A Benchmark for Heterogeneous Stereo Deblurring with Physically- and Epipolar-constrained Cross Attention

Hoju Shin[1*], Jiah Kim[1*], Seung-Wook Kim[1†], and Seowon Ji[2†]

[1] Department of Intelligent Robot Engineering, Pukyong National University, Busan, Republic of Korea
[2] Department of Computer Science and Engineering, Konkuk University, Seoul, Republic of Korea
{hjshin, whatime08}@pukyong.ac.kr, swkim@pknu.ac.kr, seowonji@konkuk.ac.kr

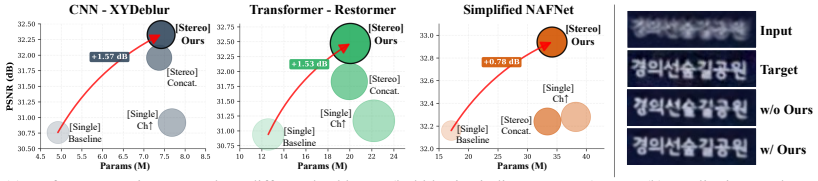
Abstract. Modern stereo-capable smartphones enable immersive XR content capture. However, hardware heterogeneity across camera modules often causes severe asymmetric blur artifacts. Existing methods and benchmarks largely assume homogeneous stereo setups and therefore do not explicitly address such asymmetric degradation. To bridge this gap, we present a dedicated framework for heterogeneous stereo deblurring. First, we introduce the heterogeneous stereo deblurring (HSD) dataset, constructed from real smartphone stereo captures via multi-frame integration. Second, we propose physically- and epipolar-constrained cross attention (PECA), a lightweight module that restricts cross-view matching to an epipolar search window bounded by a optics-derived disparity upper bound. By enforcing physically valid disparity constraints, PECA enables efficient and reliable cross-view feature fusion. Moreover, our confidence-weighted attention with residual fusion emphasizes cross-guided deblurring when correspondences are reliable, while naturally falling back to self-deblurring in occluded or unreliable regions. PECA is architecture-agnostic and consistently improves CNN-, Transformer-, and NAFNet-based baselines. Extensive experiments on HSD show that PECA-enhanced models achieve improved restoration performance with favorable efficiency. The dataset and source code are publicly released at https://github.com/shinhoju/PECA.

Keywords: Heterogeneous stereo-camera systems · Stereo deblurring · Geometry-constrained cross-view attention

1 Introduction

Modern smartphones have evolved into sophisticated imaging devices by adopting heterogeneous multi-camera systems as a standard configuration. These systems generally combine a primary wide camera with auxiliary modules, such as

* Equal contribution.
† Corresponding author.



(a) Performance gains across three different backbones (bubble size indicates MACs). (b) Qualitative result.

Fig. 1: Overview of PECA performance and visual quality on the HSD dataset. (a) Accuracy-efficiency trade-off across three representative backbones (XYDeblur, Restormer, and NAFNet). (b) Qualitative restoration results on a challenging scene.

ultra-wide or telephoto lenses [4]. Beyond enhancing single-view photography, this configuration enables stereoscopic capture and other emerging 3D applications that require synchronized stereo observations to reconstruct content for binocular displays and head-mounted devices [8].

Although this multi-camera configuration supports stereoscopic capture, modern smartphones are not built as calibrated, homogeneous stereo rigs. Instead, they integrate heterogeneous camera modules, resulting in systematic asymmetry in image quality across views. In common devices, the primary wide camera typically benefits from a bright aperture and optical image stabilization (OIS), whereas the ultra-wide camera is often constrained by a narrower aperture and lacks dedicated stabilization. These camera properties make the ultra-wide stream more susceptible to motion. Consequently, stereoscopic captures frequently exhibit a sharp wide view alongside a blur-degraded ultra-wide view, reflecting hardware-induced blur asymmetry.

The use of synchronized stereo cameras naturally suggests using a sharper view to restore a degraded one. However, most existing restoration approaches are not designed with asymmetric hardware degradation as the primary objective. Many stereo restoration methods are developed under homogeneous camera assumptions [19, 20, 23, 26, 28, 34, 36], where both views share comparable imaging characteristics and restoration is formulated symmetrically. In such settings, cross-view correspondence is mainly exploited to resolve geometric misalignment or transfer details between similarly degraded views. This symmetric formulation does not reflect heterogeneous stereo-camera systems, where one view often serves as a high-quality reference while the other suffers from systematic blur.

Recent works have begun exploring heterogeneous multi-camera systems [7, 9, 11, 12, 22, 24, 27, 29, 30, 35]. Although these approaches exploit cross-camera complementarity, many rely on unrectified stereo data and global dense correspondence estimation or iterative alignment, which may incur significant computational overhead and can be sensitive to sensor heterogeneity. Moreover, asymmetric blur is rarely formulated as the primary restoration target under device-level constraints. Consequently, there remains a lack of a dedicated framework and benchmark that directly deal with hardware-induced blur asymmetry under rectified heterogeneous stereo settings.

In this work, we study heterogeneous stereo deblurring in practical smartphone settings, where hardware-induced blur asymmetry across views is the

primary restoration challenge. We introduce a real-world benchmark, the heterogeneous stereo deblurring (HSD) dataset, constructed from synchronized wide and ultra-wide video sequences with device-provided geometric rectification to reflect real mobile pipelines. In HSD, the wide view serves as a sharp reference, while the ultra-wide blur is approximated by temporal integration to emulate exposure-induced degradation. This design decouples geometric alignment from blur restoration, enabling controlled evaluation under practical device environments. To establish a physically grounded and efficient reference model for HSD, we propose physically- and epipolar-constrained cross attention (PECA), a lightweight cross-view fusion module. Leveraging rectified geometry and an optics-based disparity upper bound derived from the minimum focus distance of the stereo pair, PECA restricts cross-view attention to a localized 1D window along the epipolar line. The disparity bound determines the maximum physically feasible correspondence, while rectified stereo geometry further imposes a directional constraint on the search. Together, these constraints confine attention to a physically valid and computationally efficient region. The retrieved cross-view features are then integrated into the target stream via residual fusion, which preserves view-specific characteristics. As illustrated in Fig. 1, PECA consistently improves performance on the HSD dataset across diverse backbone architectures. Our contributions are summarized as follows:

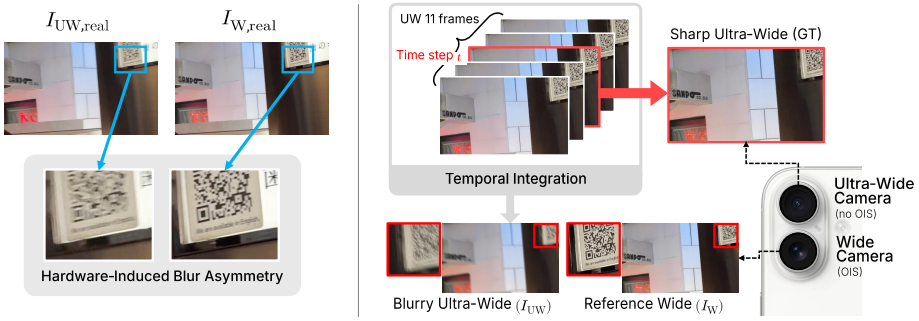
1. We introduce **HSD**, a real-world benchmark of synchronized wide/ultra-wide smartphone stereo captures with device-provided rectification, designed to study hardware-induced asymmetric blur under practical mobile camera pipelines.
2. We propose **PECA**, a lightweight cross-view fusion module that constraints attention to a directional 1D epipolar window whose extent is bounded based on optics, enabling efficient and robust cross-view feature retrieval without global dense matching.
3. We demonstrate the **backbone-agnostic effectiveness** of PECA across CNN-, Transformer-, and modern simplified NAFNet-based models, consistently improving restoration quality with favorable efficiency on HSD.

2 Related Work

2.1 Single-Image Deblurring

Single-image deblurring has been widely studied, evolving from kernel-estimation approaches [33] to modern learning-based models such as CNNs, Transformers, and activation-free networks [1, 2, 6, 10, 13, 14]. Representative datasets such as GoPro [18], REDS [17], RealBlur [21], and HIDE [25] have played a crucial role in advancing single-view deblurring by providing paired blurry-sharp data under diverse motion patterns and camera settings.

However, these methods operate on a single observation and cannot exploit cross-view redundancy available in multi-camera systems. Consequently, they are not designed to address asymmetric degradation across synchronized views,



(a) Real Heterogeneous Stereoscopic Capture (b) HSD Construction Protocol from Real Capture

Fig. 2: Overview of the HSD benchmark construction from real heterogeneous stereo capture. (a) Real smartphone stereoscopic capture, where hardware-induced blur asymmetry naturally arises between wide and ultra-wide modules. (b) Construction protocol applied to real synchronized stereo sequences.

nor are existing single-image benchmarks suitable for evaluating heterogeneous stereo deblurring.

2.2 Stereo Image Restoration

Homogeneous stereo restoration Stereo restoration methods have explored cross-view correspondence to improve reconstruction quality [26,28,36]. In stereo deblurring, prior works leverage scene flow estimation, joint optimization, or disparity-aware feature aggregation to restore latent sharp images [19,20,23,34]. Related approaches in stereo super-resolution (SR) similarly exploit cross-view correspondences to transfer fine-grained details.

These methods are typically developed under homogeneous camera assumptions, where both views share similar sensor characteristics and image quality. Restoration is therefore formulated symmetrically, treating each view as both source and target. While effective in controlled homogeneous stereo setups, this assumption does not hold in heterogeneous smartphone systems, where auxiliary cameras often exhibit systematic quality degradation.

Heterogeneous stereo restoration Recent works have explored heterogeneous multi-camera systems, leveraging cross-camera complementarity for SR and deblurring [7, 9, 11, 12, 24, 27, 29, 30, 35]. In mobile imaging, hybrid-camera deblurring has also been studied by combining complementary captures across smartphone modules and introducing datasets for evaluation [22]. These lines of work highlight the potential of cross-camera information in asymmetric settings.

However, many approaches operate on unrectified stereo data and rely on global dense correspondence estimation or iterative alignment, incurring significant computational overhead and sensitivity to sensor heterogeneity. Although device-provided rectification reduces correspondence search to a 1D epipolar space, this structural prior remains under-exploited for efficient asymmetric

restoration. Furthermore, the lack of a dedicated benchmark tailored to rectified heterogeneous stereo with controlled blur asymmetry has limited systematic evaluation under realistic device pipelines.

Epipolar-aware attention In stereo SR, attention-based stereo fusion has been adopted to enhance cross-view feature interaction. Some methods such as PASSRnet [26], iPASSR [28], and NAFSSR [3] restrict correspondence search to horizontal epipolar lines, reducing complexity compared to global 2D attention.

