ASpanFormer: Detector-Free Image Matching with Adaptive Span Transformer

Hongkai Chen[2], Zixin Luo[1], Lei Zhou[1], Yurun Tian[1], Mingmin Zhen[1], Tian Fang[1], David McKinnon[1], Yanghai Tsin[1], and Long Quan[1]

[1] Apple Inc.
{zixin_luo, zhou_lei, tian_ray, mingmin.zhen, fangtian, dmckinnon, ytsin, quan.long}@apple.com
[2] HKUST
hchencf@cse.ust.hk

Abstract. Generating robust and reliable correspondences across images is a fundamental task for a diversity of applications. To capture context at both global and local granularity, we propose ASpanFormer, a Transformer-based detector-free matcher that is built on hierarchical attention structure, adopting a novel attention operation which is capable of adjusting attention span in a self-adaptive manner. To achieve this goal, first, flow maps are regressed in each cross attention phase to locate the center of search region. Next, a sampling grid is generated around the center, whose size, instead of being empirically configured as fixed, is adaptively computed from a pixel uncertainty estimated along with the flow map. Finally, attention is computed across two images within derived regions, referred to as attention span. By these means, we are able to not only maintain long-range dependencies, but also enable fine-grained attention among pixels of high relevance that compensates essential locality and piece-wise smoothness in matching tasks. State-of-the-art accuracy on a wide range of evaluation benchmarks validates the strong matching capability of our method.

Keywords: Image Matching, Visual Localization, Pose Estimation, Transformer

1 Introduction

Image matching lays the foundation for various geometric computer vision tasks, including Structure from Motion (SfM) [1, 2], visual localization [3], and Simultaneous Localization And Mapping (SLAM) [4, 5]. As a widely accepted pipeline for image matching, cross-image correspondences are usually established by matching a set of detected and described sparse keypoints, such as SIFT [6], ORB [7], or their learning-based counterparts [8–12]. Despite its general effectiveness, this detector-based matching pipeline struggles in extreme situations, including large view point changes and textureless areas, due to the reliance on keypoint detector and context loss in feature description.

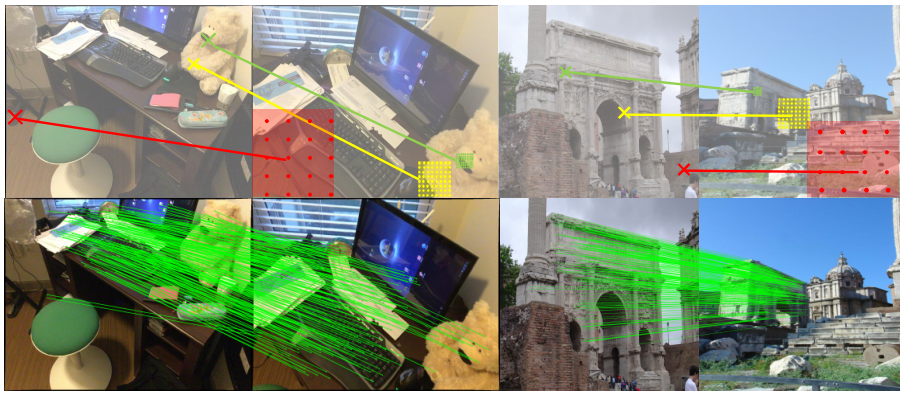

Fig. 1. An illustration of the proposed adaptive attention span (top row) and final dense matching results (bottom row). Particularly, in the top row, a rectangle with 8×8 uniform sampling grid is drawn to explain the position and size of adaptive attention span. In addition, three typical types of correspondences are visualized. Easy match in green with rich texture, which can be well localized and matched with small local contexts. Hard match in yellow with little texture, which requires larger contexts to guide matching. Impossible match in red in non-overlapping or occluded region, which has a very large attention span to avoid falsely fitting to certain regions. With this design, we enable Transformer to adaptively capture necessary context according to matching difficulty.

Concurrent with detector-based matching, another line of works [13–22] focus on generating correspondences directly from raw images, where richer context can be leveraged while keypoint detection step can be eschewed. Earlier works [16–18] in detector-free matching often rely on iterative convolution upon correlation or cost volume to discover potential neighbourhood consensus. Recently, some works [13, 14, 22] base their methods on Transformer [23, 24] backbone for better modeling of long-range dependencies. As a representative work, LoFTR utilizes self and cross attention blocks to update cross-view features, where Linear Transformer [25] is adopted to replace global full attention in order to achieve manageable computation cost. Although proven effective, a concern about LoFTR is the lack of fine-level local interaction among pixel tokens, which could limit its capability to extract highly accurate and well-localized correspondences. This concern is deepened by the findings [22] of Tang et al., which reveals that the cross attention map generated by LoFTR's Linear Transformer tends to diffuse among large areas instead of sharply focusing on actual corresponding regions.

To capture both global context and local details, we propose a Transformer-based detector-free matcher, equipped with a hierarchical attention framework. Our foundation processing blocks, referred to as Global-Local Attention (GLA) block, performs a coarse-level global attention at low resolution to acquire long-range dependencies, meanwhile, conducts fine-level local attention at high reso-

lution within only a concentrated region around a current correspondence found through dense flow prediction.

