CLOSING DEGRADATION LOOP: A STRUCTURED INFERENCE-RESTORATION-VERIFICATION FRAMEWORK FOR ALL-IN-ONE INFRARED IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

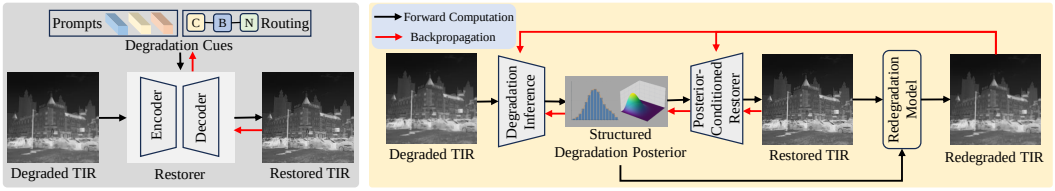
All-in-one thermal infrared image restoration recovers clean images from observations with unknown single or compound degradations. Existing methods mainly use predicted degradation cues as one-way prompts or routing signals, without explicitly reasoning about how degradations form the observation or verifying whether restoration is consistent with that process. We propose **SIRV**[1], a **S**tructured **I**nference-**R**estoration-**V**erification framework, to close this loop. Specifically, structured degradation inference first estimates a posterior over degradation states, retaining alternatives to avoid premature decisions and provide composition and severity cues. Then, posterior-conditioned restoration integrates cues from the inferred degradation posterior to modulate a shared restorer, adapting restoration to degradation composition, strength, and spatial distribution. Finally, degradation-consistent verification reuses sampled degradation parameters from the posterior in a differentiable redegradation model to reconstruct the input observation, penalizing restorations inconsistent with it. Together, SIRV transforms one-way guidance into a coherent loop in which uncertain degradation explanations guide restoration and the resulting restorations are verified against the input during training, improving robustness to ambiguous compound degradations and preserving structural details. Experiments across infrared benchmarks demonstrate SIRV's state-of-the-art restoration performance and strong generalization. On M^3FD, SIRV achieves 28.329 dB PSNR and 0.918 SSIM, exceeding the best previous scores by 2.302 dB and 0.038, respectively.

1 INTRODUCTION

Thermal infrared (TIR) imaging captures object-emitted radiation, enabling reliable perception in darkness and visually obscured environments [55]. However, limited temperature differences, atmospheric propagation, optical defocus, and detector imperfections produce degradations including low contrast [2], blur [3], spatial non-uniformity [54], stripe interference [43], and Gaussian noise [35, 28]. These degradations often coexist and obscure weak structures, making recovery under unknown compound conditions challenging. All-in-one infrared restoration addresses these degradations using a single model without requiring externally provided degradation information at inference. Recent methods exploit prompts, parameter-efficient adaptation, and infrared-specific modeling [45, 63, 35, 28]. PPFN [35] combines degradation-specific and type-specific prompts with selective progressive training, whereas SEGD [28] uses degradation evidence and structural entropy to select restoration paths. Despite this progress, predicted degradation cues primarily provides one-way guidance through prompts or routing decisions (Fig. 1 (a)). Such one-way guidance indicates what degradations may be present, but offers limited insight into how they jointly formed the degradation observation and provides no feedback on whether the resulting restoration remains consistent with the observation. Thus, ambiguous mixtures can trigger unsuitable operations, leaving residual degradation or distorting image structures.

[1]Code is available at https://anonymous.4open.science/r/SIRV-4FD2.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) One-way Guidance **(b) Our Framework (SIRV)**

Figure 1: Comparison of restoration paradigms. (a) One-way guidance uses degradation cues to guide image restoration. (b) SIRV reuses degradation parameters sampled from the posterior for restoration and redegradation, verifying consistency with the input observation during training.

To address these limitations, we propose **SIRV**, a **S**tructured **I**nference-**R**estoration-**V**erification framework for all-in-one infrared image restoration. SIRV closes the degradation loop by inferring how an observation was formed, conditioning restoration on that explanation, and checking the result through redegradation, as shown in Fig. 1 (b). First, **structured degradation inference** estimates a posterior over degradation states, thereby retaining plausible alternatives and supplying explicit composition and severity cues. Then, **posterior-conditioned restoration** integrates cues from the inferred degradation posterior through adaptive feature modulation, enabling a shared restorer to adapt to degradation composition, strength, and spatial distribution. Finally, **degradation-consistent verification** reuses the same degradation-parameter sample for restoration and differentiable redegradation, enforcing observation consistency during training. Overall, SIRV transforms one-way guidance into a coherent inference-restoration-verification loop, improving robustness to ambiguous compound degradations and preserving structural details. Experiments across infrared benchmarks demonstrate SIRV's state-of-the-art restoration performance and strong generalization. On M^3FD, SIRV achieves 28.329 dB PSNR and 0.918 SSIM, exceeding the best previous results by 2.302 dB and 0.038, respectively. Our contributions are summarized as follows:

- We formulate all-in-one infrared image restoration as a structured inference-restoration-verification framework that couples degradation inference and restoration by requiring the restored image and sampled parameters to jointly reproduce the observed input.
- We develop structured degradation inference and posterior-conditioned restoration to convert uncertainty over degradation states into composition-, severity-, and location-adaptive operations within a shared restorer.
- We introduce degradation-consistent verification with coupled posterior sampling, reusing sampled degradation parameters for restoration and redegradation to enforce consistency with the degraded observation.
- Extensive experiments demonstrate that SIRV achieves state-of-the-art performance across infrared benchmarks, with improved degradation removal and structural preservation.

2 Related Work

2.1 Infrared Image Restoration

Conventional infrared restoration usually targets a known degradation. Non-uniformity correction estimates detector responses or learns residual mappings [5, 7, 11], while stripe removal exploits decomposition, attention, sparsity, or gradient priors [6, 10, 36, 48, 60]. Infrared denoising methods similarly employ statistical estimation and learned feature aggregation to address specific degradations [13, 32, 16, 4, 19]. Infrared deblurring research covers learned inversion, detector-aware motion modeling, and dedicated benchmarks [18, 25, 52, 56, 22]. Other studies address low contrast [23, 40, 44, 46] or insufficient resolution [17, 30, 31]. These methods provide strong task-specific priors and effective architectures. However, these methods typically target predefined restoration tasks, limiting their applicability to unknown degradation types and mixtures.

2.2 All-in-One Image Restoration

All-in-one image restoration aims to restore images affected by diverse unknown degradations and their mixtures using a single model. RGB all-in-one restoration uses unified representations, transferable priors, and adaptive prompts [27, 37, 61, 45, 1, 39]. Vision-language guidance, diffusion, pretraining, and routing further improve generality [9, 12, 20, 63, 15, 58]. Yet RGB priors may erase weak targets or introduce unsupported textures in infrared images [4, 44]. Infrared all-in-one restoration has emerged more recently [35, 38, 8, 28, 29]. PPFN fuses degradation prompts, and SEGD selects restoration paths using degradation evidence and structural entropy [35, 28]. In many of these methods, predicted degradation cues guide restoration through prompts, strength modulation, or path selection, without explicitly verifying whether the restored image and inferred degradation state jointly explain the observation. SIRV instead infers a structured degradation posterior, conditions restoration on it, and checks the restored image for consistency with the observation through redegradation during training, forming a coherent loop to improve robustness to ambiguous compound degradations.

3 Methodology

3.1 Problem Formulation

All-in-one infrared restoration recovers a clean image without knowing the degradation type or severity at inference. Let $x, y \in [0,1]^{1\times H\times W}$ denote a normalized single-channel clean image and its degraded observation, where H and W are the spatial dimensions. We consider low contrast (C), Gaussian blur (B), and compound sensor noise (N), and the noise family contains optical nonuniformity (O), stripe interference (S), and signal-independent Gaussian noise (G). Let $\mathcal{R}$ be the set of admissible non-empty ordered degradation sequences. A sequence $r = (r_1, \ldots, r_L) \in \mathcal{R}$ contains L operators, where $r_l \in \{C, B, N\}$, and $\theta = (\theta_{r_1}, \ldots, \theta_{r_L})$ denotes their degradation parameters. Together, r and θ constitute the degradation state. Applying the operators to x in the order specified by r produces y, and different orders may yield different observations. Given only y, the goal is to estimate $\widehat{x} \approx x$ while r and θ remain unknown.

3.2 Motivation

Existing methods mainly use predicted degradation cues as one-way prompts or routing signals [35, 28], without explicitly reasoning about how multiple degradations jointly form the observation or verifying whether restoration agrees with that process. We instead treat the unknown degradation state as a latent explanation connecting restoration with redegradation. Let $z = (r, \theta, x)$ collect the degradation sequence, its parameters, and the clean image. Motivated by variational approximation [51], we define and factorize the approximate posterior as:

$$q_{\phi,\psi}(z \mid y) = q_\phi(r \mid y)\, q_\phi(\theta \mid y, r)\, q_\psi(x \mid y, r, \theta), \tag{1}$$

where $q_\phi(r \mid y)$ and $q_\phi(\theta \mid y, r)$ infer the structured degradation state, $q_\psi(x \mid y, r, \theta)$ estimates the clean image conditioned on that state, and ϕ and ψ denote the learnable parameters of the degradation-inference network and shared restorer, respectively. For each supervised training sample i, we introduce the sample-conditioned reference prior $p_i(z) = p_i(r)p_i(\theta \mid r)p_i(x)$. We omit the sample index i for notational simplicity. Its factors encode the available sequence, parameter, and clean-image supervision and regularize posterior learning only during training. Together with the redegradation model $p(y \mid z)$, it induces $p(y) = \mathbb{E}_{p(z)}[p(y \mid z)]$. This construction defines a supervised conditional variational objective by applying Jensen's inequality [41]:

$$\log p(y) \geq \mathcal{L}_{\mathrm{cELBO}} \equiv \mathbb{E}_{q_{\phi,\psi}(z|y)} \left[\log p(y \mid z)\right] - D_{\mathrm{KL}} \left[q_{\phi,\psi}(z \mid y) \,\|\, p(z)\right], \tag{2}$$

where D_{KL} denotes the Kullback-Leibler (KL) divergence [24], and $\mathcal{L}_{\mathrm{cELBO}}$ is the conditional evidence lower bound [14]. Using the posterior and prior factorizations, its negative form becomes

$$\begin{aligned} -\mathcal{L}_{\mathrm{cELBO}} = {} & D_{\mathrm{KL}} \left[q_\phi(r \mid y) \,\|\, p(r)\right] \\ & + \mathbb{E}_{q_\phi(r|y)} D_{\mathrm{KL}} \left[q_\phi(\theta \mid y, r) \,\|\, p(\theta \mid r)\right] \\ & + \mathbb{E}_{q_\phi(r,\theta|y)} D_{\mathrm{KL}} \left[q_\psi(x \mid y, r, \theta) \,\|\, p(x)\right] \\ & + \mathbb{E}_{q_{\phi,\psi}(z|y)} \left[-\log p(y \mid z)\right]. \end{aligned} \tag{3}$$

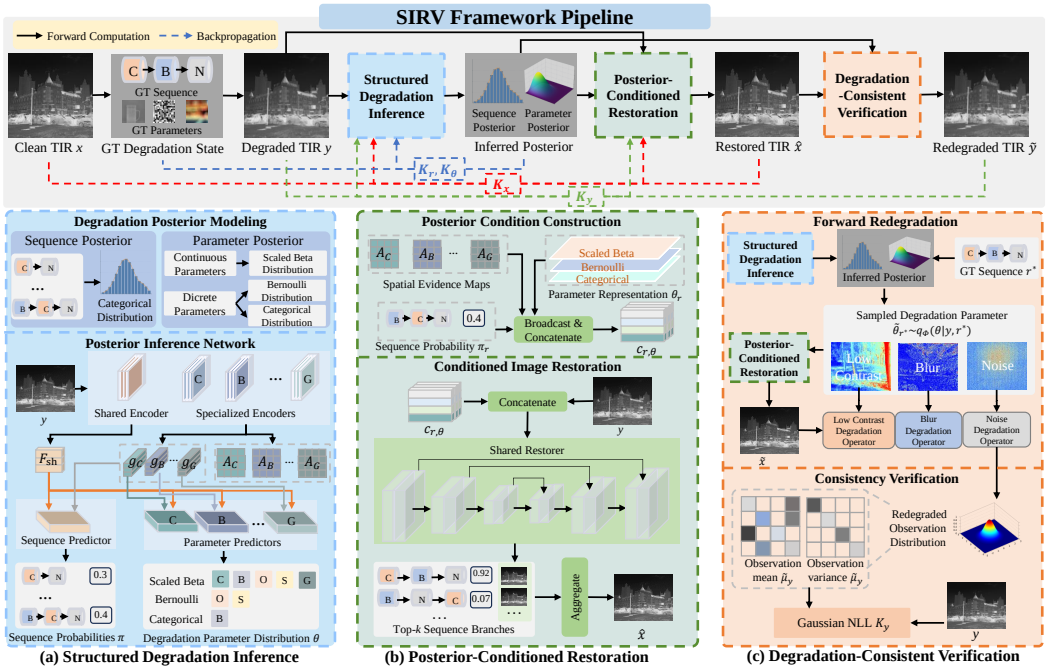


Figure 2: **Illustration of SIRV**. SIRV infers a structured degradation posterior over degradation states, conditions a shared restorer on posterior cues, and verifies consistency with the input observation through differentiable redegradation during training, improving robustness to ambiguous compound degradations and preserving structural details.

The first three terms regularize sequence inference, parameter inference, and clean-image estimation, respectively. The final term verifies whether the restored image and inferred degradation parameters reproduce the observation under the modeled degradation process. Prior definitions, closed-form KL divergences, dimensional normalization, and the supervised-branch estimator are detailed in Appendix A.

