https://doi.org/10.31436/iiumej.v27i3.4662

DesnowOracle: Quantifying Mask-Oracle Gap Recovery by a Training-Free Snow Prior in Compact Image Desnowing

TEDDY SURYA GUNAWAN[1*], MIRA KARTIWI[2]

[1]*Department of Electrical and Computer Engineering, Kulliyyah of Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia*

[2]*Department of Information Systems, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia*

**Corresponding author: tsgunawan@iium.edu.my*

(Received: 21 July 2026; Accepted: 19 August 2026; Published online: 11 September 2026)

***ABSTRACT*:** Snow occludes visual information needed by outdoor perception systems. This study evaluates whether a deterministic, training-free spatial snow-response map improves compact image desnowing beyond the effect of its conditioning branch. All conditions use a fixed approximately 2-million-parameter backbone and matched controls. Under the original 2,500-image, 100-epoch protocol, estimated-map SSM improves mean PSNR by 1.004 dB over Base and by 0.490 and 0.517 dB over zero- and shuffled-input SSM. A paired analysis averaging predictions across three seeds gives median gains of 0.387 and 0.420 dB, respectively. Under an extended protocol that also changes training-set size and learning-rate scheduling, this ordering reverses: estimated-map SSM trails zero and shuffled controls by 0.109 and 0.160 dB in mean PSNR. Thus, the experiments demonstrate protocol sensitivity, not an isolated convergence effect. A mask oracle remains an architecture-specific, synthesis-privileged diagnostic reference; the estimated map recovers 12.4% of its short-protocol PSNR headroom. Results on 1,329 unpaired real images are mixed and indirect: no-reference indices favor the estimated map, whereas unlabelled detector counts do not distinguish its content from the controls. These findings show why spatial-prior claims require matched controls, correct repeated-measures analysis, and protocol checks.

***ABSTRAK*:** Salji menghalang maklumat visual yang diperlukan oleh sistem persepsi luar bangunan. Kajian ini menilai sama ada peta tindak balas salji spatial yang deterministik dan tanpa latihan meningkatkan penyahsaljian imej padat melebihi kesan cabang pengkondisian. Semua keadaan menggunakan rangkaian asas tetap dengan kira-kira 2 juta parameter dan kawalan yang sepadan. Di bawah protokol asal 2,500 imej dan 100 epok, SSM dengan peta anggaran meningkatkan purata PSNR sebanyak 1.004 dB berbanding Base dan sebanyak 0.490 serta 0.517 dB berbanding SSM input sifar dan teracak. Purata analisis ramalan berpasangan merentas tiga benih memberikan peningkatan median masing-masing sebanyak 0.387 dan 0.420 dB. Di bawah protokol lanjutan yang turut mengubah saiz set latihan dan penjadualan kadar pembelajaran, susunan ini berbalik kepada SSM dengan peta anggaran ketinggalan berbanding kawalan sifar dan teracak sebanyak 0.109 dan 0.160 dB dalam purata PSNR. Oleh itu, eksperimen menunjukkan kepekaan terhadap protokol, bukan kesan penumpuan terasing. Topeng Oracle kekal sebagai rujukan diagnostik khusus seni bina dan keistimewaan sintesis; peta anggaran memulihkan 12.4% ruang PSNR protokol pendek. Dapatan daripada 1,329 imej sebenar tanpa pasangan adalah bercampur dan tidak langsung: indeks tanpa rujukan memihak kepada peta anggaran, manakala kiraan pengesan tanpa label tidak membezakan kandungannya daripada kawalan. Dapatan ini menunjukkan bahawa tuntutan sebelum spatial memerlukan kawalan sepadan, analisis ukuran berulang yang betul, dan semakan protokol.

https://doi.org/10.31436/iiumej.v27i3.4662

KEYWORDS: *Image Desnowing; Adverse Weather Perception; Spatial Conditioning; Snow Severity Prior; Compact Restoration*

1. INTRODUCTION

Winter storms create a compound threat to transportation and machine perception in countries with recurrent seasonal snowfall. In the United States alone, an estimated 101,390 police-reported crashes occurred under snow or sleet conditions in 2023, including 22,293 injury crashes and 320 fatal crashes, as reported by the National Highway Traffic Safety Administration (NHTSA) in 2025. During a snowstorm, airborne particles reduce visibility and contrast, obscure pedestrians, vehicles, traffic signals, and road markings, and produce veiling and reflected-light artifacts. Accumulated snow can further conceal lane boundaries and alter the appearance of familiar objects. These effects degrade the visual evidence required by driver-assistance systems, autonomous vehicles, aerial platforms, outdoor robots, and surveillance systems precisely when reliable perception is most critical. Desnowing is therefore not merely a visual-enhancement task. Restoration-aware snowy-weather detection has produced mean average precision gains of 3.80, 2.10, and 1.86 percentage points across three real-world datasets, demonstrating that recovering useful image structure can improve downstream perception [1].

Mask-guided desnowing can improve restoration, but attribution remains difficult: a conditioned model may benefit from extra capacity or optimization even when its spatial input is uninformative. The relevant comparison is therefore not only against an unconditioned network, but also against the same conditioning branch supplied with zero and shuffled maps. This study measures that distinction and tests whether the resulting ordering changes under a substantially extended protocol.

Snow is especially difficult to remove because opaque, irregular particles can replace local scene structure rather than merely attenuating it. DesnowNet introduced Snow100K and used a learned intermediate mask to guide restoration [2]. Subsequent methods modeled contradict-channel information, particle size, transparency, semantic context, depth, and scale-aware interactions [3-6]. Lightweight architectures have also addressed the tension between restoration quality and network capacity [7-9]. More recently, a snow-shape prior was introduced by [10] as surrogate supervision when real snow masks are unavailable. Their prior replaces mask supervision during training, whereas the present prior replaces a mask input at inference. Neither purpose is interchangeable, and the earlier study does not quantify recovery relative to a ground-truth-mask oracle.

This study introduces DesnowOracle, a controlled framework for measuring spatial-prior effects within a fixed compact backbone. An identity-initialized spatial modulation head receives an estimated map, an all-zero map, a map from another image, a dilated ground-truth mask, or the exact ground-truth mask. The design separates the conditioning-branch effect from the effect of prior content. A second protocol tests whether the ordering is stable when training-set size, training duration, and learning-rate scheduling change together.

The study makes four contributions. First, it specifies a deterministic, training-free spatial snow-response map and evaluates its mask agreement by particle scale. Second, it compares estimated, zero, shuffled, dilated-mask, and exact-mask inputs through the same spatial modulation branch, with the exact mask treated as a synthesis-privileged diagnostic reference. Third, it reports scale-stratified restoration and an exploratory degradation-ladder analysis showing that mask AUC correlates with gain within that ladder but does not predict the estimated map out of family. Fourth, it shows that the estimated map ranks above the

https://doi.org/10.31436/iiumej.v27i3.4662

uninformative controls under the original protocol but below them under an extended protocol that changes data volume, optimizer steps, and learning-rate scheduling. Real-image no-reference indices, unlabelled detector responses, and measured computational costs provide supporting evidence with explicit limitations.

2. RELATED WORK

The literature has established that snow masks, spatial priors, multiscale representations, and degradation-aware conditioning can improve image restoration, but it has not yet determined how much of that benefit arises from valid spatial information rather than additional model capacity.

2.1. Desnowing and Snow Priors

DesnowNet established Snow100K with paired synthetic snowy and clean images, snow masks, and an unpaired real subset, and used a learned intermediate mask with translucency recovery [2]. JSTASR modeled particle size and transparency through modified partial convolution [4], while HDCW-Net used a handcrafted contrast-channel signal as a training loss rather than an explicit network input [5]. DDMSNet introduced semantic and depth priors [6], and SnowFormer modeled scale-dependent context using a transformer [11]. These methods show that snow location, scale, and opacity are useful, but they do not determine how much of the improvement comes from accurate spatial guidance rather than architectural capacity.

Compact desnowing introduces a second challenge: preserving sufficient contextual reasoning without relying on a large model. The trade-off between network capacity and restoration performance was explicitly studied by [8]. LMQFormer combined a learned coarse Laplace-prior mask with a 2.181-million-parameter query transformer [9], while an efficient multiscale context-aware architecture was developed by [7]. These methods establish that compact desnowing is feasible, but baseline comparisons alone cannot establish whether a prior-guided model actually uses the prior spatially.

Recent prior-based methods sharpen this distinction. A snow-shape prior was embedded in a deep unfolding framework to replace impractical real-mask supervision [10]. By contrast, the present study supplies a deterministic prior directly to the restoration network and tests whether its spatial correspondence is causally useful. BAT-Net jointly predicts a snow mask and clean image through coupled transformer decoders, allowing the two tasks to refine each other [12]. Its mask remains a learned internal representation rather than a controllable input that can be removed, shuffled, spatially degraded, or replaced by ground truth. SnowMamba reports PSNR, SSIM, LPIPS, and DISTS separately across Snow100K-S, Snow100K-M, and Snow100K-L [13], confirming the value of scale-stratified evaluation. The distinction in the present study is therefore not stratification itself, but using stratification to measure how much mask-oracle headroom an estimated prior recovers.