The key challenge for fine-level local attention is to determine the size of the attention span. A naive approach is to regard its size as a fixed hyper parameter, which, however, neglects the intrinsic matchability of different regions where the dependency of context varies. As shown in Figure 1, regions in rich texture areas can be easily matched within a small neighbourhood, while the textureless areas are more uncertain about their correspondences and require larger context for matching, not to mention areas that lie out of overlapping regions and are impossible to be matched. To mitigate this issue, we introduce an adaptive attention span driven by probabilistic modelling, which can be adjusted for different locations based on underlying matching difficulty. We summarize our contributions in three aspects:

- A hierarchical attention framework is proposed for feature matching, where attention operations are performed at different scales to enable both global context awareness and fine-grained matching.
- A novel uncertainty-driven scheme, based on probabilistic modelling of flow prediction, is proposed to adaptively adjust local attention span. Through this design, our network assigns varying size of contexts to different locations according to their essential matchability and context richness.
- State-of-the-arts results on extensive set of benchmarks are achieved. Our method outperforms both detector-free and detector-based matching baselines in two-view pose estimation. Further experiments on challenging visual localization also proves our method's potential to be integrated into c omplicated down-stream applications.

2 Related Works

2.1 Detector-Free Image Matching

Differing from detector-based matching methods, which typical involve detecting [11, 9, 10, 8], describing [26–28, 12, 29] and matching [30–36] a set of keypoints, detector-free matching consumes a pair of images and output correspondences in one shot. Thanks to the removal of keypoint detection stage, detector-free matching is able to capture richer contexts from original images, thus exhibits strong potential to match in extreme situations, such as low texture areas and repetitive patterns.

Despite the potential merits of detector-free matching, its popularity hardly outperforms detector-based methods during early deep learning times due to the intrinsic difficulties in robust and distinctive features. Recently, with the help of deep neural network, possibility is explored to build high performance detector-free matching frameworks based on deep features, which can roughly be classified into two categories: cost volume-based methods [16, 19, 15, 17, 18, 37] and Transformer-based methods [13, 14, 22]. Both kinds of methods leverage strong signals in deep features' correlation, either in form of correlation layer or

cross attention, to guide correspondence search and feature update. Our method follows works on Transformer-based methods and employs multilevel cross attention for mutual feature update, encoding two-view contexts into original features for both global and local consensus.

2.2 Global-Local Structure

Balancing receptive field and interaction granularity is a long-standing issue for both cost volume-based and Transformer-based matching. To ensure global receptive field, cost volume based methods are often designed to perform convolution on large global correlation volume, while Transformer-based methods need to conduct attention among all pixel tokens in image pairs. Due to the high cost of global interaction, the input features are usually downsampled into coarse resolution [18, 19, 14] or being projected into low rank [13], which to some degree limits the networks' capability for fined grained feature update.

Complementary to global interaction, some methods propose to perform local interaction only within a certain region instead of a global field, enabling to process fine level features given a limited computation budget. This practice is especially common in cost volume based methods and are referred to as local correlation layer [19, 38, 15, 39], where the cost volumes/vectors are only constructed around neighbourhood of each correspondence estimation. Intuitively, the idea of complementary global-local interaction can also be introduced to Transformer-based matcher. In our method, a global-local attention block is proposed for message passing across images, ensuring both global receptive field and fine level feature processing. Specially, instead of fixing span for local attention, we design an adaptive mechanism to determine the size of area that each pixel should attend to.

2.3 Flow Regression and Uncertainty Modeling

Flow maps depicts correspondence coordinates, which can either be absolute or relative, for each location in an image. Predicting correspondence coordinates from an image pair has been intensively investigated by works in optical flow estimation [40, 38, 41, 39] and general dense image matching [19, 15, 37]. In these works, the flow maps are regressed from structured correlation volumes which are implicitly position-aware. Recently, a Transformer-based method, COTR [14], proves that flows can also be retrieved from positional-embedded features after several turns of attention update.

Naturally, the reliability of flow estimation in each location is not equal and predicting associated confidence scores is essential for many scenarios. As an elegant framework for uncertainty prediction, some works [15, 42–45] propose to use probabilistic model to jointly explain both flows estimations and their confidence. Inspired by above works, we propose a network that regresses a flow map for each attention block to guide local attention region and adjust the attention span adaptively based on uncertainty prediction.

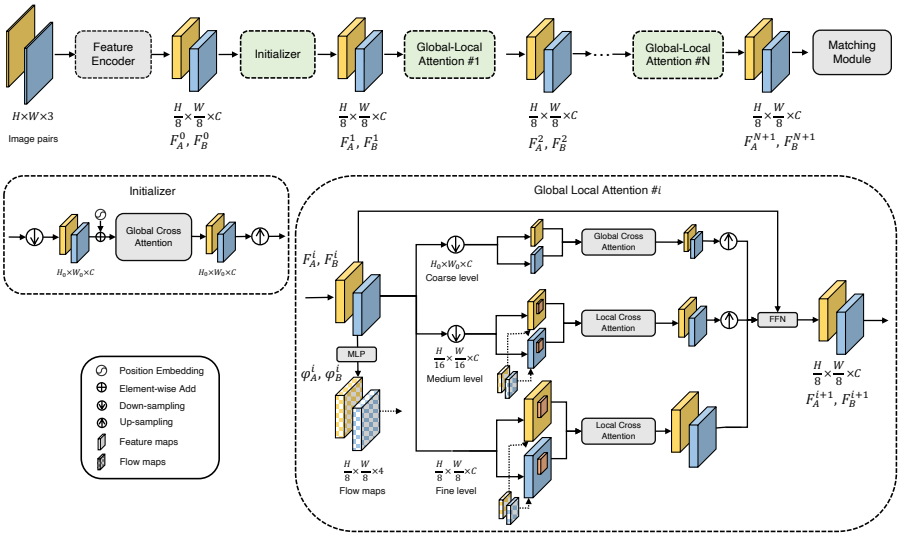

Fig. 2. We use CNN backbone to extract initial features. After initialization, the features are fed into iterative GLA blocks for updating. A matching module is used to determine final matches.

