Affine-based Deformable Attention and Selective Fusion for Semi-dense Matching

Hongkai Chen[1] Zixin Luo[1] Yurun Tian[1] Xuyang Bai[1] Ziyu Wang[1] Lei Zhou[1] Mingmin Zhen
Tian Fang[1] David McKinnon[1] Yanghai Tsin[1] Long Quan[2]
[1]Apple Inc. [2]Hong Kong University of Science and Technology

Abstract

Identifying robust and accurate correspondences across images is a fundamental problem in computer vision that enables various downstream tasks. Recent semi-dense matching methods emphasize the effectiveness of fusing relevant cross-view information through Transformer. In this paper, we propose several improvements upon this paradigm. Firstly, we introduce affine-based local attention to model cross-view deformations. Secondly, we present selective fusion to merge local and global messages from cross attention. Apart from network structure, we also identify the importance of enforcing spatial smoothness in loss design, which has been omitted by previous works. Based on these augmentations, our network demonstrate strong matching capacity under different settings. The full version of our network achieves state-of-the-art performance among semi-dense matching methods at a similar cost to LoFTR, while the slim version reaches LoFTR baseline's performance with only 15% computation cost and 18% parameters.

1. Introduction

Robust and accurate image matching serves as a critical front-end task for a wide range of applications that require estimating geometry from RGB input, such as Structure-from-Motion (SfM) [14, 34], Simultaneous Localization And Mapping (SLAM) [26, 27] and Visual Localization [33]. Conventionally, image matching is comprised of several individual stages, including keypoint extraction, feature description, and feature matching. In the past few years, the research community has observed a remarkable progress in replacing traditional steps with their learning-based counterparts [6, 9, 23, 24, 36, 46, 51], leading to promising improvements. More recently, increasing efforts have been made towards more unified and end-to-end image matching systems, which are typically implemented in dense or semi-dense fashion by incorporating Transformer structure [3, 12, 35, 38, 42], cost-volume regularization [18, 28, 29, 39, 40] and coarse-to-fine scheme [10, 11, 35], which process image pairs directly and bypass limitations imposed by pre-extracted keypoints, such as repeatability issue and inability to handle low-texture areas.

Specifically, in Transformer-based matchers, striking a balance between token granularity and computation efficiency is crucial. To address this challenge, recent works [3, 50] propose global-local attention frameworks which utilizes global attention at a coarse level to model long-range dependencies, while local attention facilitates fine-level message exchange. Although these methods demonstrate the ability to concentrate attention span into specific areas, we consider below limitations still persist.

On the one hand, the local attention usually adopts rectangular grid areas to sample tokens, disregarding complex local deformation. Due to the nature of two-view matching tasks, corresponding regions in image pairs are usually related by some extent of deformations, including scaling, shearing and rotation. Consequently, a rectangular patch in the source view will be projected to a deformed area in the target image, which should be considered in token sampling for local attention to maximize the overlap ratio between sampling areas in source/target feature maps.

On the other hand, the global-local message fusion in previous works are usually conducted through learned priors [3, 38, 50], without considering the differences in reliability among sampled patches. Due to the imprecision in intermediate flow estimation and the existence of non-overlapping areas, accepting all local message equally introduces noise for feature update, not to mention the local message from non-overlapped regions. In this regard, an ideal global-local message fusion should suppress local information from unreliable or non-related local areas.

Apart from network designs, we also review the defactor loss design for semi-dense matching methods [3, 35, 38, 42, 44]. In this series of works, classification-based loss (such as focal loss) is applied to assignment matrix to

