Local Representation is Not Enough: Soft Point-wise Transformer for Descriptor and Detector of Local Features

Zihao Wang[1] , **Xueyi Li**[2] and **Zhen Li**[1*]

[1] School of Automation, Beijing Institute of Technology, China
[2] School of Computer Science and Technology, Beijing Institute of Technology, China
{zhwang1694, xueyili, zhenli}@bit.edu.cn

Abstract

Significant progress has been witnessed for the descriptor and detector of local features, but there still exist several challenging and intractable limitations, such as insufficient localization accuracy and non-discriminative description, especially in repetitive- or blank-texture regions, which haven't be well addressed. The coarse feature representation and limited receptive field are considered as the main issues for these limitations. To address these issues, we propose a novel Soft Point-Wise Transformer for Descriptor and Detector, simultaneously mining long-range intrinsic and cross-scale dependencies of local features. Furthermore, our model leverages the distinct transformers based on the soft point-wise attention, substantially decreasing the memory and computation complexity, especially for high-resolution feature maps. In addition, multi-level decoder is constructed to guarantee the high detection accuracy and discriminative description. Extensive experiments demonstrate that our model outperforms the existing state-of-the-art methods on the image matching and visual localization benchmarks.

1 Introduction

Establishing accurate correspondences among images plays a crucial role in many Computer Vision tasks, including but not limited to wide-baseline stereo, image retrieval, visual localization, Structure-from-Motion and 3D construction. Such correspondences are generally estimated by matching local features, which comprise keypoints detection and description. Keypoints detection is to predict the coordinate of the keypoint in the image, and the description is to generate a vector describing the image patch around the keypoint. However, environmental changes, including viewpoint and illumination, make the pipeline particularly challenging.

One of the key challenges in keypoints detection and description is the local representation short of identification which is derived from limited receptive field, especially in the blank ground or repetitive texture such as white-black chessboard. Many existing methods design an extra block to identify such regions and filter them when detecting [Revaud *et al.*, 2019]. While the paradox is that the typical repetitive substance, i.e., chessboard is widely applied in the camera calibration [Zhang, 2000], which requires rigorous accurate correspondences. The critical difference is that the keypoints description for calibration is the relative representation based on other adjacent-to-remote keypoints and global information.

Therefore, it is conceptually considered that the limited local representation for description of local features is not enough and global contextual information is as important to descriptor. Inspired by the Transformer's success in NLP [Vaswani *et al.*, 2017], we propose an elaborate Transformer structure to capture long-range dependencies, enriching the representation of local features and fixing the matching issues.

Another key challenge of keypoints detection and description is to coordinately solve two subtasks, *i.e.*, keypoints localization and classification. The former requires the model to capture keypoints position accurately, while the latter expects the model to extract high-level semantic information of the keypoints. Recent joint detection and description methods extract keypoints from the deep but coarse feature maps, leading to defective localization accuracy. Therefore a transformer pyramid is conducted to fix such issues. Three kinds of attention modules are developed to mine cross-level and intrinsic dependencies, enabling interacting features across space. The multi-level descriptor and detector based on the pyramid promises reliable pixel-level prediction.

Furthermore, the high-resolution feature maps in the shallow levels require heavy computation and memory cost, limiting the potential benefit of transformer in practical application. So most methods usually adopt the attention operation on the deeper coarse feature maps to economize the computation sources. Beneficial from the soft point-wise selection module, we take the detected keypoints as the Keys set in our soft point-wise transformer, so as to decrease the dense affinity matrix complexity from $O(n^2)$ to $O(const \times n)$, squeezing the main cost in the transformer.

The main contributions of this paper are summarized as follows. **Firstly**, an attention-based transformer is developed to capture long-range dependencies, which is crucial for

*Corresponding author.

generating discriminative description. **Secondly**, the cross-scale attention module and multi-level decoder are conducted to predict more accurate pixel-level scale-invariant keypoints detection. **Thirdly**, we propose the novel soft point-wise transformer, leveraging the detected keypoints to decrease the memory and computation complexity remarkably. **Lastly**, the learned network significantly outperforms prior state-of-the-art methods.

2 Related Works

In this section, we give a brief review of local features learning based on CNNs and computer vision transformer.

Joint local features learning. Recently, the increasing attention has been focused on the joint learning of feature descriptor and detector. In terms of descriptor learning, the ranking loss [Tian *et al.*, 2017; He *et al.*, 2018] has been primarily used as a *de-facto* standard. However, there exist some conflicts between descriptor and detector such as big-or-small receptive field and deep-or-shallow features.

