

CasP: Improving Semi-Dense Feature Matching Pipeline Leveraging Cascaded Correspondence Priors for Guidance

Peiqi Chen^{1*} Lei Yu^{2*} Yi Wan^{1†} Yingying Pei¹ Xinyi Liu¹ Yongxiang Yao¹
Yingying Zhang² Lixiang Ru² Liheng Zhong² Jingdong Chen² Ming Yang² Yongjun Zhang^{1†}
¹Wuhan University ²Ant Group

Abstract

Semi-dense feature matching methods have shown strong performance in challenging scenarios. However, the existing pipeline relies on a global search across the entire feature map to establish coarse matches, limiting further improvements in accuracy and efficiency. Motivated by this limitation, we propose a novel pipeline, CasP, which leverages cascaded correspondence priors for guidance. Specifically, the matching stage is decomposed into two progressive phases, bridged by a region-based selective cross-attention mechanism designed to enhance feature discriminability. In the second phase, one-to-one matches are determined by restricting the search range to the one-to-many prior areas identified in the first phase. Additionally, this pipeline benefits from incorporating high-level features, which helps reduce the computational costs of low-level feature extraction. The acceleration gains of CasP increase with higher resolution, and our lite model achieves a speedup of $\sim 2.2\times$ at a resolution of 1152 compared to the most efficient method, ELoFTR. Furthermore, extensive experiments demonstrate its superiority in geometric estimation, particularly with impressive cross-domain generalization. These advantages highlight its potential for latency-sensitive and high-robustness applications, such as SLAM and UAV systems. Code is available at <https://github.com/pq-chen/CasP>.

1. Introduction

Local feature matching is a fundamental task in 3D computer vision that aims to establish correspondences within each image pair. This technique is crucial for accurate geometric estimation and supports a wide range of downstream applications, including structure-from-motion [15, 19, 28] and visual localization [24, 26, 35]. In particular, real-time processing tasks, such as SLAM and UAV systems, demand high computational efficiency and robustness. The classi-

*Equal contribution. †Corresponding author.

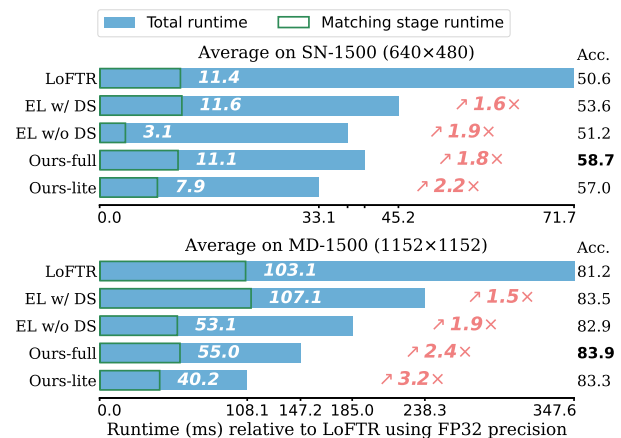


Figure 1. **Matching Accuracy and Efficiency Comparisons.** The runtime and AUC@20° accuracy are measured across two benchmarks and compared with LoFTR and ELoFTR (EL). The acceleration gains of ELoFTR diminish with increasing resolution since the matching stage occupies a significant portion of runtime. The novel cascaded matching pipeline is more efficient and robust than merely removing the dual-softmax (DS) operator.

cal pipeline adopted by sparse methods consists of feature detection followed by feature description. However, its success largely depends on the detector’s ability [9, 21, 37] to generate repeatable key points. Despite proposals to employ graph neural networks [20, 25] to enhance off-the-shelf local features, reliable detection remains unguaranteed, especially in areas with low texture and repetitive patterns.

To tackle this issue, LoFTR [32] proposes a semi-dense feature matching pipeline that treats each token in the coarse feature map as a potential matching candidate, thereby replacing the feature detection stage. LoFTR enhances robustness under such challenging scenarios by leveraging texture and relative position cues. As a trade-off, the dense interactions among numerous tokens lead to substantial computational costs compared to sparse methods.

Follow-up work [5, 6, 34] primarily focused on addressing the limited representational capacity of LoFTR by intro-

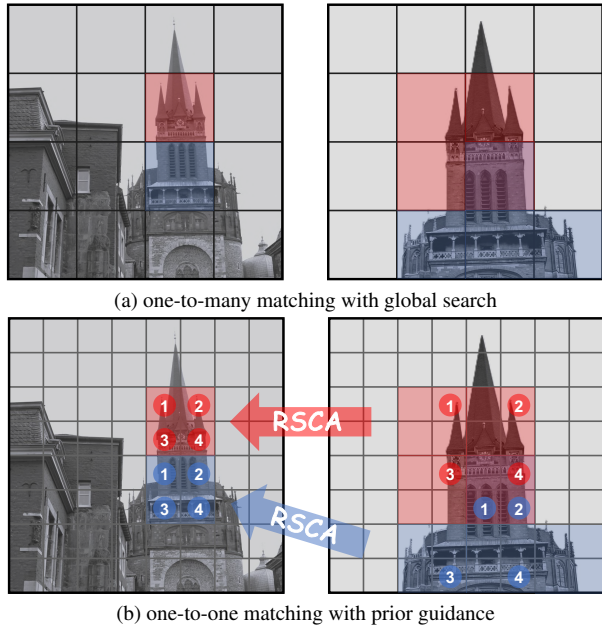


Figure 2. **Schematic Diagram of Cascaded Matching.** (a) One-to-many correspondence priors are selected with global search at a coarse scale and shown as same-colored patches (purple denotes potential common priors). (b) One-to-one matches are determined with prior guidance at the target scale and shown as same-numbered patches, with a region-based attention mechanism, RSCA, applied only at prior token positions.

ducing enhanced feature interaction modules, often at the cost of increased runtime. Recently, ELoFTR [40] incorporated an aggregated attention mechanism that operates on adaptively selected tokens to improve efficiency. However, we observe that the matching pipeline used by these methods may impose a bottleneck on further advancement. Specifically, the feature maps processed during the matching stage contain excessive tokens, leading to potential latency issues. As shown in Fig. 1, we evaluate the runtime of our model across two benchmarks at different resolutions and compare it with LoFTR and ELoFTR, with the accuracy of $AUC@20^\circ$ reported. As the resolution increases, the matching stage (□) of the two methods consistently occupies a substantial portion of the total runtime (●), which diminishes the expected acceleration gains of ELoFTR. ELoFTR provides a solution by removing the dual-softmax operator, but this leads to a notable accuracy drop across both benchmarks. In contrast, our method achieves significant speedup across all resolutions while also delivering enhanced accuracy, driven by the shift to a more efficient pipeline.

