# Asymmetric Dual-Lens Video Deblurring

**Zeyu Xiao** **Xinchao Wang***
National University of Singapore
zeyuxiao@nus.edu.sg, xinchao@nus.edu.sg

## Abstract

Modern smartphones often feature asymmetric dual-lens systems, capturing wide-angle and ultra-wide views with complementary perspectives and details. Motion and shake can blur the wide lens, while the ultra-wide lens, despite lower resolution, retains sharper details. This natural complementarity offers valuable cues for video deblurring. However, existing methods focus mainly on single-camera inputs or symmetric stereo pairs, neglecting the cross-lens redundancy in mobile dual-camera systems. In this paper, we propose a practical video deblurring method, AsLeD-Net, which recurrently aligns and propagates temporal reference features from ultra-wide views fused with features extracted from wide-angle blurry frames. AsLeD-Net consists of two key modules: the adaptive local matching (ALM) module, which refines blurry features using $K$-nearest neighbor reference features, and the difference compensation (DC) module, which ensures spatial consistency and reduces misalignment. Additionally, AsLeD-Net uses the reference-guided motion compensation (RMC) module for temporal alignment, further improving frame-to-frame consistency in the deblurring process. We validate the effectiveness of AsLeD-Net through extensive experiments, benchmarking it against potential solutions for asymmetric lens deblurring.

## 1 Introduction

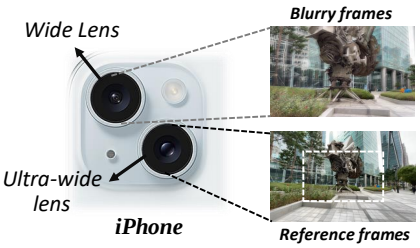


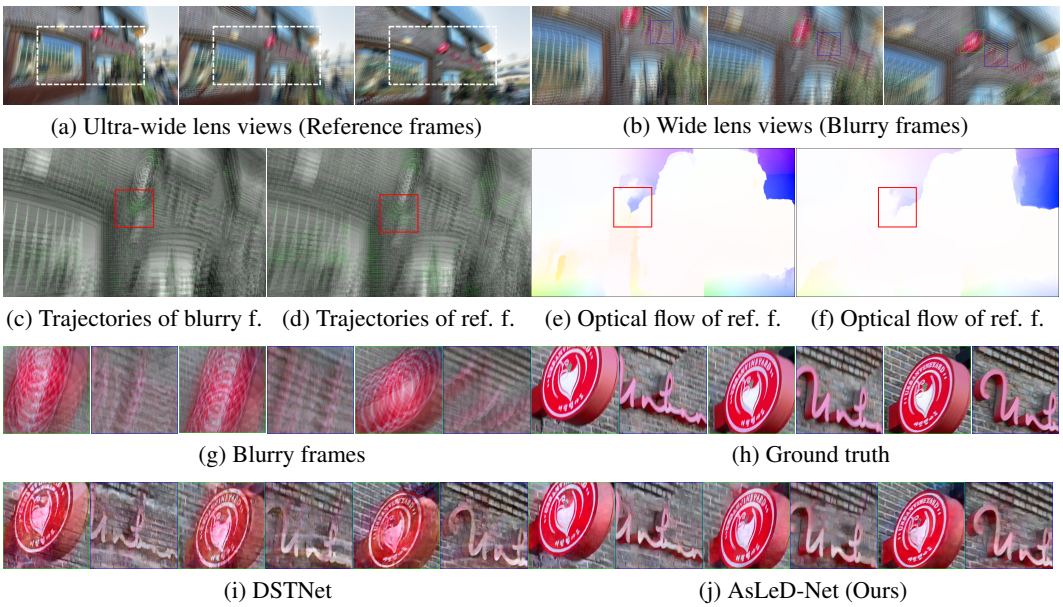

Figure 1: A typical dual-lens camera system on iPhone in practice.

Modern smartphones increasingly feature asymmetric dual-lens systems, combining wide-angle and ultra-wide lenses to capture scenes with complementary perspectives and details (see Figure 1). However, during dynamic scenes or camera shake, focal length and aperture differences introduce blur discrepancies. For example, on the iPhone 14 Plus, ultra-wide lenses typically have shorter focal lengths (*e.g.*, 13mm) and smaller apertures (*e.g.*, f/2.4) compared to wide-angle lenses (*e.g.*, 26mm, f/1.5). The shorter focal length of ultra-wide lenses results in a deeper depth of field, reducing defocus blur but making them more prone to motion blur due to the longer exposure time required for sufficient light intake. In contrast, wide-angle lenses with a larger aperture capture more light, allowing for faster shutter speeds and reducing motion blur while maintaining a shallower depth of field that enhances subject-background separation. These complementary characteristics provide valuable cues for video deblurring, as shown in Figure 2.

Existing video deblurring methods primarily focus on single-camera inputs or symmetric stereo pairs. For instance, flow-guided bidirectional propagation methods [8, 21, 40] align features using deformable convolutions and attention mechanisms, while VRT and RVRT [37, 38] employ spatio-temporal self-attention to aggregate information across video frames. However, these methods do not leverage the cross-lens redundancy inherent in mobile dual-camera systems. Although several

*Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Ultra-wide lens views (Reference frames) (b) Wide lens views (Blurry frames)

(c) Trajectories of blurry f. (d) Trajectories of ref. f. (e) Optical flow of ref. f. (f) Optical flow of ref. f.

(g) Blurry frames (h) Ground truth

(i) DSTNet (j) AsLeD-Net (Ours)

Figure 2: Examples of asymmetric dual-lens video deblurring. (a) Reference frames captured using the ultra-wide lens. (b) Blurry frames captured using the wide-angle lens. (c-d) Motion discrepancy visualization through trajectory comparisons (Please zoom in and best viewed on the screen). (e-f) Optical flow visualization of blurry and reference frames. (g) Input blurry sequences. (h) Ground truth. (i) Results from a state-of-the-art method (DSTNet) [52]. (j) Our AsLeD-Net results. In this figure, "f." is short for frames, and "ref." is short for reference, respectively. We propose the *first* deep learning method for *asymmetric dual-lens video deblurring*, where ultra-wide lens videos (shorter focal length) provide motion priors to restore clear wide-lens blurry sequences. This is motivated by the observation that ultra-wide lenses inherently suffer *less motion blur* due to their focal properties: shorter focal lengths reduce perceived motion parallax, enabling more stable scene capture despite the lower resolution. AsLeD-Net delivers high deblurring accuracy with strong temporal coherence.

image restoration techniques have been developed for dual-camera setups, including dual-lens super-resolution [70, 94, 31, 84, 28, 69, 7, 76, 85, 95, 53] and dual-lens stereo matching [12, 64, 90], as well as reference-based image super-resolution [82, 62, 44, 6, 93, 56, 23, 24, 20, 29, 92, 25], video deblurring in an *asymmetric dual-lens* setting remains unexplored. In this paper, we introduce Asymmetric dual-Lens video Deblurring (AsLeD), which exploits the complementary information between blurry and reference videos captured from asymmetric dual-camera systems. Unlike traditional video deblurring approaches, AsLeD explicitly considers the structural disparities between the two views to enhance the restoration process.

AsLeD effectively models the relationship between blurry and reference frames in asymmetric dual-lens settings. At each time step, the two frames share nearly identical content within their overlapping field of view (FoV) (top and middle rows of the leftmost column in Figure 2). As the video progresses, neighboring reference frames provide sharp details that help recover regions beyond the overlapped FoV (bottom row of the leftmost column in Figure 2). By leveraging these cross-frame correspondences, AsLeD enhances video restoration by utilizing multi-frame redundancy in asymmetric multi-camera setups.

