2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

BlurPoint: Efficient Motion Blur Aware Student-Teacher Local Feature Learning

Wenting Wang[1,2†], Zhenjun Zhao[1,3†], Jiaxin Guo[1], Yun-Hui Liu[1], Charlie C.L. Wang[4], Yeung Yam[1,2]

***Abstract*— Local feature detection and description serve as the foundation for many 3D vision tasks. However, most existing algorithms rely on sharp images, resulting in degraded performance when motion blur occurs due to long exposure. To tackle this challenge, we propose an effective end-to-end model that jointly learns feature detection and description from blurred images in a self-supervised manner, without requiring any additional labeled data. Rather than simply mixing sharp and blurred samples during training, we design a student–teacher framework to explicitly transfer knowledge from sharp to blurred domains. The teacher model extracts local features from sharp images and enforces photometric consistency in feature space, which is then distilled to the student model trained on blurred inputs. To facilitate this knowledge transfer, we introduce two tailored loss functions, feature divergence loss and triplet knowledge distillation loss, both aimed at aligning feature representations under motion blur. Extensive experiments on homography estimation, relative pose estimation, and visual localization demonstrate that our method achieves state-of-the-art performance on blurred images, while maintaining competitive accuracy on sharp images.**

I. Introduction

Detecting and describing local features are crucial tasks in various computer vision and robotics applications, including Structure-from-Motion (SfM), Simultaneous Localization and Mapping (SLAM), camera calibration, object detection, image retrieval, and visual localization [1]–[15]. Numerous algorithms have been proposed, ranging from classical geometric methods [16], [17] to deep learning-based approaches [18], [19]. Despite significant advances, motion blur remains a major challenge for local feature detection and description. Motion blur is one of the most common artifacts affecting image quality, often occurring in low-light conditions that require long exposure times. Motion blur is also commonly observed in everyday videos due to rapid handheld movement or unintended camera shake. Such degradation severely impacts feature-based methods, which struggle to detect and describe repeatable and discriminative local features necessary for correspondence estimation.

A common two-stage pipeline for handling motion blur is the *deblur-then-extract* approach as shown in Fig. 2, which first restores sharp images using a separate deblurring network [20]–[22], followed by local feature extraction. However, this strategy suffers from two major limitations. First, deblurring networks often incur high computational cost and memory overhead, which hinders efficient descriptor extraction and negatively affects downstream tasks. Second, deblurring can introduce unpredictable artifacts that degrade the quality of extracted descriptors. These challenges motivate a one-stage solution that directly extracts local features from motion blurred images.

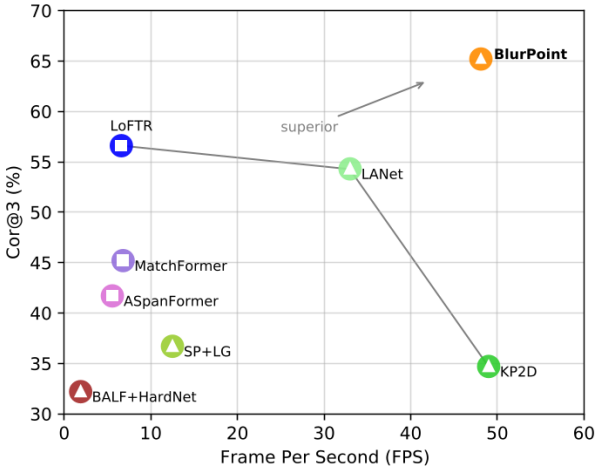

Fig. 1. **Efficiency and performance comparison of local feature methods on blurred images**. Sparse methods (triangles) and semi-dense methods (squares) are evaluated for correctness@3pixel on the Blur-HPatches dataset. The proposed BlurPoint demonstrates a clear advantage in achieving a superior balance between performance and computational efficiency.

In recent years, deep learning has shown great potential for improving local feature extraction [23], [24]. Numerous state-of-the-art methods have been proposed, achieving notable progress in improving robustness to challenges such as viewpoint and illumination variations. Nonetheless, little attention has been paid to extracting local features from motion blurred images, which is crucial for enabling robust 3D vision in low-light environments or under sudden and unpredictable camera motion. BALF [25] proposes a keypoint detector for blurred images using an MLP-based network. However, it only predicts keypoint locations and lacks feature descriptors, which are essential for image matching. To address this issue, we propose a unified network that jointly learns keypoint detection and description on blurred images. However, their effectiveness on motion blurred images remains unexplored. Our approach fills this gap by enabling more robust and accurate relative pose estimation in blurred scenarios, achieving superior accuracy and efficiency on

[1] Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong SAR.
[2] Centre for Perceptual and Interactive Intelligence (CPII) Limited, Hong Kong SAR.
[3] Formerly with The Chinese University of Hong Kong; now with the Department of Computer Science and Systems Engineering, University of Zaragoza, Zaragoza, Spain.
[4] Department of Mechanical, Aerospace and Civil Engineering, The University of Manchester, Manchester, UK.
† Corresponding author: Zhenjun Zhao (zhenjunz@unizar.es), Wenting Wang (wtwang@mae.cuhk.edu.hk)

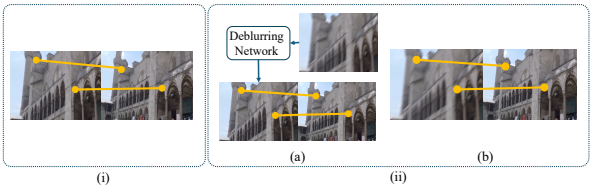

Fig. 2. **Local feature learning on blurred images.** (i) *Sharp-to-Sharp* image matching. (ii) *Sharp-to-Blur* image matching, which includes: (a) a two-stage pipeline — *deblur-then-match*; and (b) a one-stage approach — *match-on-blur*.

motion blurred data as illustrated in Fig. 1. Another practical challenge is acquiring data with ground truth correspondences and pose labels. Fully supervised learning requires camera intrinsics, extrinsics, and depth to obtain ground truth correspondences. However, acquiring such data is often infeasible in real-world or publicly available synthetic datasets. Therefore, we adopt a self-supervised local feature learning strategy [24], which requires only the blurred image itself. Despite this, a significant performance gap persists between sharp and blurred images in downstream tasks. To address this gap, we propose leveraging photometric knowledge learned from sharp images to implicitly guide local feature learning on blurred inputs. Specifically, a teacher-student framework is well suited for this purpose, where a pre-trained teacher model guides the student model in detecting and describing keypoints on blurred images.