Real-world evaluation has also advanced beyond isolated qualitative examples. SnowMaster uses multimodal large language model feedback and multiple restoration models to address real snowy scenes [14], while WeatherBench provides aligned real degraded and clean pairs for rain, snow, and haze [15]. These developments expose the limitation of the unpaired Snow100K real subset. Accordingly, the present study uses no-reference metrics on that subset as conservative supporting evidence and does not treat them as a substitute for paired real-world evaluation.

https://doi.org/10.31436/iiumej.v27i3.4662

2.2. General Adverse-Weather Restoration and Conditioning Mechanisms

All-in-one restoration seeks to handle multiple degradations within a shared model, with recent work organized around degradation representations, prompts, adaptive pathways, and modulation [16]. AirNet learns contrastive degradation representations [17], TransWeather uses weather-type queries [18], and TKL distills degradation-specific teachers into a unified student [3]. More recent architectures use self-estimated degradation prompts, efficient convolutional prompts, and diffusion guidance [19-21]. These approaches primarily condition restoration through latent degradation representations or prompts rather than through an externally testable per-pixel prior.

AdaIR is the closest recent conceptual comparator because it mines frequency components and uses them to modulate restoration features [22]. The present study instead conditions features using an explicit spatial severity map. This design follows the broader principle of feature-wise linear modulation [23] and spatial feature transformation [24], but differs in two respects: the conditioning map is deterministic and training-free, and its value is evaluated through matched falsification and mask-oracle controls.

Taken together, prior work has established the effectiveness of learned masks, surrogate priors, multiscale representations, degradation prompts, and feature modulation. What remains unmeasured is how much mask-oracle headroom a training-free spatial prior recovers under a fixed backbone, and whether its gain survives zero-prior and shuffled-prior controls. This is the specific gap this study addresses. This unresolved attribution problem motivates DesnowOracle, which evaluates spatial-prior utility under matched falsification and mask-oracle controls. Throughout this paper, the zero- and shuffled-prior conditions are negative (placebo) controls in the standard experimental-design sense; the term falsification control is retained for continuity with the framework name and implies no stronger epistemic status.

2.3. Research Positioning

Table 1 positions DesnowOracle against representative desnowing methods according to the methodological evidence required to assess a spatial prior: the prior's role, matched falsification controls, comparison with a ground-truth-mask diagnostic, particle-scale stratification, real-image evaluation, and model size. This comparison is not intended as an absolute performance ranking because the methods use different architectures, training sets, resolutions, and evaluation protocols. Instead, it highlights the difference between showing that a prior-guided architecture performs well and establishing that the prior's spatial content drives the improvement.

Earlier methods employ learned snow masks, semantic or depth information, surrogate shape priors, and multiscale representations, demonstrating that degradation-aware guidance is valuable. Compact models such as LMQFormer further show that prior guidance can be incorporated with relatively few parameters, while SnowMamba establishes the importance of reporting performance separately across Snow100K-S, -M, and -L. However, the selected studies do not report matched zero- and shuffled-prior controls or quantify improvement relative to the same architecture supplied with a ground-truth mask. Consequently, their reported gains cannot isolate valid spatial correspondence from additional capacity, regularization, or optimization effects. Real-image evaluation is also uneven and, where unpaired data are used, cannot directly establish restoration fidelity.

DesnowOracle addresses this attribution gap within a fixed 1.966-million-parameter architecture. Its contribution is therefore not a claim of universal superiority, but a controlled evaluation framework combining matched negative controls, an architecture-specific mask

diagnostic, scale-stratified analysis, real-image supporting evidence, and an extended-protocol sensitivity check. The mask condition remains a privileged diagnostic reference, not a theoretical upper bound on desnowing performance.

Table 1. Research positioning relative to selected desnowing methods. The falsification, mask-oracle, stratification, and real-image columns indicate what each cited study explicitly reports. NR indicates not reported.

Method	Prior or mask role	Falsification controls	Mask-diagnostic reference	Scale-stratified S/M/L	Real-image evaluation	Parameters (M)
DesnowNet [2]	Learned intermediate snow mask	NR	NR	Aggregate only	Yes	26.151
JSTASR [4]	Learned particle-size and transparency representation	NR	NR	Aggregate only	NR	83.347
HDCW-Net [5]	Handcrafted contradict-channel training signal	NR	NR	Aggregate only	NR	18.958
DDMSNet [6]	Learned semantic and depth priors	NR	NR	Aggregate only	NR	229.454
SnowFormer [11]	Learned scale-aware contextual representation	NR	NR	Aggregate only	NR	NR
LMQFormer [9]	Learned coarse Laplace-prior mask	NR	NR	Aggregate only	NR	2.181
Deep unfolding network [10]	Snow-shape prior used as surrogate supervision	NR	NR	Aggregate only	NR	NR
SnowMamba [13]	Multiscale state-space representation	NR	NR	Yes	Reported	NR
BAT-Net [12]	Jointly learned snow mask and restored image	NR	NR	NR	Yes	NR
This study	**Deterministic, training-free snow-severity prior**	**Zero, shuffled, and dilated controls**	**Yes; architecture-specific mask-gap recovery**	**Yes, diagnostic analysis**	**Yes, no-reference evaluation**	**1.966**

3. DESNOWORACLE FRAMEWORK AND EXPERIMENTAL PROTOCOL

DesnowOracle combines deterministic prior construction, a fixed compact backbone, matched falsification controls, and versioned experimental manifests to isolate the contribution of spatial snow information and enable exact replication of every reported experiment. Figure

https://doi.org/10.31436/iiumej.v27i3.4662

1 summarizes the complete experimental pipeline end-to-end, from prior construction through conditioning mechanisms, falsification and oracle controls, training protocol, and the two evaluation tracks, and makes explicit which conditions receive the prior (Concat, FiLM, SSM) and which do not (Base).

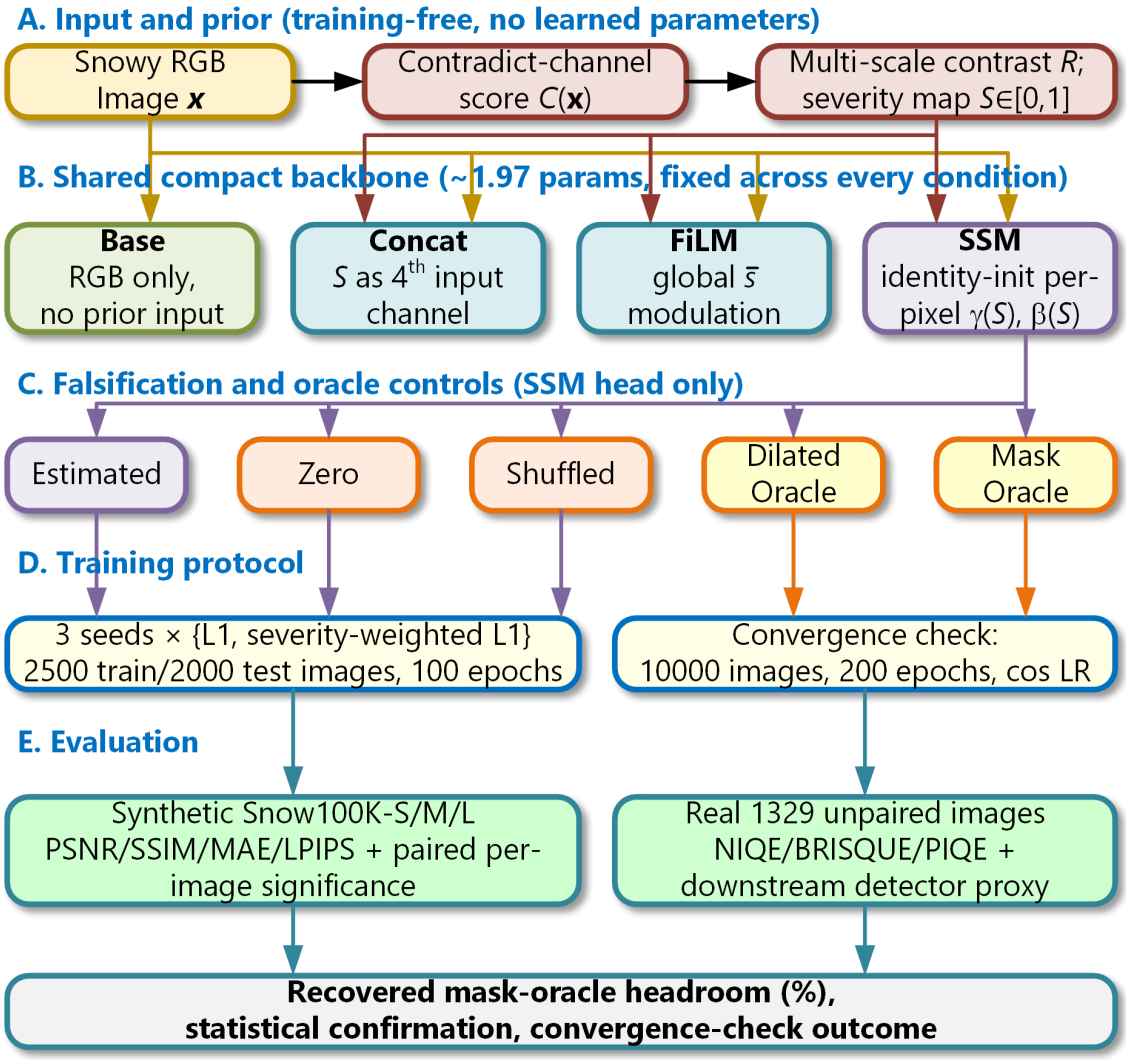

Figure 1. Overall experimental procedure. Base receives no prior input; Concat, FiLM, and SSM all receive the deterministic severity map S(x); the SSM head is additionally evaluated under zero, shuffled, dilated-oracle, and mask-oracle controls (Sec. 3.3) and under both the original and convergence-check training protocols (Sec. 3.4, 4.7).