AsLeD-Net employs a dual-branch architecture built upon the BasicVSR framework to exploit asymmetric dual-lens characteristics. The base branch uses a bidirectional recurrent design [8, 21] for temporal feature propagation, while the reference branch processes ultra-wide lens inputs with reduced motion blur. This asymmetric design incorporates three key modules in the reference branch: (1) The adaptive local matching (ALM) module: Establishes semantic-aware correspondences between blurry frames (base branch) and clear reference frames (ultra-wide branch) using $K$-nearest neighbor feature aggregation, effectively transferring structural details. (2) The difference compensation (DC) module: Bridges feature discrepancies between the branches, preserving edge sharpness in fast-moving regions and enhancing spatial-temporal consistency. (3) The reference-guided motion compensation (RMC) module: Aligns features across time steps using optical flow guided by ultra-wide reference frames, eliminating cumulative errors in traditional flow-based warping. The base branch progressively

fuses ALM-refined features, DC-compensated residuals, and RMC-aligned references via attention-guided fusion blocks. This hierarchical integration enables the model to leverage the base branch's reconstruction capabilities and the reference branch's blur-insensitive motion priors. As shown in Figure 2, AsLeD-Net outperforms DSTNet [52], enhancing texture details like the text on the wall and the coffee shop logo while preserving strong temporal coherence.

The contributions of this work are summarized as follows: (1) We formalize asymmetric dual-lens video deblurring (AsLeD), leveraging cross-lens redundancy from a sharper ultra-wide reference to assist a blurry wide view, extending prior dual-lens *image* deblurring to the video setting. (2) We present AsLeD-Net with three task-specific modules: ALM for structure-aware K-nearest-neighbor reference aggregation, DC for cross-view spatial consistency, and RMC for temporal alignment. (3) AsLeD-Net achieves superior quantitative and qualitative results on AsLeD.

## 2 Related Works

**Video deblurring.** Unlike image deblurring [48, 67, 87, 79, 77, 35], video deblurring benefits from spatio-temporal cues to enhance restoration quality [51, 33, 99, 98, 88, 72, 66, 80, 10, 21, 22, 49, 65, 81, 37, 58, 89]. Deep learning-based methods [3, 2, 97, 96, 54, 55, 36, 78] have become the dominant approach. Recurrent video deblurring models exploit temporal dependencies for progressive feature propagation [98, 91, 72, 10, 39, 49, 52, 32]. For example, STRCNN [27], RDN [74], and IFRNN [49] adopt recurrent architectures for sequential feature refinement. STFAN [100] employs dynamic filters, while PVDNet [63] integrates a blur-invariant flow estimator. Recent methods leverage bidirectional propagation for improved restoration [8, 101, 21, 40, 88]. BasicVSR++ [8] introduces aggressive bidirectional propagation, while RNN-MBP [101] incorporates multi-scale bidirectional updates. However, error accumulation remains challenging for long-range temporal modeling [21]. Spatio-temporal transformers enhance video deblurring by capturing long-range dependencies [37, 40, 39, 37, 38]. Recently, Zhang *et al.* [89] propose a spatio-temporal sparse Transformer for efficient video deblurring.

**Dual-lens image restoration.** Asymmetric dual-lens systems, commonly found in smartphones, consist of an ultra-wide lens and a wide lens with different focal lengths. These systems capture the same scene with varying FoVs. Typically, the ultra-wide lens (short focal length, large FoV) is the main lens, while the wide lens (longer focal length, narrower FoV) provides higher resolution within the overlapped FoV. This configuration enables various vision tasks by leveraging the complementary imaging capabilities of the two lenses. Previous works have explored dual-lens systems for image refocusing [1], correspondence estimation [12, 64, 90] and image super-resolution [70, 94, 31, 84, 28, 69, 7, 76, 85, 95, 53], where telephoto images serve as high-resolution references to enhance wide-angle images. Beyond image super-resolution, other tasks such as image deblurring [46, 61, 59], novel view synthesis [75, 83, 45], high-dynamic-range imaging [34, 60], and image colorization [15] have also been investigated, demonstrating the practical benefits of cross-lens redundancy. Among prior studies, the most similar to ours are Mohan *et al.* [46], who deblur static dual-lens image pairs by enforcing cross-view and depth consistency in unconstrained capture settings. Lai *et al.* [30] deblur face regions in still photos using a synchronized sharp ultra-wide reference. We extend this line from images to video by integrating reference-guided cross-view alignment with recurrent temporal propagation to preserve frame-to-frame coherence.

**Reference-based image restoration.** Another related topic is reference-based image restoration, which includes tasks such as reference-based image super-resolution [82, 62, 44, 6, 93, 56, 23, 24, 20, 29, 92, 25], image deblurring [41, 42] and burst image restoration [14, 19, 4, 26, 16, 5, 17, 73, 18, 43, 29]. These tasks leverage auxiliary high-quality images to enhance the restoration of a degraded input. In this work, we address reference-based video deblurring in the context of asymmetric dual-lens smartphone cameras. Unlike existing approaches focusing on still images, our method utilizes cross-lens redundancy in videos to achieve high-quality deblurring, effectively integrating information from both lenses.

## 3 Method

### 3.1 Overview

The proposed AsLeD-Net reconstructs a high-quality video $\hat{\mathcal{I}} \in \mathbb{R}^{T\times H\times W\times 3}$ from a blurry input $\mathcal{I}^B \in \mathbb{R}^{T\times H\times W\times 3}$ captured with a wide lens, leveraging a relatively sharper reference video

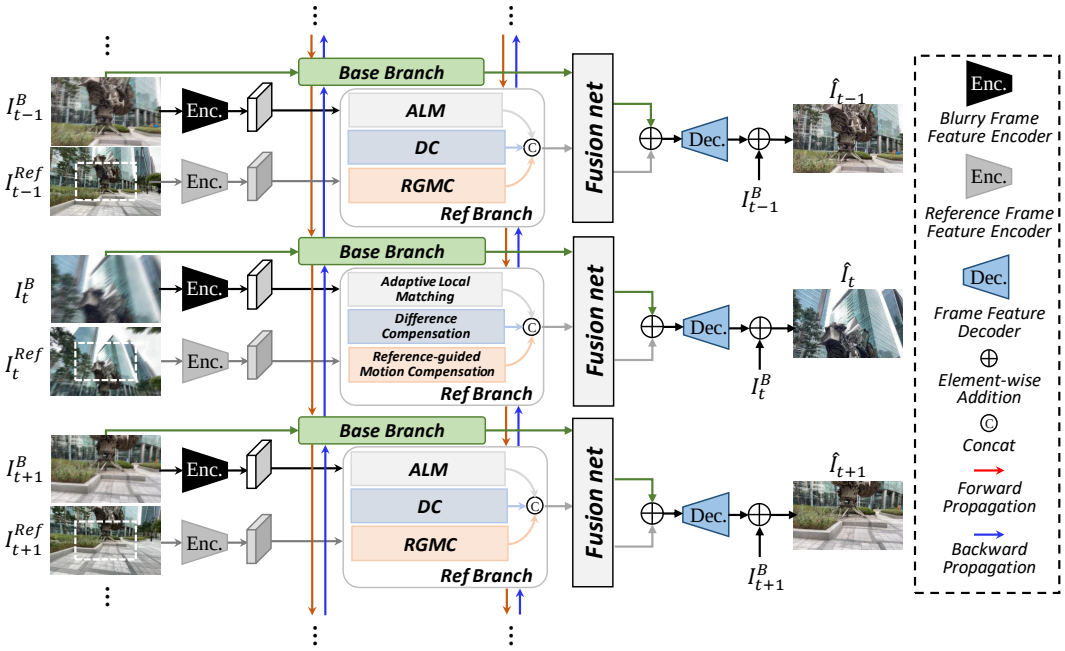



Figure 3: Overview of the proposed AsLeD-Net. AsLeD-Net processes blurry video frames from a wide lens and reference frames from an ultra-wide lens using a dual-branch architecture. The base branch reconstructs the blurry frames with bidirectional frame propagation [9], while the reference branch extracts asymmetric information from the ultra-wide frames. Features from both branches are encoded separately and then refined by ALM, DC, and RMC modules. The concatenated features are merged through residual blocks and fed to the frame decoder to generate the final clear output.

