DECOUPLE TO RECONCILE: EXPLICIT INTER- AND INTRA-EXPOSURE MOTION MODELING FOR NIGHT IMAGE RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Photon scarcity in low-light environments makes high-quality night image restoration from a single exposure inherently challenging. Dual-exposure imaging offers an effective solution by exploiting the complementary characteristics of short- and long-exposure images. However, motion-induced inter-frame misalignment and intra-exposure blur hinder their effective collaborative restoration. To this end, we propose an explicit inter- and intra-exposure motion decoupling framework that separately models cross-exposure displacement and within-exposure motion blur, enabling accurate alignment and collaborative restoration of complementary dual-exposure information. Specifically, I^2EM follows a two-stage decouple-to-reconcile paradigm. The decoupling stage separately models cross-exposure displacement and intra-exposure blur using bidirectional optical flow and exposure-specific motion-conditioned kernels. The reconciliation stage restores and aligns long-exposure features against the sharper short-exposure reference, then integrates complementary information through confidence-guided multi-scale fusion and residual denoising. On D2, I^2EM achieves state-of-the-art performance among the compared methods, demonstrating the effectiveness of explicit inter- and intra-exposure motion modeling for dual-exposure night-image restoration.

1 INTRODUCTION

Night-image restoration aims to recover a clean, sharp, and visually faithful scene from observations captured under insufficient illumination. It provides essential support for the reliable operation of mobile photography (Hasinoff et al., 2016), video surveillance (Li et al., 2020), autonomous driving (Li et al., 2024), and robotic perception (Wang et al., 2023) under low-visibility conditions. To meet these demands, early studies focused primarily on restoration under a single exposure setting, including single-frame RAW enhancement for photon-limited observations (Chen et al., 2018), short-exposure burst restoration using temporal redundancy (Mildenhall et al., 2018), and learning-based single-image deblurring (Nah et al., 2017, Tao et al., 2018, Zamir et al., 2021, Cho et al., 2021). However, limited to a single exposure condition, these methods cannot simultaneously exploit the sharp structures of short exposures and the high-SNR information of long exposures.

Recently, short–long exposure restoration has emerged as a promising direction for overcoming the information limitations of a single exposure by jointly exploiting the sharp structures of short exposures and the high-SNR information of long exposures (Zhang et al., 2025, Rim et al., 2024, Shekarforoush et al., 2023, Meng et al., 2026, Zhao et al., 2025, Liu et al., 2026). Representative methods generally follow two paradigms. One adopts a sequential restoration pipeline (Zhao et al., 2022, Liu et al., 2026) that successively performs long-exposure deblurring (Nah et al., 2017, Zamir et al., 2021, Yuan et al., 2020, Fang et al., 2023), cross-exposure alignment, and complementary information fusion. The other employs learned sampling offsets to implicitly absorb motion, enabling image reconstruction through joint feature alignment and aggregation (Zhao et al., 2022, Yuan et al., 2020). Nevertheless, existing dual-exposure methods entangle inter-frame displacement with intra-exposure blur, leading to erroneous information transfer, residual blur, and ghosting.

To elucidate this issue, we make two observations from Fig. 1. First, the temporal gap between successive captures induces inter-frame misalignment, causing the same scene content to appear at

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.

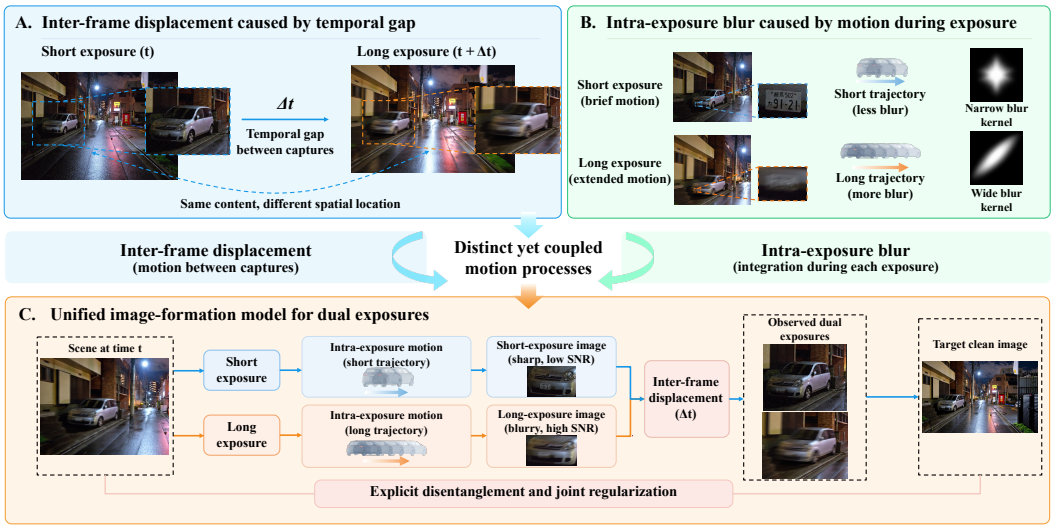

Figure 1: **Dual-Exposure Motion.** (A) The short exposure is captured first at center time t_s, followed by the long exposure at $t_l = t_s + \Delta t$. Scene motion over Δt shifts corresponding content between the two observations. (B) Motion integrated during the short exposure traces a shorter trajectory and produces a narrower blur kernel, while the longer integration interval produces a wider trajectory-dependent kernel. (C) Both observations arise from the same time-varying scene. Their relation combines displacement between exposure centers with radiance accumulation within each exposure, linking the noisy sharp short exposure and the high-SNR blurred long exposure to the clean short-reference target.

different spatial locations. Second, motion during each finite exposure integrates scene radiance along local trajectories, producing intra-exposure blur that is particularly pronounced in the long-exposure image. These observations reveal that inter-frame displacement and intra-exposure blur arise from distinct yet coupled motion processes. Reliable dual-exposure restoration therefore requires their explicit disentanglement and joint regularization within a unified image-formation model.

To this end, we propose an **I**nter- and **I**ntra-**E**xposure **M**otion **M**odeling framework (I^{2}EM) that follows a decouple-to-reconcile paradigm for dual-exposure restoration. During decoupling, two independently parameterized exposure-specific encoders extract multi-scale features from the short- and long-exposure observations. A bidirectional motion estimator predicts cross-exposure sampling fields and confidence maps, while a motion-conditioned kernel estimator predicts exposure-specific kernels for within-exposure integration. Observation reconstruction and centered-kernel regularization connect the two representations while encouraging distinct roles. During reconciliation, long-exposure features are restored with the long-exposure kernel and sampled onto the short-exposure reference coordinates. Confidence-aware multi-scale gating combines the aligned long-exposure evidence with short-exposure features. A short-reference residual decoder produces an initial estimate, which is further refined by an image-space residual head using the aligned long exposure and its motion guidance.

The main contributions are as follows.

- We reformulate successive dual-exposure image restoration as a problem of factorized modeling and coordinated fusion of cross-exposure displacement and intra-exposure motion integration.
- We propose a decouple-to-reconcile framework that assigns the two motion effects to bidirectional optical flow and exposure-specific motion-conditioned kernels, yielding physically interpretable representations with clearly delineated roles. The disentangled representations further coordinate feature restoration, reference-centered alignment, and confidence-aware fusion for reliable integration of complementary dual-exposure information.
- On the D2 benchmark, our method achieves the best PSNR and SSIM among the compared methods, with 40.78 dB PSNR, 0.9783 SSIM, and a competitive LPIPS of 0.054. Extensive experiments further validate the effectiveness of the factorized motion modeling and dual-exposure

reconciliation mechanisms. Reconstruction consistency and kernel-centering constraints preserve the distinct semantics of the two representations.

2 Related Work

2.1 Low-light and multi-exposure restoration

Low-light restoration is constrained by the evidence retained during acquisition. Single-frame RAW enhancement and physically informed unprocessing recover photon-limited measurements from one observation (Chen et al., 2018, Brooks et al., 2019), while short-exposure bursts exploit temporal redundancy for denoising and local alignment (Mildenhall et al., 2018). Dual- and multi-exposure systems instead combine observations with complementary signal and blur characteristics. Their acquisition settings range from successive short–long pairs and exposure brackets to synchronized optical paths, complementary sensors, and interleaved exposure patterns (Zhao et al., 2022, Zhang et al., 2025, Rim et al., 2024, Shekarforoush et al., 2023, Meng et al., 2026, Zhao et al., 2025, Liu et al., 2026). These settings differ in temporal offset, viewpoint, sampling pattern, and observation count, so they do not share the same alignment or fusion assumptions. Consequently, existing formulations do not directly resolve the successive-pair setting because they generally neither separate displacement between exposure centers from blur accumulated within each exposure nor constrain their interaction through a unified image-formation process.

2.2 Image deblurring and blur modeling

Image deblurring methods can be organized by how explicitly they represent degradation. Direct restoration networks absorb motion, blur, and image priors into an end-to-end mapping. This family spans convolutional single-image models (Nah et al., 2017, Kupyn et al., 2019, Zamir et al., 2021, Cho et al., 2021, Chen et al., 2021, 2022), recurrent video models (Zhong et al., 2020, Pan et al., 2020), and Transformer backbones (Liang et al., 2021, Wang et al., 2022, Zamir et al., 2022). Prompt-conditioned restoration and attention architectures adapted from broader low-level vision provide further architectural comparisons (Potlapalli et al., 2023, Chen et al., 2023), while state-space and diffusion models introduce long-range or generative priors (Kong et al., 2025a,b). Explicit approaches instead condition restoration on learned degradation descriptors, flow-guided filters, or spatially varying kernels (Li et al., 2022, Yuan et al., 2020, Fang et al., 2023). These formulations primarily address a single degraded image or a same-modality sequence. Applied directly to successive dual exposures, a shared latent motion representation provides no explicit separation between cross-exposure displacement and exposure-time integration, whereas unconstrained independent estimates may reduce the same reconstruction residual without preserving distinct functions. The required representation must therefore separate the two effects and couple them through the observation process.

2.3 Optical flow and feature alignment

Optical-flow methods estimate correspondence through local or global feature correlation. PWC-Net combines feature pyramids, warping, and local cost volumes (Sun et al., 2018), whereas RAFT uses all-pairs correlation with recurrent refinement (Teed & Deng, 2020). Both rely on cross-frame similarity, which is weakened in successive short and long exposures by noise, saturation, radiometric differences, and temporal integration. Optical flow can locate scene content between exposure centers but cannot represent the set of positions integrated within one exposure. A spatial kernel can approximate that local accumulation but does not determine cross-exposure correspondence. This complementary limitation motivates estimating correspondence from exposure-specific features, representing exposure-time integration with spatial kernels, and constraining both variables within a shared reconstruction process.

3 Method

Figure 2 presents I^2EM as a decouple-to-reconcile framework. Decoupling preserves exposure-specific evidence while factorizing motion into bidirectional sampling fields for cross-exposure

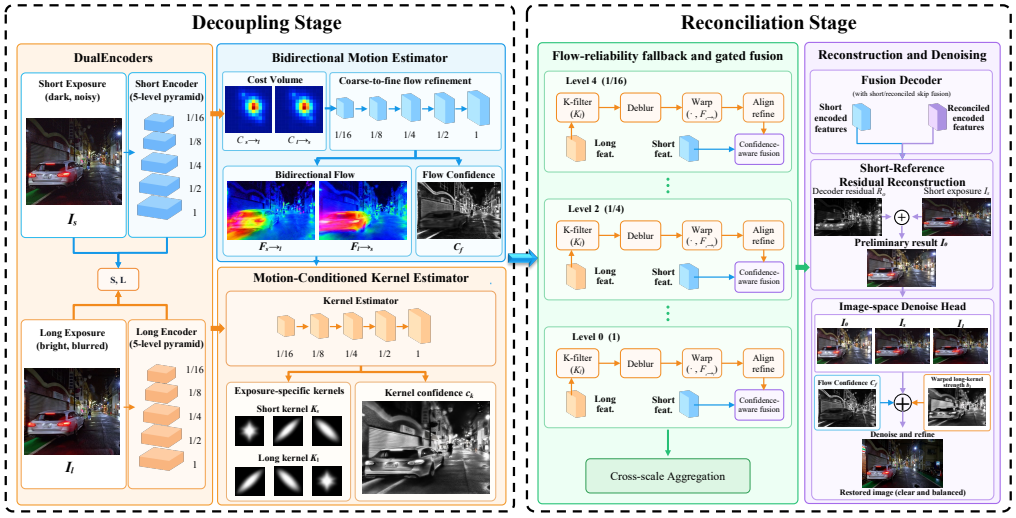

Figure 2: Overview of I^{2}EM. In the **decoupling stage** (left), exposure-specific encoders form five-level short- and long-exposure pyramids (S, L). Bidirectional cost volumes and coarse-to-fine refinement estimate $F_{s\to l}$ and $F_{l\to s}$ with flow confidence C_f, while the motion-conditioned kernel estimator predicts exposure-specific kernels (K_s, K_l) and kernel confidence C_k. The blur-strength map b_l is derived from the second moment of K_l and is distinct from C_k. In the **reconciliation stage** (right), the illustrated levels 4, 2, and 0 represent the full pyramid. At each scale, long-exposure features are restored with K_l, sampled by $F_{s\to l}$, refined, and fused with short-exposure features according to correspondence reliability. Cross-scale aggregation and the fusion decoder predict the short-reference residual R_0 and preliminary result $I_0 = I_s + R_0$. The image-space denoise head then combines I_0, I_s, aligned I_l, C_f, and transported b_l to produce the restored image.

displacement and spatially varying kernels for exposure-time integration. Reconciliation then uses these representations to restore and transfer long-exposure evidence to the short-exposure coordinates, where reliability-aware multi-scale fusion and residual refinement recover the output.