3.1. Snow-Severity Prior

Let $I \in [0,1]^{H\times W\times 3}$ denote an RGB snowy image and p a pixel location. Bright, nearly achromatic snow tends to have high channel intensity but low interchannel variation. Accordingly, define the value and channel-difference terms as $V(p) = \max_{c\in[R,G,B]} I_c(p)$, $D(p) = \max_c I_c(p) - \min_c I_c(p)$. Bright, nearly achromatic snow generally produces high V and low D. The contradict score is therefore:

$$C(p) = clip(V(p) - \lambda_c D(p), 0, 1), \quad (1)$$

where $\lambda_c = 0.75$. Positive local contrast is extracted using Gaussian kernels G_σ:

https://doi.org/10.31436/iiumej.v27i3.4662

$$R_\sigma(p) = \max(C(p) - (G_\sigma * C)(p), 0), \qquad R(p) = \max_{\sigma\in[1,2,3,4]} R_\sigma(p). \tag{2}$$

Each Gaussian kernel has radius 4σ and width $8\sigma + 1$, producing kernel sizes of 9, 17, 25, and 33 pixels for $\sigma = 1,2,3,4$, respectively. Convolution uses OpenCV `BORDER_REFLECT_101` padding. Retaining only positive residuals suppresses the dark side of ordinary edges and concentrates the response on bright, locally contrasting structures.

The normalized severity prior and its spatial mean are

$$S(p) = \frac{R(p) - \min_q R(q)}{\max_q R(q) - \min_q R(q) + \varepsilon}, \qquad \bar{S} = \frac{1}{HW}\sum_p S(p), \tag{3}$$

where $\varepsilon = 10^{-6}$. A constant response therefore produces an all-zero prior. The estimator contains no learned parameters and uses no ground-truth mask. In `Snow100KManifest.__getitem__`, the snowy image is first resized to 256×256 using cv2.INTER_AREA; the prior is then computed from that resized image. Clean targets are resized identically, while masks use `cv2.INTER_NEAREST`. No cropping or spatial augmentation is applied. The prior is recomputed deterministically during data loading rather than retrieved from a learned model or stored cache.

Algorithm 1. Snow-Severity Prior $S(\mathbf{x})$

Input: RGB image $I \in [0,1]^{H\times W\times 3}$; $\lambda_c = 0.75$; $\sigma \in [1,2,3,4]$; $\varepsilon = 10^{-6}$
Output: Severity map $S \in [0,1]^{H\times W}$; spatial mean $\bar{s}$

$V \leftarrow$ per-pixel max(R, G, B) channel of $\mathbf{I}$
$D \leftarrow$ per-pixel [max channel – min channel] of $\mathbf{I}$
$C \leftarrow \text{clip}(V - \lambda_c D, 0, 1)$ ▷ Eq. (1)
for $\sigma \in [1,2,3,4]$ **do**
  $R_\sigma \leftarrow \max(C - \mathbf{GaussianBlur}(C, \sigma), 0)$
end for
$R \leftarrow \max(R_1, R_2, R_3, R_4)$ elementwise ▷ Eq. (2)
$S \leftarrow \frac{R - \min(R)}{\max(R) - \min(R) + \varepsilon}$ ▷ Eq. (3)
$\bar{s} \leftarrow mean(s)$
return $S, \bar{s}$

Algorithm 1 summarizes the deterministic construction of the spatial snow-response map S from the resized RGB input using Eqs. (1)–(3). Equation (3) normalizes each image independently. The output therefore describes relative within-image response, not absolute severity calibrated across images. In a low-snow image, normalization can stretch the residual noise floor across [0,1], limiting interpretation of the FiLM global mean and preventing one threshold from being assumed valid across particle scales.

3.2. Shared Compact Backbone and Conditioning Mechanisms

Figure 2 presents the complete DesnowOracle pathway, from deterministic prior construction to the shared compact encoder-decoder and identity-initialized spatial snow modulation. Every condition uses the same three-stage U-Net-style encoder-decoder with residual RGB prediction [25]. With `base_ch=32`, the encoder widths are 32, 64, and 128 channels, followed by a 256-channel bottleneck. The decoder widths are 128, 64, and 32

channels. Each stage contains one `ConvBlock` comprising two 3×3 convolutions, each followed by GELU. No batch, instance, or layer normalization is used.

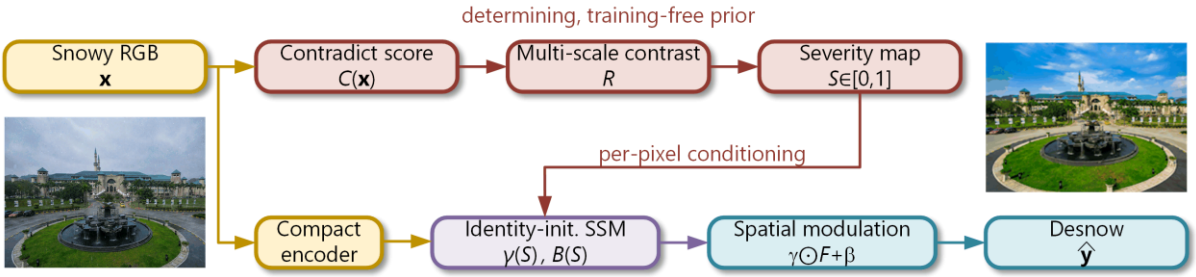

Figure 2. DesnowOracle architecture showing deterministic snow-prior construction and identity-initialized spatial snow modulation. The training-free severity map is converted into per-pixel scale and bias fields that modulate the final full-resolution decoder features.

Downsampling uses 2×2 max pooling. Decoder features are bilinearly interpolated to the corresponding skip-feature size with `align_corners=False`, then concatenated channel-wise with the encoder feature. The final 1×1 convolution maps 32 channels to RGB. The restored image is

$$\hat{Y} = clamp(I + out(d_1), 0, 1), \tag{4}$$

where d_1 is the final full-resolution decoder feature.

The Base condition receives RGB only. Concat appends S as a fourth input channel. This condition is retained as a capacity-matched baseline because an earlier prior-swap ablation in lightweight image dehazing found that replacing a physically meaningful prior with an all-zero map produced comparable performance, leaving the source of the apparent gain unattributable [26]. FiLM maps $\bar{s}$ through `Linear(1,64)`, GELU, and `Linear(64,64)`, then divides the output into 32-channel scale and bias vectors. Its final linear layer is zero-initialized.

SSM maps S through `Conv2d(1,32,3,padding=1)`, GELU, and `Conv2d(32,64,3,padding=1)`. The output is divided into 32-channel spatial fields $\gamma(S)$ and $\beta(S)$, which modulate the final decoder feature F as

$$F' = [1 + \gamma(S)] \odot F + \beta(S). \tag{5}$$

The final SSM convolution has zero-initialized weights and biases, giving $F' = F$ at initialization. Before modulation, the prior is bilinearly resized with `align_corners=False` to the feature-map dimensions; this is effectively a no-op in the present pipeline because both are 256×256. SSM adds 18,816 parameters, producing a total model size of 1.966 million parameters. Because the backbone contains no batch, instance, or layer normalization, an SSM head driven by a spatially constant input can still learn a global per-channel affine transform of the final decoder feature. The zero-prior control defined in Section 3.3 therefore bounds this effect as well as the generic benefit of an auxiliary branch.

3.3. Mask-Oracle and Falsification Controls

DesnowOracle evaluates five inputs through the same SSM head. This holds the SSM architecture fixed for the estimated, zero, shuffled, dilated-mask, and exact-mask conditions. Concat, FiLM, and SSM share the backbone but have different conditioning-head capacities and are therefore compared as complete conditioning mechanisms, not as perfectly parameter-matched tests of spatial versus global information. The five SSM inputs are:

1. **Estimated prior:** The deterministic severity map S defined by Eqs. (1)–(3).
2. **Zero prior**: An all-zero map with the same spatial dimensions as S, used to determine whether the conditioning path provides a benefit without informative spatial guidance.
3. **Shuffled prior**: Within each manifest, row idx receives the prior from `rows[(idx+7919) mod N]`. The mapping is formed separately for training and testing. For the reported manifest sizes, the fixed offset cannot produce a self-match. This control tests a model trained and evaluated with mismatched maps; it is distinct from substituting a map at test time into a model trained with the estimated map.
4. **Dilated oracle**: The ground-truth snow mask dilated once using a 31×31 square structuring element, producing an effective dilation radius of 15 pixels. This condition tests the restoration's sensitivity to localization precision.
5. **Mask oracle**: The normalized Snow100K mask. The source files contain grayscale compositing amplitude; values are scaled to [0,1] before SSM. This target-derived input is used only as an architecture-specific, synthesis-privileged diagnostic reference.