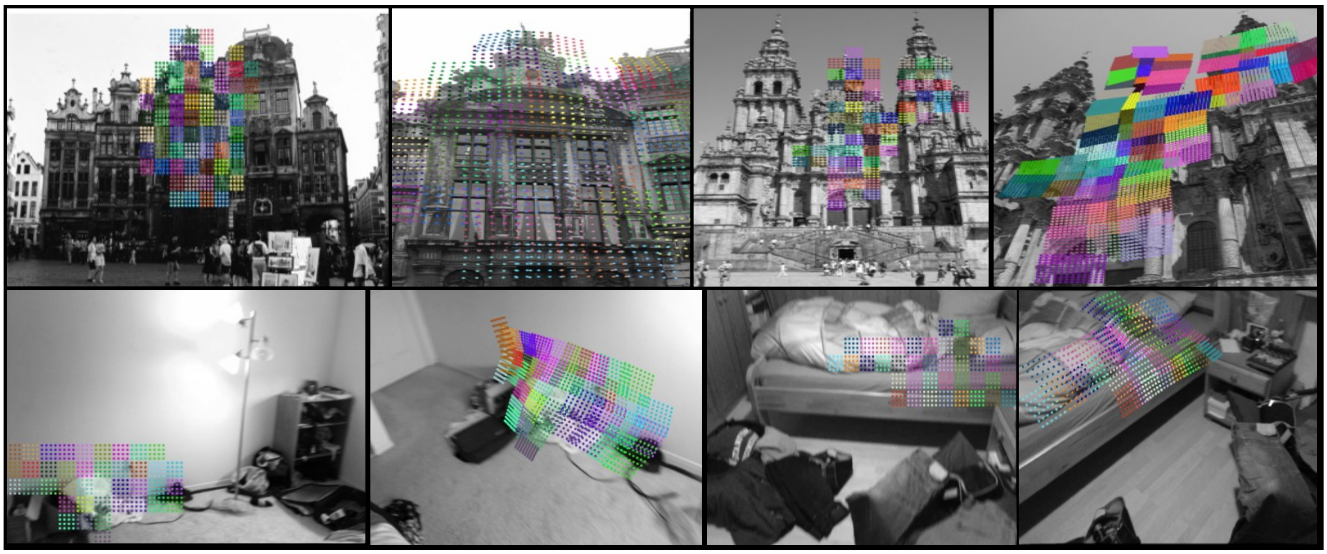

Figure 1. Visualization of the proposed deformable attention. Through piece-wise deformation estimation, we project source patches (left) to the target image (right) to sample tokens in local attention.

maximize (minimize) positive (negative) entries. Although classification-based loss is essential for learning distinctive features, it should be noticed that spatial relationship is not considered in such supervision since all entries are treated equally. To compensate for this lack of spatial supervision, we propose to additionally apply spatial softmax-based loss upon assignment matrix.

Based on aforementioned observations, this paper introduces AffineFormer, a novel cross-view Transformer equipped with geometric-aware deformable local attention and selective global-local message fusion, supervised by hybrid spatial softmax/classification-based loss. Following previous work [3], we regress intermediate flow map during attention process, from which an affine-based deformation field is estimated. This deformation field is then utilized to project the local attention span from the source view to the target view. In parallel with local attention, we also incorporate coarse global attention, where the local and global messages are fused based on the reliability of the flow. As supervision, we apply a Extensive experiments on both two-view pose estimation and visual localization demonstrate the effectiveness of our method.

2. Related Works

2.1. Global-Local Attention for Image Matching

Employing Transformer in dense/semi-dense matching boosts the matching capability of original features extracted from a single view, as has been studied by previous works [2, 3, 35, 38, 42, 50], yet the quadratic complexity of vanilla Transformer imposes challenges. Recently, some researchers [3, 38, 50] propose to combine coarse global attention and sparse local attention to handle high-resolution input while maintaining modest computation costs. Specifically, global attention establishes correlation across source and target tokens to guide local attention. In QuadTree Attention [38], full token sets are gradually tailored into different groups where sparse attention is performed only at more related target tokens at a fine level. AspanFormer [3] regresses intermediate flow maps with uncertainty to adaptively adjust attention span. ASTR [49] utilizes neighborhood around intermediate matching points to generate local attention region for better local consistency.

Sharing a similar practice with ASpanFormer, we regress intermediate flow field during cross attention process. However, instead of relying on learned uncertainty to determine attention pattern, we base our method on a geometric ground, where piece-wise affine deformation field are estimated to shape local attention span.

2.2. Deformable Attention

In the context of the general Vision Transformer [8], deformable attention [45, 56] is introduced as an augmentation to the vanilla attention framework. It dynamically shapes attention span based on local features, similar to the concept of deformable convolution in CNNs [15, 16]. Deformable attention predicts a set of offsets from sampled tokens to modify the sampling position. While it has shown effectiveness in various applications, plain deformable attention is not directly applicable to cross-view attention in image matching tasks. Additionally, the current deformable attention approach follows a fully data-driven method, making it challenging to interpret the learned deformation clearly. In contrast to the free-form deformable attention, we propose a

geometry-driven deformable operation that explicitly models local deformation introduced by view changes, as shown in Fig. 1.

2.3. Local Deformation Estimation

Estimating local deformation patterns is a highly focused subject in the field of image matching. In the pre-deep learning era, traditional descriptors relied on manually designed shape detectors to guide the generation of local patches [1, 22]. With rapid advancements of deep learning techniques, numerous studies have explored learning-based approaches for estimating deformations. OriNet [48] and LIFT [47] proposed to learn a canonical orientation for feature points, AffNet [25] predicts additional affine parameters to enhance modeling capabilities. UCN [4] and LF-Net [47] take images as input and apply spatial transformation networks to intermediate features. Some CNN-based local features [24, 53, 54] employs deformable convolution to generate dense deformation fields.

Unlike the aforementioned works that focus solely on single-view feature descriptions, we embed local shape estimation into a cross-view attention framework in a more principled manner.

3. Methodology

In Fig. 2, we provide an overview of our network architecture, which inherits the paradigm of semi-dense matching [35]. Taking an image pair I_A, I_B as input, our network generates coarse correspondences and then refine them. The nework begins with a CNN-based encoder to extract initial features in $\frac{1}{8}$ resolution for both images. These features are position-encoded through element-wise summation of sinusodinal signals [41] and passed through iterative self/cross attention blocks for enhancing. All self attention blocks are conducted at $\frac{1}{32}$ scale. For each cross attention, global attention is conducted at $\frac{1}{32}$ resolution, while affine-based local deformable attention is conducted at $\frac{1}{8}$ resolution. Local and global messages are then fused based on uncertainty of intermediate flow estimation to suppress unreliable local information. Self attention is also conducted at $\frac{1}{32}$ resolution between two cross attention blocks.