Nevertheless, these approaches typically search across the full image width without explicit physical constraints. Without camera-optics-derived disparity bounds, attention may aggregate features from geometrically implausible regions along the epipolar line. Moreover, correspondence search is often performed symmetrically along the scanline, ignoring the physically valid disparity direction imposed by rectified stereo geometry. In contrast, our work incorporates an optics-based disparity upper bound and a directional epipolar constraint to confine cross-view matching to a physically plausible and computationally efficient search region.

3 HSD Benchmark Construction

As shown in Fig. 2(a), heterogeneous smartphone cameras often exhibit systematic blur asymmetry between wide and ultra-wide modules. To study this phenomenon, we construct a real-world benchmark for heterogeneous stereo deblurring from synchronized stereo captures. The benchmark is built from real heterogeneous stereo capture, preserving device-level characteristics while introducing controlled exposure simulation. It is designed to (i) reflect device-level camera processing used in commodity smartphones, (ii) approximate physically plausible motion blur, and (iii) avoid trivial or near-static cases that would otherwise obscure the asymmetric deblurring problem.

3.1 Data Acquisition and Camera Configuration

All data are captured using the stereoscopic capture mode of commercial smartphones (e.g., iPhone 15 Pro Max, iPhone 16 series, and iPhone 17) equipped with heterogeneous stereo systems. Each recording provides synchronized wide and ultra-wide video streams observing the same scene. The captured videos are rectified through the native imaging pipeline, which dynamically compensates for variations in focal length and principal points induced by the auto focus (AF) and OIS, producing epipolar-aligned stereo frames. We directly adopt the device-provided rectified outputs without additional calibration to reflect practical deployment scenarios. The final collection contains diverse indoor and outdoor environments with varying lighting and depth distributions.

3.2 Blur Generation via Temporal Integration

A key challenge in stereo deblurring benchmarks is obtaining blur–sharp pairs that reflect real-world image formation. In heterogeneous smartphone systems,

motion blur in the ultra-wide stream primarily arises from longer effective exposure times and weaker stabilization relative to the primary wide camera. To model this exposure-induced degradation, we adopt the temporal integration protocol established by Nah *et al.* [18] while adapting it to the constraints of a synchronized stereo capture system.

Unlike conventional benchmarks that rely on high-speed cameras (e.g., 240 fps), such synchronized capture is impractical for commercial smartphones. Instead, we adopt a scene-diversity-centric acquisition strategy. We collect a large set of 384 scenes with controlled and stable camera motion, ensuring moderate inter-frame displacement and avoiding unnatural artifacts during temporal integration.

As shown in Fig. 2(b), for each sequence, we generate blurred ultra-wide frames by averaging consecutive frames centered at a target time step t. This accumulation mimics a longer exposure interval in which a sensor integrates light over time. Since our raw sequences are captured with stable motion, this integration results in a smooth, continuous blur that faithfully approximates real-world degradation without the discrete step artifacts typical of low-FPS averaging with fast motion. Because the integration follows real handheld camera trajectories, the resulting blur is depth-dependent and spatially varying rather than uniform. The central frame is treated as the ground truth, while the synchronized wide-view frame at the same timestamp serves as the sharp cross-view reference. We do not explicitly model the rolling shutter effect, as this would require assumptions about device-specific readout patterns and ISP patterns. Consequently, the HSD dataset does not fully account for rolling-shutter distortions. To mitigate noise artifacts while maintaining tractable computational complexity, we downscale the synthesized ultra-wide blurry frames as well as the recorded sharp ultra-wide and wide frames from 1080×1920 to 720×1280 using bicubic interpolation.

3.3 Dataset Curation and Partitioning

To ensure that the benchmark presents a non-trivial challenge and maintains evaluation validity, we curate the dataset before final partitioning. Since temporal integration does not always guarantee sufficiently severe degradation, we filter out near-static samples where integration yields negligible blur. Specifically, we compute the peak signal-to-noise ratio (PSNR) between each integrated blurry image and its corresponding sharp ultra-wide frame, and discard pairs whose PSNR exceeds 33dB. This filtering step ensures that the benchmark emphasizes scenarios where asymmetric deblurring provides meaningful gains.

After curation, we split the 384 sequences at the sequence level to prevent data leakage. We randomly assign 256 sequences for training and 128 for testing. From these splits, we sample 2,200 training pairs and 1,100 testing pairs to form the final benchmark. This protocol ensures diverse motion patterns and scene structures while enabling reproducible quantitative evaluation. Detailed dataset statistics and qualitative examples are provided in Sec. A of the supplementary document.

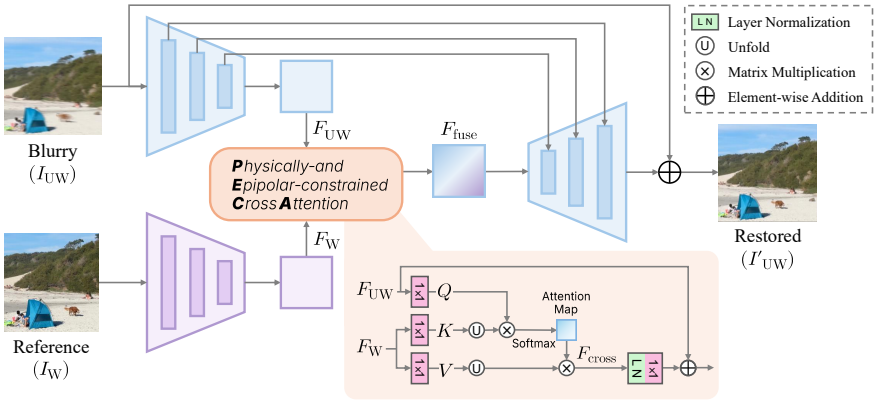


Fig. 3: Overview of the proposed physically- and epipolar-constrained cross attention (PECA). The framework consists of dual encoders for the blurry ultra-wide input and the sharp wide reference, followed by the PECA module and residual feature fusion.

4 Proposed Method

4.1 Design Principles

We design PECA to address heterogeneous stereo deblurring in practical mobile capture, where blur is systematically asymmetric across views, and stereo pairs are provided through a device-level rectification pipeline, producing approximately epipolar-aligned observations. PECA is guided by following three principles. **(i) Asymmetric restoration:** The wide view serves as a sharp reference while the ultra-wide view is blur-degraded; cross-view interaction should therefore be selective to enhance the target view without enforcing symmetric restoration. **(ii) Geometry- and physics-constrained correspondence:** Rectification restricts valid matches to epipolar scanlines, stereo geometry determines the disparity direction, and camera parameters provide a physically feasible disparity bound. PECA uses these priors to confine attention to a compact and feasible epipolar window. **(iii) Robust fusion without explicit matching or masks:** Dense matching or occlusion masks can be costly and sensitive to sensor heterogeneity and residual rectification errors. Instead, PECA relies on confidence-weighted attention within the constrained window and residual fusion to naturally attenuate unreliable regions. This design produces an efficient and practical cross-view module that exploits rectified stereo geometry while remaining robust to partial visibility and device-level imperfections.

4.2 Overall Framework

The goal of our framework is to restore a sharp ultra-wide image I'_{UW} from a blurry ultra-wide observation I_{UW} by leveraging a synchronized sharp wide-view reference I_W. In a heterogeneous stereo camera system, blur is typically concentrated in the ultra-wide stream, whereas the wide stream retains higher

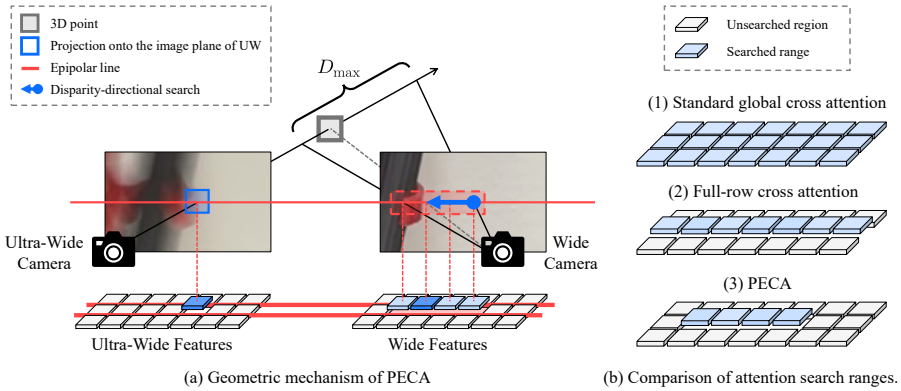


Fig. 4: An illustration of the proposed PECA module. (a) The physically derived disparity bound $D_{\max}$ restricts correspondence search along the epipolar line. (b) Progressive restriction of attention search space: global cross attention (full 2D), full-row cross attention (full 1D scanline), and PECA (constrained 1D disparity window).

spatial fidelity. We therefore formulate restoration asymmetrically, restoring I_{UW} while using I_{W} as a guidance source.

As illustrated in Fig. 3, we adopt a dual-branch encoder architecture. The primary (top) branch encodes I_{UW} into self-view features F_{UW}, while the reference (bottom) branch encodes I_{W} into reference features F_{W} that preserve sharp structural cues. The proposed PECA module retrieves cross-view features within a physically and geometrically constrained epipolar window, so the guidance signal is formed only from correspondence candidates that are plausible under rectified stereo geometry.

We handle partial visibility through confidence-modulated fusion. Some locations admit reliable cross-view correspondence, while others suffer from occlusions, texture ambiguity, or residual rectification errors. PECA captures this variation through its attention responses within the constrained window. When the attention weights concentrate on a few candidates, the module produces a strong cross-view guidance signal. When the weights spread out across the window, the guidance signal becomes weak and less informative. Residual fusion then injects the cross-view signal into the primary features while preserving the self-view pathway as an identity residual, enabling reference-guided enhancement in well-matched regions and relying on the self-view residual path in regions with weak or ambiguous correspondence.

Building on this overview, we next present PECA as a Transformer-style cross attention mechanism tailored to rectified stereo, where F_{UW} and F_{W} are mapped to query–key–value embeddings and attention is evaluated only within a physically bounded epipolar window to produce cross-view guidance features.

4.3 Physically- and epipolar-constrained cross attention (PECA)

Feature embeddings Let $F_{\mathrm{UW}} \in \mathbb{R}^{H\times W\times C}$ denote the latent feature from the primary (ultra-wide) branch and $F_{\mathrm{W}} \in \mathbb{R}^{H\times W\times C}$ denote the latent feature from the reference (wide) branch at the same spatial scale. To perform cross attention, we map them into a shared embedding space via pointwise projections as follows:

$$Q = \mathrm{Proj}(F_{\mathrm{UW}}; W_q),\ \ K = \mathrm{Proj}(F_{\mathrm{W}}; W_k),\ \ \text{and}\ \ V = \mathrm{Proj}(F_{\mathrm{W}}; W_v), \tag{1}$$

where $\mathrm{Proj}(\cdot; W)$ denotes a pointwise convolution parameterized by W, and W_q, W_k, and W_v are the projection weights for query, key, value, respectively. This design forms the queries from the degraded ultra-wide features, while extracting keys and values from the sharp wide reference, thereby enabling asymmetric cross-view retrieval.

