE-like designs [41]. This formulation aligns better with spatially varying mixed-degradation behavior because different locations can receive different restoration behaviors. The main limitation is reliability: many AiOIR systems do not have explicit supervision for spatial degradation states, so the predicted local conditioning maps may be only weakly grounded and can hallucinate local corrections under severe or unseen degradations [21].

Token-based guidance (prompts, cross-attention, and CLIP/VLM tokens). Transformer-based and prompt-based restorers often use token-level conditioning, where degradation cues are injected as prompts or semantic tokens and fused through attention mechanisms [64, 19, 31, 21]. This family is powerful because it enables selective conditioning: different spatial positions can attend to different condition tokens, which is useful for composite and spatially varying degradations. It is also a natural interface for CLIP/VLM embeddings and prompt pools [19, 96, 31]. The limitation is robustness

and cost: token conditioning increases memory/compute and remains sensitive to noisy or semantically entangled tokens, especially when the semantic encoder is not adapted to degraded inputs.

Routing and expert selection. Routing or mixture-of-experts (MoE) conditioning allocates capacity conditionally based on estimated degradation characteristics [98, 99, 26]. This is attractive for mixed degradations because different experts can specialize in different corruption regimes or granularity levels. However, routing depends critically on gate calibration. Under domain shift or ambiguous mixed degradations, incorrect routing can produce brittle artifacts and unstable behavior. Moreover, routing decisions are often based on global or coarse cues, which can limit fine-grained spatial specificity in complex scenes [26].

Frequency-aware guidance. Frequency-aware guidance exploits the fact that degradations often have distinctive spectral footprints (e.g., blur attenuates high frequencies, noise alters high-frequency energy, and compression introduces structured artifacts) [100, 24]. This makes frequency-guided restoration a strong complement to semantic and spatial cues, particularly for detail preservation and blur-sensitive restoration. The limitation is that many implementations use global spectral summaries or coarse LF/HF conditioning, which may not fully capture spatially varying composite degradations. Therefore, frequency-aware guidance is most effective when coupled with representations and conditioning paths that also preserve local spatial specificity.

2.3.2.3 Degradation Representation and Conditioning in implications for mixed-degradation AiOIR design

The literature indicates that degradation representation and conditioning should be designed jointly rather than treated as independent components. Single-label representations are simple and stable in closed-set settings, but they are fundamentally limited for mixed degradations because softmax competition emphasizes dominant degradations and suppresses minor ones. Multi-label representations better match non-exclusive composites, yet they require calibrated inference and spatial grounding to be effective in practice [18, 96, 26]. Physical and statistics-based descriptors improve interpretability or practi-

cality, but they are often insufficient for open-world mixed degradations when used alone [90, 91, 22]. CLIP/VLM-guided and prompt-based representations are especially promising for blind restoration and open-set generalization, but they need explicit mechanisms to separate degradation semantics from normal image content semantics and to stabilize conditioning under corruption [30, 31, 21, 22].

On the conditioning side, global fusion/modulation mechanisms are efficient but weak for localized mixtures, while token-based, spatial, and routing-based mechanisms improve expressiveness at the cost of robustness or computational complexity [40, 41, 64, 98, 99]. Frequency-aware guidance is an important complementary cue, but it should be integrated with spatially grounded conditioning for real mixed degradations [100, 24]. These observations support a mixed-degradation AiOIR design in which compositional degradation semantics, spatially grounded conditioning, and frequency-aware restoration are coordinated within a single framework.

2.4 Related Work

Real-world image restoration is naturally formulated as recovering a latent clean image $x \in \mathbb{R}^{H \times W \times 3}$ from a degraded observation y , where the degradation process may be composite, non-exclusive, and spatially varying. As already discussed in Section 2.3.2, mixed degradations are difficult not only because multiple operators may co-exist, but also because their local interactions create ambiguity in both representation and restoration control. This motivates the following related-work discussion around three connected questions: what restoration backbones are used, how degradation states are represented, and how those states are injected into the restoration network.

Early progress in restoration largely came from strong task-specific or single-degradation backbones, including multi-stage CNN pipelines and transformer restorers, which demonstrated that expressive feature hierarchies can invert complex operators when θ is fixed within a task. Representative baselines include MPRNet, SwinIR, Restormer, NAFNet, and DGUNet [36, 11, 13, 14, 101]. However, these methods generally optimize a task-conditioned mapping $f_\phi(y)$ under a narrow training distribution; under composite degradation, their failures often arise because the model lacks explicit control of **which** degradation is present and **where** it is acting, so the learned estimator tends to collapse toward

a compromise solution. This gap motivated universal/all-in-one restoration (AiOIR), which tries to cover multiple degradations with a single model by introducing a degradation representation z and a guidance mechanism that injects z into the network.

A first line of AiOIR focuses on learning a degradation descriptor z (embedding/prior) that summarizes θ and then conditions a shared restorer. AirNet learns an explicit degradation representation to steer restoration [77]; OKNet advances a kernel/operator perspective for restoration [102]; and SRUDC illustrates how specialized physical structure (e.g., scattering) can be modeled with dedicated branches [103]. In parallel, adverse-weather universal restoration (e.g., haze/rain/snow/raindrops) is developed with unified encoders and decoders, as in TransWeather [85].

Prompt-driven AiOIR replaces explicit physical parameters with learned prompts or tokens as an implicit degradation representation. PromptIR uses learnable prompts to guide a unified restorer [19], and ProRes introduces degradation-aware visual prompts that can be combined for controllability [104]. Sequential and prompt-learning strategies (e.g., MiOIR) explore ordered restoration under multiple degradations [105]. These approaches improve **guidance interaction mechanisms** compared to naive concatenation by letting prompts modulate features through attention. Yet two limitations remain central for mixed degradations. First, prompts are often **weakly interpretable** they improve performance, but the mapping from a prompt dimension to a human-meaningful degradation factor is unclear and Second, prompts are typically broadcast or injected at a few layers, so **spatial specificity** is limited unless prompts are converted into spatial maps. In mixed images, this can look like prompt-induced “global biasing”—helpful when the whole image is rainy, harmful when only some regions are.

A complementary direction tackles heterogeneity by allocating computation conditionally rather than relying on a single shared transformation for all inputs. MoCE-IR implements this idea through mixture-of-complexity experts, where a learned router selects experts of different capacity so that difficult degradations activate heavier computation while easier cases bypass unnecessary experts [42]. The authors report that this routing improves restoration performance across diverse degradations (including composite settings) while reducing redundant computation via conditional execution. Nevertheless, the routing signal is learned implicitly and is not a human-readable multi-label

degradation state, and the routing decisions are not calibrated as a dense spatial mixture field $m_k(u)$; thus, interpretability and explicit region-wise verification of “which degradation is handled where” remain limited. Related AiOIR improvements include privilege/self-improvement training signals and feature perturbation for robustness across tasks [106, 46].

More explicit region-aware reasoning has recently been pursued via latent-prior formulations that infer degradation cues directly from the corrupted input. DAIR-LPE proposes a structured “which–where–what” pipeline: it estimates latent priors using a VAE-like encoder, injects these priors into the restoration encoder to form degradation-aware features, and predicts compact degradation maps that guide localized reconstruction in the decoder [107]. The authors report improved all-in-one restoration quality and robustness, particularly because the explicit map pathway strengthens “where/how” guidance compared to purely global prompts or single latent descriptors. However, the inferred degradation state remains primarily encoded in latent variables and internal maps rather than an explicit human-readable non-exclusive multi-label “what” representation, so semantic interpretability is only partial. In addition, to avoid unsupported claims, we now justify computational cost using the metrics reported in [107] (e.g., published Params/FLOPs and memory/speed analysis) and we describe its added pipeline complexity (latent prior pretraining followed by restoration training), instead of using vague wording.

Rather than modeling degradation semantics explicitly, EvoIR improves all-in-one restoration by strengthening frequency-aware restoration and by adapting training objectives over time. It introduces a Frequency-Modulated Module that separates low/high-frequency processing and applies modulation to better preserve structures and textures, and it couples this with an evolutionary optimization strategy that adjusts loss weights during training to mitigate objective conflicts [47]. This design is particularly relevant for composite cases involving blur, where recovering high-frequency detail and balancing fidelity versus perceptual objectives are both challenging. Despite these gains, EvoIR’s degradation reasoning remains implicit: the method improves restoration behavior through frequency modulation and training-time weight adaptation, but it does not expose an explicit semantic multi-label degradation state or a calibrated dense spatial

representation of region-wise mixed degradation [47].

Diffusion-based universal restoration addresses restoration ambiguity more directly by modeling a generative reverse process instead of a single deterministic inverse. Patch-based diffusion for adverse weather restoration demonstrates that localized priors and iterative refinement can improve generalization [108]. WaveDM integrates wavelet structure into diffusion, aiming to better preserve textures and high-frequency details [100]. AutoDIR combines automatic degradation identification (via BIQA/VLM cues) with latent diffusion restoration and optional text control [22]. DiffUIR introduces a selective/shared-distribution mapping strategy within conditional diffusion for universal restoration [23], and mixed-degradation diffusion variants (e.g., joint-conditional designs) explore stronger conditioning of the reverse process [109]. Very recent directions attempt to unlock diffusion priors for all-in-one settings at scale [110]. Despite strong realism, diffusion methods face practical and methodological constraints. First, multi-step sampling cost complicates deployment. Second, conditioning is often global or prompt-based, so spatially heterogeneous mixtures can still be under-specified. Third, without an explicit interpretable degradation decomposition, it is harder to attribute gains (or failures) to specific corruption factors.

Vision-language-guided restoration pushes **semantic awareness** into AiOIR by using language or CLIP-like embeddings as a high-level degradation descriptor [30]. Controlling vision-language models for universal restoration (often implemented with CLIP-aligned guidance) demonstrates that semantic descriptors can steer restoration without hard task IDs [32, 31]. InstructIR further explores instruction-driven restoration as a generic interface [82], and text-aware diffusion restoration extends this coupling to diffusion conditioning [43]. Large-scale restoration in the wild (e.g., SUPIR, InstantIR) uses powerful generative priors and reference-like guidance for photo-realism beyond synthetic benchmarks [111, 112]. However, semantic guidance introduces a new failure mode: VLM embeddings can entangle **content semantics** with **degradation cues**, so the guidance may shift colors, textures, or structure in ways not attributable to a clean inverse operator. In mixed degradation, this ambiguity is amplified unless the system can (a) separate “what is degraded” from “what is depicted” and (b) localize guidance spatially to match the region-wise layout of degradation across the image. Thus, semantic awareness alone

is insufficient; it must be paired with a spatially granular interaction mechanism and an interpretable representation of the composite degradation state.

Composite-degradation-centric frameworks combine unified modeling with stronger conditioning, but they still reveal a coverage-and-representation gap when degradations are diverse and spatially heterogeneous. OneRestore is an important reference point because it explicitly targets composite restoration using scene/degradation descriptors and strong descriptor-guided interaction within a unified transformer, and it reports strong results on the CDD-11 single and composite benchmarks [21]. At the same time, the descriptor set is closed and the conditioning is fundamentally global, so weak factors and region-wise mixtures are not represented as a calibrated spatial field $m_k(u)$ and the reasoning is only partially verifiable at pixel level. Moreover, OneRestore’s CDD-11 benchmark formulation is centered on adverse-weather/illumination primitives (low-light, haze, rain, snow) and their composites; thus, blur-centric composites are outside that benchmark scope [21]. This distinction matters for our setting because blur is a non-invertible low-pass degradation that removes high-frequency structure, so mixtures such as blur+rain or blur+haze require stronger frequency/structure recovery than weather-only compositions.

Legend: ▲: explicitly designed and validated; ▲: partial/implicit support; ▲: absent or very low.

Desirable Properties for Table 2.3: **G (Spatial specificity)** a method can represent spatial mixtures (e.g., different degradations in different regions) rather than only a global descriptor; **S (Semantic awareness)** check degradation cues are language/CLIP/instruction grounded (human-aligned “what to fix”); **C (Composite robustness)** credits explicit mixed/composite training/evaluation; **I (Interaction)** distinguishes shallow conditioning (weak influence) from deep control (modulation, cross-attn, routing, diffusion conditioning); **T (Interpretability)** check if guidance signal is readable as an explicit multi-label degradation state (and ideally supports “what+where”). For clarity, the table scores are assigned using the above rubric definitions: in particular, **G** evaluates whether a method supports a dense spatially aligned degradation representation (not only descriptor-guided interaction), while **T** evaluates whether the degradation guidance/state

Table 2.3: Comparative analysis on Related Work

Paper	G	S	C	I	T	Summary of limitations
MPRNet(2021) [36]	▲	▲	▲	▲	▲	No explicit degradation state, weak under mixtures
NAFNet(2022) [14]	▲	▲	▲	▲	▲	Strong backbone, but global compromise on mixed scenes
Restormer(2022) [13]	▲	▲	▲	▲	▲	No controllable/inspectable z ; spatial mix under-specified
Fourmer(2023) [113]	▲	▲	▲	▲	▲	Frequency prior helps, but lacks explicit mixture semantics
AirNet(2022) [77]	▲	▲	▲	▲	▲	z mostly global; limited “where” for spatial mixtures
TransWeather(2022) [85]	▲	▲	▲	▲	▲	Weather scope; no explicit multi-label representation
PromptIR(2023) [19]	▲	▲	▲	▲	▲	Prompts typically global; partial interpretability of mixtures
WeatherDiff(2023) [108]	▲	▲	▲	▲	▲	Iterative cost + opaque control across steps
AutoDIR(2024) [22]	▲	▲	▲	▲	▲	z often global; diffusion cost/opacity remains
DA-CLIP(2024) [31]	▲	▲	▲	▲	▲	Semantic “what” improves, but “where” is ambiguous
InstructIR(2024) [82]	▲	▲	▲	▲	▲	Instructions global; spatial heterogeneity under-constrained
OneRestore(2024) [21]	▲	▲	▲	▲	▲	Primitive coverage gap (e.g., blur-centric mixtures)
DAIR-LPE(2025) [107]	▲	▲	▲	▲	▲	Explicit latent “which–where–what” localization, but degradation semantics remain latent (not a readable multi-label state)
EvoIR(2025) [47]	▲	▲	▲	▲	▲	Strong frequency-aware modulation and adaptive optimization, but no explicit semantic/spatial degradation state
MoCE-IR(2025) [42]	▲	▲	▲	▲	▲	Routing not semantic; spatially dense routing often limited

is readable or inspectable from the method outputs/controls.

Table 2.3 presents our synthesis of the cited literature under the G/S/C/I/T rubric, based on each paper’s reported design choices and validation evidence, and exposes a consistent structural trade-off across restoration families. Strong backbones (MPRNet/-NAFNet/Restormer/Fourmer) can score well on distortion metrics when trained broadly, yet they remain low in **T** because they do not expose any explicit degradation representation; consequently, mixed scenes are handled by a global compromise rather than localized corrections. Prompt and factorization methods (AirNet, PromptIR, ProRes, WGWS-Net) improve **I** and sometimes **C**, but their control variables are often global tokens or latent codes, so **G** remains limited when degradations are spatially heterogeneous. Diffusion-based lines (WeatherDiff, WaveDM, AutoDIR) increase **I** and robustness

by modeling multi-modality, but this comes with iterative cost and weaker interpretability of “which degradation was fixed where.” Vision-language control (DA-CLIP, InstructIR) substantially improves $\mathbf{S}$, yet without spatial grounding it still cannot fully represent region-wise mixtures, so semantics alone does not solve mixed degradation restoration.

Taken together, these recent composite-aware AiOIR frameworks make clear progress, but they also expose four recurring gaps that remain incompletely addressed in combination [21, 42, 107, 47]. First, non-exclusive degradation representation is still limited: OneRestore relies on a closed descriptor set, MoCE-IR and EvoIR adapt implicitly via routing or frequency behavior, and DAIR-LPE encodes degradation cues in latent variables; consequently, none of them provides a readable multi-label degradation state that can be directly verified under arbitrary mixtures [21, 42, 107, 47]. Second, spatial specificity of conditioning is uneven: DAIR-LPE advances localization through explicit degradation maps, whereas OneRestore remains largely global at the descriptor level, MoCE-IR routes computation without producing a calibrated dense spatial decomposition, and EvoIR’s spatial effects are mainly frequency-driven rather than mixture-calibrated [21, 42, 107, 47].

Interpretability of degradation reasoning remains partial: descriptors in OneRestore and maps in DAIR-LPE provide some interpretability signals, but the control variables in MoCE-IR and EvoIR are not aligned to named degradations and therefore offer limited degradation-level explainability [21, 42, 107, 47]. Fourth, multi-objective conflict under composite degradation is addressed only selectively: EvoIR explicitly adapts objective weighting during training, while OneRestore and DAIR-LPE improve composite performance primarily through architectural conditioning; however, these gains are not coupled with an explicit readable non-exclusive representation and fully verifiable region-wise reasoning across mixtures [21, 107, 47]. These open gaps motivate designs that jointly support readable non-exclusive mixture representation, spatially specific conditioning, interpretable degradation reasoning, and stable learning under competing objectives in composite restoration.

Chapter 3

Methodology

This chapter presents the methodology used to build and evaluate the proposed restoration system. It first introduces the Design Science Research Process (DSRP) adopted in the study, then describes the dataset protocol, pre-processing pipeline, and the implementation details of the proposed model. Mathematical expressions are included only where they define implemented operators, conditioning functions, and optimization objectives.

3.1 Research Methodology

This study adopts the **Design Science Research Process (DSRP)** as the guiding research methodology [114]. In this thesis, DSRP is used to organize the study from problem definition to artifact construction, demonstration, evaluation, and communication. Figure 3.1 summarizes the six DSRP stages followed in this work.

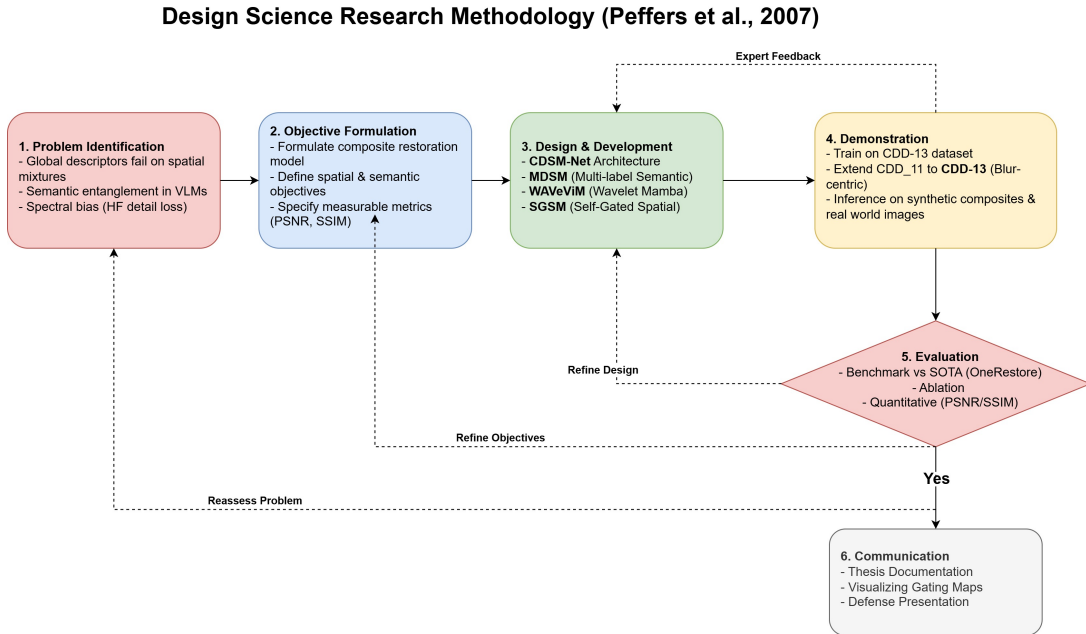


Figure 3.1: Research Design Workflow[1].

Problem Identification and Motivation: In the DSRP problem-identification stage, the study focuses on the limitations of existing all-in-one restoration methods under mixed degradations, particularly when multiple corruptions co-exist and vary spatially within the same image.

Objective Formulation: In the objective-formulation stage, the study defines the design goal as developing a restoration artifact that can represent non-exclusive degradations, guide restoration spatially, and preserve both structural and detail information under composite corruption.

Design and Development: In the design-and-development stage, the proposed artifact is constructed as CDSM-Net. The detailed architecture, conditioning strategy, and learning objective are presented later in Section 3.5.

Demonstration: In the demonstration stage, the implemented artifact is trained and applied on the selected benchmark protocols to verify that the full pipeline operates under controlled mixed-degradation settings.

Evaluation: In the evaluation stage, the artifact is assessed using quantitative metrics, ablation studies, robustness analysis, and real-world no-reference evaluation.

Communication: In the communication stage, the complete methodology, implementation details, experimental results, and interpretations are documented in this thesis.

3.2 Data Acquisition

A central hypothesis of this thesis is that All-in-One Image Restoration (AiOIR) models trained on narrowly defined benchmarks can struggle to generalize to real-world **composite** degradations, particularly when motion blur co-occurs with weather effects. Existing benchmarks provide a strong foundation for weather-based restoration, but they offer limited **blur-aware composite coverage** within the same paired protocol [21]. In this thesis, camera-shake blur is treated as a motivating *blur model* for synthetic stress-testing, while the rolling-shutter component is used only as a lightweight dataset-synthesis

heuristic to diversify blur structure; it is **not** claimed as an independently benchmarked restoration target. The purpose of adding blur-centric synthesis is therefore not to prove standalone rolling-shutter recovery, but to test whether the proposed frequency-aware restoration pipeline remains robust when motion blur suppresses high-frequency detail, a difficulty well established in the image deblurring literature [115, 5, 39]. To address this gap, we construct the **Composite Degradation Dataset-13 (CDD-13)**, a physics-based extension of the standard benchmark setting [21] that expands blur-centric composites and supports controlled robustness stress-testing.