3 Methodology

We present an overview of our network structure in Figure 2. Taking an image pair I_A, I_B as input, our network produces reliable correspondences across images. The matching process starts with a CNN-based encoder to extract initial features F_A^0, F_B^0 for both images separately. After initialization, these features are turned into F_A^1, F_B^1 and fed into the proposed Adaptive Span Transformer (ASpanFormer) module for updating, which is composed of iterative global-local attention (**GLA**) blocks with hierarchical structure. Particularly, for each GLA block, we regress auxiliary flow maps ϕ_A, ϕ_B describing correspondence coordinates (flows) and their uncertainty. Instead of adopting these flow maps as our correspondence output, we use them to guide local cross attention, enabling adaptive local attention span according to matching uncertainty. After N GLA blocks, the updated features $F_A^{(N+1)}, F_B^{(N+1)}$ are use to construct coarse level matches, which will be further refined into final correspondences.

In the following part, we demonstrate the details of each individual block as well as the underlying insights.

3.1 Preliminary

Before introducing the structure of our network, we first clarify necessary notations and concepts.

Attention. As the key operation in Vision Transformer, attention is defined over a set of query (Q), key (K) and value (V) vectors as

$$M = \text{Att}(Q, K, V) = \text{softmax}(QK^T)V, \tag{1}$$

where Q, K, V are linear projections of upstream features F and M is retrieved message. More specially, in the context of cross attention, Q are derived from source features F_s and K, V vectors are derived from target features F_t. M is used to update source features F_s through a feed forward network (FFN), which involves concatenation, layer normalization and linear layers.

$$F_s^{i+1} = \mathbf{FFN}(F_S^i, M). \tag{2}$$

Typically, in each pass, the position of source/target features can be switched and cross attention is performed symmetrically.

Flow map. Flow maps $\phi_A, \phi_B \in R^{H\times W\times 2}$ depict the correspondence relationship between an image pair $I_A, I_B \in R^{H\times W}$, such that for any location (i, j) in an image, $I_A[i, j] \leftrightarrow I_B[\phi_A[i, j]], I_B[i, j] \leftrightarrow I_A[\phi_B[i, j]]$. Here, $\leftrightarrow$ denotes that the points on two sides are correspondences.

Instead of depicting simple correspondences, a stream of works [15, 42–45] proposes to model flow field with a probabilistic framework. Following these works, we model the flow field as a Gaussian distribution defined by a set of parameters. More specifically, assuming conditional independence among pixels, two flow maps $\phi_A, \phi_B \in R^{H,W,4}$ are predicted, such that $\phi[i, j] = [u_x^{ij}, u_y^{ij}, \sigma_x^{ij}, \sigma_y^{ij}]$, where (u_x^{ij}, u_y^{ij}) are estimated correspondence coordinates and $(\sigma_x^{ij}, \sigma_y^{ij})$ are standard deviations. The probability for $I_A[i, j] \leftrightarrow I_B[x, y]$ is given by

$$P(x, y|\phi_A[i, j]) = \frac{1}{2\pi\sigma_x^{ij}\sigma_y^{ij}}\exp(-\frac{(x-u_x^{ij})^2}{2{\sigma_x^{ij}}^2} - \frac{(y-u_y^{ij})^2}{2{\sigma_y^{ij}}^2}) \tag{3}$$

Instead of thresholding flow estimation with uncertainty, we use it to adjust the search region for subsequent network interaction, as described in later sections.

3.2 Feature Extractor

As the first part of our network, a convolutional neural network (CNN) is used to extract 1/8 down-sampled initial features $F_A, F_B \in R^{\frac{H}{8}\times\frac{W}{8}}$ for each image. As is shown in previous works [10, 8, 9, 12, 28, 27, 26, 29], CNN exhibits strong capability to capture local context and generates high-level features, which can be directly used to perform nearest neighbour matching. However, since these features are processed independently for each image and critical cross view contexts are missed. To enrich features with long range and cross view contexts, the initial features are further fed into our proposed Transformer module for updating.

3.3 Initialization

Our Transformer-module starts with a fast initialization block, which conducts (1) positional encoding and (2) two-view contexts initialization.

Positional encoding. As validated in Transformer networks [13, 30, 14], positional encoding is critical in maintaining spatial information for the flattened tokens. Following the same formulation in LoFTR [13], 2D sinusoidal signals in different frequencies are used to encode position information and are added to initial features. Specially, we apply normalization when testing resolution differs from training resolution. We provide more details in Appendix A.5.

Two-view contexts initialization. At each local attention phase, our network requires regressing an auxiliary flow map as guidance, which requires cross view contexts. To this end, we pass positional embedded features to a light-weight cross attention block. More specifically, These features are downsampled to low resolution H_0, W_0 and two global cross attention blocks are used for feature processing. After initialization, the features are upsampled back to original input resolution, denoted as F_A^1, F_B^1, and sent to iterative global-local attention blocks for further processing.