The core guideline for acceleration is to defer primary operations to a coarser scale wherever feasible, thereby reducing the number of tokens processed. To achieve this, we propose a cascaded matching pipeline, CasP, which decomposes the matching stage into two progressive phases.

As shown in Fig. 2, the pipeline first establishes one-to-many correspondences at a coarser scale as cascaded priors. Then, one-to-one matches at the target scale are determined by leveraging these priors for guidance. The acceleration gains stem from two key factors: **1)** Instead of conducting a global search across the entire feature map, the second phase focuses only on tokens within the prior areas, eliminating irrelevant computations outside these areas. **2)** Incorporating high-level features helps reduce the computational costs associated with low-level feature extraction. To ensure more reliable matching confidence, we introduce a region-based selective cross-attention (RSCA) mechanism between the two phases of cascaded matching to enhance feature discriminability among prior candidates. Furthermore, our pipeline adopts a training-inference decoupling strategy, which enhances model representational capacity during training and maximizes inference efficiency.

Building upon the cascaded matching pipeline, we propose a novel semi-dense method integrating advanced modules for enhanced matching accuracy and efficiency. We present two versions of this method, which differ in the number of channels used for low-level feature extraction. Our lite model achieves a speedup of $\sim 2.2\times$ and $3.2\times$ at a resolution of 1152 compared to ELoFTR and LoFTR, respectively. Furthermore, an additional boost is attainable by using FP16 precision. In terms of accuracy, our full model achieves state-of-the-art performance in extensive experiments. In particular, the ablation study demonstrates the significant improvement of CasP in cross-domain generalization, which underscores the practical effectiveness of our methods.

Our contributions are summarized as follows:

- A novel pipeline that leverages cascaded correspondence priors to address the existing efficiency bottleneck.
- A novel attention mechanism that focuses on prior areas to bridge the two phases of cascaded matching.
- A novel semi-dense method that integrates advanced modules to deliver superior performance, with strong efficiency and cross-domain generalization for practical applications.

2. Related Work

Efficient Matching Strategy. Sparse methods [4, 25, 31] control computational efficiency by adjusting the number of key points extracted by detectors [9, 14, 21, 37, 41, 42]. LightGlue [20] proposed a pruning scheme that adapts to the varying complexities of image pairs. As the first semi-dense method, LoFTR [32] employed linear attention [17] to ensure manageable computational costs for dense interactions. TopicFM [13] introduced a topic-assisted approach, enabling indirect interactions between tokens and fixed-size latent topics. EcoMatcher [7], a recently proposed non-transformer-based method, leveraged context clusters to facilitate point-wise interactions with selected anchors. Notably, ELoFTR [40] adopted a lightweight convolutional

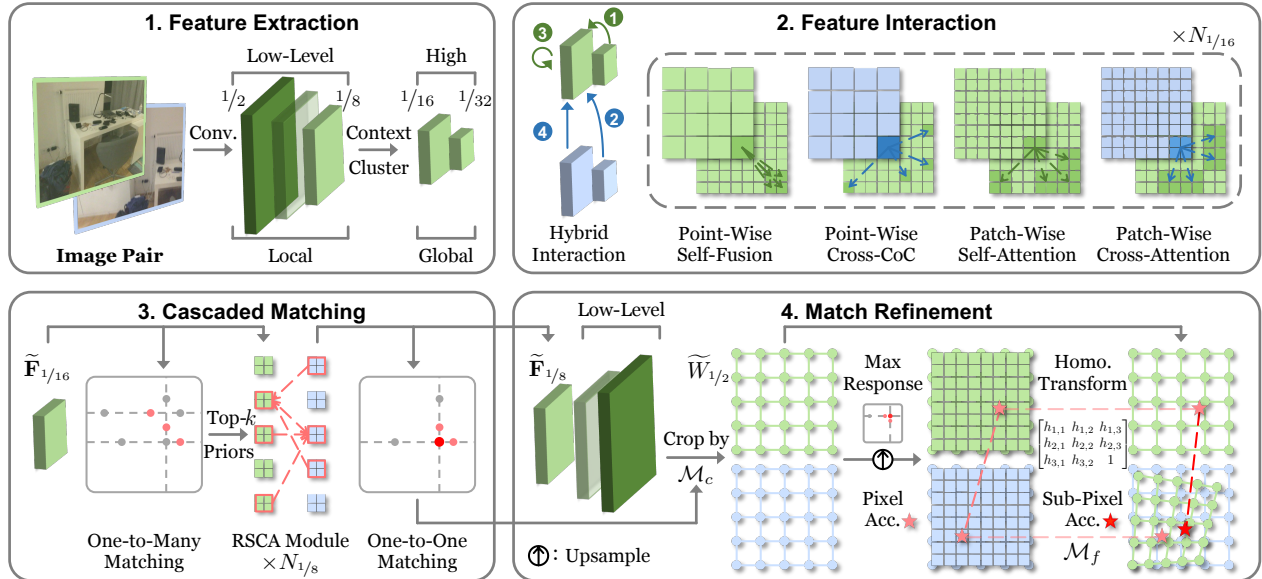


Figure 3. **Pipeline Overview.** (1) Low-level ($1/2$ to $1/8$) and high-level ($1/16$ to $1/32$) feature maps are extracted for local and global descriptions, respectively. (2) High-level feature maps are transformed by a hybrid interaction module composed of sub-modules that operate at different scales to enhance complementarity. (3) A cascaded matching module establishes one-to-one correspondences at the $1/8$ scale, with the search range constrained by one-to-many priors identified at the $1/16$ scale. (4) A two-step homography-based refinement module is applied progressively to reach pixel-level and subpixel-level accuracy.

neural network (CNN), RepVGG [10], for feature extraction and incorporated an aggregated attention module to perform transformers on reduced tokens.