Physically derived disparity constraint As illustrated in 4(a), instead of searching across the entire image or full epipolar line, we use stereo geometry to obtain a physics-guided upper bound on disparity and restrict correspondence search to a plausible range. Let $d_{\mathrm{min}}^{\mathrm{W}}$ and $d_{\mathrm{min}}^{\mathrm{UW}}$ denote the minimum focus distances of the wide and ultra-wide cameras, respectively, and define the effective minimum acquisition distance as

$$d_{\mathrm{eff}} = \max(d_{\mathrm{min}}^{\mathrm{W}}, d_{\mathrm{min}}^{\mathrm{UW}}). \tag{2}$$

Given the baseline distance B between camera centers and the focal length f (in pixel units), the maximum physically feasible disparity is approximately bounded by $f \cdot B/d_{\mathrm{eff}}$. When PECA operates at a reduced feature resolution, we can map this bound to the feature grid using a down-scaling factor S, which yields

$$D_{\mathrm{phys}} = \left\lceil \frac{f \cdot B}{d_{\mathrm{eff}}} \times \frac{1}{S} \right\rceil, \tag{3}$$

where S accounts for input resizing and architectural down-sampling. This quantity provides a geometry-consistent ceiling on the correspondence search range in the feature space. For our W/UW pair, the input-level disparity bound $f \cdot B/d_{\mathrm{eff}}$ is approximately 70 pixels, which corresponds to $D_{\mathrm{phys}} \approx 17$ when $S = 4$.

It is worth noting that D_{phys} reflects only the geometric scaling induced by S, with coarser feature grids yielding proportionally smaller bounds. Since PECA matches encoded features, each feature embedding summarizes a spatial neighborhood whose extent depends on the encoder design, including network depth and kernel configuration, *i.e.*, the effective receptive field. As the receptive field grows, fewer discrete offsets along the epipolar line can be sufficient to cover the physically feasible disparity range. We therefore treat D_{phys} as an upper bound and choose the operational window size D_{max} enforcing $D_{\mathrm{max}} \leq D_{\mathrm{phys}}$.

Epipolar-constrained attention Since device-provided rectification aligns stereo pairs along epipolar scanlines, valid correspondences for a query location (i, j) may lie on the same row i. Under the stereo-camera configuration,

disparity is horizontal and bounded in a single direction along the scanline. Using the operational window size $D_{\max}$ chosen under the upper bound D_{phys}, we define the localized 1D search window as

$$\Omega_{i,j} = \{(i,k) \mid j - D_{\max} \leq k \leq j\}. \tag{4}$$

The cross-view feature at (i,j) is then computed by attending to the candidates in $\Omega_{i,j}$ as

$$F_{\text{cross}}(i,j) = \sum_{(i,k)\in\Omega_{i,j}} \text{Softmax}_k \left(\frac{\mathcal{D}_{\cos}(Q_{i,j}, K_{i,k})}{\tau} \right) \cdot V_{i,k}, \tag{5}$$

where $\mathcal{D}_{\cos}(\cdot,\cdot)$ denotes the cosine similarity, and τ is the softmax temperature controlling the sharpness of the attention distribution.

In regions with reliable cross-view correspondence, attention weights concentrate around geometrically consistent matches, enabling effective transfer of sharp textures. In regions affected by occlusions, ambiguous textures, or residual rectification errors, the similarity scores become less distinctive within $\Omega_{i,j}$, producing diffuse weights and weaker guidance. This confidence-modulated response, together with residual fusion, allows PECA to emphasize reference information when correspondence is reliable and to rely on the self-view pathway otherwise, without explicit matching or occlusion masks.

Computational complexity As shown in Fig. 4(b-1), standard global cross attention performs an unconstrained 2D search over the spatial domain, incurring $\mathcal{O}((HW)^2)$ complexity. By exploiting stereo geometry, prior stereo SR methods [3, 26, 28] reduce this to $\mathcal{O}(HW^2)$ by restricting correspondence search to horizontal scanlines, which we refer to as full-row cross attention (Fig. 4(b-2)). PECA further constrains the search to a physically bounded 1D disparity window (Fig. 4(b-3)), leading to a complexity of

$$\mathcal{O}(HWD_{\max}),$$

where $D_{\max} \ll W$ in practical mobile stereo configurations. This restriction significantly reduces computational overhead while preserving geometrically valid correspondence.

4.4 Complementary Feature Fusion

Finally, cross-view features are integrated with self-view features via residual summation:

$$F_{\text{fuse}} = F_{\text{UW}} + \text{Proj}(F_{\text{cross}}; W_{\text{cross}}), \tag{6}$$

where W_{cross} is the learnable projection weights for cross-view feature aligning feature dimensions. Since the magnitude of F_{cross} is governed by attention confidence within the physically constrained window, residual fusion effectively acts

as a soft, data-adaptive gating mechanism. When confident matches exist, cross-view enhancement becomes dominant. Otherwise, when attention responses are weak or inconsistent, the identity pathway ensures stable self-deblurring.

Through this complementary interaction, PECA exploits cross-view redundancy while maintaining robustness under occlusion and geometric uncertainty, providing an efficient and physically grounded solution for heterogeneous stereo deblurring.

5 Experiments

5.1 Experimental Setup

All models were implemented in PyTorch and trained on NVIDIA GPUs. We evaluated all methods on the proposed HSD dataset (see Sec. 3 for details). To demonstrate its architectural generality, PECA was incorporated into three representative restoration backbones: CNN-based XYDeblur [6], Transformer-based Restormer [31], and the simplified NAFNet [1]. Each backbone was trained for 400K iterations with a batch size of 8 using the AdamW [15] optimizer and a cosine annealing learning rate scheduler.

Optimization hyperparameters follow the original settings of each backbone, with the initial learning rate adjusted for our batch size under the standard training heuristics [5]. Specifically, the learning rate, weight decay, and AdamW β parameters were set to 1×10^{-4}, 1×10^{-3}, and $(0.9, 0.9)$ for XYDeblur, 3×10^{-4}, 1×10^{-4}, and $(0.9, 0.999)$ for Restormer, and 5×10^{-4}, 1×10^{-3}, and $(0.9, 0.9)$ for NAFNet, respectively.

We applied random cropping, gamma, and shot/read noise augmentation. Specifically, input images are randomly cropped into patches of size 128×128 for XYDeblur and Restormer, and 256×256 for NAFNet. To ensure a fair comparison of PECA across architectures, we fixed the spatial resolution of the deepest latent features to 32×32 for all backbones, following the default configuration of XYDeblur. Accordingly, NAFNet was configured as a three-level architecture with [1, 1, 28] and [1, 1, 1] NAFBlocks in the encoder and decoder, respectively. For Restormer, the downsampling operation preceding the fourth encoder stage was removed.

Note that computational complexity (MACs) was measured at an input resolution of 720×1280. Unless otherwise specified, we fix the softmax temperature to $\tau = 0.01$ and use a single default window size $D_{\max} = 5$ across all backbones to avoid backbone-specific tuning. Sensitivity analyses are provided in Sec. 5.4, with additional qualitative results in Supplementary Sec. H.

5.2 Comparison on the HSD Dataset

We evaluate PECA on three representative restoration backbones, XYDeblur [6], Restormer [31], and NAFNet [1]. For each backbone, we compare four variants: the single-image baseline, a channel-expansion model that increases capacity in the single-view setting, a naive stereo concatenation model, and the proposed

Table 1: Performance and complexity comparison on HSD. Four model variants are evaluated across three backbones representing distinct architectural paradigms.

Backbone	Variant	Input	PSNR↑	SSIM↑	Params. (M)↓	MACs (G) ↓
XYDeblur [6]	baseline	Single	30.76	0.9444	4.92	1056.5
	+ channel-expansion	Single	30.92	0.9462	7.68	1650.1
	+ stereo concatenation	Stereo	31.96	0.9595	7.37	1373.5
	+ **PECA (Ours)**	Stereo	**32.22**	**0.9620**	7.42	1376.4
Restormer [31]	baseline	Single	30.94	0.9444	12.59	2139.8
	+ channel-expansion	Single	31.17	0.9471	22.17	3754.7
	+ stereo concatenation	Stereo	31.83	0.9581	19.94	2851.3
	+ **PECA (Ours)**	Stereo	**32.47**	**0.9636**	20.05	2857.8
NAFNet [1]	baseline	Single	32.16	0.9579	17.08	832.6
	+ channel-expansion	Single	32.28	0.9590	38.22	1860.5
	+ stereo concatenation	Stereo	32.24	0.9604	33.38	1567.4
	+ **PECA (Ours)**	Stereo	**32.92**	**0.9669**	34.17	1589.6

PECA module. Parameter counts are kept comparable whenever possible for fair comparison.

Table 1 reports quantitative results on the HSD dataset. Incorporating the synchronized wide-view reference consistently improves performance over single-image baselines across all backbones. The gains from PECA are not simply explained by increased model capacity. Despite using more parameters and higher computational cost, the channel-expansion models underperform PECA. For example, with Restormer, PECA improves PSNR by +1.30 dB over channel-expansion while requiring lower computational cost. This suggests that the improvement of PECA stems from structured cross-view feature retrieval rather than simply scaling up the network. Even naive stereo concatenation yields noticeable gains (*e.g.*, +1.20 dB for XYDeblur), confirming that cross-view information is beneficial for asymmetric deblurring. Compared with stereo concatenation, PECA consistently achieves higher restoration quality (*e.g.*, +0.68 dB on NAFNet), indicating that explicitly modeling cross-view correspondence is critical beyond simple feature fusion. Additional analysis in Supplementary Sec. B examines this behavior under severe occlusion, where PECA maintains positive gains while naive stereo concatenation can fall below the single-view baseline. Qualitative results in Fig. 5 further show that PECA restores sharper edges and fine textures, particularly in severely blurred regions and small text patterns, while avoiding over-smoothed reconstructions. Overall, the consistent gains across CNN-, Transformer-, and NAFNet-based architectures demonstrate that PECA provides a robust and backbone-agnostic mechanism for heterogeneous stereo deblurring.

5.3 Real-world Qualitative Evaluation

Beyond the temporally integrated pairs in HSD, we additionally provide qualitative results on real handheld heterogeneous stereo captures, where the target blur arises directly from natural capture dynamics rather than the dataset construction protocol. These sequences contain severe motion blur with depth variation and frequent occlusions, and no ground-truth sharp images are available

Fig. 5: Qualitative results on the HSD dataset for three backbones, evaluated based on the presence of PECA. Three zoomed regions focusing on large-scale blur, transparent vinyl, and fine-grained text are highlighted.