To break through the limitation of restricted receptive field, D2Net [Dusmanu *et al.*, 2019] used the deep stacked convolutional network as backbone and detected-and-described upon the last feature maps. R2D2 [Revaud *et al.*, 2019] utilized dilated convolutions to improve the keypoints localization accuracy and generate pixel-level description, while limited in the mutual-vision boundary area. More recent ASLFeat [Luo *et al.*, 2020] used multi-level keypoints predictions to restore spatial resolution and low-level details.

Visual transformer. Transformer [Vaswani *et al.*, 2017] and its variants have proven its success of unsupervised or self-supervised pertaining frameworks in various NLP tasks. Therefore, there are many attempts to explore the benefits of Transformer in computer vision tasks [Li *et al.*, 2020]. DANet [Fu *et al.*, 2019] developed the context information by combining spatial and channel attention in the scene segmentation. Non-local Networks [Wang *et al.*, 2018] utilized a self-attention mechanism, enabling a single feature from any position to perceive features of all the other positions, thus harvesting full-image contextual information. Recent methods also attempt to replace the convolutional neural network with transformer pipeline, like ViT [Dosovitskiy *et al.*, 2020] in image classification, DETR [Carion *et al.*, 2020] in object detection and SETR [Zheng *et al.*, 2020] in semantic segmentation. While there exists few related work in the descriptor and detector of local features.

3 Methodology

Two ingredients are essential for adopting transformer on joint local feature learning: (1) an architecture that outputs keypoints detection and description simultaneously; (2) attention optimization for efficient contextual information capture.

3.1 Architecture

As illuminated in Figure 1, the overall architecture of soft point-wise transformer for description and detection of local features is designed as an encoder-decoder pipeline.

Feature Uniformization. To exploit the inter-dependencies between channel maps, a feature uniformization module is built at first. Given a local feature $F \in \mathbb{R}^{C\times H\times W}$, we first reshape F to $\mathbb{R}^{C\times N}$, and then perform a matrix multiplication between the F and the transpose of F to compute the channel attention map as:

$$X_{ji} = \frac{exp(F_i \cdot F_j^\top)}{\sum_{i=1}^{C} exp(F_i \cdot F_j^\top)}, \tag{1}$$

in which X_{ji} measures the i^{th} channel's impact on the j^{th} channel. Then we perform the weighted element-wise sum operation and 1×1 convolution to obtain the fixed dimension feature map $e \in \mathbb{R}^{C\times H\times W}$:

$$e_j = conv\left(\left(\gamma \sum_{i=1}^{C} X_{ji}F_i\right) + F_j\right), \tag{2}$$

where the γ is a learning scale parameter. The final feature of each channel is a weighted sum of features of all channels and original features, which models the long-range dependencies between feature maps, improving feature discriminability.

Cross-scale Attention. By mapping each point's representation into a latent fixed dimensional embedding space, we obtain a 1D sequence of point embeddings for a certain scale of the input image I. To encode the point spatial information, we learn a specific embedding p_i for every location i with a Multi-Layer Perception, which is added to e_i to form the final sequence input $E = \{e_1 + p_1, e_2 + p_2, ..., e_N + p_N\}$. Therefore, the spatial information is kept through the orderless self-attention and residual fusion.

Following the non-local operation [Wang *et al.*, 2018], we define the generic attention operation as:

$$y_i = \frac{\sum_j exp(Q_i \cdot K_j^\top)V_j}{\sum_j exp(Q_i \cdot K_j^\top)}. \tag{3}$$

Here i is the index of the output position and j is the index that enumerates all possible positions. The $\{Q, K, V\}$ represents the query, key and value for the attention, computed as $\{EW_Q, EW_K, EW_V\}$. $\{W_Q, W_K, W_V\} \in \mathbb{R}^{C*d}$ are the learnable parameters of three linear projection layers and d is the dimension of $\{Q, K, V\}$.

When the $\{Q, K\}$ comes from the same feature map, we call it as In-Scale attention. We further extend the K from the deeper or shallower feature maps, and we call it as Up-Scale and Down-Scale attention, respectively. The Up-Scale attention is developed to enrich the high-level feature representation of "patch" with the lower-level feature representation of "point". And the Down-Scale attention is developed in the opposite direction. The Up-Scale and Down-Scale attention jointly mine the cross-level dependencies to enrich the local feature maps from shallow to deep layers.