$\mathcal{I}^{Ref} \in \mathbb{R}^{T\times H\times W\times 3}$ taken simultaneously by an ultra-wide len in a dual-lens camera (*e.g.*, iPhone). The objective is to restore $\hat{\mathcal{I}}$ to closely match the ground truth $\mathcal{I}^{GT}$, effectively exploiting the complementary information from the reference view. Here, $T$, $H$, and $W$ denote the number of frames, height, and width, respectively. Following previous video deblurring and restoration methods [8, 9], we adopt a bi-directional recurrent architecture for its simplicity and efficiency. The overall framework of our proposed AsLeD-Net is shown in Figure 3.

The base branch, identical to BasicVSR, reconstructs blurry videos $\mathcal{I}^B$ captured with a wide lens, using bidirectional frame propagation based on optical flow to extract features $\boldsymbol{F}^{Base}$. To exploit the unique properties of the AsLeD task, we introduce a second reference branch, which processes ultra-wide lens inputs with reduced motion blur. Since reference and blurry frames have different FoV and contain varying information, we center-crop and resize the reference frames to match the blurry frame size. We leverage the reference branch to extract asymmetric reference information despite potential colour discrepancies due to lens differences.

For the blurry frame $I_t^B$ and its corresponding reference frame $I_t^{Ref}$ at time $t$, we first feed them to separate feature encoders with non-shared weights to obtain feature representations $F_t^B$ and $F_t^{Ref}$, each consisting of $N_1$ residual blocks. These features are then processed by the ALM, DC, and RMC modules, which transfer structural details from the reference frames, bridge feature discrepancies while preserving edge sharpness in fast-moving regions, and align features across time steps using optical flow guided by the ultra-wide reference frames. After concatenation, the outputs $F_t^{ALM}$, $F_t^{DC}$, and $F_t^{RMC}$ are merged with the base branch's output $F_t^{Base}$ by feeding them to residual blocks, resulting in $F_t$. Finally, $F_t$ is input into a frame decoder, composed of $N_2$ residual blocks, to generate the final reconstructed result $\hat{I}_t$, with a residual connection from the input added.

Our approach does not assume perfect geometric calibration or strict cross-view alignment. Instead, AsLeD-Net learns content-aware feature alignment that tolerates cross-lens baselines and photometric shifts, enabling reference guidance under realistic dual-lens configurations. In practice, we retain cross-view geometric and photometric discrepancies and rely on content-aware, data-driven feature alignment, while ensuring frame-level temporal synchronization.

### 3.2 Adaptive Local Matching

Motion blur erases high-frequency details and disrupts spatial coherence, leading to texture ambiguity and structural distortions. Recovering these details is particularly challenging due to the ill-posed nature of the problem. To address this issue, we exploit the complementary information available in reference frames, which are captured under similar conditions but with less or no blur. These frames retain rich structural cues that can guide the restoration of the blurry frame, providing a more robust mechanism for handling local details. The core idea is to adaptively leverage the reference frame's features based on local context, enabling more accurate restoration of the blurred image.

Without loss of generality, we focus on the time step $t$ to illustrate the ALM module. Given the extracted blurry frame feature $F_t^B \in \mathbb{R}^{C\times H\times W}$ at timestamp $t$, the ALM module aims to refine these blurry local features by incorporating the complementary reference feature $F_t^{Ref} \in \mathbb{R}^{C\times H\times W}$ captured under similar conditions but with less blur. These deep features can be represented as a set of $H \times W$ local representations, each of size $C$-dimensional. To match the blurry frame with the reference features, we first compute a similarity map $S_t$ to measure the cosine similarity between each blurry local feature $F_t^{B(i)}$ and each reference feature $F_t^{Ref(j)}$

$$S_t^{(i,j)} = \cos\big(F_t^{B(i)}, F_t^{Ref(j)}\big) = \frac{{F_t^{B(i)}}^\top F_t^{Ref(j)}}{\|F_t^{B(i)}\| \cdot \|F_t^{Ref(j)}\|}, \tag{1}$$

where $i \in \{1,\dots,N_t^B\}$ and $j \in \{1,\dots,N_t^{Ref}\}$, and $N_t^B$ and $N_t^{Ref}$ represent the number of local representations in the blurry and reference frames, respectively. $S_t^{(i,j)}$ represents the similarity between the $i$-th blurry local representation and the $j$-th reference local representation.

For each blurry local representation, we select its $K$-nearest neighbors from the reference features and fuse these local representations into a reference representation $\phi_{t,ref}^{(i)}$

$$\phi_{t,ref}^{(i)} = \sum_{k=1}^{K} \alpha_t^{(k)} \times F_t^{Ref(k)}, \tag{2}$$

where $\sum_{k=1}^{K} \alpha_t^{(k)} = 1$, $\alpha_t^{(k)} > 0$, and $\alpha_t^{(k)}$ is obtained using the softmax function. The reference representations then adaptively refine the corresponding blurry features. Finally, the blurry and reference representations are concatenated and processed by an adaptive layer $g_\theta$ to obtain the fused local representations

$$F_t^{ALM} = g_\theta([(\phi_{t,ref}, F_t^B)]), \tag{3}$$

where $\phi_{t,fuse}$ represents the fused local representations, and $g_\theta$ denotes the adaptive fusion layer. The cosine similarities in Eq. (1) form a query-reference affinity used to rank reference tokens; the top-$K$ are aggregated with softmax-normalized weights. The adaptive layer $g_\theta$ in Eq. (3) is a lightweight $1\times1$ convolution applied to the concatenated $[\phi_{t,\text{ref}},\ F_t^B]$, enabling content-aware fusion with low overhead. This process facilitates information fusion by incorporating relevant reference features to enhance the blurry frame representation.

### 3.3 Difference Compensation

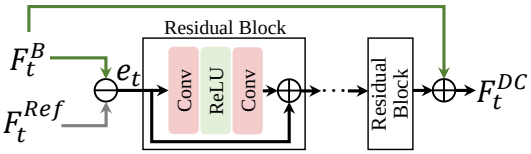


Figure 4: Overview of the difference compensation (DC) module.

In AsLeD, while blurry and reference images share structural consistency, focus, exposure, and sensor processing differences create subtle feature mismatches, making direct concatenation suboptimal. The DC module provides a simple yet effective solution by adaptively modeling residual information. It ensures that only beneficial details are aggregated while preserving spatial consistency. Unlike complex alignment-based approaches, the DC module efficiently refines blurry features with relevant reference information, preventing artifacts and enhancing deblurring with minimal computational overhead.