3.1. Global Attention

In this part, we introduce global attention at $\frac{1}{32}$ resolution, which is used in self attention blocks and global branch in cross attention blocks.Formally, the input source/target features F_s, F_t at $\frac{1}{8}$ scale are downsampled to $\frac{1}{32}$ through strided 2 average pooling. Vanilla dot-product attention is then performed to retrieve coarse message m_c, which is bilinear upsampled back to $\frac{1}{8}$ and fused with F_s through a feed-forward network (FFN).:

$$F'_s = \textbf{Avgpool}_{2\times2}(F_s) \tag{1}$$
$$F'_t = \textbf{Avgpool}_{2\times2}(F_t) \tag{2}$$
$$m' = \textbf{Attn}(W_q F'_s, W_k F'_t, W_v F'_t) \tag{3}$$
$$m = \textbf{Up}_{2\times2}(m') \tag{4}$$
$$\hat{F}_s = F_s + \textbf{FFN}(F_s, m) \tag{5}$$
$$= F_s + \textbf{LN}(\textbf{DWConv}(F_s + \textbf{MLP}(m)) \tag{6}$$

Here, $W_{(q/k/v)}$ denotes the linear transformation matrix for query/key/values vectors, **DWConv** means depth-wise convolution, **LN** means layer normalization. $\hat{F}_s$ is the updated source features, which is fed into following attention layers.

Note that global msg will additionally be combined with local message in cross attention blocks, which will be introduced in the next section. We also would like to mention that even without local branch in cross attention, this basic global attention blocks has been a very competitive baseline, which is validated both in our experiment in Sec. 4.3 and a concurrent LoFTR's follow up work [44].

3.2. Local Deformable Attention

Global attention ensures long-range dependency, yet inevitably lost fine-level information due to lack of concentration and downsampling. To alleviate this issue, previous works [3, 38, 50] propose to enhance global attention with parallel fine-level attention. However, these works either adopts irregular sparse sampling or data-driven rectangular sampling, neglecting the importance to align local deformation across two-views. Further more, fixed learned prior are used to fuse global and local message, yet not all local message are equally reliable due to inherent uncertainty.

To address these issues, we propose affine-based deformable attention and selective message fusion. Concretely, we estimate intermediate deformation filed to align the focusing region for each token group. The retrieved message in local attention are further fused with global message based on estimated uncertainty. In the following part, we introduce the workflow and insights of this operation.

3.2.1 Intermediate Flow Regression

As mentioned in Sec. 3.1, global cross attention is conducted parallel with each local attention, which outputs coarse retrieved message $m_c \in R^{\frac{H}{32}\times\frac{W}{32}\times C}$ and attention matrix $A \in R^{N\times M\times H}$. Here N, M denotes the number of source and target tokens, and H denotes the head number in multi-head attention. Naturally, the attention matrix reflects the similarity between cross-view features and thus can be used to decode a rough intermediate flow map. Inspired by the global decoder in dense matching methods [10, 11], we utilize weighted sum of position embeddings and a decoder to regress intermediate flows. Formally, mean of A

