CONTROLIR: LEARNING TO CONTROL STATE SPACE MODELS VIA SPATIAL EVIDENCE FOR IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image restoration aims to remove degradations and reconstruct clear scene content, but complex scene structures and spatially varying degradations make it difficult to suppress interference while preserving details. Existing methods use scene structure or degradation cues to guide restoration, but how to jointly describe the structural and degradation properties of different regions and use them to regulate the restoration process remains a key challenge. We propose ControlIR, a compact restoration network that extracts multidimensional spatial evidence from the input image to describe regional scene structure and degradation characteristics while preserving the spatial correspondence among evidence fields. It then uses these fields to condition the state-space restoration process according to their roles, allowing each location to adapt how it uses cross-region information while removing local interference and preserving both fine details and overall structure. Experiments show that ControlIR achieves higher image restoration quality than the compared methods. Controlled studies verify the contributions of multidimensional spatial evidence and operation-level control while requiring only small increases in parameters and computation.

1 INTRODUCTION

Image restoration aims to remove degradation and reconstruct clear scene content, while preserving the details that remain visible. However, scene structures and degradation levels often vary across an image, making it difficult to remove interference and retain details at the same time. Regions with clear structures need to preserve their local details, whereas heavily degraded regions may require more information from surrounding areas. Restoration therefore needs to adjust the use of local observations and cross-region context according to regional conditions.

Existing methods use scene structure and degradation information to adapt restoration. DAF-Net and DGNL-Net use depth to guide attention and non-local feature aggregation, respectively (Hu et al., 2019; 2021), while UDR-S^{2}Former uses uncertainty to guide sparse sampling (Chen et al., 2023a). AirNet, PromptIR, and AdaIR guide restoration through degradation representations, input-conditioned prompts, and frequency information, respectively (Li et al., 2022; Potlapalli et al., 2023; Cui et al., 2025). DPMambaIR further introduces degradation information into state-space parameter modulation (Liu et al., 2025). These studies show that both the construction and use of guidance information affect restoration quality. For images with pronounced spatial variation, a further question is how to describe each region using information about structure, local variation, and degradation, and enable

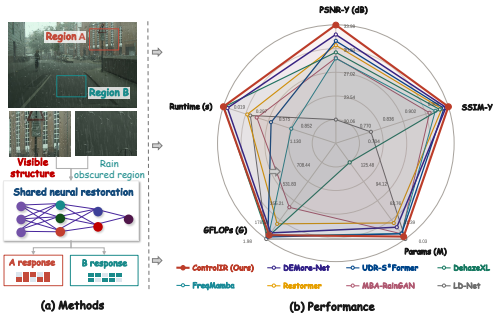


Figure 1: Spatial evidence guides restoration operations in ControlIR. Top: spatial evidence maps extracted from the rainy input. Middle: operation schematics showing how selected evidence fields condition restoration operations. Bottom: corresponding response visualizations and restored output.

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.

distinct restoration operations to use evidence from corresponding locations. Addressing this question requires jointly considering spatial evidence construction, alignment, and operation-specific assignment, so that regional descriptions can directly guide restoration computation.

To address this problem, we propose ControlIR, a compact restoration network that constructs multidimensional spatial evidence and uses it to learn to regulate state-space restoration. From a single input image, ControlIR derives spatially aligned evidence describing relative geometry, directional and local variations, degradation responses, and reliability. These fields preserve their distinct information and spatial correspondence, and are assigned to receiving operations according to their roles. This gives each restoration location corresponding structural and degradation cues for subsequent computation.

ControlSSM uses spatial evidence to regulate context formation and its contribution to local restoration. The unit first extracts local features and then aggregates cross-region information through bidirectional scanning. During scanning, spatial evidence helps regulate the propagation of the existing state and the current token's contribution to the state, allowing context formation to account for structural and degradation characteristics at different locations. After scanning, spatial evidence and the magnitude of the local–global feature difference jointly determine a correction weight, regulating how strongly the aggregated context corrects local features. This allows the model to coordinate local details and cross-region information according to regional conditions.

The main contributions are as follows. (1) We introduce a multidimensional spatial-evidence construction and organization scheme that extracts scene structure and degradation information from a single input image and provides region-specific cues for different restoration operations. (2) We develop ControlSSM, an operation-coupled state-space unit that jointly controls how contextual features are formed and how they correct local features, allowing local details and cross-region information to adapt to regional differences. (3) We propose ControlIR, a compact restoration network that uses multidimensional spatial evidence to drive operation-level state-space control and adapt restoration to regional differences. Experiments show that ControlIR achieves better image restoration quality than the compared methods while maintaining fast inference. Controlled studies verify the roles of multidimensional spatial evidence and operation-level control and show that these gains require only small increases in parameters and computation.

2 Related Work

Image Restoration with Spatial Guidance. Existing restoration methods address degradation differences by distinguishing degradation components, adjusting regional attention, and selecting contextual information. In image deraining, DAiAM uses attention to distinguish rain types and intensity regions (Zhang et al., 2021), while MBA-RainGAN jointly handles drops, streaks, and haze through multi-branch attention (Shen et al., 2022). CCN processes raindrops and streaks in complementary orders before fusing the results (Quan et al., 2021). UDR-S^{2}Former uses uncertainty-driven ranking to guide sparse sampling and connect irregularly distributed degradation regions (Chen et al., 2023a).

Scene geometry provides structural cues for guiding restoration. DAF-Net and DGNL-Net incorporate depth into attention and non-local feature aggregation, respectively (Hu et al., 2019; 2021). DG-RainDiff introduces depth-guided dynamic message passing into diffusion-based rain removal (Yu et al., 2024), while DEMore-Net jointly trains depth estimation and mixture-of-rain restoration through depth-guided adversarial learning (Wang et al., 2025b). T^3-DiffWeather combines weather and scene prompts and constrains scene-prompt learning with Depth-Anything features (Chen et al., 2024b). These works explore the role of geometry through feature aggregation, prompt learning, and auxiliary supervision.

Image Restoration with Model Modulation. Conditional restoration uses input-related information to adapt to different degradation conditions. AirNet learns degradation representations through contrastive learning (Li et al., 2022), PromptIR generates input-conditioned prompts (Potlapalli et al., 2023), and OneRestore incorporates scene descriptors from text or visual attributes through cross-attention (Guo et al., 2025b). AdaIR extracts low- and high-frequency features under image-spectrum guidance and exchanges information through bidirectional modulation (Cui et al., 2025). Histoformer groups spatial locations by feature intensity and establishes interactions through within- and across-

group attention (Sun et al., 2024). CoIC jointly learns rain- and detail-aware representations, uses global average pooling to obtain an instance-level guidance vector, and uses this vector to modulate convolutional or self-attention layers (Ran et al., 2024).

State-Space Models for Image Restoration. Mamba aggregates long-range information through input-dependent selection (Gu & Dao, 2024). Mamba2 further connects state-space models with attention and improves the efficiency of the core computation (Dao & Gu, 2024). For image restoration, existing studies improve state-space modeling through spatial structure and feature interactions. MambaIR combines local enhancement and channel attention to strengthen local modeling and channel interactions (Guo et al., 2024). MambaIRv2 strengthens interactions among related pixels through attentive state-space modeling and semantic-guided neighboring (Guo et al., 2025a). MaIR preserves spatial continuity through nested S-shaped scanning and aggregates information from different directions using sequence-shuffle attention (Li et al., 2025).

Beyond feature modeling and scan organization, existing methods also use degradation information to modulate state-space computation. DPMambaIR uses degradation embeddings to modulate the discretization step, state-input map, and state-readout map, introducing degradation conditioning into state-space computation (Liu et al., 2025). ControlIR jointly considers the construction of multidimensional spatial evidence and its use in distinct operations, preserving spatial correspondence and assigning evidence by restoration operation. During scanning, evidence modulates state evolution and current-token write strength. After scanning, evidence and the magnitude of the local–global feature difference jointly determine correction weights. This design allows spatial evidence to guide cross-region information aggregation and regulate how the aggregated information corrects local features.