Multi-Level Matching Strategy. GLU-Net [36] combined global and local correlation layers to achieve robust and accurate dense correspondence predictions. Building on this global-local architecture, ASpanFormer [5] and AffineFormer [6] introduced a multi-level cross-attention mechanism into the semi-dense pipeline. MatchFormer [38] proposed an extract-and-match approach that interleaves self- and cross-attention layers at each feature extraction stage. QuadTree [34] employed a quadtree-based attention mechanism to focus on relevant regions. PATS [23] adopted a hierarchical framework that sequentially adds and trains each corresponding network at different resolutions. CasMTR [3] incorporated cascade modules after freezing the existing method’s feature encoder and coarse attention modules. However, both QuadTree and CasMTR aimed to obtain finer-grained matches based on established coarse correspondences, albeit at the cost of increased runtime.

3. Method

Given a pair of grayscale images, I^A and I^B , the existing semi-dense pipeline directly establishes one-to-one matches at a scale of $1/8$ and refines them to sub-pixel accuracy. By contrast, our proposed cascaded pipeline improves efficiency by performing primary operations at a coarser scale of

$1/16$ to establish one-to-many correspondence priors, which then guide the subsequent one-to-one matching stage. An overview of our pipeline is shown in Fig. 3.

3.1. Feature Extraction

Low-Level Local Features. A lightweight CNN initially extracts low-level feature maps at scales ranging from $1/2$ to $1/8$ to capture local cues. The existing pipeline performs a global search across feature maps at the $1/8$ scale to determine coarse matches, which requires a sufficient number of channels to ensure global feature discriminability. However, this strategy imposes significant throughput bottlenecks for high-resolution inputs because of rapidly increasing computational costs. Leveraging the proposed cascaded pipeline, our models adopt a modified RepVGG [10] architecture with a reduced number of parameters, as shown in Tab. 1.

High-Level Global Features. Since we defer primary operations to a coarser scale, additional down-sampled feature maps, $\mathbf{F}_{1/16}^A$, $\mathbf{F}_{1/16}^B$, $\mathbf{F}_{1/32}^A$, and $\mathbf{F}_{1/32}^B$, are required for subsequent interaction and matching stages. Rather than applying convolutions for local description, we follow EcoMatcher [7] and employ the context cluster mechanism [22], referred to as self-CoC, to extract high-level features and enhance contextual understanding with a global receptive field.

The self-CoC module utilizes selected anchors A as proxies to enable indirect point-wise interactions among all feature points P . Specifically, it consists of three main stages:

Method	Type	#Channels	#Blocks	#Params (M)
LoFTR [32]	ResNet	[128,196,256]	[2,2,2]	5.9
ELoFTR [40]	RepVGG	[64,128,256]	[2,4,14]	9.5
Ours-full	RepVGG	[64,128,192]	[2,4,4]	2.0
Ours-lite	RepVGG	[64,64,128]	[2,4,4]	0.8

Table 1. **Comparison of Low-Level Feature Extraction.** Our models adopt an efficient design benefiting from the novel pipeline.

Clustering, Aggregating, and Dispatching. In the *Clustering* stage, each point is allocated to the most similar anchor, forming corresponding clusters C . The *Aggregating* stage then updates anchors by aggregating the points within the same clusters. Finally, the *Dispatching* stage propagates contextual information back from anchors to points, completing a round of message exchange. Formally, both points P and anchors A are linearly projected into the similarity and value spaces. The operations for each stage are given by:

$$S = \text{sim}(P^s, A^s), \quad i \in C[j] \Leftrightarrow j = \arg \max S[i, :], \quad (1)$$

$$\hat{A}^v[j] = \frac{A^v[j] + \sum_{k \in C[j]} S[k, j] \cdot P^v[k]}{1 + \sum_{k \in C[j]} S[k, j]}, \quad (2)$$

$$\hat{P}^v[i] = \text{sigmoid}(S[i, j]) \cdot \hat{A}^v[j], \quad (3)$$

where superscripts s and v denote the respective spaces, and $\text{sim}(\cdot, \cdot)$ measures pair-wise similarity. The computational cost remains manageable by controlling the number of A .

3.2. Feature Interaction

Following multi-level feature extraction, the interaction stage incorporates cross-view cues to strengthen the similarity of each corresponding token pair. We introduce a hybrid module that comprises two complementary mechanisms.

Attention Mechanism. As the core mechanism of transformers, attention models the point-wise relationships among all involved tokens by measuring the similarities between queries Q and keys K and then obtaining a weighted average of values V , which can be generally formulated as:

$$\text{Attn}(Q, K, V) = \text{Softmax}(\sigma_q(Q)\sigma_k(K)^T)\sigma_v(V). \quad (4)$$

Although the interaction stage is deferred to the feature maps at the $1/16$ scale, vanilla attention, where σ_q , σ_k and σ_v are identity mappings, still incurs high computational costs. Inspired by ELoFTR [40], aggregated attention is employed to down-sample tokens into patches by setting σ_q as a depth-wise convolution layer and σ_k and σ_v as max-pooling layers, with kernel size and stride both set to 2. Consequently, the actual interaction is conducted at the $1/32$ scale.

Cross-CoC Mechanism. The down-sampling of the original tokens in aggregated attention serves as an effective strategy for reducing complexity. However, it sacrifices point-to-point modeling in favor of a patch-to-patch approach.

This trade-off may compromise the ability to capture token-level details and potentially impact performance. To address this limitation, we adopt the cross-CoC mechanism from EcoMatcher [7], which utilizes coarser-grained tokens from $\mathbf{F}_{1/32}^A$ and $\mathbf{F}_{1/32}^B$ as the selected anchors \hat{A}^v in Eq. (3). This process facilitates indirect point-wise interactions at the $1/16$ scale, thereby complementing aggregated attention at that processing scale. Moreover, a fusion module is incorporated to enable the exchange of local information across feature maps at the $1/16$ and $1/32$ scales.

Hybrid Interaction Module. The hybrid interaction module is constructed according to the order specified in Fig. 3 and is repeated $N_{1/16}$ times to generate the transformed feature maps, $\tilde{\mathbf{F}}_{1/16}^A$ and $\tilde{\mathbf{F}}_{1/16}^B$. This module not only maximizes computational efficiency but also enhances representational capacity by enabling interactions across different scales.