3.2.1 Datasets and category definitions

CDD-11 (baseline). CDD-11 provides clean images and 11 degraded categories spanning four primary degradations (haze, rain, low-light, snow) and their composites [21]. We denote its category set by $\mathcal{C}_{11}$.

CDD-13 (paired experimental dataset used in this thesis). The primary dataset in this thesis is **CDD-13**, defined as

$$\mathcal{C}_{13} = \mathcal{C}_{11} \cup \{\text{blur}, \text{blur_haze}\}. \quad (3.1)$$

CDD-13 preserves the paired clean/degraded supervision protocol while extending $\mathcal{C}_{11}$ with two blur-centric categories, $\{\text{blur}, \text{blur_haze}\}$. In CDD-11, there are **4 single** and **7 composite** categories (**5 double** + **2 triple**), and we follow the published split of $(n_{\text{tr}}, n_{\text{val}}, n_{\text{te}}) = (1183, 0, 200)$ per category (no validation), totaling **13,013** training pairs and **2,200** testing pairs. For CDD-13, each category contains **1,183** paired samples (total **15,379** pairs across 13 categories; **5 single** and **8 composite**, where the additional composite is **blur_haze**), and we use a standard 80/20 split with no validation (**946** train / **237** test per category), resulting in **12,298** training pairs and **3,081** testing pairs. This thesis treats CDD-13 as an **experimental paired dataset protocol** used for model development and evaluation; it is not presented as a community-standard benchmark.

This thesis uses both **CDD-11** and **CDD-13** in complementary roles under a paired evaluation setting. **CDD-11** is used as the primary external benchmark to evaluate restoration performance under the standard composite-degradation protocol and to support fair comparison with prior methods reported on the same categories. **CDD-13** is

used to assess composite robustness under blur-aware mixed degradations and to provide an additional stress-test setting beyond the original CDD-11 category space. This separation allows the thesis to report standardized benchmark performance while also examining robustness to blur-centric composites within the same paired evaluation framework.

3.2.2 Dataset synthesis / generation (CDD-11 $\rightarrow$ CDD-13)

3.2.2.1 Synthesis principle and pairing

Given a clean image $\mathbf{y} \in [0, 1]^{H \times W \times 3}$, each synthesized degraded image is generated by composing a sequence of degradation operators with sampled severity parameters. One-to-one pairing is preserved by exporting synthesized images with the same filename stem and split membership as their clean counterparts, ensuring aligned supervised pairs $(\mathbf{x}, \mathbf{y})$.

(1) Blur category (blur)

The **blur** category is generated to represent motion-blur-dominant degradations that are under-represented in the original CDD-11 taxonomy. The primary blur process is modeled as convolution with a point spread function (PSF), following standard image deblurring formulations for camera-shake blur [115, 5, 39]:

$$\mathbf{x}_b(u) = \sum_{v \in \Omega} \mathbf{y}(v) K(v - u; k_b), \quad [115, 5] \quad (3.2)$$

where $K(\cdot; k_b)$ is a trajectory-based PSF sampled per image. To introduce additional non-uniform blur variation beyond a single spatially uniform PSF, we use the following **dataset-synthesis heuristic**, in which several blurred variants are blended across vertical bands to create a rolling-shutter-like effect:

$$\mathbf{x}_{rs}(u) = \sum_{j=1}^B w_j(u) (\mathbf{y} * K_j)(u), \quad \sum_j w_j(u) = 1, \quad (3.3)$$

with B bands and a small jitter controlling band transitions. This band-wise blending is used only to diversify the synthetic blur structure within the **blur** category; it is not treated as a standalone calibrated rolling-shutter benchmark. With a small probability, local object motion blur is added by blending a soft mask region with an alternative PSF-blurred version.

(2) Composite blur+haze category (blur_haze)

For `blur_haze`, the blur image $\mathbf{x}_{rs}$ is further degraded using the atmospheric scattering model:

$$\mathbf{x}(u) = \mathbf{x}_{rs}(u) t(u) + \mathbf{A} (1 - t(u)), \quad [116, 33] \quad (3.4)$$

where $\mathbf{A}$ is airlight and $t(u)$ is transmission [116]. Transmission is generated using depth guidance:

$$t_{\text{raw}}(u) = \exp\left(-\beta(u) d_{\text{far}}(u)\right), \quad [116, 33] \quad (3.5)$$

where d_{far} is a far-depth proxy derived from a monocular depth estimator (MiDaS) [117]. If depth is unavailable, a dark-channel prior estimate is used as fallback [90]. To increase realism, we define an implementation-specific spatially varying scattering coefficient:

$$\beta(u) = \beta_0(1 + \lambda n(u)), \quad n(u) \in [-1, 1], \quad (3.6)$$

where $n(u)$ is low-frequency noise and λ controls variation. An edge-aware guided filter is then applied to refine transmission and reduce halo artifacts.

Severity ranges. For `blur`, we sample PSF length $L \in \{9, 13, 17\}$ pixels and direction $\theta \sim \mathcal{U}(0^\circ, 180^\circ)$ to cover mild, medium, and strong motion blur. Thus, the **main controlled blur severity** in this thesis is determined by PSF extent and direction diversity, not by a separately calibrated rolling-shutter parameter. For the rolling-shutter-like blurring in Eq. (3.3), we use a small number of vertical bands and mild transition jitter so that this component remains a secondary source of non-uniformity rather than the dominant degradation. For `blur_haze`, we apply haze on top of blur and sample $\beta_0 \sim \mathcal{U}(1.0, 2.0)$ and airlight $\mathbf{A} \sim \mathcal{U}(0.6, 0.9)$ (with the spatially varying $\beta(u)$ defined in Eq. (3.6)).

The CDD-13 protocol provides three properties required by this thesis: paired supervision, blur-aware category coverage, and controlled split consistency for quantitative comparison. These properties are sufficient for the main training and evaluation pipeline of **CDSM-Net**.

3.3 Benchmark Datasets and Evaluation Protocol

Our evaluation is designed to answer two complementary questions. First, can the model restore known composite degradations faithfully when pixel-aligned ground truth is available. Second, does the same model generalize to uncontrolled real scenes where mixed degradations arise naturally and no reference image is used during evaluation. To cover both, we report paired full-reference results on synthetic composite benchmarks (CDD family) and no-reference results on real-world datasets (LOLv2, NH-HAZE, Snow100K-Real).

3.3.1 Paired Synthetic Composite Benchmarks (Primary)

3.3.1.1 CDD-11 benchmark protocol

We evaluate composite restoration using the **CDD-11 benchmark protocol** introduced with OneRestore [21], which is built by synthesizing degradations from clean images sourced from the RAISE dataset [118]. CDD-11 contains 11 categories consisting of four single degradations and their canonical mixtures: {low, haze, rain, snow, low_haze, low_rain, low_snow, haze_rain, haze_snow, low_haze_rain, low_haze_snow}. We use CDD-11 as an **external benchmark** to compare fairly against prior composite-restoration work under the same protocol.

3.3.1.2 CDD-13 (ours): CDD-11 + two blur-centric categories

To study blur-centric composites in the same CDD setting, we *extended* the original CDD-11 category space by adding 2 blur-related categories in an internal extension that we refer to as **CDD-13**. We use only the original CDD-11 categories plus **two** additional categories from the extension: {blur, blur_haze}. This yields **13** total categories, and we denote the resulting experimental set as **CDD-13**. CDD-13 is the primary training set used in this thesis and serves as an internal blur-aware extension for model development and auxiliary stress analysis, while the main cross-method quantitative benchmark is kept on **CDD-11** [21] for fair comparison with prior work.

3.3.2 Real-World Benchmarks (No-Reference Setting, NIQE-only)

Real mixed degradations rarely provide pixel-aligned ground truth. Therefore, for real-world testing we deliberately adopt a **no-reference** evaluation protocol and report **NIQE** [119] on restored outputs. We evaluate on three real datasets: **LOLv2** (real low-light images) [120], **NH-HAZE** [121], and **Snow100K-Real** [122]. Although NH-HAZE is released as paired (hazy / haze-free), the haze in NH-HAZE is *not digitally added*; it is produced on-location using a professional haze generator to create *non-homogeneous* haze [121]. In this thesis we use **only the hazy images** and report NIQE to stress-test generalization under realistic haze that is spatially varying and not synthetic. Snow100K-Real provides real snowy scenes without ground truth [122]. In practice, these real scenes often include additional uncontrolled artifacts (e.g., low-light exposure or atmospheric effects), making it a useful stress-test for mixed real degradations under NIQE-only evaluation. The final real-world analysis in this thesis is restricted to NIQE-based evaluation on LOLv2, NH-HAZE, and Snow100K-Real..

Dataset	Image nature	GT used	Metrics	Use in Chapter 4
CDD-13 (ours)	Synthetic composite (13 cats)	Yes	PSNR/SSIM	Primary training set; paired quantitative evaluation and blur-centric robustness analysis.
CDD-11 [21]	Synthetic composite (11 cats)	Yes	PSNR/SSIM	External benchmark for fair composite comparison.
LOLv2 [120]	Real low-light images	No (GT not used)	NIQE	Real-world generalization (NIQE-only).
NH-HAZE [121]	Real non-homogeneous haze	No (GT not used)	NIQE	Real haze generalization using hazy-only split (NIQE-only).
Snow100K-Real [122]	Real snowy scenes (no GT)	No	NIQE	Real snowy / mixed real-artifact stress test (NIQE-only).

Table 3.1: Datasets and metrics used for training and evaluation.

3.3.3 Evaluation Metrics and Reporting Protocol

Evaluation region and normalization. All full-reference metrics are computed on $[0, 1]$ -normalized images (outputs clipped to this range) after mapping each restored output back to the corresponding ground-truth resolution. When padding/letterboxing is used, PSNR/SSIM are computed only on the valid scene region Ω (mask bounding-box crop), ensuring that padded borders never influence the reported scores.

Full-reference metrics (paired). For paired benchmarks (CDD-11 and CDD-13), we report PSNR and SSIM. Let $x \in [0, 1]^{3 \times H \times W}$ be the ground truth and $\hat{x} \in [0, 1]^{3 \times H \times W}$ be the restored output. Let Ω denote the evaluated valid region (the full image if no padding is present, otherwise the mask bounding-box crop). We compute mean squared error over Ω and convert it to PSNR using the standard full-reference definition from image processing [2]. PSNR measures pixel-wise reconstruction fidelity in decibels; higher PSNR indicates lower reconstruction error. SSIM complements PSNR by measuring perceived structural similarity (luminance, contrast, and structural correlation) between x and $\hat{x}$; higher SSIM indicates better preservation of image structure. We compute SSIM using the standard local-window formulation of Wang et al. [123] over the same evaluated region Ω , and report dataset means for both PSNR and SSIM.

No-reference metric (real-world). For real-world datasets without ground-truth usage, we report NIQE [119] computed on restored images only. NIQE measures deviation from natural image statistics; lower NIQE indicates higher perceptual naturalness under the model assumptions. We use NIQE consistently for LOLv2, NH-HAZE (hazy-only), and Snow100K-Real. NIQE is computed using the standard PyIQA (IQA-PyTorch) no-reference implementation (default settings) on each restored output image and reported as the dataset mean.

Efficiency metrics. To quantify deployability, we report parameter count, multiply-accumulate operations (MACs) for a 256×256 input, and single-image inference latency on a single GPU.

3.4 Data Preprocessing

This section describes how raw paired images are converted into fixed-size training tensors for CDSM-Net. The preprocessing pipeline is designed to satisfy three requirements of mixed-degradation restoration: (i) **shape standardization** for efficient batched training at fixed resolution, (ii) **geometry preservation** so degradation cues (e.g., streak direction, blur extent, haze structure) are not distorted by anisotropic resizing, and (iii) **valid optimization/evaluation** so padded pixels do not bias losses and image-quality metrics. Figure 3.2 illustrates the full pipeline.

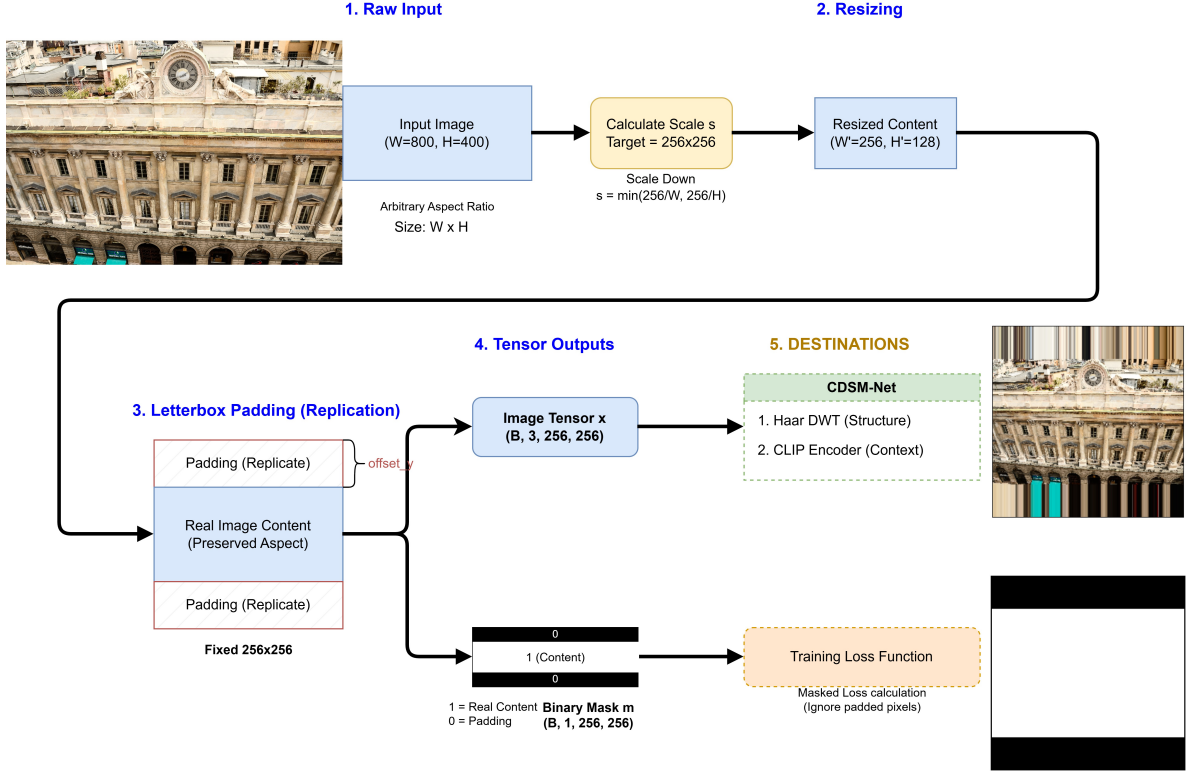


Figure 3.2: Data preprocessing pipeline.

These preprocessing operations were selected as methodological controls rather than as cosmetic image manipulations. Letterbox resizing is used instead of stretched resizing to preserve the original aspect ratio and avoid geometric distortion during resampling, since anisotropic scaling can alter image structure and change the spatial appearance of degradation cues [49, 2]. Replicate padding is preferred over zero padding because padding values can introduce artificial boundary statistics and spatial bias in convolutional feature maps, especially near image borders [124]. A validity mask is therefore computed so losses and full-reference metrics are restricted to real image content, following the masked-processing principle that invalid regions should not drive feature responses or optimization [125]. Finally, the negative sample is used only during training because it supports the auxiliary composite-separation objective, whereas inference requires only the degraded input and should not depend on any additional supervisory signal. The empirical effect of these preprocessing choices is evaluated separately in Chapter 4.

3.4.1 Letterbox resizing with replicate padding

Given a paired degraded image $x \in \mathbb{R}^{3 \times h \times w}$ and clean target $y \in \mathbb{R}^{3 \times h \times w}$, we map both to a fixed training size $(H, W) = (256, 256)$ using letterbox resizing (aspect ratio preserved).

Let

$$s = \min\left(\frac{W}{w}, \frac{H}{h}\right), \quad w' = \text{round}(sw), \quad h' = \text{round}(sh). \quad (3.7)$$

We resize to (h', w') (bicubic) and place the result at the center of an (H, W) canvas. The remaining border region is filled using **replicate padding** (edge-value repetition), which prevents the model from learning artifacts caused by constant-color padding.

3.4.2 Validity mask for content-only supervision

Padding creates pixels that do not correspond to real image content. We therefore compute a binary validity mask $m \in \{0, 1\}^{1 \times H \times W}$ indicating the resized content region:

$$m(u, v) = \begin{cases} 1, & (u, v) \text{ falls inside the placed resized region,} \\ 0, & \text{otherwise.} \end{cases} \quad (3.8)$$

During training and evaluation, losses and metrics are computed on the valid region only, i.e., using $m \odot (\cdot)$, which removes any bias from padded pixels.

3.4.3 Negative sample (training only)

For objectives that require a mismatched reference (described in the loss section), we additionally sample a **negative** image y^- from a different scene in the same split. The dataloader therefore returns $(\tilde{x}, \tilde{y}, y^-, m)$ during training; the negative is **not** used during testing.

3.5 Design and Development

The notation in this section follows standard image restoration and composite-degradation conventions used in AiOIR literature [21]. Let $y \in \mathbb{R}^{H \times W \times 3}$ denote the latent clean image and $x^{(0)} \in \mathbb{R}^{H \times W \times 3}$ the observed degraded image. We use the standard composite-

operator notation

$$x^{(0)} = \left(\mathcal{D}_k \circ \mathcal{D}_{k-1} \circ \dots \circ \mathcal{D}_1 \right)(y) + \varepsilon, \quad (3.9)$$

where $\mathcal{D}_i$ denotes a degradation operator (e.g., blur, haze, rain, low-light effects) and ε is residual noise. Standard formulations (e.g., affine modulation, attention operators, and wavelet-domain transforms) are adopted where appropriate. Unless an explicit citation is attached to an equation, the equations that follow in the MDSM, SGSM, and WAVEViM subsections are implementation definitions of the proposed model in this thesis rather than formulas copied from a single prior source.

The proposed restoration system is formalized in the following subsections.

3.5.1 Proposed Architecture Overview: CDSM-Net

CDSM-Net is organized as a **controller–interface–backbone** pipeline: MDSM predicts a multi-label degradation mixture and emits a compact global code C ; SGSM translates C into spatially selective modulation parameters; and WAVEViM performs wavelet-domain restoration by separately processing low-frequency (LF) structure and high-frequency (HF) details before reconstructing the final image. Figure 3.3 summarizes the complete end-to-end data flow.

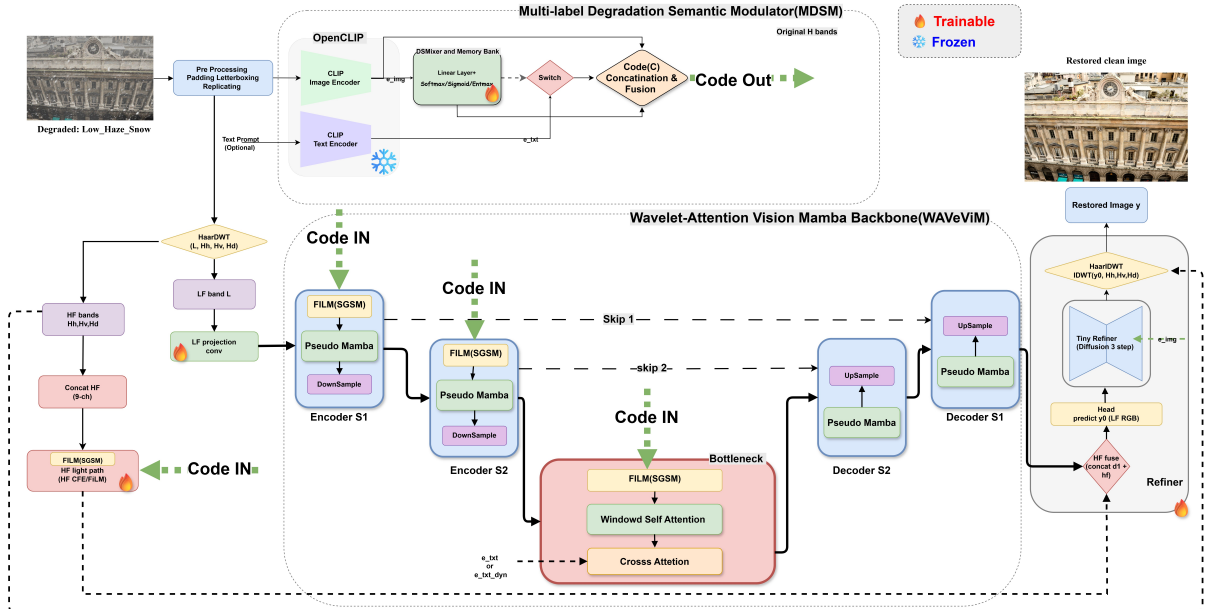


Figure 3.3: Overview of the proposed architecture network: CDSM-Net

At a high level, Figure 3.3 can be read from left to right: we first standardize the

input resolution **without** geometric distortion (avoiding stretch-based resizing), then infer a **multi-label** degradation state, and finally apply a wavelet-conditioned restoration backbone that treats structure and texture with dedicated pathways. The key integration points are (i) the conditioning code C produced by MDSM and (ii) the SGSM operator that injects C into multiple backbone stages to guide restoration under spatially varying composites.