3 METHOD

3.1 OVERALL FRAMEWORK

Given an input image I^r, ControlIR uses multidimensional spatial evidence to guide feature extraction and processing for image restoration (the overview). The normalized input enters two parallel paths: the evidence path describes scene structure, degradation-related variations, and geometry-prediction reliability, while the restoration path extracts image features for reconstruction. Evidence is aligned with these features and supplied to the control interfaces in ControlSSM.

Restoration features pass through two independently parameterized ControlSSM units, with local feature refinement between them. Both units use the same evidence bank and field-assignment rules and learn their own modulation. Within each unit, spatial evidence modulates scan-level state updates and guides post-scan local feature correction.

After the second ControlSSM unit, local and channel processing is followed by a reconstruction head that predicts an RGB residual. The residual is added to the normalized input. Inverse normalization and clipping to $[0, 1]$ produce the restored image $\widehat{I}$.

3.2 MULTIDIMENSIONAL SPATIAL EVIDENCE CONSTRUCTION AND ASSIGNMENT

ControlIR constructs six spatial evidence fields from a single input image. D represents the relative depth of the scene; G_x and G_y describe horizontal and vertical geometric variations, and L describes local geometric variation. R^{deg} captures local image-intensity variations arising from both degradation and scene content. C^{rel} estimates the reliability of the geometry prediction.

The depth branch predicts relative depth $Z = f_{\mathrm{geo}}(I^n)$ from the normalized input. For reliability estimation, the absolute deviation of Z from its local 5×5 mean gives the depth-variation map Q_Z. Applying the same operation to each input channel and averaging over RGB gives the intensity-variation map Q_I. The reliability head receives concatenated Z, Q_Z, Q_I and outputs C^{rel} through sigmoid.

Spatial normalization of Z gives the base depth field D_0. Horizontal and vertical absolute neighbor differences of D_0 yield $G_{x,0}, G_{y,0}$, and its absolute deviation from the local 3×3 mean yields L_0. Normalizing Q_I gives R^{deg}. A pointwise refinement head receives the concatenation of these four

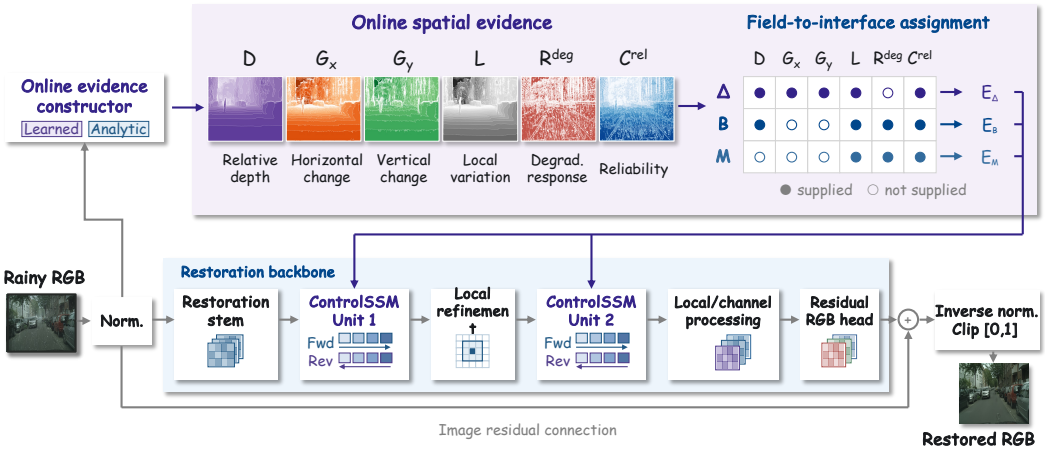


Figure 2: ControlIR inference and training supervision. The normalized rainy image drives the spatial-evidence constructor and restoration encoder in parallel. The co-registered evidence bank conditions two independently parameterized ControlSSM units, and the restoration path predicts an RGB residual for reconstruction.

base geometry fields and C^{rel} to predict residual adjustments. In compact form, the refined geometry fields are $\mathbf{e}_{\mathrm{geo}} = (D, G_x, G_y, L) = \big(\mathrm{SN}(D_0 + r_D), \mathrm{SN}(|G_{x,0} + r_x|) \odot C^{\mathrm{rel}}, \mathrm{SN}(|G_{y,0} + r_y|) \odot C^{\mathrm{rel}}, \mathrm{SN}(|L_0 + r_L|) \odot C^{\mathrm{rel}}\big)$. Here, SN applies spatial min–max normalization independently to each sample and channel, and $\odot$ denotes element-wise multiplication. The resulting evidence bank is $\mathbf{e} = [D, G_x, G_y, L, R^{\mathrm{deg}}, C^{\mathrm{rel}}]$.

Evidence alignment and assignment. ControlIR aligns evidence to the feature grid of each receiving interface and selects fields according to a predefined assignment (the topology description). Before scanning, evidence and restoration features are flattened in the same row-major order. The post-scan correction interface uses evidence aligned to the local-feature grid.

The discretization interface uses $D, G_x, G_y, L, C^{\mathrm{rel}}$ to modulate state evolution with relative depth, geometric variation, and reliability. The write interface uses $D, L, R^{\mathrm{deg}}, C^{\mathrm{rel}}$ to regulate current-token write strength with structural and degradation-related cues. The post-scan correction interface receives $L, R^{\mathrm{deg}}, C^{\mathrm{rel}}$ for local variation, degradation response, and reliability information. Both ControlSSM units use the same field assignment. Each interface reads selected fields in their original order in the evidence bank. The corresponding inputs are denoted by $\mathbf{E}_{\Delta}^{(k)}$, $\mathbf{E}_{B}^{(k)}$, and $\mathbf{E}_{M}^{(k)}$, where k indexes the unit. Field-assignment comparisons are reported in Table 5.

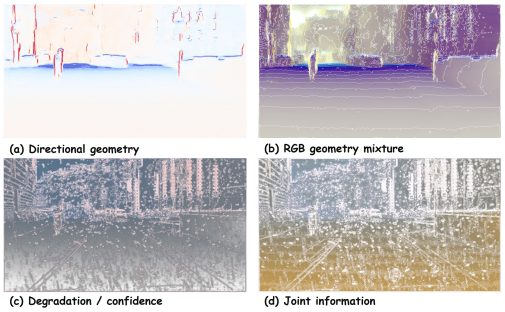


Figure 3: Spatial evidence and its interpretable visualization. ControlIR extracts directional geometry, local geometry, degradation, and reliability signals while preserving their spatial correspondence. (a) Directional geometry response; (b) RGB geometry mixture of G_x, L, and G_y; (c) joint display of degradation cues and low-confidence geometry; (d) joint information from degradation, reliable geometry, and low-confidence regions. RGB colors indicate relative combinations of evidence channels, whereas the other panels show normalized relative responses. Geometry fields are normalized residuals after removing components linearly correlated with D.

3.3 Operation-Coupled ControlSSM Units

The kth ControlSSM unit extracts local features $F^{\ell,(k)}$ from its input $F^{(k)}$. The global branch applies projection, depthwise convolution, and pooling, then flattens the features in row-major order into the sequence $x^{(k)}$ supplied to Mamba2 (Dao & Gu,

2024). Inside this module, the input projection generates the baseline discretization prediction $s_{\Delta}^{(k)}$. The projected convolutional branch, followed by SiLU activation, produces recurrence inputs $u^{(k)}$, state-input maps $B^{(k)}$, and readout maps $C_{\mathrm{ssm}}^{(k)}$. Spatial evidence modulates discretization and write strength; the resulting context is used to correct $F^{\ell,(k)}$.