3.3. Cascaded Matching

Existing semi-dense methods [5–7, 34, 38] typically follow the LoFTR [32] pipeline and apply a dual-softmax (DS) operator across both dimensions of the score matrix to filter out low-confidence matches. As evidenced by Fig. 1, this stage significantly increases the runtime, particularly for high-resolution inputs. ELoFTR [40] offers a simple solution by directly using the raw score matrix. While this strategy improves efficiency, it compromises robustness and generalizability. AdaMatcher [16] introduces a many-to-one assignment mechanism to address scale inconsistencies, but not for efficiency. To pursue a comprehensive solution, we propose a cascaded matching module, as illustrated in Fig. 4.

One-to-Many Matching. We first construct the score matrix $S_{1/16}$ from the correlations between $\tilde{\mathbf{F}}_{1/16}^A$ and $\tilde{\mathbf{F}}_{1/16}^B$. Our objective is to derive the top- k correspondence priors $\pi_{1/16}$ for each token in both views, which are defined as:

$$\pi_{1/16}^A = \arg \max_k (S_{1/16}), \quad \pi_{1/16}^B = \arg \max_k (S_{1/16}^T). \quad (5)$$

Assuming one-to-one matches at the $1/8$ scale, we set $k \geq 4$ because each token at the $1/16$ scale may correspond to at most $(16/8)^2$ tokens in the cross-view feature map. During **training**, we apply a DS operator as a differentiable matching layer to $S_{1/16}$, yielding distinctive feature representations and the confidence matrix $P_{1/16}$ for supervision. In addition, we inject one-to-many ground-truth correspondences into $P_{1/16}$ to accelerate convergence in the subsequent one-to-one matching. During **inference**, the DS operator is omitted because $S_{1/16}$ alone suffices for the top- k selection.

Region-Based Selective Cross-Attention Mechanism. Prior to the one-to-one matching stage, the previously extracted feature map at the $1/8$ scale is fused with the transformed feature map $\tilde{\mathbf{F}}_{1/16}$ to inherit cross-view cues, thereby producing $\tilde{\mathbf{F}}_{1/8}$. Subsequently, we introduce a region-based

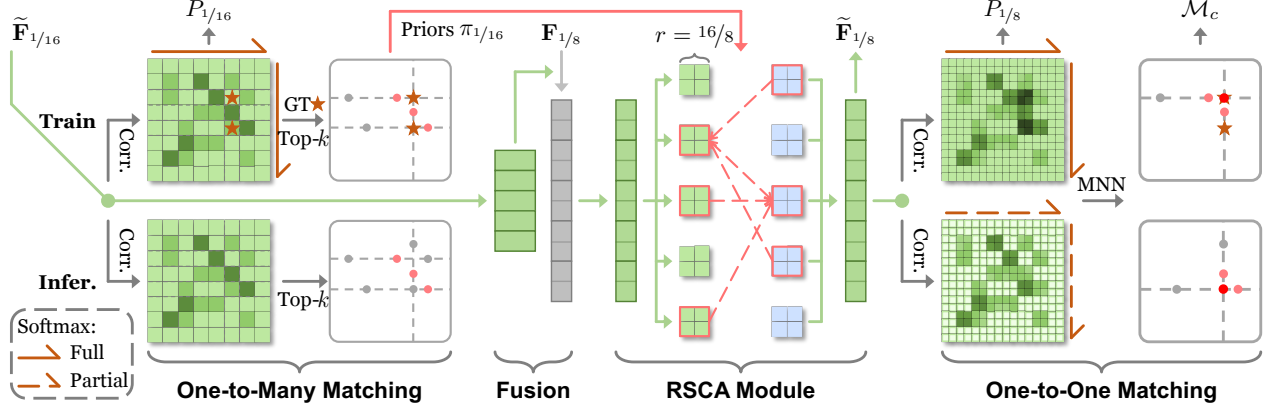


Figure 4. **Training-Inference Decoupled Cascaded Matching.** During **training**, the top- k priors selected by one-to-many matching include ground-truth correspondences for RSCA learning, and the DS operator is applied in both matching stages for supervision. During **inference**, one-to-many matching omits this step, whereas one-to-one matching applies partial softmax instead of the full version.

attention mechanism, RSCA, which allows each token to attend selectively to its correspondence priors in $\pi_{1/16}$.

Specifically, we illustrate this by computing attention from $\tilde{\mathbf{F}}^B \in \mathbb{R}^{H^B \times W^B \times c}$ to $\tilde{\mathbf{F}}^A \in \mathbb{R}^{H^A \times W^A \times c}$, together with the correspondence priors $\pi^A \in \mathbb{N}^{(H^A W^A / r^2) \times k}$ ($r = 16/8$ in our case). Here we omit the subscript for simplicity, and the messages from RSCA are computed as follows:

$$\hat{\mathbf{F}}^l = \text{Split}_r(\tilde{\mathbf{F}}^l) \in \mathbb{R}^{(H^l W^l / r^2) \times r^2 \times c}, \quad l \in \{A, B\}, \quad (6)$$

$$\hat{\mathbf{F}}^{A \leftarrow B} = \hat{\mathbf{F}}^B[\pi^A] \in \mathbb{R}^{(H^A W^A / r^2) \times k r^2 \times c}, \quad (7)$$

$$\hat{\mathbf{m}}^{A \leftarrow B} = \text{Attn}(\hat{\mathbf{F}}^A, \hat{\mathbf{F}}^{A \leftarrow B}, \hat{\mathbf{F}}^{A \leftarrow B}), \quad (8)$$

$$\mathbf{m}^{A \leftarrow B} = \text{Merge}_r(\hat{\mathbf{m}}^{A \leftarrow B}) \in \mathbb{R}^{H^A \times W^A \times c}, \quad (9)$$

where $\text{Split}_r(\cdot)$ partitions the input into cells of size $r \times r$, and $\text{Merge}_r(\cdot)$ performs the reverse operation. In Eq. (8), the length of each query is r^2 , and that of each key/value is $k \cdot r^2$. $\tilde{\mathbf{F}}^A$ is then updated using a feed-forward network that incorporates a convolution for locality modeling, which compensates for the absence of self-attention mechanism:

$$\tilde{\mathbf{F}}^A = \tilde{\mathbf{F}}^A + \text{FFN}(\tilde{\mathbf{F}}^A, \mathbf{m}^{A \leftarrow B}), \quad (10)$$

$$\text{FFN}(\tilde{\mathbf{F}}, \mathbf{m}) = \text{Conv}(\text{GeLU}(\text{Linear}([\tilde{\mathbf{F}} \parallel \mathbf{m}]))), \quad (11)$$

where $[\cdot \parallel \cdot]$ represents the concatenation operation. We repeat the RSCA module $N_{1/8}$ times, as shown in Fig. 3, to enhance feature discriminability among tokens at prior positions.