3.4 Global-Local Attention Block

The basic structure of our Transformer network is global-local attention (**GLA**) block. As is shown in Figure 2, for each GLA block, attention is performed upon a 3-level coarse-to-fine feature pyramid built by strided average pooling.

For the i-th GLA block, **global attention** is conducted on coarsest downsampled features in resolution $[H_0, W_0]$, while **local attention** with adaptive span is used to pass message between medium-resolution features in resolution $[\frac{H}{16}, \frac{W}{16}]$ and fine level features in resolution $[\frac{H}{8}, \frac{W}{8}]$. Note that we keep the coarsest resolution as a constant, making the complexity of global full attention unaffected by input size. Retrieved messages M^c, M^m, M^f from coarse/medium/fine level are upsampled to same $[\frac{H}{8}, \frac{W}{8}]$ resolution, concatenated and fused with an MLP to update source features.

$$M = \mathbf{MLP}(M^c || M^m || M^f), \tag{4}$$

$$F^{i+1} = \mathbf{FFN}(M, F^i). \tag{5}$$

The FFN in our network is defined as

$$\mathbf{FFN}(M, F) = F + \mathbf{LN}(\mathbf{Conv}_3(F||M)). \tag{6}$$

LN stands for layer normalization. Specially, we adopt a 3×3 convolution $\mathbf{Conv}_3$ in FFN for locality modeling, which compensate for the absence of self attention within each feature map. Empirically, we find 3×3 convolution in FFN works better than the combination of linear projection FFN and self attention, more details can be found in Appendix A.5.

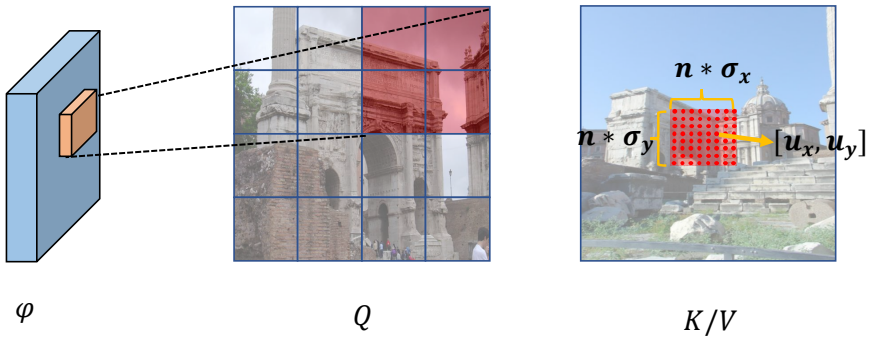

Fig. 3. Illustration Local cross attention. Query map Q are partitioned into cells in size $S \times S (S = 2$ in this case), retrieving prediction from flow map ϕ and generate attention span. Here we only show attention span for one cell(marked in red).

Local cross attention with adaptive attention span. To facilitate fine-grained attention with modest cost, we adopt local attention on medium and fine level feature maps, where attention span focuses on the neighbourhood regions around current correspondences estimation.

A key problem for local attention is how to define the size of neighbourhood region. A naive approach is to define neighbourhood with a fixed radius r for all pixels, neglecting the fact that the optimal attention span for different regions varies. For instance, it is sufficient to match regions with distinctive features using small contexts, while regions that are harder to match require larger contexts. Instead of using fixed attention span for all pixels, we propose to adaptively adjust the attention span according to the uncertainty of flow estimation. This design lets each area balance their local receptive fields with uncertainty awareness. Regions with high confidence in flow estimation can sharply focus on a small region for fine level matching, while larger contexts are extracted in low confidence areas for better convergence.

Formally, for the i-th GLA block, we first regress flow maps ϕ_A^i, ϕ_B^i from input features F_A^i, F_B^i in fine level with an MLP, while the medium level flow map are obtained by strided average pooling. For each scale level, we partition the corresponding query map Q into cells with size $S \times S$. For each cell, we use the mean flow estimation to generate a rectangle region upon K, V map and uniformly sample a fixed number of tokens. Attention is performed between each cell and the sampled tokens. The detailed process is defined in Algorithm 1. An illustration for local attention is given in Fig. 3. Since number of sampled tokens for each location is fixed, the whole process retains linear complexity.

3.5 Matches Determination

We inherit the scheme in LoFTR [13] to generate final correspondences, including a coarse matching stage and a sub-pixel refinement stage.

Algorithm 1 Local Cross Attention

Input: $Q, K, V \in R^{H\times W\times C}, \phi \in R^{H\times 4}$, span coefficient n, sample number w, window size S

Output: Retrieved message $M \in R^{H\times W\times C}$

Partition Q into cells set Q_p with window size $S \times S$, there will be $\frac{H}{S} \times \frac{W}{S}$ cells in total
$M = [\,]$
for each cell $Q_{pi} \in R^{S^2\times C}$ in Q_p **do in parallel**
    Retrieving flow $\phi_p \in R^{S^2\times 4}$ from flow map ϕ according to the location of Q_{pi}
    Let $[u_x, u_y, \sigma_x, \sigma_y] = \overline{\phi_p} = \sum_j \phi_p[j, :]$
    Let Γ be a rectangle area with center $[u_x, u_y]$, width $n * \sigma_x$ and height $n * \sigma_y$
    Uniformly sample w^2 tokens in Γ region from K, V, denoted as $K_\Gamma, V_\Gamma \in R^{w^2\times C}$
    Attention $m_i = \mathrm{Att}(Q_{pi}, K_\Gamma, V_\Gamma)$
    Append m_i to M
Reshaping M into $R^{H\times W\times C}$
return M