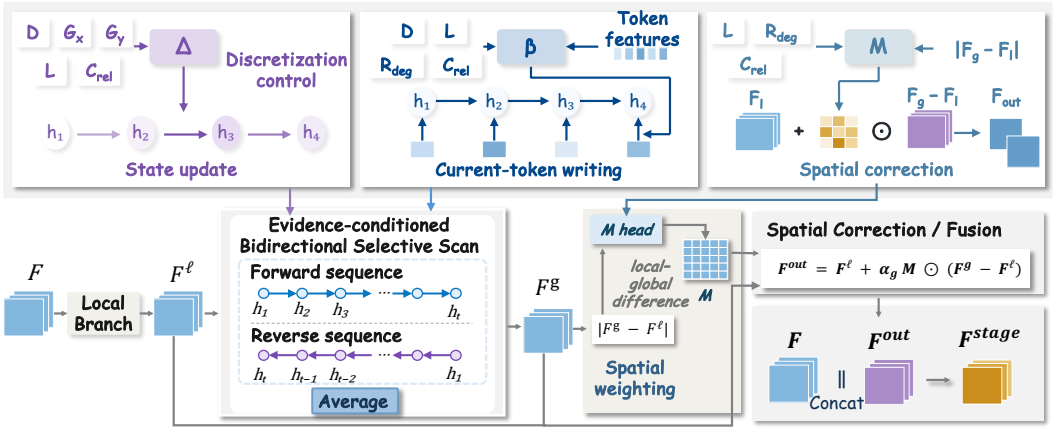


Figure 4: Operation-coupled ControlSSM unit. Evidence modulates effective discretization and current-token state writing within the bidirectional selective scan. The post-scan interface predicts a spatial weight from the port-specific evidence and the local–global feature difference.

3.3.1 Scan-level conditioning

We omit batch and unit indices below, with t denoting the scan position. The discretization controller adjusts the baseline prediction using evidence at the corresponding location:

$$\Delta'_t = \mathrm{softplus}\big(s_{\Delta,t} + b_\Delta + C_t^{\mathrm{rel}} \alpha_\Delta \tanh\big(h_\Delta(\mathbf{E}_{\Delta,t})\big)\big)\,. \tag{1}$$

Here, b_Δ is a bias, h_Δ is the control head, and α_Δ is a learnable modulation strength. Reliability C_t^{rel} scales the adjustment, and softplus makes Δ'_t positive. The write controller combines the channel mean $\bar{x}_t$ of the Mamba2 input token x_t with interface evidence $\mathbf{E}_{B,t}$ to generate a scale around one:

$$\beta_t = 1 + C_t^{\mathrm{rel}} \alpha_B \tanh\big(h_B([\bar{x}_t; \mathbf{E}_{B,t}])\big), \qquad B'_t = \beta_t B_t. \tag{2}$$

Here, h_B is the write-control head and α_B is a learnable strength. The scale is broadcast over the state dimension. Substituting Δ'_t and B'_t into the recurrence gives, for a single scan head,

$$H_t = \exp(\Delta'_t a) H_{t-1} + \Delta'_t u_t (B'_t)^\top. \tag{3}$$

Here, $H_t \in \mathbb{R}^{d_h \times n_s}$ is the state, with head width d_h and state dimension n_s; $u_t \in \mathbb{R}^{d_h}$ is the recurrence input after the internal projection, convolution, and activation, and $B'_t \in \mathbb{R}^{n_s}$. The learnable transition parameter satisfies $a < 0$. The first term propagates the previous state and the second writes the current input. Discretization Δ'_t affects both terms, while β_t further regulates writing through $B'_t = \beta_t B_t$.

Forward and reverse scans share scan and controller parameters, with features and evidence reversed together. The reverse output is restored to the original order and averaged with the forward output. The result is reshaped into a two-dimensional feature map, projected, and aligned to the local-feature grid to obtain contextual features F^g. The forward and reverse scans share parameters and reverse the evidence with the feature sequence. Their outputs are restored to the original order and averaged:

$$y^{(k)} = \tfrac{1}{2}\Big[\mathcal{S}_k(x^{(k)}; \mathbf{E}_{\mathrm{scan}}^{(k)}) + \mathrm{Rev}\Big(\mathcal{S}_k(\mathrm{Rev}(x^{(k)}); \mathrm{Rev}(\mathbf{E}_{\mathrm{scan}}^{(k)}))\Big)\Big]\,. \tag{4}$$

3.3.2 Post-scan spatial correction

After scanning, the correction interface predicts a single-channel spatial weight from interface evidence and the magnitude of the contextual–local feature difference:

$$M = \mathrm{sigmoid}\big(h_M([|F^g - F^{\ell}|; \mathbf{E}_M])\big)\,. \tag{5}$$

Here, h_M is the correction-weight head. The weight M is broadcast over feature channels and scales the contextual–local difference before it is added to the local feature:

$$F^{\text{out}} = F^{\ell} + \alpha_g M \odot (F^{g} - F^{\ell}), \tag{6}$$

where α_g is a learnable correction-strength coefficient. Weight prediction uses the absolute difference, while correction retains its sign, allowing context to increase or decrease local feature values. Finally, F^{out} is concatenated with the unit input F along channels for subsequent restoration layers.

4 Experiments

4.1 Experimental Settings

We instantiate and evaluate the spatial-evidence control framework on RainCityscapes++, a mixture-of-rain restoration benchmark. RainCityscapes++ images are resized to 720×480 pixels (width $\times$ height), with scene-disjoint training and validation splits. We report PSNR and SSIM (Wang et al., 2004) on the luminance channel and AlexNet-based LPIPS (Zhang et al., 2018) on RGB images. Each ablation is repeated three times and reported as mean $\pm$ sample standard deviation. Full training settings and supplementary experiments on other mixture-of-rain benchmarks are provided in Appendices B and A, respectively.

4.2 Comparison with Existing Methods

Table 1 reports the quantitative comparison on RainCityscapes++. ControlIR achieves the highest PSNR-Y and SSIM-Y among the compared methods, improving over DEMore-Net by 1.41 dB and 0.007, respectively.

The qualitative examples include adherent raindrops, rain streaks, and rainy haze. In the marked regions, ControlIR removes visible rain interference while retaining clearer edges and local details. ControlIR has an average inference time of 0.019 s per 720×480 image on an NVIDIA RTX 3090.

4.3 Ablation Studies

Overall control design. Table 2 compares the restoration backbone and the evidence-control design. A uses DenseRR convolutional blocks with a depth branch, B replaces DenseRR with

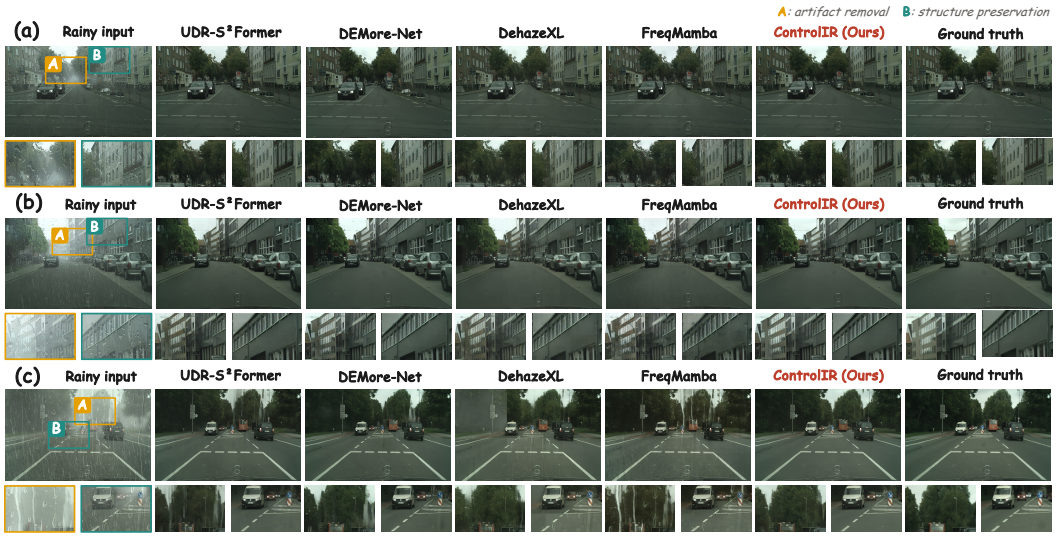


Figure 5: Qualitative comparison on RainCityscapes++. Columns show the rainy input, DehazeXL, DEMore-Net, UDR-S^{2}Former, FreqMamba, ControlIR, and ground truth. Rows (a)–(c) focus on adherent raindrops, rain streaks, and rainy haze, respectively. Marked regions are enlarged for aligned comparison.