One-to-One Matching. Similarly, the score matrix $S_{1/8}$ is obtained from the correlations, and a DS operator is applied to provide supervision during **training**. During **inference**, the softmax operator is applied only to the key/value tokens that are attended to in the RSCA module for each query token, while the remaining ones are omitted and set to zero. This process, referred to as partial softmax, significantly reduces computational costs. The resulting confidence matrix

$P_{1/8}$ is defined as follows:

$$\text{PartialSoftmax}(\mathbf{x}, \pi)[i] = \frac{\chi_\pi(i) \exp(\mathbf{x}[i])}{\sum_{k \in \pi} \exp(\mathbf{x}[k])}, \quad (12)$$

$$P_{1/8}[i, j] = \text{PartialSoftmax}(S_{1/8}[i], \phi_r(\pi_{1/16}^A)[i])[j] \odot \text{PartialSoftmax}(S_{1/8}^T[j], \phi_r(\pi_{1/16}^B)[j])[i], \quad (13)$$

where $\chi_\pi(\cdot)$ denotes the indicator function and $\phi_r(\cdot)$ maps each position at the $1/16$ scale to r^2 corresponding positions at the $1/8$ scale. Coarse matches are filtered based on a pre-defined threshold θ over $P_{1/8}$. The mutual-nearest-neighbor (MNN) criterion is then applied for one-to-one matching, forming \mathcal{M}_c . Note that Eq. (12) implies that each selected match (i, j) must be derived from the correspondence priors on both sides, which can be formulated as follows:

$$j \in \phi_r(\pi_{1/16}^A)[i] \quad \text{and} \quad i \in \phi_r(\pi_{1/16}^B)[j]. \quad (14)$$

3.4. Match Refinement

For each match in \mathcal{M}_c , a point-to-point correspondence is established at the $1/8$ scale, along with a patch-to-patch correspondence at the original resolution. To refine coarse matches, local patches are first extracted, followed by a two-stage homography-based module for sub-pixel accuracy.

Local Patch Extraction. As with obtaining $\tilde{\mathbf{F}}_{1/8}$, the feature maps at the $1/2$ and $1/4$ scales are progressively fused in an FPN-like manner. Local patches at the $1/2$ scale, denoted as $\tilde{\mathbf{W}}_{1/2}$, are then cropped using a $w \times w$ window centered on each coarse match for subsequent refinement.

Two-Stage Homography. In the first stage, we upsample $\tilde{\mathbf{W}}_{1/2}$ to the original resolution to establish pixel-level correspondences by selecting the maximum response from the correlations. Rather than employing the regression-by-expectation strategy [32, 40] to achieve sub-pixel accuracy,

Category	Method	AUC on MD-1500 \uparrow			AUC on SN-1500 \uparrow			Average Runtime (ms) \downarrow	
		@5°	@10°	@20°	@5°	@10°	@20°	on MD-1500	on SN-1500
Sparse	SP [9] + SG [25]	49.7	67.1	80.6	17.4	33.9	49.5	51.9 + 72.0	36.7 + 72.0
	SP [9] + LG [20]	49.9	67.0	80.1	17.7	34.6	51.2	51.9 + 44.0	36.7 + 44.0
Semi-Dense	LoFTR [32]	52.8	69.2	81.2	16.9	33.6	50.6	347.6	71.7
	QuadTree [34]	54.6	70.5	82.2	19.0	37.3	53.5	506.4	128.5
	ASpanFormer [5]	55.3	71.5	83.1	19.6	37.7	54.4	414.0	92.6
	ELoFTR [40]	56.4	72.2	83.5	19.2	37.0	53.6	238.3 / 158.8	45.2 / 36.5
	AffineFormer \dagger [6]	57.3	72.8	84.0	22.0	40.9	58.0	≥ 347.6	≥ 71.7
	Ours-full	57.1	72.7	83.9	23.0	41.6	58.7	147.2 / 83.8	40.1 / 32.3
	Ours-lite	55.6	71.7	83.3	21.6	40.1	57.0	108.1 / 67.7	33.1 / 30.7
Dense	DKM [11]	60.4	74.9	85.1	26.6	47.1	64.2	1355.6	414.8
	ROMA [12]	62.6	76.7	86.3	28.9	50.4	68.3	1482.5	493.2

Table 2. **Relative Pose Estimation Results on MD-1500 and SN-1500 with Standard RANSAC.** All methods are evaluated using a model trained on outdoor scenes. The AUCs of errors up to 5°, 10°, and 20°, and the average runtime, are reported. For ELoFTR, we compare the runtime using FP32/FP16 precision with its full model. \dagger denotes that the runtime is inferred from the paper since the code is unavailable.

we draw inspiration from HomoMatcher [39] and model the transformations between $\widetilde{\mathbf{W}}_{1/2}^A$ and $\widetilde{\mathbf{W}}_{1/2}^B$ as rigid homographies, disregarding deformable regions. Unlike HomoMatcher’s use of a fixed central location, our method leverages pixel-level correspondences as well-estimated initial positions to enhance accuracy.

3.5. Supervision

Coarse Supervision. We first construct the one-hot 4D ground truth matrix $M_{1/8}^{gt}$ at the $1/8$ scale by establishing one-to-one correspondences between I^A and I^B using camera poses and depth maps. We then extract the supervision set $\mathcal{M}_{1/8}^{gt}$ by selecting non-zero element positions and converting them into index pairs. Next, we downsample $M_{1/8}^{gt}$ via max-pooling across each dimension to obtain $M_{1/16}^{gt}$ and form $\mathcal{M}_{1/16}^{gt}$ in the same manner. Finally, we define the coarse loss as the negative log-likelihood on the confidence matrix $P_{1/s}$, where $s \in \{8, 16\}$, as follows:

$$L_{1/s}^c = -\frac{1}{\#\mathcal{M}_{1/s}^{gt}} \sum_{(i,j) \in \mathcal{M}_{1/s}^{gt}} \log(P_{1/s}[i, j]). \quad (15)$$

Fine Supervision. The pixel-level supervision set $\mathcal{M}_{1/1}^{gt}$ and loss $L_{1/1}^f$ can be defined similarly as above. The sub-pixel loss L_{sub}^f is calculated as a ℓ_2 loss between the warped positions and the ground truth.