Fig. 6 compares each baseline backbone with its PECA-enhanced counterparts. The baseline models often produce over-smoothed textures or unstable edge reconstruction under strong blur. In contrast, the PECA exhibits sharper structures and clearer fine details across all three architectures, indicating that the proposed reference-guided fusion remains effective under real handheld blur. Although quantitative evaluation is not available for these real samples, the consistent qualitative improvements complement the paired benchmark results and support the practical applicability of PECA in realistic heterogeneous smartphone imaging conditions.

5.4 Ablation Study

Effect of attention search ranges Table 2 compares three cross-view attention designs that differ only in their correspondence search space. Global cross attention performs exhaustive 2D matching (HW), which incurs high computational cost (852.2G MACs for the attention module). Under device-provided rectification, valid correspondences are strongly concentrated on the epipolar scanline; hence, if similarity matching is effective, the off-row candidates in global search are unlikely to receive high attention weights, making global and full-row-based search yield similar restoration behavior in practice. Full-row attention explicitly restricts the search to the epipolar line, reducing the number of MACs by 112×, but it still considers the range along the scanline. In asymmetric deblurring, blur suppresses discriminative high-frequency cues and increases pattern repetition along a row, so many distant locations become competitive, leading to diffuse attention and limited gains (+0.03 dB) over the global attention model.

In contrast, PECA further constrains matching to a geometrically plausible disparity window, filtering out scanline candidates that are physically unlikely to correspond to the query. This reduces ambiguity and computation costs, result-

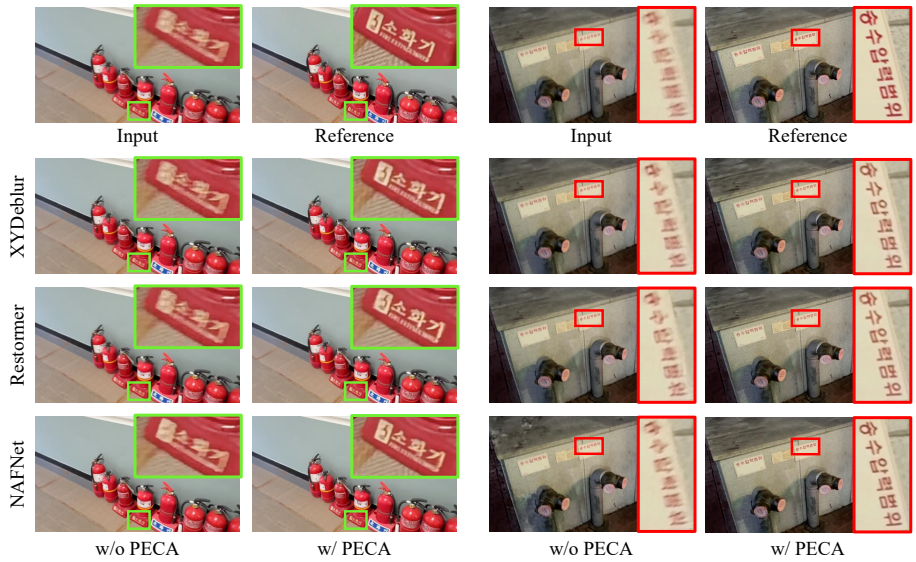


Fig. 6: Qualitative comparison on real handheld stereo captures without ground truth.

Table 2: Effect of search-space design on deblurring performance. All attention variants use identical 1×1 projection layers and cosine similarity with $\tau = 0.01$.

Method	Search space	PSNR ↑	SSIM ↑	Module MACs (G) ↓	Total MACs (G) ↓
Global	2D (HW)	30.72	0.9439	852.2	2225.7
Full-row	1D (W)	30.75	0.9438	7.6	1381.0
PECA	1D ($D_{\max}$)	**32.22**	**0.9620**	**2.9**	**1376.4**

ing in a PSNR of 32.22 dB (+1.50 dB/+1.47 dB over the global/full-row models) with only 2.9G MACs for the attention module. Additional qualitative comparisons and the top-3 attention distribution analysis corresponding to these results are provided in the supplementary document, in Sec. F and Sec. G, respectively.

Parameter sensitivity of $D_{\max}$ and τ Fig. 7 analyzes how the epipolar window size $D_{\max}$ and the softmax temperature τ interact in PECA using the XYDeblur [6] backbone. When the attention distribution is sufficiently sharp (*e.g.*, τ=0.01), PECA achieves consistently strong performance across a range of window sizes, from $D_{\max}$=3 up to $D_{\max}$=16 (which is close to the physically feasible disparity range D_{phys}). We observe the best PSNR at $D_{\max}$=9 for XYDeblur when τ=0.01, while smaller windows deliver comparable accuracy with reduced attention cost. As discussed in Sec. 4.3, the preferred $D_{\max}$ can vary across backbones due to differences in the effective receptive field, which depends on kernel configuration and network depth. Since our goal is to verify the general improvement of PECA rather than to tune $D_{\max}$ for each specific architecture, unless otherwise specified, we use a single default window size $D_{\max}$=5 in our main experiments.

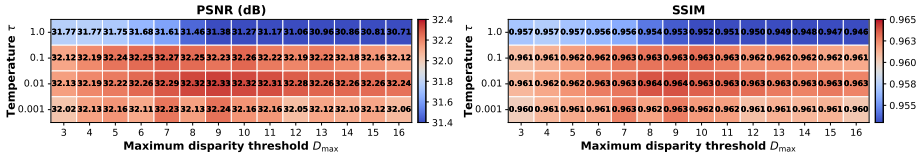


Fig. 7: PSNR and SSIM heatmaps over $D_{\max}$ and τ with the XYDeblur backbone.

Regarding τ, a small temperature yields a peaky attention distribution that better isolates reliable correspondences, whereas a large temperature (*e.g.*, τ=1.0) produces relatively diffuse weights due to the bounded cosine similarity, which may weaken discrimination and degrades restoration quality. In this diffuse regime, expanding $D_{\max}$ can be less beneficial since a wider window mainly introduces additional ambiguous candidates along the scanline.

6 Conclusion

In this paper, we addressed asymmetric stereo deblurring in heterogeneous stereo camera systems, where hardware-induced quality discrepancies naturally arise between views. We proposed PECA, which embeds physical geometric priors into cross-view attention. This approach enables efficient and reliable information transfer by replacing unconstrained global search with localized and directional epipolar-aware matching. We also introduced the HSD Dataset, a real-world benchmark designed to evaluate asymmetric stereo deblurring in heterogeneous stereo camera systems. Extensive experiments demonstrate that PECA consistently improves deblurring performance across diverse network architectures while maintaining computational efficiency suitable for practical deployment. These results validate the effectiveness of geometry-constrained cross-view modeling in addressing asymmetric degradation under practical heterogeneous camera settings. We believe that incorporating physically grounded constraints into cross-view attention provides a principled foundation for future asymmetric multi-camera restoration methods.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (RS-2025-16067383). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the virtual convergence support program to nurture the best talents (IITP-2026-RS-2023-00256615) grant funded by the Korea government (MSIT).

References

1. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: European conference on computer vision. pp. 17–33. Springer (2022)
2. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4641–4650 (2021)
3. Chu, X., Chen, L., Yu, W.: Nafssr: Stereo image super-resolution using nafnet. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1239–1248 (2022)
4. Delbracio, M., Kelly, D., Brown, M.S., Milanfar, P.: Mobile computational photography: A tour. Annual review of vision science **7**(1), 571–604 (2021)
5. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
6. Ji, S.W., Lee, J., Kim, S.W., Hong, J.P., Baek, S.J., Jung, S.W., Ko, S.J.: Xydeblur: Divide and conquer for single image deblurring. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17421–17430 (2022)
7. Kang, J., Yao, R., Zhu, H., Sun, K., Li, X., Zhao, J., Zhou, Y.: Dual-lens super-resolution with semantic-enhanced feature matching and adaptive texture transfer. In: Journal of Physics: Conference Series. vol. 3108, p. 012021. IOP Publishing (2025)
8. Kim, S.: Generation of stereo images from the heterogeneous cameras. Journal homepage: http://iieta. org/journals/i2m **20**(2), 73–78 (2021)
9. Kim, Y., Lim, J., Cho, H., Lee, M., Lee, D., Yoon, K.J., Choi, H.J.: Efficient reference-based video super-resolution (ervsr): Single reference image is all you need. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1828–1837 (2023)
10. Kong, L., Dong, J., Ge, J., Li, M., Pan, J.: Efficient frequency domain-based transformers for high-quality image deblurring. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5886–5895 (2023)
11. Lee, J., Lee, M., Cho, S., Lee, S.: Reference-based video super-resolution using multi-camera video triplets. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17824–17833 (2022)
12. Lin, M., Zhang, C., He, C., Yu, L.: Learning parallax for stereo event-based motion deblurring. IEEE Transactions on Circuits and Systems for Video Technology (2025)
13. Liu, H., Li, B., Lu, M., Wu, Y.: Real-world image deblurring via unsupervised domain adaptation. In: International Symposium on Visual Computing. pp. 148–159. Springer (2023)
14. Liu, H., Liu, C., Xu, J., Jiang, P., Lu, M.: Xyscannet: A state space model for single image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 779–789 (2025)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
16. Min, J., Jeon, Y., Kim, J., Choi, M.: S2m2: Scalable stereo matching model for reliable depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26729–26739 (2025)
17. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: 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. 0–0 (2019)
18. Nah, S., Hyun Kim, T., Mu Lee, K.: 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)
19. Pan, L., Dai, Y., Liu, M., Porikli, F.: Simultaneous stereo video deblurring and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4382–4391 (2017)
20. Pan, L., Dai, Y., Liu, M., Porikli, F., Pan, Q.: Joint stereo video deblurring, scene flow estimation and moving object segmentation. IEEE Transactions on Image Processing **29**, 1748–1761 (2019)
21. Rim, J., Lee, H., Won, J., Cho, S.: Real-world blur dataset for learning and benchmarking deblurring algorithms. In: European conference on computer vision. pp. 184–201. Springer (2020)
22. Rim, J., Lee, J., Yang, H., Cho, S.: Deep hybrid camera deblurring for smartphone cameras. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
23. Sellent, A., Rother, C., Roth, S.: Stereo video deblurring. In: European conference on computer vision. pp. 558–575. Springer (2016)
24. Shekarforoush, S., Walia, A., Brubaker, M.A., Derpanis, K.G., Levinshtein, A.: Dual-camera joint deblurring-denoising. arXiv preprint arXiv:2309.08826 (2023)
25. Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., Shao, L.: Human-aware motion deblurring. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5572–5581 (2019)
26. Wang, L., Wang, Y., Liang, Z., Lin, Z., Yang, J., An, W., Guo, Y.: Learning parallax attention for stereo image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12250–12259 (2019)
27. Wang, T., Xie, J., Sun, W., Yan, Q., Chen, Q.: Dual-camera super-resolution with aligned attention modules. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2001–2010 (2021)
28. Wang, Y., Ying, X., Wang, L., Yang, J., An, W., Guo, Y.: Symmetric parallax attention for stereo image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 766–775 (2021)
29. Xiao, Z., Wang, X.: Asymmetric dual-lens video deblurring. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
30. Yue, H., Cui, Z., Li, K., Yang, J.: Kedusr: Real-world dual-lens super-resolution via kernel-free matching. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6881–6889 (2024)
31. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: 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)
32. Zhang, S., Yu, W., Jiang, F., Nie, L., Yao, H., Huang, Q., Tao, D.: Stereo image restoration via attention-guided correspondence learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(7), 4850–4865 (2024)
33. Zhang, T., Lu, J., Jin, Q., Zeng, T.: A survey of single image blind motion deblurring from traditional to deep learning. ACM Computing Surveys (2025)
34. Zhou, S., Zhang, J., Zuo, W., Xie, H., Pan, J., Ren, J.S.: Davanet: Stereo deblurring with view aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10996–11005 (2019)