Table 1: Complete comparison on RainCityscapes++. The best restoration scores are highlighted in bold.

Method	Target	Complexity		Restoration quality	
		Params (M)↓	FLOPs (G)↓	PSNR-Y↑	SSIM-Y↑
AttentiveGAN Qian et al. (2018)	Rain D	6.00	353.08	24.03	0.842
CMFNet Fan et al. (2023)	Rain D	17.24	638.96	28.59	0.891
MSPFN Jiang et al. (2020)	Rain S	21.00	708.44	22.19	0.704
Syn2Real Yasarla et al. (2020)	Rain S	2.61	20.92	27.15	0.898
MPRNet Zamir et al. (2021)	Rain S	3.64	148.55	22.87	0.762
Restormer Zamir et al. (2022)	Rain S	26.13	140.99	31.00	0.949
HCT-FFN Chen et al. (2023b)	Rain S	0.87	80.25	26.65	0.816
DAWN Jiang et al. (2023)	Rain S	3.36	199.89	28.79	0.864
MSDT Chen et al. (2024a)	Rain S	16.60	129.90	31.74	0.911
FreqMamba Zou et al. (2024)	Rain S	8.93	30.18	29.06	0.930
LD-Net Ullah et al. (2021)	Rain H	0.03	1.98	20.06	0.741
SGID Bai et al. (2022)	Rain H	13.87	156.40	21.87	0.761
DehazeFormer-B Song et al. (2023)	Rain H	2.51	25.79	27.24	0.824
UCL-Dehaze Wang et al. (2024)	Rain H	19.45	78.80	20.17	0.802
DehazeXL Chen et al. (2025)	Rain H	125.48	22.79	29.95	0.943
DAF-Net Hu et al. (2019)	Rain S/H	1.72	13.15	25.21	0.852
DGNL-Net Hu et al. (2021)	Rain S/H	4.03	19.73	26.89	0.864
CCN Quan et al. (2021)	Rain D/S	3.75	245.85	27.51	0.901
UDR-S^{2}Former Chen et al. (2023a)	Rain D/S	8.53	21.58	31.64	0.955
MBA-RainGAN Shen et al. (2022)	Rain D/S/H	8.62	294.32	29.16	0.913
DEMore-Net Wang et al. (2025b)	Rain D/S/H	19.25	58.01	32.57	0.961
ControlIR (Ours)	Rain D/S/H	3.90	36.02	**33.98**	**0.968**

Note: Rain D, Rain S, and Rain H denote raindrops, rain streaks, and rainy haze, respectively. A slash denotes a mixture of rain types; a plus sign between method names denotes sequential application. Published comparison scores are retained where used in the benchmark summary. External complexity values follow their source protocols, whereas ControlIR reports active parameters and internally estimated FLOPs.

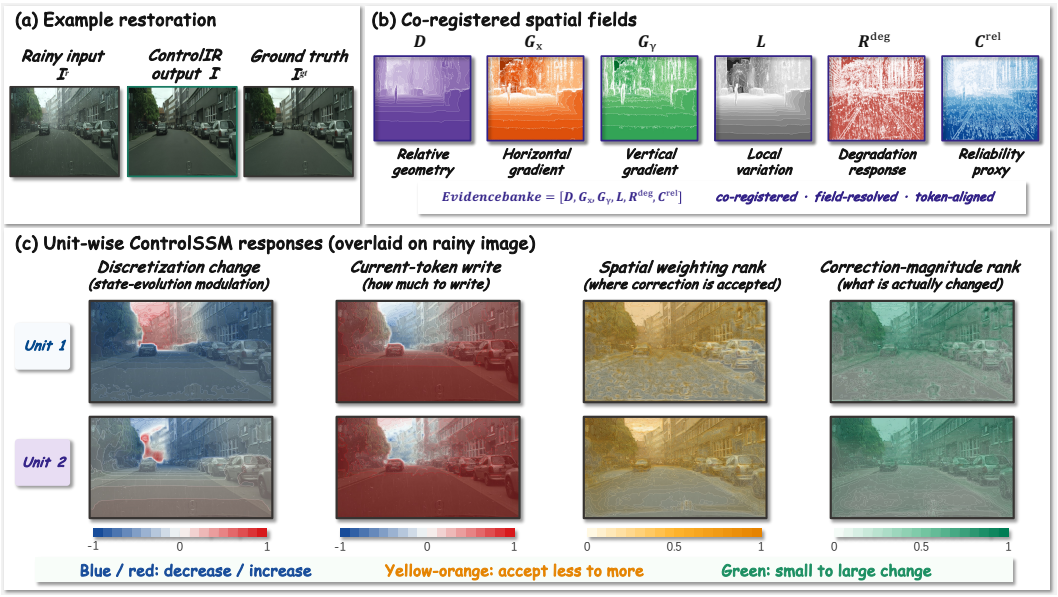


Figure 6: Spatial evidence and ControlSSM responses. The figure shows an example restoration, the six co-registered evidence fields, and the location-dependent responses of the two ControlSSM units.

bidirectional state-space processing, C (StateConcat) introduces the six evidence fields through pre-scan concatenation, and Full uses operation-coupled control during scanning and post-scan correction.

Replacing DenseRR with bidirectional state-space processing improves restoration quality while reducing model complexity. Full achieves the best scores on all three metrics. Relative to B, it improves PSNR-Y by 1.15 dB, SSIM-Y by 0.0060, and LPIPS-RGB by 0.0125, with increases of 0.0400 M parameters and 0.6733 G FLOPs.

Evidence delivery and supervision. Table 3 compares evidence delivery and auxiliary prior supervision. U uses neither evidence conditioning nor auxiliary prior supervision. StateConcat-noPrior and C0 introduce evidence through pre-scan feature concatenation and the three operation-specific control interfaces, respectively, using only the restoration loss. StateConcat and Full additionally use the auxiliary prior supervision.

Table 2: Overall control design.

Variant	PSNR-Y (dB)↑	SSIM-Y↑	LPIPS-RGB↓	Params (M)↓	FLOPs (G)↓
A: CNN + Depth	32.0125 ± 0.0984	0.9573 ± 0.0003	0.0641 ± 0.0046	4.3469	44.4810
B: BiSSM + Depth	32.8163 ± 0.0797	0.9615 ± 0.0005	0.0581 ± 0.0034	3.8638	35.3440
C: StateConcat	33.3065 ± 0.0386	0.9653 ± 0.0007	0.0491 ± 0.0006	3.8734	35.3919
D: Full	$\mathbf{33.9681 \pm 0.0143}$	$\mathbf{0.9675 \pm 0.0002}$	$\mathbf{0.0456 \pm 0.0014}$	3.9039	36.0173

A uses DenseRR convolutional blocks. A/B retain the depth branch; C uses evidence concatenation and D uses three control interfaces. Values are means ± sample standard deviations across three runs. Parameters count active parameters; FLOPs use the common $1 \times 3 \times 128 \times 256$ profiling input.