Total Loss. The total loss L is formulated as a linear combination of each term mentioned above:

$$L = \lambda_1 L_{1/16}^c + \lambda_2 L_{1/8}^c + \lambda_3 L_{1/1}^f + \lambda_4 L_{sub}^f. \quad (16)$$

4. Experiments

Unless otherwise stated, all methods are evaluated by default on a single NVIDIA V100 GPU using FP32 precision. The **first** and **second** results are highlighted.

4.1. Implementation Details

As shown in Tab. 1, the only difference between our full and lite models is the number of channels for low-level feature extraction. The number of channels for high-level feature extraction and one-to-many matching is set to 256. $k = 8$ correspondence priors are selected for one-to-one matching. The hybrid interaction module and the RSCA module are repeated $N_{1/16} = 2$ and $N_{1/8} = 2$ times, respectively. The window size w for local patch extraction is set to 5. The loss weights are set to $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\lambda_3 = 0.25$, and $\lambda_4 = 1.0$. Both models are trained on MegaDepth [18] using 8 NVIDIA V100 GPUs with a batch size of 8 for 30 epochs.

4.2. Relative Pose Estimation

Datasets. MD-1500 and SN-1500 are widely adopted benchmarks for estimating relative pose in outdoor and indoor scenes, respectively. The MegaDepth [18] dataset contains around 130K images that correspond to 196 sparse 3D models reconstructed using COLMAP [28]. MD-1500, selected by LoFTR [32], contains image pairs exhibiting changes in viewpoint and illumination from the “St. Peter’s Square” and “Brandenburger Tor”. The ScanNet [8] dataset is richly annotated with 3D camera poses and contains 2.5M views from 1,613 indoor scans. SN-1500, selected by SuperGlue [25], consists of image pairs from scenes with low texture and repetitive patterns.

Evaluation Protocol. We evaluate all methods on both benchmarks using a model trained on outdoor scenes, assessing in-domain and cross-domain performance. Matching accuracy is reported as the area under the curve (AUC) of the relative pose error at various thresholds, and efficiency is measured by the average runtime across benchmarks to reveal resolution effects. All images are resized to align with the corresponding accuracy metrics. Standard RANSAC is used as a common estimator, with a uniform threshold of 0.5 pixels applied in both accuracy and efficiency evaluations.

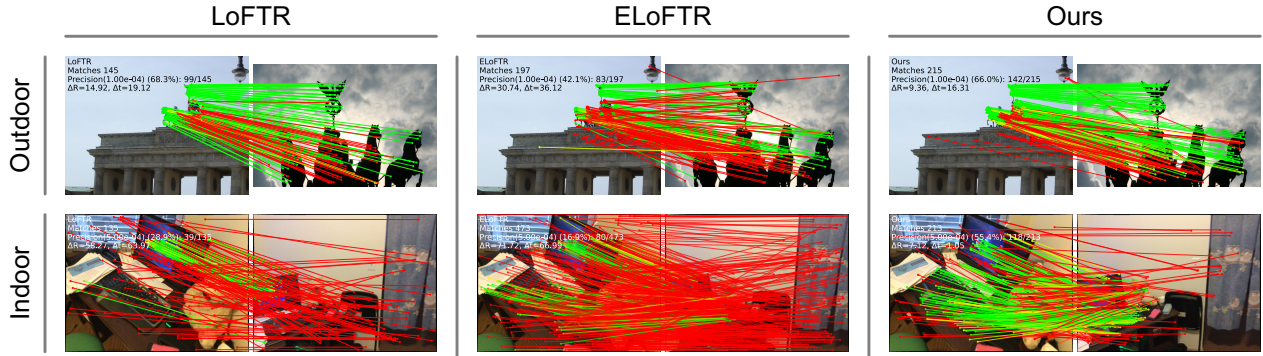


Figure 5. **Qualitative Results.** Two challenging image pairs are selected for qualitative analysis and compared with LoFTR and ELoFTR. One pair lacks texture details due to illumination changes, while the other undergoes significant viewpoint variations.

Method	AUC on HPatches \uparrow		
	@3px	@5px	@10px
SP [9] + SG [25]	53.9	68.3	81.7
LoFTR [32]	65.9	75.6	84.6
ELoFTR [40]	66.5	76.4	85.5
DKM [11]	71.3	80.6	88.5
Ours-full	71.8	80.6	88.0

Table 3. **Homography Estimation Results on HPatches.**

Results. As shown in Tab. 2, the proposed method demonstrates highly competitive performance on both in-domain and cross-domain benchmarks across all categories. In terms of accuracy, our full model achieves results comparable to AffineFormer [6], the best-performing semi-dense method. However, AffineFormer significantly lags behind our method in runtime. Notably, the marked improvement on SN-1500 highlights the strong cross-domain generalization capability of our method, with further validation provided in the ablation study. Regarding efficiency, our lite model achieves a speedup of $\sim 2.2/3.2\times$ on MD-1500 and $\sim 1.4/2.2\times$ on SN-1500 compared to ELoFTR/LoFTR using FP32 precision, with additional speedup under FP16 precision.

4.3. Homography Estimation

Datasets. HPatches [1] is a well-established benchmark for homography estimation. It contains 108 sequences, each consisting of 5 image pairs with viewpoint or illumination changes, along with their corresponding ground truth.

Evaluation Protocol. We follow previous work by reporting the AUCs of the mean reprojection error for the four corner points warped by the estimated homography at different thresholds. Similarly, the standard RANSAC solver with a threshold of 2 pixels is used to estimate the homography.