35. Zou, H., Suganuma, M., Okatani, T.: Refvsr++: Exploiting reference inputs for reference-based video super-resolution. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2756–2765. IEEE (2025)
36. Zou, W., Gao, H., Chen, L., Zhang, Y., Jiang, M., Yu, Z., Tan, M.: Cross-view hierarchy network for stereo image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1396–1405 (2023)

A Benchmark for Heterogeneous Stereo Deblurring with PECA
–*Supplementary Document*–

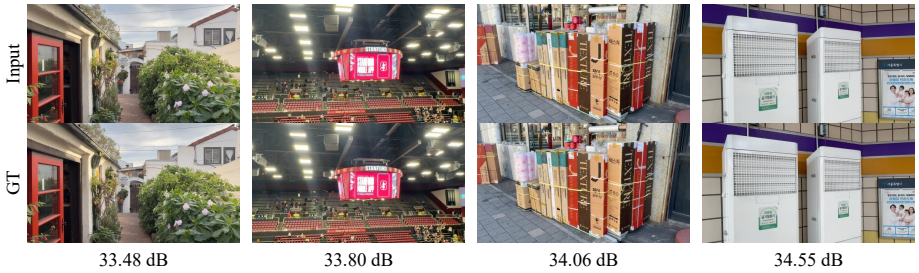
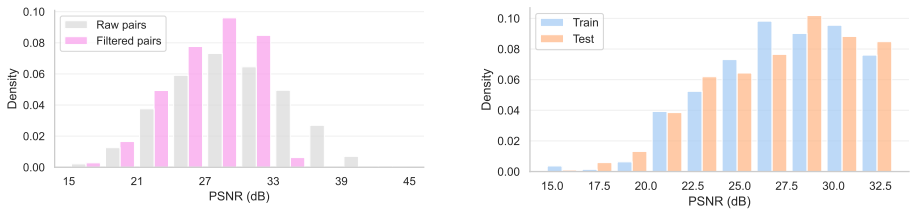


Fig. A: Examples of image pairs with PSNR above 33 dB.

A Dataset Statistics and Qualitative Examples

To analyze dataset characteristics and identify nearly static scenes, we compute the PSNR between blurry and sharp ultra-wide image pairs. Through visual inspection, we observe that pairs with PSNR above approximately 33 dB correspond to scenes with negligible motion blur. Since such samples provide insufficient supervision for motion deblurring, we exclude pairs whose PSNR exceeds this threshold. Examples of such nearly static scenes are shown in Fig. A.

After filtering, we obtain 37,623 image pairs from 384 sequences. We randomly split the dataset into training and test sets at the sequence level to avoid data leakage between the two splits. Specifically, we use 256 sequences for training and 128 sequences for testing. We then randomly sample 2,200 and 1,100 image pairs from the training and test sequences, respectively. Figs. B(a) and (b) illustrate the PSNR distributions before and after filtering, and those of the training and test splits, respectively. The two splits exhibit similar distributions, suggesting that the partitioning does not introduce significant statistical bias.



(a) Before and after filtering.

(b) Training and test splits.

Fig. B: Distribution of PSNR in the HSD dataset.

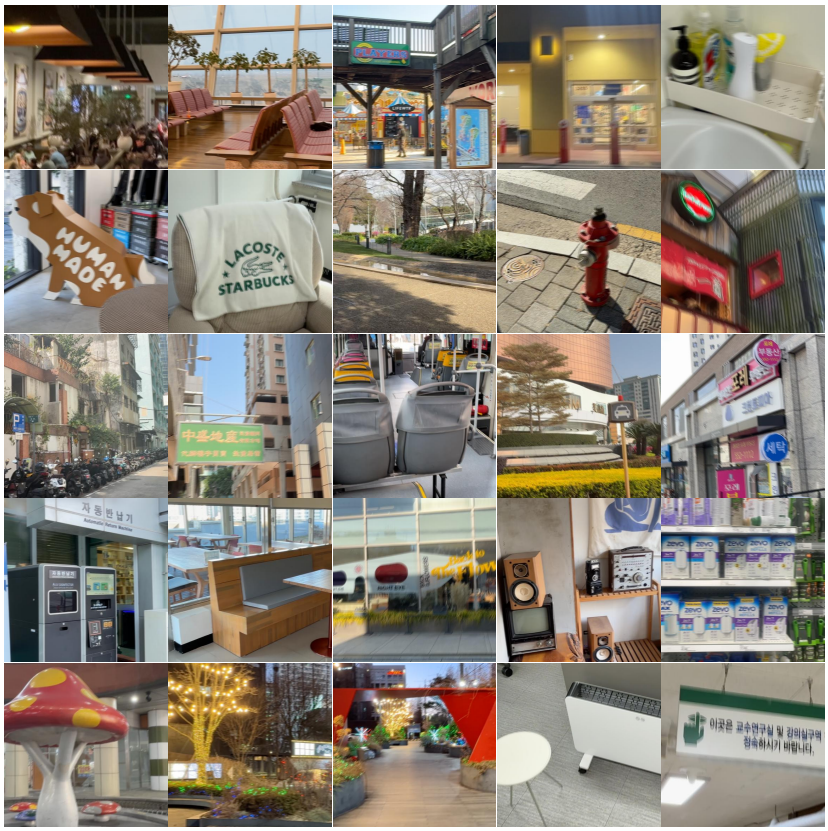


Fig. C: Representative blurry image patches from the HSD dataset, illustrating diverse scenarios such as indoor/outdoor scenes, varying lighting and blur levels, and different object densities and depth structures.

The HSD dataset covers a wide variety of scenes captured across multiple cities, including Seoul, Busan, Macau, Tokyo, and San Francisco, with diverse objects, environments, and motion patterns. Representative examples are shown in Fig. C.

B Analysis under Occlusion and Depth Variation

We analyze PECA under occlusion and depth variation using XYDeblur as a representative backbone. We estimate disparity with S^2M^2 [16] stereo matching on each rectified wide/ultra-wide pair and treat pixels with failed matching as occluded (hole) regions. Table A compares the single-view baseline, naive stereo concatenation, and PECA along three axes.

Region-wise. In non-occluded regions, PECA achieves a larger gain over the single-view baseline (+1.48 dB), indicating that it effectively exploits the reference view when correspondence is reliable. In occluded regions, the gain becomes smaller but remains positive (+0.20 dB), suggesting that PECA avoids severe degradation when reference cues are unreliable.

Hole ratio. Most images contain few hole pixels (mean 1.04%, median 0.57%, max 10.4%), so we construct the most-occluded Top 30% and Top 10% subsets. Naive concatenation degrades as occlusion severity increases and falls below the single-view baseline on the Top 10% subset (−0.10 dB). In contrast, PECA maintains positive gains across all subsets, including +0.32 dB on the Top 10% subset. This suggests that PECA is more robust than naive concatenation in highly occluded regions, where unreliable reference cues can otherwise degrade performance.

Relative depth. We group pixels by mean disparity into near, intermediate, and far ranges. The gain over naive concatenation is largest in the near range (+0.33 dB), suggesting that bounded directional search is particularly effective when disparity is large and correspondence is ambiguous. In the intermediate and far ranges, PECA remains comparable to naive concatenation, indicating that both stereo fusion strategies can effectively exploit the reference view when correspondence is relatively reliable.

Overall, PECA exploits the reference view when reliable correspondence is available and remains stable when cross-view matching becomes unreliable. These results support that bounded epipolar attention and residual fusion suppress unreliable cross-view aggregation without requiring explicit occlusion masks.

Table A: PSNR comparison on the XYDeblur backbone for occluded (OCC) and non-occluded (Non-OCC) regions, hole-severity subsets, and relative-depth subsets.

	Region-wise		Hole ratio			Relative depth		
Method	OCC	Non-OCC	Full	Top 30%	Top 10%	Near	Intermediate	Far
Single baseline	29.74	30.80	30.76	29.55	28.03	30.89	31.93	31.34
Stereo concatenation	29.78	32.02	31.96	29.86	27.93	31.68	33.28	**32.94**
PECA (ours)	**29.94**	**32.28**	**32.22**	**30.53**	**28.35**	**32.01**	**33.31**	32.93

C Qualitative Comparison of PECA and Model Variants

Figs. D-F present qualitative comparisons across three backbones, supporting the quantitative results in Table 1.

Stereo information improves restoration quality. As shown in Fig. D, single-view settings (baseline and channel-expansion) are inherently limited under severe motion blur, where complex spatial layouts around the vehicle remain difficult to recover faithfully. In contrast, stereo-view settings (stereo concatenation and PECA) leverage complementary cross-view information, enabling clearer reconstruction of object boundaries and overlapping structures.

Naive stereo fusion provides limited improvement. While naive stereo concatenation can outperform single-view settings, its ability to exploit cross-view information remains limited. In Fig. E, stereo concatenation partially improves the restoration quality compared with single-view variants, yet still fails to recover fine structures. In contrast, PECA produces noticeably clearer results by establishing more reliable cross-view correspondences.

Naive stereo fusion can even degrade performance. However, the effectiveness of stereo information depends strongly on the interaction mechanism. As illustrated in Fig. F, naive stereo concatenation does not consistently outperform single-view variants and may even yield inferior results compared to channel-expansion. In contrast, PECA maintains robust restoration performance across all backbones. By constraining the correspondence search space to a geometrically valid disparity range, PECA suppresses ambiguous matches and enables more accurate cross-view feature aggregation.