Table 3: Evidence delivery and supervision.

Variant	PSNR-Y (dB)↑	SSIM-Y↑	LPIPS-RGB↓	Params (M)↓	FLOPs (G)↓
U	32.5587 ± 0.0209	0.9605 ± 0.0003	0.0541 ± 0.0027	2.1667	34.3055
StateConcat-noPrior	32.8003 ± 0.0353	0.9643 ± 0.0002	0.0517 ± 0.0007	3.8734	35.3919
StateConcat	33.3065 ± 0.0386	0.9653 ± 0.0007	0.0491 ± 0.0006	3.8734	35.3919
StateConcat+M	33.4394 ± 0.0029	0.9660 ± 0.0004	0.0480 ± 0.0012	3.9121	36.0253
SpatialAffine+M	33.3266 ± 0.0053	0.9653 ± 0.0007	0.0493 ± 0.0010	3.9040	36.0167
C0	33.6846 ± 0.0195	0.9670 ± 0.0002	0.0465 ± 0.0012	3.9039	36.0173
Full	$\mathbf{33.9681 \pm 0.0143}$	$\mathbf{0.9675 \pm 0.0002}$	$\mathbf{0.0456 \pm 0.0014}$	3.9039	36.0173

Quality scores are three-run means ± sample standard deviations. U uses neither evidence conditioning nor auxiliary prior supervision; StateConcat-noPrior and C0 omit only the auxiliary prior loss. Complexity follows the common internal counting protocol.

Both evidence delivery designs improve PSNR-Y over U without auxiliary supervision, and both benefit from adding it. Under either supervision setting, operation-coupled control achieves higher PSNR-Y and SSIM-Y and lower LPIPS-RGB than pre-scan concatenation.

StateConcat+M retains pre-scan concatenation, native discretization, and $\beta = 1$, and adds the same post-scan M controller and local–global fusion used by Full. Its PSNR-Y increases to 33.4394 dB, while Full remains 0.53 dB higher with better SSIM-Y and LPIPS-RGB. SpatialAffine+M provides a stronger feature-level comparison by adding learned spatial affine modulation to the concatenation scheme while retaining the post-scan correction. It obtains 33.3266 PSNR-Y, 0.9653 SSIM-Y, and 0.0493 LPIPS-RGB, remaining close to StateConcat and below Full.

Control routes and evidence assignment. Table 4 removes one control route at a time while retaining the other two. Without Δ control, the scan uses the feature-dependent baseline discretization. Without B control, the additional current-token write scaling is removed by setting $\beta = 1$. For M-identity, the learned spatial weight is replaced with $M = 1$, while the signed local–global correction is retained.

Removing any route lowers PSNR-Y and SSIM-Y relative to Full. The largest PSNR-Y drop occurs without Δ control (0.4964 dB), followed by removing B control (0.3346 dB) and replacing M with the identity (0.3681 dB). The B-removed variant has a slightly lower LPIPS-RGB than Full, so the three routes do not improve every metric in the same direction. Overall, the results support the complementary roles of context propagation, current-token writing, and spatially weighted local correction.

Table 5 examines whether giving additional evidence fields to the control interfaces improves restoration. All-to-all provides all six fields to every control head. Full+$D \to M$ additionally provides relative depth to the post-scan correction interface. Full+$G_x, G_y \to B, M$ additionally provides both

directional variation fields to the write and post-scan correction interfaces. All other settings remain unchanged.

Table 4: Control routes.

Variant	PSNR-Y (dB)↑	SSIM-Y↑	LPIPS-RGB↓	Params (M)↓	FLOPs (G)↓
w/o Δ	33.4717 ± 0.0217	0.9663 ± 0.0007	0.0473 ± 0.0009	3.9034	36.0165
w/o B	33.6335 ± 0.0022	0.9665 ± 0.0005	$\mathbf{0.0455 \pm 0.0009}$	3.9037	36.0170
$M = 1$	33.6000 ± 0.0210	0.9664 ± 0.0008	0.0464 ± 0.0007	3.8652	35.3840
Full	$\mathbf{33.9681 \pm 0.0143}$	$\mathbf{0.9675 \pm 0.0002}$	0.0456 ± 0.0014	3.9039	36.0173

Quality scores are three-run means $\pm$ sample standard deviations. w/o Δ uses native discretization, w/o B sets $\beta = 1$, and $M = 1$ replaces the learned spatial weight with the identity.

Table 5: Evidence assignment.

Variant	PSNR-Y (dB)↑	SSIM-Y↑	LPIPS-RGB↓	Params (M)↓	FLOPs (G)↓
All-to-all	33.4088 ± 0.0401	0.9658 ± 0.0005	0.0473 ± 0.0003	3.9057	36.0459
Full+$D \to M$	33.5700 ± 0.0033	0.9666 ± 0.0005	0.0471 ± 0.0007	3.9045	36.0268
Full+$G_x, G_y \to B, M$	33.6516 ± 0.0353	0.9670 ± 0.0003	0.0473 ± 0.0017	3.9051	36.0364
Full	$\mathbf{33.9681 \pm 0.0143}$	$\mathbf{0.9675 \pm 0.0002}$	$\mathbf{0.0456 \pm 0.0014}$	3.9039	36.0173

Quality scores are three-run means $\pm$ sample standard deviations. The expanded-input variants retain three separate control heads; Full uses the role-specific field assignment.

Full achieves higher mean PSNR-Y, SSIM-Y, and LPIPS-RGB than the three tested expanded-input variants, with slightly fewer parameters and FLOPs. Under these comparisons, assigning fields according to the role of each operation is more effective than uniformly expanding its input.

5 Conclusion

We presented ControlIR, a compact framework that learns to control state space restoration via spatial evidence. From a single input, it constructs distinct, spatially corresponding evidence fields and converts them into controls for context formation and post-scan local correction. ControlSSM therefore links the evidence bank to the state-space operations that propagate, write, and use contextual information. Instantiated for mixture-of-rain restoration, ControlIR improves image restoration quality while maintaining compact model size and efficient inference. Controlled experiments show that preserving evidence identity and spatial correspondence, and assigning evidence to specific operations, make the learned controls more effective. The results support spatial-evidence conditioned control as an effective design for the tested state-space restoration setting.

References

Haoran Bai, Jinshan Pan, Xinguang Xiang, and Jinhui Tang. Self-guided image dehazing using progressive feature fusion. *IEEE Transactions on Image Processing*, 31:1217–1229, 2022.

Hongming Chen, Xiang Chen, Jiyang Lu, and Yufeng Li. Rethinking multi-scale representations in deep deraining transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1046–1053, 2024a. doi: 10.1609/aaai.v38i2.27865.

Jiuchen Chen, Xinyu Yan, Qizhi Xu, and Kaiqi Li. Tokenize image patches: Global context fusion for effective haze removal in large images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2258–2268, 2025.

Sixiang Chen, Tian Ye, Jinbin Bai, Erkang Chen, Jun Shi, and Lei Zhu. Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13106–13117, 2023a. doi: 10.1109/ICCV51070.2023.01205.

Sixiang Chen, Tian Ye, Kai Zhang, Zhaohu Xing, Yunlong Lin, and Lei Zhu. Teaching tailored to talent: Adverse weather restoration via prompt pool and depth-anything constraint. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024b. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/1377_ECCV_2024_paper.php.

Xiang Chen, Jinshan Pan, Jiyang Lu, Zhentao Fan, and Hao Li. Hybrid CNN-Transformer feature fusion for single image deraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 378–386, 2023b. doi: 10.1609/aaai.v37i1.25111.

Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. AdaIR: Adaptive all-in-one image restoration via frequency mining and modulation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fb44a668c2d4bc984e9d6ca261262cbb-Abstract-Conference.html.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu. Compound multi-branch feature fusion for image deraindrop. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3399–3403, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. In *Computer Vision – ECCV 2024*, pp. 222–241. Springer, 2024. doi: 10.1007/978-3-031-72649-1_13.

Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. MambaIRv2: Attentive state space restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28124–28133, 2025a. doi: 10.1109/CVPR52734.2025.02619.

Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. OneRestore: A universal restoration framework for composite degradation. In *Computer Vision – ECCV 2024*, volume 15077 of *Lecture Notes in Computer Science*, pp. 255–272. Springer, 2025b. doi: 10.1007/978-3-031-72655-2_15.

Yun Guo, Xueyao Xiao, Yi Chang, Shumin Deng, and Luxin Yan. From sky to the ground: A large-scale benchmark and simple baseline towards real rain removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12097–12107, 2023.

Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8022–8031, 2019. doi: 10.1109/CVPR.2019.00821.

Xiaowei Hu, Lei Zhu, Tianyu Wang, Chi-Wing Fu, and Pheng-Ann Heng. Single-image real-time rain removal based on depth-guided non-local features. *IEEE Transactions on Image Processing*, 30:1759–1770, 2021. doi: 10.1109/TIP.2020.3048625.

Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8346–8355, 2020. doi: 10.1109/CVPR42600.2020.00837.

Kui Jiang, Wenxuan Liu, Zheng Wang, Xian Zhong, Junjun Jiang, and Chia-Wen Lin. DAWN: Direction-aware attention wavelet network for image deraining. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7065–7074. Association for Computing Machinery, 2023. doi: 10.1145/3581783.3611697.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17452–17462, June 2022. doi: 10.1109/CVPR52688.2022.01693.

Boyun Li, Haiyu Zhao, Wenxin Wang, Peng Hu, Yuanbiao Gou, and Xi Peng. MaIR: A locality- and continuity-preserving Mamba for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7491–7501, 2025. doi: 10.1109/CVPR52734.2025.00702.

Zhanwen Liu, Sai Zhou, Yuchao Dai, Yang Wang, Yisheng An, and Xiangmo Zhao. DPMambaIR: All-in-one image restoration via degradation-aware prompt state space model, 2025. URL https://arxiv.org/abs/2504.17732. Version 3, revised February 2026.

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H. Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71275–71293. Curran Associates, Inc., 2023. doi: 10.52202/075280-3121.

Rui Qian, Robby T. Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2482–2491, 2018. doi: 10.1109/CVPR.2018.00263.

Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9147–9156, 2021. doi: 10.1109/CVPR46437.2021.00903.

Wu Ran, Peirong Ma, Zhiquan He, Hao Ren, and Hong Lu. Harnessing joint rain-/detail-aware representations to eliminate intricate rains. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 8861–8881, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/24e4e3234178a836b70e0aa48827e0ff-Paper-Conference.pdf.

Yiyang Shen, Yidan Feng, Weiming Wang, Dong Liang, Jing Qin, Haoran Xie, and Mingqiang Wei. MBA-RainGAN: A multi-branch attention generative adversarial network for mixture of rain removal. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3418–3422, 2022. doi: 10.1109/ICASSP43922.2022.9746588.

Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. doi: 10.1109/TIP.2023.3256763.

Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring images in adverse weather conditions via histogram transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/3249_ECCV_2024_paper.php.

Hayat Ullah, Khan Muhammad, Muhammad Irfan, Saeed Anwar, Muhammad Sajjad, Ali Shariq Imran, and Victor Hugo C. de Albuquerque. Light-dehazenet: A novel lightweight cnn architecture for single image dehazing. *IEEE Transactions on Image Processing*, 30:8968–8982, 2021. doi: 10.1109/TIP.2021.3116790.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025a.

Yongzhen Wang, Xuefeng Yan, Fu Lee Wang, Haoran Xie, Wenhan Yang, Xiao-Ping Zhang, Jing Qin, and Mingqiang Wei. UCL-Dehaze: Toward real-world image dehazing via unsupervised contrastive learning. *IEEE Transactions on Image Processing*, 33:1361–1374, 2024. doi: 10.1109/TIP.2024.3362153.

Yongzhen Wang, Xuefeng Yan, Yanbiao Niu, Lina Gong, Yanwen Guo, and Mingqiang Wei. Rethinking mixture of rain removal via depth-guided adversarial learning. *Neural Networks*, 191, 2025b. doi: 10.1016/j.neunet.2025.107739. Art. no. 107739.

Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior. *arXiv preprint arXiv:2505.10565*, 2025c.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, volume 2, pp. 1398–1402, 2003. doi: 10.1109/ACSSC.2003.1292216.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems*, 2024.

Rajeev Yasarla, Vishwanath A. Sindagi, and Vishal M. Patel. Syn2Real transfer learning for image deraining using gaussian processes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2726–2736, 2020. doi: 10.1109/CVPR42600.2020.00280.

Rongwei Yu, Peihao Zhang, and Jingyi Xiang. DG-RainDiff: Depth-guided dynamic message passing diffusion model for mixture of rain removal. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6280–6284, 2024. doi: 10.1109/ICASSP48485.2024.10445956.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14821–14831, 2021. doi: 10.1109/CVPR46437.2021.01458.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5728–5739, 2022. doi: 10.1109/CVPR52688.2022.00564.

Kaihao Zhang, Dongxu Li, Wenhan Luo, and Wenqi Ren. Dual attention-in-attention model for joint rain streak and raindrop removal. *IEEE Transactions on Image Processing*, 30:7608–7619, 2021. doi: 10.1109/TIP.2021.3108019.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Zhen Zou, Hu Yu, Jie Huang, and Feng Zhao. FreqMamba: Viewing Mamba from a frequency perspective for image deraining. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1905–1914, 2024. doi: 10.1145/3664647.3680862.

A Evaluation on LHP-Rain

We evaluate ControlIR on the official LHP-Rain (Guo et al., 2023) test set, with all compared methods trained from scratch under the same protocol. ControlIR achieves the best PSNR-Y, SSIM-Y, and LPIPS-RGB among the compared methods. It improves PSNR-Y over DehazeXL by 0.30 dB, and improves SSIM-Y and LPIPS-RGB over DEMore-Net by 0.0031 and 0.0075, respectively.

Table 6: Restoration results on LHP-Rain after training from scratch. The best and second-best scores are bold and underlined, respectively.

Method	PSNR-Y↑	SSIM-Y↑	LPIPS-RGB↓
ControlIR (Ours)	**36.5209**	**0.9581**	**0.0745**
DehazeXL	<u>36.2195</u>	0.9518	0.0888
DEMore-Net	36.0100	<u>0.9550</u>	<u>0.0820</u>
UDR-S^{2}Former	35.9445	0.9517	0.0880
FreqMamba	35.9311	0.9505	0.0919

The qualitative comparison uses ROI A to highlight rain removal and ROI B to highlight detail recovery. ControlIR removes more visible rain streaks in ROI A and preserves clearer roof edges and foreground railings in ROI B.