Residual Fusion. The In-Scale and Cross-scale attention generate three intensive feature maps with the same dimension, exploiting different scale information independently. Then we fuse these separate feature maps into a comprehensive feature map. Different from simply adding or concentrating them together [Lin *et al.*, 2017], we propose the residual fusion block to better combine features.

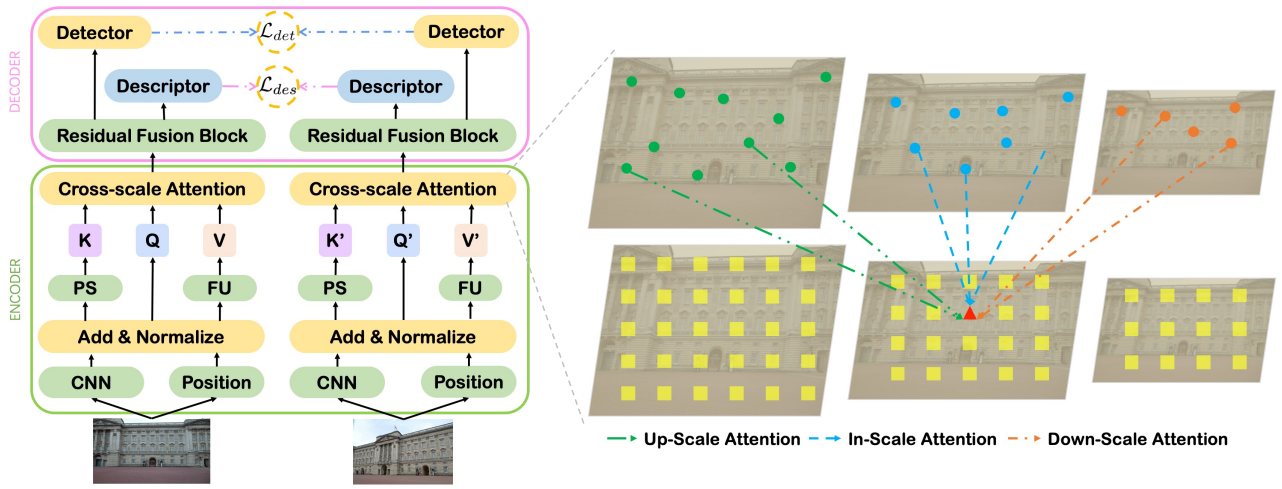

Figure 1: **Overall structure of our proposed SPTD2 network.** The network is extended into a Siamese network with the image pair as input during the training time. All points are taken as the Q set, while only the top-k keypoints are selected as the K set. PS and FU represents the point-wise selection and feature uniformization. CNN and Position is the feature embedding and position embedding for local points.

To allow the network to concentrate on more discriminative features, we first compute the residual between features $F_D \in \mathbb{R}^{d\times H\times W}$ from down-scale attention and features $F_I \in \mathbb{R}^{d\times H\times W}$ from in-scale attention. Then we can obtain the synthetic bottom-to-top representation $\hat{F}_I$:

$$\hat{F}_I = conv(F_I - F_D) + F_I. \tag{4}$$

Intuitively, the residual feature represents the abundant shape details like corner, edge and blob existing in shallow layer while degraded in the deep representation.

Similar operation is also adopt between the updated features $\hat{F}_I$ and features $F_U \in \mathbb{R}^{d\times H\times W}$ from up-scale attention:

$$P = conv(\hat{F}_I - F_U) + \hat{F}_I. \tag{5}$$

Finally we obtain the fusion feature maps $P \in \mathbb{R}^{d\times H\times W}$ fed into the decoder to output the keypoints detection and description. The residual block allows the network to focus on only the distinct information among different levels, while passing the common knowledge, enabling a more discriminative residual feature learning compared with trivial adding or concatenating.

Multi-level Decoder. Above feature uniformization, cross-scale attention and residual fusion modules make up the encoder of the transformer structure. As illuminated in the Figure 1, a dual-head decoder, *i.e.*, descriptor and detector is adopt on multi-level feature maps to extract multi-scale descriptions and keypoints. The detector and descriptor simultaneously output the 3D description $D \in \mathbb{R}^{d\times H\times W}$ and the detection score map $S \in [0,1]^{H\times W}$.