This formulation motivates the three components of SIRV, as shown in Fig. 2. First, **structured degradation inference** realizes $q_\phi(r \mid y)$ and $q_\phi(\theta \mid y, r)$ and corresponds to the first two KL terms in Eq. 3. By retaining plausible ordered alternatives, it reduces premature decisions under ambiguous compound degradations, as shown in Fig. 2 (a). **Posterior-conditioned restoration** realizes $q_\psi(x \mid y, r, \theta)$ and corresponds to the third KL term, converting the inferred degradation states into adaptive restoration operations, as displayed in Fig. 2 (b). **Degradation-consistent verification** instantiates $p(y \mid x, r, \theta)$ and corresponds to the final expected negative log-likelihood (Fig. 2 (c)), reusing the sampled degradation parameters to verify whether the conditioned restoration is consistent with the observation, directly constraining parameter inference and restoration. The four terms in Eq. 3 form the training objective of SIRV. Together, the components form a coherent inference-restoration-verification loop, improving robustness to ambiguous compound degradations and preserving structural details without verification cost at inference.

3.3 Structured Degradation Inference

Degradation posterior modeling. We introduce a structured degradation posterior as an image-conditioned distribution over plausible degradation states to represent uncertainty about how the observation was formed and guide restoration. The structured degradation posterior includes a sequence posterior describing which degradations occur and in what order, and a parameter posterior describing their severity and configuration. We factorize this posterior as $q_\phi(r, \theta \mid y) = q_\phi(r \mid y) q_\phi(\theta \mid y, r)$, with both factors inferred by a network parameterized by ϕ. The sequence posterior is $q_\phi(r \mid y) = \mathrm{Cat}(r \mid \pi_\phi(y))$, where Cat denotes the categorical distribution and $\pi_\phi(y)$ contains the predicted probabilities over all admissible ordered degradation sequences $\mathcal{R}$. We abbreviate $\pi_r = \pi_{\phi,r}(y)$, with $\sum_{r \in \mathcal{R}} \pi_r = 1$. For each sequence, we factorize the parameter posterior as

$q_\phi(\theta \mid y, r) = \prod_{t \in \mathcal{I}_r^{\text{cont}}} q_{\phi,t}(\theta_t \mid y, r) \cdot \prod_{d \in \mathcal{I}_r^{\text{disc}}} q_{\phi,d}(\theta_d \mid y, r)$, where $\mathcal{I}_r^{\text{cont}}$ and $\mathcal{I}_r^{\text{disc}}$ index its continuous and discrete parameters. Continuous factors use scaled Beta distributions [47] over valid physical ranges, and discrete factors use Categorical or Bernoulli distributions [49]. This representation retains alternative degradation explanations and their probabilities to guide restoration.

Posterior Inference Network. The posterior inference network estimates the sequence and parameter posteriors from the degraded TIR image y. A shared encoder extracts a global context descriptor F_{sh} that captures cross-degradation information. Specialized encoders produce a spatial evidence map A_j and a global degradation feature g_j for each degradation type $j \in \{C, B, O, S, G\}$, capturing local and global degradation-specific evidence, respectively. For the sequence posterior, a sequence predictor combines F_{sh} and $\{g_j\}_{j \in \{C,B,O,S,G\}}$ to predict sequence probabilities through softmax. For the parameter posterior, parameter predictors combine F_{sh} and the corresponding g_j to predict the distributions of degradation parameters, while A_j is retained for restoration conditioning. For training sample i, the posterior inference losses $K_r = D_{\text{KL}}[q_\phi(r \mid y_i) \| p_i(r)]$ and $K_\theta = D_{\text{KL}}[q_\phi(\theta \mid y_i, r_i^\star) \| p_i(\theta \mid r_i^\star)]$ align predictions with reference priors constructed from the ground-truth (GT) degradation state. They provide sequence and parameter supervision associated with the first two terms of Eq. 3, with K_θ evaluated on the ground-truth sequence $r_i^\star$. These priors are used only during training. Architectural details, prior definitions, and KL expressions are provided in Appendix B.

3.4 Posterior-Conditioned Restoration

Posterior condition construction. Posterior-conditioned restoration translates inferred degradation states into guidance for a shared restorer. For a candidate sequence r, we construct a condition feature $c_{r,\theta}$ by broadcasting spatially and concatenating sequence posterior probability π_r, parameter representation θ_r, and spatial evidence maps $\{A_j\}_{j \in \{C,B,O,S,G\}}$, and components associated with absent degradation families are masked. The sequence and parameter cues describe degradation composition, order, and severity, while the spatial maps indicate where restoration should adapt.

Conditioned image restoration. Given the degradation condition, a shared restorer network processes y together with $c_{r,\theta}$ to adapt restoration to the inferred degradation state. The restorer uses a lightweight encoder-decoder with conditional residual blocks and a Restormer bottleneck [59]. The encoder extracts multiscale features from the channel-wise concatenation of y and $c_{r,\theta}$, while the decoder progressively upsamples these features using additive skip connections from corresponding encoder stages. The output layer predicts an image residual, which is added to y to obtain the restored image. We denote this restoration mapping by μ_ψ, yielding the candidate image $\widehat{x}_r = \mu_\psi(y, c_{r,\theta})$. Within the variational formulation, this output defines the posterior $q_\psi(x \mid y, r, \theta) = \mathcal{N}(x \mid \widehat{x}_r, \varepsilon_x^2 I)$, where $\varepsilon_x^2 = 10^{-5}$ is fixed and I is the identity matrix for the vectorized image. The clean-image KL loss becomes $K_x = \|\widehat{x}_{r^\star} - x^\star\|_2^2 / (2\varepsilon_x^2)$, directly penalizing reconstruction errors against the clean target $x^\star$. Its pixelwise contributions are averaged during optimization. Architectural and KL details are provided in Appendix C.

At inference, given only y, we retain the k most probable sequences, denoted by $\mathcal{R}_k$. For each retained sequence branch, the inferred parameter representation $\widehat{\theta}_r$ combines continuous posterior modes with discrete posterior probabilities. The final restoration is $\widehat{x} = \sum_{r \in \mathcal{R}_k} \bar{\pi}_r \, \mu_\psi(y, c_{r,\widehat{\theta}_r})$, where $\bar{\pi}_r = \pi_r / \sum_{s \in \mathcal{R}_k} \pi_s$ normalizes the retained probabilities. This aggregation allows alternative degradation explanations to contribute while limiting the number of restoration passes.

3.5 Degradation-Consistent Verification

Degradation-consistent verification checks whether a restored image and its degradation explanation jointly account for the input observation. It complements clean-image and degradation-state supervision by requiring their predictions to remain consistent through the redegradation process.

Forward redegradation. We introduce a differentiable forward model $\mathcal{U}$ to map the restored image back to observation space. It computes an approximate observation mean and pixelwise variance, describing the reconstructed observation and the uncertainty introduced by the modeled degradations. During training, we use the GT sequence $r^\star = (r_1^\star, \ldots, r_L^\star)$ and obtain a parameter sample $\widetilde{\theta}_{r^\star}$ from $q_\phi(\theta \mid y, r^\star)$. This sample conditions the restoration $\widetilde{x} = \mu_\psi(y, c_{r^\star, \widetilde{\theta}_{r^\star}})$ and is reused for

redegradation. Starting from $(\mu^{(0)}, v^{(0)}) = (\widetilde{x}, \tau_y^2 \mathbf{1})$, with $\tau_y^2 = 10^{-4}$, we sequentially update

$$\left(\mu^{(l)}, v^{(l)}\right) = \mathcal{U}_{r_l^\star}\left(\mu^{(l-1)}, v^{(l-1)}; \widetilde{\theta}_{r_l^\star}\right), \quad l = 1, \ldots, L, \tag{4}$$

where $\mathcal{U}_{r_l^\star}$ updates the mean μ and variance v of the redegraded observation for the l-th operator using its parameters $\widetilde{\theta}_{r_l^\star}$. The final outputs $\widetilde{\mu}_y = \mu^{(L)}$ and $\widetilde{v}_y = \max(v^{(L)}, \tau_y^2 \mathbf{1})$ define an approximate Gaussian observation distribution, where the variance floor ensures numerical stability. Operator-specific redegradation updates are given in Appendix D.

Consistency verification. Using the propagated mean and variance, we define the observation-consistency loss as the Gaussian negative log-likelihood: $K_y = -\log \mathcal{N}(y \mid \widetilde{\mu}_y, \mathrm{Diag}(\widetilde{v}_y))$, where Diag constructs a diagonal covariance matrix. By integrating sampled degradation parameters and restoration, K_y jointly constrains restoration and parameter inference, providing the practical observation supervision associated with the fourth term in Eq. 3. Its penalty accounts for uncertainty in the modeled degradation process, discouraging restorations that appear visually plausible but fail to explain the input observation. The forward verification branch is removed at inference.

4 Experiments

4.1 Experimental Settings

Datasets and metrics. We evaluate paired restoration on HM-TIR [35], Night-TIR [28], and M^{3}FD [33] using PSNR and SSIM, and real-world restoration on the 1,000-image TIIQAD [57] using MUSIQ [21], NIMA [50], and NIQE [42]. Higher MUSIQ and NIMA and lower NIQE indicate better quality. HM-TIR contains 1,503 images and follows an 8:2 split, and cross-dataset tests use 647 Night-TIR and 300 M^{3}FD images. We use mAP [62] to evaluate the effect of restoration on downstream object detection and also report parameter count and latency. Data construction and evaluation protocols are provided in Appendix F, with training and inference settings in Appendix E.

Comparison methods. We compare SIRV with degradation-specific infrared methods WFAF [43], LRSID [6], and TSIRIE [44], RGB all-in-one methods DA-CLIP [39], DiffUIR [63], PromptIR [45], and DRNet [26], and recent infrared all-in-one methods ECMRNet [29], PPFN [35], SEGD [28], and HiDRA [8].

4.2 Comparison With State-of-the-Art Methods

Performance Comparison on HM-TIR. Table 1 reports the comparison on HM-TIR. SIRV achieves the best average PSNR and SSIM of 28.331 dB and 0.892, surpassing the strongest previous results by 1.552 dB and 0.041, respectively. It ranks first in seven of the eight reported metrics. Under triple degradation, SIRV obtains the second-best PSNR but the highest SSIM, improving the previous best by 0.033. This behavior indicates that explicitly inferring degradation states is particularly beneficial for preserving thermal structures when several degradations interact. In the HM-TIR row of Fig. 3, SIRV reconstructs building contours and fine edge transitions more faithfully, whereas competing outputs either lose these structures, blur their boundaries, or introduce artifacts.

Cross-dataset Generalization. Table 3 evaluates all methods on M^{3}FD and Night-TIR without target-domain fine-tuning. On M^{3}FD, SIRV reaches 28.329 dB PSNR and 0.918 SSIM, improving the best previous averages by 2.302 dB and 0.038. On Night-TIR, it obtains 24.235 dB and 0.862, with gains of 1.174 dB and 0.026. Notably, SIRV achieves the highest SSIM for every degradation level on both datasets, although its PSNR is second best in several more complex subsets. This consistent structural advantage suggests that posterior-conditioned restoration transfers more reliably across scene content and sensor domains than fixed prompting or routing. In the Night-TIR and M^{3}FD rows of Fig. 3, SIRV resolves thermal variations around human targets more precisely and better preserves high-contrast local details, consistent with its systematic SSIM advantage.

Generalization to real-world degradations. SIRV also generalizes to authentic degradations that are not paired with clean references. As shown in Table 2, it achieves the highest MUSIQ and NIMA scores of 51.675 and 5.061, exceeding the respective second-best results by 1.367 and 0.039, while attaining the second-best NIQE of 5.611. The agreement between two learned perceptual metrics and a competitive naturalness score indicates a favorable balance between enhancement and artifact

Table 1: Quantitative comparison on HM-TIR.

Method	Single		Double		Triple		Avg.	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
WFAF	22.572	0.733	19.209	0.622	16.893	0.532	19.558	0.629
LRSID	20.463	0.599	18.602	0.587	16.635	0.565	18.566	0.584
TSIRIE	27.169	0.888	24.702	0.826	22.483	0.762	24.785	0.826
DA-CLIP	16.635	0.645	16.533	0.583	15.947	0.532	16.372	0.587
DiffUIR	21.781	0.838	20.475	0.761	18.314	0.680	20.190	0.760
PromptIR	25.855	0.840	22.350	0.748	19.420	0.676	22.541	0.755
DRNet	28.830	0.897	25.537	0.822	23.238	0.768	25.868	0.829
ECMRNet	24.390	0.881	23.438	0.840	22.346	0.789	23.391	0.837
PPFN	25.287	0.870	24.904	0.827	23.196	0.767	24.463	0.822
SEGD	32.347	0.936	25.882	0.848	22.107	0.770	26.779	0.851
HiDRA	27.605	0.866	26.971	0.831	25.541	0.787	26.705	0.828
SIRV (Ours)	**33.219**	**0.960**	**27.809**	**0.893**	**23.966**	**0.822**	**28.331**	**0.892**

Table 2: Quantitative comparison on the TIIQAD dataset.

Method	MUSIQ↑	NIMA↑	NIQE↓
WFAF	49.730	4.935	6.068
LRSID	30.906	4.528	6.092
TSIRIE	45.521	4.583	5.800
DA-CLIP	46.108	4.855	10.779
DiffUIR	47.792	4.443	6.663
PromptIR	44.176	4.561	6.414
DRNet	46.791	4.685	6.507
ECMRNet	46.785	4.817	6.080
PPFN	50.308	5.022	6.414
SEGD	48.277	4.840	6.244
HiDRA	50.188	4.865	5.052
SIRV (Ours)	**51.675**	**5.061**	**5.611**

Table 3: Quantitative cross-dataset results on the M^{3}FD and Night-TIR datasets.