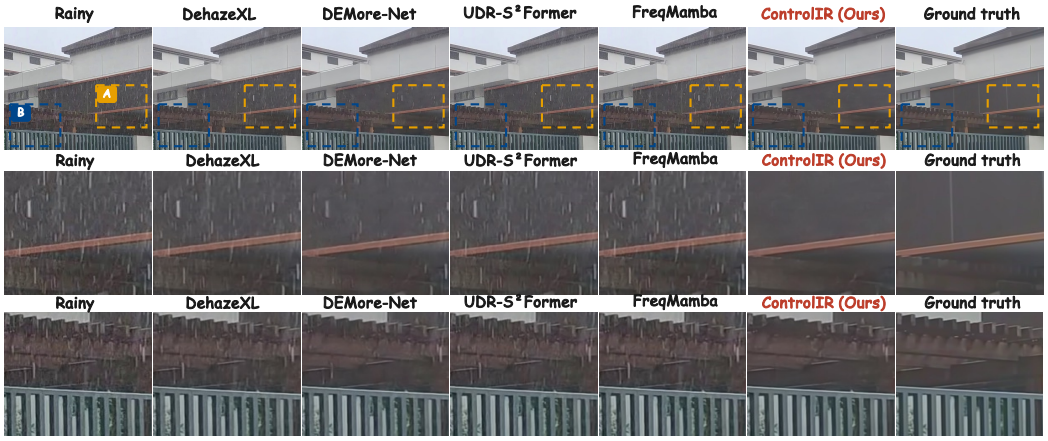


Figure 7: Qualitative comparison on LHP-Rain after training from scratch. Columns show the rainy input, DehazeXL, DEMore-Net, UDR-S^{2}Former, FreqMamba, ControlIR, and ground truth. Rows show the full image and aligned regions A and B, highlighting rain removal and detail recovery, respectively.

B Training, Evaluation, and Complexity Details

On RainCityscapes++, ControlIR is trained using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and no weight decay. The initial learning rate is 5×10^{-4} and follows a polynomial decay schedule with power 0.9. The depth branch uses one quarter of the base learning rate, while bias parameters use twice the base rate. Training uses a batch size of 2 for 150,000 updates. All rainy realizations of the same clean scene are assigned to the same training or validation split.

Internal quality results are summarized over three runs using seeds 3407, 2025, and 2026, with sample standard deviations. RainCityscapes++ evaluation excludes a four-pixel image border; LHP-Rain evaluation uses no additional border crop. SSIM uses an 11×11 Gaussian window with standard deviation 1.5. AlexNet-based LPIPS receives RGB images scaled to $[-1, 1]$.

Complexity estimates use an FP32 input of size $1 \times 3 \times 128 \times 256$ and count one multiply–accumulate as two FLOPs. The reported FLOPs combine THOP counts with a recurrence-based estimate of selective scanning. Parameter counts refer to active execution paths; for architecture variants A and B, these include the executed depth branch. Average inference time is measured on 720×480

RainCityscapes++ images using an NVIDIA RTX 3090. Some functional operations are outside the FLOPs profiling scope.

For the control-response visualization, discretization and write responses are independently min–max normalized, while spatial weights and correction magnitudes use independent percentile-rank displays. The colors therefore describe relative response levels within each displayed map.

C Comparison Protocol

The complete RainCityscapes++ comparison is reported in Table 1. Published comparison scores and source-protocol complexity values are retained where applicable. ControlIR reports active parameters and FLOPs under the internal profiling protocol.

D Inference Efficiency

Table 7: Model complexity and average inference time. Complexity is profiled with a $1\times3\times128\times256$ FP32 input. Inference time is measured on 720×480 RainCityscapes++ images using an NVIDIA RTX 3090 GPU.

Method	Params (M)↓	FLOPs (G)↓	Avg. Time (s)↓
AttentiveGAN	6.00	353.08	0.257
CMFNet	17.24	638.96	0.223
MSPFN	21.00	708.44	0.681
Syn2Real	2.61	20.92	0.378
MPRNet	3.64	148.55	0.230
Restormer	26.13	140.99	0.313
FreqMamba	8.93	30.18	0.854
LD-Net	0.03	1.98	0.356
SGID	13.87	156.40	0.525
UCL-Dehaze	19.45	78.80	0.035
DehazeXL	125.48	22.79	0.027
DAF-Net	1.72	13.15	0.057
DGNL-Net	4.03	19.73	0.093
CCN	3.75	245.85	0.369
UDR-S^{2}Former	8.53	21.58	0.605
MBA-RainGAN	8.62	294.32	0.433
DEMore-Net	19.25	58.01	0.073
ControlIR (Ours)	3.90	36.02	**0.019**

Average inference times are measured on 720×480 RainCityscapes++ images using an NVIDIA RTX 3090 GPU. The parameter count for ControlIR includes parameters active in the forward pass. Parameter and FLOPs values are reproduced from Table 1.

E Branch and Conditioning-Head Configuration

In the topology diagram, FR and TA denote field-resolved representation and token-aligned delivery. Its interface labels $\mathbf{E}_{p_\Delta}$, $\mathbf{E}_{p_B}$, and $\mathbf{E}_{p_M}$ correspond to $\mathbf{E}_\Delta$, $\mathbf{E}_B$, and $\mathbf{E}_M$ in the main text, with the unit index omitted in the diagram. The reliability target shown as $\widetilde{C}^{\mathrm{rel},T}$ in the overview is denoted by C^{target} in the training objective. The input normalization and its inverse use the fixed RGB vectors

$$\mu_{\mathrm{rgb}} = (0.485, 0.456, 0.406), \qquad \sigma_{\mathrm{rgb}} = (0.229, 0.224, 0.225). \tag{7}$$

They are broadcast over spatial locations, with $I^n = (I^r - \mu_{\mathrm{rgb}})/\sigma_{\mathrm{rgb}}$. For the predicted normalized-space residual R^n, reconstruction is

$$\widehat{I} = \mathrm{clip}((I^n + R^n) \odot \sigma_{\mathrm{rgb}} + \mu_{\mathrm{rgb}}, 0, 1)\,. \tag{8}$$

The global branch pools features onto its scan grid. Evidence maps are aligned to that grid for discretization and state writing, and to the local-feature grid for spatial correction. When source and target sizes differ, alignment uses bilinear interpolation with corner alignment disabled; maps already at the target size are retained. The global scan output is also interpolated to the local-feature resolution when needed. Directional base fields use forward differences with zero padding at the terminal boundaries. The 5×5 pooling in the image restoration objective has unit stride and zero padding.

Within the state-space module, the input projection generates the baseline discretization prediction, while the convolutional branch produces the state-input and state-readout maps. The input sequence

is denoted by x in the main text. Its channel mean $\bar{x}_t$ enters the write controller before the internal input projection. The recurrence input u_t is the per-head feature produced after that projection, the internal convolution, and SiLU activation.

The local branch of each ControlSSM unit applies two 3×3 convolutions, each followed by ReLU. Both convolutions use 64 output channels. Their dilation is 2 in unit 1 and 8 in unit 2. The global branch receives the local feature. Channel concatenation of the fused feature and the original unit input changes the widths from 192 to 256 in unit 1 and from 320 to 384 in unit 2.

Table 8: Evidence and conditioning heads. Channel and feature widths are listed as input $\rightarrow$ output. h_k and g_k denote the SSM head and group counts in unit k.

Head	Layers	Output operation
Reliability h_C	Conv 3×3, $3 \rightarrow 16$; GELU; Conv 1×1, $16 \rightarrow 1$	Sigmoid
Refinement h_{ref}	Conv 1×1, $5 \rightarrow 16$; GELU; Conv 1×1, $16 \rightarrow 4$	Field refinement
Discretization h_Δ	Linear $5 \rightarrow 16$; SiLU; Linear $16 \rightarrow h_k$	Tanh
State writing h_B	Linear $5 \rightarrow 16$; SiLU; Linear $16 \rightarrow g_k$	Tanh
Spatial weight h_M	Conv 3×3, $67 \rightarrow 32$; ReLU; Conv 1×1, $32 \rightarrow 1$	Sigmoid

The five inputs to h_B comprise the scalar current-token channel mean and the four selected evidence fields. The 67 inputs to h_M comprise the 64-channel absolute local–global difference and the three selected fields. The spatial weight has one channel and is broadcast across fusion channels. The current configuration uses one SSM group, so the write coefficient β_t in the main text is scalar at each token and is shared across the state dimension.

The scalar strengths α_Δ and α_B are initialized to 0.01, and α_g to 0.1. They are optimized directly as learnable scalars. The last layers of h_Δ, h_B, and h_M use zero biases and normally distributed weights with zero mean and standard deviation 0.001.