In this paper, we introduce ***BlurPoint***, a self-supervised framework for motion blur aware local feature learning. Our network consists of two key components: a self-supervised student model that learns to detect and describe local features from motion blurred images, and a knowledge distillation process that transfers knowledge learned from sharp images to blurred images. The teacher model supervises the student via two proposed losses: a feature divergence loss, which guides the student's detector in the feature space, and a triplet knowledge distillation loss, which transfers descriptor-level knowledge from the teacher to the student. This design significantly enhances robustness and accuracy in local feature learning under motion blur. Extensive evaluations on both synthetic and real datasets demonstrate that our method consistently outperforms existing approaches in homography and relative pose estimation tasks on blurred images.

In summary, our **contributions** are as follows:

- We propose a novel joint learning network for local feature detection and description on motion blurred images, an area that remains largely unexplored.
- We employ a self-supervised strategy to optimize the student model while distilling knowledge of local features from sharp images to motion blurred images via a teacher network.
- Extensive experiments demonstrate that our method outperforms prior works in local feature extraction under motion blur, while maintaining competitive performance on sharp images.

II. Related Work

A. Local Feature Extraction on Blurred Images

Existing data-driven local feature methods [24], [26] perform well on sharp images but often struggle to handle motion blur, which is prevalent in real-world scenarios. Keypoint extraction under motion blur is particularly challenging because feature distortion caused by relative motion between the camera and scene significantly degrades detection and description accuracy. A straightforward solution is the *deblur-then-match* pipeline, illustrated in Fig. 2. It first deblurs the image using a separate deblurring network [21], [22], then applies a keypoint detector to the restored image. However, this approach inherits the limitations of both stages: the quality of deblurring directly impacts keypoint detection reliability, and artifacts introduced during deblurring can degrade feature accuracy.

Deblur-then-match. Deblurring algorithms can be broadly categorized into classical optimization-based and deep learning-based methods. Early learning-based methods [27] train neural networks to estimate blur kernels. These kernels are then used in traditional deconvolution procedures. Later works such as DeblurGAN-v2 [21] employ adversarial loss to improve motion deblurring results. Tao *et al.* [22] propose SRN-DeblurNet, a multi-scale deblurring network based on a scale-recurrent architecture. Although these methods generate visually sharp results, they often introduce artifacts. Moreover, they require substantial computational resources, making real-time deployment challenging.

Match-on-blur. Rather than deblurring each image and then extracting features, a more efficient and promising alternative is an end-to-end approach that directly extracts features from motion blurred images. BALF [25] proposes a keypoint detection method specifically designed for blurred images using an MLP-based network. However, it focuses solely on keypoint localization and lacks descriptors, limiting its applicability in downstream tasks like relative pose estimation and visual localization. To address this, we propose a joint motion blur-aware detector and descriptor that operates directly on blurred images.

B. Knowledge Distillation in Local Feature Learning

Student-teacher learning has been widely explored in computer vision tasks such as object detection [28], segmentation [29], and tracking [30]. Previous works have used student-teacher paradigms to perform cross-modal knowledge distillation for local features, for example by transferring depth [31] or semantic [32] information into feature representations on sharp images. However, these approaches assume ideal high-quality image conditions and often fail in real-world settings with motion blur due to long exposure or fast motion. In contrast, we propose a novel strategy to distill knowledge from sharp to blurred images without relying on additional supervision. Our framework leverages a pre-trained teacher model on sharp images to guide the student model trained on blurred images, effectively transferring robust local feature representations via self-supervision.

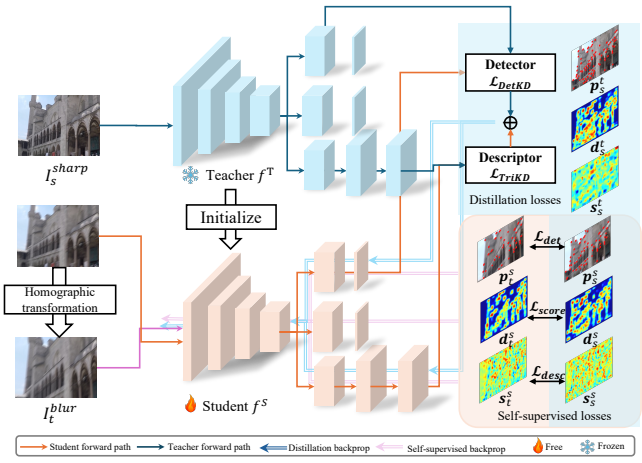

Fig. 3. **Overview of the proposed method.** BlurPoint has two main training schemes: 1) A self-supervised learning module to train student model $f^{\mathcal{S}}$ by using $\mathcal{L}_{det}$, $\mathcal{L}_{desc}$ and $\mathcal{L}_{score}$ (Section III-A); 2) A knowledge distillation regime to learn keypoint location features and descriptor feature distributions from sharp images based on feature divergence loss $\mathcal{L}_{DetKD}$ and triplet knowledge distillation loss $\mathcal{L}_{TriKD}$ (Section III-B).

III. Method

Our proposed method, BlurPoint, is designed to train a motion blur-aware local feature learning model (Student) by distilling knowledge from a pre-trained model on sharp images (Teacher). As illustrated in Fig. 3, we first train a Teacher model on sharp images. Then, we train the Student model using a self-supervised paradigm combined with knowledge distillation. The self-supervised paradigm jointly trains detection and description of salient points without relying on ground truth labels. Knowledge is then transferred to the Student model through two key strategies: 1) **feature divergence loss**, which guides the Student to perform the Teacher's detection responses; and 2) **triplet knowledge distillation loss**, which distills descriptor knowledge from Teacher to Student. These mechanisms collectively enhance the Student model's capacity to detect and describe features under motion blur, facilitating more robust matching in challenging blurred scenarios.