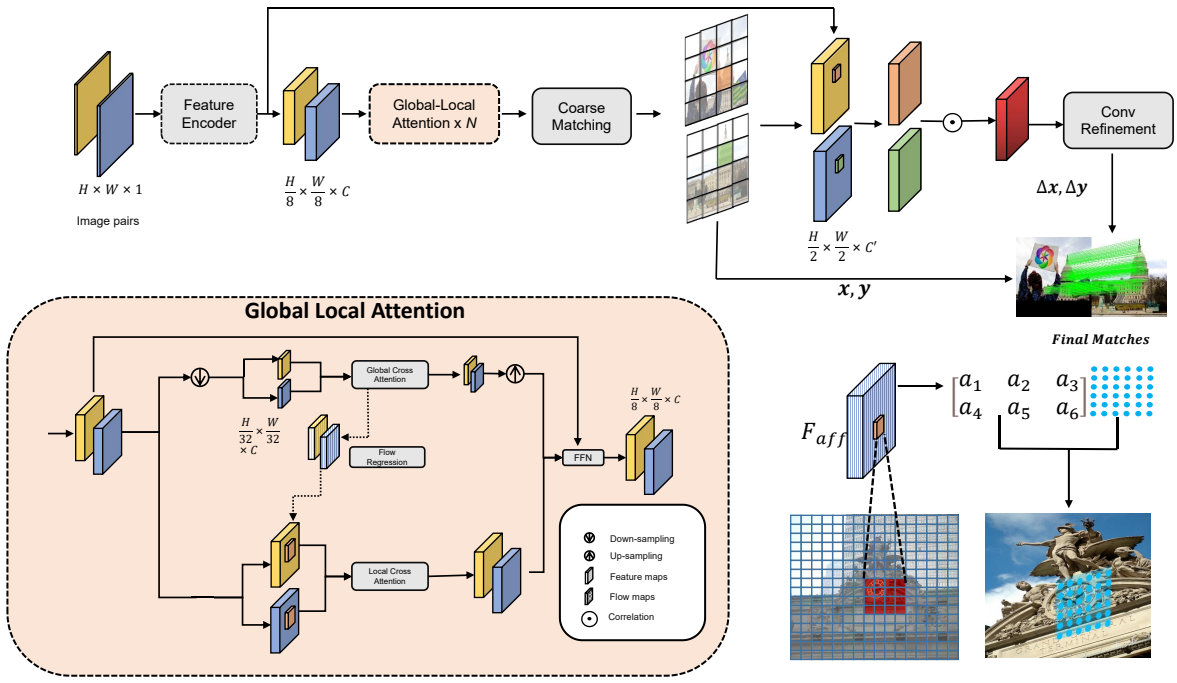

Figure 2. The overall structure of our proposed network. The network adopts iterative global-local attention operations to pass cross-view messages at both global and local scales. After identifying coarse level matches at $1/8$ resolution, a convolution refiner follows to predict correspondence residuals.

along the head dimension is computed as $\hat{A} \in R^{N\times M}$, which is used to aggregate weighted positional embeddings $P_t \in R^{M\times D}$ from the target image. Here, D is the positional embeddings' channel number. A convolution-based decoder is followed to regress flows and uncertainty:

$$\Phi = \textbf{Conv}(\hat{A}P), \quad \Phi \in R^{\frac{H}{32}\times\frac{W}{32}\times 4}. \tag{7}$$

Each element $\phi_i = [u_{xi}, u_{yi}, \sigma_{xi}, \sigma_{yi}]$ from Φ indicates the corresponding flow coordinates u_{xi}, u_{yi} and uncertainty σ_{xi}, σ_{yi} in each location. Drawing inspiration from previous works [3, 40, 55] that adopts a probabilistic framework to model flow uncertainty, we take uncertainty σ_{xi}, σ_{yi} as stand deviation in a two-dimensional Gaussian distribution and train them in a self-supervised manner. The estimated flow map is bilinear upsampled to $\frac{1}{8}$ resolution and is used to estimate patch-wise affine field.

3.2.2 Affine Field Estimation

Flow maps recovered by coarse attention matrix roughly reflect corresponding regions for each location in the source feature map, yet flow in free-form is inevitably corrupted by outliers and doesn't reflect priors in two-view geometry. For example, point correspondences from a local area without large depth fluctuation can be well approximated by an affine transformation.

To embed geometric priors into deformation field, we estimate piece-wise affine parameters from intermediate flow map Φ. Concretely, source flow map Φ is grouped by non-overlapping windows with size l. For each $l \times l$ window, a set of affine parameters is estimated from the local flow. Formally, we denote coordinates of each point i in a local window as $[x_i, y_i]$, while its corresponding flow as $[\hat{x}_i, \hat{y}_i]$. The affine to be estimated is denoted as $A \in R^{2\times 3}$. We set the last column of A as difference between mean of $[x_i, y_i]$, $[\hat{x}_i, \hat{y}_i]$,

$$\begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix} = \frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} x_i - \hat{x}_i \\ y_i - \hat{y}_i \end{bmatrix} \tag{8}$$

The rest of elements in A are estimated through linear least squares,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^T = (C^T C)^{-1} C^T \begin{bmatrix} \hat{x}'_1 & \hat{y}'_1 \\ \hat{x}'_2 & \hat{y}'_2 \\ \cdots & \cdots \\ \hat{x}'_N & \hat{y}'_N \end{bmatrix} \tag{9}$$

where

$$C = \begin{bmatrix} x'_1 & y'_1 \\ x'_2 & y'_2 \\ \cdots & \cdots \\ x'_N & y'_N \end{bmatrix} \tag{10}$$

$$\tag{11}$$

and

$$x_i' = x_i - a_{13}, y_i' = y_i - a_{23} \tag{12}$$

$$\hat{x}_i' = \hat{y}_i - a_{13}, \hat{y}_i' = \hat{y}_i - a_{23} \tag{13}$$

To reduce misalignment caused by noisy flow, we further regularize the estimated A through a series of operations, including constraining the underlying scale, rotation and shearing. More details about affine regularization are provided in supplementary materials.

As a result, the flow field $\Phi \in R^{\frac{H}{8} \times \frac{W}{8} \times 4}$ is converted to affine parameters field $F_{aff} \times R^{\frac{H}{8l} \times \frac{W}{8l} \times 6}$.

3.2.3 Deformable Attention

Ideally, the source/target patch in local attention should be aligned to capture the most relevant features. However, achieving this alignment with a rectangular attention pattern is often infeasible due to complex local deformations.