3.1 Decoupling

The decoupling stage maps (I_s, I_l) to exposure-specific feature pyramids and a factorized motion representation. The five-scale fields $\{F^r_{s\to l}, F^r_{l\to s}\}_{r=0}^4$ and kernels $\{K^r_s, K^r_l\}_{r=0}^4$ are passed jointly to reconciliation, where $r = 0$ denotes the input grid and $r = 4$ the coarsest grid. The sampling fields transfer evidence between exposure centers, whereas the spatial kernels characterize local accumulation over each exposure interval, thereby preserving complementary physical roles.

3.1.1 Exposure-specific encoding and correspondence

Independent encoders construct short- and long-exposure pyramids $\{S^r\}$ and $\{L^r\}$ that retain their distinct degradation statistics while supplying shared scene evidence to motion estimation and restoration. Bidirectional fields account for displacement between exposure centers. Projected features are first compared over a local displacement set at the coarsest level. Each finer level then refines the field from the source feature, the sampled target feature, and the current estimate. This coarse-to-fine construction assigns cross-exposure correspondence to flow rather than allowing it to be absorbed by the exposure kernels.

For the two directions, the coarsest-level correlation is defined as

$$\mathcal{C}^4_{a\to b}(\mathbf{p}, \boldsymbol{\delta}) = \frac{1}{D_4} \left\langle \phi^4_a(\mathbf{p}), \phi^4_b(\mathbf{p} + \boldsymbol{\delta}) \right\rangle, \qquad \boldsymbol{\delta} \in [-d, d]^2, \quad (a, b) \in \{(s, l), (l, s)\}. \tag{1}$$

Here ϕ^4_a and ϕ^4_b are the projected coarsest-level features and D_4 is their channel dimension. Shared prediction operators parameterize the two directions, while the resulting fields remain distinct. Their

scale-specific confidence C_f^r subsequently regulates how much transferred long-exposure evidence contributes to reconciliation.

3.1.2 EXPOSURE-TIME INTEGRATION

Exposure-time integration is assigned to two exposure-specific spatial kernel fields. The kernel estimator takes the paired feature pyramids and both directional fields as its input, so the kernels can use correspondence context without becoming an alternative correspondence variable. A bidirectional feature-pyramid interaction supplies scale context, and separate heads produce the short- and long-exposure kernels together with a kernel confidence map C_k^r. Each kernel is a nonnegative distribution over a local support. Specifically, its predicted logits q_e^r are normalized over Ω_r as

$$K_{e,u,v}^r(\mathbf{p}) = \frac{\exp q_{e,u,v}^r(\mathbf{p})}{\sum_{(u',v')\in\Omega_r} \exp q_{e,u',v'}^r(\mathbf{p})}, \qquad e \in \{s,l\},\ (u,v) \in \Omega_r. \tag{2}$$

The centroid of the normalized kernel is

$$\boldsymbol{\mu}_e^r(\mathbf{p}) = \sum_{(u,v)\in\Omega_r} K_{e,u,v}^r(\mathbf{p}) \begin{bmatrix} u - u_0 \\ v - v_0 \end{bmatrix}, \qquad e \in \{s,l\}, \tag{3}$$

where (u_0, v_0) is the support center. The training objective applies the centered-kernel penalty to the full-resolution kernels, with a larger weight on the long-exposure kernel.

The centroid is the first spatial moment of the kernel. Penalizing it discourages the exposure kernels from reproducing translation already represented by the sampling fields. The normalized kernels therefore describe local exposure integration, while the centered second moment of the long-exposure kernel yields a normalized blur-strength map b_l^r for restoration conditioning. Their coupling through feature estimation and observation reconstruction preserves this division of roles while allowing displacement and blur to explain the observations jointly.

3.2 RECONCILIATION

The reconciliation stage converts the factorized motion representation into a reliability-weighted reconstruction on the short-exposure coordinate system. It first restores the long-exposure branch under kernel-derived conditioning, then transfers and gates complementary evidence according to correspondence confidence, and finally performs short-reference residual reconstruction followed by image-space refinement.

3.2.1 KERNEL-CONDITIONED RESTORATION AND TRANSFER

Before cross-exposure transfer, the long-exposure path is conditioned on the long-exposure kernel, the independently predicted kernel confidence, and the kernel-derived blur strength. The restored long-exposure feature is

$$\widetilde{L}^r = P^r(L^r + R^r[L^r, \mathcal{B}(L^r, K_l^r), C_k^r, b_l^r]), \tag{4}$$

where R^r and P^r are learned residual and projection blocks. The blurred feature in the condition is generated with the same spatially varying kernel that is used to describe the long exposure, making the degradation representation available to the feature restoration path.

The restored long-exposure feature is sampled with $F_{s\to l}^r$ on the short grid. A valid-region mask removes out-of-frame samples, and the flow confidence is converted into a soft reliability weight with stabilizer $\tau_c = 0.3$. This weight provides a continuous fallback to the short-exposure feature when the transferred evidence is unreliable. A learned gate then selects between the short feature and the reliability-weighted long feature,

$$Z^r = Q^r(G^r \odot S^r + (1 - G^r) \odot A^r). \tag{5}$$

3.2.2 REFERENCE RECONSTRUCTION AND REFINEMENT

Residual cross-scale blocks propagate the reconciled features, and a decoder with short-branch skip connections predicts $I_0 - I_s$ in the short-exposure coordinates. Anchoring I_0 to the sharper observation preserves its structural evidence. An image-space refinement path then conditions its residual on I_0, I_s, the valid aligned long image, flow confidence, and the transported long-kernel blur strength, producing $\widehat{X}$ from both feature-level coordination and image-level reliability cues.

3.3 TRAINING OBJECTIVE

The objective gives the factorized variables distinct training roles. Pixel and structural losses anchor the restored image, flow cycle and edge-aware smoothness terms constrain correspondence, and observation reconstruction ties the exposure kernels to the two measured images. Kernel centering supplies the first-moment prior that prevents the kernels from reintroducing cross-exposure translation. The observation terms are defined on the predicted reference image, so the same output is responsible for both restoration fidelity and observation reconstruction.

The exposure kernels are linked to the predicted short-reference image through two RGB observation regularizers. The short-exposure term is

$$\mathcal{L}_s^{\mathrm{obs}} = \rho\Big(\mathcal{B}(\widehat{X}, K_s^0), I_s\Big). \tag{6}$$

The long-exposure term is

$$\mathcal{L}_l^{\mathrm{obs}} = \rho\Big(\mathcal{B}(\mathcal{W}(\widehat{X}, F_{l\to s}^0), K_l^0), I_l\Big). \tag{7}$$

The losses are evaluated on valid kernel support and valid flow samples. The same observation construction is applied to the four coarser levels after matching the image resolution, and their mean is added with weight 0.5. The weighted objective is

$$\mathcal{L} = \lambda_{\mathrm{pix}}\mathcal{L}_1 + \lambda_{\mathrm{str}}\mathcal{L}_{\mathrm{SSIM}} + \lambda_{\mathrm{cyc}}\mathcal{L}_{\text{flow-cyc}} + \lambda_{\mathrm{sm}}\mathcal{L}_{\mathrm{smooth}} + \lambda_s^{\mathrm{obs}}\mathcal{L}_s^{\mathrm{obs}} + \lambda_l^{\mathrm{obs}}\mathcal{L}_l^{\mathrm{obs}} + \lambda_{\mathrm{ctr}}\mathcal{L}_{\mathrm{ctr}}. \tag{8}$$

The coefficients are treated as fixed hyperparameters, with their numerical values specified in the implementation details. Together, the reconstruction, correspondence, and centering terms require the output and the two motion representations to explain the observations through complementary roles rather than through an unconstrained shared residual.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

The primary evaluation uses the public D2 benchmark released with D2HNet (Zhao et al., 2022), which provides successive short- and long-exposure observations with paired sharp references. Following the official D2HNet data-generation rules, we construct a non-overlapping validation set of 1,192 scenes at 5120×2880 for all main comparisons and controlled ablations. Peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) measure pixel fidelity, structural agreement, and perceptual distance. All methods are evaluated on the same platform with an NVIDIA RTX 5090 GPU with 32 GB memory, a 16-core Intel Core i9 CPU, and 128 GB DDR5 system memory. Runtime measurement discards 10 warm-up runs and reports statistics over 100 synchronized CUDA-event trials. The methods in Table 1 retain their stated input settings. Single-image methods receive the long exposure, video methods receive their required neighboring observations, and dual-exposure methods receive the same short–long pair. STGDNet uses complementary sensor measurements and is included as a heterogeneous reference, whereas D2HNet provides the task-matched dual-exposure comparison.

The network is trained for 100 epochs with Adam (Kingma & Ba, 2015). Training applies calibrated noise synthesis and photometric augmentation to random 1024^2 source crops before resizing them to the configured 512^2 network input. Evaluation instead starts from the native 5120×2880 validation images and uses the same full-resolution inference and metric aggregation for every main comparison and controlled ablation. The checkpoint with the highest validation PSNR is retained.

Robustness experiments use self-collected Xiaomi Mi Note 10 sequences acquired following the public D2HNet capture procedure, and vary sensor noise, exposure ratio, and resolution while retaining the metric definitions above. The cross-resolution results use this Xiaomi dataset.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Table 1: Quantitative comparison on the generated D2 2880p validation set. ΔPSNR denotes the PSNR of I^{2}EM minus that of each comparison method, in dB. Parameter counts are in millions (M), and runtime is in milliseconds (ms) at 2880p. The best and second-best restoration results are shown in bold and underlined.

Method	Input	Restoration quality				Params	Time
		PSNR↑	ΔPSNR	SSIM↑	LPIPS↓		
DeepDeblur (Nah et al., 2017)	Single	36.55	+4.23	0.9636	0.095	42.3	14
DeblurGAN-v2 (Kupyn et al., 2019)	Single	36.82	+3.96	0.9655	0.090	60.9	2
CDVD-TSP (Pan et al., 2020)	Multi	37.00	+3.78	0.9668	0.087	16.6	23
ESTRNN (Zhong et al., 2020)	Multi	37.10	+3.68	0.9675	0.088	20.5	25
HINet (Chen et al., 2021)	Multi	37.22	+3.56	0.9685	0.083	88.7	6
MPRNet (Zamir et al., 2021)	Single	37.08	+3.70	0.9684	0.086	23.5	10
SwinIR (Liang et al., 2021)	Single	37.70	+3.08	0.9712	0.079	12.1	14
MIMO-UNet+ (Cho et al., 2021)	Single	37.31	+3.47	0.9690	0.083	16.1	9
Restormer (Zamir et al., 2022)	Single	37.87	+2.91	0.9717	0.077	26.1	12
D2HNet (Zhao et al., 2022)	Dual	37.95	+2.83	0.9723	0.076	35.2	12
NAFNet (Chen et al., 2022)	Single	37.70	+3.08	0.9710	0.079	23.4	11
Uformer (Wang et al., 2022)	Single	37.55	+3.23	0.9702	0.081	28.3	18
UFPNet (Fang et al., 2023)	Single	37.72	+3.06	0.9712	0.079	80.3	29
HAT (Chen et al., 2023)	Single	37.85	+2.93	0.9715	0.078	12.4	15
PromptIR (Potlapalli et al., 2023)	Single	37.45	+3.33	0.9695	0.082	87.2	45
EVSSM (Kong et al., 2025a)	Single	38.09	+2.69	0.9727	0.072	17.1	3
DeblurDiff (Kong et al., 2025b)	Single	37.55	+3.23	0.9703	**0.050**	386.2	304
STGDNet (Meng et al., 2026)	Multi-sensor	<u>38.12</u>	+2.66	<u>0.9729</u>	0.072	28.4	15
I^{2}EM	Dual	**40.78**	–	**0.9783**	<u>0.054</u>	133.79	78

Table 1 compares convolutional, recurrent, Transformer, kernel-based, state-space, diffusion, dual-exposure, and complementary-sensor baselines under the same 2880p evaluation. I^{2}EM obtains the highest PSNR and SSIM. Relative to the strongest competing PSNR result from STGDNet, it gains 2.66 dB and 0.0054 SSIM while reducing LPIPS by 0.018. D2HNet provides the input-matched comparison on the same short–long pair. I^{2}EM improves it by 2.83 dB and 0.0060 SSIM while reducing LPIPS by 0.022, a 28.9% relative LPIPS reduction. DeblurDiff reaches the lowest LPIPS of 0.050 but remains 3.23 dB below I^{2}EM in PSNR, indicating a different balance between perceptual similarity and pixel fidelity.