Architectural contribution. CDSM-Net integrates established building blocks (frozen CLIP feature extraction [30], wavelet decomposition, attention-based interactions, and an optional lightweight diffusion-style refiner) within a unified mixed-degradation restoration framework. The contribution of this thesis is the **problem-specific integration and conditioning design**: (i) **MDSM**, which models non-exclusive degradation co-occurrence through multi-label routing and produces a global conditioning code; (ii) **SGSM**, which converts global semantic conditioning into feature-dependent spatial modulation without external masks; and (iii) **WAVEViM integration**, which couples semantic control and spatial modulation with explicit LF/HF restoration pathways for composite degradations.

The contribution of this thesis does not lie in introducing CLIP, Mamba, or diffusion as standalone modules, since these are established components in the literature. Instead, the contribution lies in the interpretable semantic controller (MDSM), the self-gated spatial conditioning interface (SGSM), and their integration with a wavelet-based hybrid restoration backbone for mixed-degradation restoration under practical memory constraints.

The following subsections formalize the three principal components of CDSM-Net and their interaction: the MDSM semantic controller (Section 3.5.2), the SGSM conditioning interface (Section 3.5.3), and the WAVEViM restoration backbone (Section 3.5.4).

3.5.2 Multi-label Degradation Semantic Modulator (MDSM)

The restoration backbone in CDSM-Net requires an explicit control signal that specifies **what degradations are present** in the observed input, including **multiple co-occurring** degradations, while remaining stable and interpretable. Vision-language encoders (e.g., CLIP) provide strong global representations, but their embeddings are pri-

marily optimized for semantic alignment and therefore may entangle scene semantics with degradation cues [30]. The **Multi-label Degradation Semantic Modulator (MDSM)** addresses this by converting frozen VLM features into (i) an **interpretable** multi-label degradation mixture vector α over a predefined degradation vocabulary, (ii) a dynamic degradation text embedding $e_{\text{txt}}^{\text{dyn}}$ synthesized from a learnable degradation prototype memory bank, and (iii) a compact global conditioning code C that summarizes the degradation state for downstream conditioning. The CLIP encoder notation and similarity computations used below follow standard VLM formulations [30]; the proposed component in this stage is the **DSMixer-based routing and memory-mixing mechanism** for composite degradation modeling. In our implementation, the **DSMixer** is the routing/mixing component **inside** MDSM: it computes the mixture weights α and mixes the degradation memory to form $e_{\text{txt}}^{\text{dyn}}$ (see Figure 3.4).

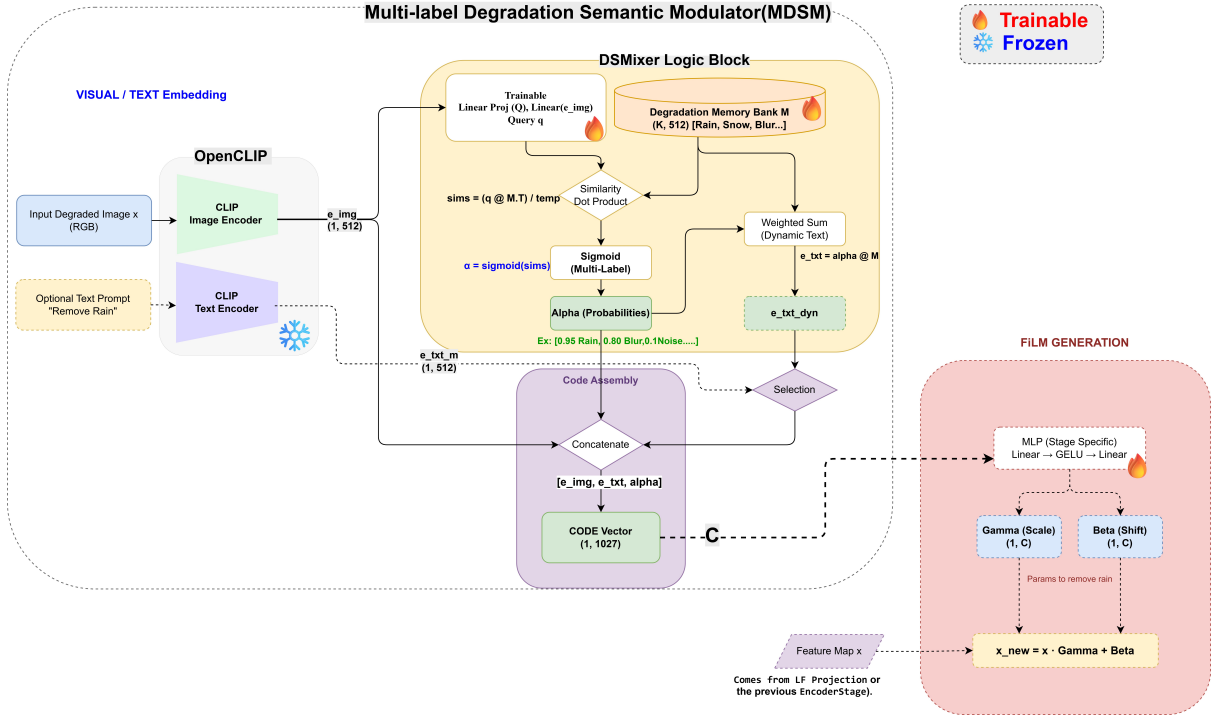


Figure 3.4: MDSM block diagram

Input, VLM image encoding, and optional user text. Let $x \in \mathbb{R}^{B \times 3 \times 256 \times 256}$ be the letterboxed input (Eq. (3.8)). The controller branch follows the VLM/CLIP preprocessing requirements; we resize **only for the controller**:

$$x_{224} = \text{Resize}_{224}(x), \quad \tilde{x} = \frac{x_{224} - \mu}{\sigma}, \quad (3.10)$$

where $\mu, \sigma \in \mathbb{R}^3$ are the VLM normalization constants. A frozen VLM image encoder then returns an ℓ_2 -normalized embedding

$$e_{\text{img}} = \frac{\phi_I^{\text{VL}}(\tilde{x})}{\|\phi_I^{\text{VL}}(\tilde{x})\|_2} \in \mathbb{R}^{B \times D}, \quad (D = 512 \text{ for CLIP ViT-B/32}), \quad (3.11)$$

so in the running batch example $B = 2$ we have $e_{\text{img}} \in \mathbb{R}^{2 \times 512}$. For intuition, one embedding row may begin as

$$e_{\text{img}}^{(1)} = [0.042, -0.158, 0.891, -0.003, 0.112, \dots], \quad \|e_{\text{img}}^{(1)}\|_2 = 1.$$

In addition, MDSM supports an optional user-provided text input (scene description and/or explicit degradation descriptor). Given a text string s (e.g., “night street view with heavy rain and fog” or “degradation: rain, haze, low-light”), a frozen VLM text encoder produces

$$e_{\text{txt}}^{\text{usr}} = \frac{\phi_T^{\text{VL}}(s)}{\|\phi_T^{\text{VL}}(s)\|_2} \in \mathbb{R}^{B \times D}. \quad (3.12)$$

When no user text is available, we treat $e_{\text{txt}}^{\text{usr}}$ as absent and instead construct a dynamic text embedding from the memory bank (Eq. (3.18)).

Degradation memory bank. We maintain a degradation memory bank that stores semantic prototypes for each degradation category:

$$M \in \mathbb{R}^{K \times D}, \quad (3.13)$$

where K is the number of degradations used by the controller (e.g., $K = 13$ for CDD-13), and each row $M_k \in \mathbb{R}^D$ represents a prototype direction in the VLM text space for degradation k . We initialize M using the VLM text encoder applied to canonical prompts such as “a photo with rain degradation” or “a photo with low-light degradation”, providing a semantically meaningful starting point. To adapt prototypes to dataset-specific appearances, we optionally allow M to be trainable; to avoid unstable drift, an exponential moving average (EMA) update can be used:

$$M \leftarrow \rho M + (1 - \rho) M_{\text{learned}}, \quad \rho \in [0, 1). \quad (3.14)$$

DSMixer: computing multi-label mixture weights. DSMixer predicts which degradations are present by comparing the image embedding to the memory prototypes. We first project and normalize the image embedding into the memory space:

$$q = \frac{W_Q e_{\text{img}}}{\|W_Q e_{\text{img}}\|_2} \in \mathbb{R}^{B \times D}, \quad W_Q \in \mathbb{R}^{D \times D}, \quad (3.15)$$

then compute temperature-scaled similarity logits to all prototypes:

$$z = \frac{q M^\top}{\tau} \in \mathbb{R}^{B \times K}, \quad \tau > 0. \quad (3.16)$$

Each logit $z_{b,k}$ is directly associated with a named degradation k ; for instance, with an ordering

$$[k] = [\text{blur}, \text{haze}, \text{rain}, \text{low}, \dots],$$

the mixture vector is interpretable as

$$\alpha_b = [\alpha_{b,\text{blur}}, \alpha_{b,\text{haze}}, \alpha_{b,\text{rain}}, \alpha_{b,\text{low}}, \dots].$$

To allow multiple degradations to be active simultaneously, we use a multi-label activation:

$$\alpha = \sigma(z) \in [0, 1]^{B \times K}, \quad \alpha_{b,k} = \frac{1}{1 + \exp(-z_{b,k})}, \quad (3.17)$$

(optionally replaced by a sparse normalizer such as entmax to encourage selective activations while retaining multi-label behavior). This design avoids the mutual competition imposed by softmax normalization. For example, if a composite input produces high logits for several degradations, e.g.,

$$z_b = [z_{\text{haze}} = 7.8, z_{\text{rain}} = 7.5, z_{\text{low}} = 6.9, z_{\text{blur}} = 1.2, \dots],$$

then sigmoid yields strong simultaneous activations

$$\alpha_b = [\alpha_{\text{haze}} \approx 0.9996, \alpha_{\text{rain}} \approx 0.9994, \alpha_{\text{low}} \approx 0.9990, \alpha_{\text{blur}} \approx 0.77, \dots],$$

whereas a softmax would necessarily redistribute probability mass and reduce at least

one active component, even when the degradations truly co-occur. Consequently, α is directly readable as per-degradation presence/severity evidence (e.g., α_{haze} close to 1 indicates strong haze evidence in the input).

Dynamic text embedding from the degradation cocktail. Given the multi-label mixture α , we synthesize an image-specific degradation embedding by mixing the memory prototypes:

$$e_{\text{txt}}^{\text{dyn}} = \alpha M \in \mathbb{R}^{B \times D}. \quad (3.18)$$

This operation turns the interpretable mixture weights into a continuous, VLM-aligned text vector describing the degradation cocktail. Using the same named ordering, if one sample has

$$\alpha_b = [\alpha_{\text{blur}} = 0.12, \alpha_{\text{haze}} = 0.92, \alpha_{\text{rain}} = 0.81, \alpha_{\text{low}} = 0.64, \dots],$$

then $e_{\text{txt}}^{\text{dyn}}$ is dominated by the haze/rain/low-light prototype directions while still retaining blur influence, producing a single vector that remains compatible with VLM conditioning.

Global conditioning code C . MDSM aggregates the controller outputs into a compact global conditioning code. We first select the text embedding used by the controller:

$$e_{\text{txt}} = \begin{cases} e_{\text{txt}}^{\text{usr}}, & \text{if user text } s \text{ is provided,} \\ e_{\text{txt}}^{\text{dyn}}, & \text{otherwise,} \end{cases} \quad e_{\text{txt}} \in \mathbb{R}^{B \times D}. \quad (3.19)$$

We then form the global conditioning code by concatenating image, text, and multi-label mixture:

$$C = \text{Concat}(e_{\text{img}}, e_{\text{txt}}, \alpha) \in \mathbb{R}^{B \times (2D+K)}. \quad (3.20)$$

For CDD-13, $K = 13$ and $D = 512$, hence $C \in \mathbb{R}^{B \times 1037}$ (and in the running example, $C \in \mathbb{R}^{2 \times 1037}$). This code compactly summarizes **what to restore**: it contains the global semantic context (e_{img}), an explicit degradation descriptor in VLM space (e_{txt}), and the interpretable multi-label degradation state (α). Having produced a global multi-label state C , we still need an operator that injects it spatially and stably into deep feature maps using **SGSM**.

3.5.3 Self-Gated Spatial Modulation (SGSM)

Having derived the multi-label degradation state and global conditioning code C from MDSM (Section 3.5.2), we still require a mechanism that injects this **global** semantics into the **spatially structured** restoration process. This linking step is essential under mixed degradations because the effective corruption is often spatially heterogeneous: some effects are broadly global (e.g., haze/illumination shift) while others are sparse and localized (e.g., rain streaks, raindrops, motion blur bands). If the backbone is conditioned only by a spatially uniform signal, it tends to apply a global compromise, which is precisely the “limited spatial specificity” gap highlighted in our related-work analysis. To address this, we propose **Self-Gated Spatial Modulation (SGSM)**, a specialized spatial FiLM operator that transforms the global code C into **location-dependent** feature modulation **without** requiring external masks, segmentation maps, or coordinate channels.

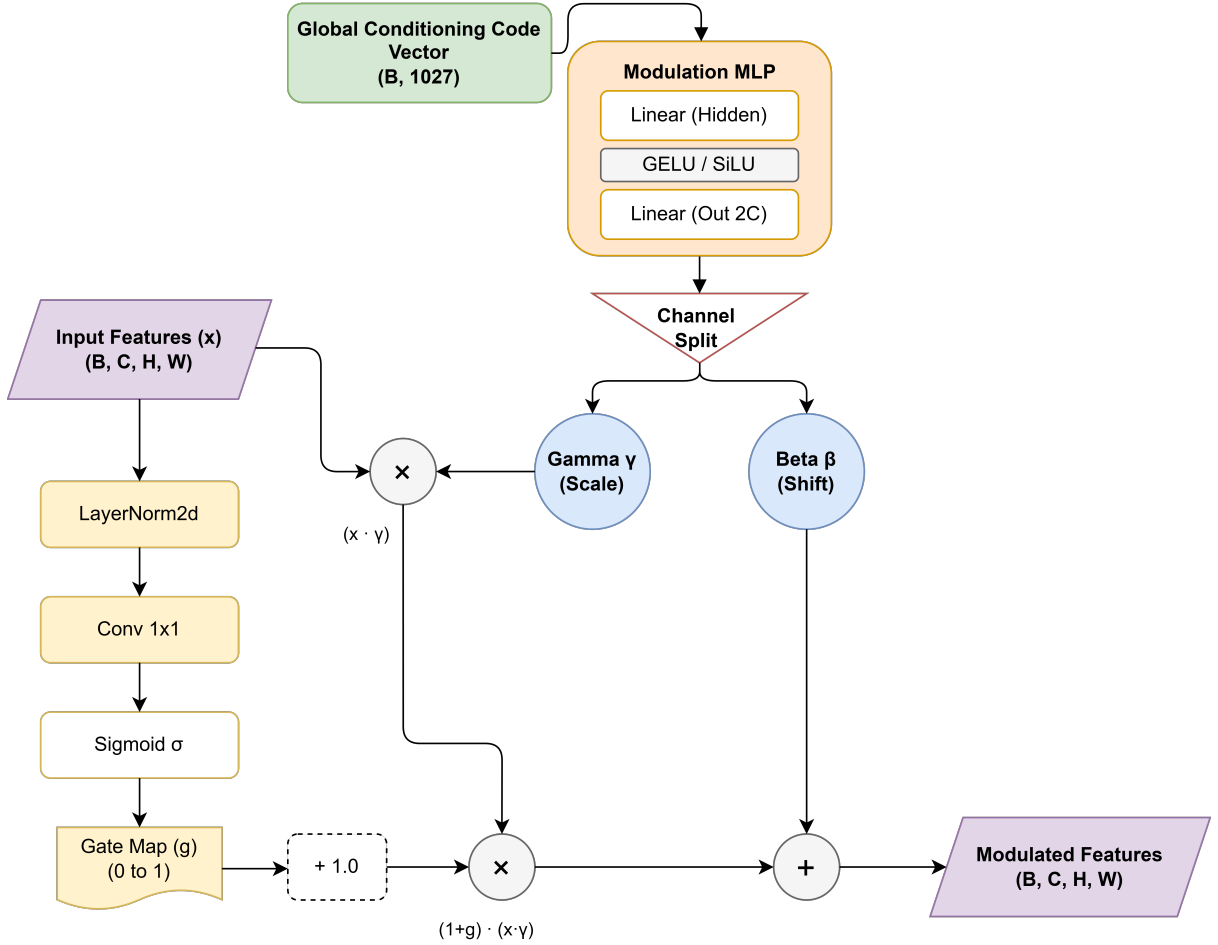


Figure 3.5: Self-Gated Spatial Modulation (SGSM)

Figure 3.5 summarizes the computation: C provides the **intent** through (γ_s, β_s) , while the current features provide the **location** through a self-generated gate G_s . This is favorable in blind restoration because explicit region annotations are typically unavailable, yet the network must still localize corrections when degradations are interleaved within the same image.

From FiLM to spatially selective modulation. Let $F_s \in \mathbb{R}^{B \times C_s \times H_s \times W_s}$ denote the backbone feature tensor at stage s , and let $C \in \mathbb{R}^{B \times d_C}$ be the MDSM global code. We first recall the channel-wise affine conditioning form used in FiLM [40], then define the proposed self-gated spatial extension used in CDSM-Net.

Channel-wise FiLM (spatially uniform). A standard FiLM layer conditions F_s using channel-wise affine parameters generated from C [40]:

$$\gamma_s = \Gamma_s(C) \in \mathbb{R}^{B \times C_s}, \quad \beta_s = \mathcal{B}_s(C) \in \mathbb{R}^{B \times C_s}, \quad (3.21)$$

$$\text{FiLM}(F_s; C) = \gamma_s \odot F_s + \beta_s, \quad \gamma_s, \beta_s \text{ broadcast to } \mathbb{R}^{B \times C_s \times H_s \times W_s}, \quad [40] \quad (3.22)$$

This design is efficient and effective when the degradation is globally consistent, but it cannot represent spatial mixtures because the same correction is applied everywhere.

Spatial FiLM with external spatial priors. A direct extension is to make γ_s, β_s spatial maps (size $H_s \times W_s$), typically requiring a spatial condition signal (e.g., a segmentation/degradation map):

$$\gamma_s = \Gamma_s(\Pi) \in \mathbb{R}^{B \times C_s \times H_s \times W_s}, \quad \beta_s = \mathcal{B}_s(\Pi) \in \mathbb{R}^{B \times C_s \times H_s \times W_s}, \quad (3.23)$$

$$\text{SpatialFiLM}(F_s; \Pi) = \gamma_s \odot F_s + \beta_s, \quad [41] \quad (3.24)$$

where Π is an external spatial prior (mask/segmentation/coordinate-conditioned map). While powerful, this assumption is misaligned with blind mixed-degradation restoration because Π is not available at inference time.

Self-Gated Spatial Modulation (SGSM). SGSM removes the need for external spatial priors by generating the spatial selectivity directly from the feature statistics. Concretely, SGSM keeps the **intent** (what) global via C and derives the **location** (where) from F_s .

(i) **Intent from C : stage-wise channel parameters.** We map C to channel-wise affine parameters for stage s :

$$[\gamma_s, \beta_s] = h_s(C) \in \mathbb{R}^{B \times 2C_s}, \quad \gamma_s, \beta_s \rightarrow \mathbb{R}^{B \times C_s \times 1 \times 1}. \quad (3.25)$$

(ii) **Location from F_s : self-gating map.** We compute a spatial gate from F_s itself:

$$G_s = \sigma\left(\text{Conv}_{1 \times 1}(\text{Norm}(F_s))\right) \in (0, 1)^{B \times C_s \times H_s \times W_s}, \quad (3.26)$$

where $\text{Norm}(\cdot)$ denotes a per-channel normalization and $\sigma(\cdot)$ is the sigmoid. Because G_s is derived from the current features, it adapts to the evolving representation and can highlight regions whose statistics suggest localized corruption.

(iii) **Self-gated modulation.** SGSM then applies a spatially selective affine transformation:

$$\text{SGSM}(F_s; C) = \underbrace{(1 + G_s)}_{\substack{\text{Local Intensity} \\ \text{(Spatial Attention)}}} \odot \underbrace{(\gamma_s \odot F_s)}_{\text{Global Style}} + \beta_s, \quad (3.27)$$

We formalize the complete modulation mechanism by applying the global style parameters weighted by the local spatial attention gate

with broadcasting of γ_s, β_s over (H_s, W_s) . Equation (4.9) makes the gate act as a continuous controller of modulation strength: if $G_s(u) \approx 0$ at location u , then $\text{SGSM}(F_s; C)(u) \approx \gamma_s \odot F_s(u) + \beta_s$ (weak spatial emphasis); if $G_s(u) \approx 1$, the modulation is amplified. For example, for a single channel/location with $F_s(u) = 0.50$, $\gamma_s = 0.80$, $\beta_s = -0.10$, we obtain

$$G_s(u) = 0 \Rightarrow F'_s(u) = 0.80 \cdot 0.50 - 0.10 = 0.30, \quad G_s(u) = 1 \Rightarrow F'_s(u) = 2 \cdot 0.80 \cdot 0.50 - 0.10 = 0.70,$$

illustrating how the same global intent can be applied conservatively or strongly depending on the local gate.