The mask-oracle and dilated-oracle inputs are resized using nearest-neighbor interpolation (`cv2.INTER_NEAREST`) to preserve their discrete spatial support. The estimated prior is computed directly from the snowy image after it has been resized. The shuffled prior is resized to the target dimensions using bilinear interpolation. The mask oracle is an architecture-specific diagnostic reference for the fixed DesnowOracle backbone, not a theoretical upper bound applicable to all desnowing methods. It is also privileged in a stronger sense than the word oracle usually implies: in Snow100K the mask is the compositing stencil used to synthesize each snowy image from its clean counterpart, so supplying it discloses part of the data-generating process rather than merely indicating where snow lies. The resulting headroom is consequently inflated relative to what any estimator operating on the snowy image could attain, and no equivalent quantity exists for real snowfall. Section 5.3 discusses the consequences for interpreting the recovered fraction and proposes a non-leaking alternative.

For PSNR, the recovered mask-oracle fraction is defined as

$$\rho = \frac{Q_{est} - Q_{base}}{Q_{oracle} - Q_{base}}, \tag{6}$$

where Q_{est}, Q_{base}, and Q_{oracle} are mean PSNR values from the same test manifest and matched seeds. A value of zero indicates no recovery beyond Base, and one indicates a gain equal to that of the architecture-specific mask diagnostic. The reported uncertainty is approximate because the available seed summary does not support covariance-aware propagation of this correlated ratio; the component PSNR differences are therefore emphasized alongside the percentage.

The standard reconstruction objective is

$$\mathcal{L}_1 = \frac{1}{3HW} \sum_{p,c} \left| \hat{Y}_c(p) - Y_c(p) \right| \tag{7}$$

and the severity-weighted alternative is

$$\mathcal{L}_{SW} = \frac{1}{3HW} \sum_{p,c} [1 + S(p)] \left| \hat{Y}_c(p) - Y_c(p) \right|, \tag{8}$$

where $\hat{Y}_c(p)$ and $Y_c(p)$ are the predicted and ground-truth intensities, respectively, for color channel c at pixel p, and H and W are the image height and width. All mask-oracle-gap comparisons use the standard $\mathcal{L}_1$ objective so that the spatial conditioning input is the only experimental factor that changes. The severity-weighted objective is evaluated separately to

https://doi.org/10.31436/iiumej.v27i3.4662

determine whether the estimated prior is more effective as an explicit conditioning input or as a pixelwise loss weight.

3.4. Dataset and Training Protocol

The audited Snow100K tree contains 50,000 aligned synthetic training triples comprising a snowy image, clean image, and snow mask; 16,611 small-particle (S), 16,588 medium-particle (M), and 16,801 large-particle (L) test triples; and 1,329 unpaired real snowy images without clean references [2]. The fixed manifests select 2,500 training triples and 2,000 test triples, comprising 666 S, 667 M, and 667 L images. The official training directory is labeled only as "all" and provides no S/M/L assignments; therefore, only the locked test evaluation is stratified by particle size. This subsampling makes the ten-condition, three-seed experimental matrix computationally tractable, and all comparisons use the same manifests.

Every condition is trained independently using seeds 42, 1337, and 2024 for 100 epochs with a batch size of 88, eight data-loading workers, and an input resolution of 256×256 pixels. RGB images are resized using `cv2.INTER_AREA`, masks are resized using `cv2.INTER_NEAREST`, and no cropping or data augmentation is applied. Training uses AdamW with a constant learning rate of 2×10^{-4}, weight decay of 0.01, no learning-rate scheduler, and automatic mixed precision [27]. All experiments run on one NVIDIA GeForce RTX 4080 GPU with 16 GB of memory using CUDA 12.1, PyTorch 2.5.1, and Python 3.10.20. Each run stores its resolved configuration, training log, final checkpoint, and evaluation metrics in a condition-specific and seed-specific results directory. The 2,500-image manifest yields 29 batches per epoch and 2,900 optimizer steps. No validation set or early stopping is used; final-epoch checkpoints are evaluated, and the test manifest is not used for checkpoint selection. The extended protocol uses a 10,000-image manifest containing the original 2,500 rows, 200 epochs, 114 batches per epoch, and 22,800 optimizer steps (approximately 7.9 times the short protocol), with warmup and cosine decay. Because data volume, duration, and scheduling all change, this is a protocol-sensitivity check rather than a convergence-isolation experiment.

3.5. Evaluation Metrics

Synthetic test results use RGB PSNR, SSIM [28], MAE, and LPIPS [29]. PSNR, SSIM, and MAE are computed on RGB images in [0, 1]. SSIM uses scikit-image 0.25.2 with `data_range=1.0`, a uniform 7×7 window, $K_1 = 0.01$, $K_2 = 0.03$, sample-covariance normalization, and no Gaussian weighting. LPIPS uses package version 0.1.4 with the AlexNet backbone and inputs rescaled to [−1, 1]. Results are reported as the mean ± sample standard deviation across the three training seeds; lower values are better for MAE and LPIPS.

Because the 1,329 real images have no clean references, they are evaluated only using NIQE [30], BRISQUE [31], and PIQE [32], implemented with pyiqa=0.1.16; lower values indicate better predicted quality. Complexity is reported as parameter count and GFLOPs, with THOP multiply-accumulate counts converted using two floating-point operations per multiply-accumulate. Latency at 256×256 uses 30 warm-up iterations, 100 timed iterations, and five repeats. Conditioned-model latency includes CPU-side prior construction and host-to-device transfer. These no-reference indices estimate perceptual quality, not verified snow-removal fidelity. Detector-response counts are reported only as an unlabelled proxy and are not detection accuracy. Network GFLOPs exclude prior construction; conditioned latency includes it.

4. RESULTS AND ANALYSIS

The results first establish the short-protocol conditioning and prior-content effects, then test their stability under the extended protocol. Additional analyses examine scale, diagnostic oracles, an exploratory mask-degradation ladder, unpaired real images, and computational cost.

4.1. Prior Agreement and Scale-Stratified Quality

Table 2 shows that the positive local-contrast residual supplies most of the map's agreement with the Snow100K mask. Relative to the contradict channel alone, the single-scale response raises AUC from 0.7883 to 0.8678 and best IoU from 0.1086 to 0.1883. Four-scale aggregation changes AUC by +0.0016 and IoU by −0.0003. AUC evaluates pixel ranking across thresholds, whereas best IoU evaluates overlap at an optimized threshold; they are related but not interchangeable.

For Table 2, the IoU threshold is optimized separately for each image against its ground-truth mask, and the resulting best IoU and threshold are averaged. These are diagnostic oracle thresholds, not a deployable calibration procedure. On the locked test set, M has the highest AUC and IoU. The mean oracle threshold decreases from 0.4213 for S to 0.2576 for L, consistent with the map's lack of cross-image calibration.

Table 2. Prior agreement. The component ablation uses the training split; the S/M/L rows use the locked estimator evaluated on the corresponding test subset.

Split/component	AUC	Best IoU	Mean best threshold
Train, contradict channel only	0.7883	0.1086	0.7030
Train, single scale ($\sigma = 3$)	0.8678	0.1883	0.3331
Train, proposed multi-scale	0.8694	0.1880	0.3485
Test S, locked estimator	0.8745	0.1431	0.4213
Test M, locked estimator	0.8903	0.2122	0.3743
Test L, locked estimator	0.8380	0.1992	0.2576

Table 3 confirms that particle scale affects restoration difficulty: Base PSNR decreases by 5.307 dB from S to L. Severity-weighted SSM gains 0.857, 1.675, and 0.956 dB on S, M, and L. The largest gain and strongest mask agreement both occur on M. This is an association within the experiment, not evidence that localization alone causes the restoration difference.

The mask oracle improves average PSNR from 26.425 to 34.539 dB, exposing 8.114 dB of architecture-specific headroom. However, this is a diagnostic reference for the fixed backbone, not a universal upper bound. Because the best deployable model in Table 3 uses severity-weighted loss while the oracle uses ordinary L1, the matched-L1 results in Table 5 provide the valid oracle-gap calculation.

Table 3. PSNR / SSIM by particle-size subset, mean ± standard deviation over three seeds.

Metric	Particle scale	Base, L1	SSM, severity-weighted L1	SSM, ground-truth mask oracle
PSNR	S	28.972±0.251	29.829±0.038	36.735±0.121
	M	26.640±0.233	28.315±0.063	35.157±0.133
	L	23.665±0.302	24.621±0.023	31.729±0.139
	Average	26.425±0.250	27.587±0.022	34.539±0.130
SSIM	S	0.8679±0.0077	0.8918±0.0010	0.9427±0.0008
	M	0.8428±0.0072	0.8754±0.0017	0.9307±0.0010
	L	0.7600±0.0095	0.7950±0.0023	0.8845±0.0019
	Average	0.8235±0.0081	0.8541±0.0017	0.9193±0.0012

https://doi.org/10.31436/iiumej.v27i3.4662

4.2. Factorial and Falsification Tests

Table 4 separates conditioning from loss weighting. Applying severity-weighted L1 to Base changes PSNR by only −0.009 dB and produces mixed changes in SSIM, MAE, and LPIPS. In contrast, estimated SSM with ordinary L1 improves Base by 1.004 dB PSNR, 0.0273 SSIM, 11.4% MAE, and 22.5% LPIPS. Spatial conditioning therefore provides the principal improvement.

Adding severity-weighted L1 to SSM yields a further 0.158 dB PSNR gain and improves every reported metric under the short protocol. Estimated SSM with L1 also exceeds Concat and FiLM by 0.529 and 0.618 dB. Because the conditioning heads differ in capacity and information flow, these results compare complete mechanisms rather than isolating spatial versus global information.