Descriptor. We set the 3D tensor P as a dense set of descriptor vectors D. These descriptor vectors can be readily compared between images to establish correspondences using the Euclidian distance with the hypothesis that the same keypoints will produce similar descriptors even in different conditions. In practice, a channel-wise L2-Normalization is applied to generate more robust feature presentation prior to comparing them.

$$D_{ij} = \frac{P_{ij}}{\|P_{ij}\|_2}, \tag{6}$$

with $i = 1\ldots H$ and $j = 1\ldots W$.

Detector. We also suppose that the 3D tensor P as a collection of 2D response maps at different channels. These detection score maps are analogous to the Difference-of-Gaussian (DoG) response maps obtained in Scale Invariant Feature Transform (SIFT). In practice, an element-wise square operation followed by a 1×1 convolution and softmax function are adopt to obtain the detection response score S of each descriptor.

$$S_{ij} = \theta\left(conv(P_{ij}^2)\right), \tag{7}$$

with $i = 1\ldots H$ and $j = 1\ldots W$, where the $\theta(*)$ represents the softmax operation. Only the locations with high confidence are selected as keypoints. Similar to multi-scale object detection, a non-maximum suppression (NMS) is applied to remove the detection points that are spatially too close.

Note that we build the feature maps from the encoder as a feature pyramid. The dual-head multi-level decoder extracts the final results upon the multi-level pyramid.

3.2 Soft Point-wise Attention

Recent works show that the keypoints located in the uniform and even well textured regions like tree leafages or ocean waves, could lead to bad matching[Revaud *et al.*, 2019]. So some works learn to distinguish such regions and filter them as less discriminative keypoints. While directly deleting such keypoints will result in defective detection accuracy and discontinuities especially in large repetitive regions.

The attention module and position encoding will improve the discrimination of the local features representation. While the original attention module needs to generate enormous affinity matrix to measure the relationships with the complexity of $O(N^2)$, where N is the number of input points. High-

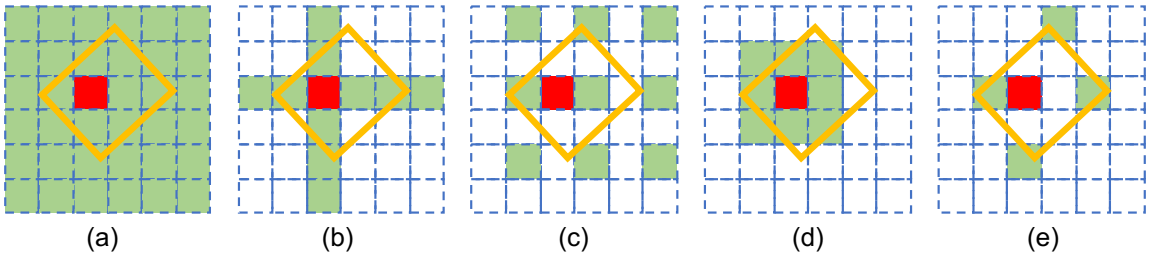

Figure 2: **Comparisons with other efficient attention module.** From left to right: (a) Original attention module, (b) CCNet, (c) ISSA, (d) Locality-constrained FPT, (e) Our SPT. The red grid is the query, the green grids represent the keys set and the yellow is the instance in the image. Our soft point-wise transformer can choose the most representative keypoints as the Keys set compared to other sota methods.

resolution feature maps are essential for accurate keypoints localization, taking heavy computation and memory cost.

To address the above mentioned issue, our motivation is to replace the common single dense feature maps with sparse representative features. Without loss of generality, we propose the soft point-wise attention module which aggregates contextual information with the most representative keypoints, greatly reducing the complexity from $O(N^2)$ to $O(const \times N)$.

The key of the soft point-wise attention is the soft keypoints selection, similar to our detector head but more efficient. We require a point (i, j) being selected by a hard dual-maximum strategy, *i.e.*, the feature value in the location (i, j) of channel k is the local maximum both in spatial-wise and channel-wise. To be amenable for back-propagation during the training procedure, the above selection procedure is softened as follows.