It generates a pixel-wise attention mask (g) directly from its own input features (F), rather than relying solely on the global conditioning code (C). This allows the restoration intensity to vary spatially across the image. The multiplication by $(1 + g)$ is the "Gate." It allows the network to say: "Apply the Rain Removal style strongly here (1.0), but do not touch the background here (0.0). We insert SGSM at multiple points in the WAVEViM backbone (encoder stages, bottleneck, decoder stages, and optionally the HF correction branch). Because C is global while G_s is spatial, SGSM provides a principled interface between MDSM and WAVEViM: the controller communicates **what** degradations are present via C , and the backbone learns **where** to apply this guidance via self-gating. This spatial specificity reduces the need for a global compromise under heterogeneous mixtures and supports stable multi-task interactions by preventing unnecessary corrections in already-clean regions.

3.5.4 Wavelet-Attention Vision Mamba Backbone (WAVEViM)

The **Wavelet-Attention Vision Mamba (WAVEViM)** backbone $\mathcal{B}_\theta$ is the restoration core of CDSM-Net (Figure 3.3). Its design follows a frequency-aware hybrid principle: we explicitly separate **low-frequency (LF)** structure from **high-frequency (HF)** details, then restore LF through a conditioned encoder–bottleneck–decoder while using an auxiliary HF light-path to stabilize texture reconstruction under blur-centric and composite degradations. This separation is crucial because different degradations concentrate their energy in different bands (e.g., haze predominantly affects LF contrast, while blur and rain corrupt HF edges and textures), so entangling them in a single feature stream often yields over-smoothing or artifact amplification.

Wavelet factorization (LF/HF separation). Given the padded degraded input $y \in \mathbb{R}^{B \times 3 \times 256 \times 256}$ (Section 3.5.1), we apply a one-level Haar discrete wavelet transform (DWT) to obtain

$$(L, H_h, H_v, H_d) = \text{DWT}(y), \quad L, H_h, H_v, H_d \in \mathbb{R}^{B \times 3 \times 128 \times 128}, \quad (3.28)$$

where L captures coarse structure and $\{H_h, H_v, H_d\}$ capture directional details. WAVEViM predicts a corrected LF band $\hat{L}$ while using HF features as an auxiliary guidance

signal, and reconstructs the full-resolution output via inverse DWT:

$$\hat{y} = \text{IDWT}(\hat{L}, H_h, H_v, H_d) \in \mathbb{R}^{B \times 3 \times 256 \times 256}. \quad (3.29)$$

LF path: projection + conditioned encoder–bottleneck–decoder. We first lift the LF band into a feature space with a shallow projection:

$$z_0 = \text{Conv}_{3 \times 3}(L) \in \mathbb{R}^{B \times C_0 \times 128 \times 128}. \quad (3.30)$$

Let the global multi-label controller outputs from MDSM be $(e_{\text{img}}, e_{\text{txt}}, \alpha)$ and the global code be C (Section 3.5.2). In the implementation, the backbone forms a conditioning vector by concatenating these quantities,

$$c = [e_{\text{img}}, e_{\text{txt}}, \alpha] \in \mathbb{R}^{B \times d_c}, \quad (3.31)$$

and injects it **deeply** (at multiple stages) using feature modulation. At a given stage ℓ , the modulated feature is

$$\tilde{z}_\ell = \gamma_\ell(c) \odot z_\ell + \beta_\ell(c), \quad (3.32)$$

where $\gamma_\ell(\cdot), \beta_\ell(\cdot)$ are learned affine predictors and $\odot$ denotes channel-wise multiplication (the spatially-gated variant is handled by SGSIM in Section 3.5.3).

The LF encoder is a 3-stage hierarchy:

$$(z_1, s_1) = \text{Enc}_1(\tilde{z}_0), \quad (z_2, s_2) = \text{Enc}_2(\tilde{z}_1), \quad (z_3, s_3) = \text{Enc}_3(\tilde{z}_2), \quad (3.33)$$

where s_1, s_2 are skip features forwarded to the decoder. Each stage stacks a small number of local-mixing blocks and applies down-sampling in early stages to increase receptive field efficiently.

Context modeling at the bottleneck: window attention + cross-attention. At the lowest spatial resolution (bottleneck), WAVEViM can optionally employ **windowed self-attention** to strengthen long-range interactions at reduced cost. For a window $\mathcal{W}$

with tokens $U \in \mathbb{R}^{|\mathcal{W}| \times d}$, the attention update is

$$\text{Attn}(U) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad Q = UW_Q, \quad K = UW_K, \quad V = UW_V, \quad [64] \quad (3.34)$$

applied independently per window; this maintains local detail modeling while avoiding the quadratic cost of global attention at higher resolutions.

In addition, we include an optional **single bottleneck cross-attention** to the semantic text embedding e_{txt} , which provides a direct (yet controlled) semantic injection point:

$$z'_3 = z_3 + \text{Attn}\left(Q = z_3W_Q, \quad K = tW_K, \quad V = tW_V\right), \quad t = e_{\text{txt}} \in \mathbb{R}^{B \times d}, \quad (3.35)$$

so the LF features can align with the degradation semantics without turning the backbone into a fully prompt-driven Transformer.

PseudoMamba: linear-time mixing under restoration constraints. Where full state-space (Mamba) selective-scan operators can be heavy or implementation-sensitive for our mixed-degradation setting, we use a Mamba-inspired (PseudoMamba) token mixer inside encoder/decoder blocks. The goal is to approximate the useful bias of state-space models—**long-range, linear-time mixing**—while preserving stable feature statistics for restoration. Concretely, for a feature map $z \in \mathbb{R}^{B \times C \times h \times w}$, PseudoMamba performs (i) depthwise local mixing with kernel size k , and (ii) gated channel interaction:

$$u = \text{DWConv}_k(z), \quad g = \sigma(\text{Conv}_{1 \times 1}(z)), \quad z^+ = z + \text{Conv}_{1 \times 1}(g \odot u), \quad (3.36)$$

yielding linear complexity in the number of spatial tokens while remaining compatible with multi-scale U-shaped restoration. In practice, this hybrid (PseudoMamba + selective window attention at the bottleneck) balances global structure recovery with fine texture preservation.

Decoder and HF light-path fusion. The decoder mirrors the encoder with two up-sampling stages and skip fusion:

$$d_2 = \text{Dec}_2(z'_3, s_2), \quad d_1 = \text{Dec}_1(d_2, s_1). \quad (3.37)$$

In parallel, we construct an HF guidance tensor by concatenating the three HF subbands and processing them with a lightweight conv stack:

$$h = \text{Conv}(\text{Concat}[H_h, H_v, H_d]) \in \mathbb{R}^{B \times C_h \times 128 \times 128}, \quad (3.38)$$

then apply spatially-aware conditional enhancement (via SGSM-style gating, Section 3.5.3) before fusing with the decoder feature:

$$\tilde{h} = \text{SGSM}(h; c), \quad d_1 \leftarrow \text{Conv}_{1 \times 1}(\text{Concat}[d_1, \tilde{h}]). \quad (3.39)$$

Finally, the backbone predicts a corrected LF band and reconstructs the output:

$$\hat{L} = \text{Head}(d_1) \in \mathbb{R}^{B \times 3 \times 128 \times 128}, \quad \hat{y} = \text{IDWT}(\hat{L}, H_h, H_v, H_d). \quad (3.40)$$

Optionally, a lightweight refiner can operate on the reconstructed LF (or its wavelet LF component) to further reduce residual artifacts while keeping inference cost bounded.

Overall, WAVEViM operationalizes the key premise of Figure 3.3: **frequency factorization** reduces LF/HF entanglement, while **hybrid context modeling** (PseudoMamba + window attention + optional cross-attention) improves global consistency without the cost of full-resolution global Transformers.

3.5.5 Learning Objective and Optimization

We train CDSM-Net in a supervised setting with paired samples. Let $x \in \mathbb{R}^{B \times 3 \times 256 \times 256}$ denote the degraded input after letterbox replicate padding, $y \in \mathbb{R}^{B \times 3 \times 256 \times 256}$ the clean target, and $m \in \{0, 1\}^{B \times 1 \times 256 \times 256}$ the validity mask that excludes padded pixels. Given the controller output $C = \text{MDSM}(x, \text{text})$, the network predicts $\hat{y} = f_\theta(x; C)$. All pixel-domain objectives are computed only on valid pixels by masking with m , ensuring the model does not learn padding artifacts and that reported metrics reflect true content

restoration.

Mixed-degradation restoration is inherently multi-objective: different degradations impose competing corrections (e.g., sharpening for deblurring may amplify noise, while denoising may suppress texture). To stabilize learning under such conflicts, we use a compact set of complementary losses that constrain fidelity in image space, wavelet subbands, and frequency space, and optionally add regularizers for composite separation and perceptual consistency. When a loss is adapted from a standard formulation, its equation is cited explicitly; otherwise, the equation is stated as an implementation definition of the CDSM-Net training objective.

Masked Charbonnier reconstruction loss. We enforce robust photometric fidelity on valid pixels using the Charbonnier penalty [126]:

$$\mathcal{L}_{\text{rec}} = \frac{1}{\|m\|_1} \sum_u m(u) \sqrt{(\hat{y}(u) - y(u))^2 + \varepsilon^2}, \quad [126] \quad (3.41)$$

where u indexes spatial locations (channels summed implicitly), $\varepsilon > 0$ is a small constant, and $\|m\|_1 = \sum_u m(u)$.

Wavelet subband (LF/HF) consistency loss. To explicitly supervise structure and detail under blur-centric composites, we compare Haar-DWT subbands:

$$(\hat{L}, \hat{H}_h, \hat{H}_v, \hat{H}_d) = \text{DWT}(\hat{y}), \quad (L, H_h, H_v, H_d) = \text{DWT}(y), \quad (3.42)$$

and use a downsampled mask $m^{(2)} = \text{AvgPool}_{2 \times 2}(m) \in [0, 1]^{B \times 1 \times 128 \times 128}$. The wavelet loss is

$$\mathcal{L}_{\text{wav}} = \lambda_L \frac{\|m^{(2)} \odot (\hat{L} - L)\|_1}{\|m^{(2)}\|_1} + \lambda_H \sum_{t \in \{h, v, d\}} \frac{\|m^{(2)} \odot (\hat{H}_t - H_t)\|_1}{\|m^{(2)}\|_1}. \quad (3.43)$$

Frequency focal (FFT-domain) loss. We further regularize residual frequency mismatch that often manifests as ringing or over-smoothing. Let $\mathcal{F}(\cdot)$ denote the 2D FFT (per channel) and $\Delta(\omega) = |\mathcal{F}(\hat{y})(\omega) - \mathcal{F}(y)(\omega)|$. We define

$$\mathcal{L}_{\text{freq}} = \mathbb{E}_{\omega} \left[(1 + \tanh(\alpha \Delta(\omega))) \Delta(\omega)^2 \right], \quad (3.44)$$

where $\alpha > 0$ increases emphasis on large spectral errors.

Composite separation loss. To discourage “average” restorations under mixed labels, we use a negative sample y^- and impose a margin:

$$\mathcal{L}_{\text{sep}} = \max\left(0, \delta - \frac{\|m \odot (\hat{y} - y^-)\|_1}{\|m\|_1}\right), \quad (3.45)$$

with margin $\delta > 0$. This encourages $\hat{y}$ to remain discriminative relative to mismatched targets.

Blur-consistency / data-fidelity loss. For blur-containing samples, we add a lightweight measurement-consistency constraint using a blur operator k :

$$\mathcal{L}_{\text{blur}} = \frac{1}{\|m\|_1} \|m \odot ((k * \hat{y}) - x)\|_1, \quad (3.46)$$

and set $\mathcal{L}_{\text{blur}} = 0$ otherwise.

CLIP feature consistency loss. To encourage perceptual faithfulness, we optionally match frozen VLM (CLIP) image features:

$$\mathcal{L}_{\text{clip}} = \|\Phi_I(m_3 \odot \hat{y}) - \Phi_I(m_3 \odot y)\|_2^2, \quad (3.47)$$

where $\Phi_I(\cdot)$ is the frozen image encoder and m_3 is m broadcast to three channels.

Unified objective. The total training loss is the weighted sum of enabled terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{wav}}\mathcal{L}_{\text{wav}} + \lambda_{\text{freq}}\mathcal{L}_{\text{freq}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}} + \lambda_{\text{blur}}\mathcal{L}_{\text{blur}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}}, \quad (3.48)$$

where coefficients are set by configuration; optional terms are disabled by setting their weights to zero.

3.5.5.1 Loss balancing under multi-objective conflicts

Where full state-space selective-scan operators from Mamba-family models [127] can be heavy or implementation-sensitive for our mixed-degradation setting, we use a Mamba-

inspired (PseudoMamba) token mixer inside encoder/decoder blocks. The goal is to retain the useful bias of state-space models—**long-range, linear-time mixing**—while preserving stable feature statistics for restoration under practical memory constraints. Concretely, for a feature map $z \in \mathbb{R}^{B \times C \times h \times w}$, PseudoMamba performs (i) depthwise local mixing with kernel size k , and (ii) gated channel interaction:

$$\mathcal{L}_{\text{bal}} = \sum_j (\exp(-s_j) \mathcal{L}_j + s_j), \quad [45] \quad (3.49)$$

which down-weights high-variance objectives while preventing trivial collapse through the $+s_j$ regularizer. In practice, this stabilizes training when multiple degradations co-occur and different correction cues compete.

We intentionally kept the Frequency Focal Loss as a fixed regularizer ($\lambda = 0.1$) rather than including it in the uncertainty balancing. This prevents the uncertainty mechanism from down-weighting high-frequency penalties, ensuring the model prioritizes structural sharpness and spectral fidelity even when the restoration task is difficult. It acts as a spectral anchor while the other objectives adapt to the degradation mix

3.5.6 Training Process

Training CDSM-Net follows the end-to-end pipeline in Figure 3.3. Each iteration samples paired data $(x^{(0)}, y^{(0)}, \mathcal{S})$ from the dataset, where $x^{(0)}$ is the raw degraded image, $y^{(0)}$ is the clean target, and $\mathcal{S} \subseteq \mathcal{K}$ is the (multi-label) degradation set. To standardize resolution without geometric distortion, we apply letterbox replicate padding **with shared padding parameters** for the input–target pair:

$$(x, y, m) = \text{LetterboxReplicatePair}(x^{(0)}, y^{(0)}; 256), \quad (3.50)$$

where $x, y \in \mathbb{R}^{B \times 3 \times 256 \times 256}$ and $m \in \{0, 1\}^{B \times 1 \times 256 \times 256}$ is a validity mask ensuring that pixel-domain losses and metrics are computed only on valid (unpadded) regions (Section 3.5.5). The controller branch then computes conditioning signals with MDSM: the frozen VLM encoders produce e_{img} and (optionally) a user text embedding $e_{\text{txt}}^{\text{usr}}$, while DSMixer predicts a multi-label mixture α and synthesizes a dynamic degradation text embedding when user text is absent. These components are fused into a global condition-

ing code C , which is constant across spatial positions but is injected into the restoration backbone through SGSIM at multiple stages.

The restoration backbone (WAVEViM) operates in the wavelet domain. We decompose the degraded input via Haar DWT,

$$(L, H_h, H_v, H_d) = \text{DWT}(x), \quad (3.51)$$

restore low-frequency structure using a conditioned encoder–bottleneck–decoder, and use the high-frequency light-path to stabilize edge and texture recovery under composite degradations. The image is reconstructed via inverse DWT, and an optional lightweight refiner may be applied (depending on configuration) to reduce residual artifacts while keeping inference cost bounded. The forward prediction is summarized as

$$\hat{y} = f_{\theta}(x; C). \quad (3.52)$$

We then compute the multi-objective loss described in Section 3.5.5, including masked reconstruction, wavelet subband consistency, frequency focal regularization, and optional terms (composite separation, blur-consistency, CLIP feature loss). When enabled, uncertainty-based balancing mitigates gradient conflicts among competing objectives. Parameters are updated using AdamW with mixed precision (AMP) and gradient scaling; VLM encoders remain frozen, while WAVEViM (and optional trainable memory components) are optimized end-to-end. Checkpointing and EMA (if enabled) are updated per iteration to improve stability.

Algorithm 1 CDSM-Net Training Procedure

```
1: Input: Dataset  $\mathcal{D}$  yielding  $(x^{(0)}, y^{(0)}, y^-, \text{meta})$ , frozen CLIP encoders  $\Phi^{img}$  (and  $\Phi^{txt}$ ), max epochs  $E$ .
2: Output: Trained model parameters  $\theta$  (all learnable weights of CDSM-Net  $f_\theta$ : WAVE-ViM backbone + SGSM + MDSM, including trainable memory/refiner when enabled).
3: Initialize: CDSM-Net  $f_\theta$  (WAVEViM + SGSM + MDSM), degradation memory bank  $\mathcal{M}_{bank}$ , optimizer AdamW, scheduler.
4: Freeze: CLIP encoders  $\Phi^{img}$  (and  $\Phi^{txt}$ ). Set: loss weights  $\lambda_{rec}, \lambda_{wav}, \lambda_{freq}, \lambda_{comp}$ .
5: for  $e = 1$  to  $E$  do
6:   for all  $(x^{(0)}, y^{(0)}, y^-, \text{meta}) \in \mathcal{D}$  do
7:     // 1) Geometry alignment (no stretching)
8:      $(x, y, m) \leftarrow \text{LetterboxReplicatePair}(x^{(0)}, y^{(0)}; 256)$   $\triangleright m$  masks padded pixels
9:     // 2) Semantic guidance (MDSM)
10:     $e_{img} \leftarrow \Phi^{img}(x)$ 
11:     $e_{txt} \leftarrow \mathbf{0}$   $\triangleright$  blind training; user prompt can be used at inference
12:     $\alpha \leftarrow \text{DSMixer}(e_{img}; \mathcal{M}_{bank})$   $\triangleright$  multi-label degradation mixture
13:     $e_{txt}^{dyn} \leftarrow \alpha \cdot \mathcal{M}_{bank}$   $\triangleright$  dynamic degradation text embedding
14:     $C \leftarrow \text{CodeHead}([e_{img}, e_{txt}^{dyn}, \alpha])$   $\triangleright$  global conditioning code
15:    // 3) Forward restoration (SGSM-conditioned WAVEViM)
16:     $\hat{y} \leftarrow f_\theta(x; C)$   $\triangleright \text{DWT} \rightarrow \text{LF/HF paths} \rightarrow \text{IDWT}$ 
17:    // 4) Multi-objective loss (masked where applicable)
18:     $\mathcal{L}_{rec} \leftarrow \text{Charbonnier}(\hat{y} \odot m, y \odot m)$ 
19:     $\mathcal{L}_{wav} \leftarrow \text{WaveletL1}(\text{DWT}(\hat{y}), \text{DWT}(y), \text{Down}(m))$ 
20:     $\mathcal{L}_{freq} \leftarrow \text{FrequencyFocalLoss}(\hat{y}, y)$ 
21:     $\mathcal{L}_{comp} \leftarrow \text{CompositeSeparation}(\hat{y} \odot m, y \odot m, y^- \odot m)$ 
22:     $\mathcal{L}_{total} \leftarrow \lambda_{rec}\mathcal{L}_{rec} + \lambda_{wav}\mathcal{L}_{wav} + \lambda_{freq}\mathcal{L}_{freq} + \lambda_{comp}\mathcal{L}_{comp}$ 
23:    // 5) Optimization (end-to-end update)
24:    Backpropagate  $\mathcal{L}_{total}$  and update  $\theta$  with AdamW; step scheduler; zero gradients.
25:   end for
26: end for
27: return  $\theta$ 
```

Computational feasibility and reliability of the conditioning pipeline. CDSM-Net integrates a vision–language encoder, a hybrid restoration backbone, and an optional lightweight refiner within a single restoration pipeline, but these modules do not incur comparable training cost. In the proposed implementation, CLIP is used as a **frozen semantic feature extractor** that provides conditioning features rather than a trainable image generator. The primary trainable restoration path is the **WAVEViM** backbone conditioned through MDSM and SGSM, while the refiner is **optional**, lightweight, and applied after the main reconstruction stage.

Practical feasibility is supported by fixed-resolution training (256×256), mixed-precision

computation, a small batch configuration, frozen VLM encoders, and checkpoint-based training/resumption. These choices reduce memory consumption and runtime while preserving the conditioning functionality required for mixed-degradation restoration.

Reliability is maintained by supervised constraints on the final restoration output. CLIP contributes conditioning features only and does not synthesize pixels. The optional refiner is also trained under paired supervision rather than unconstrained generation. Consequently, restoration outputs remain bounded by pixel-domain, wavelet-domain, and frequency-domain objectives, which reduces semantic drift and hallucination-like artifacts.

3.5.7 Inference Process

Inference follows the same structural pipeline as training (Figure 3.3) to avoid train–test mismatch, while omitting loss computation and parameter updates. Given a raw degraded image $x^{(0)}$, we apply the identical letterbox replicate padding and validity mask,

$$(x, m, \pi) = \text{LetterboxReplicate}(x^{(0)}; 256), \quad (3.53)$$

then compute the controller code using MDSM. If a user-provided prompt is available (scene description and/or degradation descriptor), we encode it to $e_{\text{txt}}^{\text{usr}}$ and fuse it in the controller; otherwise the dynamic degradation text embedding is synthesized from the memory bank using DSMixer. The resulting global code C is injected into WAVEViM via SGSM, producing $\hat{y}$ through wavelet LF restoration and HF correction, followed by inverse DWT reconstruction and an optional lightweight refiner. Finally, we remove padded regions using the stored padding parameters π (crop-back) to obtain the restored image on the original content area; evaluation metrics are computed on the valid region only.