Given the blurry frame feature $F_t^B$ and its corresponding reference frame feature $F_t^{Ref}$ at time $t$, the goal of the DC module is to refine the blurry features by effectively leveraging reference information. As shown in Figure 4, we predict the difference between the blurry feature $F_t^B$ and the reference

feature $F_t^{Ref}$, enhancing the residual information $e_t$ to facilitate effective feature aggregation and adaptive refinement. The difference compensation between $F_t^{Ref}$ and $F_t^{Ref}$ can be summarized as

$$e_t = F_t^B - F_t^{Ref}, \quad \hat{e}_t = f(e_t), \quad F_t^{DC} = F_t^B + \hat{e}_t, \tag{4}$$

where $e_t$ represents the residual feature that captures useful differences between the blurry and reference features, and $f(\cdot)$ denotes the cascaded residual blocks. This progressive refinement adaptively integrates reference information while preserving spatial consistency, ensuring effective feature aggregation for improved deblurring.

### 3.4 Reference-Guided Motion Compensation

In the base branch of AsLeD-Net, optical flow is computed between the current and previous blurry frames for warping. While this approach has its merits, it often introduces limitations that degrade deblurring performance. Specifically, both the current and previous blurry frames may contain significant motion blur and noise, leading to inaccuracies in optical flow estimation. These imperfections can cause misalignment during warping, resulting in artifacts or an inability to capture fine motion details, ultimately reducing the quality of deblurring.

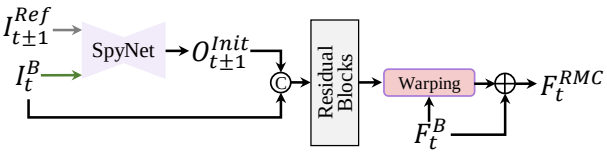


Figure 5: Overview of the reference-guided motion compensation (RMC) module.

To address these issues, we propose the RMC module, which estimates optical flow between the current blurry frame and a sharper reference frame (previous or next) rather than between consecutive blurry frames. It leverages two complementary motions: temporal flow within the base view and cross view flow guided by the sharper reference; the latter accounts for differences in field of view, perspective, and resolution, is more reliable under heavy blur, and reduces misalignment to improve motion tracking. As is shown in Figure 5, given the current blurry frame $I_t^B$ and the previous reference frame $I_{t-1}^{Ref}$, we first use SpyNet [57] to estimate the initial optical flow $O_{t-1}^{Init}$ between these two frames. However, this initial flow estimate may not be sufficiently accurate due to factors such as noise, motion blur, or misalignment, which can affect the precision of the optical flow calculation. To improve the flow estimate, we concatenate the current blurry frame $I_t^B$ and the initial flow estimate $O_{t-1}^{Init}$, then feed this concatenated pair to residual blocks to refine the flow. This refined flow is expected to capture fine details and correct the misalignment introduced by the initial estimation

$$O_{t-1}^{Init} = \text{SpyNet}(I_t^B, I_{t-1}^{Ref}), \quad O_{t-1}^{Refined} = f\left([I_t^B, O_{t-1}^{Init}]\right). \tag{5}$$

The refined flow is not directly supervised; it is learned implicitly via end-to-end reconstruction loss, in line with prior video restoration designs where flow acts as an auxiliary representation to facilitate motion compensation rather than a target itself.

After refining the optical flow, we perform warping using the refined flow and the current blurry frame $I_t^B$. The result is then combined with the residual information to obtain the motion-compensated feature $F_t^{RMC}$, which is more accurate in capturing the motion and structural details of the scene. This process ensures that the reference frame's motion is effectively transferred to the blurry frame, improving alignment and reducing artifacts. The above process can be formalized as

$$F_t^{RMC} = \text{warp}(I_t^B, O_{t-1}^{Refined}) + f(I_t^B). \tag{6}$$

By refining the optical flow and leveraging the residual information, the RMC module ensures a more precise alignment between the current blurry frame and the reference frame, leading to improved deblurring performance.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets.** We use the RealMCVSR dataset [31] for our experiments. Originally designed for multi-view video super-resolution, RealMCVSR consists of triplets captured with ultra-wide, wide-angle, and telephoto lenses. In this setup, the wide-angle and telephoto videos share the same spatial dimensions as the ultra-wide video but have $\times 2$ and $\times 4$ higher resolutions, respectively. For the

AsLeD task, we adapt the dataset by using the ultra-wide video as the reference input and the wide-angle video as the blurry input. Since the wide-angle video has twice the resolution of the ultra-wide video, we center-crop and upsample the ultra-wide frames to match its resolution. We follow the original data split of the RealMCVSR dataset. To simulate motion blur, we generate blurred frames by the widely used technique of averaging multiple consecutive frames of the video captured by different lenses [65, 48, 47, 71, 50]. We generate blurry frames for training by averaging every 7 consecutive frames, simulating varying motion blur intensities. We synthesize motion blur by averaging seven consecutive frames for both training and testing, a widely used approximation that does not fully model real-world blur. For real captures, we use an iPhone 14 Plus whose two lenses introduce a physical baseline and distinct imaging characteristics, yielding cross-view shifts and photometric differences. We intentionally avoid spatial or color pre-alignment and ensure frame-level temporal synchronization with the DoubleTake APP. Note that RealMCVSR is captured on iPhone 12 Pro Max, whereas our real tests use iPhone 14 Plus, introducing cross-device differences in optics and ISP to assess generalization. Evaluations on these real videos demonstrate robust generalization beyond the averaging assumption.

**Implementation details.** We use 7 frames as input during training, with a mini-batch size of 4 and an input frame resolution of $128 \times 128$. We apply data augmentation techniques to the training data, including horizontal flips and random rotations of $90°$, $180°$, and $270°$. AsLeD-Net is trained for 300K iterations using the Adam optimizer with a Cosine Annealing learning rate scheduler. Network architecture parameters are set to $N_1 = 1$ and $N_2 = 30$. The number of channels is 64, and $K$ in ALM is set to 3. Supervision is enforced using the Charbonnier loss [71] via $\mathcal{L} = \sqrt{\|\hat{\mathcal{I}} - \mathcal{I}^{GT}\|^2 + \varepsilon^2}$, where $\varepsilon$ is set to $1 \times 10^{-3}$ in our experiments. We omit the subscript $t$ for simplicity. The initial learning rate for AsLeD-Net is $1 \times 10^{-4}$. Training is conducted on an NVIDIA RTX 3090 GPU.

**Inference settings.** We evaluate the reconstructed results using PSNR, SSIM, and LPIPS on the RGB channels to assess fidelity and perceptual quality.

### 4.2 Quantitative and Qualitative Comparisons

We compare the proposed AsLeD-Net against a diverse set of baseline methods to evaluate its effectiveness for the AsLeD task. These baselines encompass a wide range of potential approaches, aiming to cover as many varied and rich methodologies as possible: (1) Single-image deblurring and restoration methods: including MIMOUNet [13], MIMOUNet++ [13], NAFNet [11], and Restormer [86]. (2) Video deblurring and restoration methods: including IFIRNN [49], DBN [65], EDVR [71], BasicVSR [9], and BasicVSR++ [8]. (3) Reference-based video restoration method: specifically RefVSR [31]. Notably, some of these methods are not initially designed for deblurring. For approaches primarily intended for super-resolution tasks, we remove the upsampling operation and instead apply a convolutional layer to generate the final output directly. We meticulously re-train all baseline methods on our dataset to ensure a fair and comprehensive comparison using the publicly available code. However, due to *limited computational resources* (with only a 3090 GPU and 1080 Ti GPUs available), we cannot reproduce specific potential video deblurring methods. We plan to incorporate more advanced and computationally demanding methods when sufficient computational resources are available.