To address this issue, we leverage the estimated affine field to sample a deformed target patch for each source patch. Concretely, we reuse the non-overlapping windows with size l in previous affine field estimation as source patches. For each source patch S, we use the corresponding affine parameters $f_{aff} \in R^6$ from F_{aff} to project an affine patch $\hat{S}$ in target feature map. To ensure better coverage, the size of target patch is set as αl, which is α times larger than the source token patch. Illustration of this process can be seen Fig. 2.

Given query and key/value feature maps $F_q, F_k/v$, we sample tokens uniformly in $S, \hat{S}$, which yields tokens $f_q \in R^{l^2 \times D}, f_k/v \in R^{\alpha^2 l^2 \times D}$ for each patch. Local attention is performed within each patch pair to generate local message.

3.3. Selective Message Fusion

Simply averaging or concatenating global and local message m_g, m_l is a straightforward way for message fusion, which, however, is problematic since m_l may come from inaccurate flow estimation or non-overlapping regions. To address this issue, we use predicted uncertainty to weight local message. Concretely, for each position in the source feature map, we index the corresponding uncertainty σ_x, σ_y from the intermediate flow map. The fused message is calculated as a weighted sum of m_g, m_l:

$$m = p_1 m_g + p_2 m_l, \tag{14}$$

$$[p_1, p_2] = \textbf{softmax}(\alpha, \beta[(1 + \gamma \textbf{ReLu}(\sigma_x + \sigma_y)]^{-1}) \tag{15}$$

α, β are learnable parameters to balance prior weight for global and local messages, which are modified by uncertainty-related factor. A learnable parameter γ controls its sensitivity. Through this formulation, large uncertainty results in lower weight in message fusion. The obtained fuse message m is used to update source features through a feed-forward network as introduced in Sec. 3.1. An illustration of learned fusion heatmap can be seen in Fig 4. The produced score map helps our network to sharply focus on co-visible and salient regions, discarding the non-relevant areas in message fusion.

3.4. Match Determination

After all attentional blocks, the enhanced features $\tilde{F}_A \in R^{n \times c}, \tilde{F}_B \in R^{m \times c}$ are fisrt used to generate correlation matrix $C = \tau \tilde{F}_A \tilde{F}_B^T \in R^{n \times m}$, where τ is a temperature parameter, followed by dual-direction softmax operation to produce assignment matrix S. We retain coarse-level correspondences M_c by mutual nearest neighbor (MNN) and threshold of 0.2 on dual-softmax score.

To refine coarse matches M_c are in $1/8$ resolution, a local correlation-based refinement block is adopted. For each coarse match, we sample a local window with size w from the source and target feature map in $1/2$ resolution, which yield local patches $p_s, p_t \in R^{w \times w \times D}$. For feature in source patch p_s, we calculate its correlation score with target patch p_t, which are flattened into correlation feature and fed into a convolutional refiner to predict fine level residuals conditioned on coarse match.

3.5. Loss Formulation

Loss of our method consists of three parts, (1) coarse-level loss, (2) fine-level loss, (3) flow estimation loss.

Coarse level loss includes two terms, the classification loss L_{ce} and spatial softmax loss L_{cs}. In cogruent with previous works, we re-reproject points in each image pair using ground-truth depth and camera poses, where the mutual nearest match M_{gt} are considered as ground truth match. The classification loss L_{ce} is defined as a focal loss using assignment matrix S:

$$\begin{aligned} L_{ce} = &- \sum_{(i,j) \in M_{gt}} (1 - S(i,j))^{\gamma} \log(S(i,j)) \\ &- \sum_{(i,j) \notin M_{gt}} S(i,j)^{\gamma} \log(1 - S(i,j)). \end{aligned} \tag{16}$$

One limitation of classification loss is that all mismatches suffer the same loss penalty no matter how far they are from the ground truth. Taking inspiration from previous works on learned descriptor [43], we adopt an additional spatial softmax loss as compensation, where close mismatch should produce lower loss than 'distant' ones.

$$L_{cs} = \frac{1}{|M_{gt}|} \sum_{i \in M_{gt}[:,0]} [\sum_j S(i,j) P_{ij} - P_i^{gt}]^2 \tag{17}$$

Here, P_{ij} denotes coordinates for each entry in the assignment matrix, while P_i^{gt} denotes the ground truth match

LoFTR ASpanFormer Ours

Figure 3. Visualization of image matches from LoFTR [35], ASpanFormer [3] and our method, where green lines represent inlier matches, while red lines represent outlier matches.

coordinate for left point i. Spatial softmax loss encourage maximization of positive entries in assignment matrix, which is in the same direction of classification loss. Additionally, it also considers spatial relationship and penalizes mismatch according to their distance to the ground truth. In our ablation study in Sec. 4.3, we fine the simple modification on loss term largely benefit overall performance.

For intermediate flow, we follow previous works [3, 55] to minimize the log-likelihood for the estimated Gaussian distribution. Formally, given flow estimation Φ from each layer and ground truth flow D^{gt}, L_{flow} is defined as:

$$L_{flow} = -\frac{1}{|D^{gt}|}\sum_{ij}\log(P(D^{gt}_{ij}|\Phi_{ij})). \tag{18}$$

More details about flow regression are provided in the supplementary material.