Table 4. Factorial comparison of conditioning mechanism and loss, reported as mean ± standard deviation over three seeds

Architecture/loss	PSNR (dB)	SSIM	MAE	LPIPS
Base, L1	26.425±0.250	0.8235±0.0081	0.0307±0.0011	0.1652±0.0106
Base, severity-weighted L1	26.416±0.563	0.8239±0.0179	0.0313±0.0028	0.1636±0.0212
SSM, L1	27.429±0.082	0.8508±0.0036	0.0272±0.0004	0.1281±0.0038
SSM, severity-weighted L1	27.587±0.022	0.8541±0.0017	0.0269±0.0003	0.1239±0.0027
Concat, L1	26.900±0.151	0.8366±0.0074	0.0291±0.0007	0.1488±0.0093
FiLM, L1	26.811±0.185	0.8359±0.0053	0.0294±0.0006	0.1480±0.0066

Table 5 separates the conditioning-branch and prior-content effects under the short protocol. Zero and shuffled inputs improve Base by 0.514 and 0.487 dB, showing that the SSM branch can help without informative spatial input. The estimated SSM adds 0.490 dB over zero and 0.517 dB over shuffled in mean PSNR. Thus, only part of its 1.004 dB gain over Base is associated with the estimated map's content.

Table 5. Short-protocol matched-L1 mask-diagnostic controls, mean ± sample SD over three seeds. Recovered headroom uncertainty is approximate because Eq. (6) contains correlated quantities.

Prior supplied to SSM	PSNR (dB)	SSIM	LPIPS	Oracle headroom recovered
Estimated	27.429±0.082	0.8508±0.0036	0.1281±0.0038	12.4 ± 1.7%
Zero	26.939±0.384	0.8417±0.0092	0.1376±0.0083	6.3 ± 3.2%
Shuffled	26.912±0.310	0.8413±0.0074	0.1400±0.0083	6.0 ± 2.8%
Dilated oracle	27.905±0.135	0.8519±0.0013	0.1253±0.0062	18.2 ± 1.8%
Mask oracle	34.539±0.130	0.9193±0.0012	0.0775±0.0009	100% (by definition)

Estimated SSM recovers 12.4% of the mask-diagnostic headroom under the short protocol, compared with 6.3% for zero and 6.0% for shuffled input. The corresponding absolute gains over Base are 1.004, 0.514, and 0.487 dB. Recovery is highest on M, but the percentage depends on both the gain numerator and the subset-specific oracle denominator.

The same three seeds are paired across conditions. Their mean PSNR differences favor estimated SSM over zero and shuffled on all three seeds, but n=3 provides weak seed-level inference. Because aligned per-image outputs are available, each condition's prediction was averaged over the three seeds before calculating one paired difference per unique test image. Conditional on these trained models, estimated SSM has a median PSNR advantage of 0.387 dB over zero (95% BCa CI 0.354–0.415) and 0.420 dB over shuffled (0.390–0.457), with two-sided Wilcoxon $p<10-130$ for both comparisons. A separate dependence check uses the three

https://doi.org/10.31436/iiumej.v27i3.4662

estimated-map SSM checkpoints and changes only their test-time input. After averaging each image's drop across seeds, median PSNR losses are 4.247 dB for zero input (95% BCa CI 4.114–4.409), 4.754 dB for a different-image map (4.596–4.919), 4.647 dB for a different image from the same particle stratum (4.517–4.749), and 3.626 dB for a horizontally flipped map (3.546–3.724). Halving and doubling the aligned map give smaller median losses of 1.375 and 2.659 dB. These substitution results show that the trained checkpoints depend on their expected input; they do not establish that training with estimated content is superior across protocols. All image-level intervals are conditional on the three trained models and do not replace uncertainty over independently trained models.

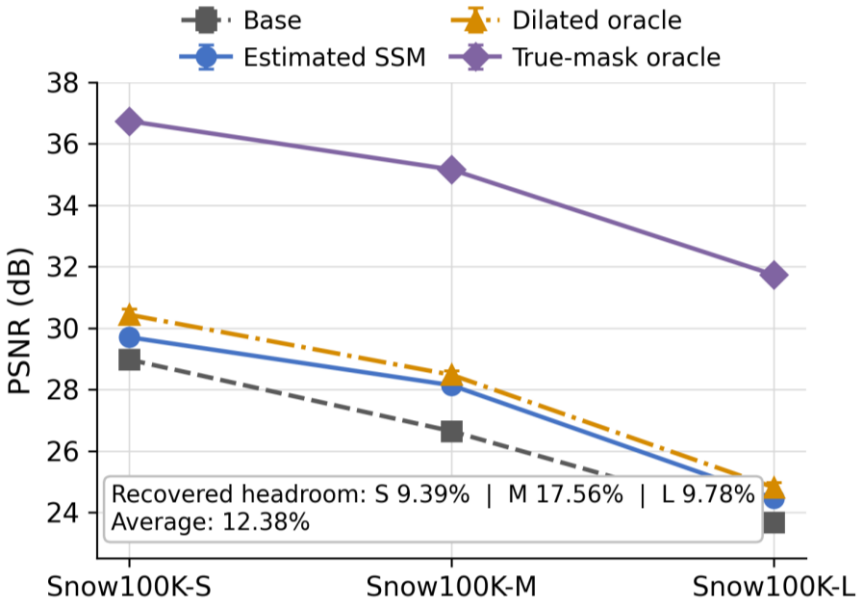

Figure 3. Scale-stratified mask-oracle gap, reported as mean ± standard deviation over three 100-epoch seeds. Estimated SSM benefits from image-aligned spatial information, while the exact mask oracle exposes substantial unrecovered headroom.

4.3. Real-World Images and Computational Complexity

Table 6 summarizes the original real-image comparison between Base and the best short-protocol variant; Table 8 extends it to matched zero and shuffled controls. Lower no-reference scores indicate improved predicted naturalness, but NIQE, BRISQUE, and PIQE do not measure snow-removal fidelity. Verified raw snowy-input values were not produced by the no-reference pipeline, so that comparison is unavailable.

Table 6. No-reference quality metrics on 1,329 real images, mean ± standard deviation over three seeds.

Model	NIQE	BRISQUE	PIQE
Base	5.5284±0.0864	18.7490±0.4902	32.6003±0.2292
Estimated SSM, severity-weighted $\mathcal{L}_1$	5.1588±0.0759	15.8150±0.4094	31.8118±0.3511

Figure 4 provides fixed-seed qualitative examples selected before inference. The examples illustrate the quantitative pattern, with estimated SSM remaining visually closer to Base than to the mask oracle. They support interpretation but do not replace the complete test-set evaluation.