Results. As shown in Tab. 3, the proposed method demonstrates significant improvements in matching accuracy compared to all baseline methods. Notably, it achieves results

Method	Day	Night
	(0.25m,2 $^\circ$) / (0.5m,5 $^\circ$) / (1.0m,10 $^\circ$) \uparrow	
LoFTR [32]	88.7 / 95.6 / 99.0	78.5 / 90.6 / 99.0
TopicFM [13]	90.2 / 95.9 / 98.9	77.5 / 91.1 / 99.5
ASpanFormer [5]	89.4 / 95.6 / 99.0	77.5 / 91.6 / 99.5
ELoFTR [40]	89.6 / 96.2 / 99.0	77.0 / 91.1 / 99.5
Ours-full	89.2 / 96.1 / 98.9	78.0 / 91.6 / 99.5

Table 4. **Visual Localization Results on Aachen Day-Night v1.1.**

Method	DUC1	DUC2
	(0.25m,2 $^\circ$) / (0.5m,5 $^\circ$) / (1.0m,10 $^\circ$) \uparrow	
LoFTR [32]	47.5 / 72.2 / 84.8	54.2 / 74.8 / 85.5
TopicFM [13]	52.0 / 74.7 / 87.4	53.4 / 74.8 / 83.2
ASpanFormer [5]	51.5 / 73.7 / 86.0	55.0 / 74.0 / 81.7
ELoFTR [40]	52.0 / 74.7 / 86.9	58.0 / 80.9 / 89.3
Ours-full	52.0 / 77.3 / 86.4	55.0 / 80.2 / 84.0

Table 5. **Visual Localization Results on InLoc.**

comparable to the dense method DKM [11], highlighting the superiority of the two-stage homography-based refinement module in achieving sub-pixel accuracy.

4.4. Visual Localization

Datasets. Another major application is estimating 6-DoF camera poses relative to a known 3D scene, commonly referred to as visual localization. The Aachen Day-Night v1.1 [27] is a challenging outdoor dataset that involves significant illumination changes, while InLoc [33] is an indoor dataset characterized by viewpoint changes and occlusions.

Evaluation Protocol. Following prior work, we adopt the feature-based framework HLoc [24] to evaluate the accuracy of multi-view matching in visual localization. We report the percentage of query images with localization errors below the specified angular and distance thresholds.

Method	AUC on ETH3D[O]-3438 \uparrow						AUC on ETH3D[I]-2131 \uparrow					
	RANSAC			MAGSAC++			RANSAC			MAGSAC++		
	@5°	@10°	@20°	@5°	@10°	@20°	@5°	@10°	@20°	@5°	@10°	@20°
EL w/ DS	56.7	63.2	69.1	58.2	64.5	70.2	49.1	55.0	59.4	51.3	56.8	61.1
EL w/o DS	53.4 _{-3.3}	60.1 _{-3.1}	66.3 _{-2.8}	54.7 _{-3.5}	61.1 _{-3.4}	67.3 _{-2.9}	44.7 _{-4.4}	50.8 _{-4.2}	55.7 _{-3.7}	46.2 _{-5.1}	52.2 _{-4.6}	57.0 _{-4.1}
EL+CM-full	60.1 _{+3.4}	65.6 _{+2.4}	70.6 _{+1.5}	61.8 _{+3.6}	67.1 _{+2.6}	71.8 _{+1.6}	52.3 _{+3.2}	57.2 _{+2.2}	60.7 _{+1.3}	54.3 _{+3.0}	58.8 _{+2.0}	62.3 _{+1.2}
EL+CM-lite	58.3 _{+1.6}	64.1 _{+0.9}	69.4 _{+0.3}	60.5 _{+2.3}	66.0 _{+1.5}	71.1 _{+0.9}	50.1 _{+1.0}	54.9 _{-0.1}	58.6 _{-0.8}	52.5 _{+1.2}	57.0 _{+0.2}	60.3 _{-0.8}
Ours-full	61.8 _{+5.1}	66.8 _{+3.6}	71.5 _{+2.4}	63.2 _{+5.0}	68.0 _{+3.5}	72.6 _{+2.4}	56.1 _{+7.0}	60.7 _{+5.7}	64.0 _{+4.6}	58.2 _{+6.9}	62.6 _{+5.8}	65.6 _{+4.5}
Ours-lite	60.3 _{+3.6}	65.9 _{+2.7}	71.2 _{+2.1}	62.2 _{+4.0}	67.5 _{+3.0}	72.6 _{+2.4}	54.3 _{+5.2}	59.2 _{+4.2}	62.7 _{+3.3}	56.6 _{+5.3}	61.2 _{+4.4}	64.5 _{+3.4}

Table 6. **Ablation Study on Cross-Domain Relative Pose Estimation.** EL refers to ELoFTR, and CM denotes the cascaded matching.

Method	#Params (M)	GMACs \downarrow	Runtime (ms) \downarrow		Mem. (GB) \downarrow	
			FP32	FP16	FP32	FP16
EL w/ DS	16.0	909.1	238.3	158.8	13.4	13.6
EL w/o DS	16.0	909.1	185.3	92.8	10.1	10.3
EL+CM-full	17.6	708.1	144.6	79.4	12.6	7.4
EL+CM-lite	14.5	382.1	107.1	64.0	11.5	6.7
Ours-full	16.3	691.0	147.2	83.8	9.9	5.9
Ours-lite	13.2	365.1	108.1	67.7	9.0	5.9

Table 7. **Ablation Study on Efficiency.**

Results. As shown in Tab. 4 and 5, the proposed method achieves competitive results compared to methods that prioritize accuracy. Considered the most efficient semi-dense method, our method can accelerate the matching stage of this framework by ~ 2 to $3\times$ compared to other methods.

4.5. Understanding CasP

Ablation Study. The ablation study primarily addresses two concerns: **a)** How does ELoFTR [40] perform when the DS operator is removed to speed up inference? **b)** As shown in Tab. 2, CasP achieves a more significant accuracy gain on SN-1500 than on MD-1500. How can this cross-domain generalization be further validated? To investigate these issues, we select two additional datasets, ETH3D[O] and ETH3D[I], from the zero-shot evaluation benchmark proposed by GIM [30]. These benchmarks represent real-world outdoor and indoor scenes in the ETH3D [29], respectively. The longer side of each image is resized to 1152 pixels, and standard RANSAC and MAGSAC++ [2] are employed as estimators. We draw the following conclusions from the results shown in Tab. 6 and 7: **1)** Removing the DS operator is a trade-off that compromises accuracy, as the matching stage relies solely on descriptor similarity and ignores global confidence. **2)** To examine the core design of our method, we replace the DS operator in ELoFTR with the cascaded matching module described in Sec. 3.3. Even the lite model performs comparably to or better than the original full model. **3)** Building upon the novel matching pipeline, integrating

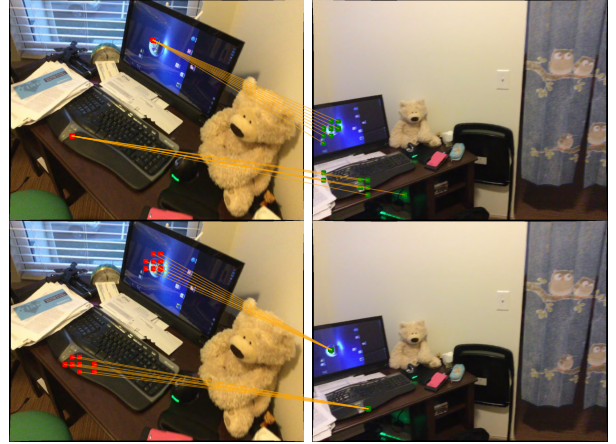


Figure 6. **Visualization of One-to-Many Matching.** For each token, $k = 8$ priors are displayed in the corresponding image.

additional advanced modules in our method further enhances accuracy. **4)** Our pipeline requires fewer GMACs, delivers faster runtime, and uses less memory on MD-1500.