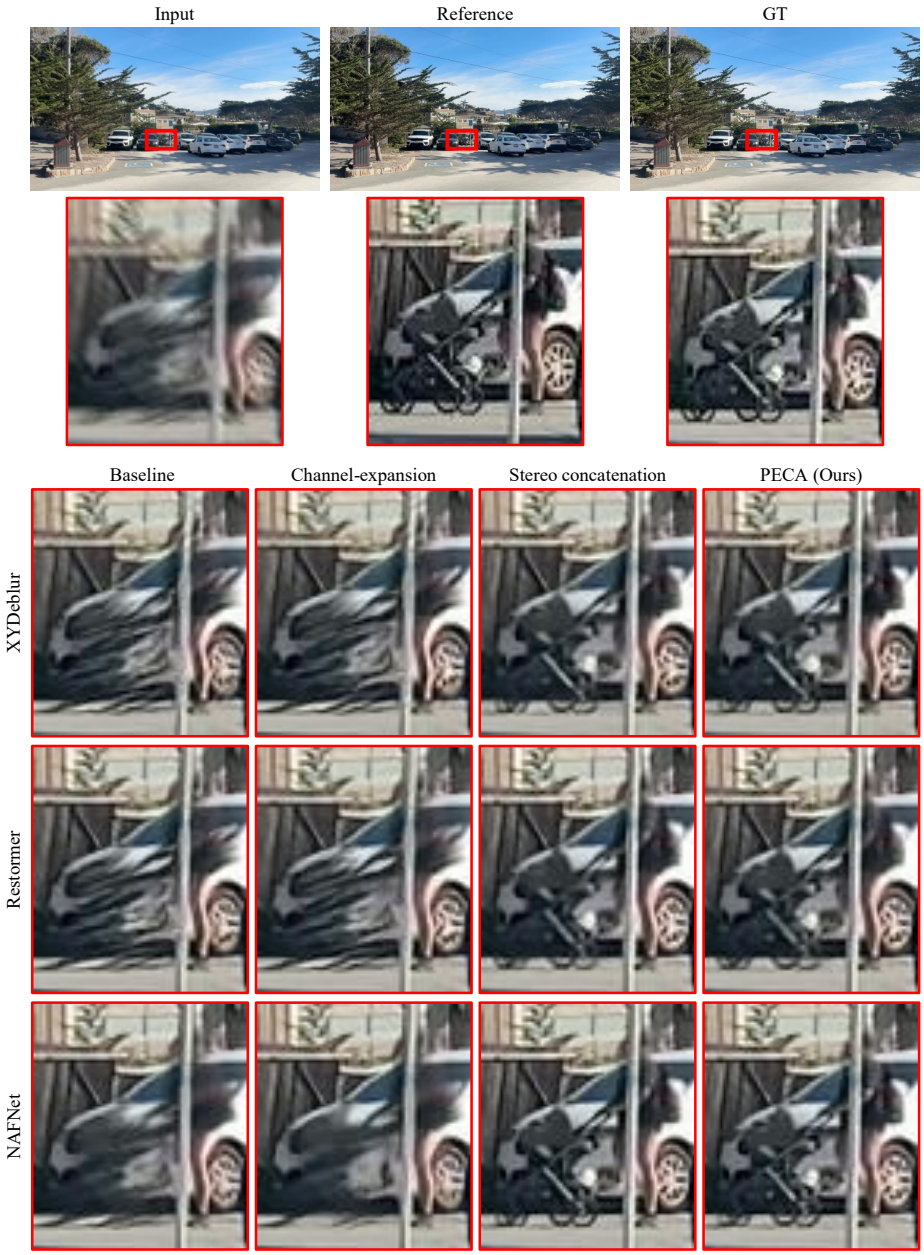


Fig. D: Qualitative results on the HSD dataset. A representative region is zoomed in to highlight large-scale motion blur and transparent structures around the vehicle.

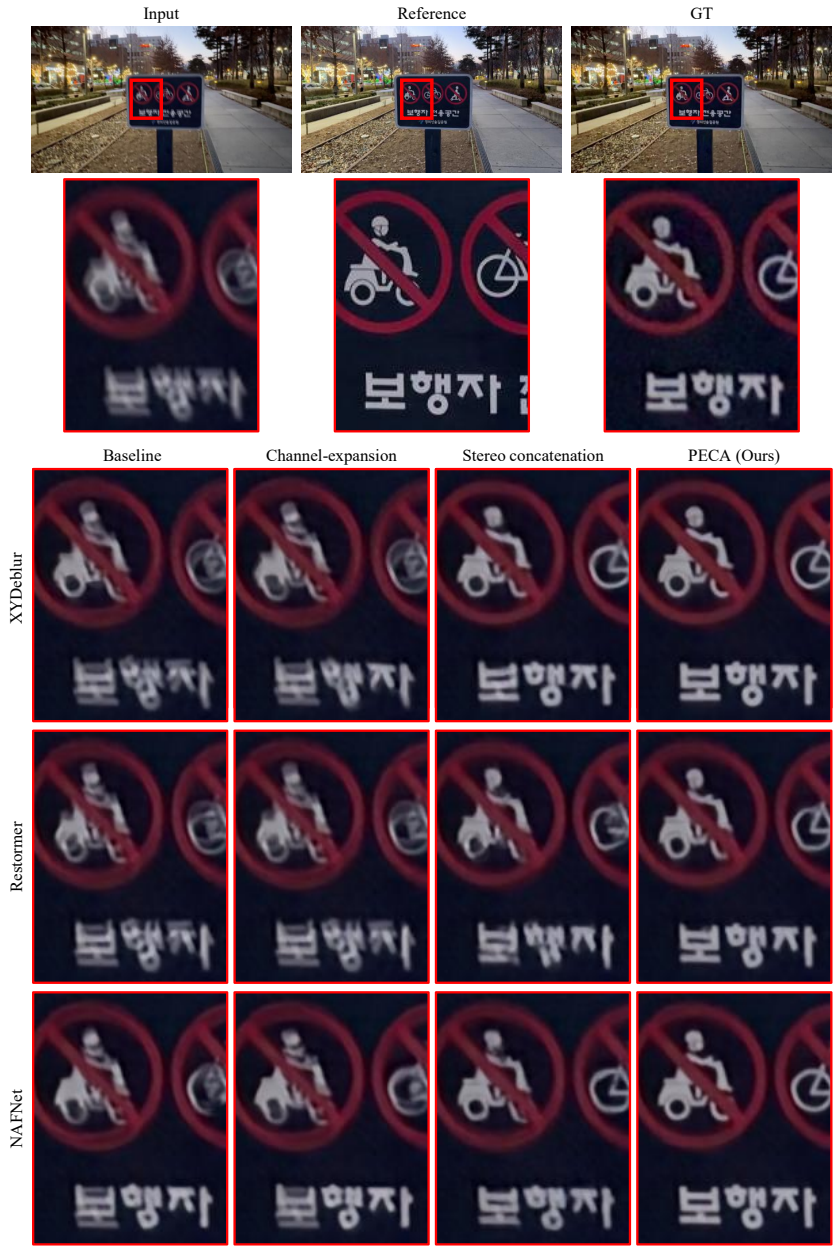


Fig. E: Qualitative results on the HSD dataset. Two regions are zoomed in for comparison, highlighting fine text and layered paper textures.

Fig. F: Qualitative results on the HSD dataset. A representative region is zoomed in to highlight fine-grained text on the box.

D Comparison with Recent Stereo Fusion Methods

We conduct a controlled comparison by replacing only the stereo fusion module while keeping the backbone fixed. As shown in Table B, PECA consistently outperforms stereo fusion modules such as SPAM [32] and ALM [29] across all three backbones. ALM incurs high computational cost due to dense alignment and falls below stereo concatenation on XYDeblur and Restormer, suggesting that dense matching can be less suitable for asymmetric deblurring. SPAM restricts matching to epipolar scanlines but lacks an explicit physical disparity bound. In contrast, PECA enforces both the disparity direction and a bounded search range, reducing ambiguous candidates under asymmetric blur. As a result, PECA achieves the best restoration accuracy with favorable module complexity.

Table B: Quantitative comparison with stereo fusion modules

	XYDeblur [6]		Restormer [31]		NAFNet [1]		Module	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	MACs(G)	Params(K)
Stereo concatenation	31.95	0.9590	31.83	0.9581	32.23	0.9601	-	-
SPAM [32]	32.02	0.9601	32.04	0.9580	32.24	0.9612	11.3	188.6
ALM [29]	30.81	0.9448	31.07	0.9460	32.05	0.9574	426.6	32.9
PECA (ours)	**32.22**	**0.9622**	**32.47**	**0.9636**	**32.92**	**0.9669**	**4.8**	**83.1**

E Qualitative Cross-device Evaluation

To further assess the cross-device generalization of PECA, we provide additional qualitative results on real-blur examples captured by an unseen Galaxy smartphone. Fig. G shows that PECA yields clearer restoration than the baseline without PECA. Specifically, PECA better preserves the boundary and improves the legibility of text details in the first example, and recovers sharper object structures around the bicycles in the second example. While a comprehensive quantitative cross-device evaluation remains future work, these results suggest that PECA generalizes to real blur from unseen devices.

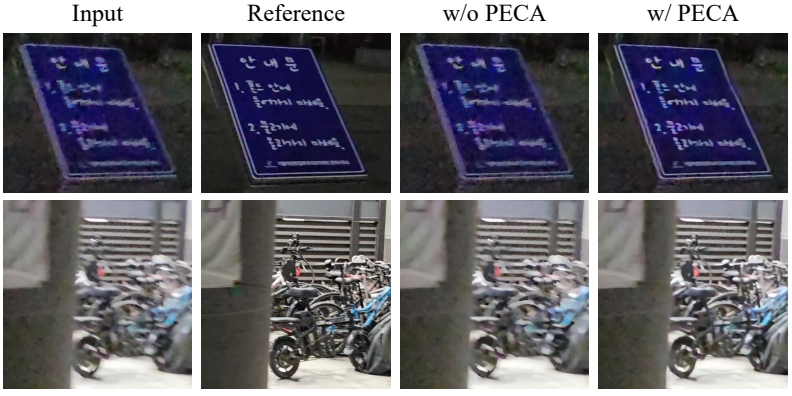


Fig. G: Qualitative results on real blur captured by Galaxy S25.