Method	M^{3}FD								Night-TIR							
	Single		Double		Triple		Avg.		Single		Double		Triple		Avg.	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
WFAF	24.577	0.768	19.242	0.671	17.063	0.621	20.294	0.687	23.134	0.770	18.117	0.621	16.825	0.560	19.359	0.650
LRSID	22.767	0.761	19.047	0.724	16.918	0.668	19.578	0.718	19.585	0.555	17.072	0.535	16.309	0.542	17.656	0.544
TSIRIE	24.780	0.891	22.644	0.847	21.074	0.794	22.832	0.844	23.796	0.864	20.816	0.794	19.328	0.728	21.313	0.795
DA-CLIP	17.905	0.636	17.269	0.601	16.225	0.560	17.133	0.599	14.096	0.662	13.790	0.596	13.670	0.543	13.851	0.600
DiffUIR	21.428	0.876	20.248	0.822	18.896	0.746	20.190	0.815	21.202	0.846	19.740	0.759	18.512	0.684	19.818	0.763
PromptIR	26.429	0.850	21.596	0.780	19.212	0.726	22.412	0.785	25.125	0.835	21.139	0.737	19.348	0.667	21.871	0.746
DRNet	25.830	0.902	22.909	0.857	22.079	0.815	23.606	0.858	24.370	0.873	20.794	0.797	19.441	0.736	21.535	0.802
ECMRNet	23.174	0.896	22.410	0.877	21.969	0.827	22.517	0.867	21.803	0.867	21.773	0.840	20.795	0.779	21.457	0.829
PPFN	25.061	0.897	24.653	0.862	23.104	0.814	24.273	0.858	22.350	0.867	22.004	0.824	20.820	0.757	21.725	0.816
SEGD	30.289	0.942	25.065	0.881	22.194	0.817	25.849	0.880	27.370	0.921	22.139	0.831	19.675	0.755	23.061	0.836
HiDRA	27.756	0.913	25.744	0.880	24.581	0.834	26.027	0.875	24.390	0.862	23.077	0.820	21.675	0.771	23.047	0.817
SIRV (Ours)	**33.885**	**0.966**	**26.951**	**0.920**	**24.151**	**0.869**	**28.329**	**0.918**	**29.589**	**0.949**	**22.143**	**0.847**	**20.973**	**0.791**	**24.235**	**0.862**

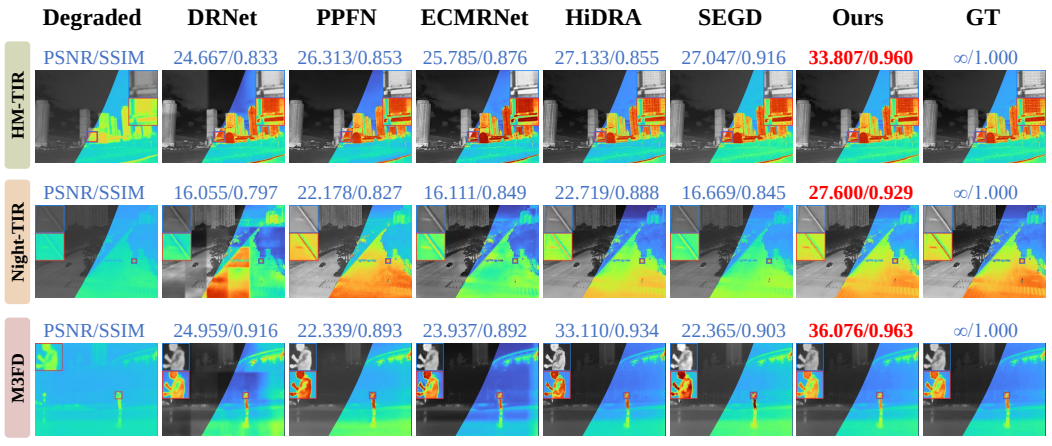


Figure 3: Qualitative comparisons on HM-TIR, Night-TIR, and M^{3}FD. The highest PSNR and SSIM values among restoration methods are highlighted in red.

suppression. Figure 4 further shows that SIRV recovers clearer object contours and thermal textures while producing stronger, well-balanced contrast. Together, these results suggest that the inferred degradation explanation remains useful beyond the synthetic training setting.

Downstream object detection. We evaluate all restored images using the same frozen YOLOv8s detector [53], fine-tuned on M^{3}FD with the 300 test source images excluded. SIRV obtains 0.697 mAP$_{50}$ and 0.447 mAP$_{50:95}$ in Table 4, improving the strongest restored-image baseline by 0.009 and 0.008, respectively. Among restoration methods, SIRV ranks first in four classes and second in the remaining two. Relative to the degraded inputs, the gains are 0.653 and 0.423, showing that

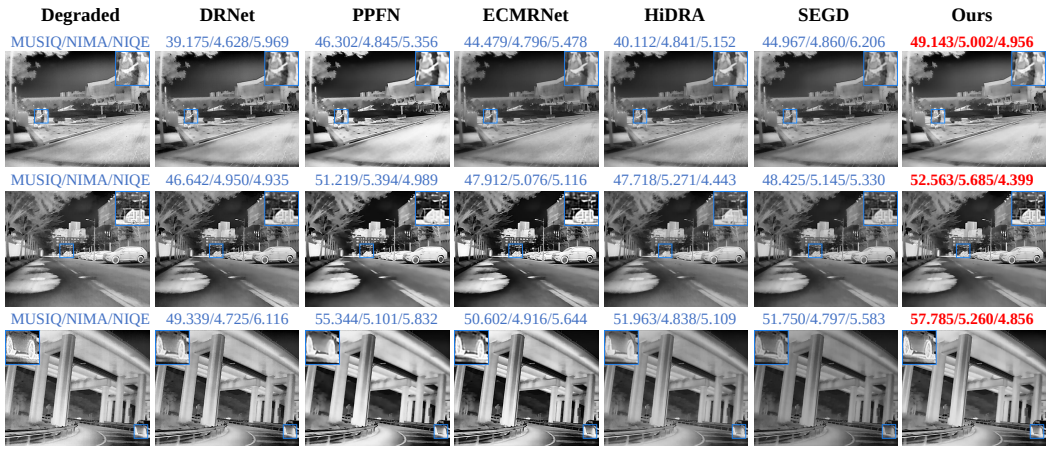


Figure 4: Qualitative comparison on real-world TIIQAD images.

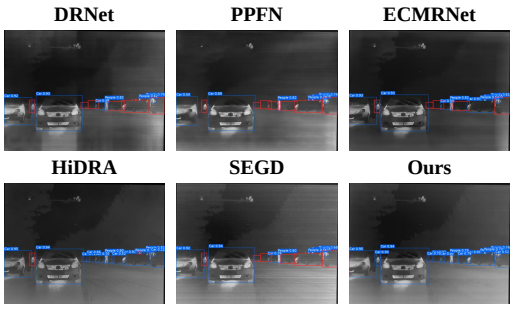


Figure 5: Qualitative comparison of downstream object detection results on M^{3}FD. Red boxes indicate representative missed or false detections in competing results.

Table 4: Quantitative comparison of different TIR image restoration methods on downstream object detection using the M^{3}FD dataset.

Method	Bus	Car	Lamp	Motor	Person	Truck	mAP$_{50}$	mAP$_{50:95}$
Degraded	0.001	0.040	0.013	0.035	0.054	0.000	0.044	0.024
WFAF	0.387	0.402	0.123	0.252	0.356	0.333	0.480	0.309
LRSID	0.227	0.291	0.066	0.142	0.164	0.194	0.315	0.181
TSIRIE	0.471	0.497	0.128	0.299	0.452	0.400	0.593	0.374
DA-CLIP	0.455	0.430	0.111	0.203	0.386	0.357	0.528	0.324
DiffUIR	0.429	0.473	0.137	0.238	0.432	0.381	0.552	0.348
PromptIR	0.366	0.445	0.147	0.241	0.417	0.373	0.514	0.332
DRNet	0.459	0.519	0.177	0.305	0.456	0.418	0.617	0.389
ECMRNet	0.563	0.554	0.172	0.312	0.463	0.459	0.665	0.421
PPFN	0.485	0.480	0.123	0.267	0.439	0.392	0.584	0.364
SEGD	0.518	0.530	0.192	0.316	0.457	0.415	0.643	0.405
HiDRA	0.638	0.571	0.205	0.340	0.451	0.425	0.688	0.439
SIRV (Ours)	0.602	0.580	0.220	0.357	0.477	0.449	0.697	0.447
Clean	0.709	0.626	0.301	0.400	0.504	0.517	0.774	0.509

the recovered contrast and structures are directly useful to a downstream detector. The red boxes in Fig. 5 mark missed or false detections in competing results, where restoration artifacts or weakened infrared edges obscure target cues. By better preserving these cues, SIRV produces more complete detections.

Model Efficiency Comparison. Table 5 compares parameter count and inference latency of models. SIRV contains only 3.6M parameters, making it substantially smaller than most learned all-in-one baselines, and only TSIRIE and SEGD are smaller among methods with reported model sizes. Its 149.6 ms latency is also lower than seven of the nine neural baselines in the table. Thus, the restoration gains do not arise from excessive model scaling, and degradation-consistent verification adds no computation at inference because the forward branch is discarded after training.

4.3 Ablation Studies

Effect of core components. Table 6 progressively evaluates the inference-conditioned restoration and verification stages of SIRV. Introducing structured degradation inference together with posterior conditioning improves the shared-restorer baseline by 0.718 dB PSNR and 0.010 SSIM. Adding degradation-consistent verification further provides gains of 0.869 dB and 0.031, resulting in total improvements of 1.587 dB and 0.041. Together with the dedicated condition ablation below, these results demonstrate the complementary roles of posterior-guided restoration and observation-space consistency.

Effect of ordered degradation inference. All variants in Table 7 use the shared restorer and posterior supervision without degradation-consistent verification. The order-agnostic variant predicts only

Table 5: Comparison of parameter count and per-image inference latency on HM-TIR.

Methods	WFAF	LRSID	TSIRIE	DA-CLIP	DiffUIR	PromptIR	DRNet	ECMRNet	PPFN	SEGD	HiDRA	Ours
Params (M)	-	-	2.5	295.2	36.2	35.6	16.6	17.4	26.6	2.6	1414.2	3.6
Latency (ms)	24.7	2489.8	8.2	1509.5	410.4	278.9	582.3	95.4	1339.3	177.2	378.2	149.6

Table 6: Incremental ablation of core components in SIRV.

Configuration	PSNR↑	SSIM↑
Shared restorer	26.744	0.851
+ Posterior inference	27.462	0.861
+ Consistency verification	28.331	0.892

Table 7: Ablation of ordered degradation-sequence inference.

Setting	PSNR↑	SSIM↑
No order	26.882	0.856
Ordered, top-1	27.242	0.854
Ordered, top-k	27.462	0.861

Table 8: Ablation of posterior conditions on HM-TIR.

Condition	PSNR↑	SSIM↑
Sequence	27.849	0.879
+ Parameters	28.195	0.883
Full condition (Ours)	28.331	0.892

the degradation set, whereas the others infer ordered sequences. Ordered top-1 improves PSNR but slightly reduces SSIM, while top-k achieves the best restoration of 27.462 dB and 0.861. It outperforms the order-agnostic variant by 0.580 dB and 0.005, showing that retaining finer ordered alternatives provides more effective restoration guidance despite the harder sequence-prediction task.

Effect of posterior-conditioned restoration. Table 8 progressively enriches the condition supplied to the shared restorer. All variants use the same inference network, restorer, and Gaussian-NLL verification objective, differing only in the condition supplied to the restorer. Adding estimated parameters to the sequence improves PSNR by 0.346 dB and SSIM by 0.004. The complete condition, which further incorporates posterior probability and degradation-specific spatial evidence, achieves 28.331 dB and 0.892, yielding total gains of 0.482 dB and 0.013 over sequence-only conditioning. Thus, the sequence provides degradation composition, while parameter and spatial posterior cues adapt restoration to its severity and location.

Table 9: Ablation of degradation consistency on HM-TIR.

Objective	PSNR↑	SSIM↑
None	27.462	0.861
L_1	28.187	0.882
Gaussian NLL (Ours)	28.331	0.892

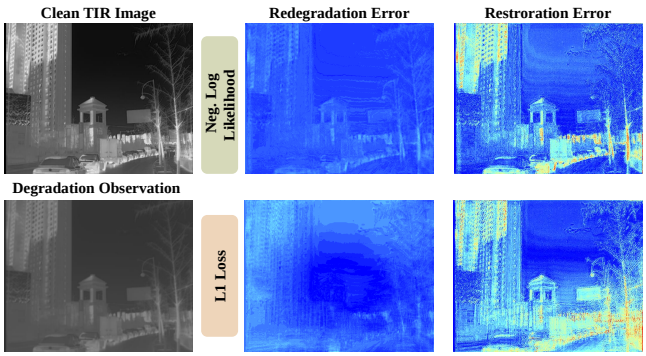


Figure 6: Redegradation and restoration error maps, where bluer regions indicate lower errors.

Effect of degradation-consistent verification. All variants in Table 9 use the same structured degradation inference and posterior-conditioned restorer, differing only in the consistency objective. Adding an L_1 redegradation loss improves the no-verification baseline by 0.725 dB PSNR and 0.021 SSIM, confirming the value of observation-space supervision. Using the Gaussian-NLL verification loss K_y yields 28.331 dB PSNR and 0.892 SSIM, outperforming L_1 by 0.144 dB and 0.010, respectively. Unlike a uniform absolute penalty, the NLL evaluates residuals under propagated degradation variance. In Fig. 6, it produces broader blue regions and fewer high-error responses around object textures and structural boundaries, indicating lower redegradation and restoration errors together with clearer recovered details.