Figure 3 relates the visual differences to the two motion effects modeled by I^{2}EM. In panel (a), several baselines retain softened or displaced strokes because long-exposure integration changes both edge extent and apparent correspondence. Mapping restored long-exposure evidence to the short-reference grid preserves the letter spacing and circular boundary. In panel (b), the spatially varying long-exposure kernel contracts motion trails before confidence-aware alignment, which reduces duplicated vehicle edges and localizes the highlights. The two scenes show that correspondence and exposure integration affect different image structures and must be coordinated in the reconstruction path.

4.3 ROBUSTNESS ACROSS ACQUISITION CONDITIONS

The visual trends in Fig. 4 connect the main result to controlled changes in acquisition and scale. Within the calibrated noise range, all methods vary little. Beyond that range, the proposed representation degrades more slowly because the fusion gate can preserve high-SNR long-exposure evidence where correspondence remains reliable while retaining short-exposure structure around uncertain motion. Increasing the exposure ratio lengthens radiance integration and weakens local geometric evidence in the long branch, producing the gradual decline in panel (b). Multi-scale training reduces the 2880p-to-1440p drop by exposing the flow and kernel predictors to different pixel-domain motion supports. The three trends attribute robustness to complementary sensing, reliability-aware reconciliation, and scale-diverse training at their respective stages.

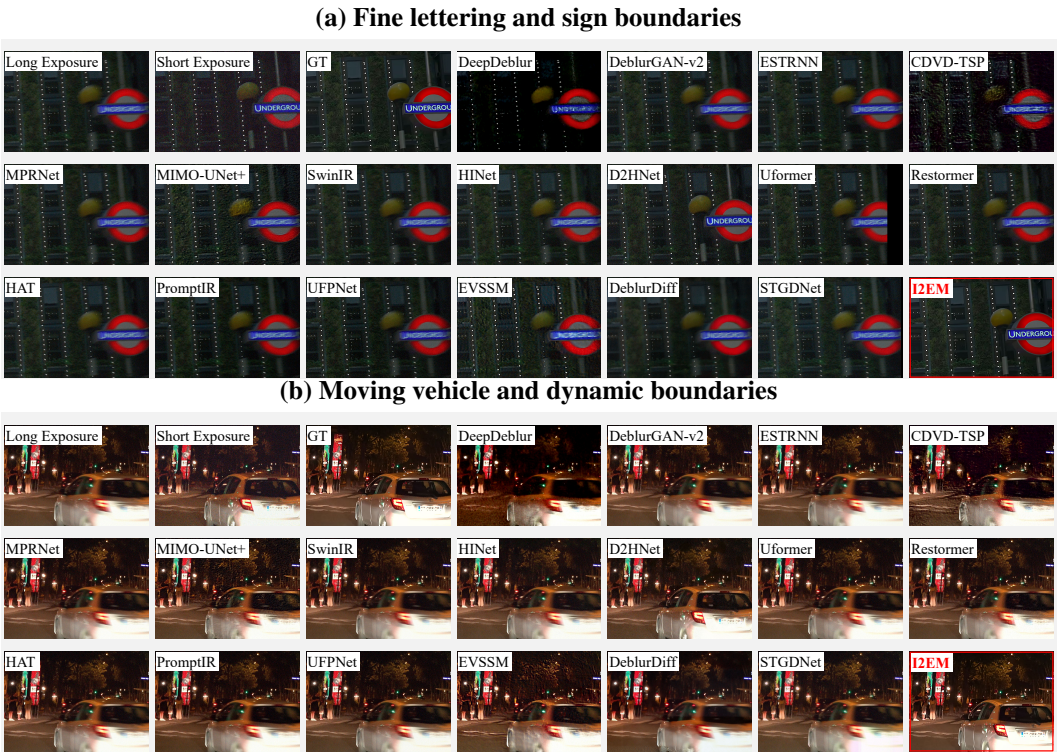

Figure 3: Qualitative comparison on two D2 2880p scenes. Each grid uses the common input, reference, baseline, and output order. The red box marks I^{2}EM. Panel (a) isolates thin lettering and circular sign boundaries affected by cross-exposure displacement, whereas panel (b) isolates moving vehicle contours and local highlights affected by exposure-time integration. I^{2}EM preserves the reference geometry while suppressing both displacement-induced duplication and integration-induced trails.

4.4 Ablation Studies

Table 2 groups factorization, decoupling, and reconciliation controls under the same full-resolution protocol. Parameter counts identify capacity changes.

The implicit joint-motion variant uses 68.42M parameters and reaches 39.29 dB. At the full 133.79M capacity, one-way flow and sequential flow–kernel estimation still reduce PSNR by 0.58 and 0.42 dB. Bidirectional fields provide reciprocal geometric context to kernel prediction and cycle consistency, while joint estimation lets correspondence condition the exposure-integration support. Removing kernel centering costs 0.37 dB because kernel displacement can then absorb part of the cross-exposure translation. Applying the short-to-long blur loss or blurring before warping costs 1.24 and 0.79 dB, respectively. Both changes violate the formation order in which the short-reference estimate is first transported to the long-exposure coordinates and then integrated over the long exposure. Predicting a residual from the long exposure produces the largest decrease of 1.78 dB because the reconstruction target is defined on the short-reference grid. Finally, the slightly larger 134.51M concatenation variant remains 0.31 dB below confidence-aware gating, showing that spatial reliability is the effective control signal for selecting complementary evidence.

5 Conclusion

This paper introduced I^{2}EM, a decouple-to-reconcile framework designed to address the entanglement of cross-exposure displacement and intra-exposure blur in dual-exposure night-image restoration. By factorizing motion into bidirectional optical flow and exposure-specific motion-conditioned kernels, I^{2}EM establishes physically interpretable representations for inter- and intra-exposure motion. We further couple these representations through observation reconstruction and decoupling constraints,

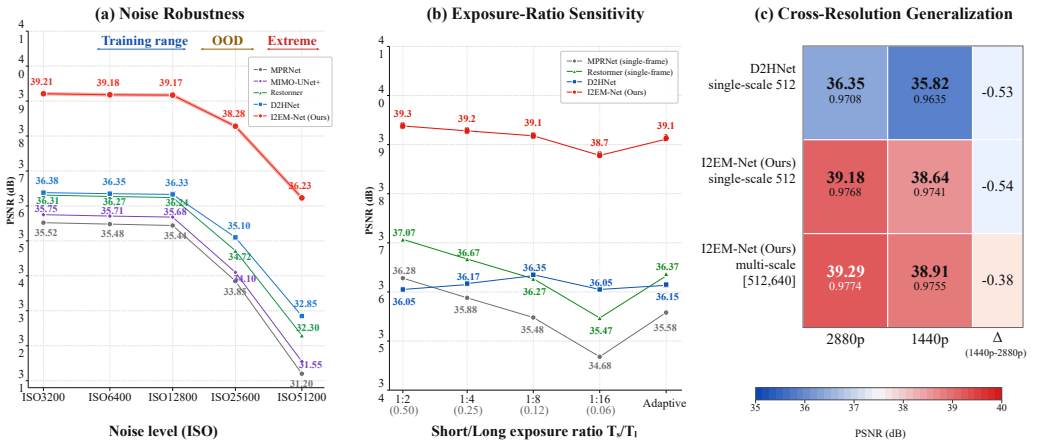

Figure 4: Robustness to acquisition and scale shifts on self-collected Xiaomi Mi Note 10 sequences acquired following the public D2HNet capture procedure. The panels report PSNR across three settings: (a) calibrated, out-of-distribution, and extreme ISO noise, (b) fixed and adaptive exposure ratios, and (c) 2880p-to-1440p transfer under single- and multi-scale training. Slower noise degradation and a smaller multi-scale resolution drop separate exposure complementarity from scale-diverse training.

Table 2: Core ablations on the generated D2 2880p validation set. Each row changes one stated factor relative to the full model. ΔPSNR is measured in dB relative to the full model.

Controlled design change		Complexity	D2 2880p restoration quality			
Mechanism	Intervention	Params (M)	PSNR↑	ΔPSNR	SSIM↑	LPIPS↓
Factorization	One-way flow	133.79	40.20	−0.58	0.9756	0.062
	Sequential flow–kernel	133.79	40.36	−0.42	0.9763	0.059
	Implicit joint motion	68.42	39.29	−1.49	0.9689	0.077
Decoupling constraints	No kernel centering	133.79	40.41	−0.37	0.9769	0.058
	Short-to-long blur loss	133.79	39.54	−1.24	0.9712	0.070
	Blur before warp	133.79	39.99	−0.79	0.9743	0.065
Reconciliation	Long-exposure residual	133.79	39.00	−1.78	0.9646	0.084
	No flow-conditioned kernels	133.79	40.32	−0.46	0.9761	0.060
	No residual denoising	121.64	40.09	−0.69	0.9750	0.066
	Concatenation instead of gating	134.51	40.47	−0.31	0.9772	0.057
Complete design	Full model	133.79	**40.78**	–	**0.9783**	**0.054**

Loss weighting controls how strongly the latent motion variables are tied to image formation. Table 3 separates adaptive optimization from a common scale applied to the physical constraints. Uncertainty weighting improves PSNR by 0.08 dB by adjusting term magnitudes during training. Under fixed weights, the selected $1.0\times$ scale gives the highest PSNR. Reducing it to $0.5\times$ weakens the coupling between correspondence, kernels, and reconstructed observations and costs 0.28 dB. Increasing it to $2.0\times$ costs 0.12 dB, while $5.0\times$ costs 0.69 dB because latent regularization increasingly restricts the image-reconstruction solution. The selected coefficients therefore balance data fidelity with identifiable roles for displacement and exposure integration.

Group	Setting	PSNR↑	ΔPSNR
Strategy	Fixed weights	40.78	0.00
	Uncertainty weighting	**40.86**	+0.08
Physical scale	$0.5\times$	40.50	−0.28
	$1.0\times$	**40.78**	0.00
	$2.0\times$	40.66	−0.12
	$5.0\times$	40.09	−0.69

Table 3: Objective weighting and physical-constraint sensitivity on D2 2880p. Values are PSNR (dB). ΔPSNR is relative to fixed weights.

and leverage them to guide confidence-aware feature restoration, alignment, and complementary information fusion. On the D2 2880p validation set, I^{2}EM achieves 40.78 dB PSNR, outperforming the input-matched D2HNet by 2.83 dB while reducing LPIPS by 28.9%. These results validate that explicitly separating where scene content moves across exposures from how it accumulates within an exposure enables more reliable coordination of complementary short- and long-exposure information, offering a principled foundation for dual-exposure image restoration.

AI USE STATEMENT

Generative AI tools were used to improve the clarity, grammar, and presentation of the manuscript. The authors reviewed all AI-assisted edits and verified the technical descriptions, citations, equations, and reported results. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This study concerns image restoration and involves neither human-subject experiments nor biometric inference. Restoration may nevertheless alter visible evidence or introduce plausible details that are absent from the captured signal. Outputs should therefore be treated as reconstructed estimates and should not be used for forensic, legal, or safety-critical factual determination without access to the original observations and independent validation.