A. Self-supervised Local Feature Learning

In this work, both the Teacher and Student models are trained to regress a function that takes an image as input and simultaneously outputs keypoints, descriptors, and scores. Specifically, we define a function $f : I \rightarrow \boldsymbol{p}, \boldsymbol{d}, \boldsymbol{s}$, where the input image $I \in \mathbb{R}^{3\times H\times W}$, and the outputs are keypoints $\boldsymbol{p} = \{(u_i, v_i) \mid i = 1, \ldots, N\} \in \mathbb{R}^{2\times N}$, descriptors $\boldsymbol{d} \in \mathbb{R}^{C\times N}$, and keypoint scores $\boldsymbol{s} \in \mathbb{R}^{1\times N}$. Here, N represents the total number of keypoints extracted, which varies with the input image resolution. The Teacher model $f^{\mathcal{T}}$ takes sharp images I^{sharp} as input, while the Student model $f^{\mathcal{S}}$ takes blurred images I^{blur} as input. Following [24], we train both the Teacher and Student models in a self-supervised manner by applying randomly sampled but known homography transformations to the original image. Specifically, we treat the original image as the source image I_s, resulting in $f(I_s) = \{\boldsymbol{p}_s, \boldsymbol{d}_s, \boldsymbol{s}_s\}$, and the warped image as the target image I_t, giving $f(I_t) = \{\boldsymbol{p}_t, \boldsymbol{d}_t, \boldsymbol{s}_t\}$. The images I_s and I_t are related through a known homography transformation $\mathbf{H}$, which maps pixels from the source to the target image, such that $I_t = \mathbf{H}(I_s)$. We use superscripts stu and tch to denote student and teacher models, and subscripts s and t to denote source and target images, respectively. The primary goal of self-supervised learning is to ensure geometric consistency in keypoint locations and descriptors between the source and target images. The overall architecture of our self-supervised local feature learning follows an encoder-decoder network consisting of a shared encoder and three separate decoder heads for detection, description, and scoring, closely following the structure of [24].

Detection learning. The detection head predicts an offset for each entry $\tilde{\boldsymbol{p}} = (\tilde{u}, \tilde{v}) \in \mathbb{R}^{2\times \frac{H}{w}\times \frac{W}{w}}$ within a window, which is then mapped to original-resolution pixel coordinates. We optimize keypoint locations using the following self-supervised loss formulation, enforcing keypoint location consistency across different views of the same scene:

$$\mathcal{L}_{det} = \sum_i \|H(\tilde{p}_{s,i}) - \tilde{p}_{t,i}\|_2 \tag{1}$$

Description learning. We adopt a per-pixel triplet loss [33] with nested hardest negative sampling [34] to supervise descriptor learning. The objective is to minimize the distance between *anchor* $d_{s,i}$ and *positive* $d^+_{t,i}$ descriptors while maximizing the distance between anchor and *negative* descriptors $d^-_{t,i}$ in the embedding space. This hardest negative sample contributes most to the loss, thereby accelerating metric learning. The descriptor loss is therefore defined as:

$$\mathcal{L}_{desc} = \sum_i \max\left(\|d_{s,i} - d^+_{t,i}\|_2 - \|d_{s,i} - d^-_{t,i}\|_2 + \epsilon_1, 0\right) \tag{2}$$

where ϵ_1 denotes the margin that controls how far apart the anchor and hardest negative should be in the descriptor space.

Score learning. The score head in the decoder outputs a confidence value for each descriptor, indicating the likelihood of the corresponding point being a keypoint. The objective of score learning is twofold: 1) to enforce score consistency between matched point pairs, and 2) to assign higher scores to geometrically accurate correspondences:

$$\mathcal{L}_{score} = \sum_i \|s_{s,i} - s_{t,i}\|_2 + \frac{(s_{s,i} + s_{t,i})}{2} \cdot \left(\|H(p_{s,i}) - p_{t,i}\|_2 - \bar{l}\right) \tag{3}$$

where $\bar{l}$ denotes the average reprojection error across all point pairs between the source and target images.

B. Motion Blur Knowledge Distillation

Self-supervised learning performs well in-domain but degrades on motion-blurred images due to a domain gap between sharp and blurred feature representations (Table V). Mixed training on both domains confuses the network and weakens features for each. We address this with a student–teacher framework: a teacher, pre-trained on sharp images and then frozen, guides a student trained on blurred

images to jointly detect and describe features. The student is optimized with distillation losses from the teacher alongside self-supervised objectives on blurred data. This separation preserves domain-specific representations and improves overall performance across sharp and blurred imagery.

Detection feature distillation. We propose to reduce the difference between the probability distributions of the teacher and student detector feature maps in the feature space, rather than directly regressing the positions of sharp keypoints on blurred images. Specifically, we use the Kullback–Leibler (KL) divergence to distill soft targets for the student detector and guide it to learn the underlying distribution from the teacher. We extract features from the penultimate layer of the teacher detector head $x^{\mathcal{T}}$ and the corresponding layer of the student detector head $x^{\mathcal{S}}$. In particular, the feature divergence loss function $\mathcal{L}_{DetKD}$ is set to be:

$$\mathcal{L}_{DetKD} = \frac{1}{N}\sum_{i} f_{KL}\big(\sigma(x_i^{\mathcal{S}}), \sigma(x_i^{\mathcal{T}})\big) \tag{4}$$

where $\sigma(\cdot)$ denotes the *Softmax* function, which transforms a feature vector into a probability distribution. In this context, we treat each feature vector as a distribution. The KL divergence function f_{KL} is defined as:

$$f_{KL}(p, q) = \sum_{k=1}^{c} \sigma(p_k) log\Big(\frac{p_k}{q_k}\Big) \tag{5}$$

where p, q are two distributions. $\mathcal{L}_{DetKD}$ encourages the student blur-detector to produce feature distributions similar to those of the teacher sharp-detector in the feature space.