The fine-level loss is defined as the l2-distance between regress refine offset and ground truth.

In summary, we formulate the loss as:

$$L = L_{ce} + L_f + \lambda_1 L_{cs} + \lambda_2 L_{flow}. \tag{19}$$

3.6. Implementation Details

Our network uses ResNet-18 [13] CNN backbone. We use 4 interleaved self/cross attention blocks for feature update. For local cross attention, we set window size $l = 4$. For fine-level refinement, we set local window size $w = 5$. For supervision, we set λ_1 as 1 for outdoor model and 5 for indoor model, λ_2 is set as 0.1.

We train two different models for indoor and outdoor scenes respectively on ScanNet and Megadepth. Both models follows the same training scheme in previous works [3, 35, 38], which lasts for 30 epochs on 8 V-100 GPUs. More details of implementation and training are available in the supplementary material.

4. Experiments

4.1. Two-view Pose Estimation

Datasets. We use ScanNet [5] and MegaDepth [19] datasets to evaluate the matching ability of our method in indoor scenes and outdoor scenes. We follow evaluation protocols to select 1500 image pairs from two datasets respectively, where the relative poses are recovered through OpenCV ransac, as is done in previous works [3, 35, 38, 50]. For ScanNet, we resize all images to [640,480] resolutions. For MegaDepth, we resize all images to [1152,1152]. LoFTR(E) and TopicFM only reports outdoor trained model, thus the corresponding number for indoor evaluation are omitted.

Comparative methods. We compare the proposed method with 1) sparse approaches, 2) semi-dense approaches that outputs $1/8$ resolution coarse matches with local refinement, including LoFTR [35], QuadTree Attention [38], ASpanFormer [3], TopicFM [12] and efficient LoFTR [44]. We also include a series fully dense methods producing dense warp for reference.

Results. As presented in Table 1, our method outperforms all sparse and semi-dense methods on both indoor and out-

Figure 4. Heatmap for selective local fusion score.

Method	ScanNet			MegaDepth		
	AUC@5	AUC@10	AUC@20	AUC@5	AUC@10	AUC@20
SP [6]+SuperGlue [31]	16.2	33.8	51.8	49.7	67.1	80.6
SP [6]+LightGlue [21]	14.8	30.8	47.5	49.9	67.0	80.1
LoFTR [35]	22.0 (16.9)	40.8 (33.6)	57.6 (50.6)	52.8	69.2	81.2
QuadTree [38]	24.9 (19.0)	44.7 (37.3)	61.8 (53.5)	54.6	70.5	82.2
MatchFormer [42]	24.3 (15.8)	43.9 (32.0)	61.4 (48.0)	53.3	69.7	81.8
ASpanFormer [3]	25.6 (19.6)	46.0 (37.7)	63.3 (54.4)	55.3	71.5	83.1
TopicFM [12]	- (17.3)	- (35.5)	- (50.9)	54.1	70.1	81.6
LoFTR (E) [44]	- (19.2)	- (37.0)	- (53.6)	56.4	72.2	83.5
Ours	**27.1 (22.0)**	**47.5 (40.9)**	**64.8 (58.0)**	**57.3**	**72.8**	**84.0**
PDCNet+(H) [40]	20.3	39.4	57.1	51.5	67.2	78.5
CasMTR [2]	27.1	47.0	64.4	59.1	74.3	84.8
DKM [11]	29.4	50.7	68.3	60.4	74.9	85.1
RoMa [10]	31.8	53.4	70.9	62.6	76.7	86.3

Table 1. Two-view pose estimation results on ScanNet dataset [5] in indoor scenes and MegaDepth dataset [20] in outdoor scenes. Figures in bracket are results of evaluating outdoor-trained model on ScanNet dataset.

door dataset. We also report cross-dataset generalization results by evaluating MegaDepth model on ScanNet. Even without training on any indoor scenes, our outdoor model demonstrates high accuracy on indoor scenerios.

The speed-optimized efficient loftr demonstrates impressive performance with several advanced network designs, including rotary positional encoding, two stage refinement and RepVGG backbone, which can also be adopted to enhance our method (the reported results don't use these enhancements).

4.2. Visual Localization

Datasets. Apart from two-view pose estimation, we further evaluate our network in the visual localization pipeline, where two stand benchmarks InLoc [37] and Aachen Day-Night v1.1 [32, 33, 52] are used to demonstrate the performance in indoor and outdoor scenes. We embed our method to Hloc [30] pipeline for evaluation. We use the model trained on the MegaDepth to localize both inloc and Aachen datasets. All input images are resized so that the longest dimension is 1024.

Results. The results are reported on Table 2. Our method demonstrates highest accuracy on Aachen dataset, and similar performance with a concurrent work efficient LoFTR [44] on InLoc dataset. Noted that the improvements proposed in our method is orthogonal to efficient LoFTR. Generally, our method shows strong generalization ability in visual localization settings.