REPRODUCIBILITY STATEMENT

The task definition, model interfaces, central image-formation constraints, evaluation protocol, and principal controlled comparisons are specified explicitly in the paper.

REFERENCES

Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11036–11045, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Brooks_Unprocessing_Images_for_Learned_Raw_Denoising_CVPR_2019_paper.html.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3291–3300, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Chen_Learning_to_See_CVPR_2018_paper.html.

Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. HINet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 182–192, 2021. URL https://openaccess.thecvf.com/content/CVPR2021W/NTIRE/html/Chen_HINet_Half_Instance_Normalization_Network_for_Image_Restoration_CVPRW_2021_paper.html.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Proceedings of the European Conference on Computer Vision*, pp. 17–33, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/3043_ECCV_2022_paper.php.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22367–22377, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Chen_Activating_More_Pixels_in_Image_Super-Resolution_Transformer_CVPR_2023_paper.html.

Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4641–4650, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Cho_Rethinking_Coarse-To-Fine_Approach_in_Single_Image_Deblurring_ICCV_2021_paper.html.

Zhenxuan Fang, Fangfang Wu, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Self-supervised non-uniform kernel estimation with flow-based motion prior for blind image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.

18105–18114, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Fang_Self-Supervised_Non-Uniform_Kernel_Estimation_With_Flow-Based_Motion_Prior_for_Blind_CVPR_2023_paper.html.

Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics*, 35(6):192:1–192:12, 2016. doi: 10.1145/2980179.2980254.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Rolf Köhler, Michael Hirsch, Betty J. Mohler, Bernhard Schölkopf, and Stefan Harmeling. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *European Conference on Computer Vision*, pp. 27–40, 2012. doi: 10.1007/978-3-642-33786-4_3. URL https://is.mpg.de/publications/kohlerhmshc2012.

Lingshun Kong, Jiangxin Dong, Jinhui Tang, Ming-Hsuan Yang, and Jinshan Pan. Efficient visual state space model for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12710–12719, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Kong_Efficient_Visual_State_Space_Model_for_Image_Deblurring_CVPR_2025_paper.html.

Lingshun Kong, Jiawei Zhang, Dongqing Zou, Fu Lee Wang, Jimmy S. Ren, Xiaohe Wu, Jiangxin Dong, and Jinshan Pan. DeblurDiff: Real-word image deblurring with generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 171890–171908, 2025b. doi: 10.52202/085713-5179. URL https://papers.nips.cc/paper_files/paper/2025/hash/e393677793767624f2821cec8bdd02f1-Abstract-Conference.html.

Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8878–8887, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Kupyn_DeblurGAN-v2_Deblurring_Orders-of-Magnitude_Faster_and_Better_ICCV_2019_paper.html.

Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *European Conference on Computer Vision*, pp. 736–753, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/4261_ECCV_2022_paper.php.

Jinlong Li, Baolu Li, Zhengzhong Tu, Xinyu Liu, Qing Guo, Felix Juefei-Xu, Runsheng Xu, and Hongkai Yu. Light the night: A multi-condition diffusion framework for unpaired low-light enhancement in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15205–15215, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_Light_the_Night_A_Multi-Condition_Diffusion_Framework_for_Unpaired_Low-Light_CVPR_2024_paper.html.

Zhi Li, Zhenhong Jia, Jie Yang, and Nikola Kasabov. Low illumination video image enhancement. *IEEE Photonics Journal*, 12(4):1–13, 2020. doi: 10.1109/JPHOT.2020.3010966.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1833–1844, 2021. URL https://openaccess.thecvf.com/content/ICCV2021W/AIM/html/Liang_SwinIR_Image_Restoration_Using_Swin_Transformer_ICCVW_2021_paper.html.

Panjun Liu, Jiyuan Xia, Yuanshen Guan, Yong Li, Zhiqiang Lang, Ruikang Xu, Chang Chen, Dehua Song, Fenglong Song, and Zhiwei Xiong. RawMetaDiff: Unlocking extreme darkness from dual-exposure RAW with meta-guided diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5617–5626,

2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_RawMetaDiff_Unlocking_Extreme_Darkness_from_Dual-Exposure_RAW_with_Meta-Guided_Diffusion_CVPR_2026_paper.html.

Yapeng Meng, Lin Yang, Yuguo Chen, Xiangru Chen, Taoyi Wang, Lijian Wang, Zheyu Yang, Yihan Lin, and Rong Zhao. Spatio-temporal difference guided motion deblurring with the complementary vision sensor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37496–37506, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Meng_Spatio-Temporal_Difference_Guided_Motion_Deblurring_with_the_Complementary_Vision_Sensor_CVPR_2026_paper.html.

Ben Mildenhall, Jonathan T. Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2502–2510, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Mildenhall_Burst_Denoising_With_CVPR_2018_paper.html.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3883–3891, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Nah_Deep_Multi-Scale_Convolutional_CVPR_2017_paper.html.

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1996–2005, 2019. doi: 10.1109/CVPRW.2019.00251. URL https://openaccess.thecvf.com/content_CVPRW_2019/html/NTIRE/Nah_NTIRE_2019_Challenge_on_Video_Deblurring_and_Super-Resolution_Dataset_and_CVPRW_2019_paper.html.

Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3043–3051, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Pan_Cascaded_Deep_Video_Deblurring_Using_Temporal_Sharpness_Prior_CVPR_2020_paper.html.

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, 2023. doi: 10.52202/075280-3121. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e187897ed7780a579a0d76fd4a35d107-Abstract-Conference.html.

Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pp. 184–201, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/4924_ECCV_2020_paper.php.

Jaesung Rim, Junyong Lee, Heemin Yang, and Sunghyun Cho. Deep hybrid camera deblurring for smartphone cameras. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024. doi: 10.1145/3641519.3657507.

Shayan Shekarforoush, Amanpreet Walia, Marcus A. Brubaker, Konstantinos G. Derpanis, and Alex Levinshtein. Dual-camera joint deblurring-denoising. *arXiv preprint arXiv:2309.08826*, 2023. URL https://arxiv.org/abs/2309.08826.

Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5572–5581, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Shen_Human-Aware_Motion_Deblurring_ICCV_2019_paper.html.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Sun_PWC-Net_CNNs_for_CVPR_2018_paper.html.

Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8174–8182, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Tao_Scale-Recurrent_Network_for_CVPR_2018_paper.html.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419, 2020. doi: 10.1007/978-3-030-58536-5_24. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/3526_ECCV_2020_paper.php.

Jikun Wang, Weixiang Liang, Jiangang Yang, Shizheng Wang, and Zhi-Xin Yang. An adaptive image enhancement approach for safety monitoring robot under insufficient illumination condition. *Computers in Industry*, 147:103862, 2023. doi: 10.1016/j.compind.2023.103862.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17683–17693, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Uformer_A_General_U-Shaped_Transformer_for_Image_Restoration_CVPR_2022_paper.html.

Yuan Yuan, Wei Su, and Dandan Ma. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3555–3564, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yuan_Efficient_Dynamic_Scene_Deblurring_Using_Spatially_Variant_Deconvolution_Network_With_CVPR_2020_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zamir_Multi-Stage_Progressive_Image_Restoration_CVPR_2021_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zamir_Restormer_Efficient_Transformer_for_High-Resolution_Image_Restoration_CVPR_2022_paper.html.

Zhilu Zhang, Shuohao Zhang, Renlong Wu, Zifei Yan, and Wangmeng Zuo. Exposure bracketing is all you need for a high-quality image. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rDIf6NA5mj.

Yuzhi Zhao, Yongzhe Xu, Qiong Yan, Dingdong Yang, Xuehui Wang, and Lai-Man Po. D2HNet: Joint denoising and deblurring with hierarchical network for robust night image restoration. In *European Conference on Computer Vision*, pp. 91–110, 2022. doi: 10.1007/978-3-031-20071-7_6. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136670089.pdf.

Yuzhi Zhao, Lai-Man Po, Xin Ye, Yongzhe Xu, and Qiong Yan. Modeling dual-exposure quad-bayer patterns for joint denoising and deblurring. *IEEE Transactions on Image Processing*, 34:350–364, 2025. doi: 10.1109/TIP.2024.3515873.

Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *Proceedings of the European Conference on Computer Vision*, pp. 191–207, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/5116_ECCV_2020_paper.php.

Supplementary Material

A Image Formation and Notation

Let $X(\mathbf{p}, t)$ denote scene radiance projected to pixel $\mathbf{p}$ at time t. For exposure $e \in \{s, l\}$ with center time t_e and duration T_e, the short- and long-exposure observations are represented as exposure-time averages with additive noise:

$$I_s(\mathbf{p}) = \frac{1}{T_s} \int_{t_s - T_s/2}^{t_s + T_s/2} X(\mathbf{p}, t)\, dt + n_s(\mathbf{p}). \tag{9}$$

$$I_l(\mathbf{p}) = \frac{1}{T_l} \int_{t_l - T_l/2}^{t_l + T_l/2} X(\mathbf{p}, t)\, dt + n_l(\mathbf{p}). \tag{10}$$

Here $T_s \ll T_l$, the short exposure precedes the long exposure, and $\Delta t = t_l - t_s > 0$ is the interval between their center times. The terms n_s and n_l collect signal-dependent and read noise. This ordering matches the acquisition sequence in Fig. 1 and supplies the temporal structure used by the network, while camera-dependent nonlinearities are handled by the learned restoration process.

Given (I_s, I_l), the network predicts $\widehat{X}$ in the short-exposure reference coordinates together with five-scale sampling fields $\{F^r_{s\to l}, F^r_{l\to s}\}_{r=0}^4$ and exposure-specific kernel fields $\{K^r_s, K^r_l\}_{r=0}^4$. Scale $r = 0$ denotes the input grid and $r = 4$ the coarsest grid. The bilinear sampling operator is $\mathcal{W}(I, F)[\mathbf{p}] = I[\mathbf{p} + F(\mathbf{p})]$. Hence, $F_{s\to l}$ transfers long-exposure evidence to the short grid and $F_{l\to s}$ transfers the short-reference estimate to the long grid. The operator $\mathcal{B}(I, K)$ applies a normalized spatially varying kernel at each pixel.

B Architecture Details

Figure 5 summarizes the exposure model and its local-kernel approximation. Table 4 maps the five functional blocks in Figure 2 to their implemented feature scales, prediction heads, and reconstruction interfaces. The surrounding text specifies the complete feature, motion, reconciliation, decoding, and training paths.

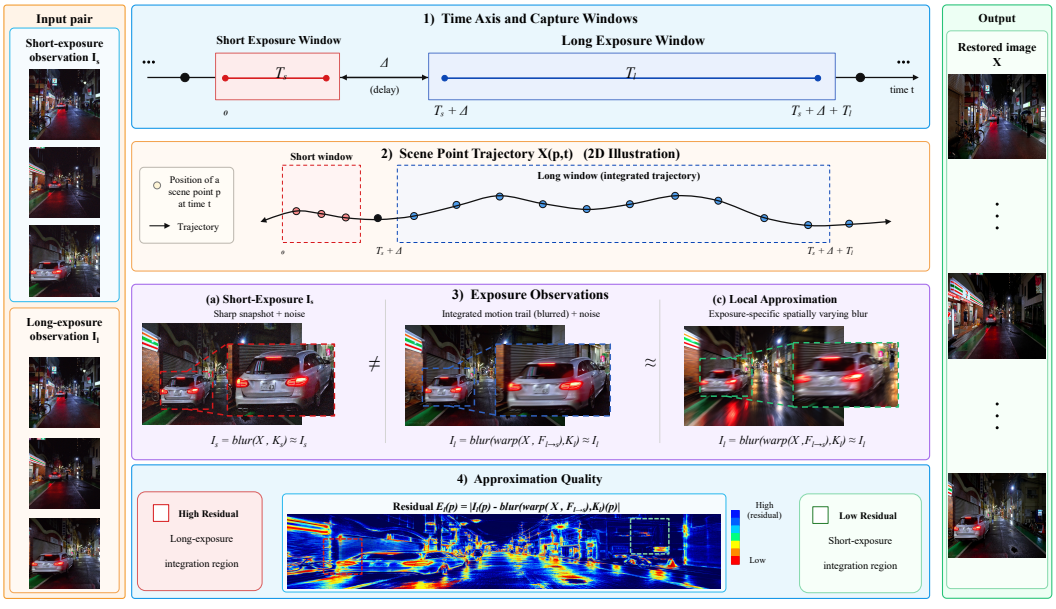

Figure 5: Exposure formation and the local-kernel approximation of long-exposure integration.