Description triplet knowledge distillation. We adopt a response-based knowledge distillation strategy to help the student model learn keypoint descriptors from the teacher model. Inspired by the concept of Knowledge Distillation (KD) loss [35], we introduce a novel triplet loss, denoted as $\mathcal{L}_{TriKD}$, for student-teacher learning. In $\mathcal{L}_{TriKD}$, we define the loss using a triplet structure. For each keypoint, we assign the *anchor* descriptor $d_i^{\mathcal{S}}$ from the student branch and the corresponding *positive* descriptor $d_i^{\mathcal{T},+}$ from the teacher branch. These descriptors correspond to the same spatial locations in both sharp and blurred images. In addition, the *negative* descriptor $d_i^{\mathcal{T},-}$ is selected as the non-matching descriptor in the teacher's descriptor head that is closest to the positive descriptor in feature space. The objective of the triplet loss is twofold: to minimize the distance between the anchor descriptor from the student's blur branch and the positive descriptor from the teacher's sharp branch, while maximizing the distance to the negative descriptor. This approach ensures that the student model learns descriptors for blurred images that are consistent with those for sharp images. The formulation of $\mathcal{L}_{TriKD}$ is as follows:

$$\mathcal{L}_{TriKD} = \sum_{i} \max\big(\|d_i^{\mathcal{S}} - d_i^{\mathcal{T},+}\|_2 - \|d_i^{\mathcal{S}} - d_i^{\mathcal{T},-}\|_2 + \epsilon_2, 0\big) \tag{6}$$

where ϵ_2 denotes the triplet margin. This loss encourages the student model to generate descriptors for blurred images that closely resemble those of sharp images, thereby enhancing the consistency and quality of feature representations across different domains.

C. Supervision

Both the teacher and student models are trained under a self-supervised setting, where ground-truth keypoint correspondences are not available. The teacher model is pre-trained exclusively on sharp images and serves as an additional source of supervision for detection and description learning during the student model's training. Conversely, the student model is trained solely on blurred images and receives supervision from the teacher model's sharp detector and descriptor, in addition to the homography-based self-supervision. The overall loss for the teacher and student models is as follows:

$$\mathcal{L}_{tea} = \lambda_1\mathcal{L}_{det} + \lambda_2\mathcal{L}_{desc} + \lambda_3\mathcal{L}_{score} \tag{7}$$

$$\mathcal{L}_{stu} = \lambda_1\mathcal{L}_{det} + \lambda_2\mathcal{L}_{desc} + \lambda_3\mathcal{L}_{score} + \lambda_4\mathcal{L}_{DetKD} + \lambda_5\mathcal{L}_{TriKD} \tag{8}$$

where $\mathcal{L}_{det}$, $\mathcal{L}_{desc}$, $\mathcal{L}_{score}$, $\mathcal{L}_{DetKD}$, and $\mathcal{L}_{TriKD}$ are defined by Eqs. 1, 2, 3, 4, and 6, respectively. λ_1, λ_2, λ_3, λ_4, and λ_5 are hyperparameters that balance the contribution of each loss term. These losses facilitate the transfer of detection and description knowledge from sharp to blurred images, thus improving the student model performance under motion blur.

D. Implementation Details

Training data. We train our network on the GoPro dataset [36], which is widely used to evaluate image deblurring algorithms. Following their protocol, we use 3,214 images from 28 video sequences for training. Note that we solely use images, without any additional depth and camera labels, as our method is completely self-supervised. We apply data augmentation to the paired sharp and blurred images following the protocol of [24]. Specifically, we apply intensity perturbations and geometric transformations, and then resize all images to a resolution of 640×480.

Training details. We pre-train the teacher model on sharp images from GoPro and freeze its weights before training the student model on blurred images. We use the Adam optimizer [37] with an initial learning rate of 1×10^{-3} and train the model for 50 epochs, halving the learning rate after 30 epochs. The batch size is set to 32, and training is conducted on a single NVIDIA RTX 4090 GPU. The teacher model is randomly initialized, while the student model is initialized with the weights of the pre-trained teacher model on sharp GoPro images. We set the weights for the total training loss to $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 1$, $\lambda_4 = 1$, and $\lambda_5 = 2$ as defined in Eq. 8. The margins in the triplet losses, ϵ_1 and ϵ_2, are both set to 0.2.

IV. Experiments

Baselines. We categorize the baselines into four groups for performance comparison: (a) traditional hand-crafted local feature method SIFT [16]; (b) learning-based sparse local feature methods designed for sharp images SuperPoint [18],

TABLE I

QUANTITATIVE RESULTS ON THE HPATCHES [42] AND BLUR-HPATCHES [25] DATASETS. THE PROPOSED METHOD ACHIEVES THE BEST PERFORMANCE ON BOTH SHARP-TO-BLUR AND BLUR-TO-BLUR CONFIGURATIONS. THE 1ST, 2ND, AND 3RD -BEST METHODS ARE HIGHLIGHTED.

Methods	*Reference: Sharp* *Target: Sharp*				*Reference: Sharp* *Target: Blur*				*Reference: Blur* *Target: Blur*			
	Repeat. ↑	Cor-1 ↑	Cor-3 ↑	Cor-5 ↑	Repeat. ↑	Cor-1 ↑	Cor-3 ↑	Cor-5 ↑	Repeat. ↑	Cor-1 ↑	Cor-3 ↑	Cor-5 ↑
Sparse												
SIFT [16]	0.404	0.483	0.759	0.855	0.222	0.128	0.410	0.545	0.323	0.121	0.412	0.536
SuperPoint [18] + NN	0.503	0.202	0.428	0.478	0.351	0.024	0.181	0.333	0.335	0.019	0.150	0.297
SuperPoint [18] + LightGlue [40]	0.593	0.409	0.821	0.909	0.443	0.059	0.367	0.621	0.405	0.041	0.319	0.564
KP2D [24]	0.644	0.455	0.791	0.893	0.533	0.126	0.347	0.462	0.536	0.166	0.395	0.514
LANet [26]	0.666	0.500	0.686	0.762	0.523	0.209	0.543	0.597	0.549	0.244	0.545	0.619
BALF [25] + HardNet [41] + NN	0.488	0.255	0.667	0.822	0.323	0.026	0.322	0.584	0.346	0.055	0.331	0.548
Semi-Dense												
LoFTR [23]	0.935	0.597	0.852	0.914	0.548	0.240	0.566	0.659	0.552	0.241	0.581	0.686
AspanFormer [38]	0.589	0.379	0.479	0.484	0.552	0.176	0.417	0.450	0.557	0.172	0.409	0.445
MatchFormer [39]	0.476	0.398	0.466	0.481	0.430	0.216	0.452	0.467	0.444	0.226	0.453	0.472
BlurPoint (Ours)	**0.675**	**0.490**	**0.847**	**0.910**	**0.573**	**0.212**	**0.652**	**0.809**	**0.592**	**0.252**	**0.669**	**0.798**

TABLE II

QUANTITATIVE RESULTS ON THE DEBLUR-HPATCHES DATASET [25]. THE PROPOSED ONE-STAGE FRAMEWORK DIRECTLY ON BLURRED IMAGES HAS SUPERIOR PERFORMANCE COMPARED TO DELUR-THEN-EXTRACT METHODS.