Table 1: Quantitative evaluation on the RealMCVSR testset. We mark the best and the second best results in **bold** and underline, respectively. #Params means the number of network parameters (M). Time costs (ms) are measured on blurred frames with a resolution of $256 \times 256$ using an NVIDIA GTX 1080 Ti GPU.

| Method | RealMCVSR | | | Costs | |
|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | #Params | Time |
| Blur frame | 18.22 | 0.8009 | 0.4319 | - | - |
| MIMOUNet | 24.50 | 0.8937 | 0.4042 | 6.8 | 23.1 |
| MIMOUNet++ | 24.70 | 0.8894 | 0.4074 | 16.1 | 47.4 |
| NAFNet | 24.80 | 0.9047 | 0.4109 | 67.9 | 52.9 |
| Restormer | 24.90 | 0.9045 | 0.3875 | 26.1 | 116.9 |
| IFIRNN | 24.75 | 0.8984 | 0.3434 | 4.1 | 7.6 |
| DBN | 25.05 | 0.9050 | 0.3763 | 15.3 | 4.2 |
| EDVR | 25.15 | 0.9088 | 0.3773 | 23.6 | 320.1 |
| BasicVSR | 25.10 | 0.9117 | 0.2885 | 6.2 | 55.8 |
| BasicVSR++ | 25.60 | 0.9121 | 0.2207 | 9.5 | 57.1 |
| DSTNet | <u>25.70</u> | <u>0.9134</u> | 0.2604 | 7.5 | 22.1 |
| RefVSR | 25.58 | 0.9124 | <u>0.2156</u> | 4.8 | 268.3 |
| Ours | **26.34** | **0.9167** | **0.1614** | 8.8 | 248.7 |

**Quantitative results.** As shown in Table 1, our proposed AsLeD-Net consistently achieves the best performance across PSNR, SSIM, and LPIPS metrics. Our method outperforms the second-best DSTNet by 0.64 dB in PSNR (26.34 vs. 25.70 dB), demonstrating superior noise reduction and detail preservation. Compared to widely-used methods like BasicVSR++ (25.60 dB) and EDVR

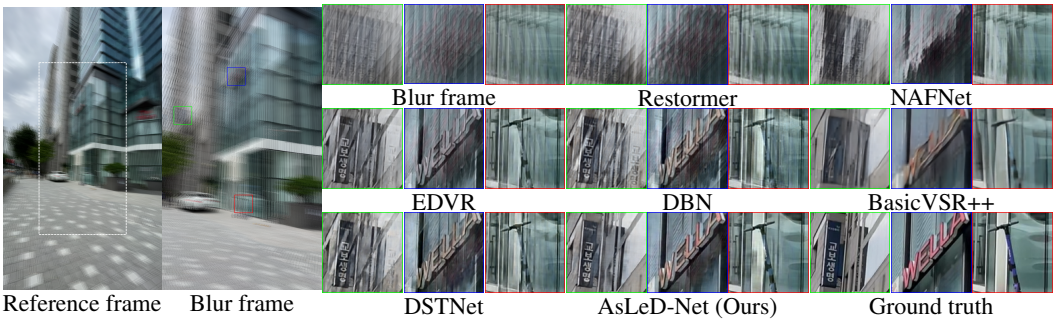



Figure 6: Qualitative comparison of AsLeD performance on the RealMCVSR dataset.

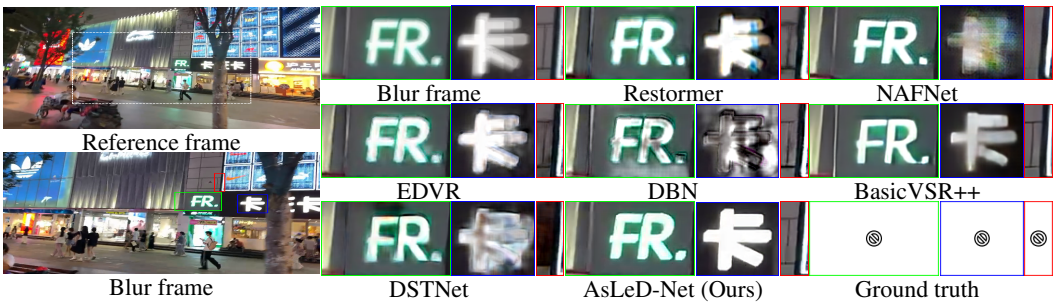



Figure 7: Qualitative comparison of AsLeD performance on real-world blurry scenes captured using an iPhone 14 Plus. *More results can be found in the supplementary material.*

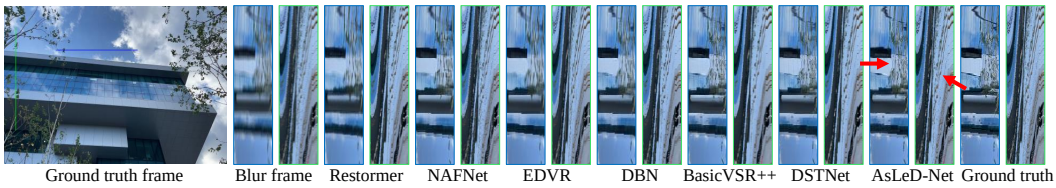



Figure 8: Temporal consistency comparison on the AsLeD task, where temporal profiles are extracted along horizontal (in blue) and vertical (in green) directions of the reconstructed frames.

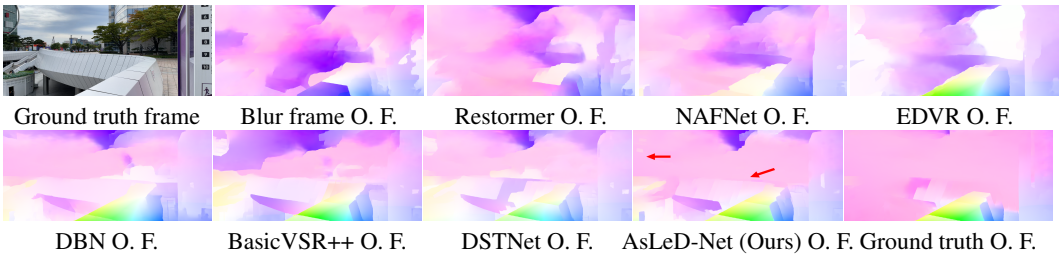



Figure 9: Temporal consistency comparison on the AsLeD task. The optical flow (O. F.) are estimated using the pre-trained RAFT [68].

(25.15 dB), our approach achieves gains of 0.74 dB and 1.19 dB, respectively. For SSIM, AsLeD-Net achieves 0.9167, improving upon DSTNet (0.9134), BasicVSR++ (0.9121), and RefVSR (0.9124), indicating enhanced structural consistency. Moreover, our method sets a new benchmark in LPIPS (0.1614), significantly surpassing the closest competitor, RefVSR (0.2156), with a 0.0542 reduction, demonstrating the perceptual quality of our restored frames.