Method	DUC1	DUC2	Mean	Day	Night	Mean
	(0.25m,2) / (0.5m,5) / (1m,10)					
SP [7]+SuperGlue [31]	47.0 / 69.2 / 79.8	53.4 / 77.1 / 80.9	67.90	89.8 / 96.1 / **99.4**	77.0 / 90.6 / **100.0**	92.15
SP [7]+LightGlue [21]	49.0 / 68.7 / 80.8	55.0 / 74.8 / 79.4	67.95	**90.2** / 96.0 / **99.4**	77.0 / 91.1 / **100.0**	92.28
LoFTR [35]	47.5 / 72.2 / 84.8	54.2 / 74.8 / 85.5	69.83	88.7 / 95.6 / 99.0	78.5 / 90.6 / 99.0	91.90
MatchFormer [42]	46.5 / 73.2/ 85.9	55.7 /71.8 / 81.7	69.13	- /- / -	- / - / -	-
ASpanFormer [3]	51.5 / 73.7 / 86.4	55.0 / 74.0 / 81.7	70.38	89.4 / 95.6 / 99.0	77.5 / 91.6 / 99.5	92.10
PATS [17]	55.6 / 71.2 / 81.0	**58.8** / **80.9** / 85.5	71.45	89.6 / 95.8 / 99.3	73.8 / **92.1** / 99.5	91.68
TopicFM [12]	52.0 / 74.7 / **87.4**	53.4 / 74.8 / 83.2	72.17	90.2 / 95.9 / 98.9	77.5 / 91.1 / 99.5	92.18
LoFTR (E) [44]	52.0 / 74.7 / 86.9	58.0 / **80.9** / **89.3**	**73.63**	89.6 / **96.2** / 99.0	77.0 / 91.1 / 99.5	92.06
Ours	**56.1** / 74.7 / 86.9	55.0 /79.4 / 87.0	73.18	89.9 / **96.2** / 98.9	**79.1** / 91.1 / 99.5	**92.45**

Table 2. Visual localization results on InLoc [37] and Aachen Day-Night v1.1 dataset.

4.3. Ablation Study

We conducted an ablation study on the ScanNet dataset, following the protocol outlined in Section 4.1. The results in shown in Table 3. Our baseline uses 4 interleaved self/cross attention, where the cross attention is only conducted at $\frac{1}{32}$ scale without fine-level local attention.

To study the effect of loss design, we test two settings: (1) replace focal loss with spatial softmax loss (row 2), (2) use spatial softmax loss as an additional term (row 3). We observe that direct replacement results in worse results, indicating focal loss is essential for learning distinctive features. Adding spatial softmax loss brings considerable improvement, reflecting the importance to enforce spatial smoothness in loss design.

We then sequentially add local attention with fixed-size rectangular span (row 4), add uncertainty-based selective fusion in Sec. 3.3 (row 5), and apply affine-based deformation (row 6). All proposed components make notable contributions over baseline, validating the effectiveness of our method designs.

Design	Pose Estimation AUC		
	@5	@10	@20
baseline	24.0	44.4	61.5
replace focal loss with s.s. loss	21.5	40.1	57.9
+s.s. loss	25.1	45.6	62.8
+local attn.	26.0	46.1	63.4
+selective fusion	26.6	46.8	63.9
+affine estimation.	27.1	47.5	64.8

Table 3. Ablation study on ScanNet dataset [5].

4.4. Efficiency Evaluation

In this section, we conduct a comparison of different semi-dense methods on their model size and inference cost. In addition to our normal setting, we also provide a light variant with very slim backbone (600k parameters) and 2 attention layers (denoted as ours-L). Details for the light setting can be found in supplementary materials. We apply the similar light-weight modifications to LoFTR (denoted as LoFTR-L).

Method	AUC@5/10/20	#Parameters(M)	GFLOPs	Latency(ms)
LoFTR	52.8 / 69.2 / 81.2	11.1	1767	281.6
LoFTR-L	49.6 / 66.7 / 79.6	2.3	235	89.3
Ours-L	52.4 / 69.8 / 81.7	2.1	265	98.5
QuadTree	54.6 / 70.5 / 82.2	13.2	1792	335.2
ASpanFormer	55.3 / 71.5 / 83.1	15.5	1855	312.3
Ours	57.3 / 72.8 / 84.0	12.8	1678	296.4

Table 4. Performance-cost trade-off for different methods, we report auc on megadepth dataset. Ours-L denotes our light variants. Flops and latency for all methods are measured with image resize to 1200/1152 resolution. We use one V100 GPU for testing.

As can be seen in Tab. 4, our normal version network shares similar cost and model size with LoFTR while delivers significant performance gain. Our light version network uses only 15% flops and 18% parameters to reach LoFTR's performance, while reducting LoFTR to the same level results in largely degenerated performance. Overall, our method shows good performance under different cost budget.

5. Conclusion

In this paper, we introduce AffineFormer, a novel semi-dense matcher equipped with affine-based deformable local attention and selective message fusion. We capture local deformation caused by viewpoint changes through the estimation of affine transformation field, which is used to shape local attention patterns. We then propose to fuse global-local message robustly through adaptive fusion. The effectiveness of spatial softmax-based loss is also studied, which is neglected in previous works. Extensive experiments demonstrates our method's effectiveness in geometry estimation.