5 Conclusion

We presented SIRV, a structured inference-restoration-verification framework for all-in-one TIR image restoration. SIRV infers a structured degradation posterior over degradation states, conditions a shared restorer on these explanations, and verifies observation consistency through coupled forward redegradation during training. This loop improves restoration fidelity and structural preservation under ambiguous compound degradations without adding verification cost at inference. Experiments across in-domain, cross-dataset, and real-world benchmarks demonstrate strong restoration performance and generalization.

AI USE STATEMENT

We used OpenAI ChatGPT to assist with LaTeX formatting, language polishing, literature retrieval and discovery, research ideation or execution, and the development of mathematical proofs. We also used it to prepare submission metadata and draft non-technical disclosure and reproducibility statements. We did not use generative AI tools to generate or preprocess datasets. All AI-assisted material, including literature suggestions, research contributions, and mathematical reasoning, was critically reviewed, checked for correctness, and edited by the authors. The authors take full responsibility for the final content of this work, including all text, claims, proofs, code, and other artifacts.

REPRODUCIBILITY STATEMENT

We support reproducibility by reporting the model architecture, training and evaluation protocols, datasets, preprocessing procedures, hyperparameter settings, and implementation details in the main paper and appendix. The appendix provides additional experimental details and results. An anonymized implementation is provided in the supplementary package and through the anonymous repository referenced in the paper. The released materials include setup instructions, dependency specifications, configuration files, and scripts for training and evaluation. Together, these materials are intended to facilitate independent reproduction of the reported results.

REFERENCES

[1] Y. Ai, H. Huang, X. Zhou, J. Wang, and R. He. Multimodal prompt perceiver: Empower adaptiveness, generalizability and fidelity for all-in-one image restoration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25432–25444. IEEE, 2024.

[2] X. Bai, F. Zhou, and B. Xue. Infrared image enhancement through contrast enhancement by using multiscale new top-hat transform. *Infrared Physics & Technology*, 54(2):61–69, 2011.

[3] F. Bao, S. Jape, A. Schramka, J. Wang, T. E. McGraw, and Z. Jacob. Why thermal images are blurry. *Optics Express*, 32(3):3852–3865, 2024.

[4] L. Cai, X. Dong, K. Zhou, and X. Cao. Exploring video denoising in thermal infrared imaging: Physics-inspired noise generator, dataset, and model. *IEEE Transactions on Image Processing*, 33:3839–3854, 2024.

[5] Y. Cao and C.-L. Tisse. Single-image-based solution for optics temperature-dependent nonuniformity correction in an uncooled long-wave infrared camera. *Optics letters*, 39(3):646–648, 2014.

[6] Y. Chang, L. Yan, T. Wu, and S. Zhong. Remote sensing image stripe noise removal: From image decomposition perspective. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (12):7018–7031, 2016.

[7] Y. Chang, L. Yan, L. Liu, H. Fang, and S. Zhong. Infrared aerothermal nonuniform correction via deep multiscale residual network. *IEEE geoscience and remote sensing letters*, 16(7):1120–1124, 2019.

[8] Z. Chen, Z. Liu, C. Yan, J. Liu, and R. Liu. Hidra: Hierarchical degradation representation and adaptation with generative priors for enhancing infrared vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 37434–37444, 2026.

[9] M. V. Conde, G. Geigle, and R. Timofte. Instructir: High-quality image restoration following human instructions. In *European Conference on Computer Vision*, pages 1–21. Springer, 2024.

[10] D. Ding, Y. Li, P. Zhao, K. Li, S. Jiang, and Y. Liu. Single infrared image stripe removal via residual attention network. *Sensors*, 22(22):8734, 2022.

[11] J. Guan, R. Lai, A. Xiong, Z. Liu, and L. Gu. Fixed pattern noise reduction for infrared images based on cascade residual attention cnn. *Neurocomputing*, 377:301–313, 2020.

[12] Y. Guo, Y. Gao, Y. Lu, H. Zhu, R. W. Liu, and S. He. Onerestore: A universal restoration framework for composite degradation. In *European conference on computer vision*, pages 255–272. Springer, 2024.

[13] H. He, W.-J. Lee, D. Luo, and Y. Cao. Insulator infrared image denoising method based on wavelet generic gaussian distribution and map estimation. *IEEE transactions on industry applications*, 53(4):3279–3284, 2017.

[14] M. D. Hoffman, M. J. Johnson, et al. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in advances in approximate Bayesian inference, NIPS*, volume 1, 2016.

[15] J. Hu, L. Jin, Z. Yao, and Y. Lu. Universal image restoration pre-training via degradation classification. In *International Conference on Learning Representations*, volume 2025, pages 64647–64668, 2025.

[16] X. Hu, S. Luo, C. He, W. Wu, and H. Wu. Infrared thermal image denoising with symmetric multi-scale sampling network. *Infrared Physics & Technology*, 134:104909, 2023.

[17] Y. Huang, Z. Jiang, R. Lan, S. Zhang, and K. Pi. Infrared image super-resolution via transfer learning and psrgan. *IEEE Signal Processing Letters*, 28:982–986, 2021.

[18] H. Jiang, F. Chen, X. Liu, J. Chen, K. Zhang, and L. Chen. Thermal wave image deblurring based on depth residual network. *Infrared Physics & Technology*, 117:103847, 2021.

[19] N. Jiang, Y. Zhang, Q. Li, and F. Yan. An infrared thermal image denoising method focusing on noise feature learning. *Optics & Laser Technology*, 184:112475, 2025.

[20] 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*, pages 340–359. Springer, 2024.

[21] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5128–5137. IEEE, 2021.

[22] K. Ko, K. Shim, K. Lee, and C. Kim. Large-scale benchmark for uncooled infrared image deblurring. *IEEE Sensors Journal*, 23(24):30119–30128, 2023.

[23] X. Kuang, X. Sui, Y. Liu, Q. Chen, and G. Gu. Single infrared image enhancement using a deep convolutional neural network. *Neurocomputing*, 332:119–128, 2019.

[24] S. Kullback. Kullback-leibler divergence. *Encyclopedia of Machine Learning*, pages 581–583, 1951.

[25] K. Lee, Y. Ban, and C. Kim. Motion blur kernel rendering using an inertial sensor: Interpreting the mechanism of a thermal detector. *Sensors*, 22(5):1893, 2022.

[26] A. Li, X. Liu, S. Li, Y. Du, Z. Long, L. Luo, L. Zhang, and C. Zhu. Drnet: All-in-one image restoration via prior-guided dynamic reparameterization. *IEEE Transactions on Multimedia*, pages 1–14, 2026.

[27] 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)*, pages 17431–17441. IEEE, 2022.

[28] P. Li, H. Li, Y. Zhang, Y. Liu, and W. Wang. Breaking degradation coupling: A structural entropy-guided decoupled framework and benchmark for infrared enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1304–1313, 2026.

[29] P. Li, H. Li, Y. Zhang, W. Wang, N. Dong, and J. Wen. Expandable, compressible, mineable: Open-world thermal infrared image restoration. In *Forty-third International Conference on Machine Learning*, 2026.

[30] X. Li, J. Liu, Z. Chen, Y. Zou, L. Ma, X. Fan, and R. Liu. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, pages 270–288. Springer, 2024.

[31] X. Li, Z. Wang, Y. Zou, Z. Chen, J. Ma, Z. Jiang, L. Ma, and J. Liu. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7534–7544. IEEE, 2025.

[32] Z. Li, S. Luo, M. Chen, H. Wu, T. Wang, and L. Cheng. Infrared thermal imaging denoising method based on second-order channel attention mechanism. *Infrared Physics & Technology*, 116:103789, 2021.

[33] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5792–5801. IEEE, 2022.

[34] J. Liu, H. Chen, S. Li, and J. Li. Generation of upright panoramic image from non-upright panoramic image. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5249–5258. IEEE, 2024.

[35] J. Liu, Z. Chen, Z. Liu, Z. Jiang, L. Ma, X. Fan, and R. Liu. Enhancing infrared vision: progressive prompt fusion network and benchmark. *Advances in Neural Information Processing Systems*, 38:96850–96875, 2025.

[36] L. Liu, L. Xu, and H. Fang. Simultaneous intensity bias estimation and stripe noise removal in infrared images using the global and local sparsity constraints. *IEEE Transactions on Geoscience and Remote Sensing*, 58(3):1777–1789, 2019.

[37] L. Liu, L. Xie, X. Zhang, S. Yuan, X. Chen, W. Zhou, H. Li, and Q. Tian. Tape: Task-agnostic prior embedding for image restoration. In *European conference on computer vision*, pages 447–464. Springer, 2022.

[38] Z. Liu, Z. Wang, J. Liu, F. Meng, L. Ma, and R. Liu. Deal: Data-efficient adversarial learning for high-quality infrared imaging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28198–28207. IEEE, 2025.

[39] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. Schön. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations*, volume 2024, pages 16226–16246, 2024.

[40] H. Lv, P. Shan, H. Shi, and L. Zhao. An adaptive bilateral filtering method based on improved convolution kernel used for infrared image enhancement. *Signal, Image and Video Processing*, 16(8):2231–2237, 2022.

[41] E. J. McShane. Jensens inequality. *Bulletin of the American Mathematical Society*, 43(8): 521–527, 1937.

[42] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[43] B. Münch, P. Trtik, F. Marone, and M. Stampanoni. Stripe and ring artifact removal with combined waveletfourier filtering. *Optics express*, 17(10):8567–8591, 2009.

[44] Z. Pang, G. Liu, G. Li, J. Gong, C. Chen, and C. Yao. An infrared image enhancement method via content and detail two-stream deep convolutional neural network. *Infrared Physics & Technology*, 132:104761, 2023.

[45] V. Potlapalli, S. W. Zamir, S. H. Khan, and F. Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in neural information processing systems*, 36:71275–71293, 2023.

[46] J. Qiu, Z. Wang, Y. Huang, and H. Huang. Infrared image dynamic range compression based on adaptive contrast adjustment and structure preservation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

[47] M. Raschke. Empirical behaviour of tests for the beta distribution and their application in environmental research. *Stochastic Environmental Research and Risk Assessment*, 25(1):79–89, 2011.

[48] Y. Shao, Y. Sun, M. Zhao, Y. Chang, Z. Zheng, C. Tian, and Y. Zhang. Infrared image stripe noise removing using least squares and gradient domain guided filtering. *Infrared Physics & Technology*, 119:103968, 2021.

[49] S. S. Shapiro and H. Zahedi. Bernoulli trials and discrete distributions. *Journal of Quality Technology*, 22(3):193–205, 1990.

[50] H. Talebi and P. Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018.

[51] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos. The variational approximation for bayesian inference. *IEEE Signal Processing Magazine*, 25(6):131–146, 2008.

[52] N. Varghese, M. M. MR, and A. Rajagopalan. Fast motion-deblurring of ir images. *IEEE Signal Processing Letters*, 29:459–463, 2022.

[53] R. Varghese and M. Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*, pages 1–6. IEEE, 2024.

[54] L. Wang, J. Zhou, S. Liao, S. Li, and Y. Wang. Infrared multi-detector non-uniformity images simulation method. *Infrared Physics & Technology*, 148:105829, 2025.

[55] T. Ye, Q. Wu, J. Deng, G. Liu, L. Liu, S. Xia, L. Pang, W. Yu, and L. Pei. Thermal-nerf: Neural radiance fields from an infrared camera. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1046–1053. IEEE, 2024.

[56] S. Yi, L. Li, X. Liu, J. Li, and L. Chen. Hctirdeblur: A hybrid convolution-transformer network for single infrared image deblurring. *Infrared Physics & Technology*, 131:104640, 2023.

[57] G. Yue, L. Zhang, J. Zhang, Z. Xu, S. Wang, T. Zhou, Y. Gong, and W. Zhou. Subjective quality assessment of thermal infrared images. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 1212–1217. IEEE, 2024.

[58] 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12753–12763. IEEE, 2025.

[59] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[60] J. Zhang, X. Zhou, L. Li, T. Hu, and C. Fansheng. A combined stripe noise removal and deblurring recovering method for thermal infrared remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[61] J. Zhang, J. Huang, M. Yao, Z. Yang, H. Yu, M. Zhou, and F. Zhao. Ingredient-oriented multi-degradation learning for image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5825–5835. IEEE, 2023.

[62] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[63] D. Zheng, X.-M. Wu, S. Yang, J. Zhang, J.-F. Hu, and W.-S. Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25445–25455. IEEE, 2024.