After being updated by N GLA blocks, we flatten the output features into $\tilde{F}_A \in R^{n\times c}, \tilde{F}_B \in R^{m\times c}$ and construct correlation matrix $C = \tau \tilde{F}_A \tilde{F}_B^T \in R^{n\times m}$, where τ is a temperature parameter and n, m are feature numbers of two images. By applying dual-direction softmax in both column/row dimensions, a score matrix is given by $S = \mathbf{softmax}_{row}(C) \cdot \mathbf{softmax}_{col}(C)$, from which we retain coarse-level matches M_c by mutual nearest neighbour (MNN) and filtering scores below a certain threshold θ. The coarse matches M_c are further fed into a correlation-based refinement block, which is the same with LoFTR [13], to obtain the final matching results.

3.6 Loss Formulation

We formulate the final loss from three parts, (1) coarse matches loss L_c, (2) fine-level loss L_f and (3) flow estimation loss L_{flow}

$$L = L_c + L_f + \alpha L_{flow}. \tag{7}$$

For coarse level loss L_c, the ground truth matches M_{gt} is determined by reprojection using depth and camera poses in datasets. We supervise the dual-softmax score matrix S with cross entropy loss

$$L_c = -\frac{1}{|M_{gt}|} \sum_{(i,j)\in M_{gt}} \log(S(i,j)). \tag{8}$$

The fine-level loss is supervised directly with L2-distance between each refined coordinates $M_f(i,j)$ and ground truth reprojection coordinates, which are further normalized by the coordinate variance.

For flow estimation supervision, we minimize the log-likelihood for each estimated distribution. Formally, given flow estimation Φ from each layer and ground

truth flow D^{gt}, L_{flow} is defined as

$$L_{flow} = -\frac{1}{|D^{gt}|}\sum_{ij} log(P(D_{ij}^{gt}|\Phi_{ij})). \tag{9}$$

In our implementation, this log-likelihood formulation can be further substituted and decomposed into a more compact form, which is elaborated in Appendix B.

3.7 Implementation Details

Our network shares the same ResNet-18 [46] CNN feature extractor with LoFTR. After feature extraction and flow initialization, we use 4 GLA blocks for updating. For adaptive attention span, we set $n = 5$, meaning that using 5 standard deviation to crop local neighbourhood region for each token. We uniformly sample 8×8 features in each cropped local region.

We train two different models specified for indoor and outdoor scenes respectively. Both models are optimized using Adam with learning rate 1×10^{-3} for 30 epochs on 8 V-100 GPUs. Indoor model is trained on ScanNet [47] dataset with batch size 24, where the training consumes 5 days. Outdoor model is trained on MegaDepth [48] with batch size 8, taking 2 days to converge. More details about implementation are introduced in Appendix A.3.

4 Experiments

In this section, we demonstrate the performance of our method on two-view pose estimation and visual localization tasks, among both indoor and outdoor scenes. Besides, we conduct ablation study to validate the effectiveness of key design components of our method.

4.1 Two-view Pose Estimation

We resort to two popular datasets, ScanNet [47] and MegaDepth [48], introduced below, to demonstrate the matching ability of our method in indoor scenes and outdoor scenes, respectively. We also provide additional results on YFCC100M [49] and Image Matching Challenge(IMC) 2022 in Appendix C.

Indoor two-view matching dataset. ScanNet dataset [47] is composed of 1613 sequences, each of which contains RGB images exposing large view changes and repetitive or textureless patterns, with ground-truth depth maps and camera poses associated. For fair comparison, we follow the same training and testing protocols used by SuperGlue [30] and LoFTR [13], where 230M and 1.5K image pairs are sampled for training and testing, respectively. In congruent with LoFTR, we resize all test images to 480×640.

Outdoor two-view matching dataset. MegaDepth [48] consists of 196 3D reconstructions from 1M Internet images, whose camera poses and depth maps are initially computed from COLMAP [1] and then refined as ground-truth.

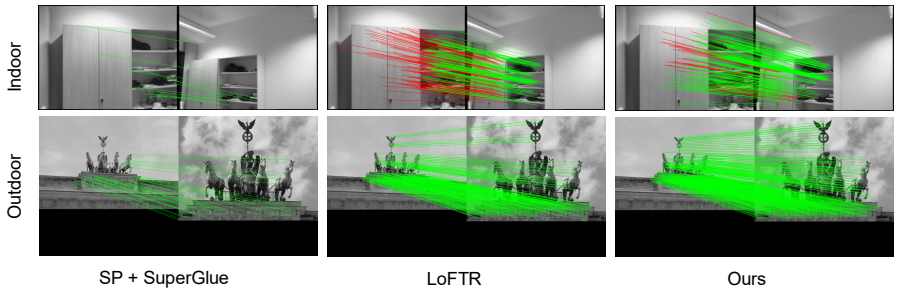

Fig. 4. Qualitative results of dense matching in different scenarios.

We perform two view pose estimation on 1.5k testing pairs. All test images are resized so that their longest dimension is 1152.

Evaluation protocols. Following previous works [30, 13], we train and evaluate our method separately on the two datasets. Two-view pose is recovered by solving essential matrix from correspondences produced, while pose accuracy is measured by AUC at multiple error thresholds (5°, 10° and 20°). A pose is only considered accurate if both its angular rotation error and translation error is under a certain threshold compared with ground-truth poses.