Table 4: Architecture specification of the default I^{2}EM configuration. S and L denote the short- and long-exposure branches. The scale ranges are ordered from the input grid to the coarsest feature level.

Block	Scale/channels	Role
Exposure-specific encoding	1–1/16, 32–512	Independent S/L pyramids
Bidirectional motion estimator	five scales, $d = 6$	Flow refinement and C_f
Motion-conditioned kernel estimator	five scales, $[15, 9, 5, 3, 3]$	K_s, K_l, and C_k
Confidence-aware reconciliation	five scales	Restore with K_l, sample, and gate
Fusion decoder	$1/16 \to 1$	Reconciled features and short skips
Image-space denoise head	five levels, 11 channels	Guided residual refinement
Deployment	overlap tiles	50% overlap, Gaussian blending

The motion cost volume contains $(2d+1)^2 = 169$ displacement channels for $d = 6$, compared with 81 channels for the narrower $d = 4$ setting. Kernel prediction uses a finest-scale support of $k = 15$, and the application, transpose, centroid, confidence, and learned-restoration paths derive their tensor shapes from the resulting sequence. The forward field transports restored long-exposure evidence into the short-reference coordinates. The reverse field closes the correspondence cycle and provides reciprocal motion context to exposure-kernel prediction. These directional roles jointly constrain geometric transport and exposure-time integration.

C Additional Formulation and Losses

The reconciliation module forms reliability-weighted alignment as

$$A^r = \rho^r \odot \mathcal{W}(\widetilde{L}^r, F^r_{s\to l}) + (1 - \rho^r) \odot S^r. \tag{11}$$

The decoder relations are

$$I_0 = I_s + R_0, \tag{12}$$

and

$$\widehat{X} = I_0 + D(X_D). \tag{13}$$

The valid confidence used for long-to-short alignment is

$$\rho^r = V^r \odot \frac{C^r_f}{C^r_f + \tau_c}. \tag{14}$$

The fusion gate is predicted from the short-reference and aligned features

$$G^r = \sigma(H^r(\mathrm{cat}[S^r, A^r])). \tag{15}$$

For a kernel with support size k_r, the normalized long-kernel blur strength used by the reconciliation and denoising stages is

$$b^r_l(\mathbf{p}) = \tanh\left(\frac{2\, m_2(K^r_l)(\mathbf{p})}{(k_r/2)^2 + \epsilon}\right), \tag{16}$$

where m_2 is the second moment about the kernel centroid. The image-space denoising head receives

$$X_D = [I_0, I_s, \widetilde{I}_l, \widetilde{C}_f, \widetilde{b}_l] \in \mathbb{R}^{11\times H\times W}. \tag{17}$$

Here, $\widetilde{I}_l$ is the valid-region-masked long exposure aligned to the short grid, $\widetilde{C}_f$ is the corresponding masked flow confidence, and $\widetilde{b}_l$ is the transported blur-strength map.

The image reconstruction terms are the pixel loss

$$\mathcal{L}_1 = \|\widehat{X} - X^*\|_1, \tag{18}$$

and the structural loss

$$\mathcal{L}_{\mathrm{SSIM}} = 1 - \mathrm{SSIM}(\widehat{X}, X^*). \tag{19}$$

The finest bidirectional fields are regularized in their valid region by

$$\mathcal{L}_{\text{flow-cyc}} = \left\|\mathcal{W}(F^0_{s\to l}, F^0_{l\to s}) + F^0_{l\to s}\right\|_{1,V}. \tag{20}$$

The edge-aware first-order operator shown in Fig. 6 is

$$\mathrm{Smooth}(F, I) = \left\langle |\partial_x F| \odot e^{-|\partial_x I|} + |\partial_y F| \odot e^{-|\partial_y I|} \right\rangle_V , \tag{21}$$

where $\langle\cdot\rangle_V$ averages valid samples and broadcasts the image-gradient weights over the flow channels. Applied to both directions at all five scales, it gives

$$\mathcal{L}_{\mathrm{smooth}} = \frac{1}{5} \sum_{r=0}^{4} \left[\mathrm{Smooth}(F^r_{s\to l}, I^r_s) + \mathrm{Smooth}(F^r_{l\to s}, I^r_l) \right] . \tag{22}$$

The centered-kernel component is

$$\mathcal{L}_{\mathrm{ctr}} = \|\boldsymbol{\mu}(K^0_s)\|_1 + 2\|\boldsymbol{\mu}(K^0_l)\|_1. \tag{23}$$

Here $\boldsymbol{\mu}(K^0_e) \equiv \boldsymbol{\mu}^0_e$ is the centroid defined in Equation 3. The coefficient two places stronger centering regularization on the long-exposure kernel. To match the reconstructed observations shown in Fig. 6, define

$$\widehat{I}_s = \mathcal{B}(\widehat{X}, K^0_s), \tag{24}$$

$$\widehat{I}_l = \mathcal{B}\Big(\mathcal{W}(\widehat{X}, F^0_{l\to s}), K^0_l\Big) . \tag{25}$$

The corresponding observation losses are

$$\mathcal{L}^{\mathrm{obs}}_s = \rho(\widehat{I}_s, I_s), \tag{26}$$

$$\mathcal{L}^{\mathrm{obs}}_l = \rho(\widehat{I}_l, I_l). \tag{27}$$

These expressions are equivalent to Equations 6 and 7 and are evaluated on valid kernel interiors and in-frame flow samples. The same construction is applied at the four coarser kernel levels, whose mean valid losses are multiplied by 0.5 and added to the corresponding finest-scale terms. Combining all seven losses yields the weighted objective in Equation 8. Its default coefficients are listed in the training configuration. Optimization retains the full magnitude of each loss term and clips the aggregate parameter-gradient norm to one.

Figure 6 summarizes the complete supervision path from final-image fidelity to correspondence, observation, and kernel constraints.

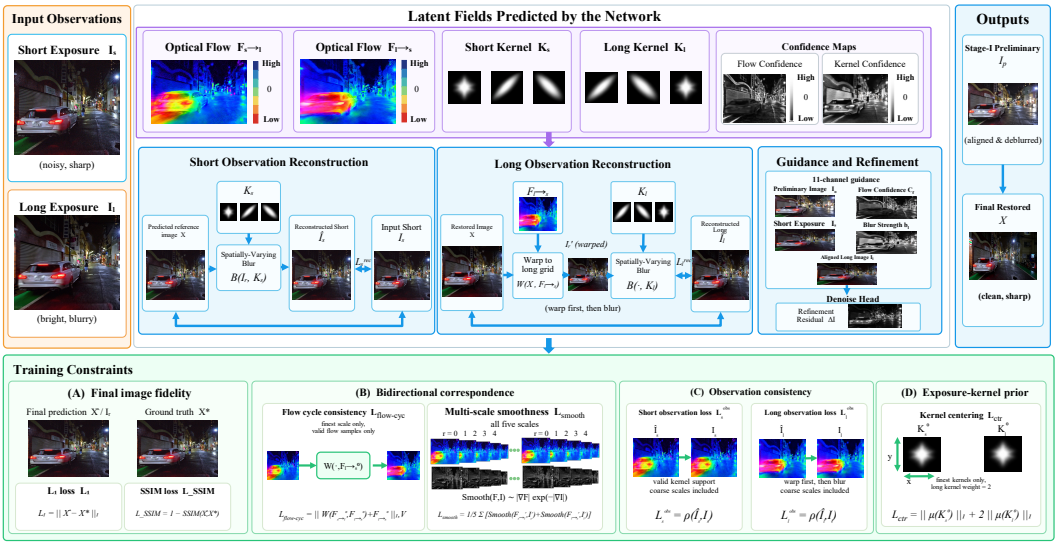

Figure 6: Complete training constraints and bidirectional observation modeling in I^{2}EM. From the short and long observations (I_s, I_l), the network predicts bidirectional fields $(F_{s\to l}, F_{l\to s})$, exposure-specific kernels (K_s, K_l), flow confidence C_f, and kernel confidence C_k. The blur strength b_l is computed from K_l. The short observation is reconstructed by blurring $\widehat{X}$ with K_s, whereas the long observation is reconstructed by first warping $\widehat{X}$ with $F_{l\to s}$ and then blurring it with K_l. The preliminary image, I_s, aligned I_l, C_f, and transported b_l form the 11-channel refinement guidance. Panels (A)–(D) group the full objective into final-image fidelity $(\mathcal{L}_1, \mathcal{L}_{\text{SSIM}})$, bidirectional correspondence $(\mathcal{L}_{\text{flow-cyc}}, \mathcal{L}_{\text{smooth}})$, observation consistency $(\mathcal{L}_s^{\text{obs}}, \mathcal{L}_l^{\text{obs}})$, and the centered exposure-kernel prior $\mathcal{L}_{\text{ctr}}$.

D Implementation Details

Dataset and splits. The public D2 release contains 5,661 training tuples (Zhao et al., 2022). Following the official data-generation rules produces 1,192 non-overlapping validation scenes at 5120×2880. Each training tuple provides one short exposure, four long-exposure durations, the corresponding reference image, and selected blurred patches accompanying the full-resolution crop.

Training configuration. Each sampled tuple pairs one randomly selected long-exposure duration with the corresponding short exposure and reference. Optimization uses Adam (Kingma & Ba, 2015) for 100 epochs with batch size one, $(\beta_1, \beta_2) = (0.9, 0.999)$, no weight decay, and a base learning rate of 10^{-4}. Five epochs of linear warm-up precede the main schedule, the learning rate is halved at epoch 55, and the gradient norm is clipped at one. The checkpoint with the highest validation PSNR is retained. Restoration values remain unclipped during optimization and are mapped to $[0, 1]$ for inference.

Loss-weight configuration. The default coefficients used in all reported comparisons are $(\lambda_{\text{pix}}, \lambda_{\text{str}}, \lambda_{\text{cyc}}, \lambda_{\text{sm}}, \lambda_s^{\text{obs}}, \lambda_l^{\text{obs}}, \lambda_{\text{ctr}}) = (120, 15, 1, 0.05, 5, 1, 0.5)$. The four coarse-scale observation terms retain the 0.5 multiplier described in Section 3.3. The coefficient scans vary one group while keeping the remaining terms at these defaults.

Noise synthesis. The active shot-noise path approximately inverts gamma and white balance, mosaics to a Bayer pattern, applies Poisson shot noise and Tukey-lambda read noise with Xiaomi Mi Note 10 calibration, demosaics, and applies a forward image-signal-processing (ISP) pipeline. The short- and long-exposure sensor sensitivities are sampled from International Organization for Standardization (ISO) sensitivity ranges $[6400, 12800]$ and $[1000, 4000]$, respectively. This procedure follows the motivation of physically informed unprocessing (Brooks et al., 2019) and begins from the available standard red–green–blue (sRGB) data.

Data augmentation. Training uses random 1024^2 crops followed by processing at 512^2. CutNoise replaces a randomly selected 120×120 region of the short-exposure input with the co-located reference region with probability 0.3. Illumination adjustment is applied with probability 0.3, and short-image color adjustment is applied with probability 0.5. The principal augmentation pipeline contains these photometric operations without geometric flips or rotations.

Full-resolution inference. For a native 4K-class pair, the inference path enumerates fully covering 1024^2 tiles at stride 512, including explicit terminal positions along the right and bottom borders. The automatic resolution rule

$$p = \mathrm{clip}\left(\left\lfloor \frac{512 \times 1920}{\max(H, W)} \right\rfloor, 256, 512 \right) \tag{28}$$

selects its lower processing bound for this input class. At the lower bound $p = 256$, the nominal ± 96-pixel coarse support at network scale corresponds to approximately ± 384 pixels within a 1024-pixel native tile before residual refinement. This mapping characterizes the sensor-coordinate coverage of the enlarged cost volume. Each prediction is resized to its native tile extent and merged with a Gaussian weight map whose minimum is clipped to 10^{-3}, preventing holes at the canvas boundary. The resulting image retains the original sensor dimensions, although each tile is restored at the lower processing resolution and resized back before blending. All network-only results use the base restoration output. The optional short-image-guided detail enhancement uses a strength of 0.3 and forms a separate output setting. Figure 7 summarizes this deployment path and its native-coordinate search interpretation.