Methods	SRN-DeblurNet [22]				DeblurGAN-v2 [21]			
	Repeat. ↑	Cor-1 ↑	Cor-3 ↑	Cor-5 ↑	Repeat. ↑	Cor-1 ↑	Cor-3 ↑	Cor-5 ↑
Sparse								
SIFT [16]	0.221	0.009	0.133	0.269	0.242	0.011	0.142	0.278
SuperPoint [18] + NN	0.215	0.000	0.026	0.131	0.226	0.000	0.055	0.172
SuperPoint [18] + LightGlue [40]	0.231	0.002	0.095	0.284	0.269	0.002	0.153	0.362
KP2D [24]	0.302	0.055	0.190	0.276	0.347	0.059	0.200	0.317
BALF [25] + HardNet [41] + NN	0.278	0.007	0.086	0.260	0.310	0.009	0.131	0.326
Semi-Dense								
LoFTR [23]	0.539	0.029	0.276	0.383	0.537	0.045	0.317	0.460
AspanFormer [38]	0.499	0.026	0.240	0.290	0.506	0.036	0.281	0.331
MatchFormer [39]	0.265	0.047	0.253	0.319 1	0.299	0.071	0.274	0.328
	Repeat. / Cor-1 / Cor-3 / Cor-5							
BlurPoint (Ours)	**0.484** / **0.098** / **0.333** / **0.541**							

KP2D [24], and LANet [26]; (c) semi-dense local feature matching methods designed for sharp images LoFTR [23], AspanFormer [38], and MatchFormer [39]; (d) learning-based local feature detection method designed for blurred images BALF [25]. For SuperPoint, we evaluate two versions using different matchers: (i) mutual nearest neighbor (NN) and (ii) the learning-based matcher LightGlue [40]. Since BALF is a detector-only method, we pair it with HardNet [41] as the descriptor for downstream evaluation.

A. Homography Estimation

We evaluate BlurPoint on homography estimation using the HPatches dataset [42], along with the synthesized Blur-HPatches and Deblur-HPatches datasets introduced in [25].

Dataset. The HPatches dataset consists of 59 sequences with viewpoint changes and 57 sequences with illumination changes. The Blur-HPatches and Deblur-HPatches datasets [25] are derived from the original HPatches dataset [42]. For detailed information on these two datasets, please refer to [25].

Evaluation protocol. Repeatability [43] is a widely used metric for evaluating the quality of keypoint detection. For descriptor evaluation, we compute the reprojected mean error (correctness) of the four image corners, using pixel thresholds of 1, 3, and 5, following [18]. All images are resized to 640 × 480 pixels, and up to 1,000 keypoints are extracted per image.

Results on the original HPatches dataset. To evaluate our method on sharp images, we compare it against other methods using the original HPatches dataset [42]. In this experiment, our network is trained on the GoPro dataset using both sharp and blurred images, while other methods use their own official pre-trained models. The results in Tab. I demonstrate that our method performs comparably to state-of-the-art local feature extraction methods on sharp images. Despite being trained on a relatively small dataset encompassing both sharp and blurred images, our network performs well on sharp images, with only ∼0.4% drop in the correctness-5 metric compared to LoFTR [23]. These findings underscore the robustness of our method, even though it is specifically designed to address motion blur.

Results on the Blur-HPatches dataset. To study the performance of our network on motion blurred images, we evaluate it against other methods using the synthesized Blur-HPatches dataset [25]. We simulate two different application scenarios for evaluation: *Sharp-to-Blur* and *Blur-to-Blur* settings using *EASY* blur subset of Blur-HPatches dataset [25]. As shown in Tab. I, baseline methods experience performance degradation on motion blurred images, with accuracy declining as the blur increases. Motion blurred images thus challenge those networks on local feature extraction task. In contrast, our method performs significantly better, benefiting from the implicit guidance of the teacher model on blurred images.

Results on the Deblur-HPatches dataset. We further examine local feature extraction performance on the Deblur-HPatches dataset with two objectives: (i) to determine whether deblurring networks can enhance feature extraction from blurred images, and (ii) to evaluate whether our one-stage method, which directly extracts local features from blurred images, outperforms two-stage approaches that first deblur and then extract. Specifically, we utilize the restored images from the Deblur-HPatches dataset, deblurred by SRN-DeblurNet [22] and DeblurGANv2 [21], to evaluate all the other local feature extraction methods. We use the *TOUGH* blur subset of Deblur-HPatches to simulate severe