**Computational cost results.** As shown in Table 1, our AsLeD-Net achieves superior performance on the RealMCVSR dataset but comes with a higher model complexity (8.8M parameters) and longer inference time (248.7 ms) compared to lightweight models like IFIRNN (4.1M, 7.6 ms) and DBN (15.3M, 4.2 ms). This limitation hinders its applicability in real-time or resource-constrained

scenarios. This limitation indicates that our method might not be optimal for real-time or resource-constrained applications. In future work, we plan to explore model compression techniques, such as knowledge distillation and lightweight architecture designs, to reduce computational costs while maintaining high restoration quality.

**Qualitative results.** We present visual comparison results on the RealMCVSR dataset in Figures 6 and 7. AsLeD-Net excels in preserving fine texture details, consistently outperforming methods such as Restormer, NAFNet, EDVR, DBN, DSTNet, and BasicVSR++. It effectively retains architectural details, such as grid-like window patterns and building facades, which are often blurred by other methods. As shown in Figure 7, AsLeD-Net produces sharper, more legible Chinese characters on signs and enhances clarity in real-world iPhone 14 Plus captures, revealing finer details in street signs and logos, thus improving text readability and recognition for practical applications.

**Temporal consistency.** We present the temporal consistency analysis in Figure 8 and Figure 9, highlighting the superior performance of our method in preserving motion coherence and fine details across consecutive frames. Compared to baselines, our method maintains smoother transitions and finer details in the temporal profiles, ensuring better motion consistency and reducing artifacts such as flickering. The estimated optical flow using RAFT [68] closely matches the ground truth and exhibits sharper, more well-defined structures than baselines. This demonstrates our model's ability to capture temporal coherence.

### 4.3 Ablation Study

We conduct experiments on RealMCVSR in terms of PSNR/SSIM. *Due to space limitations, please refer to the supplementary material.*

Table 2: Ablation study of three core components in AsLeD-Net on RealMCVSR. We replace the removed components with the residual blocks, ensuring parameter consistency.

| Method | Core Components | | | RealMCVSR | |
|---|---|---|---|---|---|
| | ALM | DC | RMC | PSNR↑ | SSIM↑ |
| (a) | ✗ | ✗ | ✗ | 25.40 | 0.9120 |
| (b) | ✓ | ✗ | ✗ | 25.72 | 0.9137 |
| (c) | ✗ | ✓ | ✗ | 25.69 | 0.9138 |
| (d) | ✗ | ✗ | ✓ | 25.79 | 0.9143 |
| (e) | ✗ | ✓ | ✓ | 25.89 | 0.9145 |
| (f) | ✓ | ✗ | ✓ | 25.84 | 0.9140 |
| (g) | ✓ | ✓ | ✗ | 25.75 | 0.9139 |
| (h) | ✓ | ✓ | ✓ | 26.34 | 0.9167 |

Table 2 ablates the three core components. Each single module improves the baseline (25.40 PSNR / 0.9120 SSIM). RMC yields the largest single module gain (25.79 / 0.9143). Adding RMC to ALM or DC brings further gains, and using all three achieves the best result (26.34 / 0.9167). This shows that temporal alignment is pivotal, while ALM and DC provide complementary spatial refinement. Within context aggregation, our ALM with $k = 3$ attains the best PSNR/SSIM and outperforms DAT and MASA SR under the same protocol. For latency sensitive settings, an ALM+RMC variant offers a favorable speed and quality tradeoff. Notably, ALM+DC without RMC (25.75 / 0.9139) yields only limited uplift, indicating that temporal alignment is the dominant bottleneck under realistic motion. Taken together with the ALM study (where $k=3$ peaks), these results suggest that structure aware reference matching and temporal propagation mitigate complementary error modes, namely cross view misalignment and blur induced motion ambiguity, and that a reduced ALM+RMC configuration is a practical choice when latency is critical.

## 5 Conclusion

In this paper, we propose AsLeD-Net, a practical video deblurring method that leverages the complementary perspectives of asymmetric dual-lens systems. By aligning and propagating temporal reference features from ultra-wide views and fusing them with blurry wide-angle frames, AsLeD-Net effectively addresses cross-lens redundancy. Key modules such as the ALM module and the DC module ensure refined feature alignment and spatial consistency, while the RMC module enhances temporal alignment. Through extensive experiments, we validate the effectiveness of AsLeD-Net, showcasing its superiority over existing methods for deblurring asymmetric lens systems.

Our method performs well across various blur levels and scenarios, but frame averaging for training has limitations in simulating real-world motion blur, especially in high-motion and high-contrast scenes, and is frame rate-dependent; *see the supplementary material for details*.

**Acknowledgment.** This project is supported by the National Research Foundation, Singapore, under its Medium Sized Center for Advanced Robotics Technology Innovation.

## References


[1] Hadi Alzayer, Abdullah Abuolaim, Leung Chun Chan, Yang Yang, Ying Chen Lou, Jia-Bin Huang, and Abhishek Kar. Dc2: Dual-camera defocus control by learning to refocus. In *CVPR*, 2023.

[2] Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *CVPR*, 2025.

[3] Haowen Bai, Zixiang Zhao, Jiangshe Zhang, Yichen Wu, Lilun Deng, Yukun Cui, Baisong Jiang, and Shuang Xu. Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning. *International Journal of Computer Vision*, 2024.

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *CVPR*, 2021.

[5] Goutam Bhat, Michaël Gharbi, Jiawen Chen, Luc Van Gool, and Zhihao Xia. Self-supervised burst super-resolution. In *ICCV*, 2023.

[6] Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *ECCV*, 2022.

[7] Tiantian Cao, Xuan Dong, Chunli Peng, Zhengqing Li, Xinyu Guo, and Weixin Li. View transition based dual camera image fusion. *arXiv preprint arXiv:2312.11184*, 2023.

[8] Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, 2022.

[9] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR*, 2021.

[10] Zhu Chao, Dong Hang, Pan Jinshan, Liang Boyang, Huang Yuhao, Fu Lean, and Wang Fei. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In *AAAI*, 2022.

[11] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, 2022.

[12] Xihao Chen, Zhiwei Xiong, Zhen Cheng, Jiayong Peng, Yueyi Zhang, and Zheng-Jun Zha. Degradation-agnostic correspondence from resolution-asymmetric stereo. In *CVPR*, 2022.

[13] 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 *ICCV*, 2021.

[14] Mauricio Delbracio and Guillermo Sapiro. Burst deblurring: Removing camera shake through fourier burst accumulation. In *CVPR*, 2015.

[15] Xuan Dong, Weixin Li, Xiaoyan Hu, Xiaojie Wang, and Yunhong Wang. A colorization framework for monochrome-color dual-lens systems using a deep convolutional network. *IEEE TVCG*, 28(3):1469–1485, 2020.

[16] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *CVPR*, 2022.

[17] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burstormer: Burst image restoration and enhancement transformer. In *CVPR*, 2023.

[18] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Husan Yang. Burst image restoration and enhancement. *IEEE TPAMI*, 2024.

[19] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising. In *ECCV*, 2018.

[20] Yixuan Huang, Xiaoyun Zhang, Yu Fu, Siheng Chen, Ya Zhang, Yan-Feng Wang, and Dazhi He. Task decoupled framework for reference-based super-resolution. In *CVPR*, 2022.

[21] Bo Ji and Angela Yao. Multi-scale memory-based video deblurring. In *CVPR*, 2022.

[22] Bangrui Jiang, Zhihuai Xie, Zhen Xia, Songnan Li, and Shan Liu. Erdn: Equivalent receptive field deformable network for video deblurring. In *ECCV*, 2022.