Supplementary Material

A Conditional ELBO Derivation

A.1 Supervised conditional evidence bound

For each supervised training sample i, the known generating annotations define the sample-conditioned reference distribution $p_i(z)$, where $z = (r, \theta, x)$. Together with the forward model $p(y_i \mid z)$, it induces the conditional evidence $p_i(y_i)$. This construction is used to derive a supervised conditional variational objective. Let $\theta = (\theta^{\text{cont}}, \theta^{\text{disc}})$ denote the continuous and discrete degradation parameters. The sample-conditioned evidence is

$$p_i(y_i) = \sum_{r \in \mathcal{R}} \sum_{\theta^{\text{disc}} \in \Theta_r^{\text{disc}}} \int_{\Theta_r^{\text{cont}}} \int_{\mathcal{X}} p(y_i \mid x, r, \theta)\, p_i(r, \theta, x)\, dx\, d\theta^{\text{cont}} = \mathbb{E}_{p_i(z)} \left[p(y_i \mid z) \right], \tag{5}$$

where $\mathcal{R}$ contains all admissible ordered degradation sequences, and $\mathcal{X}$ denotes the clean image space. The summation over sequences and the integrals over degradation parameters and clean images make direct evaluation intractable. We therefore introduce an amortized variational distribution $q_{\phi,\psi}(z \mid y_i)$ whose support covers that of $p(y_i \mid z)p_i(z)$. Multiplying and dividing by this distribution gives

$$\log p_i(y_i) = \log \mathbb{E}_{q_{\phi,\psi}(z|y_i)} \left[\frac{p(y_i \mid z) p_i(z)}{q_{\phi,\psi}(z \mid y_i)} \right]. \tag{6}$$

Since $\log(\cdot)$ is concave, Jensen's inequality [41] moves the logarithm inside the expectation:

$$\log p_i(y_i) \geq \mathbb{E}_{q_{\phi,\psi}(z|y_i)} \left[\log p(y_i \mid z) + \log p_i(z) - \log q_{\phi,\psi}(z \mid y_i) \right] \equiv \mathcal{L}_{\text{cELBO}}(y_i). \tag{7}$$

Grouping the likelihood contribution and the expected log-density ratio gives exactly the compact cELBO [14] in Eq. 2 of the main paper. Here, D_{KL} is the Kullback-Leibler divergence [24]. The likelihood rewards latent explanations that reproduce the observation, whereas the KL term keeps the inferred latent state close to the sample-conditioned reference prior. The tightness of the bound follows directly from Bayes' rule, $p_i(z \mid y_i) = p(y_i \mid z)p_i(z)/p_i(y_i)$. Substitution into the KL divergence gives

$$\begin{aligned} D_{\text{KL}}[q_{\phi,\psi}(z \mid y_i) \| p_i(z \mid y_i)] &= \mathbb{E}_{q_{\phi,\psi}(z|y_i)} \left[\log \frac{q_{\phi,\psi}(z \mid y_i) p_i(y_i)}{p(y_i \mid z) p_i(z)} \right] \\ &= \log p_i(y_i) - \mathcal{L}_{\text{cELBO}}(y_i). \end{aligned} \tag{8}$$

Consequently,

$$\log p_i(y_i) = \mathcal{L}_{\text{cELBO}}(y_i) + D_{\text{KL}}[q_{\phi,\psi}(z \mid y_i) \| p_i(z \mid y_i)] \geq \mathcal{L}_{\text{cELBO}}(y_i). \tag{9}$$

Maximizing the cELBO therefore increases a tractable lower bound on the sample-conditioned log evidence while reducing the gap between the approximate and reference posteriors. The bound supplies the theoretical organization of SIRV's training terms.

A.2 Factorization and practical estimation

Substituting the posterior and prior factorizations into Eq. 7 yields the four-term decomposition in Eq. 3 of the main paper. We denote the practical sequence, parameter, clean-image, and observation-consistency losses by K_r, K_θ, K_x, and K_y, respectively. The first three are KL terms, whereas K_y is a Gaussian negative log-likelihood. Their component-specific definitions are provided below. Their closed forms are derived with the corresponding components in Appendices B and C.1, and the observation term is instantiated by the Gaussian forward model in Appendix D. The ideal expectations would evaluate restoration under every sequence and parameter state. The practical estimator computes K_r exactly over all sequence probabilities, but evaluates K_θ, K_x, and the observation likelihood on the supervised branch $r_i^\star$, using one reparameterized sample $\widetilde{\theta}_i$ and restoration mean $\widetilde{x}_i$. This branch-truncated estimator preserves the four principal terms without exhaustively running the restorer over $\mathcal{R}$. Consequently, the verification term directly updates the parameter-inference and restoration networks on $r_i^\star$, whereas the sequence logits are optimized by the complete K_r term.

Table 10: Encoder structures used in the posterior inference network.

Encoder	Structure	Output
Shared	3×3 Conv-GroupNorm-GELU, residual stages with 1×1 channel projections, 3×3 Conv-GroupNorm-GELU, and GeM pooling	F_{sh}
Contrast	5×5 Conv-GroupNorm-GELU, three residual blocks, channel attention, and two output layers	$A_C,\ g_C$
Blur	5×5 Conv-GroupNorm-GELU, three channel-attention residual blocks, and two output layers	$A_B,\ g_B$
Optical	7×7 Conv-GroupNorm-GELU, three residual blocks, channel attention, and two output layers	$A_O,\ g_O$
Stripe	3×3 Conv-GroupNorm-GELU, parallel 1×7 and 7×1 depthwise convolutions, concatenation followed by 1×1 Conv-GroupNorm-GELU, three residual blocks, and two output layers	$A_S,\ g_S$
Gaussian	3×3 Conv-GroupNorm-GELU, three residual blocks, and two output layers	$A_G,\ g_G$

B Structured Degradation Inference

B.1 Architecture of Posterior Inference Network

Given y, the shared encoder produces a global context descriptor $F_{\mathrm{sh}} \in \mathbb{R}^{64}$ through a 3×3 Conv-GroupNorm-GELU projection followed by learnable generalized-mean (GeM) pooling. Each specialized encoder has two parallel output layers: a 3×3 convolution with 64 input channels and one output channel, followed by sigmoid activation, produces the spatial evidence map $A_j \in [0,1]^{1 \times H \times W}$, and GELU followed by learnable GeM pooling produces the global descriptor $g_j \in \mathbb{R}^{64}$, where $j \in \{C, B, O, S, G\}$. The five specialized encoders use the same output-layer architectures with separate learnable parameters. Table 10 summarizes their structures. The contrast and optical encoders use spatial convolutions and channel attention to characterize contrast variation and spatial non-uniformity, respectively. The blur and Gaussian-noise encoders supplement the image with local residual cues to capture edge attenuation and noise fluctuations, respectively, while the stripe encoder uses horizontal and vertical depthwise convolutions to capture directional stripe patterns. These complementary encoders separate global process context from degradation-specific spatial evidence. The sequence predictor concatenates $[F_{\mathrm{sh}}, g_C, g_B, g_O, g_S, g_G]$ and uses an MLP to produce $|\mathcal{R}|$ sequence logits. For a candidate r, each parameter predictor concatenates F_{sh}, the corresponding descriptor g_j, yielding a 144-dimensional input. The five two-hidden-layer MLPs jointly output nine continuous parameter logits, six blur-kernel logits, and two binary logits for each candidate sequence. These predicted parameters, together with explicit $C/B/N$ presence indicators, form part of the restoration condition, and K_θ is evaluated only for variables defined by the supervised sequence.

B.2 Sequence Posterior

Let $\ell_r(y)$ denote the logit predicted for sequence r. Following Section 3.3, the sequence posterior is

$$q_\phi(r \mid y) = \mathrm{Cat}(r \mid \pi_\phi(y)), \qquad \pi_r = \frac{\exp(\ell_r(y))}{\sum_{s \in \mathcal{R}} \exp(\ell_s(y))}. \tag{10}$$

For brevity, we write $\pi_r = \pi_{\phi,r}(y)$ when the input is clear. For sample i, the probabilities π_r below are evaluated from y_i. The reference sequence prior is a smoothed Categorical distribution centered on the known generating sequence $r_i^\star$:

$$p_i(r) = \frac{\epsilon_r}{|\mathcal{R}|} + (1 - \epsilon_r)\mathbf{1}[r = r_i^\star], \quad \epsilon_r = 0.01. \tag{11}$$

The uniform smoothing gives every sequence non-zero mass and prevents an infinite KL when the posterior assigns probability to a non-target sequence. Substituting Eqs. 10 and 11 into the definition

of KL divergence gives

$$K_r = \mathbb{E}_{q_\phi(r|y_i)}\left[\log \frac{q_\phi(r \mid y_i)}{p_i(r)}\right] = \sum_{r\in\mathcal{R}} \pi_r \log \frac{\pi_r}{p_i(r)}. \tag{12}$$

This term is evaluated exactly over the complete sequence posterior rather than on a single selected branch.

B.3 Continuous Parameter Posterior

Table 11 summarizes the continuous variables. Each continuous parameter is represented in normalized coordinates on $[0, 1]$ for posterior modeling. For an active parameter $\theta_t \in [L_t, U_t]$, we define its normalized value $\xi_t = (\theta_t - L_t)/(U_t - L_t)$. This notation avoids overloading the full latent state $z = (r, \theta, x)$ used in the main paper. The parameter predictor predicts a normalized mode $m_t = \mathrm{sigmoid}(h_t(y, r))$ and constructs

$$\begin{aligned} \alpha_{q,t} &= 1 + \kappa_q m_t, & \beta_{q,t} &= 1 + \kappa_q(1 - m_t), \\ q_{\phi,t}(\xi_t \mid y, r) = \mathrm{Beta}(\xi_t \mid \alpha_{q,t}, \beta_{q,t}), & \theta_t &= L_t + (U_t - L_t)\xi_t, \end{aligned} \qquad \kappa_q = 20. \tag{13}$$

The Beta density is

$$\mathrm{Beta}(\xi \mid a, b) = \frac{\xi^{a-1}(1-\xi)^{b-1}}{B(a, b)}, \quad 0 < \xi < 1, \tag{14}$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$. Because $\alpha_{q,t}, \beta_{q,t} > 1$, the posterior mode and mean are

$$\mathrm{mode}(\xi_t) = m_t, \quad \mathbb{E}[\xi_t] = \frac{1 + \kappa_q m_t}{2 + \kappa_q}. \tag{15}$$

Thus, the network output is the mode rather than the exact posterior mean. Fixing $\kappa_q = 20$ gives a constant total concentration $\alpha_{q,t} + \beta_{q,t} = 22$. The variance still depends on the predicted mode, but concentration is not learned independently for each sample. Sample-dependent ambiguity is represented primarily by the sequence posterior, while the scaled Beta factors provide bounded stochastic parameter estimates.

Let $\xi^\star_{i,t}$ be the normalized generating value for training sample i. Its reference prior uses the same family,

$$\begin{aligned} &p_{i,t}(\xi_t \mid r^\star_i) = \mathrm{Beta}\left(\xi_t \mid \alpha_{p,t}, \beta_{p,t}\right), \\ &\alpha_{p,t} = 1 + \kappa_p \xi^\star_{i,t}, \quad \beta_{p,t} = 1 + \kappa_p(1 - \xi^\star_{i,t}), \quad \kappa_p = 20. \end{aligned} \tag{16}$$

For $q = \mathrm{Beta}(a_q, b_q)$ and $p = \mathrm{Beta}(a_p, b_p)$, the density ratio is

$$\log \frac{q(\xi)}{p(\xi)} = \log \frac{B(a_p, b_p)}{B(a_q, b_q)} + (a_q - a_p)\log \xi + (b_q - b_p)\log(1 - \xi), \tag{17}$$

where $B(\cdot, \cdot)$ is the Beta function. Using $\mathbb{E}_q[\log \xi] = \mathrm{digamma}(a_q) - \mathrm{digamma}(a_q + b_q)$ and $\mathbb{E}_q[\log(1 - \xi)] = \mathrm{digamma}(b_q) - \mathrm{digamma}(a_q + b_q)$ gives the closed-form KL divergence, where digamma denotes the digamma function

$$\begin{aligned} D_{\mathrm{KL}}(q\|p) = \log \frac{B(a_p, b_p)}{B(a_q, b_q)} &+ (a_q - a_p)\,\mathrm{digamma}(a_q) + (b_q - b_p)\,\mathrm{digamma}(b_q) \\ &+ (a_p + b_p - a_q - b_q)\,\mathrm{digamma}(a_q + b_q). \end{aligned} \tag{18}$$

The prior and posterior share the same affine transformation from ξ_t to θ_t, so their Jacobians cancel in the density ratio. The continuous part of the parameter KL is therefore computed in normalized space:

$$K_{\mathrm{Beta}} = \sum_{t \in \mathcal{I}^{\mathrm{cont}}_{r^\star_i}} D_{\mathrm{KL}}[q_{\phi,t}(\xi_t \mid y_i, r^\star_i) \| p_{i,t}(\xi_t \mid r^\star_i)]\,. \tag{19}$$

Table 11: Continuous degradation parameters.

Family	Parameter and symbol	Physical range
Contrast	intensity scale α_C	[0.2, 0.6]
Contrast	intensity offset β_C	[0.1, 0.2]
Blur	Gaussian width σ_B	[1, 2]
Optical	field amplitude δ_O	[15/255, 55/255]
Optical	vertical center c_h	[0, 1]
Optical	horizontal center c_w	[0, 1]
Stripe	multiplicative std. σ_g	[0.03, 0.07]
Stripe	additive std. σ_b	[0, 3/255]
Gaussian	noise std. σ_n	[5/255, 15/255]

Table 12: Discrete degradation parameters.