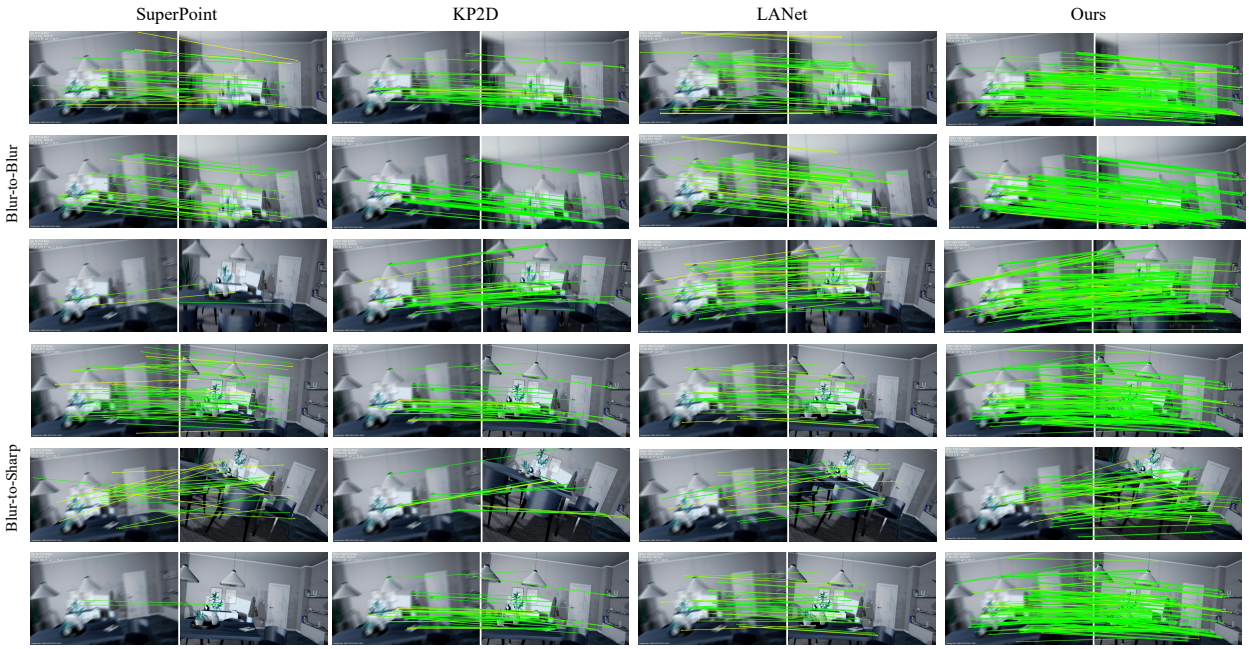

Fig. 4. **Qualitative Results for indoor relative pose estimation on the ArchVizInterior dataset [44]**. Green lines indicate epipolar error within 5×10^{-4} (normalized coordinates). BlurPoint consistently delivers more robust matching on *Blur-to-Blur* and *Blur-to-Sharp* configurations.

motion blur. Results in Tab. II indicate that deblurring networks can indeed assist in extracting keypoints from blurred images. However, they still fall short of our method, due to the limitations of image deblurring in handling severe motion blur. These results further support the effectiveness of a one-stage, motion blur aware feature extraction pipeline over two-stage *deblur-then-match* approaches.

B. Relative Pose Estimation

Dataset. The ArchVizInterior dataset [44] is rendered using the Unreal Engine and the free ArchVizInterior scene, simulating rapid back-and-forth camera shake. This dataset provides ground truth sharp images and corresponding camera poses, making it suitable for relative pose estimation tasks under motion blurred conditions. In addition, the dataset is challenging due to extensive texture-less regions and repetitive patterns. We exhaustively match all image pairs and randomly select 116 pairs for evaluation. All images are resized to 320×240 pixels, and up to 1,000 keypoints are extracted per image.

Evaluation protocol. Following [23], we report the area under the curve (AUC) of the pose error at thresholds 10°, 20°, and 30°. The predicted matches are used to estimate the essential matrix with RANSAC [45], and the relative poses are recovered using the five-point algorithm.

Results. The results in Tab. III show that our method outperforms all other state-of-the-art competitors, achieving an average AUC@30° improvement of 18%. As shown in Fig. 4, our model provides more reliable correspondences than [18], [24], [26], especially in regions affected by severe motion blur. In addition, our model produces more robust descriptors in motion blurred regions, further improving pose estimation. Both quantitative and qualitative results demonstrate the effectiveness of our approach, which learns sharp feature distributions from the teacher model and implicitly acquires motion blur aware local features.

TABLE III

EVALUATION ON THE ARCHVIZINTERIOR DATASET [44] FOR INDOOR POSE ESTIMATION. BLURPOINT IMPROVES THE STATE-OF-THE-ART METHODS BY A LARGE MARGIN.

Local features	*Reference: Blur Target: Blur*			*Reference: Blur Target: Sharp*		
	Pose Estimation AUC			Pose Estimation AUC		
	@10°	@20°	@30°	@10°	@20°	@30°
Sparse						
SIFT [16]	0.66	0.77	1.66	0.00	0.00	0.75
SuperPoint [18] + NN	0.73	1.73	3.78	0.97	2.74	4.21
SuperPoint [18] + LightGlue [40]	11.34	17.26	18.78	16.25	18.03	21.62
KP2D [24]	13.54	22.32	27.44	17.73	29.56	35.80
LANet [26]	3.12	6.47	10.48	6.34	15.04	20.09
BALF [25] + HardNet [41] + NN	0.79	3.54	5.79	0.69	2.64	4.72
Semi-Dense						
LoFTR [23]	13.25	19.96	24.05	18.28	25.73	30.07
AspanFormer [38]	2.47	18.17	29.81	3.15	20.32	31.48
JamMa [46]	12.27	21.28	25.59	24.13	33.74	37.41
BlurPoint (Ours)	**15.67**	**26.31**	**32.45**	**22.80**	**37.69**	**44.36**

C. Visual Localization

In addition to benefiting relative pose estimation, BlurPoint also achieves competitive performance in visual localization, a task that estimates the 6-DoF poses of query images with respect to a large-scale 3D scene model. We employ the Long-Term Visual Localization Benchmark [50] to assess methods under challenging conditions, such as day-night illumination changes, texture-less areas, and scene geometry changes. To further evaluate the performance of our method, we simulate motion blur on both query and database images.

Evaluation protocol. We evaluate BlurPoint on the

TABLE IV

VISUAL LOCALIZATION EVALUATION ON THE AACHEN DAY-NIGHT V1.0 BENCHMARK [47]. THE EVALUATION RESULTS ON BOTH THE LOCAL FEATURE EVALUATION TRACK AND THE FULL VISUAL LOCALIZATION TRACK ARE REPORTED.