Comparative methods. We compare the proposed method with 1) detector-based approaches, including SuperGlue [30] and SGMNet [31] that are equipped with SuperPoint(SP) [9] as local feature extractor, 2) detector-free approaches, including DRC-Net [18], PDC-Net [15, 50], LoFTR [13], QuadTree Attention [22], MatchFormer [51] and DKM [52].

Results. As presented in Table 1 and Table 2, our method consistently achieves the best accuracy in both indoor and outdoor scenes. Visualization in Figure 4 qualitatively demonstrates our method performance against other matches. More visualizations are provided in Appendix D.

Table 1. Two-view pose estimation results on ScanNet dataset [47] in indoor scenes.

Method	Pose Estimation AUC		
	@5°	@10°	@20°
SP [9]+SuperGlue [30]	16.2	33.8	51.8
SP [9]+SGMNet [31]	15.4	32.1	48.3
DRC-Net [18]	7.7	17.9	30.5
PDC-Net+(H) [50]	20.2	39.4	57.1
LoFTR [13]	22.0	40.8	57.6
QuadTree [22]	24.9	44.7	61.8
MatchFormer [51]	24.3	43.9	61.4
DKM [52]	24.8	44.4	61.9
Ours	**25.6**	**46.0**	**63.3**

Table 2. Two-view pose estimation results on MegaDepth dataset [48] in outdoor scenes.

Method	Pose Estimation AUC		
	@5°	@10°	@20°
SP [9]+SuperGlue [30]	42.2	61.2	75.9
SP [9]+SGMNet [31]	40.5	59.0	73.6
DRC-Net [18]	27.0	42.9	58.3
PDC-Net+(H) [50]	43.1	61.9	76.1
LoFTR [13]	52.8	69.2	81.2
QuadTree [22]	54.6	70.5	82.2
MatchFormer [51]	53.3	69.7	81.8
DKM [52]	54.5	70.7	82.3
Ours	**55.3**	**71.5**	**83.1**

4.2 Visual Localization

Apart from evaluation on two-view pose estimation task, we further integrate our network into a visual localization pipeline, and use two popular datasets, InLoc [53] and Aachen Day-Night v1.1 [54, 3, 55] datasets, to demonstrate performance on multi-view matching in indoor scenes and outdoor scenes, respectively.

Indoor localization dataset. InLoc dataset [53] contains a database of 9,972 RGBD indoor images that are geometrically registered to form the reference scene model, while 329 RGB query images are provided for visual localization, annotated with manually verified camera poses. Great challenge is posed in matching textureless or repetitive patterns under large perspective differences.

Outdoor localization dataset. Aachen Day-Night v1.1 dataset [54] depicts a city whose reference scene model is built upon 6,697 day-time images. For visual localization, the dataset provides another 824 day-time images and 191 night-time images as queries. Great challenge is posed in identifying correspondences from, in particular, night-time images under extremely large illumination changes.

Evaluation protocols. We follow the instructions from Long-Term Visual Localization Benchmark [56] to compute query poses. For both datasets, we use pretrained HLoc [57] to retrieve candidate pairs, and recover camera poses with the model trained on MegaDepth dataset following SuperGlue [30] and LoFTR [13]. More details on localization pipeline are elaborated in Appendix A.4.

Results. On InLoc dataset, as shown in Table 3, our methods achieves overall best results compared with multiple comparative methods. On Aachen V1.1, as shown in Table 4, our method outperforms all other methods except SuperGlue. We partially ascribe this to the fact that we use only coarse matches for database reconstruction (see Appendix A.4.), causing localization error that harms pose estimation. In general, our method generalizes well in practical pipelines.

Table 3. Visual localization results on InLoc dataset [53].

Method	DUC1	DUC2
	(0.25m,2°) / (0.5m,5°) / (1m,10°)	
HLoc [57] + SP [9]+SuperGlue [30]	49.0 / 68.7 / 80.8	53.4 / **77.1** / 82.4
HLoc [57] + LoFTR [13]	47.5 / 72.2 / 84.8	54.2 / 74.8 / **85.5**
HLoc [57] + Ours	**51.5** / **73.7** / **86.4**	**55.0** / 74.0 / 81.7

Table 4. Visual localization results on Aachen V1.1 dataset [54].

Method	Day	Night
	(0.25m,2°) / (0.5m,5°) / (1m,10°)	
Localization with matching pairs provided in dataset		
R2D2 [8] + NN	-	71.2 / 86.9 / 98.9
ASLFeat [10] + NN	-	72.3 / 86.4 / 97.9
SP [9]+SuperGlue [30]	-	73.3 / 88.0 / 98.4
SP [9]+SGMNet [31]	-	72.3 / 85.3 / 97.9
Localization with matching pairs generated by HLoc		
SP [9]+SuperGlue [30]	**89.8** / **96.1** / **99.4**	77.0 / 90.6 / **100.0**
LoFTR [13]	88.7 / 95.6 / 99.0	**78.5** / 90.6 / 99.0
Ours	89.4 / 95.6 / 99.0	77.5 / **91.6** / 99.5

4.3 Ablation Study

To validate the effectiveness of different design components of our method, we conduct ablation experiments on ScanNet dataset [47] following the same setting in Section 4.1. Specifically, we compare three designs of attention structure:

- *Single-Level Attn.*: A design with only global attention at coarsest feature maps without the need of flow estimation. In this design, global context is well captured, whereas essential locality in motion smoothness is omitted and fine-grained message exchange becomes difficult.
- *Multi-Level Attn.*: A design with the hierarchical attention framework proposed in this paper, except that the size of local attention span is fixed to 13 px, i.e., the statistical mean of the adaptive attention span used in our network.
- *Adaptive Span Attn.*: Our full design that enables hierarchical attention with adaptive attention span. By this means, the need of context for different pixels is dynamically decided regarding different matchability.