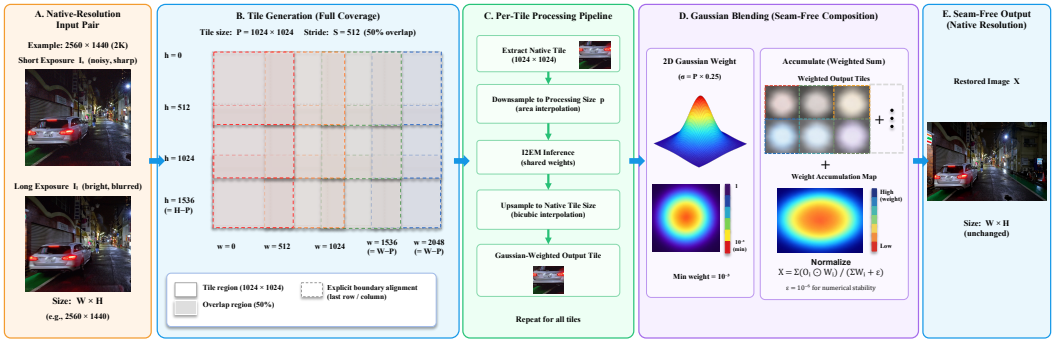

Figure 7: Native-resolution tiled inference. (A) A noisy short exposure and a brighter blurred long exposure enter at their original $W \times H$ resolution. (B) Overlapping 1024×1024 tiles are enumerated with stride 512, with the final row and column aligned to the image boundaries for complete coverage. (C) Each native tile is resized to processing size p, restored by the shared-weight I^{2}EM network, resized back, and multiplied by a Gaussian weight map. (D) Weighted predictions are accumulated and normalized over their overlaps. (E) The merged output retains the original $W \times H$ dimensions and suppresses tile-boundary discontinuities.

Table 5: Evaluation datasets and their roles. D2 provides native dual-exposure pairs.

Dataset	Split/scale	Pairing	Role
D2 (Zhao et al., 2022)	5,661 train, 1,192 generated val	native S/L	quantitative benchmark
GoPro (Nah et al., 2017)	standard train/test split	single blurred frame	cross-dataset evaluation
HIDE (Shen et al., 2019)	official test split	single blurred frame	cross-dataset evaluation
RealBlur (Rim et al., 2020)	RealBlur-J/R	single blurred frame	real-blur evaluation
Köhler (Köhler et al., 2012)	official benchmark	single blurred frame	cross-dataset evaluation
REDS (Nah et al., 2019)	official training split	video frames	cross-dataset training

Evaluation protocol. Bilinear sampling uses the unaligned-corner convention. Integer pixel centers are normalized as $2(x + 0.5)/W - 1$ and $2(y + 0.5)/H - 1$, so zero flow preserves the identity grid under `align_corners=False`. All reported configurations use this convention.

The dataset roles in Table 5 separate task-matched evaluation from modality-boundary tests. D2 supplies the complementary observations required by the method, whereas the single-image datasets test architectural transfer after both branches receive the same blurred frame.

E Additional Experiments

The supplementary evaluation examines acquisition robustness, repeated-run stability, transfer under missing exposure complementarity, computational cost, and additional design controls. The noise, cross-resolution, exposure-ratio, and repeated-run results are evaluated on the self-collected MI_NOTE10 sequences. Sensor noise is swept from ISO 3,200 to 51,200. Exposure ratios cover 1:2, 1:4, 1:8, 1:16, and adaptive sampling. Resolution is evaluated at 2880p and 1440p under single- and multi-scale training. Repeated-run statistics use three independent runs. Each table reports its evaluation dimensions and aggregation in the caption or surrounding text.

E.1 Robustness tables

Tables 6, 7, and 8 provide the exact values underlying the main-paper trends in Fig. 4 without repeating the same visual encoding.

Table 6: Robustness to calibrated sensor-noise variation. The final column measures the PSNR change from the upper training boundary at ISO 12,800 to the extreme ISO 51,200 setting. Bold values identify the best result at each ISO.

	Training range and lower			Extrapolation		
Method	3,200	6,400	12,800	25,600	51,200	$\Delta_{12.8\text{k}\to51.2\text{k}}$
MPRNet (Zamir et al., 2021)	35.52	35.48	35.44	33.85	31.20	-4.24
MIMO-UNet+ (Cho et al., 2021)	35.75	35.71	35.68	34.10	31.55	-4.13
Restormer (Zamir et al., 2022)	36.31	36.27	36.24	34.72	32.30	-3.94
D2HNet (Zhao et al., 2022)	36.38	36.35	36.33	35.10	32.85	-3.48
I^{2}EM	**39.21**	**39.18**	**39.17**	**38.28**	**36.23**	-2.94

The in-range results change by at most 0.04 dB for each method, whereas the extrapolation regime separates their degradation rates. From ISO 12,800 to 51,200, I^{2}EM loses 2.94 dB compared with 3.48 dB for D2HNet and more than 3.9 dB for the single-image baselines. This smaller decline is consistent with complementary long-exposure evidence reducing sensitivity to short-exposure noise escalation. The remaining 2.94 dB loss also defines a clear boundary. Exposure complementarity mitigates but does not remove severe sensor-noise shift.

Table 7: Cross-resolution stability on self-collected Xiaomi Mi Note 10 sequences under single-scale and multi-scale training. Change is the PSNR difference from 2880p to 1440p, and the final column is the 1440p PSNR difference from D2HNet. All differences are in dB.

Method	Training	2880p		1440p		Change	vs. D2HNet
		PSNR	SSIM	PSNR	SSIM		
D2HNet (Zhao et al., 2022)	single-scale 512	36.35	0.9708	35.82	0.9635	-0.53	–
I^{2}EM	single-scale 512	39.18	0.9768	38.64	0.9741	-0.54	$+2.82$
I^{2}EM	multi-scale $[512, 640]$	**39.29**	**0.9774**	**38.91**	**0.9755**	-0.38	$+3.09$

Single-scale I^{2}EM and D2HNet exhibit nearly identical resolution changes of 0.54 and 0.53 dB, so the base architecture retains comparable scale sensitivity. Multi-scale training improves both resolutions and reduces the change to 0.38 dB, attributing the benefit to exposure-scale diversity during training. The main comparison uses the single-scale model to isolate architectural effects from this training intervention.

Table 8: Sensitivity to the short-to-long exposure ratio. The adaptive setting samples an available long-exposure level. The final row gives the paired PSNR difference between I^{2}EM and D2HNet, in dB, computed from the preceding rows.

Method	1:2	1:4	1:8	1:16	Adaptive	Mean
MPRNet (Zamir et al., 2021)	36.28	35.88	35.48	34.68	35.58	35.58
Restormer (Zamir et al., 2022)	37.07	36.67	36.27	35.47	36.37	36.37
D2HNet (Zhao et al., 2022)	36.05	36.17	36.35	36.05	36.15	36.15
I^{2}EM	**39.38**	**39.28**	**39.18**	**38.78**	**39.13**	**39.15**
ΔPSNR	$+3.33$	$+3.11$	$+2.83$	$+2.73$	$+2.98$	$+3.00$

The single-image MPRNet and Restormer baselines each lose 1.60 dB from 1:2 to 1:16, showing that the exposure ratio materially changes the degradation difficulty for a single input. D2HNet peaks at 36.35 dB for 1:8, whereas I^{2}EM decreases gradually from 39.38 dB at 1:2 to 38.78 dB at 1:16 but retains the highest PSNR under every fixed and adaptive setting. Its margin over D2HNet ranges from 2.73 to 3.33 dB and averages 3.00 dB. The adaptive setting reaches 39.13 dB, 0.25 dB below the best fixed ratio, while exceeding D2HNet by 2.98 dB under the same policy.

E.2 Statistical analysis

Table 9 separates optimization variability from scene-level evidence, while Table 10 tests whether the aggregate margin is concentrated in one illumination group. Figure 8 visualizes the corresponding run-level, paired-effect, and illumination-stratified views. The mean and standard deviation summarize three independent runs. The Wilcoxon signed-rank test uses the paired PSNR differences of 1,192 validation scenes as its experimental units.

Table 9: Repeated-run stability and paired scene-level comparison. ΔPSNR is the mean PSNR of I^{2}EM minus that of each baseline, in dB. The Wilcoxon test uses 1,192 paired scenes as experimental units.

Method	Run 1	Run 2	Run 3	Mean $\pm$ std.	ΔPSNR	p-value
MIMO-UNet+ (Cho et al., 2021)	35.69	35.73	35.71	35.71 ± 0.02	$+3.47$	< 0.001
Restormer (Zamir et al., 2022)	36.25	36.29	36.27	36.27 ± 0.02	$+2.91$	< 0.001
D2HNet (Zhao et al., 2022)	36.33	36.37	36.35	36.35 ± 0.02	$+2.83$	< 0.001
I^{2}EM	39.16	39.19	39.20	$\mathbf{39.18 \pm 0.02}$	–	–

Table 10: Illumination-stratified D2 results. ΔPSNR is computed as I^{2}EM minus D2HNet within each scene group, in dB.

	Day			Night		
Method	PSNR↑	SSIM↑	ΔPSNR	PSNR↑	SSIM↑	ΔPSNR
D2HNet (Zhao et al., 2022)	37.12	0.9735	–	35.58	0.9662	–
I^{2}EM	**39.98**	**0.9805**	$+2.86$	**38.38**	**0.9731**	$+2.80$

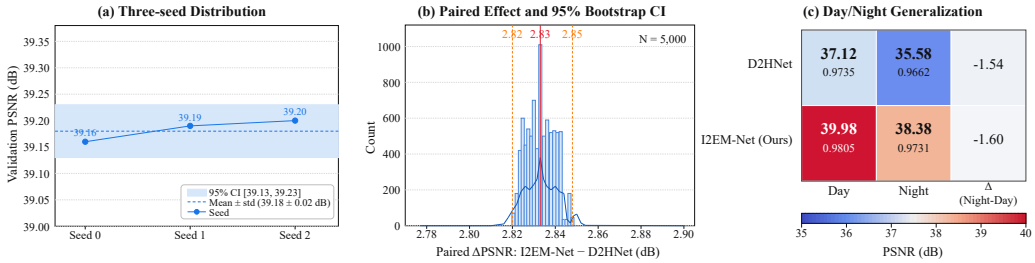

Figure 8: Statistical consistency across runs, scenes, and illumination. Panel (a) shows validation PSNR for three independent seeds with the mean, standard deviation, and 95% confidence interval. Panel (b) shows the 5,000-sample bootstrap distribution of the paired per-scene PSNR gain over D2HNet. Panel (c) compares PSNR and SSIM between day and night scenes. Stable runs, concentrated positive paired gains, and nearly matched day/night margins connect the aggregate improvement to both optimization and scene-level behavior.

The run-to-run standard deviation is 0.02 dB for every listed method, far below the 2.83 dB mean gap between I^{2}EM and D2HNet in this evaluation series. The day and night gains are also similar at 2.86 and 2.80 dB, while both methods decline by approximately 1.6 dB at night. The aggregate margin is therefore stable across optimization runs and illumination groups, although absolute restoration quality remains lower at night.

E.3 Cross-dataset generalization

GoPro, HIDE, RealBlur, Köhler, and REDS do not contain native short–long exposure pairs (Nah et al., 2017, Shen et al., 2019, Rim et al., 2020, Köhler et al., 2012, Nah et al., 2019). Dual-exposure models therefore receive the same blurred frame in both branches. This construction equalizes input information while removing exposure complementarity. Table 11 reports cross-dataset transfer, and Table 12 evaluates same-distribution training on RealBlur. Together, they characterize the architecture under modality mismatch.

Table 11: Cross-dataset PSNR under equivalent single-image input. Each row is a separate off-domain train–test protocol. The in-domain D2 evaluation is reported separately in Table 1. ΔPSNR is computed as I^{2}EM minus D2HNet under the same protocol, in dB. Best results are bold.

Train	Test	Deep	MIMO+	HINet	D2HNet	EVSSM	I^{2}EM	ΔPSNR
GoPro	GoPro	29.08	32.45	32.71	30.85	**33.89**	31.65	+0.80
GoPro	HIDE	25.73	29.98	30.50	29.20	**32.73**	29.85	+0.65
GoPro	RealBlur-J	27.87	28.15	28.30	27.70	**31.90**	28.10	+0.40
GoPro	RealBlur-R	33.90	35.30	35.50	34.80	**39.65**	35.20	+0.40
GoPro	Köhler	26.10	27.90	28.20	27.30	**29.00**	27.75	+0.45
REDS	GoPro	26.50	28.80	29.10	28.10	**30.00**	28.60	+0.50

Replicating one blurred frame across both branches removes exposure complementarity while retaining the dual-branch architecture. Under all six off-domain protocols, I^{2}EM exceeds D2HNet by 0.40–0.80 dB, indicating that reconciliation remains useful when the two inputs coincide. However, it consistently trails MIMO-UNet+, HINet, and EVSSM, so the factorized dual-exposure model is not a universal single-image deblurring leader.