Algorithm 2 CDSM-Net Inference Procedure

- 1: **Input:** Degraded image $x^{(0)}$, trained CDSM-Net f_θ , frozen CLIP encoders Φ^{img} (and Φ^{txt}), optional user text t_{usr}
 - 2: **Output:** Restored image $\hat{y}^{(0)}$ at the original resolution
 - 3: Apply letterbox replicate padding: $(x, \pi) \leftarrow \text{LetterboxReplicate}(x^{(0)}; 256) \triangleright \pi$ stores crop-back indices
 - 4: Compute CLIP image embedding: $e_{img} \leftarrow \Phi^{img}(x)$
 - 5: **if** t_{usr} is provided **then**
 - 6: Compute CLIP text embedding: $e_{txt} \leftarrow \Phi^{txt}(t_{usr})$ $\triangleright$ user-guided mode
 - 7: **else**
 - 8: Set $e_{txt} \leftarrow \mathbf{0}$ $\triangleright$ blind mode
 - 9: **end if**
 - 10: Compute routing weights: $\alpha \leftarrow \text{DSMixer}(e_{img}; \mathcal{M}_{bank})$
 - 11: **if** $e_{txt} = \mathbf{0}$ **then**
 - 12: Synthesize dynamic degradation text: $e_{txt}^{dyn} \leftarrow \alpha \cdot \mathcal{M}_{bank}$
 - 13: **else**
 - 14: Set $e_{txt}^{dyn} \leftarrow e_{txt}$
 - 15: **end if**
 - 16: Construct global conditioning code: $C \leftarrow \text{CodeHead}([e_{img}, e_{txt}^{dyn}, \alpha])$
 - 17: Predict using SGSM-conditioned WAVEViM: $\hat{y} \leftarrow f_\theta(x; C)$
 - 18: Crop back to original content: $\hat{y}^{(0)} \leftarrow \text{CropBack}(\hat{y}, \pi)$
 - 19: Clamp output: $\hat{y}^{(0)} \leftarrow \text{Clamp}(\hat{y}^{(0)}, 0, 1)$
 - 20: **return** $\hat{y}^{(0)}$
-

Chapter 4

Experimentation and Results

4.1 Experimental Setup

Table 4.1 summarizes the implementation settings used throughout this chapter. The benchmark definitions, valid-region evaluation protocol, and metric formulations are the same as those introduced in Chapter 3. In brief, CDD-11 is used as the main cross-method benchmark, CDD-13 is retained for internal blur-aware stress analysis, and real-world results are reported with NIQE on LOLv2, NH-HAZE, and Snow100K-Real.

Item	Setting
Framework	PyTorch
Hardware	Single GPU per run (Tesla P100 on Kaggle; A100 on Colab Pro for selected runs)
Precision	AMP mixed precision
Input size	256×256 patches
Batch size / Accumulation	2 / 1
Optimizer	AdamW [128] ($\beta_1=0.9$, $\beta_2=0.99$, weight decay = 0)
Learning rate	2×10^{-4}
Schedule	Warmup (1200 iters) + cosine annealing [129]
Epochs / Iterations	100 epochs / 2400 iterations per epoch
Checkpointing	Periodic checkpoint save/resume across runtime-limited sessions
Data loading	Seed = 42; workers = 2; filename-based degraded/clean pairing (threshold 0.86)

Table 4.1: Implementation and optimization settings used in the experiments.

4.2 Training Stability, Convergence, and Optimization

This section analyzes the optimization behavior of CDSM-Net under a multi-objective restoration objective. We first examine convergence using the joint dynamics of the logged total training objective and a paired PSNR trace (Fig. 4.1). We then motivate and analyze our uncertainty-weighted multi-task formulation, which balances heterogeneous objectives through learnable task uncertainties (Figs. 4.2–4.3). This analysis es-

establishes that optimization converges reliably before we report quantitative benchmark comparisons in Sec. 4.3.

4.2.1 Training Stability and Convergence

Figure 4.1 plots the **logged** training objective together with a paired PSNR trace over iterations. Here, the PSNR curve is computed online at logging steps on the same mini-batches used to log the loss, after applying the valid-content mask and clamping to $[0, 1]$. This masking is consistent with our preprocessing pipeline: inputs are padded/letterboxed to match architectural constraints, and the binary mask marks the valid content region so that padded borders do not influence the reported metric.

The faint background traces correspond to raw per-iteration measurements, which exhibit high variance due to stochastic mini-batch sampling, heterogeneous composite degradations, and the interaction of multiple objective terms. To reveal the underlying convergence behavior, we additionally plot a smoothed trend (moving-average), which shows consistent optimization progress: the logged objective decreases while PSNR increases throughout training.

A visible discontinuity occurs at the boundary between two training windows. This step is not a divergence event: it is produced by resuming training from checkpoints across separate Kaggle sessions (runtime-limited executions). At each resume, we restore model parameters together with optimizer states and AMP scaler states, then continue optimization from the same global iteration count. Small transients can still appear at the resume boundary (e.g., due to data-order re-initialization and stochasticity), but the smoothed trends recover immediately and continue improving, indicating stable checkpointed continuation.

What “fixed weights” means. In the baseline setting, the total objective is a fixed weighted sum of the constituent losses, with coefficients set a priori in the configuration:

$$\mathcal{L}_{\text{fixed}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{wav}}\mathcal{L}_{\text{wav}} + \lambda_{\text{comp}}\mathcal{L}_{\text{comp}} + \lambda_{\text{blur}}\mathcal{L}_{\text{blur}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}} (+ \lambda_{\text{freq}}\mathcal{L}_{\text{freq}}), \quad (4.1)$$

where the λ values remain constant throughout training. Although Fig. 4.1 shows overall convergence, the raw traces remain noisy and PSNR saturates earlier (high-20 dB range),

motivating adaptive balancing of heterogeneous objectives.

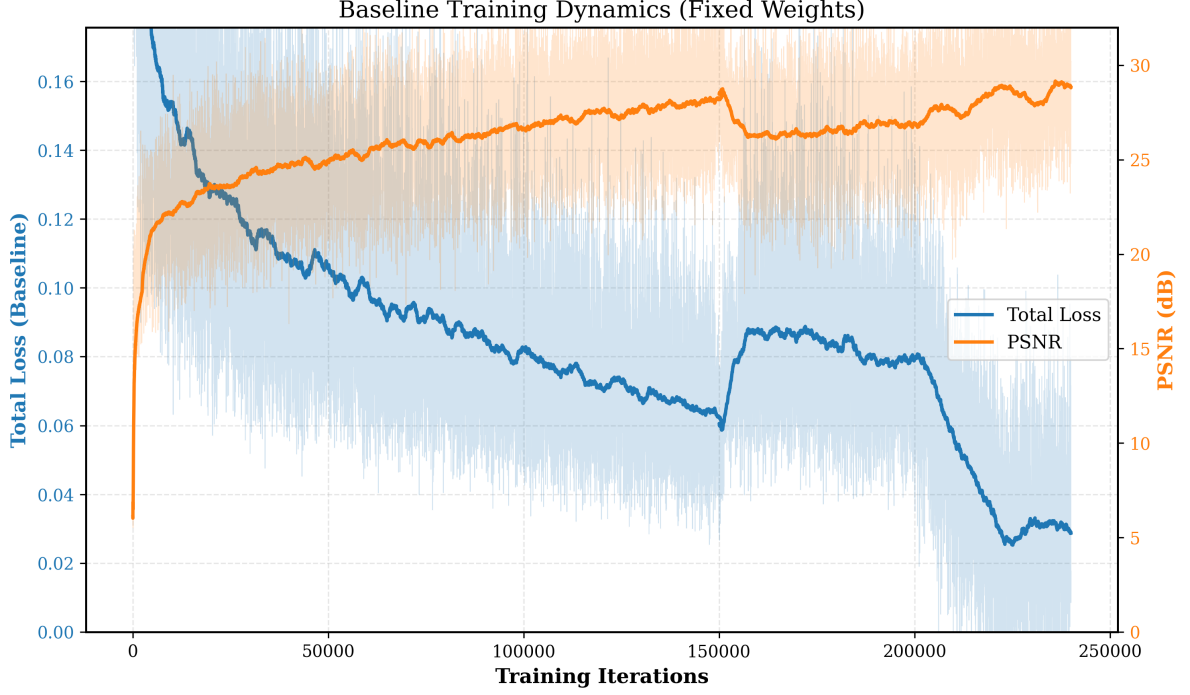


Figure 4.1: Baseline training dynamics (fixed loss weights)

The raw per-iteration traces (faint) are noisy due to stochastic training and heterogeneous composite degradations, while the smoothed trend shows overall convergence with decreasing objective and increasing PSNR. The discontinuity corresponds to checkpointed resumption across multiple Kaggle sessions. Adaptive weighting strategy analyzed next, which explicitly reduces gradient interference among heterogeneous objectives and improves late-stage restoration quality.

4.2.2 Adaptive Loss Balancing via Uncertainty-Weighted Multi-Task Learning

To stabilize optimization under heterogeneous objectives, we adopt uncertainty-weighted multi-task learning [45]. Let $\{\mathcal{L}_i\}_{i=1}^K$ denote the K primary objective terms used in our training (reconstruction fidelity, wavelet-domain supervision, composite separation, blur consistency, and CLIP-feature alignment). Following [45], we optimize a homoscedastic uncertainty-weighted objective using a numerically stable log-variance parameterization:

$$\mathcal{L}_{\text{UW}} = \sum_{i=1}^K (\exp(-s_i) \mathcal{L}_i + s_i), \quad s_i = \log \sigma_i^2, \quad [45] \quad (4.2)$$

The effective weight on task i is

$$w_i = \exp(-s_i) = \frac{1}{\sigma_i^2}. \quad (4.3)$$

Thus, objectives with larger uncertainty (larger σ_i / larger s_i) are automatically down-weighted, while more stable objectives receive relatively larger weights as training progresses. The additive s_i term prevents the trivial solution $\sigma_i \rightarrow \infty$ and yields a well-behaved trade-off.

Between Fig. 4.1 (baseline) and Fig. 4.2 (uncertainty loss), the architecture, data, and loss **definitions** are unchanged; the difference is the **balancing rule**. The baseline uses constant coefficients λ_i (Eq. 4.1), while our method replaces them with learnable s_i that adapt the effective weights w_i during training (Eq. 4.3). (When an additional frequency regularizer is used, it is kept as a fixed-weight term outside the uncertainty-weighted sum in both settings.)

Figure 4.2 shows the resulting uncertainty-weighted objective trajectory alongside PSNR. Compared with the fixed-weight dynamics in Fig. 4.1, uncertainty-weighted training exhibits smoother progress and achieves higher restoration fidelity, with PSNR exceeding 33 dB in late training. We emphasize that the absolute scale of $\mathcal{L}_{\text{UW}}$ is not directly comparable to the baseline loss: due to the added s_i terms, the uncertainty-weighted objective can take negative values while still corresponding to improving restoration quality. For this reason, convergence is primarily interpreted through PSNR/SSIM trends and benchmark results.

The uncertainty-weighted objective evolves smoothly while PSNR improves steadily, surpassing 33 dB in late training. To interpret **how** the model balances objectives, Fig. 4.3 plots the evolution of the learned uncertainty parameters per objective (visualized as $\sigma_i = \exp(\frac{1}{2}s_i)$). A critical reading rule is that the **effective contribution** is proportional to $w_i = 1/\sigma_i^2$ (Eq. 4.3), i.e., the effective weights evolve in the **opposite direction** of the plotted σ_i curves: **large** σ_i means the corresponding loss is **down-weighted**, and decreasing σ_i means that objective is progressively **up-weighted**. Under this interpretation, we observe a curriculum-like behavior: perceptual/semantic alignment and composite-related supervision are strongly down-weighted early (high σ_i), allowing the model to first fit stable reconstruction and wavelet fidelity signals. As training pro-

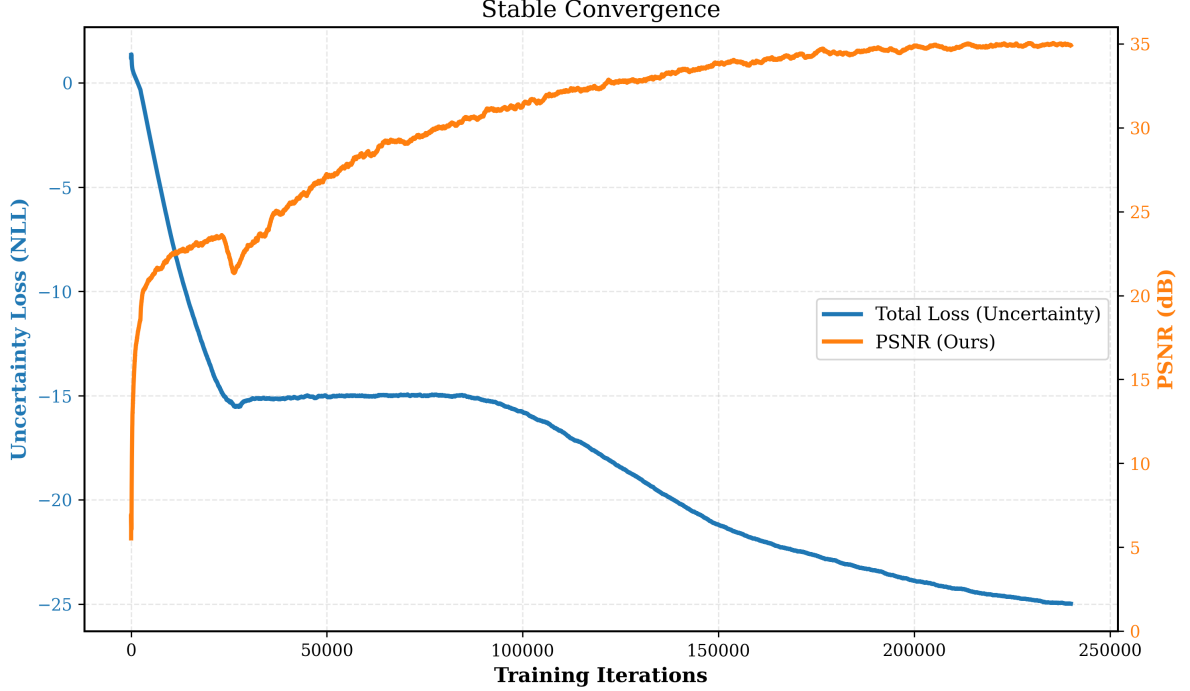


Figure 4.2: Uncertainty-weighted training dynamics

ceeds, their uncertainties decrease, increasing the contribution of auxiliary constraints and improving late-stage restoration quality. Finally, the stabilization of the uncertainty trajectories indicates that optimization reaches a consistent multi-objective equilibrium rather than oscillating between competing terms.

Visualized as σ_i over training iterations for the five primary objective terms. Since the effective weight is $w_i = 1/\sigma_i^2$, larger σ_i implies a smaller contribution of the corresponding objective to the total gradient, and decreasing σ_i implies increased emphasis during later training. Having established stable convergence and interpretable adaptive balancing under the adapted objective, we next report quantitative results across composite benchmarks, single-degradation benchmarks, real-world no-reference evaluation, and efficiency measures (Sec. 4.3).

4.3 Quantitative Results

This section reports quantitative evidence for the effectiveness of our restoration approach under standard evaluation protocols. We use full-reference metrics (PSNR/S-SIM) on paired synthetic benchmarks and a no-reference perceptual metric (NIQE; lower

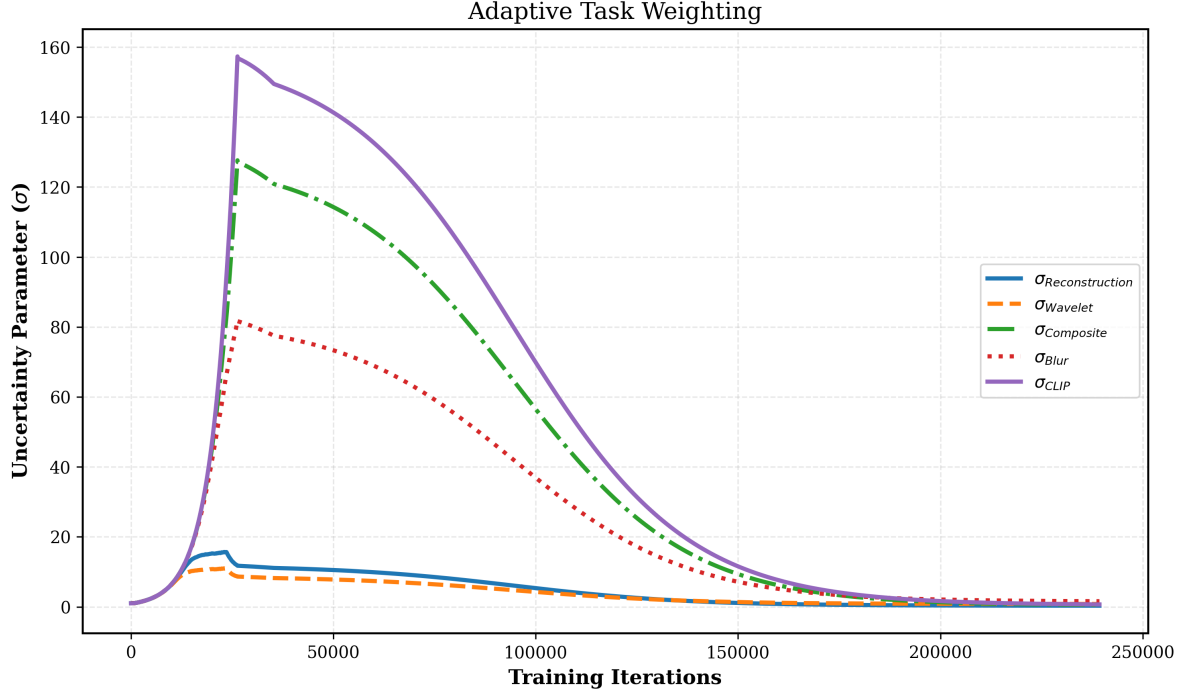


Figure 4.3: Learned task uncertainties

is better) on real-world unpaired benchmarks. While this thesis emphasizes **semantic awareness**, **spatial specificity**, **composite robustness**, and **interpretability**, these properties are not treated here as single universal scalar metrics. Instead, they are assessed operationally through targeted empirical evidence. Semantic awareness is evaluated through the semantic-controller ablations and the multi-label versus single-label routing comparison; spatial specificity is evaluated through the SGSIM ablation; composite robustness is evaluated through category-wise single/composite breakdowns and real-world stress tests; and interpretability is discussed through the structure and usefulness of the intermediate routing signal itself. Accordingly, the results are organized from overall performance to category-wise analysis, real-world evaluation, complexity analysis, and targeted ablations.

4.3.1 Overall Results

We begin with the overall performance on CDD-11, obtained by averaging PSNR/SSIM across the 11 degradation categories. This aggregate view is important because it reflects the expected operating regime of all-in-one restoration systems: a single model must remain competitive across diverse and often composite degradations. Table 4.2 summarizes

the overall CDD-11 results for all compared methods.

Category	Methods	Venue & Year	PSNR $\uparrow$	SSIM $\uparrow$	#Params
Specialized restorers	MIRNet [62]	ECCV'20	25.97	0.8474	31.79M
	MPRNet [36]	CVPR'21	25.47	0.8555	15.74M
	MIRNetv2 [130]	TPAMI'22	25.37	0.8335	5.86M
	Restormer [13]	CVPR'22	26.99	0.8646	26.13M
	DGUNet [131]	CVPR'22	26.92	0.8559	17.33M
	NAFNet [132]	ECCV'22	24.13	0.7964	17.11M
	SRUDC [133]	ICCV'23	27.64	0.8600	6.80M
	Fourmer [134]	ICML'23	23.44	0.7885	0.55M
	OKNet [135]	AAAI'24	26.33	0.8605	4.72M
All-in-One / multi-weather	AirNet [16]	CVPR'22	23.75	0.8140	8.93M
	TransWeather [136]	CVPR'22	23.13	0.7810	21.90M
	WeatherDiff [137]	TPAMI'23	22.49	0.7985	82.96M
	PromptIR [19]	NeurIPS'23	25.90	0.8499	38.45M
	WGWSNet [138]	CVPR'23	26.96	0.8626	25.76M
Composite-aware universal	OneRestore [†] [21]	ECCV'24	28.47	0.8784	5.98M
	OneRestore [21]	ECCV'24	28.72	0.8821	5.98M
	MoCE-IR [42]	CVPR'25	29.05	0.8806	2.80M
	DCPT [139]	ICLR'25	30.66	0.8912	22.0M
	CDSM-Net (Ours)	Ours	29.47	0.9883	13.17M

Table 4.2: CDD-11 overall results (average PSNR/SSIM over 11 categories). Best/second/third are highlighted in red/green/blue.

Based on Table 4.2, two trends are evident. First, approaches that explicitly target composite or broad corruption coverage (e.g., OneRestore [21], MoCE-IR [42], and DCPT [139]) consistently outperform task-specialized restorers (e.g., MIRNet [62], MPRNet [36], and Restormer [13]) in average PSNR. This supports our motivation that mixed-degradation restoration benefits from **degradation-aware control** rather than relying on a single-task inductive bias.