The soft spatial detection score and soft channel score are defined as:

$$\alpha_{ij}^k = \frac{\exp(F_{ij}^k)}{\sum_{(i',j')\in\mathcal{N}(i,j)}\exp(F_{i',j'}^k)}, \beta_{ij}^k = \frac{F_{ij}^k}{\max_t F_{ij}^t}, \quad (8)$$

where F is the feature map and $\mathcal{N}(i, j)$ is the set of 8 neighbors of the pixel (i, j). The soft selection takes both scores into account and performs an image-level normalization:

$$s_{ij} = \frac{\max_k(\alpha_{ij}^k \cdot \beta_{ij}^k)}{\sum_{(i',j')}\max_{k'}(\alpha_{i'j'}^{k'} \cdot \beta_{i'j'}^{k'})}. \quad (9)$$

Only the locations with high confidence (greater than the keypoints detection threshold) will be selected as the keys set $\hat{K} = \phi(K) \in \mathbb{R}^{C\times const}$, where $\phi(*)$ is the soft keypoint selection operation. And the attention in Eq.(3) with soft point-wise selection is computed as:

$$y_i = \frac{\sum_j exp(Q_i \cdot \phi(K)_j^\top)V_j}{\sum_j exp(Q_i \cdot \phi(K)_j^\top)}. \quad (10)$$

In addition, we compare our soft point-wise attention block with the original non-local block and other optimization methods [Huang *et al.*, 2019b; Huang *et al.*, 2019a; Zhang *et al.*, 2020] in Figure 2.

3.3 Implementation

Training. During the training time, the SPTD2 will be extended into a Siamese Network to simultaneously generate the keypoints detection score maps $\{S, S'\}$ and descriptors $\{D, D'\}$ of the correspond image pair $\{I, I'\} \in \mathbb{R}^{3\times H\times W}$. Our SPTD2 is independent of the backbone network, we use the VGG16_BN to evaluate our models during the experiments. In practice, the first feature map $F^1 \in \mathbb{R}^{C\times H/2\times W/2}$ and the feature maps smaller than $\mathbb{R}^{C\times H/16\times W/16}$ are dropped when building the transformer pyramid. For every input image pair, we select a random 200×200 crop centered around one correspondence.

Testing. During the testing time, the single image is fed into the model to generate the detection score maps and descriptors with the original resolution. All detection results will be aligned with the original image resolution and the descriptions are then bilinear interpolated at the refined positions. A non-maximum suppression is also applied on the multi-head detection score maps to remove the overlapping keypoints.

Loss design. As illuminated in the Figure 1, the loss function integrates the detection loss $\mathcal{L}_{det}$ and the description loss $\mathcal{L}_{des}$. The detection loss is formulated as:

$$\mathcal{L}_{det}(I, I') = \sum\nolimits_l w_l(\mathcal{L}_c(S^l, S'^l) + r(\mathcal{L}_p(S^l) + \mathcal{L}_p(S'^l))), \quad (11)$$

where the $\mathcal{L}_c$ computes the cosine similarity of the correspond detection score maps and the $\mathcal{L}_p$ tries to maximize the local peak of the detection score maps [Revaud *et al.*, 2019].

The description loss is written as:

$$\mathcal{L}_{des}(I, I') = \sum\nolimits_l w_l \sum_{c\in C} \frac{S_c^l S'^l_c}{\sum_{q\in C} S_q^l S'^l_q}\mathcal{M}(d_c^l, d'^l_c), \quad (12)$$

where C is the correspondences between I and I', S_c and S'^l_c are their detection scores, d_c^l and d'^l_c are their corresponding descriptors, and the $\mathcal{M}(*)$ is the circle loss [Sun *et al.*, 2020] for representation learning.

The final loss function is formulated as $\mathcal{L}_{det} + \lambda\mathcal{L}_{des}$.

4 Experiments

In this section, we evaluate the performance of the proposed model on the image matching and visual localization tasks. We show that the our model can achieve state-of-the-art performance on these tasks. Moreover, extensive experiments for ablation study show that our SPTD2 is effective.