[23] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *CVPR*, 2021.

[24] Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Reference-based image and video super-resolution via c2-matching. *IEEE TPAMI*, 45(7):8874–8887, 2022.

[25] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Turning frequency to resolution: Video super-resolution via event cameras. In *CVPR*, 2021.

[26] Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *IEEE TIP*, 30:9372–9385, 2021.

[27] Tae Hyun Kim, Kyoung Mu Lee, Bernhard Schölkopf, and Michael Hirsch. Online video deblurring via dynamic temporal blending network. In *ICCV*, 2017.

[28] Youngrae Kim, Jinsu Lim, Hoonhee Cho, Minji Lee, Dongman Lee, Kuk-Jin Yoon, and Ho-Jin Choi. Efficient reference-based video super-resolution (ervsr): Single reference image is all you need. 2023.

[29] Seonggwan Ko, Yeong Jun Koh, and Donghyeon Cho. Reference-based burst super-resolution. In *ACM MM*, 2024.

[30] Wei-Sheng Lai, Yichang Shih, Lun-Cheng Chu, Xiaotong Wu, Sung-Fang Tsai, Michael Krainin, Deqing Sun, and Chia-Kai Liang. Face deblurring using dual camera fusion on mobile phones. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022.

[31] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *CVPR*, 2022.

[32] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *CVPR*, 2023.

[33] Dongxu Li, Chenchen Xu, Kaihao Zhang, Xin Yu, Yiran Zhong, Wenqi Ren, Hanna Suominen, and Hongdong Li. Arvo: Learning all-range volumetric correspondence for video deblurring. In *CVPR*, 2021.

[34] Weixin Li, Tiantian Cao, Chang Liu, Xue Tian, Ya Li, Xiaojie Wang, and Xuan Dong. Dual-lens hdr using guided 3d exposure cnn and guided denoising transformer. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(5):1–20, 2023.

[35] Zhuoyuan Li, Jiacheng Li, Yao Li, Li Li, Dong Liu, and Feng Wu. In-loop filtering via trained look-up tables. In *VCIP*, 2024.

[36] Zhuoyuan Li, Junqi Liao, Chuanbo Tang, Haotian Zhang, Yuqi Li, Yifan Bian, Xihua Sheng, Xinmin Feng, Yao Li, Changsheng Gao, et al. Ustc-td: A test dataset and benchmark for image and video coding in 2020s. *IEEE Transactions on Multimedia*, 2025.

[37] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE TIP*, 2024.

[38] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. 2022.

[39] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. 2022.

[40] Jing Lin, Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Youliang Yan, Xueyi Zou, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Flow-guided sparse transformer for video deblurring. 2022.

[41] Cunzhe Liu, Zhen Hua, and Jinjiang Li. Deep reference-based dynamic scene deblurring. *KSII Transactions on Internet & Information Systems*, 18(3), 2024.

[42] Cunzhe Liu, Zhen Hua, and Jinjiang Li. Reference-based dual-task framework for motion deblurring. *The Visual Computer*, 40(1):137–151, 2024.

[43] Huan Liu, Mingwen Shao, Yecong Wan, Yuexian Liu, and Kai Shang. Sebir: Semantic-guided burst image restoration. *Neural Networks*, 181:106834, 2025.

[44] Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *CVPR*, 2021.

[45] Xianrui Luo, Zijin Wu, Juewen Peng, Huiqiang Sun, Zhiguo Cao, and Guosheng Lin. Dual-camera all-in-focus neural radiance fields. *IEEE TPAMI*, 2025.

[46] MR Mahesh Mohan, GK Nithin, and AN Rajagopalan. Deep dynamic scene deblurring for unconstrained dual-lens cameras. *IEEE TIP*, 30:4479–4491, 2021.

[47] 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 *CVPRW*, 2019.

[48] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017.

[49] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *CVPR*, 2019.

[50] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, Kyoung Mu Lee, Liangyu Chen, Jie Zhang, Xin Lu, Xiaojie Chu, Chengpeng Chen, et al. Ntire 2021 challenge on image deblurring. In *CVPR*, 2021.

[51] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *CVPR*, 2020.

[52] Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *CVPR*, 2023.

[53] Chunli Peng, Xuan Dong, Tiantian Cao, Zhengqing Li, Kun Dong, and Weixin Li. Rewite: Realistic wide-angle and telephoto dual camera fusion dataset via beam splitter camera rig. In *ACM MM*, 2024.

[54] Long Peng, Yang Cao, Yuejin Sun, and Yang Wang. Lightweight adaptive feature de-drifting for compressed image classification. *IEEE Transactions on Multimedia*, 26:6424–6436, 2024.

[55] Long Peng, Yang Wang, Xin Di, Xueyang Fu, Yang Cao, Zheng-Jun Zha, et al. Boosting image de-raining via central-surrounding synergistic convolution. In *AAAI*, 2025.

[56] Marco Pesavento, Marco Volino, and Adrian Hilton. Attention-based multi-reference learning for image super-resolution. In *ICCV*, 2021.

[57] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017.

[58] Chen Rao, Guangyuan Li, Zehua Lan, Jiakai Sun, Junsheng Luan, Wei Xing, Lei Zhao, Huaizhong Lin, Jianfeng Dong, and Dalong Zhang. Rethinking video deblurring with wavelet-aware dynamic transformer and diffusion model. In *ECCV*, 2024.

[59] Jaesung Rim, Junyong Lee, Heemin Yang, and Sunghyun Cho. Deep hybrid camera deblurring for smartphone cameras. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[60] Finn Russell and William JB Midgley. Asymmetric stereo high dynamic range imaging with smartphone cameras. *Sensors*, 24(18):5876, 2024.

[61] Shayan Shekarforoush, Amanpreet Walia, Marcus A Brubaker, Konstantinos G Derpanis, and Alex Levinshtein. Dual-camera joint deblurring-denoising. *arXiv preprint arXiv:2309.08826*, 2023.

[62] Gyumin Shim, Jinsun Park, and In So Kweon. Robust reference-based super-resolution with similarity-aware deformable convolution. In *CVPR*, 2020.

[63] Hyeongseok Son, Junyong Lee, Jonghyeop Lee, Sunghyun Cho, and Seungyong Lee. Recurrent video deblurring with blur-invariant motion estimation and pixel volumes. *IEEE TIP*, 40(5):185:1–185:18, 2021.

[64] Taeyong Song, Sunok Kim, and Kwanghoon Sohn. Unsupervised deep asymmetric stereo matching with spatially-adaptive self-similarity. In *CVPR*, 2023.

[65] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017.

[66] Maitreya Suin and A. N. Rajagopalan. Gated spatio-temporal attention-guided video deblurring. In *CVPR*, 2021.

[67] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018.

[68] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020.

[69] Ruohao Wang, Xiaohui Liu, Zhilu Zhang, Xiaohe Wu, Chun-Mei Feng, Lei Zhang, and Wangmeng Zuo. Benchmark dataset and effective inter-frame alignment for real-world video super-resolution. In *CVPR*, 2023.

[70] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *ICCV*, 2021.

[71] Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019.

[72] Yusheng Wang, Yunfan Lu, Ye Gao, Lin Wang, Zhihang Zhong, Yinqiang Zheng, and Atsushi Yamashita. Efficient video deblurring guided by motion magnitude. In *ECCV*, 2022.

[73] Pengxu Wei, Yujing Sun, Xingbei Guo, Chang Liu, Guanbin Li, Jie Chen, Xiangyang Ji, and Liang Lin. Towards real-world burst image super-resolution: Benchmark and method. In *ICCV*, 2023.