Second, as shown in Table 4.2, our CDSM-Net achieves **29.47 dB** average PSNR and **0.9883** average SSIM on CDD-11, ranking **second** in PSNR among all compared methods. Compared to OneRestore [21] (28.72 dB / 0.8821), CDSM-Net improves the overall average by **+0.74 dB** PSNR. We attribute this gain to our semantics-conditioned control: the multi-label semantic modulation reduces ambiguity about **what** degradations are present, while spatially gated conditioning modulates **where** restoration strength is applied, mitigating over-restoration in heterogeneous mixtures. At the same time, these averages should be interpreted carefully. In particular, the reported SSIM of **0.9883** is a dataset-average structural similarity score on the paired synthetic CDD-11 benchmark, not a claim that the restored outputs are visually indistinguishable from the ground

truth for every sample. This is consistent with our own later analysis: performance is not uniform across all categories, and snow-heavy mixtures remain challenging even when global structure is preserved. Therefore, we interpret the high average SSIM as evidence of strong structural recovery on average rather than perfect reconstruction. The concern of overfitting is also reduced by the broader evaluation protocol in this thesis, which includes training-stability analysis, category-wise decomposition, component ablations, and real-world no-reference testing under domain shift; taken together, these observations are more consistent with genuine learning than with simple memorization.

Although DCPT [139] attains the best overall PSNR, the overall average alone does not indicate which categories dominate the ranking. Therefore, Sections 4.3.2–4.3.3 decompose the results into single-degradation and composite-degradation settings to clarify the trade-offs and to connect performance changes to our intended properties.

4.3.2 Single-Degradation Benchmarks

After establishing the overall CDD-11 averages in Section 4.3.1, we next isolate the **single-degradation** settings (**low-light**, **haze**, **rain**, and **snow**). This diagnostic view verifies that the model is not only competitive on the aggregate, but also exhibits strong **base competence** on each underlying corruption mode before we analyze composite mixtures in Section 4.3.3. Table 4.3 summarizes PSNR/SSIM for representative specialized restorers (e.g., Restormer [13], PromptIR [19]) and universal/composite-aware baselines (OneRestore [21], MoCE-IR [42], DCPT [139]), alongside our CDSM-Net.

Low-light and haze. On **low-light**, CDSM-Net attains the highest PSNR among compared methods (**28.80 dB**) with a large margin over both universal and specialized baselines (Table 4.3). This indicates that the model can infer an appropriate enhancement strength from the input—a prerequisite for our broader goal of **semantic awareness** when degradations vary across images and scenes. On **haze**, DCPT [139] achieves the strongest PSNR, while CDSM-Net remains highly competitive (**34.99 dB**), outperforming OneRestore [21] and MoCE-IR [42]. This suggests that our conditioning strategy generalizes beyond illumination changes to contrast/visibility degradations, supporting robust behavior across distinct physical corruption families.

Method (Venue)	PSNR $\uparrow$				SSIM $\uparrow$			
	Low	Haze	Rain	Snow	Low	Haze	Rain	Snow
TransWeather (CVPR'22) [136]	23.39	23.95	26.69	25.74	0.7253	0.9241	0.8986	0.8903
AirNet (CVPR'22) [16]	24.83	24.21	26.55	26.79	0.7775	0.9513	0.8914	0.9193
Restormer (CVPR'22) [13]	26.29	28.35	33.10	33.43	0.8165	0.9813	0.9576	0.9675
PromptIR (NeurIPS'23) [19]	26.32	26.10	31.56	31.53	0.8054	0.9686	0.9463	0.9595
OneRestore (ECCV'24) [21]	26.55	32.71	33.48	34.50	0.8266	0.9912	0.9641	0.9742
MoCE-IR (CVPR'25) [42]	27.26	32.66	34.31	35.91	0.8240	0.9900	0.9700	0.9800
DCPT (ICLR'25) [139]	27.61	36.71	35.75	37.92	0.8330	0.9950	0.9750	0.9850
CDSM-Net (Ours)	28.80	34.99	36.74	28.54	0.9914	0.9945	0.9973	0.9933

Table 4.3: CDD-11 single-degradation results (Low/Haze/Rain/Snow). Best/second/third are highlighted in red/green/blue among the methods listed.

Rain and snow. For **rain**, CDSM-Net achieves the best PSNR (**36.74 dB**) and the best SSIM (**0.9973**) among compared methods (Table 4.3). This is consistent with our motivation that effective restoration requires identifying **what** degradation is present and modulating **how strongly** to restore, rather than applying a one-size-fits-all correction. In contrast, **snow** remains a challenging regime: DCPT [139] and MoCE-IR [42] achieve substantially higher PSNR, indicating that snow-like occlusions and dense high-frequency streaks can dominate the error profile even when overall averages are strong. We report this behavior explicitly because it clarifies why the best overall method may be driven by a subset of difficult cases, and it motivates a more fine-grained composite analysis.

Overall, Table 4.3 shows that CDSM-Net is consistently strong on low-light, haze, and rain, while exhibiting a clear weakness on snow. These single-degradation outcomes provide a necessary baseline for interpreting composite mixtures: in mixed settings, degradations interact and vary spatially, increasing ambiguity about **which** degradations coexist and **where** they occur. We therefore next analyze the full set of composite categories in Section 4.3.3 to determine whether our semantics-guided and spatially modulated conditioning yields more reliable behavior under heterogeneous mixtures.

4.3.3 All-in-One and Composite Degradation Benchmarks

Single-degradation scores (Section 4.3.2) verify base restoration competence, but they do not fully represent the practical regime of all-in-one restoration, where multiple degradations frequently co-exist and interact. Composite categories are inherently harder because they couple global effects (e.g., illumination/contrast) with localized artifacts (e.g.,

streak-like patterns), and their spatial heterogeneity increases ambiguity about **what** degradations are present and **where** restoration should be applied. Therefore, we next evaluate the **composite** subset of CDD-11 and visualize the full per-category profile to clarify where the model is strong, where it remains limited, and how those trends relate to the proposed semantics-guided and spatially modulated design.

Table 4.4 reports PSNR/SSIM on the seven composite categories of CDD-11. Although DCPT [139] attains the best composite-average PSNR, CDSM-Net wins on several mixture regimes that are central to mixed-degradation restoration: in particular, we achieve the strongest PSNR on **Low+Haze** (28.60 dB), **Low+Rain** (28.25 dB), and **Haze+Rain** (33.90 dB), outperforming OneRestore [21], MoCE-IR [42], and a strong restoration baseline (Restormer [13]). These gains are consistent with our intended design principle: such mixtures require the model to handle both global visibility/illumination correction and localized streak-like artifacts at the same time, and this is exactly where mixture-aware semantic control plus spatially selective modulation should help most. A purely global or mutually exclusive controller can over-apply one restoration mode, whereas our multi-label controller with SGSIM can distribute restoration strength more selectively across heterogeneous regions. By contrast, the weaker performance on snow-dominant mixtures indicates that the remaining gap is not in detecting mixed conditions in general, but in handling occlusion-like snow structures that are harder to recover faithfully with pixel-wise reconstruction objectives.

Method (Venue)	PSNR $\uparrow$ on composite categories							Composite Avg.	
	Low+ Haze	Low+ Rain	Low+ Snow	Haze+ Rain	Haze+ Snow	L+H +Rain	L+H +Snow	PSNR $\uparrow$	SSIM $\uparrow$
TransWeather (CVPR'22) [136]	22.24	22.62	21.80	23.10	22.34	21.55	21.01	23.13	0.7810
AirNet (CVPR'22) [16]	23.23	22.82	23.29	22.21	23.29	21.80	22.24	23.75	0.8140
Restormer (CVPR'22) [13]	24.80	25.28	24.99	26.80	26.15	23.90	23.82	25.11	0.8269
OneRestore (ECCV'24) [21]	26.15	25.83	25.56	30.27	30.46	25.18	25.28	26.96	0.8496
MoCE-IR (CVPR'25) [42]	26.24	26.25	26.04	29.93	30.19	25.41	25.39	27.06	0.8461
DCPT (ICLR'25) [139]	27.15	26.75	26.70	32.63	33.40	26.23	26.40	28.47	0.8593
CDSM-Net (Ours)	28.60	28.25	24.90	33.90	29.68	25.76	23.98	27.87	0.9851

Table 4.4: CDD-11 composite-category results. Composite-average highlighting: best/second/third in red/green/blue.

A consistent difficulty is observed on snow-involved mixtures (e.g., **Low+Snow** and **Low+Haze+Snow**), where DCPT [139] and MoCE-IR [42] provide substantially higher

PSNR (Table 4.4). Scientifically, snow is challenging because it can introduce **opaque** or near-opaque occlusions and dense high-frequency streaks that behave more like structured clutter than an additive noise process. This tends to penalize pixel-wise fidelity metrics (PSNR) strongly: even small residual snow streaks or slight intensity offsets can dominate the MSE, especially under low-light where contrast is already compressed. Notably, our SSIM remains very high on these mixtures, suggesting that the model preserves structural content while being more conservative in removing heavy snow-like occlusions. This trade-off is further supported by the category-wise quantitative results and the real-world no-reference evaluation discussed later in this chapter.

To complement the table and make category-wise trade-offs easier to interpret, Fig. 4.4 visualizes the PSNR/SSIM profile across all 11 CDD-11 categories for a compact set of representative methods. The shape of the curves highlights an important point for that the overall ranking is not driven by a uniform gain across all conditions, but by **which** categories dominate the error budget. In our case, the profile shows clear advantages on low-light and several mixture regimes (e.g., haze+rain), while snow-related categories remain the principal source of gap to the top PSNR method.

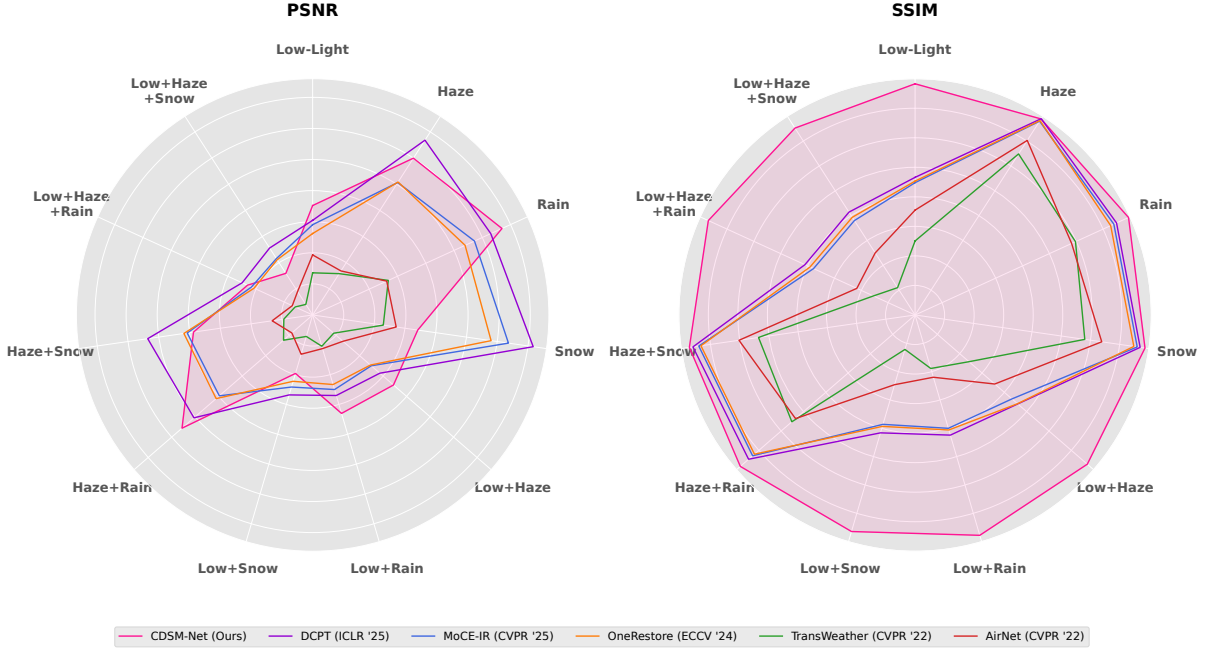


Figure 4.4: CDD-11 per-category performance profile (PSNR/SSIM radar) across 11 degradation categories, comparing representative all-in-one and composite-aware base-lines with CDSM-Net.

Taken together, the composite-only breakdown (Table 4.4) and the category-wise pro-

file (Fig. 4.4) provide a coherent picture of the model’s behavior on the established CDD-11 benchmark. Our model is most reliable on mixtures that require jointly handling illumination/visibility changes with streak-like artifacts (e.g., haze+rain and low+rain), which is consistent with a semantics-guided, mixture-aware control strategy. The primary remaining limitation is snow-heavy regimes, where occlusion-dominant patterns can disproportionately penalize PSNR even when structure is preserved (high SSIM).

4.3.4 Real-World Mixed-Degradation Images (No-Reference)

Synthetic paired benchmarks quantify fidelity, but real deployment is dominated by **domain shift** (unknown sensors, illumination, and weather statistics) and **unlabeled mixtures**. Therefore, we additionally evaluate on three real-world benchmarks using the no-reference NIQE metric (lower is better) [92]: LOL-v2 for low-light enhancement [140, 141], Snow100K-Real for real snow removal [142, 143], and NH-HAZE for non-homogeneous haze [144].

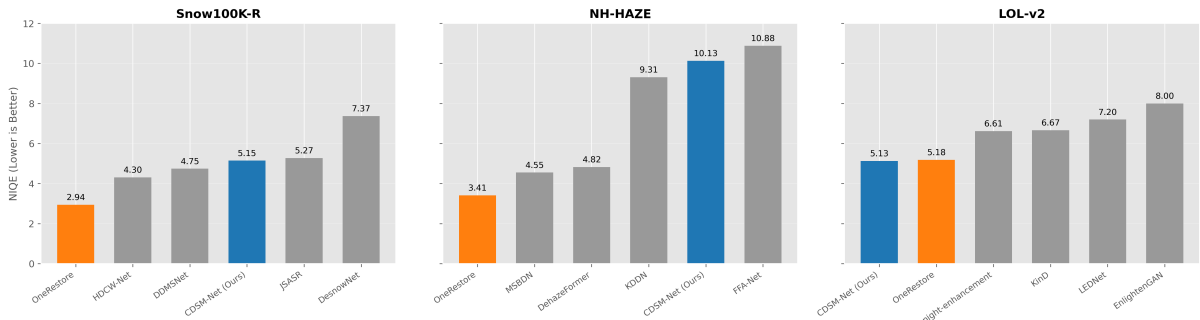


Figure 4.5: No-reference NIQE comparison on three real-world benchmarks (lower is better).

As summarized in Fig. 4.5, **CDSM-Net achieves the best NIQE on LOL-v2** (5.1309), slightly surpassing OneRestore (5.182) and clearly improving over classical low-light baselines such as KinD [145], EnlightenGAN [146], LEDNet [147], and Night Enhancement [148]. This behavior aligns with our design intent: semantic modulation helps disambiguate **what** needs enhancement under difficult illumination, while self-gated spatial modulation supports **where** to amplify visibility without uniformly boosting noise or over-saturating already-correct regions—properties that NIQE tends to reward because they preserve natural image statistics.

On **Snow100K-R**, our performance is **not yet strong** (NIQE 5.1487): although

we outperform earlier snow baselines such as DesnowNet [143] (7.374) and JSASR [149] (5.271), we remain clearly behind more recent, snow-specialized methods (HDCW-Net [142]: 4.305; DDMSNet [150]: 4.746) and OneRestore (2.935). This outcome is consistent with our paired CDD-11 results, where snow and snow-composite categories were among the weakest cases despite remaining competitive against several multi-degradation baselines (Fig. 4.4). A plausible reason is **semantic/appearance ambiguity** in Snow100K-R: many images contain ground snow accumulation and low-light/overcast illumination rather than purely airborne snow streaks, so the visual cues may be routed toward low-light or haze-like prototypes, leading to under-suppression of residual snow textures and less natural statistics (higher NIQE).

When we come to **NH-HAZE**, we observe a **clear failure mode** (NIQE 10.1284), far behind specialized dehazing baselines such as OneRestore (3.411), MSBDN [151], DehazeFormer [152], and KDDN [153]. A plausible explanation is **distribution and physics mismatch**: NH-HAZE contains strongly **non-homogeneous** and spatially varying atmospheric veils that require globally consistent color/airlight correction and careful suppression of haze-induced low-frequency veiling, whereas our training emphasis and architecture (wavelet LF/HF separation + learned semantic control) may over-compensate locally, inadvertently injecting unnatural contrast/texture statistics that NIQE penalizes. This motivates a targeted extension of the haze primitive set and/or stronger physically grounded global haze constraints in future work.

4.3.5 Efficiency and Complexity

Beyond restoration quality, deployment requires competitive **compute-accuracy** trade-offs. We therefore report model complexity and runtime in the same environment used throughout our experiments: a single NVIDIA Tesla P100-PCIE-16GB GPU on Kaggle (CUDA 12.8), single-image inference at 256×256 . We measure latency (ms/image), throughput (FPS), and peak GPU memory (VRAM), and additionally provide approximate MACs/GFLOPs using THOP at 256×256 (noting that profiling tools may undercount unsupported/custom operations such as wavelet transforms and non-standard attention kernels).

CDSM-Net is lightweight and fast in practice: FP32 inference takes **34.57 ms/im-**

age (about **28.93 FPS**) with **0.818 GB** peak VRAM, indicating feasibility for real-time or near-real-time use at fixed resolution. AMP does not improve speed in our implementation (37.10 ms/image) and slightly increases peak VRAM, consistent with mixed operator types where autocast overhead can dominate at batch size one. We emphasize that end-to-end script throughput (including OpenCLIP conditioning, disk I/O, and PNG saving) can be substantially lower than pure model latency; thus, the latency/FPS numbers below best reflect the architectural efficiency of the restoration network.

Setting	#Params (M) ↓	MACs (G) ↓	GFLOPs (G) ↓	Latency (ms) ↓	Peak VRAM (GB) ↓
CDSM-Net (FP32)	13.17	42.98*	85.96*	34.57	0.818
CDSM-Net (AMP)	13.17	42.98*	85.96*	37.10	0.994

Table 4.5: Efficiency of CDSM-Net at 256×256 on a Tesla P100-PCIE-16GB (Kaggle). *MACs/GFLOPs are THOP estimates and may under-count unsupported/custom operations.

To keep the complexity discussion aligned with our methodology, Table 4.6 groups parameters by the **restoration backbone** and the proposed **semantic/spatial control** modules. Here, the restoration backbone comprises the encoder–bottleneck–decoder pipeline (including the bottleneck interaction blocks), operating in the wavelet-separated feature space, and accounts for the majority of capacity. In contrast, our guidance mechanisms remain compact: the Multi-label Degradation Semantic Modulator (MDSM) contributes **1.5438M** parameters, the Self-Gated Spatial Modulation (SGSM; implemented via SpatialFiLM and feature-driven gating) is absorbed into the modulation layers, and the optional tiny refiner adds only **0.3275M**. This allocation matches our design intent: introduce explicit, interpretable control signals without inflating the core reconstruction cost.

Component Group (CDSM-Net)	Trainable Params (M) ↓
Restoration backbone (encoder+bottleneck+decoder)	9.6390
MDSM (semantic controller; DSM module)	1.5438
Tiny refiner (optional)	0.3275
Other (head, wavelet/HF path, SGSM/SpatialFiLM layers, fusion)	1.6587
Total	13.17

Table 4.6: Trainable parameter breakdown of CDSM-Net grouped by our methodology: restoration backbone vs. semantic/spatial control modules.

Finally, we clarify that the bottleneck operates at fixed spatial resolution: after down-sampling reaches a 64×64 latent, the bottleneck processes features at the same 64×64

grid (while increasing channel capacity), which is where global-context reasoning and the strongest semantic interactions are applied. Overall, these measurements support that CDSM-Net remains efficient while providing explicit semantic and spatial controllability. In particular, the parameter breakdown shows that most capacity remains in the restoration backbone, while the semantic/spatial control modules contribute comparatively modest overhead; therefore, the gains reported earlier are better interpreted as coming from better conditioning structure rather than from simply scaling model size.

4.4 Ablation Studies and Component Analysis

This section isolates the contribution of key design choices in our pipeline by ablating one factor at a time while keeping training and evaluation protocols unchanged. The goal is to verify that the observed gains are not incidental, but instead arise from principled components that directly address the challenges of mixed degradations: (i) stable handling of variable-resolution inputs, (ii) reliable degradation awareness under composite corruption, and (iii) avoiding restoration artifacts caused by preprocessing biases. We begin with input processing, since it affects **every** downstream module and can silently dominate both quantitative scores and qualitative behavior.

4.4.1 Impact of input data processing

Real-world and benchmark images do not always match the network input resolution; therefore, a preprocessing step is required to map arbitrary sizes into a fixed 256×256 tensor. While such resizing/padding is often treated as a minor implementation detail, it can strongly influence restoration behavior because padding introduces artificial boundaries and intensity statistics that the model may incorrectly learn as part of the degradation. Figure 4.6 visualizes this effect on a representative *low+haze* example, comparing four input modes: (a) stretched resize, (b) black letterbox padding (no mask), (c) replicate/edge letterbox padding (no mask), and (d) masked replicate padding (our default).