Visualization. We present qualitative results in Fig. 5 and visualize the one-to-many matching priors in Fig. 6. These priors significantly facilitate identifying the most probable positions for one-to-one matching in challenging scenarios.

5. Conclusion

In this paper, we propose a cascaded matching pipeline to address the efficiency bottleneck of existing methods. Building upon this pipeline, we introduce a novel semi-dense method, CasP, which integrates advanced modules to enhance both matching accuracy and efficiency. Compared to the state-of-the-art method ELoFTR, our method achieves a speedup of $\sim 2.2\times$ at a resolution of 1152. Moreover, extensive experiments demonstrate that this novel pipeline significantly contributes to cross-domain generalization. These improvements are crucial for real-world applications, particularly for latency-sensitive and high-robustness tasks.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2024YFB3909001; the National Natural Science Foundation of China under Grants 42192583, 42030102, and 42471470; the China Railway Group Laboratory Basic Research Project under Grant L2023G014; the Major Special Projects of Guizhou [2022]001; and the Ant Group Research Fund.

References

- [1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, pages 5173–5182, 2017. 7
- [2] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. In *CVPR*, pages 1304–1312, 2020. 8
- [3] Chenjie Cao and Yanwei Fu. Improving transformer-based image matching by cascaded capturing spatially informative keypoints. In *ICCV*, pages 12129–12139, 2023. 3
- [4] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *ICCV*, pages 6301–6310, 2021. 2
- [5] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *ECCV*, pages 20–36, 2022. 1, 3, 4, 6, 7
- [6] Hongkai Chen, Zixin Luo, Yurun Tian, Xuyang Bai, Ziyu Wang, Lei Zhou, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, et al. Affine-based deformable attention and selective fusion for semi-dense matching. In *CVPRW*, pages 4254–4263, 2024. 1, 3, 6, 7
- [7] Peiqi Chen, Lei Yu, Yi Wan, Yongjun Zhang, Jian Wang, Liheng Zhong, Jingdong Chen, and Ming Yang. Ecomatcher: Efficient clustering oriented matcher for detector-free image matching. In *ECCV*, pages 344–360, 2024. 2, 3, 4
- [8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 6
- [9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, pages 224–236, 2018. 1, 2, 6, 7
- [10] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *CVPR*, pages 13733–13742, 2021. 3
- [11] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *CVPR*, pages 17765–17775, 2023. 6, 7
- [12] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *CVPR*, pages 19790–19800, 2024. 6
- [13] Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm: Robust and interpretable topic-assisted feature matching. In *AAAI*, pages 2447–2455, 2023. 2, 7
- [14] Pierre Gleize, Weiyao Wang, and Matt Feiszli. Silk: Simple learned keypoints. In *ICCV*, pages 22499–22508, 2023. 2
- [15] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *CVPR*, pages 21594–21603, 2024. 1
- [16] Dihe Huang, Ying Chen, Yong Liu, Jianlin Liu, Shang Xu, Wenlong Wu, Yikang Ding, Fan Tang, and Chengjie Wang. Adaptive assignment for geometry aware local feature matching. In *CVPR*, pages 5425–5434, 2023. 4
- [17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, pages 5156–5165, 2020. 2
- [18] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 6
- [19] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *CVPR*, pages 5987–5997, 2021. 1
- [20] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV*, pages 17627–17638, 2023. 1, 2, 6
- [21] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 1, 2
- [22] Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, and Yun Fu. Image as set of points. In *ICLR*, 2023. 3
- [23] Junjie Ni, Yijin Li, Zhaoyang Huang, Hongsheng Li, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Pats: Patch area transportation with subdivision for local feature matching. In *CVPR*, pages 17776–17786, 2023. 3
- [24] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, pages 12716–12725, 2019. 1, 7
- [25] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, pages 4938–4947, 2020. 1, 2, 6, 7
- [26] Paul-Edouard Sarlin, Ajaykumar Unagar, Måns Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the feature: Learning robust camera localization from pixels to pose. In *CVPR*, pages 3247–3257, 2021. 1
- [27] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *CVPR*, pages 8601–8610, 2018. 7
- [28] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 1, 6

- [29] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 3260–3269, 2017. 8
- [30] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. In *ICLR*, 2024. 8
- [31] Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching. In *CVPR*, pages 12517–12526, 2022. 2
- [32] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, pages 8922–8931, 2021. 1, 2, 4, 5, 6, 7
- [33] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, pages 7199–7209, 2018. 7
- [34] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. In *ICLR*, 2022. 1, 3, 4, 6
- [35] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, et al. Long-term visual localization revisited. *IEEE TPAMI*, 44(4):2074–2088, 2020. 1
- [36] Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *CVPR*, pages 6258–6268, 2020. 3
- [37] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. In *NeurIPS*, pages 14254–14265, 2020. 1, 2
- [38] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *ACCV*, pages 2746–2762, 2022. 3, 4
- [39] Xiaolong Wang, Lei Yu, Yingying Zhang, Jiangwei Lao, Lixiang Ru, Liheng Zhong, Jingdong Chen, Yu Zhang, and Ming Yang. Homomatcher: Dense feature matching results with semi-dense efficiency by homography estimation. In *AAAI*, pages 7952–7960, 2025. 6
- [40] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *CVPR*, pages 21666–21675, 2024. 2, 4, 5, 6, 7, 8
- [41] Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter CY Chen, and Zhengguo Li. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE TMM*, 25:3101–3112, 2022. 2
- [42] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter CY Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE TIM*, 72:1–16, 2023. 2