F Training Objectives

ControlIR is jointly trained with an image restoration loss and an auxiliary prior loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rest}} + \mathcal{L}_{\text{prior}}. \tag{9}$$

The restoration loss equally weights the low-frequency Charbonnier error, high-frequency Charbonnier error, and MS-SSIM structural constraint (Wang et al., 2003). Gradients from this loss train the control heads through the reconstructed image. The auxiliary prior loss supervises relative depth and its directional and local variations before evidence refinement, together with reliability prediction. The reliability target is constructed from errors between current geometry predictions and teacher targets, with gradients stopped through target construction. Teacher-generated targets are used only during training. Complete loss definitions and target construction are given in Appendix G.

G Loss Definitions and Supervision Details

Applying 5×5 average pooling gives the low-frequency component I_{low}, and $I_{\text{high}} = I - I_{\text{low}}$ is the high-frequency component. The restoration loss combines Charbonnier errors in both components with an MS-SSIM structural constraint (Wang et al., 2003)

$$\begin{aligned} \mathcal{L}_{\text{rest}} = \tfrac{1}{3}\big[& \operatorname{Mean} \rho(\widehat{I}_{\text{low}}, I^{\text{gt}}_{\text{low}}) \\ & + \operatorname{Mean} \rho(\widehat{I}_{\text{high}}, I^{\text{gt}}_{\text{high}}) \\ & + 1 - \text{MS-SSIM}(\widehat{I}, I^{\text{gt}})\big]. \end{aligned} \tag{10}$$

Here, $\widehat{I}$ and I^{gt} are restored and paired clean images, $\rho(a, b) = \sqrt{(a-b)^2 + \epsilon^2}$ with $\epsilon = 10^{-3}$, and Mean averages over samples, channels, and spatial positions. The three terms are equally weighted.

The Charbonnier penalty is $\rho(a, b) = \sqrt{(a-b)^2 + 10^{-6}}$.

Auxiliary prior supervision constrains relative depth and its directional and local variations. The depth prediction Z and the offline teacher target are independently min–max normalized over their

common valid support, yielding Z^S and Z^T. Horizontal and vertical absolute differences and absolute deviations from the local 3×3 mean are then computed, with each variation component independently normalized.

We denote these supervision components by Z_q^S and Z_q^T, where $q \in \mathcal{Q} = \{z, x, y, l\}$ indexes relative depth, horizontal variation, vertical variation, and local variation, respectively, with $Z_z^S = Z^S$ and $Z_z^T = Z^T$. Their corresponding losses are

$$\mathcal{L}_q^{\text{sup}} = \text{MMean}_{\mathcal{V}_q} \rho(Z_q^S, Z_q^T), \qquad q \in \mathcal{Q}. \tag{11}$$

Here, $\text{MMean}_{\mathcal{V}_q}$ averages over the valid support $\mathcal{V}_q$ of that component. The quantities Z_q^S are derived from the depth prediction before the evidence refinement described in the main text.

The reliability target is constructed from prediction errors in depth and its variation components. At position u, averaging the absolute prediction–teacher errors of valid components gives e_u, and the target is

$$C_u^{\text{target}} = \text{stopgrad}(\text{clip}(1 - e_u, 0, 1)) . \tag{12}$$

Smaller errors correspond to higher target reliability. The target is updated with the current predictions. The operator clip restricts it to $[0, 1]$, and stopgrad prevents gradients from propagating through the target construction.

Reliability prediction error uses the Charbonnier penalty and is averaged over valid locations

$$\mathcal{L}_C^{\text{sup}} = \text{MMean}_{\mathcal{V}_C} \rho(C^{\text{rel}}, C^{\text{target}}). \tag{13}$$

Here, $\mathcal{V}_C$ contains locations with at least one valid depth or variation component. The five losses for relative depth, two directional variations, local variation, and reliability are equally averaged to form the auxiliary prior loss

$$\mathcal{L}_{\text{prior}} = \tfrac{1}{5} \left(\sum_{q \in \mathcal{Q}} \mathcal{L}_q^{\text{sup}} + \mathcal{L}_C^{\text{sup}} \right) . \tag{14}$$

Masked supervision normalization uses extrema computed independently for each sample and channel over the common finite support. Valid constant maps and sample/channel pairs with no valid positions map to zero. The validity mask is retained for subsequent reductions. For a valid support $\mathcal{V}$, the normalization at $u \in \mathcal{V}$ is

$$\text{MSN}_{\mathcal{V}}(X)_u = \frac{X_u - \min_{v \in \mathcal{V}} X_v}{\max_{v \in \mathcal{V}} X_v - \min_{v \in \mathcal{V}} X_v + 10^{-6}}. \tag{15}$$

Values outside the valid support are set to zero. The controller normalization also maps a spatially constant field to zero. For a valid support $\mathcal{V}$, the masked reduction is

$$\text{MMean}_{\mathcal{V}}(X) = \frac{\sum_u \mathbb{I}_{u \in \mathcal{V}} X_u}{\max(1, \sum_u \mathbb{I}_{u \in \mathcal{V}})}. \tag{16}$$

An empty support therefore contributes zero. Horizontal and vertical component validity requires both adjacent pixels. The 3×3 local contrast excludes invalid interior neighbors from its average and retains the zero-padding convention at the image boundary.

The reliability target averages the errors of the available normalized geometry components at each position. Its support is the union of the component-valid supports, so only positions with at least one available component contribute to its loss. The target is detached from gradient computation, and reliability predictions are clamped to $[0, 1]$ before evaluation. The input-grid controller normalization is

$$\text{SN}(X) = \text{clip}\left(\frac{X - \min_{\Omega_0} X}{\max_{\Omega_0} X - \min_{\Omega_0} X + 10^{-6}}, 0, 1 \right), \tag{17}$$

where Ω_0 is the input-image grid. Extrema are computed independently for each sample and channel. The exact reliability error over the available geometry components is

$$e_u = \frac{\sum_{\substack{q \in \mathcal{Q} \\ u \in \mathcal{V}_q}} |Z_{q,u}^S - Z_{q,u}^T|}{|\{q \in \mathcal{Q} : u \in \mathcal{V}_q\}|}, \qquad u \in \mathcal{V}_C = \bigcup_{q \in \mathcal{Q}} \mathcal{V}_q, \tag{18}$$

where $\mathcal{Q} = \{z, x, y, l\}$ and Z_q^S, Z_q^T, and $\mathcal{V}_q$ are the supervision components and valid domains defined in this section.

H Structural Variant and Evidence Transformations

StateConcat receives pooled pre-scan features X_p with 64 channels and the six aligned evidence fields E_p. It computes

$$\widetilde{X}_p = X_p + P_{1\times1}([X_p; E_p]), \tag{19}$$

where P is a bias-free $70 \to 64$ convolution initialized to zero. The result enters the bidirectional scan with native discretization and $\beta = 1$. Fusion retains the local and global branches and α_g, with $M = 1$; the final stage concatenation is unchanged.

In the representation–delivery study, each interface first aligns and selects its fields, then applies field collapse and spatial pooling when enabled. The external reliability gates are extracted after these transformations. The upstream evidence constructor, current-token input, and local–global feature-difference input remain unchanged.

I Offline Supervision

Clean targets are processed offline by VGGT (Wang et al., 2025a) followed by PriorDA (Wang et al., 2025c) ViT-B v1.0, which incorporates a frozen Depth Anything V2 (Yang et al., 2024) ViT-L model. For RainCityscapes++, the clean RGB images are resized to 720×480 before teacher processing. Refined depth and confidence outputs are cached as float32 NPY arrays. The reliability supervision target is constructed from the student–teacher geometry discrepancy and detached from gradients. For LHP-Rain, teacher targets are generated from full clean images; training crops use the normalization extrema of their source full image.