Table 4.7 reports the average PSNR/SSIM on CDD-11 under the same four preprocessing modes. Three consistent observations emerge. First, **black padding without masking is harmful** (27.29 dB / 0.873), and the failure mode is visually clear



Figure 4.6: Effect of input processing on restoration quality.

in Figure 4.6: the network tends to amplify dark padded regions into **black spots** or **intensified shadows**, especially in naturally dark areas of the image. This indicates that the model partially **learns the padding** as a signal correlated with degradation severity, creating spurious restoration behavior near borders and in low-intensity regions. Second, **stretched resizing reduces border artifacts but introduces geometric distortion**, which typically manifests as haze-like residuals and texture inconsistencies (28.46 dB / 0.926). Although stretching avoids explicit padding, it alters aspect ratio and spatial frequency content, which disrupts the restoration backbone’s assumptions about structure/texture alignment. Third, **replicate padding (no mask) is substantially**

Input processing mode	PSNR $\uparrow$	SSIM $\uparrow$
Stretched resize to 256×256	28.4580	0.9258
Black letterbox padding (no mask)	27.2888	0.8732
Replicate letterbox padding (no mask)	28.8458	0.9848
Masked replicate padding (ours)	29.4669	0.9883

Table 4.7: Ablation of input processing on CDD-11 (average over 11 categories)

better than black padding (28.85 dB / 0.985) because it avoids injecting unnatural black boundaries; however, it still yields mild **border bias**, since replicated edges are

synthetic content that the network may overfit to. Our final choice, **masked replicate padding**, achieves a large jump to **29.4669 dB / 0.9883**, and produces the cleanest output in Figure 4.6. The key advantage is that replicate padding preserves realistic boundary statistics, while the mask explicitly tells the loss (and implicitly the network behavior during training) which pixels correspond to valid image content. As a result, the model is discouraged from allocating capacity to explaining padded regions, and instead concentrates restoration capacity on meaningful areas, reducing over-restoration and suppressing padding-induced artifacts. Overall, this ablation confirms that correct input handling is not a cosmetic detail; it is a necessary condition for stable mixed-degradation restoration and for fair evaluation on variable-resolution data.

4.4.2 Semantic Controller Ablation (MDSM & OpenCLIP Guidance)

This section isolates how much of CDSM-Net’s performance comes from the **semantic controller** (OpenCLIP guidance + MDSM routing) rather than backbone capacity alone. We perform **one-factor-at-a-time** ablations: starting from the full model, we disable **exactly one** component and retrain that variant under the same protocol, then evaluate on CDD-11. The only exception is **OpenCLIP Off**, which is evaluated as an **inference-time** stress test (the model becomes unstable if OpenCLIP is removed during training; thus we directly quantify the failure mode at test time).

OpenCLIP guidance (inference-time stress test). Removing OpenCLIP guidance produces unstable restoration behavior, including color inconsistency and artifact amplification. This is consistent with the role of OpenCLIP in our design: it provides a global semantic prior that reduces ambiguity about **what** degradations are present, preventing the controller from collapsing into incorrect restoration behaviors when degradations are composite or visually confusable.

Removing DSMixer (no α , no dynamic text). DSMixer is responsible for forming a **dynamic composite descriptor** by mixing the degradation-text memory through routing weights α (multi-label). When DSMixer is removed, the model loses this mixture representation and must rely on a weaker/less adaptive conditioning signal. As shown in Table 4.8, this causes a large drop from **29.47 dB / 0.9883** to **23.95 dB / 0.8185** on average. The degradation is not limited to a single category; it is especially damaging

in composite folders (e.g., low_haze_snow, low_haze_rain), which supports our claim that **explicit mixture modeling** is essential for mixed-degradation restoration.

Single-label SoftMax routing vs. multi-label routing. To test whether **multi-label** semantics matter beyond having a semantic signal at all, we replace sigmoid multi-label routing with SoftMax (single-label), forcing the controller to choose only one dominant degradation. This creates a systematic mismatch on composite inputs: the restoration is conditioned as if the image contained a single degradation, even when multiple degradations co-occur. Quantitatively, this reduces performance to **27.81 dB / 0.9261**, i.e., **-1.64 dB** PSNR and **-0.0623** SSIM relative to the full model. This confirms that composite restoration benefits from **non-exclusive** degradation semantics rather than mutually exclusive classification.

OpenCLIP Off (inference-time). Finally, disabling OpenCLIP at inference yields a catastrophic collapse (**14.71 dB / 0.6046** on average), reflecting that the model’s controller is trained to utilize semantic guidance and becomes ill-posed when this signal is removed. We therefore treat this ablation as a diagnostic: it verifies that our gains are genuinely driven by semantic conditioning, not by incidental architectural advantages.

Variant	Avg. PSNR $\uparrow$	Avg. SSIM $\uparrow$
CDSM-Net (Full)	29.4669	0.9883
OpenCLIP Off (inference-time)	14.7054	0.6046
No DSMixer (trained ablation)	23.9461	0.8185
SoftMax routing (single-label, trained ablation)	27.8114	0.9261

Table 4.8: CDD-11 average results (mean over 11 categories) for semantic-controller ablations.

Overall, the ablations support two conclusions aligned with our methodology goals: (i) OpenCLIP provides a critical semantic anchor that stabilizes restoration under ambiguous mixtures, and (ii) MDSM must remain **multi-label and mixture-aware** (DSMixer + non-exclusive routing) to preserve performance on composite degradations.

4.4.3 Impact of SGSM (Spatial Selective Modulation)

Mixed degradations in real images are rarely spatially uniform: haze often dominates distant regions while shadows, blur, or low-light remain localized. Consequently, an effective conditioning interface must encode not only **what** degradations exist, but also **where** to

apply restoration strength. In this ablation, we evaluate a controlled evolution of conditioning expressivity **Channel FiLM** $\rightarrow$ **Multiplicative Spatial FiLM** $\rightarrow$ **SGSM (ours)**—while keeping all other components fixed.

(A) Channel FiLM (global modulation). Channel FiLM applies a global affine transform per channel, constant across all spatial locations:

$$\mathbf{y}_{c,h,w} = \gamma_c(C) \odot \mathbf{x}_{c,h,w} + \beta_c(C), \quad [40] \quad (4.4)$$

where $(\gamma, \beta) \in \mathbb{R}^{B \times C}$ are predicted from the global conditioning code C . This design is stable but implicitly assumes spatially homogeneous corruption; it risks over-correcting clean regions (e.g., clear sky) or under-correcting severe local degradations.

(B) Multiplicative Spatial FiLM (no residual). To introduce location awareness without the full SGSM mechanism, we modify the architecture to use the spatial gate purely as a mask:

$$G_s = \sigma(\text{Conv}_{1 \times 1}(\text{LN}_{2D}(F_s))) , \quad (4.5)$$

$$\mathbf{y} = G_s \odot (\gamma_s(C) \odot F_s) + \beta_s(C). \quad (4.6)$$

Because G_s is purely multiplicative ($G_s \in [0, 1]$), this variant acts as a “soft mask.” While it offers spatial selectivity, it lacks an identity path. If the gate predicts low values in clean regions, it creates feature attenuation artifacts, lowering structural fidelity (SSIM).

(C) SGSM (Self-Gated Spatial Modulation, ours). Our SGSM retains the spatial gate but applies it as a **residual amplifier** via $(1 + G_s)$. This ensures that clean regions pass through with their original statistics (via the identity term) while degraded regions receive boosted correction (via G_s):

$$[\gamma_s, \beta_s] = f_s(C), \quad \gamma_s, \beta_s \in \mathbb{R}^{B \times C}, \quad (4.7)$$

$$G_s = \sigma(\text{Conv}_{1 \times 1}(\text{LN}_{2D}(F_s))) , \quad G_s \in [0, 1]^{B \times C \times H \times W}, \quad (4.8)$$

$$\text{SGSM}(F_s; C) = \underbrace{(1 + G_s)}_{\substack{\text{Local Intensity} \\ \text{(Spatial Attention)}}} \odot \underbrace{(\gamma_s(C) \odot F_s)}_{\text{Global Style}} + \underbrace{\beta_s(C)}_{\text{Global Bias}} . \quad (4.9)$$

Notably, if $G_s \rightarrow 0$ (indicating a clean region), SGSM gracefully degrades to standard

channel-wise FiLM, preserving stability.

Table 4.9: SGSM ablation on CDD-11. (Higher is better).

Method	Metric	Low	Haze	Rain	Snow	Low+Haze	Low+Rain	Low+Snow	Haze+Rain	Haze+Snow	L+H+Rain	L+H+Snow	Average
(A) Channel FiLM	PSNR	27.99	33.76	34.99	27.92	28.48	28.07	22.00	32.74	27.98	24.00	22.97	28.26
	SSIM	.9893	.9088	.8999	.8908	.9590	.8999	.9076	.8504	.8934	.9798	.8902	.9154
(B) Multiplicative Spatial FiLM	PSNR	28.75	34.06	35.26	28.26	27.09	26.10	24.00	33.83	29.44	24.23	22.03	28.46
	SSIM	.8730	.9193	.9181	.8782	.9818	.9729	.8825	.9750	.9727	.9893	.9728	.9396
(C) SGSM (Ours)	PSNR	28.80	34.99	36.74	28.54	28.60	28.25	24.90	33.90	29.68	25.76	23.98	29.47
	SSIM	.9914	.9945	.9973	.9933	.9858	.9888	.9821	.9915	.9861	.9848	.9761	.9883

We report results in Table 4.9. The progression demonstrates that while adding spatial awareness (B) improves general restoration (PSNR $28.26 \rightarrow 28.46$), it struggles with structural preservation in complex mixtures. This is consistent with the formulation in Eq. (4.6): because the gate is purely multiplicative, low gate values can suppress useful clean-region features together with corrupted ones, which lowers structural fidelity even when average restoration strength increases. By adding the residual self-gate, SGSM (C) achieves the best of both worlds: it leverages the **intent** from the global code and the **location** from the spatial gate, while the $(1 + G_s)$ path preserves an identity route for cleaner regions. As a result, SGSM improves both fidelity (**29.47** dB) and structural similarity (**0.9883**), supporting our claim that mixed degradations require spatial selectivity without sacrificing feature stability.

4.5 Discussion

In this thesis, we aimed to improve mixed-degradation image restoration, where degradations co-exist and interact in ways that challenge many recent all-in-one restorers. Although unified backbones achieve strong results on single or narrowly-defined degradations, prior work often assumes mutually exclusive degradation control, applies guidance mostly as a global signal, or depends on explicit degradation maps/masks that are rarely available in real images. In practice, these choices make restoration less controllable under non-exclusive mixtures and particularly fragile when blur co-occurs with atmospheric or weather artifacts, where edge cues are suppressed and degradation cues become ambiguous. These gaps are evident when comparing composite-aware baselines such as OneRestore [21] against broader restoration backbones (e.g., Restormer [13], MPRNet [36],

MIRNet [62]), motivating a design that keeps composite semantics interpretable while enabling spatially selective correction under mixed degradations.

To address these gaps, we proposed the Composite Degradation Semantic Modulation Network (CDSM-Net), which separates “what degradations are present” from “how features should be restored.” We integrate vision–language priors using CLIP [30], then project the image embedding through a learnable degradation prototype memory to produce a multi-label routing signal and a compact global code, reducing irrelevant (non-degradation) content in the raw embedding. We further introduce self-gated spatial modulation to inject this global context selectively across regions without external masks, and we use explicit frequency decomposition in the backbone to strengthen detail recovery in blur-centric composites. These design choices were validated using the CDD paired composite benchmarks, blur-centric composite stress tests, real-world no-reference evaluation, and ablations that isolate each component.

Overall, the experimental evidence shows that this structured design improves both performance and controllability under composite inputs. On the CDD composite benchmark, CDSM-Net achieves state-of-the-art average SSIM and remains competitive in average PSNR, outperforming OneRestore in SSIM while staying close on PSNR [21]. Under blur-centric mixtures, explicit frequency separation combined with semantic modulation yields stronger robustness trends than prompt-only conditioning. Ablations further confirm that removing the semantic controller, spatial modulation, or frequency-aware design causes clear degradation, indicating that the gains do not come from backbone capacity alone. Importantly, the desirable properties introduced in Chapter 1 are evaluated here *operationally* rather than through a single benchmark number: semantic awareness is supported by the semantic-controller and routing ablations, spatial specificity by the SGSMS modulation ablation and localized qualitative behavior, interpretability by the inspectability of the routing weights α , composite robustness by the category-wise composite and real-world evaluations, and stable optimization by the convergence/uncertainty analysis. This makes the evaluation logic consistent with the problem statement: each desirable property is linked to a specific empirical test rather than asserted abstractly.

Based on these experimental observations, we now proceed to answer the research questions posed at the outset of this thesis. Each answer states the implementation

choice directly and then summarizes the supporting evidence from our comparisons and ablations.

- **How can vision–language priors be integrated to produce an interpretable, multi-label representation of non-exclusive mixed degradations?**

We integrate vision–language priors by extracting a CLIP image embedding and then converting these raw VLM features into an explicit degradation-focused representation using a learnable prototype memory bank. Because a raw CLIP embedding can be dominated by scene/content semantics and may carry limited direct degradation specificity, we project the image embedding through a learnable query and match it against a set of learnable degradation text prototypes, producing similarity scores that are turned into a sigmoid-based multi-label mixture vector. This multi-label mixture acts as a semantic bottleneck that retains degradation-relevant cues while suppressing unrelated content information, and it is used to synthesize a dynamic degradation descriptor that, together with the CLIP features, forms the compact global conditioning code injected into the restoration network. The effectiveness of this integration is supported by our ablations: removing the routing/memory mixing or forcing single-label control reduces composite robustness and increases degradation confusions, while the multi-label prototype-based routing yields more stable behavior on non-exclusive composite and blur-centric inputs and provides an interpretable mixture signal that can be inspected alongside the restored output.

- **How can global semantic context be spatially modulated to restore localized degradations without relying on external spatial priors (e.g., masks or degradation maps)?**

We achieve spatial modulation by injecting the global conditioning code through a self-gated spatial FiLM mechanism in which the spatial gate is predicted from intermediate feature maps, allowing the same global semantic context to be applied differently across regions without any external masks or degradation maps. This implementation makes modulation content-adaptive and spatially selective rather than uniformly global. The benefit is supported by the modulation ablations: self-gated spatial modulation reduces the common global-conditioning failure of over-correcting clean areas and under-correcting localized corruption, and it consistently improves restoration fidelity compared

to purely global or simpler spatial modulation variants. This is confirmed by the controlled FiLM ablation (Channel FiLM $\rightarrow$ Multiplicative Spatial FiLM $\rightarrow$ SGSF), where SGSF achieves the strongest average performance and the best structural preservation (Table 4.9).

• **To what extent does explicit frequency decomposition reduce spectral bias and improve detail recovery under blur-centric and composite degradations?**

We reduce spectral bias by explicitly decomposing the input into low-frequency structure and high-frequency details using wavelet factorization, then processing and fusing these bands with a backbone designed to preserve structure while recovering fine textures and edges. This implementation ensures that detail recovery is not left to implicit feature learning alone, which is typically biased toward low-frequency content under heavy blur. The impact is supported by blur-centric composite evaluations: the frequency-decomposed design recovers sharper edges and more coherent textures and avoids the oversmoothing compromise observed in backbone-only or non-decomposed variants, particularly when blur co-occurs with other degradations.

• **How does dynamic multi-objective learning affect training stability and balanced performance across competing restoration objectives?**

We implement dynamic multi-objective learning by using adaptive loss balancing (uncertainty-weighted weighting) so that competing objectives are scaled automatically during optimization, reducing gradient conflict across degradations and restoration targets. This implementation prevents a fixed set of weights from dominating training and improves stability when objectives pull the model in different directions. The effect is supported by training behavior: adaptive balancing yields smoother convergence and more consistent improvements across mixed degradations than fixed-weight training, and it reduces instability and objective collapse that appear when competing losses are not balanced appropriately.

4.5.1 Key Findings

- **Explicit degradation representation improves mixed-degradation controllability.** The controller design (multi-label degradation mixture and global

conditioning code) provides a concrete variable that can be inspected and compared across inputs. This reduces ambiguity in composite cases where multiple degradations co-activate, and it enables the model to avoid hard single-choice conditioning that is misaligned with real mixed degradations. This claim is supported by our semantic-controller ablations, where removing mixture formation or forcing single-choice routing consistently reduces composite restoration quality relative to the full multi-label controller.

- Spatially selective modulation reduces “global compromise” behavior.** The SGSM mechanism supports spatial heterogeneity by allowing the same global code to be applied differently across regions. Empirically, this improves localized artifact removal (e.g., streak-like patterns) while preserving less-degraded areas, indicating that spatial selectivity is a practical substitute for unavailable pixel-level degradation masks. This is corroborated by the controlled modulation ablation, where the full SGSM variant consistently outperforms global-only and simpler spatial modulation baselines, indicating that spatial selectivity improves both fidelity and structural preservation under mixtures.
- Frequency-aware decomposition stabilizes detail recovery under blur-centric mixtures.** The wavelet LF/HF separation introduces an inductive bias that helps preserve structural content (LF) while explicitly targeting high-frequency degradation remnants (HF). This is especially relevant when blur co-occurs with other degradations, since blur can mask the cues needed by conventional restorers to recover fine details. Consistent with this, our blur-centric robustness evaluation shows that CDSM-Net improves over a strong composite-aware baseline on blur-driven composites, indicating more reliable detail recovery when blur masks discriminative cues.
- Multi-objective balancing is necessary for mixed restoration.** Composite restoration introduces competing gradients (e.g., denoising vs. sharpening; dehazing vs. noise amplification). The training-stability results support that adaptive loss balancing is not merely an optimization trick, but a requirement to prevent collapse or oscillation when the model must satisfy multiple restoration behaviors

simultaneously. Empirically, adaptive balancing produces smoother convergence and stronger late-stage restoration trends than fixed-weight training, indicating reduced gradient interference among competing objectives.

- **The semantic controller improves composite-restoration controllability.**

The predicted multi-label degradation scores provide a structured intermediate control signal that helps the backbone avoid hard single-choice conditioning under mixed degradations. In this thesis, the main evidence for this claim comes from the semantic-controller ablations and the multi-label versus single-label comparison, where constraining the controller or removing mixture formation consistently reduces restoration quality.

4.5.2 Limitations

- The weakest real-world generalization is observed on NH-HAZE, indicating that haze correction learned from our synthetic composite protocol does not fully transfer to real, spatially varying atmospheric scattering with strong depth- and illumination-dependent effects. In such cases, locally aggressive enhancement can improve visibility but still yield non-natural statistics, suggesting that stronger physics-aware constraints or real-haze adaptation would be beneficial.
- Snow-heavy regimes remain challenging, especially when snow behaves as occlusion/accumulation rather than thin airborne streaks, which can dominate pixel-level error even when global structure is preserved. In addition, visually correlated weather cues can occasionally lead the semantic controller to partially co-activate related primitives (e.g., snow with rain), implying that more discriminative degradation prototypes and harder negative supervision are needed for robust separation under ambiguous scenes.

Chapter 5

Conclusion and recommendation

5.1 Conclusion

Mixed-degradation image restoration remains challenging because real corruptions rarely occur in isolation and often vary across regions of the same image. Although recent all-in-one restoration methods have improved generality, several limitations persist: many approaches still rely on implicitly single-label assumptions that do not reflect non-exclusive mixtures; conditioning is frequently applied globally, making it hard to handle region-wise composites without external masks; and spectral bias can limit detail recovery under blur-centric and heavily mixed conditions. In addition, distribution shift between synthetic composites and real captures (especially non-homogeneous haze and occlusion-like snow) continues to expose robustness gaps. And many existing systems provide limited interpretability for composite decisions.

To address these issues, we proposed the *Composite Degradation Semantic Modulation Network (CDSM-Net)*, which separates *degradation understanding* from *restoration execution*. Our first contribution, the *Multi-label Degradation Semantic Modulator (MDSM)*, leverages vision-language priors and a degradation-text memory to produce an interpretable, non-exclusive degradation state that can represent mixtures instead of forcing a single mode. Our second contribution, *Self-Gated Spatial Modulation (SGSM)*, injects the global semantic code selectively across spatial regions using a feature-derived gate, enabling localized restoration without requiring segmentation masks or degradation maps. Our third contribution, the *WAVeViM* backbone, applies explicit wavelet frequency decomposition to reduce spectral bias by separating low-frequency structure from high-frequency detail, strengthening restoration in blur-centric and composite regimes.

Extensive experiments show that our structured design improves stability and restoration quality under mixed degradations across both paired and real-world evaluations. On CDD-11, CDSM-Net achieves **29.47 dB** average PSNR and **0.9883** average SSIM,

delivering very strong structural preservation while remaining highly competitive in fidelity; on the blur-centric CDD-13 subset, it reaches **32.2104 dB** on **Blur** and **28.3501 dB** on **Blur+Haze** (average **30.2802 dB**), confirming the benefit of frequency-aware restoration when blur suppresses discriminative cues, and it maintains a strong full CDD-13 average of **29.5920 dB / 0.9893**. Real-world no-reference results further highlight both strengths and remaining challenges: the model achieves the best NIQE on LOL-v2 (**5.1309**).

In conclusion, this thesis shows that combining multi-label vision–language guidance, mask-free spatially selective modulation, and explicit frequency-aware restoration yields an effective and interpretable solution for mixed-degradation image restoration. A remaining limitation is reduced robustness on snow-heavy composites and non-homogeneous real haze, motivating the future work directions outlined in the next section. Overall, this thesis provides evidence that mixed-degradation restoration benefits from explicit, interpretable degradation semantics combined with spatially adaptive conditioning and frequency-aware restoration, advancing the practical reliability of all-in-one models in real-world settings.