Methods	Day	Night
	(0.25m, 2) / (0.5m, 5) / (1.0m,10)	
Sparse		
SIFT [16]	29.5 / 39.6 / 50.7	3.1 / 7.8 / 18.4
DISK [48] + NN	38.0 / 50.6 / 68.6	8.2 / 10.2 / 20.4
SuperPoint [18] + NN	35.7 / 48.7 / 71.1	4.1 / 10.2 / 26.5
SuperPoint [18] + LightGlue [40]	36.0 / 56.4 / 79.9	9.2 / 19.4 / 50.0
ALIKED [49] + LightGlue [40]	35.3 / 51.6 / 72.0	13.3 / 21.4 / 45.9
KP2D [24]	33.1 / 46.6 / 65.8	4.1 / 8.2 / 20.4
BALF [25] + HardNet [41] + NN	40.9 / 65.5 / 89.8	18.4 / 36.7 / 75.5
Semi-Dense		
LoFTR [23]	42.5 / 58.3 / 71.4	19.4 / 29.6 / 50.0
MatchFormer [39]	46.2 / 61.0 / 78.5	20.4 / 32.7 / 57.1
Aspanformer [38]	44.2 / 58.4 / 75.8	23.5 / 32.8 / 54.1
BlurPoint (Ours)	**69.2** / **82.5** / **91.1**	**35.7** / **53.1** / **76.5**

TABLE V

ABLATION STUDY. FIVE VARIANTS OF BLURPOINT ARE TRAINED AND EVALUATED ON THE BLUR-HPATCHES DATASET [25].

Methods	Repeat. ↑	Cor-1 ↑	Cor-3 ↑	Cor-5 ↑	MScore ↑
	Reference: Blur Target: Blur				
Student (Sharp & Blur)	0.541	0.169	0.610	0.750	0.380
Student (Blur only)	0.565	0.186	0.591	0.755	0.383
S-T + $\mathcal{L}_{DetKD}$	0.575	0.183	0.629	0.800	0.394
S-T + $\mathcal{L}_{TriKD}$	0.572	0.184	0.638	0.798	0.396
S-T + $\mathcal{L}_{DetKD}$+ $\mathcal{L}_{TriKD}$	**0.576**	**0.191**	**0.631**	**0.799**	**0.397**

Aachen Day-Night v1.0 benchmark [47], which includes 4,328 database images and 922 query images. We use NetVLAD [51] to generate the top 50 image pairs for each query image. We extract 400 keypoints from each image at a resolution of 320×240.

Results. As shown in Tab. IV, BlurPoint outperforms all baselines, demonstrating strong robustness under motion blur and day–night illumination changes. This is attributed to the use of knowledge distillation from sharp images, which guides the learning of blur aware local features.

D. Understanding Knowledge Transfer

Knowledge transfer visualization. To understand how the teacher model transfers knowledge to the student model, we visualize the matching details on the HPatches dataset [42], comparing the student model and the teacher model. In Fig. 5, we visualize the predicted matches from three models: the teacher model on sharp images, the baseline model on blurred images, and the teacher-guided student model on blurred images. The confidence scores for each keypoint are indicated by color, with high scores in red and low scores in blue. The results show that the student model closely follows the confidence score distribution of the teacher model, indicating that knowledge is effectively transferred and implicitly embedded into the student model.

Ablation study. To better understand the contribution of each component, we evaluate several variants of our model on the Blur-HPatches dataset [25], including: 1) a student model trained on both sharp and blurred images; 2) a student model trained on blurred images only; and 3) a full student-teacher framework (S-T) with both $\mathcal{L}_{DetKD}$ and $\mathcal{L}_{TriKD}$. As shown in Tab. V, both the feature divergence loss ($\mathcal{L}_{DetKD}$) and the triplet knowledge distillation loss ($\mathcal{L}_{TriKD}$) significantly enhance the knowledge transfer from the teacher model to the student model.

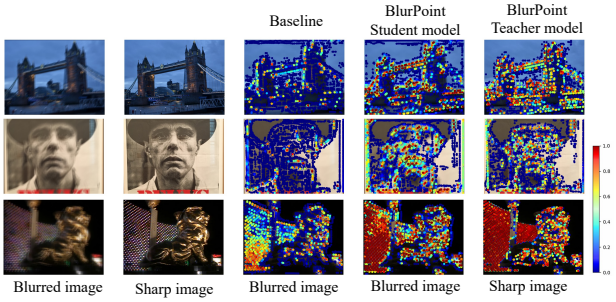

Fig. 5. **Visualization of keypoint score distribution change for better understanding student-teacher learning.** The teacher model not only guides the student model to find more salient keypoints, but also teaches the confidence score distribution to the student model.

V. CONCLUSIONS

In this paper, we introduce BlurPoint, a novel motion blur aware local feature learning method designed to improve joint detection and description performance on blurred images. We adopt a student-teacher framework, in which blurred images are implicitly supervised by both sharp images and self-supervised signals. We introduce two knowledge transfer losses: feature divergence loss, which aligns the feature location distributions from the teacher to the student detector, and triplet knowledge distillation loss, which transfers the sharp image descriptors to the blurred ones. Extensive experiments show that BlurPoint achieves superior performance on downstream tasks involving blurred images, while maintaining competitive results on sharp images.

ACKNOWLEDGMENTS

This work is supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd., a CUHK-led under the InnoHK scheme of Innovation and Technology Commission of the Hong Kong SAR. This work is also partially supported by the HK RGC AoE under AoE/E-407/24-N.

REFERENCES

[1] Z. Zhao, H. Yang, B. Liao, Y. Zeng, S. Yan, Y. Gu, P. Liu, Y. Zhou, H. Li, and J. Civera, "Advances in global solvers for 3d vision," *arXiv preprint arXiv:2602.14662*, 2026.

[2] M. Li, D. Li, S. Hu, K. Wang, Z. Zhao, and H. Wang, "Slam-x: Generalizable dynamic removal for nerf and gaussian splatting slam," in *Proc. ACM Int. Conf. Multimedia*, 2025, pp. 1132–1140.

[3] Z. Wang, Y. Sun, H. Wang, B. Jing, X. Shen, X. Dong, Z. Hao, H. Xiong, and Y. Song, "Reasoning-enhanced domain-adaptive pre-training of multimodal large language models for short video content moderation," *arXiv e-prints*, pp. arXiv–2509, 2025.

[4] F. Zhu, Y. Zhao, Z. Chen, B. Yu, and H. Zhu, "Fgo-slam: Enhancing gaussian slam with globally consistent opacity radiance field," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2025, pp. 11 075–11 081.

[5] J. Zhang and G. Lu, "Underground mapping and localization based on ground-penetrating radar," in *Proc. Asian Conf. Comput. Vis.*, 2024, pp. 2018–2033.