Table 13: Accuracy and efficiency comparison under the unified evaluation platform. PSNR, parameter count, and runtime follow the D2 2880p protocol in Table 1. Runtime is the full-image inference time in milliseconds (ms), and memory is the peak allocated GPU memory in MB.

Method	Input	PSNR (dB)↑	Params (M)	FLOPs (G)	Time (ms)	Mem. (MB)
DeepDeblur	single	36.55	42.3	156.8	14	1,245
DeblurGAN-v2	single	36.82	60.9	182.3	2	1,580
CDVD-TSP	multi	37.00	16.6	88.5	23	1,120
HINet	multi	37.22	88.7	112.3	6	1,980
MPRNet	single	37.08	23.5	98.2	10	1,450
MIMO-UNet+	single	37.31	16.1	42.1	9	980
Restormer	single	37.87	26.1	105.7	12	1,680
D2HNet	dual	37.95	35.2	165.3	12	1,520
UFPNet	single	37.72	80.3	290.4	29	2,350
EVSSM	single	38.09	17.1	126.0	3	1,100
DeblurDiff	single	37.55	386.2	2,890.0	304	12,580
STGDNet	multi-sensor	38.12	28.4	142.6	15	2,350
I^2EM	dual	**40.78**	133.79	462.8	78	10,284

Table 12: Same-distribution RealBlur evaluation under equivalent single-image input. ΔPSNR is computed as I^2EM minus D2HNet in dB, distinguishing this setting from the GoPro-to-RealBlur transfer rows in Table 11.

	RealBlur-J			RealBlur-R		
Method	PSNR↑	SSIM↑	ΔPSNR	PSNR↑	SSIM↑	ΔPSNR
DeepDeblur	28.95	0.8510	–	36.35	0.9301	–
MPRNet	31.20	0.9005	–	39.05	0.9662	–
MIMO-UNet+	31.35	0.9028	–	39.15	0.9670	–
HINet	31.50	0.9045	–	39.30	0.9682	–
Restormer	31.65	0.9060	–	39.42	0.9690	–
D2HNet	30.60	0.8952	–	38.45	0.9625	–
EVSSM	**32.50**	**0.9280**	–	**40.10**	**0.9740**	–
I^2EM	31.10	0.8990	+0.50	38.90	0.9655	+0.45

Same-distribution training raises every method shared with the GoPro-to-RealBlur transfer comparison. EVSSM, HINet, and MIMO-UNet+ remain ahead of I^2EM, while I^2EM improves over D2HNet by 0.50 dB on RealBlur-J and 0.45 dB on RealBlur-R. This pattern indicates a modest benefit from the dual-branch architecture under duplicated inputs. The substantially larger D2 margin is observed only when complementary exposures are available.

E.4 Computational efficiency

Table 13 compares restoration quality, model complexity, runtime, and memory under the unified hardware and timing protocol.

Under the Table 1 timing protocol, I^2EM reaches 40.78 dB with 78 ms runtime and 10,284 MB peak memory for one full-resolution image. Relative to the input-matched D2HNet, it gains 2.83 dB while using 6.50 times the runtime, 6.77 times the peak memory, 3.80 times the parameters, and 2.80 times the FLOPs. The additional computation supports the two exposure-specific encoders, bidirectional correspondence, and spatially varying kernel estimation that precede reconciliation. DeblurDiff remains slower and more memory intensive, whereas the other baselines operate faster but attain lower PSNR. The three-level control below reduces the parameter count by 44.44M with a 0.65 dB decrease, providing a lower-cost operating point when deployment efficiency is prioritized.

E.5 Additional ablation studies

The following controls extend the three claim groups in the main paper. Table 14 tests module substitutions, while Tables 15 and 16 examine kernel support and pyramid depth. Table 17 evaluates the cumulative augmentation pipeline. Tables 18 and 19 analyze correspondence support and objective construction. Table 20 gives the complete term-wise scans, and Figure 9 visualizes them. Figure 10

summarizes the remaining sensitivity trends. These controls distinguish representation, capacity, data processing, and optimization effects under the same full-model reference.

Table 14: Module substitutions grouped by the mechanism they test. ΔPSNR is measured in dB relative to the complete model. The groups distinguish correspondence construction, blur representation, and feature reconciliation.

Mechanism	Intervention	Params (M)	PSNR↑	ΔPSNR	SSIM↑	LPIPS↓
Correspondence	Without cost volume	127.53	40.26	−0.52	0.9759	0.061
	Global spatial kernel	133.79	39.87	−0.91	0.9729	0.069
Blur model	Kernels without flow conditioning	133.79	40.32	−0.46	0.9761	0.060
	Richardson–Lucy deblurring	118.26	39.66	−1.12	0.9719	0.072
Reconciliation	Without residual denoising	121.64	40.09	−0.69	0.9750	0.066
	Concatenation instead of gating	134.51	40.47	−0.31	0.9772	0.057
Complete design	Full model	133.79	**40.78**	–	**0.9783**	**0.054**

The largest degradations arise when Richardson–Lucy iterations or a global kernel replaces the spatially varying learned blur model. The former changes both the operator and its capacity. The global-kernel control retains the network size and more directly tests spatial variation. Its 0.91 dB loss supports location-dependent integration rather than a single image-wide blur. Removing the cost volume and flow conditioning causes smaller losses of 0.52 and 0.46 dB, showing that correspondence structure contributes to both flow estimation and kernel prediction. The reconciliation controls lose 0.69 and 0.31 dB. Because concatenation is slightly larger than the gated model, the latter difference favors reliability-aware selection over added capacity.

Table 15: Sensitivity to the finest-scale kernel support. Parameter and PSNR differences are measured relative to the default $k = 15$ configuration. ΔPSNR is reported in dB.

k	Kernel sequence	Params (M)	Param. diff. (M)	PSNR↑	ΔPSNR	SSIM↑	LPIPS↓
7	[7,5,3,3,3]	131.92	−1.87	40.30	−0.48	0.9760	0.061
15	[15,9,5,3,3]	133.79	0.00	40.78	0.00	0.9783	0.054
21	[21,11,5,3,3]	135.08	+1.29	**40.89**	+0.11	**0.9789**	**0.052**
25	[25,15,5,3,3]	136.84	+3.05	40.64	−0.14	0.9777	0.055
35	[35,21,5,3,3]	143.27	+9.48	40.23	−0.55	0.9755	0.062

Increasing k from 7 to 15 improves PSNR by 0.48 dB because the larger support can represent longer local exposure trajectories. Expanding to $k = 21$ adds 1.29M parameters and yields a further 0.11 dB, after which the scores decline. At large supports, each kernel must distribute probability over many candidate offsets, making the local integration pattern less selective and harder to estimate from noisy features. The default $k = 15$ captures most of the useful trajectory extent with lower complexity, while $k = 21$ is the accuracy-oriented configuration in the tested range.

Table 16: Multi-scale design controls. ΔPSNR is measured in dB relative to the five-level full model. Loss-side and kernel-schedule changes retain the full-model parameter count.

Factor	Configuration	Params (M)	PSNR↑	ΔPSNR	SSIM↑	LPIPS↓
Depth	Three-level pyramid	89.35	40.13	−0.65	0.9753	0.063
Depth	Five-level pyramid	133.79	**40.78**	0.00	**0.9783**	**0.054**
Supervision	Without coarse observation losses	133.79	40.44	−0.34	0.9768	0.058
Kernel schedule	Legacy [15,7,3,1,1] sequence	133.79	40.25	−0.53	0.9757	0.062

Depth, supervision, and support scheduling test different aspects of the same coarse-to-fine design. The five-level model gains 0.65 dB over the three-level alternative at a cost of 44.44M parameters, exposing the principal accuracy-efficiency trade-off. With capacity fixed, removing coarse observation losses costs 0.34 dB and collapsing the two coarsest kernels to unit support costs 0.53 dB. The results

show that coarse levels benefit not only from greater receptive-field depth but also from observation supervision and nontrivial blur support.

Table 17: Cumulative data-augmentation ablation. ΔPSNR is the conditional difference from the preceding row, in dB.

Setting	Noise	CutNoise	Illum.	Color	PSNR	ΔPSNR	SSIM	LPIPS
No augmentation	–	–	–	–	39.70	–	0.9731	0.067
Physical noise	✓	–	–	–	40.19	+0.49	0.9758	0.061
+ CutNoise	✓	✓	–	–	40.36	+0.17	0.9768	0.058
+ illumination	✓	✓	✓	–	40.50	+0.14	0.9776	0.056
+ color	✓	✓	✓	✓	**40.78**	+0.28	**0.9783**	**0.054**

Calibrated physical noise contributes the largest conditional increment of 0.49 dB in this ordering. CutNoise and illumination adjustment add 0.17 and 0.14 dB, while color adjustment contributes a further 0.28 dB. Because every row retains the preceding transformations, the table evaluates the cumulative augmentation pipeline and supports combining noise modeling with photometric diversity.

Table 18: Sensitivity to coarse flow support. Cost-volume channels equal $(2d+1)^2$. ΔPSNR is measured in dB relative to the default displacement $d = 6$.

Displacement d	Cost-volume channels	Effective support	PSNR↑	ΔPSNR	SSIM↑	LPIPS↓
4	81	±64	40.46	−0.32	0.9774	0.057
6	169	±96	40.78	0.00	0.9783	0.054
8	289	±128	**40.81**	+0.03	**0.9784**	**0.054**

Increasing d from 4 to 6 more than doubles the cost-volume channels and improves PSNR by 0.32 dB, indicating that the narrow search truncates useful correspondence. Increasing d to 8 adds another 120 channels but only 0.03 dB. The default therefore lies near the observed saturation point. This sweep supports sufficient motion coverage rather than unrestricted enlargement of the correspondence search.

Table 19: Cumulative objective construction. ΔPSNR is the conditional difference obtained by adding the listed term to the preceding objective, in dB.

Objective	PSNR↑	ΔPSNR	SSIM↑	LPIPS↓
$\mathcal{L}_1$	39.43	–	0.9553	0.086
$+\mathcal{L}_{\text{SSIM}}$	39.70	+0.27	0.9622	0.080
$+\mathcal{L}_{\text{flow-cyc}}$	39.96	+0.26	0.9661	0.073
$+\mathcal{L}_{\text{smooth}}$	40.13	+0.17	0.9691	0.068
$+\mathcal{L}_s^{\text{obs}}$	40.40	+0.27	0.9733	0.061
$+\mathcal{L}_l^{\text{obs}}$	40.60	+0.20	0.9761	0.057
$+\mathcal{L}_{\text{ctr}}$	**40.78**	+0.18	**0.9783**	**0.054**

The complete objective improves PSNR by 1.35 dB and reduces LPIPS by 0.032 relative to the pixel loss alone. Structural fidelity contributes 0.27 dB, and the flow-cycle term adds another 0.26 dB by coupling reciprocal correspondences. The short- and long-observation losses add 0.27 and 0.20 dB because they require the predicted kernels and transport field to reproduce both measured exposures. Smoothness and kernel centering provide smaller gains by regularizing local variation and fixing the kernel coordinate frame. The progression shows how image-level supervision and formation-based constraints successively stabilize the factorized representation.

Table 20 reports the complete coefficient scans recorded in the updated experiment workbook. Each scan changes one loss group while retaining the other default coefficients. The relative setting $1.0\times$ reaches the highest PSNR in the tested grid for all six groups. The corresponding coefficients are $\lambda_{\text{pix}} = 120$, $\lambda_{\text{str}} = 15$, $\lambda_s^{\text{obs}} = 5$, paired cycle weights $(1, 1)$, $\lambda_{\text{sm}} = 0.05$, and $\lambda_{\text{ctr}} = 0.5$. Pixel weighting is the most sensitive and loses 0.66 dB at $0.50\times$ because it sets the direct gradient

scale of the restoration target. Smoothness and centering remain within 0.10 and 0.16 dB of the default across their scans, consistent with their local regularization roles. The paired correspondence–observation term loses 0.20 dB at $0.20\times$ and 0.30 dB at $5.00\times$. Weak weighting leaves the two latent representations less constrained by formation consistency, while strong weighting narrows the set of image reconstructions compatible with optimization. Figure 9 visualizes these responses.