References

[1] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012. 3

[2] Chenjie Cao and Yanwei Fu. Improving transformer-based image matching by cascaded capturing spatially informative keypoints. *arXiv preprint arXiv:2303.02885*, 2023. 2, 7

[3] 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*, 2022. 1, 2, 3, 4, 6, 7, 8

[4] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *NeurIPS*, 2016. 3

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 6, 7, 8

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018. 1, 7

[7] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018. 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2

[9] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *CVPR*, 2019. 1

[10] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Revisiting robust losses for dense feature matching. *arXiv preprint arXiv:2305.15404*, 2023. 1, 3, 7

[11] Johan Edstedt, Mårten Wadenbäck, and Michael Felsberg. Deep kernelized dense geometric matching. *Preprint*, 2022. 1, 3, 7

[12] Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm: Robust and interpretable topic-assisted feature matching. In *AAAI*, 2023. 1, 6, 7, 8

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[14] J. Heinly, J. L. Schönberger, E. Dunn, and J. Frahm. Reconstructing the world* in six days. In *CVPR*, 2015. 1

[15] Kaiming He Jian Sun Jifeng Dai, Yi Li. R-FCN: Object detection via region-based fully convolutional networks. 2016. 2

[16] Yuwen Xiong Yi Li Guodong Zhang Han Hu Yichen Wei Jifeng Dai, Haozhi Qi. Deformable convolutional networks. *arXiv preprint arXiv:1703.06211*, 2017. 2

[17] Zhaoyang Huang Hongsheng Li Hujun Bao Zhaopeng Cui Guofeng Zhang Junjie Ni, Yijin Li. Pats: Patch area transportation with subdivision for local feature matching. In *The IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2023. 8

[18] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. In *NeurIPS*, 2020. 1

[19] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 6

[20] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 7

[21] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. *ICCV*, 2023. 7, 8

[22] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 3

[23] Zixin Luo, Tianwei Shen, Lei Zhou, Siyu Zhu, Runze Zhang, Yao Yao, Tian Fang, and Long Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *ECCV*, 2018. 1

[24] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Aslfeat: Learning local features of accurate shape and localization. In *CVPR*, 2020. 1, 3

[25] Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Repeatability is not enough: Learning affine regions via discriminability. In *ECCV*, 2018. 3

[26] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 2015. 1

[27] Raul Mur-Artal and Juan Tardos. ORB-SLAM2: an open-source slam system for monocular, stereo and rgb-d cameras. *IEEE Transactions on Robotics*, 2016. 1

[28] I. Rocco, R. Arandjelović, and J. Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *European Conference on Computer Vision*, 2020. 1

[29] I. Rocco, M. Cimpoi, R. Arandjelovi, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018. 1

[30] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 7

[31] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 7, 8

[32] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6dof outdoor visual localization in changing conditions. In *CVPR*, 2018. 7

[33] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, 2012. 1, 7

[34] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[35] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 1, 2, 3, 6, 7, 8

[36] Weiwei Sun, Wei Jiang, Eduard Trulls, Andrea Tagliasacchi, and Kwang Moo Yi. Acne: Attentive context normalization

for robust permutation-equivariant learning. In *CVPR*, 2020. 1

[37] 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*, 2018. 7, 8

[38] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. In *ICLR*, 2021. 1, 2, 3, 6, 7

[39] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *CVPR*, 2021. 1

[40] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. PDC-Net+: Enhanced probabilistic dense correspondence network. *Preprint*, 2021. 1, 4, 7

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPs*. 2017. 3

[42] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. *Preprint*, 2022. 1, 2, 7, 8

[43] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *ECCV*, 2020. 5

[44] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *CVPR*, 2024. 1, 3, 6, 7, 8

[45] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, 2022. 2

[46] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016. 1

[47] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016. 3

[48] Kwang Moo Yi, Yannick Verdie, Pascal Fua, and Vincent Lepetit. Learning to assign orientations to feature points. In *CVPR*, 2016. 3

[49] Jiahuan Yu, Jiahao Chang, Jianfeng He, Tianzhu Zhang, and Feng Wu. Adaptive spot-guided transformer for consistent local feature matching. *CVPR*, 2023. 2

[50] Jiahuan Yu, Jiahao Chang, Jianfeng He, Tianzhu Zhang, Jiyang Yu, and Wu Feng. ASTR: Adaptive spot-guided transformer for consistent local feature matching. *CVPR*, 2023. 1, 2, 3, 6

[51] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *ICCV*, 2019. 1

[52] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *IJCV*, 2021. 7

[53] 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 Transactions on Instrumentation and Measurement*, 2023. 3

[54] Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter CY Chen, and Zhengguo Li. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia*, 2022. 3

[55] Lei Zhou, Zixin Luo, Tianwei Shen, Jiahui Zhang, Mingmin Zhen, Yao Yao, Tian Fang, and Long Quan. Kfnet: Learning temporal camera relocalization using kalman filtering. In *CVPR*, 2020. 4, 6

[56] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2