Method	*Pub.*	*#Features*	*#Matches*	*AUC*		
				2px	5px	10px
SIFT	IJCV	4.1K	–	39.49	49.57	55.15
HesAff + RootSIFT	NIPS17	6.7K	2.9K	39.99	52.25	60.40
HAN + HN++	ECCV18	3.9K	2.0K	42.61	56.85	65.50
LF-Net	NIPS18	0.5K	0.2K	38.74	48.69	53.59
SuperPoint	CVPR18	1.7K	0.9K	44.08	59.04	68.09
DELF	ICCV17	4.6K	1.9K	44.73	49.70	58.91
SIFT + ContextDesc	CVPR19	4.1K	1.7K	47.23	58.25	65.33
D2-Net MS	CVPR19	4.9K	1.7K	19.49	37.78	56.17
R2D2 MS	NIPS19	4.9K	1.7K	43.35	64.17	75.18
SIFT + LISRD	ECCV20	4.1K	–	48.12	57.80	62.50
DISK	NIPS20	7.7K	3.9K	–	69.80	–
Key.Net	ICCV19	–	–	40.87	56.04	65.30
D2-Net + Ref	ECCV20	–	–	54.24	67.62	75.57
ASLFeat MS	CVPR20	4.8K	2.1K	50.10	66.93	76.90
Ours	–	5.0K	1.9K	**56.20**	**72.17**	**79.80**

Table 1: **Comparisons on HPatches with the area under the overall curve (AUC) up to 2, 5 and 10 pixels error threshold**. Our SPTD2 reaches the state-of-the-art on all thresholds benefiting by multi-level detection and feature fusion description. Most results are provided by the authors, which explains why some data are missing.

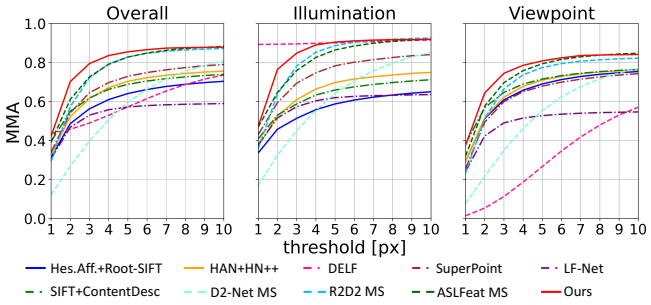

Figure 3: The curves of **mean matching accuracy (MMA)** evaluated at multiple error thresholds on *HPatches* dataset. "MS" denotes that the multi-scale inference is enabled. Note that several methods in Table 1 are not plotted here because of no code or cache file released.

baseline	FU	IA	UA	DA	RF	MMA(%)	MS(%)	REP(%)
✓						62.26	25.57	63.11
✓	✓					74.00	28.37	67.38
✓	✓	✓				76.66	39.37	71.16
✓	✓	✓	✓			77.71	43.13	73.89
✓	✓	✓	✓	✓		77.77	44.32	75.86
✓	✓	✓	✓	✓	✓	78.33	44.76	78.28

Table 2: **Ablation study on different attention modules.** The FU and RF represents the feature uniformization and residual fusion, respectively. The IA, UA and DA is the in-scale, up-scale and down-scale attention modules. The results are all with the 3px error threshold.

4.1 Training Details

Datasets. The acquisition of sufficient ground-truth supervision to train keypoints detector and descriptor has been a bottleneck over years due to the ill-defined interest points. So we treat the keypoints detection and description as a self-supervised task, to make the detector discover better and easier keypoints by defining local maxima in the detection score maps as the target. A similar pipeline as [Revaud *et al.*, 2019] is developed to obtain dense ground-truth matching data.

The image pairs are composed by two aspects: 1) using the existing image pairs extracted from the Aachen Day-Night dataset [Sattler *et al.*, 2018] about the same sceneries and 2) applying a manual transformation such as homography transform or random rotation on pascal voc and the web images [Radenović *et al.*, 2018] to obtain image pairs.

Parameters. A NVIDIA RTX 3090 card is used to train our model using Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ for 30 epochs on the datasets. The initial learning rate is set to $1e^{-4}$ and decayed to $5e^{-5}$ in 30 epoch with 8 batch size. The testing is conducted on the same machine. The r and λ in loss function are set to 0.5 and 1, respectively. The multi-level balance parameters w_l in Equ.(11) and Equ.(12) are all set to 1.

4.2 Image Matching

We first evaluate our SPTD2 on the image matching task.

Datasets. To compare with other methods fairly, our method is evaluated on the HPatches dataset [Balntas *et al.*, 2017] including 116 different sequences of 6 images with accurate homography. To compare with other methods fairly, 8 high-resolution sequences are also excluded, leaving 52 and 56 sequences with illumination or viewpoint variations respectively.

Evaluation metrics. For fair comparison, we utilize three metrics, mean matching accuracy (MMA), keypoint repeatability (Rep) and matching score (MS) as evaluation metrics following [DeTone *et al.*, 2018]. The correct match is required to be a mutual nearest neighbor during brute-force searching. To evaluate the metrics fairly and accurately, the public code from [Dusmanu *et al.*, 2019] and [Luo *et al.*, 2020] is used to compute the corresponding metric.