Table 20: Complete sensitivity scans for the principal loss coefficients on the D2 validation set. Panel (a) groups reconstruction and observation terms, while panel (b) groups correspondence and regularization terms. Relative settings are measured against each term's default while all other coefficients remain fixed. ΔPSNR is relative to the 40.78 dB default.

(a) Reconstruction and observation terms

Term	Rel.	Weight	PSNR↑	ΔPSNR	SSIM↑	LPIPS↓
λ_{pix}	$0.50\times$	60	40.12	−0.66	0.9765	0.062
	$0.75\times$	90	40.48	−0.30	0.9775	0.058
	$1.00\times$	120	**40.78**	0.00	**0.9783**	**0.054**
	$1.25\times$	150	40.72	−0.06	0.9780	0.055
	$1.50\times$	180	40.62	−0.16	0.9776	0.057
	$2.00\times$	240	40.38	−0.40	0.9768	0.060
λ_{str}	$0.33\times$	5	40.52	−0.26	0.9760	0.060
	$0.67\times$	10	40.68	−0.10	0.9775	0.056
	$1.00\times$	15	**40.78**	0.00	**0.9783**	**0.054**
	$1.33\times$	20	40.75	−0.03	0.9785	0.055
	$2.00\times$	30	40.65	−0.13	0.9782	0.058
λ_s^{obs}	$0.20\times$	1	40.55	−0.23	0.9770	0.058
	$0.50\times$	2.5	40.68	−0.10	0.9778	0.056
	$1.00\times$	5	**40.78**	0.00	**0.9783**	**0.054**
	$2.00\times$	10	40.72	−0.06	0.9780	0.055
	$3.00\times$	15	40.58	−0.20	0.9772	0.059

(b) Correspondence and regularization terms

Term	Rel.	Weight	PSNR↑	ΔPSNR	SSIM↑	LPIPS↓
λ_{cyc} $+\lambda_l^{\text{obs}}$	$0.20\times$	$0.2+0.2$	40.58	−0.20	0.9772	0.058
	$0.50\times$	$0.5+0.5$	40.70	−0.08	0.9780	0.056
	$1.00\times$	$1+1$	**40.78**	0.00	**0.9783**	**0.054**
	$2.00\times$	$2+2$	40.72	−0.06	0.9780	0.055
	$5.00\times$	$5+5$	40.48	−0.30	0.9765	0.061
λ_{sm}	$0.20\times$	0.01	40.72	−0.06	0.9778	0.056
	$0.50\times$	0.025	40.76	−0.02	0.9781	0.055
	$1.00\times$	0.05	**40.78**	0.00	**0.9783**	**0.054**
	$2.00\times$	0.1	40.75	−0.03	0.9782	0.055
	$4.00\times$	0.2	40.68	−0.10	0.9775	0.057
λ_{ctr}	$0.20\times$	0.1	40.65	−0.13	0.9775	0.057
	$0.50\times$	0.25	40.72	−0.06	0.9780	0.056
	$1.00\times$	0.5	**40.78**	0.00	**0.9783**	**0.054**
	$2.00\times$	1	40.75	−0.03	0.9781	0.055
	$5.00\times$	2.5	40.62	−0.16	0.9772	0.058

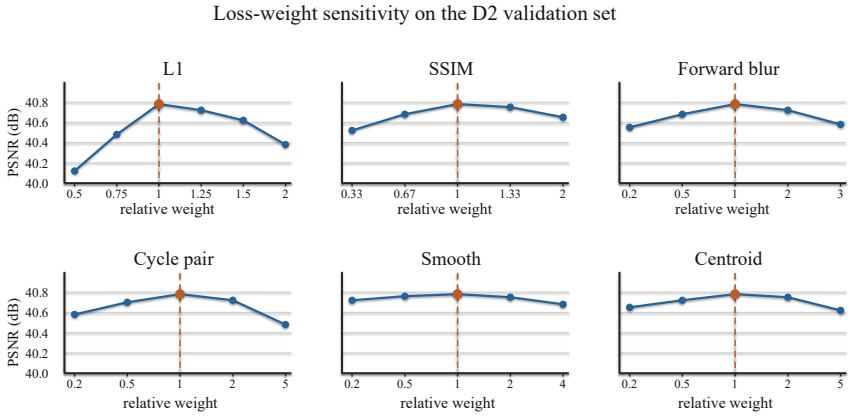

Figure 9: Sensitivity of validation PSNR to the six principal loss-weight groups. Each panel varies one group relative to its default and keeps all other coefficients fixed. The dashed line marks the selected $1.0\times$ setting. The common peak at the central setting and the gradual decline at extreme values support the reported balance between reconstruction fidelity and physical constraints.

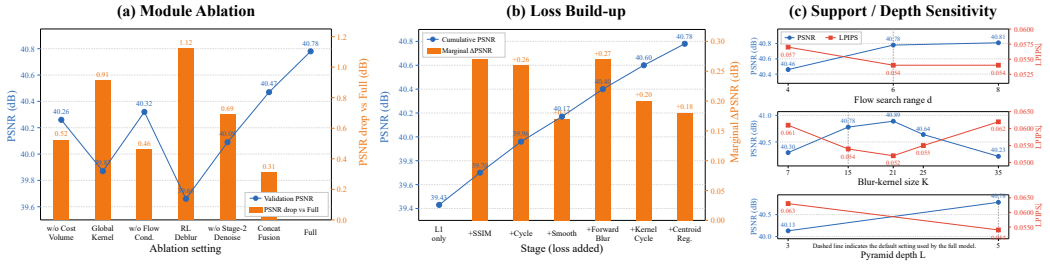

Figure 10: Mechanism ablations and support sensitivities on D2 2880p. Panel (a) reports validation PSNR and the drop from the complete model for correspondence, blur, denoising, and fusion substitutions. Panel (b) reports cumulative PSNR and the marginal gain from each added objective term. Panel (c) reports PSNR and LPIPS while varying flow support d, finest-scale kernel size k, and pyramid depth L, with dashed lines marking the default settings. The largest losses occur when the explicit motion or spatially varying blur representation is removed, while the support sweeps saturate near the selected configuration.

E.6 Qualitative comparisons

The supplementary grids add visual conditions that are not represented by the main-paper lettering and vehicle panels. The illuminated-text example is moved here from the main paper, while the additional D2 scenes emphasize diagonal light trails, repeated illuminated structures, and traffic signals. The cross-dataset grids cover all newly selected PDFs and are grouped by dataset so that scene content, rather than repeated caption wording, identifies the transfer setting. Every grid preserves the same spatial support across inputs, methods, and ground truth.

Figure 11 broadens the D2 evidence beyond the main scenes. The four panels contain distinct combinations of light accumulation, facade texture, and signal geometry. Several competing results retain elongated traces or local duplication, whereas I^{2}EM restores more coherent structures in short-exposure coordinates. The visual variation complements the main quantitative comparison without restating its lettering and vehicle observations.



Figure 11: Additional D2 comparisons. Panel (a) evaluates illuminated text whose strokes are broadened by exposure integration. Panel (b) contains diagonal light trajectories that couple displacement with elongated blur. Panel (c) tests repeated hanging lights and the surrounding thin support structure. Panel (d) combines traffic signals with facade lines at different depths. Every grid uses the same input, reference, baseline, and output order, and the red box marks I^2EM. The restored regions preserve local geometry while reducing elongated traces and duplicated edges.

Figures 12, 13, and 14 extend the visual comparison beyond D2. The transfer grids show that the same factorized representation is exposed to different motion scales, textures, and capture processes, while the method maintains sharper boundaries and fewer residual streaks in the highlighted regions. These examples complement, rather than restate, the main D2 evidence.



Figure 12: Qualitative transfer on GoPro. Panels (a)–(e) cover five sequences with different camera and object motion, edge orientation, texture density, and local contrast. The same blurred frame is supplied to both branches because GoPro contains no paired short–long observations. Every panel uses an identical comparison order, and the red box marks I^{2}EM. The highlighted regions show recovered contours and reduced residual streaks under these motion variations.

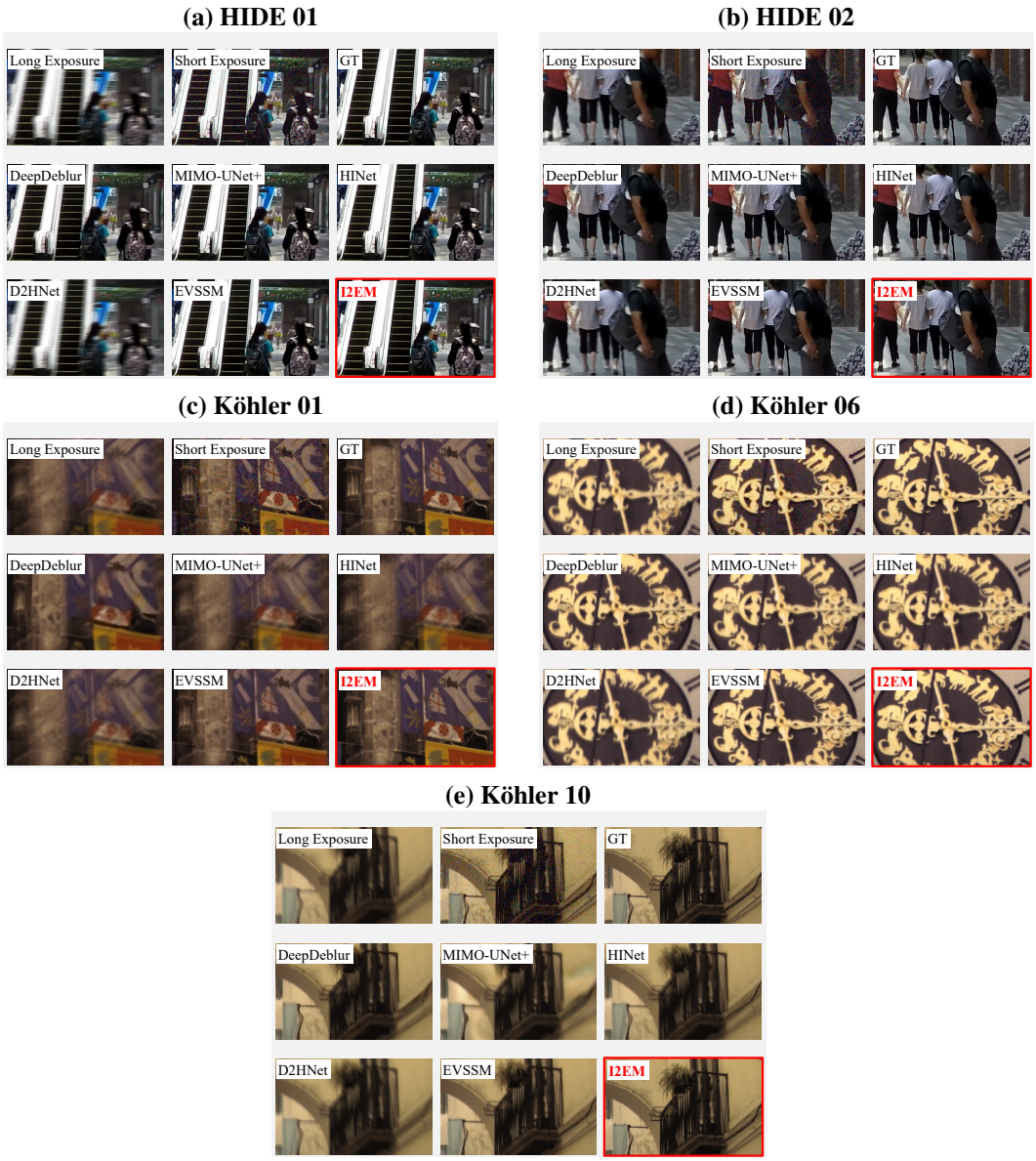

Figure 13: Qualitative transfer on HIDE and Köhler. HIDE panels contain human-scale motion and occlusion boundaries, whereas Köhler panels contain camera-shake-like blur and fine clock geometry. The common comparison order and red boxes marking I^2EM make the complementary motion scales and structures directly comparable.



Figure 14: Qualitative transfer on RealBlur-J and REDS. Panels (a)–(d) use captured RealBlur-J scenes containing text, object boundaries, and low-contrast detail under spatially varying blur. Panel (e) uses a REDS video frame with a different acquisition pipeline and motion distribution. The same blurred observation is supplied to both branches, all panels retain a common comparison order, and red boxes mark I^2EM. The highlighted regions show sharper boundaries and fewer residual streaks across the two domains.