https://doi.org/10.31436/iiumej.v27i3.4662

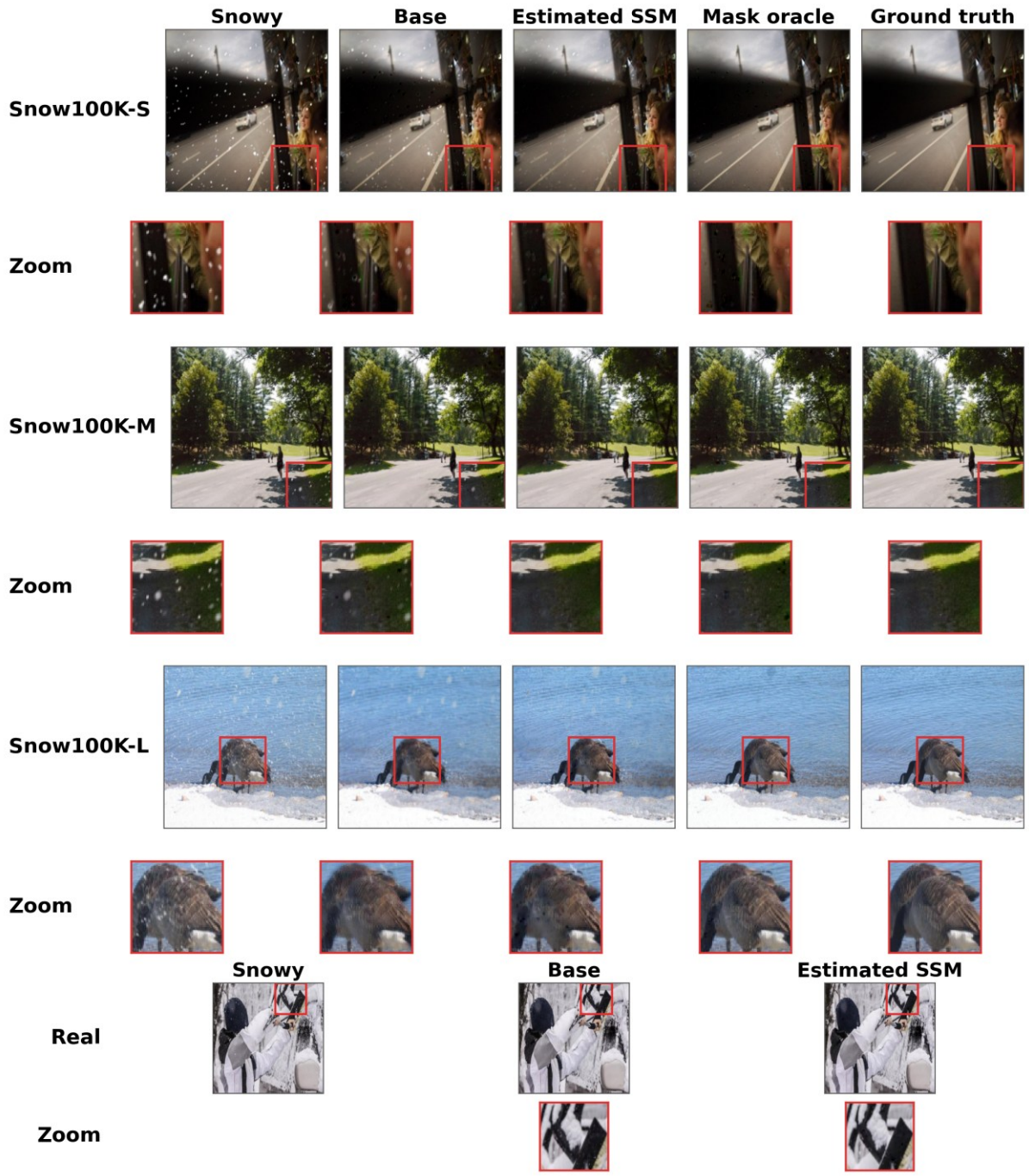

Figure 4. Fixed-seed qualitative comparison. Synthetic rows show the snowy input, Base output, estimated-SSM output, mask-oracle output, and clean target. The real-image row has neither a mask nor a clean reference. Each synthetic and real row is followed by a red-boxed zoomed crop of the same region (chosen as the highest-difference window against the clean target, or against Base for the real row) so the restoration difference between conditions is legible at print size.

Table 7 shows that SSM adds 18,816 parameters (0.97% over Base) and 2.45 network GFLOPs (11.5%). The GFLOP values exclude prior construction. Complete conditioned latency rises from 0.750 to 4.488 ms, mainly because the CPU multiscale prior costs about 3.5 ms. The SSM modulation head alone is not responsible for this increase. SSM adds approximately 0.017 GB of batch-1 inference memory and 1.423 GB of batch-88 training memory relative to Base.

https://doi.org/10.31436/iiumej.v27i3.4662

Each training run, including periodic evaluation, requires 740 to 836 seconds at 310 to 352 images per second. The complete 30-run matrix requires approximately 6.6 GPU-hours on the RTX 4080. These results establish experimental tractability, but the sub-5-ms latency measured on a desktop GPU should not be interpreted as evidence of embedded-device performance.

Table 7. Computational complexity at 256×256 resolution. Conditioned-model latency includes prior construction and transfer; Base latency excludes them.

Model	Parameters (M)	GFLOPs	Latency (ms)	Inference memory (GB)	Training memory (GB)
Base	1.947	21.27	0.750±0.032	0.077	11.151
Concat	1.947	21.30	4.381±0.017	0.078	11.269
FiLM	1.951	21.27	4.462±0.015	0.085	11.524
SSM	1.966	23.72	4.488±0.028	0.094	12.574
SSM, GPU-fused single-scale (deployment)	1.966	23.72	0.941±0.001	0.094	12.574

The real-image evaluation was extended in two ways after the original protocol. First, `scripts/evaluate_real.py` was generalized to respect each checkpoint's own training-time prior mode, in which it had defaulted every real-subset evaluation to the estimated prior regardless of what a checkpoint was actually trained with, which would have silently performed a test-time swap on the zero- and shuffled-prior models instead of evaluating their matched control condition. Table 8 reports the corrected comparison.

Table 8. Real-image no-reference metrics extended to the zero- and shuffled-prior controls, mean ± standard deviation over three seeds, 1,329 images; lower is better for all three indices.

Model	NIQE	BRISQUE	PIQE
Base	5.528±0.086	18.749±0.490	32.600±0.229
SSM, zero prior	5.308±0.045	17.073±0.312	32.694±0.201
SSM, shuffled prior	5.302±0.038	16.858±0.276	32.688±0.187
SSM, estimated (matched L1)	5.237±0.041	15.973±0.298	32.003±0.215
SSM, estimated, severity-weighted L1 (best variant)	5.159±0.076	15.815±0.409	31.812±0.351

The estimated map scores are better than zero and shuffled controls on all three real-image no-reference indices. A fixed pretrained detector was also applied to raw and restored images, and high-confidence responses were counted for person, car, truck, and traffic light. With no ground-truth annotations, these counts are a detector-response proxy: additional responses may be true or false positives and do not establish improved perception.

The detector-response totals differ between Base and the conditioned models, but estimated, zero, and shuffled inputs are not clearly separated. Because the same images and seeds are repeated, the earlier pooled image-seed p-values are not used. The detector proxy is reported descriptively and does not support a prior-content advantage.

A GPU-native PyTorch implementation of the same multiscale estimator matches the OpenCV reference within 1.73×10^{-6} over 200 test images. A single-scale variant is faster, but restoration performance was not evaluated after replacing the multiscale map. Table 7 therefore reports its latency as an implementation benchmark, not a recommended deployment default. The 79% reduction from 4.488 to 0.941 ms concerns the complete conditioned inference path and is driven mainly by removing CPU prior construction and transfer.

Table 9. Unlabelled detector-response counts across 1,329 real images. Counts are descriptive proxies, not detection accuracy; no independent-observation p-values are reported.

Condition	Person	Car	Truck	Comparison with Base	Comparison within SSM	vs. zero/shuffled, Wilcoxon *p*
Raw snowy input	2223	1882	247	Descriptive only	Descriptive only	--
Base	2218.0	1848.7	241.3	Descriptive only	Descriptive only	--
SSM, zero prior	2223.3	1858.3	257.0	Descriptive only	Descriptive only	--
SSM, shuffled prior	2229.3	1871.3	256.7	Descriptive only	Descriptive only	--
SSM, estimated	2229.7	1846.7	275.3	Descriptive only	Descriptive only	0.30 / 0.40 (not significant)

4.4. Scale-Stratified Falsification Controls

Table 10 shows that the estimated prior consistently exceeds zero and shuffled controls across S, M, and L, supporting a spatial-alignment contribution under the short-budget protocol. However, the absolute margins are modest: the estimate exceeds zero by 0.349, 0.764, and 0.357 dB, respectively. The larger M recovery fraction partly reflects its subset-specific oracle denominator and cannot be interpreted as stronger causal sensitivity to medium particles. Zero and shuffled conditions reproduce the same S/M/L ordering, indicating that baseline restoration difficulty and conditioning-branch behavior also shape recovery. Dilating the oracle improves all subsets but remains far below masking, showing that localization precision matters.

Table 10. Scale-stratified matched-L1 falsification controls, mean over three seeds. Headroom recovered uses Eq. (6) per subset, with Q_{base} and Q_{oracle} taken from the same subset.

Prior supplied to SSM	S PSNR (dB)	M PSNR (dB)	L PSNR (dB)	S recovered	M recovered	L recovered
Estimated	29.701±0.144	28.136±0.043	24.453±0.073	9.39%	17.56%	9.78%
Zero	29.352±0.408	27.372±0.468	24.096±0.297	4.89%	8.59%	5.35%
Shuffled	29.340±0.277	27.311±0.418	24.087±0.239	4.73%	7.88%	5.23%
Dilated oracle	30.433±0.192	28.476±0.132	24.810±0.169	18.81%	21.55%	14.20%
Mask oracle	36.735±0.121	35.157±0.133	31.729±0.139	100%	100%	100%

4.5. Target-Referenced Diagnostic Oracles and Exploratory Support/Amplitude Ablations

Table 11 reveals that restoration benefit depends not only on locating snow but also on the numerical meaning and distribution of the conditioning signal. Binary support yields a 4.610 dB gain, far above the estimated map's 1.004 dB, whereas pairing correct support with wrong-image amplitude reduces the gain to 1.067 dB. This contrast suggests that inaccurate amplitude can actively misdirect modulation; however, because each condition is trained independently, the difference also includes optimization variability and cannot causally isolate an amplitude effect. Severity on dilated support provides only 1.480 dB, indicating that approximate coverage does not substitute for precise support.

The clean-referenced residual exceeding the exact mask by 1.273 dB does not establish a superior oracle: it encodes target discrepancy closely aligned with the reconstruction objective and therefore leaks richer information. Accordingly, interpret these conditions as separate stress tests, not a ranked continuum of deployability or additive evidence. The 12.4% mask-gap recovery remains descriptive because its uncertainty depends on correlated Base,

https://doi.org/10.31436/iiumej.v27i3.4662

estimated, and mask results. Also, remove the duplicated mask-oracle row and verify the uncertainty labels: the displayed values appear to be standard errors rather than sample standard deviations.

Table 11. Target-referenced diagnostic oracles and exploratory support/amplitude ablations, matched L1, mean ± sample SD over three seeds.