Family	Parameter	Support	Posterior
Blur	kernel size k_B	{7, 9, 11, 13, 15, 17}	Categorical
Optical	field polarity s_O	{0, 1}	Bernoulli
Stripe	orientation o_S	{0, 1}	Bernoulli

B.4 Discrete Parameter Posterior

For a sequence containing blur, the kernel-size posterior is

$$q_{\phi,k_B}(k_B = k_c \mid y, r) = \omega_c(y, r), \quad k_c \in \{7, 9, 11, 13, 15, 17\}, \tag{20}$$

where $c \in \{1, \dots, 6\}$ indexes the six kernel sizes in Table 12. During coupled sampling, the implementation uses hard Gumbel-Softmax with temperature 1, which gives a discrete forward value and a straight-through gradient. The smoothed target prior is

$$p_i(k_B = k_c \mid r_i^\star) = \frac{\epsilon_k}{6} + (1 - \epsilon_k)\mathbf{1}[c = c_i^\star], \quad \epsilon_k = 0.01, \tag{21}$$

where $c_i^\star$ is the generating kernel-size category. Its KL is

$$K_{k_B} = \sum_{c=1}^{6} \omega_c(y_i, r_i^\star) \log \frac{\omega_c(y_i, r_i^\star)}{p_i(k_B = k_c \mid r_i^\star)}. \tag{22}$$

The optical polarity and stripe orientation use Bernoulli posteriors. For either binary variable $b \in \{s_O, o_S\}$,

$$q_{\phi,b}(b \mid y, r) = \rho_{q,b}^{b}(1 - \rho_{q,b})^{1-b}, \quad \rho_{q,b} = \mathrm{sigmoid}(h_b(y, r)). \tag{23}$$

Training uses a hard binary realization obtained by thresholding $\rho_{q,b}$ at 0.5, with a straight-through gradient. Blind inference instead supplies $\rho_{q,b}$ to the restoration condition. For target $b_i^\star \in \{0, 1\}$, the smoothed prior probability is $\rho_{p,b} = \epsilon_b + (1 - 2\epsilon_b)b_i^\star$, where $\epsilon_b = 0.01$. Summing the two Bernoulli outcomes yields

$$K_b = \rho_{q,b} \log \frac{\rho_{q,b}}{\rho_{p,b}} + (1 - \rho_{q,b}) \log \frac{1 - \rho_{q,b}}{1 - \rho_{p,b}}. \tag{24}$$

Combining the active continuous and discrete factors gives the practical degradation-parameter term

$$K_\theta = K_{\mathrm{Beta}} + \mathbf{1}[B \in r_i^\star]K_{k_B} + \mathbf{1}[N \in r_i^\star]\left(K_{s_O} + K_{o_S}\right). \tag{25}$$

Hence, variables not defined by $r_i^\star$ do not contribute to K_θ. Explicit $C/B/N$ presence indicators accompany the predicted parameter representation so that the restorer can interpret the condition according to the candidate sequence.

Table 13: Shared restorer configuration.

Stage	Resolution	Channels	Main blocks
Encoder 1	$H \times W$	48	2 conditional residual blocks
Encoder 2	$H/2 \times W/2$	96	2 conditional residual blocks
Bottleneck	$H/4 \times W/4$	144	3 Restormer blocks
Decoder	multiscale	96/48	mirrored conditional blocks

C Posterior-Conditioned Restoration

For each candidate sequence r, the implementation constructs a 42-channel condition $c_{r,\theta}$. Its parameter channels contain nine continuous factors, six blur-kernel entries, and two binary entries. Coupled posterior samples are used during training, whereas continuous posterior modes and discrete posterior probabilities are used during blind inference. The remaining channels contain three $C/B/N$ presence indicators, a 16-dimensional sequence embedding, five sequence-masked evidence maps, and the sequence probability π_r. Global quantities are broadcast spatially before concatenation. The explicit presence indicators allow the shared restorer to interpret the parameter representation according to the candidate composition, while the evidence maps retain location-dependent degradation information.

The shared restorer is a three-scale encoder-decoder with conditional residual blocks and a Restormer bottleneck [59]. Its principal configuration is summarized in Table 13. At each scale, spatial feature transformation (SFT) modulates a restoration feature F using the resized condition c:

$$\mathrm{SFT}(F, c) = \gamma(c) \odot F + \beta(c), \tag{26}$$

where γ is produced by a Sigmoid-gated convolutional projection and β is an additive convolutional projection. The same condition is concatenated with the input at the restoration head and injected into the multiscale residual and bottleneck blocks. The restorer predicts a residual image and produces

$$\mu_\psi(y, c_{r,\theta}) = \mathrm{clip}\left(y + \Delta_\psi(y, c_{r,\theta}), 0, 1\right). \tag{27}$$

This mean defines the Gaussian clean-image posterior in the main paper. At inference, SIRV retains the k most probable sequence branches. For each retained branch, the inference-time parameter representation uses the modes of the continuous scaled-Beta factors and the posterior probabilities of the discrete factors. The branch outputs are combined using renormalized sequence probabilities. We use $k = 2$ by default, and other values are evaluated only in the dedicated top-k ablation.

C.1 Clean-image Posterior

For training sample i, the clean-image posterior and its sample-conditioned reference prior are

$$\begin{aligned} q_\psi(x \mid y_i, r, \theta) &= \mathcal{N}\big(x \mid \mu_\psi(y_i, c_{r,\theta}), \varepsilon_x^2 I\big), \\ p_i(x) &= \mathcal{N}\big(x \mid x_i^\star, \varepsilon_x^2 I\big), \end{aligned} \tag{28}$$

where $x_i^\star$ is the clean target and I is the identity matrix for the vectorized image. For two P-dimensional Gaussians $q = \mathcal{N}(\mu_q, \Sigma_q)$ and $p = \mathcal{N}(\mu_p, \Sigma_p)$,

$$D_{\mathrm{KL}}(q\|p) = \frac{1}{2}\left[\mathrm{tr}(\Sigma_p^{-1}\Sigma_q) + (\mu_p - \mu_q)^\top \Sigma_p^{-1}(\mu_p - \mu_q) - P + \log\frac{\det \Sigma_p}{\det \Sigma_q}\right]. \tag{29}$$

Substituting $\Sigma_q = \Sigma_p = \varepsilon_x^2 I$ cancels the trace, dimension, and log-determinant terms. On the supervised branch, let $\widetilde{x}_i$ denote the restoration mean obtained using the shared parameter realization. The clean-image KL is

$$K_x = \frac{\|\widetilde{x}_i - x_i^\star\|_2^2}{2\varepsilon_x^2}, \qquad \varepsilon_x^2 = 10^{-5}. \tag{30}$$

For optimization, the pixelwise contributions are averaged over the image. The implementation uses the posterior mean $\widetilde{x}_i$ rather than sampling a clean image.

D DEGRADATION-CONSISTENT VERIFICATION

The differentiable forward redegradation model is a plug-in diagonal-Gaussian approximation to the offline degradation generator. It initializes the moments as $\mu^{(0)} = x$ and $v^{(0)} = \tau_y^2\mathbf{1}$, with $\tau_y^2 = 10^{-4}$, and propagates them through the operators in the order specified by a degradation sequence r. Table 14 lists the principal updates. K_B is the normalized Gaussian blur kernel, and f_O is the unit-amplitude optical field determined by (c_h, c_w) and s_O, and δ_O scales its amplitude. After the complete sequence, it imposes

$$\mu_{\mathcal{U}}(x, r, \theta) = \mu^{(L)}, \quad v_{\mathcal{U}}(x, r, \theta) = \max\left(v^{(L)}, \tau_y^2\mathbf{1}\right), \quad \tau_y^2 = 10^{-4}. \tag{31}$$

After each operator update, the propagated mean is clipped to $[0, 1]$. The differentiable implementation mirrors the quantization used by the offline generator through a straight-through approximation, while the propagated variance is floored after the complete sequence. These plug-in updates retain marginal variances only. Spatial covariances, the multiplicative interaction term $\sigma_g^2 v$, and moment corrections for clipping and quantization are omitted.

The positive initialization and final floor keep the likelihood well defined even for sequences containing only contrast and blur. The observation model is

$$p(y \mid x, r, \theta) = \mathcal{N}(y \mid \mu_{\mathcal{U}}(x, r, \theta), \mathrm{Diag}(v_{\mathcal{U}}(x, r, \theta))) . \tag{32}$$

The propagated variance represents the conditional observation variance introduced by the modeled stochastic degradations and the numerical variance floor. It is neither an independently predicted uncertainty map nor a marginalization over the degradation-parameter posterior.

For training sample i, we use the supervised sequence $r_i^\star = (r_{i,1}^\star, \ldots, r_{i,L_i}^\star)$. We construct $\widetilde{\theta}_i$ from continuous posterior samples and straight-through discrete realizations on the supervised branch. The same realization yields $\widetilde{x}_i = \mu_\psi(y_i, c_{r_i^\star, \widetilde{\theta}_i})$ and is reused by the forward model. Following the main paper, the forward model computes

$$\begin{aligned}
(\mu_i^{(0)}, v_i^{(0)}) &= (\widetilde{x}_i, \tau_y^2\mathbf{1}), \\
(\mu_i^{(l)}, v_i^{(l)}) &= \mathcal{U}_{r_{i,l}^\star}\left(\mu_i^{(l-1)}, v_i^{(l-1)}; \widetilde{\theta}_{i, r_{i,l}^\star}\right), \quad l = 1, \ldots, L_i, \\
\widetilde{\mu}_{y,i} &= \mu_i^{(L_i)}, \\
\widetilde{v}_{y,i} &= \max\left(v_i^{(L_i)}, \tau_y^2\mathbf{1}\right),
\end{aligned} \tag{33}$$

where L_i is the sequence length, and $\widetilde{\theta}_{i, r_{i,l}^\star}$ denotes the parameters of the l-th operator extracted from the same realization $\widetilde{\theta}_i$. The variance floor is $\tau_y^2 = 10^{-4}$. The observation-consistency loss follows the Gaussian negative log-likelihood defined in the main paper:

$$K_y = -\log \mathcal{N}(y_i \mid \widetilde{\mu}_{y,i}, \mathrm{Diag}(\widetilde{v}_{y,i})) . \tag{34}$$

Each pixel contributes a variance-weighted squared residual and a log-variance penalty. These contributions are averaged over pixels during optimization; the Gaussian constant, which is independent of the network parameters, may be omitted.

Reusing $\widetilde{\theta}_i$ ensures that restoration and forward redegradation evaluate the same parameter explanation. The supervised label $r_i^\star$ fixes the discrete branch used by the practical estimator, so no gradient passes through branch selection. The shared realization $\widetilde{\theta}_i$ allows K_y to constrain parameter inference and restoration jointly. Although no gradient passes through the ground-truth branch selection, K_y can update sequence logits through the softmax probability $\pi_{r_i^\star}$ in the restoration condition. K_r provides direct supervision over all sequence categories. The forward redegradation branch is used only during training and is removed at inference.

E TRAINING AND BLIND INFERENCE DETAILS

SIRV minimizes

$$\mathcal{L}_{\mathrm{total}} = K_r + K_\theta + K_x + K_y, \tag{35}$$

Table 14: Moment updates in the differentiable redegradation forward model.

Operator	Mean update	Variance update
Contrast	$\mu' = \alpha_C \mu + \beta_C$	$v' = \alpha_C^2 v$
Blur	$\mu' = K_B * \mu$	$v' = K_B^{\odot 2} * v$
Noise	$\mu' = \mu + \delta_O f_O$	$v' = v + (\mu')^2 \sigma_g^2 + \sigma_b^2 + \sigma_n^2$

Table 15: Training and inference configuration.

Setting	Value
Training patch	256×256
Epochs	200
Global batch size	4
Optimizer	AdamW
Initial learning rate	2×10^{-4}
Minimum learning rate	10^{-6}
Weight decay	10^{-5}
Learning-rate schedule	cosine annealing per iteration
Loss weights $(K_r, K_\theta, K_x, K_y)$	(1.0, 1.0, 1.0, 1.0)
Auxiliary-loss weight	0
Clean-posterior variance	$\varepsilon_x^2 = 10^{-5}$
Observation variance floor	$\tau_y^2 = 10^{-4}$
Default inference branches	top-2
Input range	[0, 1]

with unit weight for every term. The sequence KL sums over all sequence categories before batch averaging during training. The parameter KL sums the variables defined by $r_i^\star$, and K_x and K_y are averaged over $P = HW$ pixels and then over the mini-batch. The same loss symbols denote these reduced quantities in the implemented training objective. Ground-truth annotations construct the reference priors and select the supervised branches used by K_θ, K_x, and K_y, but are never inputs to the posterior encoders or blind inference. A single AdamW optimizer updates the inference and restoration networks jointly.

At test time, SIRV first predicts $q_\phi(r \mid y)$, retains the top-2 most probable sequences, obtains one parameter estimate for each branch, restores both branches with the shared network, and returns their posterior-weighted average. No clean target, degradation label, or forward verification is required.

E.1 Training and inference algorithms

Algorithm 1 gives the practical training procedure for a mini-batch of size N_b. The same posterior parameter sample $\widetilde{\theta}_i$ conditions restoration and is reapplied by the forward model, ensuring that both computations evaluate one explanation. Blind inference is summarized in Algorithm 2, and it evaluates only inference and restoration, so the forward verifier introduces no test-time computation.

E.2 Code execution

The released implementation follows four stages: data preparation, training, blind inference, and evaluation. Data preparation writes the source split, ordered sequence, active parameters, and random seed to a manifest. Training reads this manifest and the configuration in Table 15. Blind inference loads the trained checkpoint, uses $k = 2$ unless overridden for ablation, and saves restored images together with sequence probabilities and parameter estimates. Evaluation then reads the restored images and manifest to compute the required grouped metrics. Code is available at https://anonymous.4open.science/r/SIRV-4FD2.