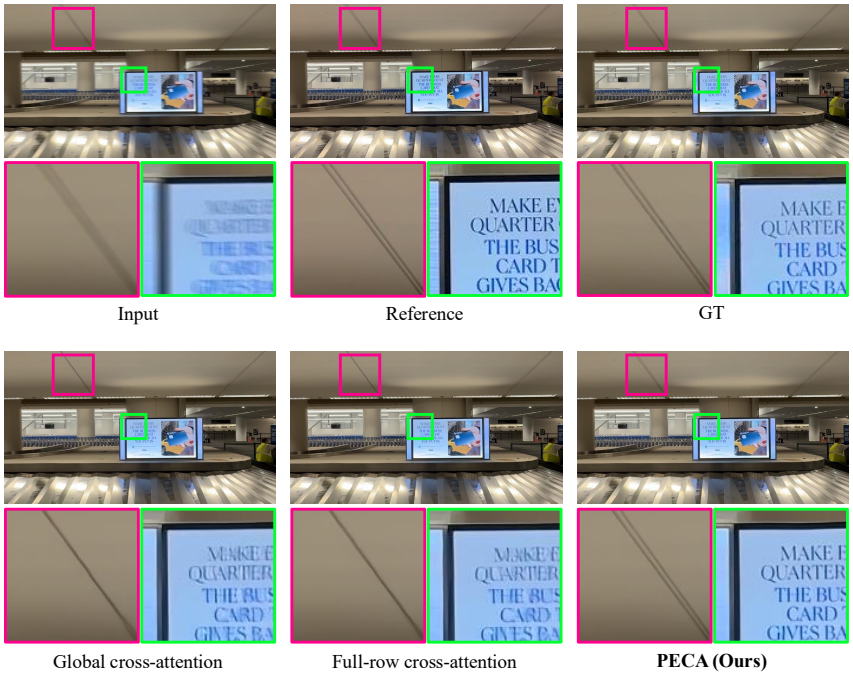


Fig. H: Qualitative comparison of global cross attention, full-row cross attention, and PECA. Two representative regions are magnified for comparison, highlighting the ceiling edge and the fine text on the signboard.

F Effect of Attention Search Ranges

Fig. H presents qualitative results corresponding to the quantitative comparison in Table 2. The highlighted regions illustrate challenging cases where conventional cross attention mechanisms fail to restore fine details under severe blur. In particular, the results show that an excessively wide correspondence search range can degrade feature aggregation in different ways. For example, it may dilute a plausible dominant match by mixing weak responses from other locations, or it may induce multi-modal matching ambiguity when repetitive fine structures produce several competing candidates.

In the red-highlighted region of Fig. H, the input contains two thin parallel line structures. Both global attention and full-row attention collapse these two lines into a single thicker edge. Under severe blur, the responses of the two lines become broadened and less sharply localized. When the search range is too wide, the number of weakly relevant or non-corresponding candidates increases substantially. Although each such location may receive only a small attention weight, their accumulated contribution becomes non-negligible during value aggregation and dilutes the dominant correspondence peak. As a result, the matched value feature no longer dominates the aggregation, and the reconstructed feature becomes less selective and more averaged. This over-smoothed

aggregation suppresses the high-frequency separation between the two lines and biases the reconstruction toward lower-frequency content, causing the two thin lines to merge into a thicker structure. In contrast, PECA constrains the correspondence search to a geometrically valid one-sided disparity interval that not only restricts the overall search range but also reduces the candidate region over which blur-broadened responses can spread. This decreases the number of irrelevant candidates that contribute to peak dilution and better preserves the separation between the two lines.

In the green-highlighted region, a different failure mode appears in fine textual patterns. Here, severe blur attenuates high-frequency stroke details, while repeated and visually similar character strokes generate multiple plausible correspondence candidates. With a wide search range, attention becomes multi-modal and aggregates features from several competing locations rather than selecting a single reliable match. This ambiguity mixes distinct stroke patterns, leading to distorted characters and reduced readability. By constraining the correspondence search space, PECA suppresses these ambiguous responses and enables more stable correspondence selection, resulting in clearer stroke boundaries.

These qualitative observations are consistent with the quantitative improvements reported in Table 2. These results demonstrate that constraining the correspondence search space is critical for reliable cross-view deblurring under asymmetric blur conditions.

G Analysis of Cross-view Correspondence in PECA

We analyze the difference between PECA and prior 1D epipolar attention mechanisms. Existing full-row epipolar attention methods compute correspondence over the entire scanline, which may include implausible candidates. In contrast, PECA jointly enforces the rectified-stereo direction and a bounded disparity range.

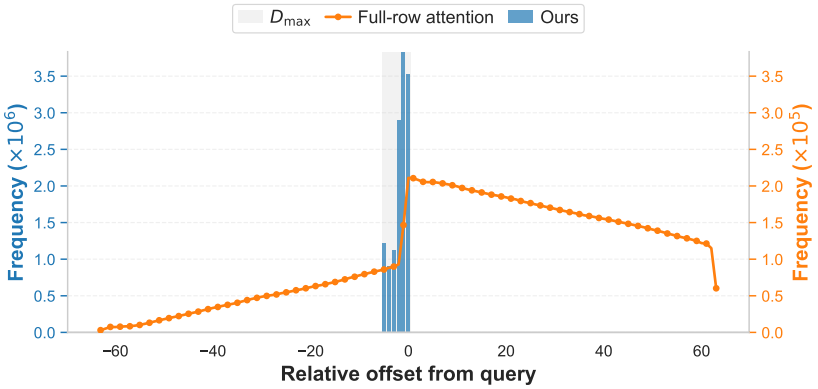


Fig. I: Top-3 attention match distribution along the epipolar line.

Fig. I visualizes the distribution of the top-3 attention positions relative to each query. Full-row attention spreads attention across the entire scanline, including large positive and negative offsets, whereas PECA concentrates its attention within the bounded disparity range. This result explains the performance gap between full-row attention and PECA.

H Qualitative Comparison on $D_{\max}$ and τ

Figs. J-L present qualitative comparisons for different epipolar window sizes $D_{\max}$ and softmax temperatures $\tau \in \{1.0, 0.1, 0.01, 0.001\}$ on the HSD dataset using the XYDeblur backbone. Two zoomed regions highlight challenging cases, including large-scale motion blur and fine-grained text.

Increasing $D_{\max}$ enlarges the search space along the epipolar direction. Although a larger window allows PECA to consider a broader set of candidates, an excessively large $D_{\max}$ introduces redundant or ambiguous correspondences that can degrade the matching reliability. In such cases, irrelevant features from distant locations are less effectively suppressed, distracting attention from the true correspondence and leading to degraded restoration quality. Notably, when $D_{\max}$ becomes large (*e.g.*, $D_{\max} \geq 12$), the foreground railings in front of the vehicle gradually disappear or merge with the background, as the enlarged search space introduces incorrect feature aggregation.

When the temperature is large (*e.g.*, τ=1.0), the attention distribution becomes overly smooth, causing the attention weights to spread across multiple candidates rather than concentrate on reliable correspondences. As a result, biased matches may arise from several locations along the scanline, which weakens the effectiveness of PECA and leads to blurry restorations. In practice, τ=1.0 yields inferior visual quality across all tested values of $D_{\max}$, producing blurrier results than smaller temperature values. This effect is particularly noticeable in structured regions such as the vehicle body and the text on the signboard.

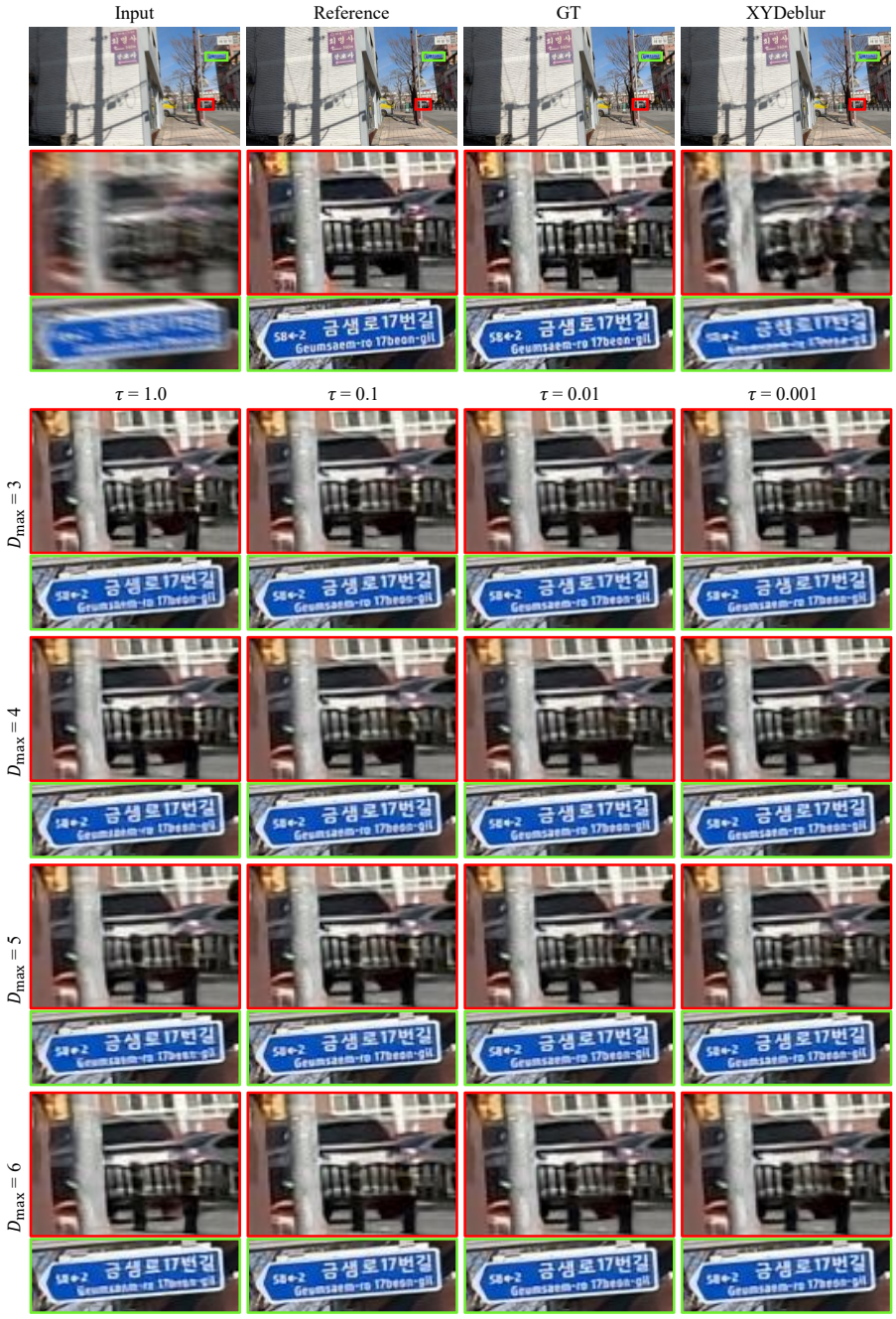


Fig. J: Qualitative results on the HSD dataset for $D_{\max} = 3$–6 under different softmax temperatures τ.