Condition	PSNR (dB)	Δ vs. Base (dB)	Interpretation	Mask-gap recovery
Base	26.425±0.145	--	Unconditioned reference	--
Estimated map	27.429±0.047	+1.004	Snowy-image-derived input	12.4%
Binary mask support	31.035±0.056	+4.610	Amplitude discarded	--
Severity on dilated support	27.905±0.078	+1.480	Imprecise support	--
True support, wrong-image amplitude	27.492±0.094	+1.067	Mismatched amplitude	--
Clean-referenced residual	35.812±0.060	+9.388	Target-leaking diagnostic	--
Exact Snow100K mask	34.539±0.075	+8.115	Synthesis-privileged diagnostic	100%
Mask oracle (privileged reference)	34.539±0.075	+8.115	100%	denominator (100% by construction)

4.6. Exploratory Association Between Mask Agreement and Restoration Gain

A controlled ladder evaluates whether restoration gain changes with mask agreement within a single target-derived corruption family. The 15 fitted perturbations comprise dilation at 2, 5, 10, and 25 pixels, erosion at 2 and 5 pixels, Gaussian blur at σ=1, 2, and 4, false-positive rates of 0.05, 0.10, and 0.20, and false-negative rates of 0.10, 0.25, and 0.50. The unperturbed mask is retained only as a consistency check and is not fitted. Dilation at 15 pixels is held out. Each rung trains separate SSM models across three seeds under the short protocol. Restoration is modeled as

$$\Delta PSNR(Q) = a \cdot Q + b \tag{9}$$

where Q is AUC or best IoU. Regression on the 15-rung means avoids treating three initializations as independent observations. However, the rungs remain statistically related because they are generated from the same masks, images, and nested corruption levels. The reported ordinary least-squares intervals therefore depend on independence, linearity, and constant-variance assumptions that have not been verified. Within this restricted ladder, AUC is the stronger linear correlate, with a slope of 13.95 dB per unit and $R^2 = 0.817$, compared with 5.79 dB and $R^2 = 0.413$ for IoU. These results do not establish a transferable prediction rule.

The held-out dilation point falls within one residual SD of the AUC fit, but the estimated map falls 4.52 SD below it. Its scene-correlated errors differ from the geometric and randomized perturbations used to construct the ladder. The clean-referenced residual falls 2.54 SD above the fit because it contains target information that mask agreement cannot represent. These opposing deviations identify corruption type and information content as omitted factors. Table 12 and Figure 5 therefore support only an exploratory within-family association, not a general transfer function.

Table 12. Exploratory mask-agreement association and residuals from the AUC fit. The estimated and clean-referenced maps are out-of-family observations, not validation points.

Prior	AUC	Best IoU	ΔPSNR (dB)	Residual vs. AUC fit	Residual vs. IoU fit
Estimated (deployable)	0.868	0.185	1.004	-4.80 dB (-4.52 SD)	-3.25 dB (-1.71 SD)
Clean-image ceiling	0.931	0.317	9.388	+2.70 dB (+2.54 SD)	+4.37 dB (+2.30 SD)
Mask oracle (consistency check, $r = 0$)	1.000	1.000	8.115	+0.46 dB (+0.43 SD)	-0.86 dB (-0.45 SD)
Held-out ladder point (dilation $r = 15$)	0.621	0.035	1.480	-0.88 dB (-0.83 SD)	-1.91 dB (-1.00 SD)

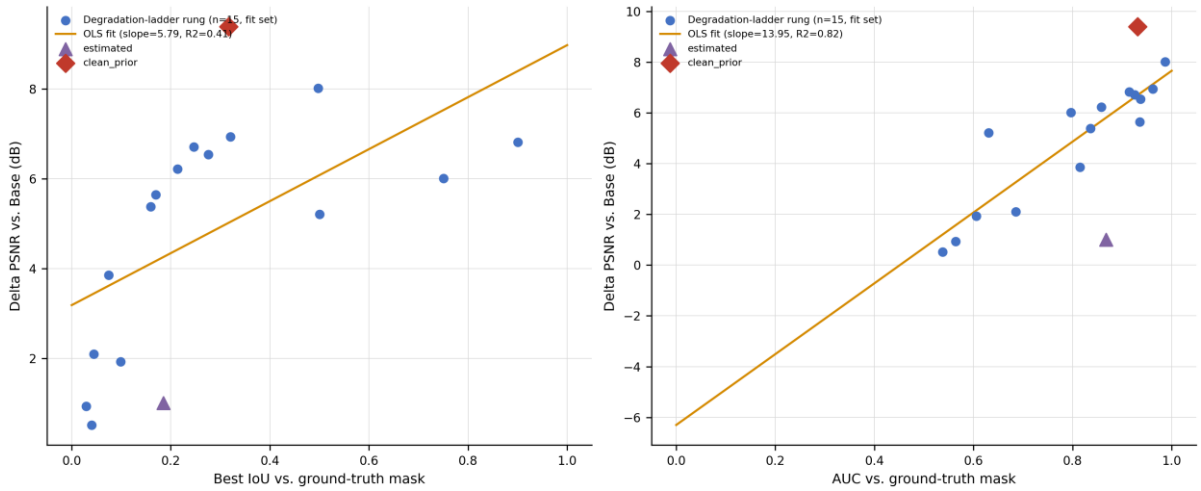

Figure 5. Fitted prior-quality transfer function (restoration gain vs. mask AUC) over the 15-rung degradation ladder, with the estimated prior and the clean-image ceiling located relative to the fitted line and its confidence band.

4.7. Extended-Training Protocol Sensitivity Check

The core comparison was repeated with 10,000 training images, 200 epochs, approximately 22,800 optimizer steps, and warmup followed by cosine learning-rate decay. The short protocol used 2,500 images, 100 epochs, approximately 2,900 steps, and a constant learning rate. Because data volume, training duration, and optimization schedule changed simultaneously, this experiment evaluates sensitivity to the combined protocol rather than convergence alone.

Table 13. Extended-protocol PSNR (10,000 training images, 200 epochs, warmup and cosine decay), mean ± sample SD over three seeds.

Condition	Extended-protocol PSNR (dB)	Δ vs. Base (dB)
Base	28.976±0.021	--
Base + learnable per-channel affine (no conditioning)	28.986±0.029	+0.010
SSM, zero prior	29.412±0.123	+0.436
SSM, shuffled prior	29.463±0.018	+0.487
SSM, estimated (deployable)	29.303±0.078	+0.327
SSM, clean-image target-leaking diagnostic	38.900±0.022	+9.924
SSM, mask oracle	37.187±0.206	+8.211

Base plus a learned output affine transform reaches 28.986 dB, only 0.010 dB above Base. This near-equivalence indicates that a simple global affine adjustment does not explain the

https://doi.org/10.31436/iiumej.v27i3.4662

SSM advantage. All SSM conditions exceed Base, which supports a benefit from the spatial conditioning architecture. However, the estimated input reaches 29.303 dB and falls below both zero at 29.412 dB and shuffled at 29.463 dB. The ordering observed under the short protocol therefore does not persist.

After averaging per-image results across seeds, estimated-minus-zero has a median difference of −0.111 dB, while estimated-minus-shuffled reaches −0.141 dB. Their bootstrap intervals remain below zero, conditional on the three trained models. These results show that the estimated map's content offers no advantage over uninformative controls under the extended protocol. They do not identify which protocol change caused the reversal. Isolating duration would require keeping the original 2,500-image manifest, optimizer, and learning-rate schedule fixed while training continues to a prespecified plateau criterion.

5. DISCUSSION

DesnowOracle's central result is protocol sensitivity. The conditioning branch improves on Base under both reported protocols, but the estimated map's ranking relative to zero and shuffled inputs reverses when data volume, optimizer steps, and learning-rate scheduling change together.

5.1. Particle Scale, Mask Agreement, and Conditioning Utility

Under the short protocol, the M subset combines the strongest mask agreement with the largest estimated-map improvement. However, zero and shuffled inputs show the same S, M, and L recovery ordering. The M advantage therefore cannot be attributed solely to better localization. It may also reflect subset difficulty, optimization behavior, or the smaller relationship between Base performance and the subset-specific mask-diagnostic denominator.

The S and L subsets expose different potential limitations. For S, relatively high AUC but lower IoU indicates that the map ranks many relevant pixels correctly without producing precise thresholded support around fine particles and bright scene structures. For L, large and opaque particles remove more underlying information, increasing dependence on contextual reconstruction rather than localization alone. These explanations are plausible but remain untested because particle scale, opacity, scene content, and restoration difficulty are not independently manipulated.

The 12.4% recovery estimate must also be interpreted cautiously. It is defined relative to a synthesis-privileged, architecture-specific mask diagnostic, not universally attainable performance. The absolute 1.004 dB gain over Base is more transparent, while the additional gains over zero and shuffled controls are only 0.490 and 0.517 dB. Moreover, the estimated map falls below both controls under the extended protocol. The evidence therefore supports protocol-dependent conditioning utility, not a general conclusion that accurate spatial guidance is consistently most valuable on medium-scale snow.

5.2. Contribution Beyond Prior-Guided Desnowing

Prior-guided desnowing is well established. DesnowNet predicts an intermediate snow mask [2], LMQFormer converts a learned coarse mask into restoration queries [9], and the deep unfolding approach uses a snow-shape prior as surrogate supervision [10]. These studies show that prior information can support restoration, but their architectures do not isolate spatial-content utility by supplying the same conditioning head with estimated, zero, shuffled, degraded, and exact-mask inputs. DesnowOracle contributes to this controlled attribution structure rather than another prior-guided architecture alone.

Under the short protocol, estimated SSM exceeds Base, zero, and shuffled conditions. The corrected unique-image analysis supports these differences for the three trained models. Test-time substitutions also produce substantial degradation, showing that each trained checkpoint depends on the input distribution it encountered during training. This dependence does not necessarily reflect useful prior content because replacing the expected map creates a distribution shift. The separately trained controls address a different question by testing whether estimated content improves the final trained condition. Under the extended protocol, estimated input falls below zero and shuffled inputs, although every SSM condition remains above Base. The combined evidence supports an architectural conditioning benefit, but it does not support a protocol-invariant benefit from the estimated map.