Algorithm 1 End-to-End Training of SIRV

Input: Training set $\mathcal{D} = \{(y_i, x_i^\star, r_i^\star, \theta_i^\star)\}_{i=1}^N$; number of epochs T, batch size N_b, image size $H \times W$ with $P = HW$, ordered forward model $\mathcal{U}$, networks parameterized by ϕ, ψ
Output: networks with trained parameters ϕ, ψ

for $t = 1, \dots, T$ **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
    for all $(y_i, x_i^\star, r_i^\star, \theta_i^\star) \in \mathcal{B}$ **do**
      Extract the shared feature F_{sh} and specialized features $\{A_j, g_j\}_{j \in \{C,B,O,S,G\}}$ from y_i
      Infer the sequence posterior $q_\phi(r \mid y_i) = \mathrm{Cat}(r \mid \pi_\phi(y_i))$ and parameter posteriors $q_\phi(\theta \mid y_i, r)$
      Construct the reference priors $p_i(r)$, $p_i(\theta \mid r)$, and $p_i(x)$ from $(r_i^\star, \theta_i^\star, x_i^\star)$
      Compute the sequence and supervised-branch parameter terms

$$K_r^{(i)} = D_{\mathrm{KL}}[q_\phi(r \mid y_i) \,\|\, p_i(r)]$$

$$K_\theta^{(i)} = D_{\mathrm{KL}}[q_\phi(\theta \mid y_i, r_i^\star) \,\|\, p_i(\theta \mid r_i^\star)]$$

      Construct $\widetilde{\theta}_i$ using continuous posterior samples and straight-through discrete realizations on $r_i^\star$, and build $c_{r_i^\star, \widetilde{\theta}_i}$
      Obtain the restoration posterior mean

$$\widetilde{x}_i = \mu_\psi\left(y_i, c_{r_i^\star, \widetilde{\theta}_i}\right)$$

      Compute the clean-image term

$$K_x^{(i)} = \frac{\|\widetilde{x}_i - x_i^\star\|_2^2}{2\varepsilon_x^2}$$

      Obtain $(\widetilde{\mu}_{y,i}, \widetilde{v}_{y,i})$ using the sequential updates and variance floor in Eq. 33
      Compute the observation-consistency term

$$K_y^{(i)} = -\log \mathcal{N}(y_i \mid \widetilde{\mu}_{y,i}, \mathrm{Diag}(\widetilde{v}_{y,i}))$$

      Apply pixel averaging to the contributions of $K_x^{(i)}$ and $K_y^{(i)}$
    end for
    Form the mini-batch objective

$$\mathcal{J}_\mathcal{B} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left(K_r^{(i)} + K_\theta^{(i)} + K_x^{(i)} + K_y^{(i)}\right)$$

    Update (ϕ, ψ) by minimizing $\mathcal{J}_\mathcal{B}$ with AdamW
  end for
end for
return networks parameterized by (ϕ, ψ)

F Data Construction and Evaluation Protocol

F.1 Datasets and Degradation Construction

HM-TIR [35] contains 1,503 thermal images and follows an 8:2 source-image split for training and testing. All trainable restoration methods, including the compared baselines, are retrained on the HM-TIR training split. The resulting checkpoints are evaluated without target-dataset fine-tuning, while non-learning baselines use the same degraded inputs. Cross-dataset evaluation uses 647 Night-TIR images [28] and 300 infrared images from M^3FD [33]. TIIQAD [57] contains 1,000 authentically distorted thermal images without paired clean references and is evaluated without additional synthetic degradation.

Algorithm 2 Blind Inference of SIRV

Input: Degraded observation y, trained networks with parameters ϕ, ψ, number of retained branches k

Output: Restored image $\widehat{x}$

1: Extract F_{sh} and $\{A_j, g_j\}_{j\in\{C,B,O,S,G\}}$ from y

2: Infer the ordered sequence posterior

$$q_\phi(r \mid y) = \mathrm{Cat}(r \mid \pi_\phi(y)) .$$

3: Select the k most probable sequences:

$$\mathcal{R}_k = \text{Top-k}\left(\{\pi_r\}_{r\in\mathcal{R}}, k\right)$$

4: **for all** $r \in \mathcal{R}_k$ **do**

5: Infer the branch-conditioned parameter posterior $q_\phi(\theta \mid y, r)$

6: Obtain the plug-in estimate $\widehat{\theta}_r$ using the continuous posterior modes and discrete posterior probabilities

7: Mask evidence maps for degradation families absent from r

8: Construct $c_{r,\widehat{\theta}_r}$ by concatenating the masked maps with spatially broadcast π_r, normalized paramete representations, and $C/B/N$ presence indicators.

9: Restore the candidate branch using the shared network:

$$\widehat{x}_r = \mu_\psi\Big(y, c_{r,\widehat{\theta}_r}\Big)$$

10: **end for**

11: Normalize the retained sequence probabilities:

$$\bar{\pi}_r = \frac{\pi_r}{\sum_{s\in\mathcal{R}_k}\pi_s}, \quad r \in \mathcal{R}_k$$

12: Compute the posterior-weighted restoration:

$$\widehat{x} = \sum_{r\in\mathcal{R}_k} \bar{\pi}_r\, \widehat{x}_r$$

13: **return** $\widehat{x}$

Source-image partitions are fixed before synthesis. For the standard paired benchmarks, each source image produces one degraded observation. Single, double, and triple degradations are selected with probabilities of 20%, 30%, and 50%, respectively, followed by an admissible sequence of the selected length and its degradation parameters. Each clean source and its degraded observation remain in the same partition, preventing source-image overlap between training and evaluation. A manifest records the source identifier, ordered sequence, active parameters, and random seed.

The generator operates on normalized grayscale images. After each active $C/B/N$ operator, the output is clipped and quantized:

$$Q(a) = \frac{\mathrm{round}(255\,\mathrm{clip}(a, 0, 1))}{255}. \tag{36}$$

Low contrast is generated by $\mathcal{D}_C(a) = Q(\alpha_C a + \beta_C)$. Gaussian blur uses a normalized kernel $K_B(k_B, \sigma_B)$ with reflection padding, giving $\mathcal{D}_B(a) = Q(K_B * a)$.

Compound sensor noise combines optical non-uniformity, stripe interference, and independent Gaussian noise. For normalized spatial coordinates $(u, v) \in [0, 1]^2$, the optical field is defined by

$$\begin{aligned} d(u,v) &= \frac{\sqrt{(u-c_h)^2 + (v-c_w)^2}}{\max_{u',v'}\sqrt{(u'-c_h)^2 + (v'-c_w)^2}}, \\ f(u,v) &= \cos^4\left(\frac{\pi d(u,v)}{2}\right), \qquad f_O = (1-s_O)f + s_O(1-f). \end{aligned} \tag{37}$$

The optical transformation is $\mathcal{O}(a) = a + \delta_O f_O$. Stripe interference is $\mathcal{S}(a) = a \odot (1 + \epsilon_g) + \epsilon_b$, where Gaussian gain and bias perturbations with standard deviations σ_g and σ_b are broadcast along

Table 16: Parameter ranges used to construct the weak and strong degradation settings.

Degradation	Parameter	Weak	Strong
Low Contrast	Scale α_C	[0.50, 0.60]	[0.20, 0.30]
	Offset β_C	[0.10, 0.20]	[0.10, 0.20]
Blur	Kernel size k_B	{7, 9}	{15, 17}
	Kernel scale σ_B	[1.00, 1.25]	[1.75, 2.00]
Noise	Optical amplitude δ_O	[15, 25]/255	[45, 55]/255
	Stripe-gain std. σ_g	[0.030, 0.040]	[0.060, 0.070]
	Stripe-bias std. σ_b	[0, 0.75]/255	[2.25, 3.00]/255
	Gaussian-noise std. σ_n	[5.0, 7.5]/255	[12.5, 15.0]/255

the orientation specified by o_S. Independent pixelwise Gaussian noise is $\mathcal{G}(a) = a + \epsilon_n$, with $\epsilon_n \sim \mathcal{N}(0, \sigma_n^2)$. The compound noise operator is

$$\mathcal{D}_N(a) = Q(\mathcal{G}(\mathcal{S}(\mathcal{O}(a)))) \,. \tag{38}$$

For $r = (r_1, \ldots, r_L)$, the observation is $y = (\mathcal{D}_{r_L} \circ \cdots \circ \mathcal{D}_{r_1})(x)$. Parameter ranges and discrete supports are listed in Tables 11 and 12. This offline process produces the stored training and benchmark pairs. The differentiable model in Appendix D approximates the same ordered process for training-time consistency verification.

F.2 Downstream Object Detection Protocol

We initialize YOLOv8s [53] from pretrained weights and fine-tune it on the annotated M^{3}FD training split for 50 epochs. Before fine-tuning, we exclude the 300 source images used in our restoration evaluation and all of their variants. The restored versions of these 300 images therefore form a strictly unseen detection test set. After fine-tuning, the detector is frozen and applied identically to degraded inputs and the outputs of every restoration method, and no method-specific detector adaptation is performed. We report mAP$_{50}$, mAP$_{50:95}$, and per-class AP$_{50:95}$ using the same annotations, detector, NMS configuration, and evaluation protocol for every restoration method. This protocol isolates the effect of restoration quality on downstream recognition rather than detector exposure to the test scenes.

F.3 Metrics and Reporting Protocol

For paired data, PSNR and SSIM are computed between the restored and clean images. We report average results together with Single, Double, and Triple subsets. On the unpaired TIIQAD set, MUSIQ [21], NIMA [50], and NIQE [42] evaluate perceptual quality. Higher MUSIQ and NIMA and lower NIQE are preferred. These generic no-reference measures are reported jointly rather than interpreted as direct substitutes for paired fidelity. Efficiency is evaluated using parameter count and per-image inference latency. Training and inference settings are provided in Table 15, and latency uses the same input resolution (512×640), device (NVIDIA RTX 4090 GPU), and repeated-run protocol (10 times) for every model. For fair comparison, competing methods use the same source splits, degraded inputs, clean targets, image range, and full-resolution evaluation procedure.

G Additional Experimental Results

G.1 Robustness to Degradation Severity

We evaluate severity robustness using paired weak and strong variants of HM-TIR. The two variants share the same source images, degradation sequences, spatial configurations, normalized parameter-sampling quantiles, and standard-normal noise draws, and only the mapping from each quantile to its severity interval is changed. The weak and strong settings use the parameter ranges summarized in Table 16. All learned methods reuse the checkpoints trained on the standard HM-TIR training split, without further training or adaptation for either severity setting. Under weak degradations, SIRV

Table 17: Restoration performance under weak degradations on HM-TIR.

Method	Single		Double		Triple		Avg.	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
WFAF	25.708	0.821	22.079	0.744	19.311	0.659	22.366	0.741
LRSID	22.335	0.610	20.478	0.597	18.543	0.573	20.452	0.593
TSIRIE	27.977	0.902	26.658	0.876	25.443	0.840	26.693	0.873
DA-CLIP	17.301	0.671	17.054	0.628	16.682	0.578	17.012	0.626
DiffUIR	20.822	0.854	20.174	0.812	18.133	0.745	19.710	0.804
PromptIR	28.653	0.891	25.847	0.842	23.490	0.796	25.997	0.843
DRNet	29.924	0.919	27.714	0.882	26.072	0.851	27.903	0.884
ECMRNet	25.541	0.894	25.177	0.881	25.731	0.871	25.483	0.882
PPFN	26.305	0.893	26.439	0.876	26.174	0.850	26.306	0.873
SEGD	33.576	0.952	28.961	0.906	26.073	0.861	29.537	0.906
HiDRA	28.966	0.876	28.547	0.860	27.416	0.830	28.310	0.855
SIRV (Ours)	**35.200**	**0.972**	**30.648**	**0.938**	**27.815**	**0.900**	**31.221**	**0.937**

Table 18: Restoration performance under strong degradations on HM-TIR.

Method	Single		Double		Triple		Avg.	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
WFAF	20.397	0.646	16.248	0.538	14.671	0.443	17.105	0.542
LRSID	19.113	0.585	16.136	0.569	14.695	0.545	16.648	0.566
TSIRIE	26.295	0.869	22.681	0.775	19.261	0.676	22.746	0.773
DA-CLIP	16.587	0.634	15.555	0.537	15.047	0.488	15.730	0.553
DiffUIR	22.713	0.818	21.152	0.716	17.685	0.590	20.517	0.708
PromptIR	23.617	0.782	18.846	0.665	15.733	0.565	19.399	0.671
DRNet	27.229	0.860	23.323	0.770	20.450	0.686	23.667	0.772
ECMRNet	23.389	0.867	20.990	0.792	19.328	0.704	21.236	0.788
PPFN	24.380	0.843	22.755	0.764	19.733	0.679	22.289	0.762
SEGD	30.286	0.916	23.716	0.796	18.600	0.676	24.201	0.796
HiDRA	26.313	0.855	25.369	0.804	**23.390**	0.736	25.024	0.798
SIRV (Ours)	**31.878**	**0.942**	**25.521**	**0.849**	20.684	**0.743**	**26.028**	**0.845**

ranks first in all eight reported entries and achieves 31.221 dB PSNR and 0.937 SSIM on Avg., outperforming SEGD by 1.684 dB and 0.031, respectively. Under strong degradations, SIRV retains the best Avg. of 26.028 dB and 0.845, exceeding HiDRA by 1.004 dB and 0.047. It also obtains the highest PSNR and SSIM on single and double degradations. Under strong triple degradation, SIRV achieves the highest SSIM of 0.743, although its PSNR is lower than HiDRA's. Overall, SIRV maintains the best average PSNR and SSIM under both severity settings.

G.2 Additional Qualitative Results

Single degradations. Figure 7 compares restoration under low contrast (C), blur (B), and compound sensor noise (N). For C, SIRV recovers the narrow horizontal patterns on the building facade and preserves their separation from the roof boundary, whereas several competing methods either retain weak contrast or distort these repetitive structures. For B, it reconstructs sharper facade contours, window boundaries, and fine foreground structures without the halos or excessive sharpening visible in other results. Under N, SIRV suppresses stripe and granular interference while retaining thin tree branches, the road boundary, and the weak pedestrian target. These observations agree with its best displayed PSNR/SSIM on all three cases: 32.811/0.979, 42.488/0.985, and 32.708/0.917.

Double degradations. As shown in Figure 8, simultaneously removing two degradations is more challenging because aggressive enhancement can amplify noise or erase already weakened structures. In the $C + B$ example, SIRV clearly separates the two pedestrians beside the vehicle and restores the surrounding road and facade boundaries. For $C + N$, it removes spatially varying interference while preserving the tree canopy, roofline, and human silhouette. In the $N + B$ case, SIRV reconstructs thin branches and small roadside targets that are blurred, fragmented, or contaminated by residual patterns in competing outputs. Its closer agreement with the reference in both structural boundaries and local thermal-intensity transitions explains the best PSNR/SSIM across all three double-degradation examples.