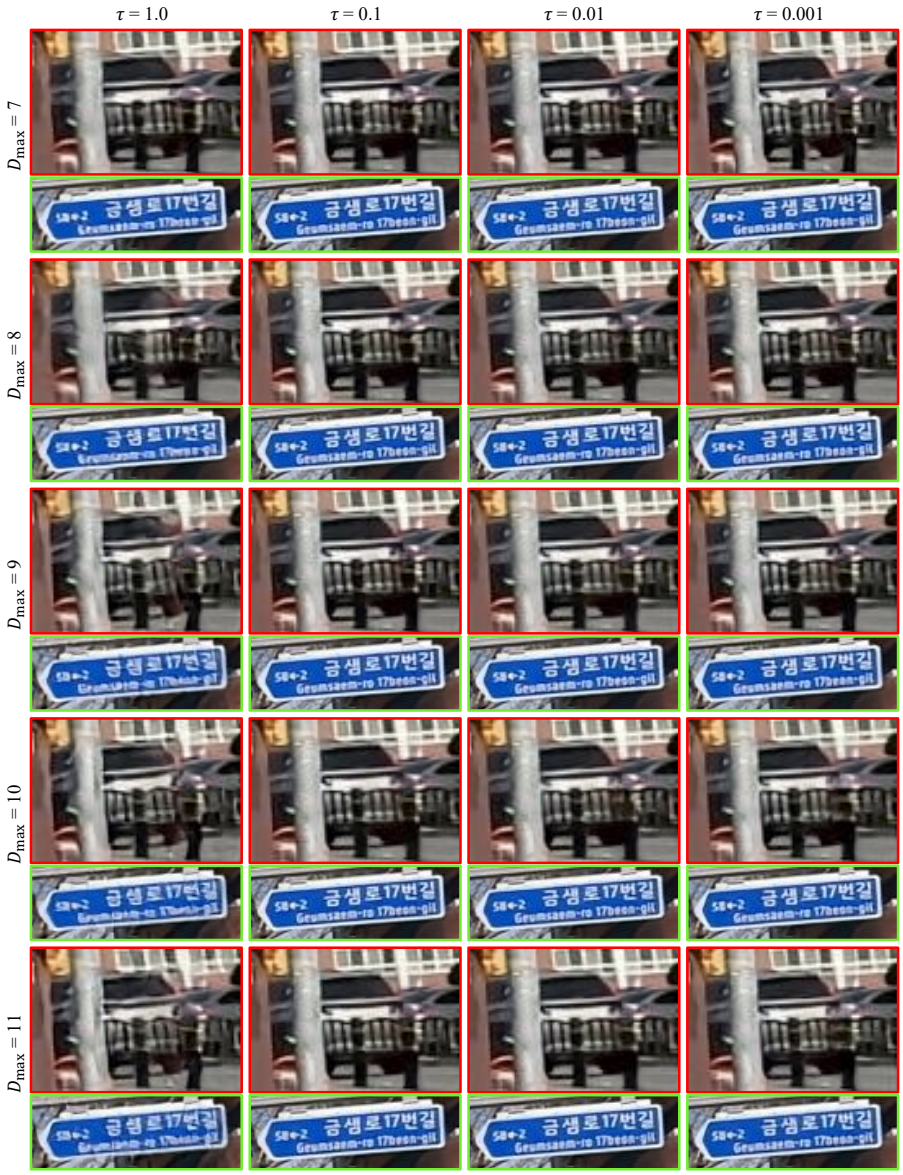


Fig. K: Qualitative results on the HSD dataset for $D_{\max} = 7$–11 under different softmax temperatures τ.

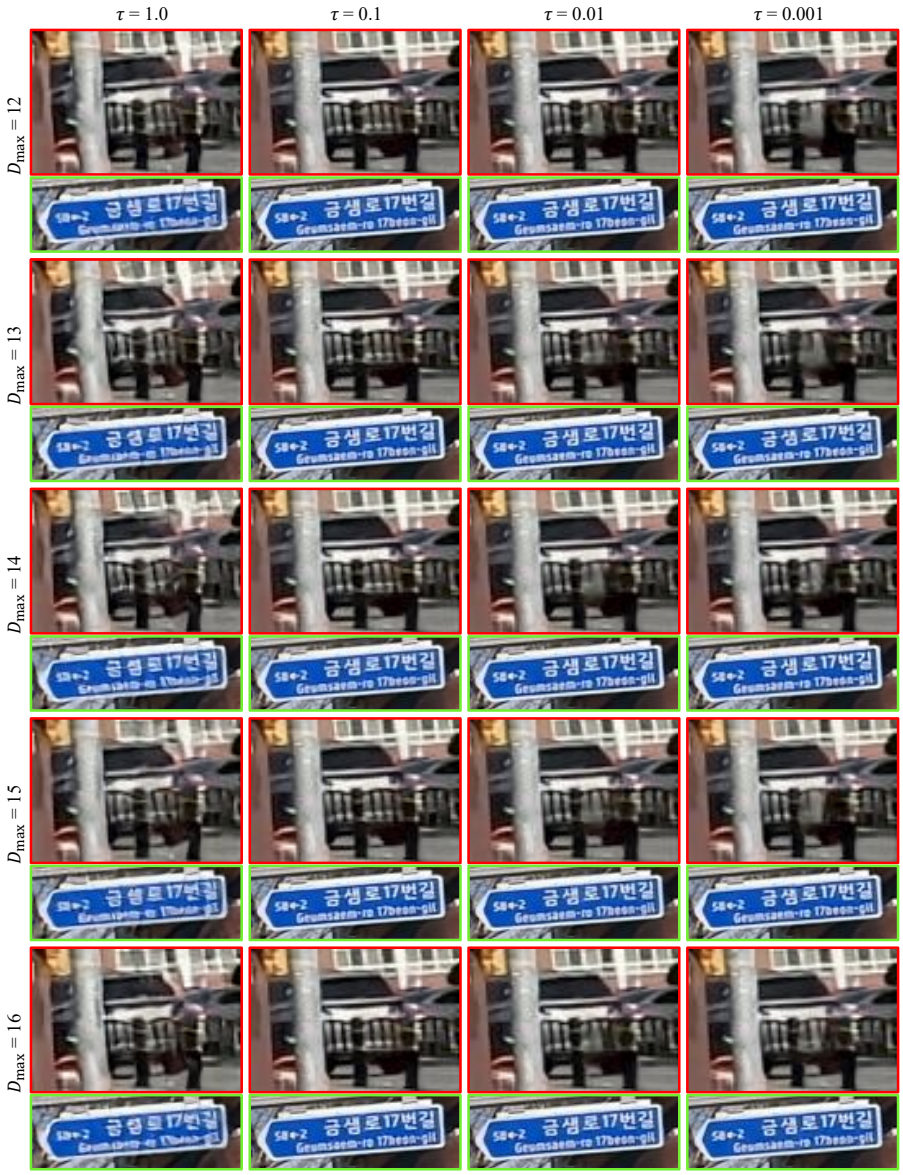


Fig. L: Qualitative results on the HSD dataset for $D_{\max} = 12$–16 under different softmax temperatures τ.

Table C: PSNR comparison on different difficulty subsets of the test set.

Method	PSNR ↑					
	Easy	(Δ)	Moderate	(Δ)	Hard	(Δ)
XYDeblur [6]	34.38	-	31.51	-	25.65	-
+ Global cross attention	34.37	(−0.01)	31.45	(−0.06)	25.62	(−0.03)
+ Full-row cross attention	34.40	(+0.02)	31.49	(−0.02)	25.62	(−0.03)
+ Stereo concatenation	34.80	(+0.42)	32.73	(+1.22)	27.58	(+1.93)
+ **PECA (Ours)**	**34.88**	(+0.50)	**32.98**	(+1.47)	**28.03**	(+2.38)
Restormer [31]	34.73	-	31.68	-	25.66	-
+ Stereo concatenation	34.98	(+0.25)	32.48	(+0.80)	27.37	(+1.71)
+ **PECA (Ours)**	**35.45**	(+0.72)	**33.08**	(+1.40)	**28.26**	(+2.60)
NAFNet [1]	35.22	-	32.88	-	27.66	-
+ Stereo concatenation	35.23	(+0.01)	33.08	(+0.20)	27.23	(−0.43)
+ **PECA (Ours)**	**35.36**	(+0.14)	**33.66**	(+0.78)	**28.99**	(+1.33)

Table D: SSIM comparison on different difficulty subsets of the test set.

Method	SSIM ↑					
	Easy	(Δ)	Moderate	(Δ)	Hard	(Δ)
XYDeblur [6]	0.9811	-	0.9626	-	0.8714	-
+ Global cross attention	0.9810	(−0.0001)	0.9622	(−0.0004)	0.8702	(−0.0012)
+ Full-row cross attention	0.9809	(−0.0002)	0.9620	(−0.0006)	0.8702	(−0.0012)
+ Stereo concatenation	0.9829	(+0.0018)	0.9717	(+0.0091)	0.9118	(+0.0404)
+ **PECA (Ours)**	**0.9830**	(+0.0019)	**0.9729**	(+0.0103)	**0.9198**	(+0.0484)
Restormer [31]	0.9817	-	0.9625	-	0.8709	-
+ Stereo concatenation	0.9826	(+0.0009)	0.9688	(+0.0063)	0.9119	(+0.0410)
+ **PECA (Ours)**	**0.9838**	(+0.0021)	**0.9723**	(+0.0098)	**0.9260**	(+0.0551)
NAFNet [1]	0.9834	-	0.9702	-	0.9077	-
+ Stereo concatenation	0.9837	(+0.0003)	0.9725	(+0.0023)	0.9125	(+0.0048)
+ **PECA (Ours)**	**0.9841**	(+0.0007)	**0.9753**	(+0.0051)	**0.9327**	(+0.0250)

I Performance Analysis on Test Subsets

To evaluate performance across different levels of degradation, we partition the test set into three difficulty-based subsets according to the PSNR between blurry and sharp image pairs. Specifically, samples in the bottom 25% (PSNR < 24.4 dB) are categorized as Hard, those in the middle 50% (24.4–30.3 dB) as Moderate, and those in the top 25% (PSNR > 30.3 dB) as Easy. These subsets contain 275, 550, and 275 image pairs, respectively.

Table C, D report quantitative results on the difficulty-based subsets. PECA consistently outperforms all compared variants across all subsets in terms of both PSNR and SSIM. Notably, the performance gains become more pronounced on the Hard subset, where severe blur and complex scene structures make correspondence estimation more challenging. Compared with stereo concatenation,

PECA improves PSNR on the Hard subset by 0.45 dB for XYDeblur, 0.89 dB for Restormer, and 1.76 dB for NAFNet. Similar trends can be observed in SSIM, with clear improvements on the most challenging samples across all three backbones. These results indicate that PECA can effectively exploit cross-view information under severe degradation, where naive stereo fusion or unconstrained matching becomes less reliable. Moreover, the results on XYDeblur show that global or full-row cross attention may even underperform the single-view baseline, further highlighting the importance of constraining the correspondence search space.