[6] Z. Ke, J. Shen, X. Zhao, X. Fu, Y. Wang, Z. Li, L. Liu, and H. Mu, "A stable technical feature with gru-cnn-ga fusion," *Applied Soft Computing*, p. 114302, 2025.

[7] S. Bao, Q. Xu, F. Li, B. Han, Z. Yang, X. Cao, and Q. Huang, "Towards size-invariant salient object detection: A generic evaluation and optimization approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025.

[8] Z. Ke, Y. Cao, Z. Chen, Y. Yin, S. He, and Y. Cheng, "Early warning of cryptocurrency reversal risks via multi-source data," *Finance Research Letters*, p. 107890, 2025.

[9] W. Li, B. Hu, R. Shao, L. Shen, and L. Nie, "Lion-fs: Fast & slow video-language thinker as online video assistant," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 3240–3251.

[10] Z. Li, Z. Chen, H. Wen, Z. Fu, Y. Hu, and W. Guan, "Encoder: Entity mining and modification relation binding for composed image retrieval," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 5, 2025, pp. 5101–5109.

[11] S. Zhou, X. Zhang, X. Chu, B. Zhang, Z. Zhao, and X. Lu, "Fastpillars: A deployment-friendly pillar-based 3d detector," *IEEE Trans. Circuits Syst. Video Technol.*, 2025.

[12] Y. Dong, C. Kang, J. Zhang, Z. Zhu, Y. Wang, X. Yang, H. Su, X. Wei, and J. Zhu, "Benchmarking robustness of 3d object detection to common corruptions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 1022–1032.

[13] X. Meng, P. Hou, Z. Zhao, J. Civera, D. Cremers, H. Wang, and H. Li, "Dream-slam: Dreaming the unseen for active slam in dynamic environments," *arXiv preprint arXiv:2602.21967*, 2026.

[14] J. Guo, J. Wang, D. Kang, W. Dong, W. Wang, and Y.-h. Liu, "Free-surgs: Sfm-free 3d gaussian splatting for surgical scene reconstruction," in *MICCAI*. Springer, 2024, pp. 350–360.

[15] J. Guo, T. Guan, W. Dong, W. Zheng, W. Wang, Y. Wang, Y. Yam, and Y.-H. Liu, "Salon3r: Structure-aware long-term generalizable 3d reconstruction from unposed images," *arXiv preprint arXiv:2510.15072*, 2025.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[17] F. Bellavia, Z. Zhao, L. Morelli, and F. Remondino, "Image matching filtering and refinement by planes and beyond," *arXiv preprint arXiv:2411.09484*, 2024.

[18] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Worksh.*, 2018, pp. 337–33 712.

[19] J. Edstedt, G. Bökman, and Z. Zhao, "Dedode v2: Analyzing and improving the dedode keypoint detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Worksh.*, 2024, pp. 4245–4253.

[20] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 8183–8192.

[21] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8877–8886.

[22] X. Tao, H. Gao, Y. Wang, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 8174–8182.

[23] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 8922–8931.

[24] J. Tang, H. Kim, V. Guizilini, S. Pillai, and R. Ambrus, "Neural outlier rejection for self-supervised keypoint learning," *arXiv preprint arXiv:1912.10615*, 2019.

[25] Z. Zhao, "Balf: Simple and efficient blur aware local feature detector," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 3362–3372.

[26] C. Wang, G. Zhang, Z. Cheng, and W. Zhou, "Rethinking low-level features for interest point detection and description," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 2059–2074.

[27] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 769–777.

[28] J. Wang, Y. Chen, Z. Zheng, X. Li, M.-M. Cheng, and Q. Hou, "Crosskd: Cross-head knowledge distillation for dense object detection," *arXiv preprint arXiv:2306.11369*, 2023.

[29] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2604–2613.

[30] C. Doersch, P. Luc, Y. Yang, D. Gokay, S. Koppula, A. Gupta, J. Heyward, I. Rocco, R. Goroshin, J. Carreira *et al.*, "Bootstap: Bootstrapped training for tracking-any-point," in *Proc. Asian Conf. Comput. Vis.*, 2024, pp. 3257–3274.

[31] R. Mao, C. Bai, Y. An, F. Zhu, and C. Lu, "3dg-stfm: 3d geometric guided student-teacher feature matching," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 125–142.

[32] J. Wu, R. Xu, Z. Wood-Doughty, and C. Wang, "Segment anything model is a good teacher for local feature learning," *arXiv preprint arXiv:2309.16992*, 2023.

[33] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 815–823.

[34] J. Tang, J. Folkesson, and P. Jensfelt, "Geometric correspondence network for camera motion estimation," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 1010–1017, 2018.

[35] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[36] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 257–265.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[38] H. Chen, Z. Luo, L. Zhou, Y. Tian, M. Zhen, T. Fang, D. Mckinnon, Y. Tsin, and L. Quan, "Aspanformer: Detector-free image matching with adaptive span transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 20–36.

[39] Q. Wang, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "Matchformer: Interleaving attention in transformers for feature matching," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 2746–2762.

[40] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17 627–17 638.

[41] P. Chao, C.-Y. Kao, Y.-S. Ruan, C.-H. Huang, and Y.-L. Lin, "Hardnet: A low memory traffic network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3552–3561.

[42] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 3852–3861.

[43] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, 2005.

[44] P. Liu, X. Zuo, V. Larsson, and M. Pollefeys, "Mba-vo: Motion blur aware visual odometry," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5550–5559.

[45] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1981.

[46] X. Lu and S. Du, "Jamma: Ultra-lightweight local feature matching with joint mamba," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 14 934–14 943.

[47] Z. Zhang, T. Sattler, and D. Scaramuzza, "Reference pose generation for long-term visual localization via learned features and view synthesis," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 821–844, 2021.

[48] M. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 14 254–14 265.

[49] X. Zhao, X. Wu, W. Chen, P. C. Chen, Q. Xu, and Z. Li, "Aliked: A lighter keypoint and descriptor extraction network via deformable transformation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–16, 2023.

[50] C. Toft, W. Maddern, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, T. Pajdla *et al.*, "Long-term visual localization revisited," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2074–2088, 2020.

[51] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 5297–5307.