Triple degradations. Figure 9 presents a challenging $N+B+C$ case, where noise, blur, and contrast reduction jointly obscure the group of people and the surrounding vegetation. Many competing methods recover global contrast but either merge adjacent human targets, smooth their boundaries, or introduce irregular high-frequency responses in the enlarged region. SIRV more clearly separates the individual silhouettes, preserves their internal thermal-intensity variations, and reconstructs the fine background branches with fewer residual artifacts. Consequently, its result is visually closest to the reference and achieves 28.141 dB PSNR and 0.904 SSIM, compared with 27.013 dB and 0.850 for the strongest displayed alternative, supporting improved restoration in this compound-degradation example.

G.3 Additional Ablation Studies

Effect of top-k posterior inference. Figure 10 shows that increasing k from 1 to 2 improves PSNR/SSIM from 28.274/0.877 to 28.331/0.892, while increasing latency from 120.9 to 149.6 ms. Although $k = 3$ gains only 0.003 dB, it reduces SSIM to 0.882 and increases latency to 178.3 ms.

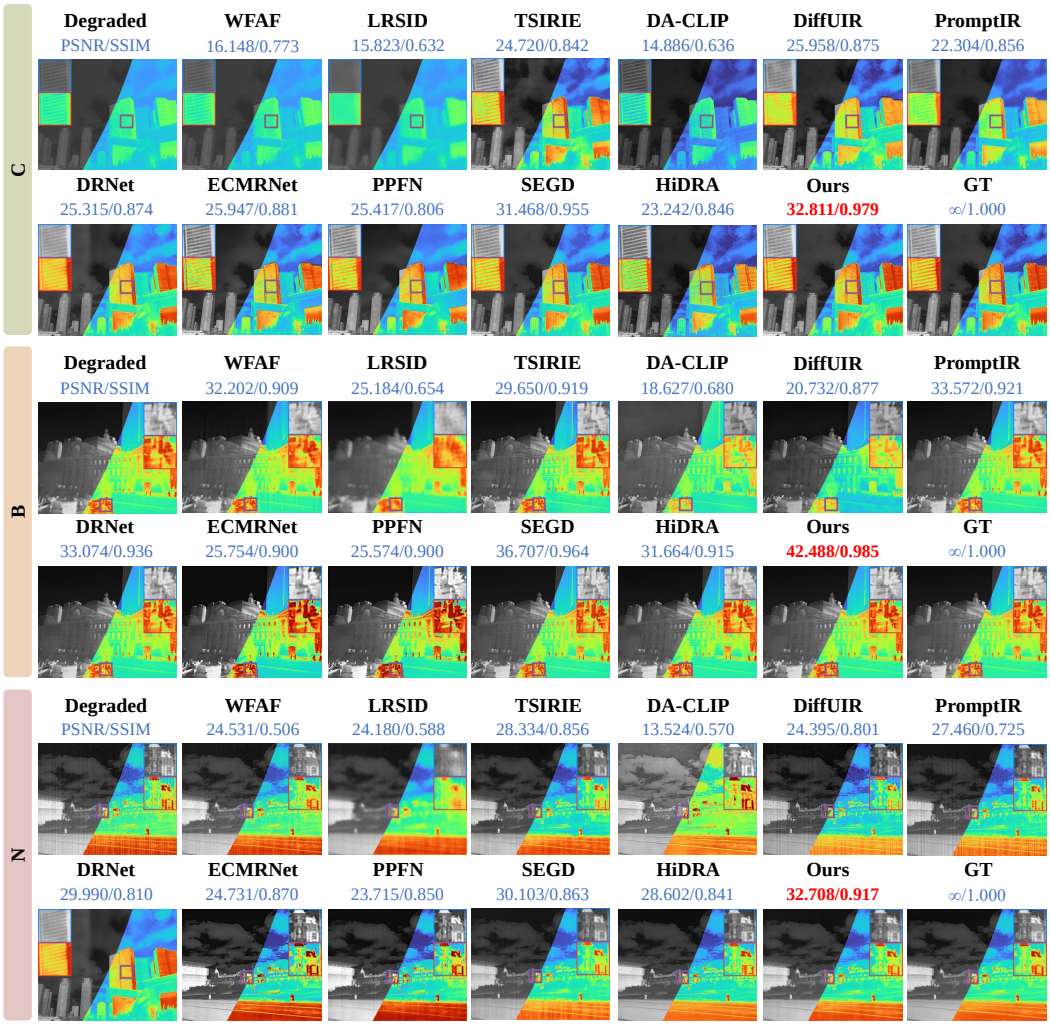


Figure 7: Qualitative comparisons of single-degradation cases on HM-TIR. The highest PSNR and SSIM values among restoration methods are highlighted in red.

Increasing k to 5 lowers restoration quality and raises latency to 235.1 ms, suggesting that retaining additional lower-probability hypotheses is not always beneficial. We therefore adopt $k = 2$ to balance restoration quality and latency among the evaluated settings.

Effect of the learned degradation posterior. We compare the learned image-conditioned posterior with image-independent baselines derived from the training manifest. The sequence baseline predicts the training-set class frequencies, while the continuous and discrete baselines use the training median and majority class, respectively. Sequence prediction is evaluated under blind inference, and parameter estimation is conditioned on the ground-truth sequence to isolate parameter-head performance. Table 19 (a) shows that the learned posterior raises Top-1 and Top-2 sequence accuracy from 8.0% and 15.3% to 71.8% and 92.4%, respectively, while reducing NLL by 65.0%, confirming that it extracts image-dependent degradation evidence rather than reproducing training frequencies. In Table 19 (b), following prior computer-vision work that reports normalized mean absolute error (NMAE) for scale-independent regression evaluation [34], each active parameter is normalized by its valid range before the absolute errors are pooled over all active parameter instances. Our posterior reduces NMAE from 0.254 to 0.183. Table 19 (c) reports per-parameter accuracy and overall accuracy pooled over all active discrete-parameter instances. Although kernel prediction remains challenging, our method substantially improves polarity, stripe-orientation, and overall accuracy, demonstrating that the learned posterior captures informative degradation states beyond image-independent statistical priors.

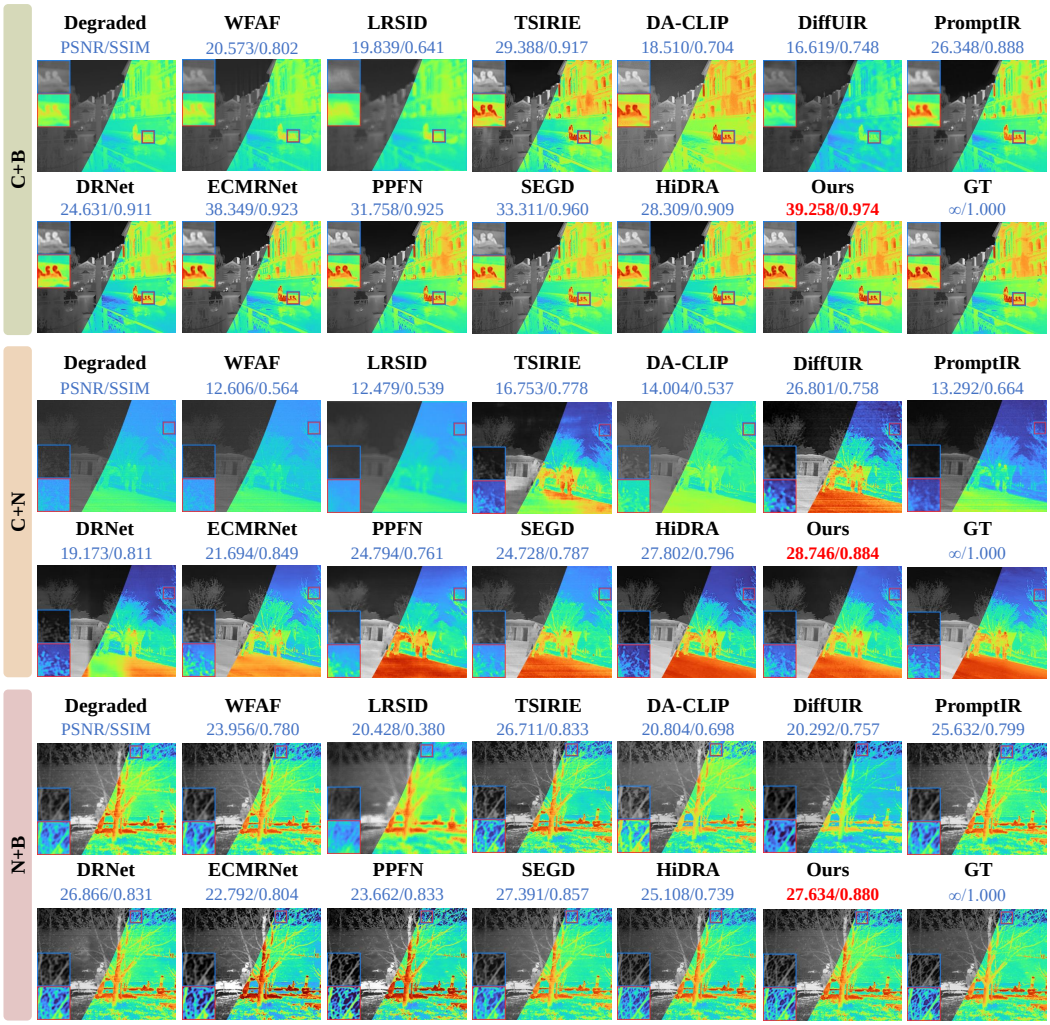


Figure 8: Qualitative comparisons of double-degradation cases on HM-TIR.

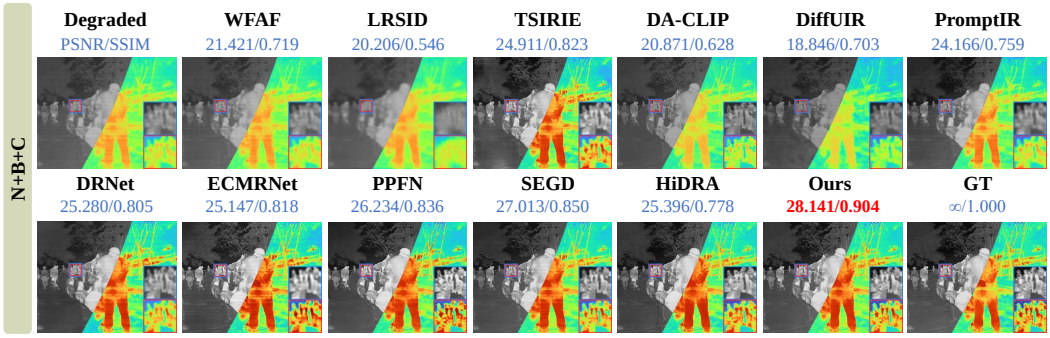


Figure 9: Qualitative comparisons of triple-degradation cases on HM-TIR.

H Limitations and Future Work

SIRV is trained with truth-centered priors derived from synthetic supervision and currently considers a closed degradation vocabulary. Its factorized parameter posterior, truncated top-k inference, and differentiable forward redegradation model provide tractable learning but simplify dependencies among degradation parameters and spatial uncertainties. Moreover, because some degradation pro-

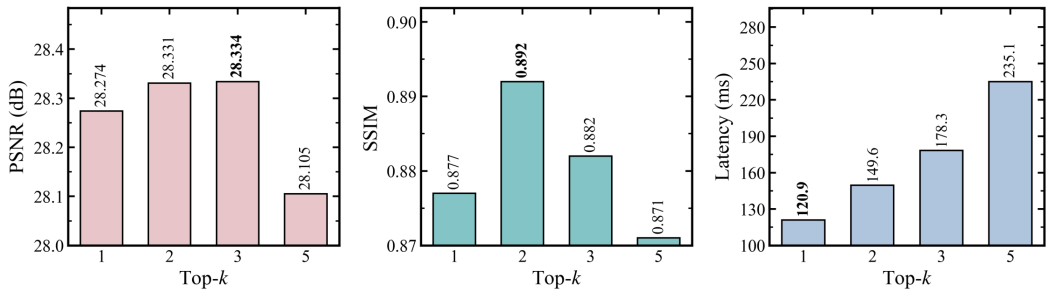


Figure 10: Effect of the number of retained posterior hypotheses on restoration performance and inference latency on HM-TIR.

(a) Sequence prediction

Method	Top-1 ↑	Top-2 ↑	NLL ↓
Training frequency prior	0.080	0.153	2.700
Learned posterior (Ours)	0.718	0.924	0.944

(b) Continuous parameter prediction

Method	NMAE ↓
Training median prior	0.254
Learned posterior (Ours)	0.183

(c) Discrete parameter prediction

Method	Kernel Acc. ↑	Polarity Acc. ↑	Stripe Acc. ↑	Overall Acc. ↑
Training majority prior	0.176	0.500	0.509	0.392
Learned posterior (Ours)	0.167	0.709	0.991	0.617

Table 19: Diagnostic comparison between the learned posterior and image-independent training-statistics baselines on HM-TIR: (a) blind degradation-sequence prediction, (b) ground-truth-order-conditioned continuous-parameter estimation, and (c) ground-truth-order-conditioned discrete-parameter prediction.

cesses are non-injective, observation consistency serves as a structured regularizer rather than a certificate of a unique correct restoration. A primary future direction is to reduce reliance on synthetic supervision by learning degradation explanations from unpaired or weakly supervised real-world TIR observations. Richer posterior models could capture multimodal and correlated degradation parameters while learning sample-adaptive uncertainty. Efficient differentiable marginalization over degradation sequences could extend observation-consistency supervision to alternative sequence branches, beyond the current evaluation on the supervised branch. The forward model could also be extended with spatially correlated uncertainty and a learnable, open degradation vocabulary to accommodate unseen sensor effects and more complex mixtures. Finally, incorporating temporal, multispectral, or downstream semantic constraints may extend the inference-restoration-verification loop to infrared video and perception-oriented restoration.