As presented in Table 5, both hierarchical global-local attention and adaptive attention span improve overall performance by a considerable margin, validating the essentiality of our network designs.

Table 5. Ablation study on ScanNet dataset [47].

Method	Pose Estimation AUC		
	@5°	@10°	@20°
Single-Level Attn.	22.65	40.72	59.06
Multi-Level Attn.	24.85	44.86	62.71
Adaptive Span Attn.	**25.61**	**46.04**	**63.33**

Table 6. Flow estimation accuracy.

Stage	<6px (%)	5σ (px)		
		Matchable	Unmatchable	Total
Iter#1	69.1	9.2	19.4	13.4
Iter#2	71.2	8.2	20.2	12.5
Iter#3	72.0	7.8	23.8	12.6
Iter#4	72.3	7.7	27.1	13.3

4.4 Understanding ASpanFormer

Flow estimation. We analyze the flow estimation through multiple iterations. As shown in Table 6, precision of flow regression is gradually improved as attention iterations are performed and converges after four iterations.

As for uncertainty estimation, we split all pixels into two categories, matchable and unmatchable pixels, identified by ground-truth camera poses and depths, and report their mean standard deviation (σ). On one hand, mean σ decreases with iterations for matchable pixels, as the network becomes more certain about its flow prediction in later stages. On the other hand, the network gradually increases uncertainty values of unmatchable pixels to prevent over-confidence to a certain region.

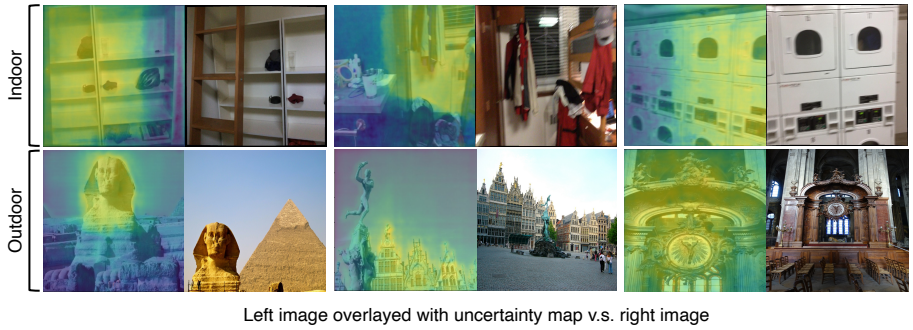

Fig. 5. Visualization of uncertainty map which is predicted along with flows, warmer color indicates smaller uncertainties.

Uncertainty map. In Figure 5, we provide visualization of uncertainty map of flow prediction. Overlapping and non-overlapping regions are firstly distinguished, while uncertainty values in textureless regions are usually larger, indicating context of larger size is required during cross attention.

Runtime evaluation. We evaluate the runtime of proposed method and compare it with LoFTR [13] where both methods apply Transformer backend. The runtime speed is tested on 100 randomly sampled ScanNet image pairs (640×480) with a NVIDIA V100 GPU. Runtime differs only on *Attention Module* compared with LoFTR, as we adopt the same Local Feature *CNN* backbone and coarse-to-fine matching module. As shown in Table 7, the proposed method is overall slightly slower than LoFTR due to the more complicated attention operation.

Table 7. Runtime speed evaluated on 640×480 images.

Stage	**Runtime (ms)**	
	LoFTR	Ours
Local Feature CNN	32.2	32.2
Attention Module	24.6	40.5
Matching	40.9	40.8
Total	97.7	113.5

5 Conclusion

In this paper, we have proposed a novel Transformer framework based on feature hierarchy, whose attention span is adaptively decided so as to acquire capabilities to capture both long-term dependencies as well as fine-grained details in local regions. State-of-the-art results validates the effectiveness of our method. With more engineering optimizations, we are looking forward to wider application of our method in real use.