5.2 Recommendations and Future Work

Although the proposed approach achieves strong performance on composite benchmarks and shows promising behavior on challenging low-light restoration, the following recommendations describe **direct continuations of our formulation** (MDSM, SGSM, WAVEViM) and the CDD protocol used in this thesis. For researchers reusing our data protocol, we recommend reporting single vs composite statistics and severity ranges and extending the protocol with new mixtures/intensity levels to stress-test generalization. For clarity, each item is stated as a compact **gap** → **extension** → **evidence** direction to support measurable improvements in mixed-degradation restoration.

- **Stronger semantic controller training (separate MDSM training).** *Evidence: improved multi-label routing reliability (e.g., calibration/ECE) and higher PSNR/SSIM on composite categories.* A practical direction is to pre-train the semantic controller independently before full end-to-end restoration training. For example, MDSM can be trained with explicit multi-label supervision on compos-

ite mixtures to stabilize routing behavior early and reduce degradation confusion. This can also include harder negative sampling that specifically targets commonly confused pairs (e.g., snow vs. rain, haze vs. smoke) to make the routing signal more discriminative.

- **Degradation-aware CLIP adaptation using degraded–clean pairs.** *Evidence: reduced content leakage in degradation embeddings and measurable gains on composite-category restoration.* The current vision–language priors can be further aligned to restoration by fine-tuning (or distilling) CLIP-style encoders on degraded/clean pairs so that embeddings encode degradation attributes more directly, rather than being dominated by scene content. A degradation-contrastive objective that pulls together embeddings of the same degradation mixture while separating different mixtures can improve prompt–degradation alignment and reduce semantic leakage from content.
- **Scale training data and composite diversity (triple mixtures and richer augmentation).** *Evidence: gains on held-out mixture severities/combinations and reduced worst-case composite error.* Future training should expand beyond limited composite combinations and include more diverse degradation orderings and intensities, including three-or-more degradation mixtures as a standard stress case. Data augmentation can be strengthened with small-angle rotations (e.g., 4–5 degrees), stronger photometric perturbations, and more realistic blur kernels and noise models, improving robustness to unseen camera motion and capture variability.
- **Improve real-world robustness with paired real datasets and physics-grounded constraints.** *Evidence: improved no-reference quality (e.g., NIQE) and fewer artifacts on non-homogeneous haze and snow-heavy scenes.* To address failures on non-homogeneous haze and snow-heavy scenes, incorporating paired real-world datasets (where available) and stronger physically grounded constraints is recommended. For haze, adding global airlight consistency and depth-aware regularization can reduce over-local contrast artifacts that no-reference metrics penalize. For snow, more diverse real snow appearances and occlusion-aware objectives can better distinguish structured snow from scene texture and illumination effects.

REFERENCES

- [1] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.
- [2] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Pearson, 4th ed., 2018.
- [3] A. Foi *et al.*, “Practical poissonian–gaussian noise modeling and fitting for single-image raw-data,” *IEEE Transactions on Image Processing*, 2008.
- [4] S. W. Hasinoff *et al.*, “Burst photography for high dynamic range and low-light imaging on mobile cameras,” *ACM Transactions on Graphics*, 2016.
- [5] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, “Removing camera shake from a single photograph,” *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 787–794, 2006.
- [6] A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
- [7] K. Dabov *et al.*, “Image denoising by sparse 3-D transform-domain collaborative filtering,” *IEEE Transactions on Image Processing*, 2007.
- [8] M. Aharon, M. Elad, and A. M. Bruckstein, “ K -SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation,” *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [9] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Transactions on Image Processing*, 2006.
- [10] H. Chen, Y. Wang, *et al.*, “Pre-trained image processing transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [11] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Swinir: Image restoration using swin transformer,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.
- [12] Z. Wang *et al.*, “Uformer: A general u-shaped transformer for image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

- [13] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, “Restormer: Efficient transformer for high-resolution image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5728–5739, 2022.
- [14] L. Chen, X. Chu, X. Zhang, and J. Sun, “Simple baselines for image restoration,” in *European Conference on Computer Vision (ECCV)*, 2022.
- [15] Z. Tu *et al.*, “Maxim: Multi-axis mlp for image processing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [16] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, “All-in-one image restoration for unknown corruption,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17431–17441, 2022.
- [17] J. Jiang *et al.*, “A survey on all-in-one image restoration: Taxonomy, evaluation and future trends,” *arXiv preprint arXiv:2410.15067*, 2025.
- [18] J. Zhang *et al.*, “Ingredient-oriented multi-degradation learning for image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [19] 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 (NeurIPS)*, vol. 36, 2023. arXiv:2306.13090.
- [20] M. V. Conde, G. Geigle, and R. Timofte, “Instructir: High-quality image restoration following human instructions,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.
- [21] Y. Guo, Y. Gao, Y. Lu, H. Zhu, R. W. Liu, and S. He, “Onerestore: A universal restoration framework for composite degradation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 255–272, 2024.
- [22] Y. Jiang, Z. Zhang, T. Xue, and J. Gu, “Autodir: Automatic all-in-one image restoration with latent diffusion,” in *European Conference on Computer Vision (ECCV)*, 2024.
- [23] D. Zheng *et al.*, “Selective hourglass mapping for universal image restoration based on diffusion model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [24] Y. Cui *et al.*, “Adair: Adaptive all-in-one image restoration via frequency mining and modulation,” 2024. arXiv:2403.14614.

- [25] M. Yao, R. Xu, Y. Guan, J. Huang, and Z. Xiong, “Neural degradation representation learning for all-in-one image restoration,” *IEEE transactions on image processing*, 2024.
- [26] J. Lin *et al.*, “Unirestorer: Universal image restoration via adaptively estimating image degradation at proper granularity,” 2024. arXiv:2412.20157.
- [27] W. Ai, Z. Zhang, *et al.*, “Lora-ir: Low-rank adaptation for efficient all-in-one image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [28] C.-J. Qin, R.-Q. Wu, Z. Liu, X. Lin, C.-L. Guo, H. H. Park, and C. Li, “Restore anything with masks: Leveraging mask image modeling for blind all-in-one image restoration,” in *European Conference on Computer Vision*, pp. 364–380, Springer, 2024.
- [29] J. Hu, L. Jin, Z. Yao, and Y. Lu, “Universal image restoration pre-training via degradation classification,” in *International Conference on Learning Representations (ICLR)*, 2025. ICLR 2025 poster.
- [30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.
- [31] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schöen, “Controlling vision-language models for multi-task image restoration,” in *International Conference on Learning Representations (ICLR)*, 2024.
- [32] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schöen, “Controlling vision-language models for multi-task image restoration,” in *International Conference on Learning Representations (ICLR)*, 2024.
- [33] S. G. Narasimhan and S. K. Nayar, “Contrast restoration of weather degraded images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 713–724, 2003.
- [34] S. M. A. Sharif, A. Rehman, F. A. Dharejo, R. Timofte, and R. A. Naqvi, “Degradation-aware all-in-one image restoration via latent prior encoding,” 2025. arXiv:2509.17792.
- [35] L. Tu, J. Wu, X. Luo, and Z. Jin, “Unifying heterogeneous degradations: Uncertainty-aware diffusion bridge model for all-in-one image restoration,” *arXiv preprint arXiv:2601.21592*, 2026.
- [36] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, “Multi-stage progressive image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

- [37] Z. Liu, S. Zhou, Y. Dai, Y. Wang, Y. An, and X. Zhao, “Dpmambair: All-in-one image restoration via degradation-aware prompt state space model,” *arXiv preprint arXiv:2504.17732*, 2025.
- [38] J. Jiang, Z. Chen, Y. Su, P. Zhang, and Y. Hu, “All-in-one image restoration via diffusion models with degradation perception and semantic enhancement,” *Neuro-computing*, vol. 656, p. 131517, 2025.
- [39] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, “Non-uniform deblurring for shaken images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 491–498, 2010.
- [40] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3942–3951, 2018.
- [41] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2337–2346, 2019.
- [42] E. Zamfir, Z. Wu, N. Mehta, Y. Tan, D. P. Paudel, Y. Zhang, and R. Timofte, “Complexity experts are task-discriminative learners for any image restoration,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12753–12763, 2025.
- [43] D. Kim *et al.*, “Text-aware image restoration with diffusion models,” 2025. [arXiv:2506.09993](#).
- [44] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [45] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [46] X. Tian *et al.*, “Degradation-aware feature perturbation for all-in-one image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [47] J. Ma, S. Hu, X. Zhang, J. Wan, J. Huang, L. Zhang, and S. Khan, “Evoir: Towards all-in-one image restoration via evolutionary frequency modulation,” 2025.
- [48] J. Hu, Z. Yao, L. Jin, and Y. Lu, “Bridging degradation discrimination and generation for universal image restoration,” *arXiv preprint arXiv:2602.00579*, 2026.
- [49] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer, 2010.

- [50] G. Healey and R. Kondepudy, “Radiometric CCD camera calibration and noise estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994.
- [51] M. R. Banham and A. K. Katsaggelos, “Digital image restoration,” *IEEE Signal Processing Magazine*, vol. 14, pp. 24–41, Mar. 1997.
- [52] S. K. Nayar and S. G. Narasimhan, “Vision in bad weather,” in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 820–827, IEEE, 1999.
- [53] A. Gupta, N. Joshi, C. Lawrence Zitnick, M. Cohen, and B. Curless, “Single image deblurring using motion density functions,” in *European conference on computer vision*, pp. 171–184, Springer, 2010.
- [54] W. Yang *et al.*, “Single image deraining: From model-based to data-driven and beyond,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [55] C. Wei, W. Wang, W. Yang, and J. Liu, “Deep retinex decomposition for low-light enhancement,” *arXiv preprint arXiv:1808.04560*, 2018.
- [56] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D*, 1992.
- [57] K. He *et al.*, “Deep residual learning for image recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [58] K. Zhang *et al.*, “Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising,” *IEEE Transactions on Image Processing*, 2017.
- [59] A. Gu, K. Goel, and C. Re, “Efficiently modeling long sequences with structured state spaces,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [60] K. Zhang *et al.*, “Learning deep CNN denoiser prior for image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [61] Y. Zhang *et al.*, “Residual dense network for image super-resolution,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [62] S. W. Zamir *et al.*, “Learning enriched features for real image restoration and enhancement,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

- [63] X. Ji *et al.*, “U²-former: A nested u-shaped transformer for image restoration,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [64] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [65] A. Gu, K. Goel, and C. Re, “Efficiently modeling long sequences with structured state spaces,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [66] Y. Liu *et al.*, “VMamba: Visual state space model,” in *arXiv preprint arXiv:2401.10166*, 2024.
- [67] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, “Mambair: A simple baseline for image restoration with state-space model,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.
- [68] Y.-C. Lin, Y.-S. Xu, H.-W. Chen, H.-K. Kuo, and C.-Y. Lee, “Eamamba: Efficient all-around vision state space model for image restoration,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11708–11719, 2025.
- [69] C. Saharia *et al.*, “Palette: Image-to-image diffusion models,” in *SIGGRAPH*, 2022.
- [70] B. Kawar *et al.*, “Denoising diffusion restoration models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [71] Y. Zhu *et al.*, “Plug-and-play denoising diffusion models for image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [72] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 6840–6851, 2020.
- [73] O. Kupyn *et al.*, “Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [74] X. Wang *et al.*, “Esrgan: Enhanced super-resolution generative adversarial networks,” in *ECCVW*, 2018.
- [75] L. Liu *et al.*, “Wavelet-based dual-branch network for image demoiréing,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [76] L. Dai *et al.*, “Awnet: Attentive wavelet network for image ISP,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

- [77] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, “All-in-one image restoration for unknown corruption,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17431–17441, 2022.
- [78] K. Zhang *et al.*, “Dpir: Deep plug-and-play image restoration,” in *CVPR Workshops*, 2021.
- [79] M. Shao, Y. Liu, Y. Cheng, Y. Wan, and C. Wang, “Adaptive fuzzy degradation perception based on clip prior for all-in-one image restoration,” *IEEE Transactions on Fuzzy Systems*, 2024.
- [80] E. Zamfir, Z. Wu, N. Mehta, D. P. Paudel, Y. Zhang, and R. Timofte, “Efficient degradation-aware any image restoration,” *arXiv preprint arXiv:2405.15475*, 2024.
- [81] Y. Liu *et al.*, “Unifying image processing as visual prompting question answering,” in *ICML*, 2024.
- [82] M. V. Conde, G. Geigle, and R. Timofte, “Instructir: High-quality image restoration following human instructions,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.
- [83] H. Duan, X. Min, S. Wu, W. Shen, and G. Zhai, “Uniprocessor: A text-induced unified low-level image processor,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 180–199, 2024.
- [84] Y. Wang, Y. Li, Z. Zheng, X.-P. Zhang, and M. Wei, “M2restore: Mixture-of-experts-based mamba-cnn fusion framework for all-in-one image restoration,” *IEEE Transactions on Image Processing*, vol. 34, pp. 8086–8100, 2025.
- [85] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, “Transweather: Transformer-based restoration of images degraded by adverse weather conditions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2353–2363, 2022.
- [86] J. Han, W. Li, P. Fang, C. Sun, J. Hong, M. A. Armin, L. Petersson, and H. Li, “Blind image decomposition,” in *European conference on computer vision*, pp. 218–237, Springer, 2022.
- [87] X. Yu, S. Zhou, H. Li, and L. Zhu, “Multi-expert adaptive selection: Task-balancing for all-in-one image restoration,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 5, pp. 4619–4634, 2025.
- [88] Z. Yue, J. Wang, and C. C. Loy, “Efficient diffusion model for image restoration by residual shifting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 1, pp. 116–130, 2025.
- [89] M.-L. Zhang and Z.-H. Zhou, “A review on multi-label learning algorithms,” *IEEE Transactions on Knowledge and Data Engineering*, 2014.

- [90] K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.
- [91] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [92] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
- [93] L. Zhang *et al.*, “A feature-enriched completely blind image quality evaluator,” *IEEE Transactions on Image Processing*, 2015.
- [94] K. Zhou *et al.*, “Learning to prompt for vision-language models,” 2021. arXiv:2109.01134.
- [95] K. Zhou *et al.*, “Conditional prompt learning for vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [96] S. Chen, T. Ye, K. Zhang, Z. Xing, Y. Lin, and L. Zhu, “Teaching tailored to talent: Adverse weather restoration via prompt pool and depth-anything constraint,” in *European conference on computer vision*, Springer, 2024.
- [97] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.
- [98] N. Shazeer *et al.*, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” 2017. arXiv:1701.06538.
- [99] W. Fedus *et al.*, “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” 2021. arXiv:2101.03961.
- [100] Y. Huang, J. Huang, J. Liu, M. Yan, Y. Dong, J. Lyu, C. Chen, and S. Chen, “Wavedm: Wavelet-based diffusion models for image restoration,” *IEEE Transactions on Multimedia*, 2023.
- [101] C. Mou *et al.*, “Deep generalized unfolding networks for image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [102] Y. Cui, W. Ren, and A. Knoll, “Oknet: Omni-kernel network for image restoration,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 1426–1434, 2024.
- [103] B. Song, X. Chen, S. Xu, and J. Zhou, “Under-display camera image restoration with scattering effect,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12546–12555, 2023.

- [104] J. Ma *et al.*, “Prores: Exploring degradation-aware prompts for universal image restoration,” 2023. arXiv:2306.13653.
- [105] X. Kong *et al.*, “Towards effective multiple-in-one image restoration: A sequential and prompt learning strategy,” 2024. arXiv:2401.03379.
- [106] G. Wu *et al.*, “Boosting all-in-one image restoration via self-improved privilege learning,” 2025. arXiv:2505.24207.
- [107] S. M. A. Sharif, A. Rehman, F. A. Dharejo, R. Timofte, and R. A. Naqvi, “Degradation-aware all-in-one image restoration via latent prior encoding,” 2025. arXiv:2509.17792.
- [108] O. Ozdenizci and R. Legenstein, “Restoring vision in adverse weather conditions with patch-based denoising diffusion models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [109] Y. Meng *et al.*, “Joint conditional diffusion model for image restoration with mixed degradations,” 2024. arXiv:2404.07770.
- [110] Y. Zhou *et al.*, “Unildiff: Unlocking the power of diffusion priors for all-in-one image restoration,” 2025. arXiv:2507.23685.
- [111] F. Yu *et al.*, “Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild,” 2024. arXiv:2401.13627.
- [112] I. Research *et al.*, “Instantir: Blind image restoration with instant generative reference,” 2024. arXiv:2410.06551.
- [113] M. Zhou *et al.*, “An efficient global modeling paradigm for image restoration,” in *ICML*, 2023.
- [114] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.
- [115] M. R. Banham and A. K. Katsaggelos, “Digital image restoration,” *IEEE Signal Processing Magazine*, vol. 14, pp. 24–41, Mar. 1997.
- [116] S. G. Narasimhan and S. K. Nayar, “Vision and the atmosphere,” *IJCV*, 2002.
- [117] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.
- [118] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, “Raise: A raw images dataset for digital image forensics,” in *Proceedings of the ACM Multimedia Systems Conference (MMSys)*, 2015.

- [119] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
- [120] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, “Sparse gradient regularized deep retinex network for robust low-light image enhancement,” *IEEE Transactions on Image Processing*, vol. 30, pp. 2072–2086, 2021.
- [121] C. O. Ancuti, C. Ancuti, and R. Timofte, “Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.
- [122] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, “Desnownet: Context-aware deep network for snow removal,” *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3064–3073, 2018.
- [123] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [124] B. Alsallakh, N. Kokhlikyan, V. Miglani, J. Yuan, and O. Reblitz-Richardson, “Mind the pad – cnns can develop blind spots,” in *International Conference on Learning Representations (ICLR)*, 2021.
- [125] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 85–100, 2018.
- [126] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, “Deterministic edge-preserving regularization in computed imaging,” *IEEE Transactions on Image Processing*, vol. 6, no. 2, pp. 298–311, 1997.
- [127] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” *arXiv preprint arXiv:2312.00752*, 2023.
- [128] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations (ICLR)*, 2019.
- [129] I. Loshchilov and F. Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [130] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, “Learning enriched features for fast image restoration and enhancement,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.
- [131] C. Mou, Q. Wang, J. Zhang, *et al.*, “Dgunet: Deep generalized unfolding network for image restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

- [132] L. Chen, X. Chu, X. Zhang, and J. Sun, “Simple baselines for image restoration,” in *European Conference on Computer Vision (ECCV)*, 2022.
- [133] B. Song, X. Chen, S. Xu, and J. Zhou, “Under-display camera image restoration with scattering effect,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12546–12555, 2023.
- [134] M. Zhou, J. Huang, C. Guo, and C. Li, “Fourmer: A universal fourier-guided transformer for image restoration,” in *International Conference on Machine Learning (ICML)*, 2023.
- [135] Y. Cui, W. Ren, and A. Knoll, “Omni-kernel network for image restoration,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 1426–1434, 2024.
- [136] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, “Transweather: Transformer-based restoration of images degraded by adverse weather conditions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2353–2363, 2022.
- [137] O. Ozdenizci and R. Legenstein, “Restoring images degraded by adverse weather conditions via a hybrid diffusion model,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.
- [138] Y. Zhu, Y. Xie, K. Zhou, Z. Jiang, C. Li, and M.-M. Cheng, “Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [139] J. Hu, L. Jin, Z. Yao, and Y. Lu, “Universal image restoration pre-training via degradation classification,” in *International Conference on Learning Representations (ICLR)*, 2025. ICLR 2025 poster.
- [140] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, “From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3063–3072, 2020.
- [141] C. Wei, W. Wang, W. Yang, and J. Liu, “Deep retinex decomposition for low-light enhancement,” *arXiv preprint arXiv:1808.04560*, 2018.
- [142] W.-T. Chen, X. Liu, J. Yang, *et al.*, “All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

- [143] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, “Desnownet: Context-aware deep network for snow removal,” *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3064–3073, 2018.
- [144] C. O. Ancuti, C. Ancuti, R. Timofte, *et al.*, “Nh-haze: An image dehazing benchmark with non-homogeneous haze,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.
- [145] Y. Zhang, J. Zhang, and X. Guo, “Kindling the darkness: A practical low-light image enhancer,” in *ACM International Conference on Multimedia (ACM MM)*, 2019.
- [146] Y. Jiang, X. Gong, D. Liu, *et al.*, “Enlightengan: Deep light enhancement without paired supervision,” *arXiv preprint arXiv:1906.06972*, 2019.
- [147] S. Zhou, C. Li, and C. Change Loy, “Lednet: Joint low-light enhancement and deblurring in the dark,” in *European conference on computer vision*, pp. 573–589, Springer, 2022.
- [148] A. Sharma and R. T. Tan, “Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11977–11986, 2021.
- [149] W.-T. Chen, H.-Y. Fang, J.-J. Ding, C.-C. Tsai, and S.-Y. Kuo, “JSTASR: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal,” in *Computer Vision – ECCV 2020*, pp. 754–770, 2020.
- [150] K. Zhang, X. Li, Y. Zhang, *et al.*, “Deep dense multi-scale network for snow removal using semantic and depth priors,” *IEEE Transactions on Image Processing*, 2021.
- [151] Y. Dong, Y. Liu, H. Chen, *et al.*, “Multi-scale boosted dehazing network with dense feature fusion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [152] Y. Song, Y. He, C. Qian, *et al.*, “Dehazformer: Vision transformers for single image dehazing,” *arXiv preprint arXiv:2204.03836*, 2022.
- [153] M. Hong, Y. Xie, C. Li, and Y. Qu, “Distilling image dehazing with heterogeneous task imitation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.