[74] Patrick Wieschollek, Michael Hirsch, Bernhard Schölkopf, and Hendrik P. A. Lensch. Learning blind motion deblurring. In *ICCV*, 2017.

[75] Renlong Wu, Zhilu Zhang, Yu Yang, and Wangmeng Zuo. Dual-camera smooth zoom on mobile phones. In *ECCV*, 2024.

[76] Xiaotong Wu, Wei-Sheng Lai, Yichang Shih, Charles Herrmann, Michael Krainin, Deqing Sun, and Chia-Kai Liang. Efficient hybrid zoom using camera fusion on mobile phones. *ACM Transactions on Graphics (TOG)*, 42(6):1–12, 2023.

[77] Zeyu Xiao, Jiawang Bai, Zhihe Lu, and Zhiwei Xiong. A dive into sam prior in image restoration. *arXiv preprint arXiv:2305.13620*, 2023.

[78] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong. Space-time distillation for video super-resolution. In *CVPR*, 2021.

[79] Zeyu Xiao, Zhihe Lu, Michael Bi Mi, Zhiwei Xiong, and Xinchao Wang. Unraveling motion uncertainty for local motion deblurring. In *ACMMM*, 2024.

[80] Zeyu Xiao and Zhiwei Xiong. Incorporating degradation estimation in light field spatial super-resolution. *CVIU*, 252:104295, 2025.

[81] Zeyu Xiao, Zhiwei Xiong, Xueyang Fu, Dong Liu, and Zheng-Jun Zha. Space-time video super-resolution using temporal profiles. In *ACMMM*, 2020.

[82] Yanchun Xie, Jimin Xiao, Mingjie Sun, Chao Yao, and Kaizhu Huang. Feature representation matters: End-to-end learning for reference-based image super-resolution. In *ECCV*, 2020.

[83] Ruikang Xu, Mingde Yao, Yue Li, Yueyi Zhang, and Zhiwei Xiong. High-resolution and few-shot view synthesis from asymmetric dual-lens inputs. In *ECCV*, 2024.

[84] Ruikang Xu, Mingde Yao, and Zhiwei Xiong. Zero-shot dual-lens super-resolution. In *CVPR*, 2023.

[85] Huanjing Yue, Zifan Cui, Kun Li, and Jingyu Yang. Kedusr: Real-world dual-lens super-resolution via kernel-free matching. In *AAAI*, 2024.

[86] 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 *CVPR*, 2022.

[87] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, 2019.

[88] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Spatio-temporal deformable attention network for video deblurring. In *ECCV*, 2022.

[89] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Blur-aware spatio-temporal sparse transformer for video deblurring. In *CVPR*, 2024.

[90] Jinrui Zhang, Huan Yang, Ju Ren, Deyu Zhang, Bangwen He, Ting Cao, Yuanchun Li, Yaoxue Zhang, and Yunxin Liu. Mobidepth: Real-time depth estimation using on-device dual cameras. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022.

[91] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Wei Liu, and Hongdong Li. Adversarial spatio-temporal learning for video deblurring. *IEEE TIP*, 28(1):291–301, 2019.

[92] Lin Zhang, Xin Li, Dongliang He, Fu Li, Errui Ding, and Zhaoxiang Zhang. Lmr: A large-scale multi-reference dataset for reference-based super-resolution. In *ICCV*, 2023.

[93] Lin Zhang, Xin Li, Dongliang He, Fu Li, Yili Wang, and Zhaoxiang Zhang. Rrsr: Reciprocal reference-based image super-resolution with progressive feature alignment and selection. In *ECCV*, 2022.

[94] Zhilu Zhang, Ruohao Wang, Hongzhi Zhang, Yunjin Chen, and Wangmeng Zuo. Self-supervised learning for real-world super-resolution from dual zoomed observations. In *ECCV*, 2022.

[95] Zhilu Zhang, Ruohao Wang, Hongzhi Zhang, and Wangmeng Zuo. Self-supervised learning for real-world super-resolution from dual and multiple zoomed observations. *IEEE TPAMI*, 2024.

[96] Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. A unified solution to video fusion: From multi-frame learning to benchmarking. *arXiv preprint arXiv:2505.19858*, 2025.

[97] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *CVPR*, 2023.

[98] Zhihang Zhong, Ye Gao, Yinqiang Zheng, Bo Zheng, and Imari Sato. Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network. *IJCV*, 2022.

[99] Kun Zhou, Wenbo Li, Liying Lu, Xiaoguang Han, and Jiangbo Lu. Revisiting temporal alignment for video restoration. In *CVPR*, 2022.

[100] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Wangmeng Zuo, Haozhe Xie, and Jimmy S. J. Ren. Spatio-temporal filter adaptive network for video deblurring. In *ICCV*, 2019.

[101] Chao Zhu, Hang Dong, Jinshan Pan, Boyang Liang, Yuhao Huang, Lean Fu, and Fei Wang. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In *AAAI*, 2022.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the contributions of AsLeD-Net. The paper introduces the first framework for asymmetric dual-lens video deblurring (AsLeD), proposes key modules (ALM, DC, RMC), and validates superior performance through experiments (Section 4). Contributions are explicitly outlined in Section 1 and Section 3.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 ("Discussions") explicitly addresses limitations: (1) training data generated via frame averaging may not fully simulate real-world motion blur, especially in high-motion scenarios; (2) computational inefficiency (248.7 ms inference time, 8.8M parameters) limits real-time applicability. Future directions (e.g., model compression) are suggested.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: The paper focuses on algorithmic design and empirical validation, with no theoretical derivations or mathematical proofs. Methodology is described through architectural diagrams (Figures 3-5) and equations for feature fusion (Eqs. 1-7).

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4.1 details datasets (RealMCVSR), training protocols (Adam, 300K iterations), and hyperparameters.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper references public datasets (RealMCVSR) but provides no links to code repositories or pre-trained models. Reproducibility relies on re-implementation from textual descriptions.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 comprehensively describes experimental settings: datasets, blur simulation (7-frame averaging), input resolution ($128 \times 128$), optimizer (Adam with Cosine Annealing), loss function (Charbonnier), hardware (NVIDIA RTX 3090), and evaluation metrics (PSNR, SSIM, LPIPS).

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Results in Tables 1-3 report PSNR/SSIM/LPIPS values but lack error bars, confidence intervals, or statistical significance tests. Variability factors (e.g., data splits, initialization) are not discussed.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper states training hardware (RTX 3090) but omits per-experiment memory usage, total compute costs, or training time. Inference time (248.7 ms) is reported.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The "Broader impact" section addresses ethical concerns: privacy risks (enhanced surveillance), dual-use potential (misuse in synthetic media), and model bias. The authors align with NeurIPS ethics guidelines.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive impacts (healthcare, autonomous systems) and negative risks (privacy violations, misuse). While mitigation strategies are not detailed, ethical trade-offs are acknowledged.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release high-risk assets (models, datasets). AsLeD-Net is presented as a research framework without deployment, and no safeguards are discussed.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: Public datasets (RealMCVSR) and methods (e.g., BasicVSR, EDVR) are cited, but licenses are not explicitly mentioned (e.g., RealMCVSR [30] terms of use are omitted).

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce new datasets, codebases, or models.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The work involves no human subjects or crowdsourcing.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects research is conducted.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs are not used in methodology, experiments, or analysis.