References

1. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR. (2016)
2. Resindra, A., Torii, A., Okutomi, M.: Structure from motion using dense cnn features with keypoint relocalization. IPSJ Transactions on Computer Vision and Applications (2018)
3. Sattler, T., Weyand, T., Leibe, B., Kobbelt, L.: Image retrieval for image-based localization revisited. In: BMVC. (2012)
4. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular slam system. IEEE transactions on robotics (2015)
5. Mur-Artal, R., Tardos, J.: ORB-SLAM2: an open-source slam system for monocular, stereo and rgb-d cameras. IEEE Transactions on Robotics (2016)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
7. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.R.: ORB: An efficient alternative to sift or surf. In: ICCV. (2011)
8. Revaud, J., Weinzaepfel, P., De Souza, C., Pion, N., Csurka, G., Cabon, Y., Humenberger, M.: R2D2: repeatable and reliable detector and descriptor. In: NeurIPS. (2019)
9. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: CVPRW. (2018)
10. Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L.: Aslfeat: Learning local features of accurate shape and localization. In: CVPR. (2020)
11. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint description and detection of local features. In: CVPR. (2019)
12. Luo, Z., Shen, T., Zhou, L., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L.: Contextdesc: Local descriptor augmentation with cross-modality context. In: CVPR. (2019)
13. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: CVPR. (2021)
14. Jiang, W., Trulls, E., Hosang, J., Tagliasacchi, A., Yi, K.M.: COTR: Correspondence transformer for matching across images. In: CVPR. (2021)
15. Truong, P., Danelljan, M., Gool, L.V., Timofte, R.: Learning accurate dense correspondences and when to trust them. In: CVPR. (2021)
16. Rocco, I., Cimpoi, M., Arandjelovi, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. In: NeurIPS. (2018)
17. Rocco, I., Arandjelović, R., Sivic, J.: Efficient neighbourhood consensus networks via submanifold sparse convolutions. In: ECCV. (2020)
18. Li, X., Han, K., Li, S., Prisacariu, V.: Dual-resolution correspondence networks. In: NeurIPS. (2020)
19. Truong, P., Danelljan, M., Timofte, R.: GLU-Net: Global-local universal network for dense flow and correspondences. In: CVPR. (2020)
20. Min, J., Cho, M.: Convolutional hough matching networks. In: CVPR. (2021)
21. Shen, X., Darmon, F., Efros, A., Aubry, M.: Ransac-flow: Generic two-stage image alignment. In: ECCV. (2020)
22. Tang, S., Zhang, J., Zhu, S., Tan, P.: Quadtree attention for vision transformers. In: ICLR. (2021)

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. (2017)
24. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR. (2020)
25. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: ICML. (2020)
26. Mishchuk, A., Mishkin, D., Radenović, F., Matas, J.: Working hard to know your neighbor's margins:local descriptor learning loss. In: NeurIPS. (2017)
27. Tian, Y., Fan, B., Wu, F.: L2-net: Deep learning of discriminative patch descriptor in euclidean space. In: CVPR. (2017)
28. Luo, Z., Shen, T., Zhou, L., Zhu, S., Zhang, R., Yao, Y., Fang, T., Quan, L.: Geodesc: Learning local descriptors by integrating geometry constraints. In: ECCV. (2018)
29. Wang, Q., Zhou, X., Hariharan, B., Snavely, N.: Learning feature descriptors using camera pose supervision. In: ECCV. (2020)
30. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR. (2020)
31. Chen, H., Luo, Z., Zhang, J., Zhou, L., Bai, X., Hu, Z., Tai, C.L., Quan, L.: Learning to match features with seeded graph matching network. In: ICCV. (2021)
32. Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., Chen, Y., Quan, L., Liao, H.: Learning two-view correspondences and geometry using order-aware network. In: ICCV. (2019)
33. Yi*, K.M., Trulls*, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. In: CVPR. (2018)
34. Sun, W., Jiang, W., Tagliasacchi, A., Trulls, E., Yi, K.M.: Attentive context normalization for robust permutation-equivariant learning. In: CVPR. (2020)
35. Cavalli, L., Larsson, V., Oswald, M.R., Sattler, T., Pollefeys, M.: Handcrafted outlier detection revisited. In: ECCV. (2020)
36. Bian, J., Lin, W.Y., Liu, Y., Zhang, L., Yeung, S.K., Cheng, M.M., Reid, I.: GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence. IJCV (2020)
37. Truong, P., Danelljan, M., Gool, L., Timofte, R.: Gocor: Bringing globally optimized correspondence volumes into your neural network. In: NeurIPS. (2020)
38. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR. (2017)
39. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. (2020)
40. Fischer, P., Dosovitskiy, A., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV. (2015)
41. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: CVPR. (2018)
42. Zhou, L., Luo, Z., Shen, T., Zhang, J., Zhen, M., Yao, Y., Fang, T., Quan, L.: Kfnet: Learning temporal camera relocalization using kalman filtering. In: CVPR. (2020)
43. Gast, J., Roth, S.: Lightweight probabilistic deep networks. In: CVPR. (2018)
44. Ilg, E., iek, z., Galesso, S., Klein, A., Makansi, O., Hutter, F., Brox, T.: Uncertainty estimates and multi-hypotheses networks for optical flow. In: ECCV. (2018)
45. Danelljan, M., Gool, L., Timofte, R.: Probabilistic regression for visual tracking. In: CVPR. (2020)

46. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)
47. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. (2017)
48. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: CVPR. (2018)
49. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: The new data in multimedia research. Communications of the ACM (2016)
50. Truong, P., Danelljan, M., Timofte, R., Van Gool, L.: PDC-Net+: Enhanced probabilistic dense correspondence network. Preprint (2021)
51. Wang, Q., Zhang, J., Yang, K., Peng, K., Stiefelhagen, R.: Matchformer: Interleaving attention in transformers for feature matching. Preprint (2022)
52. Edstedt, J., Wadenbäck, M., Felsberg, M.: Deep kernelized dense geometric matching. Preprint (2022)
53. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: Inloc: Indoor visual localization with dense matching and view synthesis. In: CVPR. (2018)
54. Zhang, Z., Sattler, T., Scaramuzza, D.: Reference pose generation for long-term visual localization via learned features and view synthesis. IJCV (2021)
55. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T.: Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In: CVPR. (2018)
56. Toft, C., Maddern, W., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Pajdla, T., et al.: Long-term visual localization revisited. TPAMI (2020)
57. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: CVPR. (2019)