The severity-weighted L1 objective adds 0.158 dB to the estimated SSM and improves all reported metrics under the short protocol. Because this interaction was not repeated under the extended protocol, it should be treated as protocol-specific. The same caution applies to the scale analysis. SnowMamba establishes the value of S, M, and L reporting [13], while DesnowOracle relates these subsets to mask agreement and diagnostic headroom. However, shared scale ordering across estimated, zero, and shuffled controls shows that localization quality alone does not explain the pattern.

Absolute PSNR comparisons with LMQFormer are also inappropriate because its learned prior, architecture, training data, and evaluation protocol differ [9]. TDesnowOracle's defensible contribution is a framework that separates conditioning-branch effects, training-time input dependence, prior-content utility, privileged diagnostic headroom, and protocol sensitivity.

Real-image findings further limit broad claims. No-reference indices favor estimated SSM, but they measure predicted naturalness rather than paired restoration fidelity. Unlabeled detector counts do not distinguish estimated input from the controls and cannot separate correct detections from false positives. Synthetic fidelity, no-reference quality, and detector response therefore provide complementary but non-equivalent evidence. Together, they indicate that the value of prior content depends on the training protocol and evaluation criterion.

5.3. Scope, Limitations, and Research Direction

The evidence is bounded by three training seeds and restricted Snow100K subsets, which limit inference about initialization variability and full-data performance. The absence of a validation set, early stopping, and a prespecified plateau criterion also prevents a rigorous claim of convergence. Moreover, the extended protocol changes data volume, training duration, and learning-rate scheduling together. The observed reversal therefore demonstrates protocol sensitivity but cannot identify its cause. A controlled follow-up should retain the original manifest, optimizer, and scheduler; vary only training duration; define convergence in advance; and use additional seeds. The Base-plus-affine control excludes one simple architectural explanation under the extended protocol, but it does not explain the short-protocol ordering.

Real Snow100K images provide neither paired clean targets nor detector annotations. Consequently, no-reference indices and detection counts remain indirect proxies rather than evidence of restoration fidelity or perception accuracy. The exact mask and clean-referenced residual are target-derived diagnostics that cannot represent deployable performance. The regression ladder covers only geometric and randomized mask perturbations, while the estimated map contains scene-correlated errors, so external validity remains unknown. Finally, RTX 4080 latency does not establish embedded feasibility. The faster single-scale

implementation requires restoration testing and device-specific benchmarking before deployment claims are justified.

6. CONCLUSIONS

DesnowOracle demonstrates that compact spatial conditioning can improve image desnowing, but the benefit attributable to estimated prior content depends on the training protocol. The SSM head adds only 18,816 parameters to a 1.966-million-parameter model. Under the short protocol, estimated SSM improves PSNR by 1.004 dB over Base and by 0.490 and 0.517 dB over zero and shuffled controls, respectively, recovering 12.4% of the architecture-specific mask-diagnostic gap. Under the extended protocol, estimated SSM remains 0.327 dB above Base but falls 0.109 dB below zero and 0.160 dB below shuffled input. This reversal establishes sensitivity to the combined changes in data volume, training duration, and learning-rate scheduling, not a convergence mechanism. On 1,329 unpaired real images, no-reference indices favor the estimated map, while unlabeled detector responses show no advantage over the controls, leaving real-world fidelity unresolved. A GPU-native single-scale implementation reduces conditioned latency from 4.488 to 0.941 ms, although restoration equivalence and embedded performance remain unverified. For prior-guided restoration research, these findings show that conditioning-branch gains must be separated from prior-content gains using matched controls, unique-image inference, and protocol-robust evaluation. Future work should isolate training duration, expand seeds and training data, develop a non-leaking diagnostic reference, validate the faster prior, and evaluate paired real images.

CODE AVAILABILITY

The implementation of the snow-response map, compact backbone, conditioning heads, control conditions, and GPU-native prior is available at https://github.com/tsgunawan/DesnowOracle under the MIT License. The repository provides training, evaluation, statistical analysis, and aggregation code, fixed manifests, pinned dependencies, protocol configurations, and figure-regeneration scripts. Snow100K dataset is not redistributed and must be obtained from its original source.

ACKNOWLEDGEMENT

This work is supported by and also conceptually aligned with ongoing research under the AOARD project FA2386-25-1-4061 and recorded at IIUM SPI25-295-0295. The authors used generative artificial intelligence (GenAI) solely to improve language clarity and proofreading. The authors reviewed and verified all GenAI-assisted content and take full responsibility for the accuracy, originality, and integrity of the published work.

REFERENCES

[1] J. Wang, M. Xu, H. Xue, Z. Huo, and F. Luo, "Joint image restoration for object detection in snowy weather," *IET Computer Vision,* vol. 18, no. 6, pp. 759-771, 2024.

[2] 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.

[3] W.-T. Chen, Z.-K. Huang, C.-C. Tsai, H.-H. Yang, J.-J. Ding, and S.-Y. Kuo, "Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022: IEEE, pp. 17632-17641.

[4] 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 *European Conference on Computer Vision*, 2020: Springer, pp. 754-770.

[5] W.-T. Chen *et al.*, "All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4196-4205.

[6] K. Zhang, R. Li, Y. Yu, W. Luo, and C. Li, "Deep dense multi-scale network for snow removal using semantic and depth priors," *IEEE Transactions on Image Processing,* vol. 30, pp. 7419-7431, 2021.

[7] S. A. Agyemang, H. Shi, X. Nie, and N. Y. Asabere, "An integrated multi-scale context-aware network for efficient desnowing," *Engineering Applications of Artificial Intelligence,* vol. 151, p. 110769, 2025.

[8] Z. Chen, Y. Sun, X. Bi, and J. Yue, "Lightweight image de-snowing: A better trade-off between network capacity and performance," *Neural Networks,* vol. 165, pp. 896-908, 2023.

[9] J. Lin, N. Jiang, Z. Zhang, W. Chen, and T. Zhao, "LMQFormer: A laplace-prior-guided mask query transformer for lightweight snow removal," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 33, no. 11, pp. 6225-6235, 2023.

[10] X. Guo, X. Wang, X. Fu, and Z.-J. Zha, "Deep unfolding network for image desnowing with snow shape prior," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 35, no. 5, pp. 4740-4752, 2025.

[11] S. Chen, T. Ye, Y. Liu, E. Chen, J. Shi, and J. Zhou, "SnowFormer: Scale-aware transformer via context interaction for single image desnowing," *arXiv preprint arXiv:2208.09703,* 2022.

[12] Y. Zhang, "BAT-Net: Bidirectional Attention Transformer Network for Joint Single-Image Desnowing and Snow Mask Prediction," *Information,* vol. 16, no. 11, p. 966, 2025.

[13] G. Wang, Y. Zhou, F. Shi, and Z. Jia, "SnowMamba: Achieving more precise snow removal with Mamba," *Applied Sciences,* vol. 15, no. 10, p. 5404, 2025.

[14] J. Lai *et al.*, "Snowmaster: Comprehensive real-world image desnowing via mllm with multi-model feedback optimization," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025: IEEE, pp. 4302-4312.

[15] Q. Guan, Q. Yang, X. Chen, T. Song, G. Jin, and J. Jin, "Weatherbench: A real-world benchmark dataset for all-in-one adverse weather image restoration," in *Proceedings of the 33rd ACM international conference on multimedia*, 2025, pp. 12607-12613.

[16] J. Jiang, Z. Zuo, G. Wu, K. Jiang, and X. Liu, "A survey on all-in-one image restoration: Taxonomy, evaluation and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2025.

[17] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022: IEEE, pp. 17431-17441.

[18] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022: IEEE, pp. 2343-2353.

[19] C. Li, F. Sun, H. Zhou, Y. Xie, Z. Li, and L. Zhu, "Multi-weather restoration: An efficient prompt-guided convolution architecture," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 35, no. 2, pp. 1436-1450, 2024.

[20] Y. Wen, T. Gao, Z. Li, J. Zhang, K. Zhang, and T. Chen, "All-in-one weather-degraded image restoration via adaptive degradation-aware self-prompting model," *IEEE Transactions on Multimedia,* vol. 27, pp. 3343-3355, 2025.

[21] D. Xie, X. Hu, Y. Zhou, and S. Duan, "All-in-one adverse weather removal via dual state space-based diffusion model with degradation-aware guidance," *Pattern Recognition,* vol. 171, p. 112081, 2026.

[22] Y. Cui, S. W. Zamir, S. Khan, A. Knoll, M. Shah, and F. Khan, "Adair: Adaptive all-in-one image restoration via frequency mining and modulation," in *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

[23] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32, no. 1.

[24] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: IEEE, pp. 606-615.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015: Springer, pp. 234-241.

[26] T. S. Gunawan and M. Kartiwi, "Stability-Aware Physics-Prior Conditioning for Lightweight Image Dehazing," presented at the 2026 IEEE 12th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA 2026), Kuching, Sarawak, Malaysia, 28-30 July 2026, 2026.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," presented at the Proceedings of the 13th International Conference on Learning Representations (ICLR), 2019.

[28] 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.

[29] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: IEEE, pp. 586-595.

[30] 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.

[31] 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.

[32] N. Venkatanath, D. Praneeth, S. C. Sumohana, and S. M. Swarup, "Blind image quality evaluation using perception based features," in *2015 Twenty First National Conference on Communications (NCC)*, 2015: IEEE, pp. 1-6.