Comparisons with other methods. We compare the mean matching accuracy with the state-of-the-art methods, namely DELF [Noh *et al.*, 2017], SuperPoint [DeTone *et al.*, 2018], multi-scale D2-Net [Dusmanu *et al.*, 2019], R2D2 [Revaud *et al.*, 2019], ASLFeat [Luo *et al.*, 2020], LISRD descriptors with SIFT detector [Pautrat *et al.*, 2020], DISK [Tyszkiewicz *et al.*, 2020], Key.Net [Barroso-Laguna *et al.*, 2019], D2-Net and Refinement [Dusmanu *et al.*, 2020], HardNet++ descriptors with HesAFFNet regions and [Mishkin *et al.*, 2018] (HAN + HN++), etc. Unless otherwise specified, we report either results reported in original papers, or derived from authors' public implementations with default parameters. We limit the maximum number of features of our method to 5K.

As shown in the Table 1 and Figure 3, SPTD2 achieves overall the best results regarding both illumination and viewpoint variations at different thresholds. Specifically, SPTD2 delivers remarkable improvements upon other methods especially for low range error thresholds, which in particular demonstrates that the keypoints localization error has been largely reduced. Besides, our method notably outperforms the more recent ASLFeat (78.33 vs 72.64 for MMA@3 overall), which also applied multi-level detection.

Method	Resolution	Memory(MB)	GFLOPs
SA		2168	619
DANet		2339	1110
RCCA	*2048×128×128*	427	804
ISSA		252	386
Ours		364	359

Table 3: **Efficiency comparison** given input feature map of size $2048 \times 128 \times 128$ in inference stage.

Ablations on Attention module. We further conduct diagnostic analysis to verify the effectiveness of the essential modules in our approach. We use the VGG-structure as the default backbone for all the studies. The performance of our baseline model with default parameters is given in the first row of the Table 2. The effect of each essential component of our SPTD2 on image matching task is shown as follows.

Ablations on soft point-wise transformer. As shown in Table 3, applying the soft point-wise attention module reduces the computation complexity and GPU memory compared to original attention module [Wang *et al.*, 2018], RCCA [Huang *et al.*, 2019b], DANet [Fu *et al.*, 2019], ISSA [Huang *et al.*, 2019a]. We further verify the impact of the detection threshold in the soft point-wise transformer. The detection threshold determines the number of keypoints which will be kept in the Keys set. The computation complexity will be higher with lower threshold. While derisory keypoints will influence the performance of the attention module to capture enough information. It's interesting that the soft point-wise attention module will degrade into the original attention module when the threshold is set to 0. We then choose the keypoints with the top-2K scores when the threshold is lower than 0.8 because of "CUDA out of memory". We set the threshold to $\{0.6, 0.8, 0.9, 0.95\}$ and get corresponding MMA@3 at $\{78.33, 78.31, 74.56, 72.37\}$. So we set the threshold at 0.8 to reach the best balance of performance and computation load.

4.3 Visual Localization

To further verify the effectiveness of the novel SPTD2, we evaluate it on the task of visual localization, which aims to estimate the camera pose within a given scene using images sequence. The task was proposed in [Sattler *et al.*, 2018] to evaluate the performance of local features in the context of localization. To evaluate our method fairly, we also produce the public format of keypoints and compare with other methods on the official evaluation server.

Datasets. We resort the Aachen Day-Night dataset [Sattler *et al.*, 2018] to demonstrate the effect on visual localization tasks, which contains images from the old inner city of Aachen, Germany. The key challenge in the dataset lies on matching images with extreme day-night changes.

Evaluation metrics. The evaluation is done using *The Visual Localization Benchmark*, which takes a pre-defined visual localization pipeline based on COMLAP [Schonberger and Frahm, 2016]. The successfully localized images are counted within three error tolerances (0.25m, 2°) / (0.5m, 5°) / (5m, 10°), representing the maximum position error in meters and degrees, respectively.

	Method	#Features	Dim	0.25m, 2°	0.5m, 5°	5m, 10°
Day	D2-Net	19.3K	512	83.7	91.6	96.5
	R2D2	10K	128	86.9	94.3	97.2
	ASLFeat	10K	128	85.2	93.2	96.1
	Ours	10K	128	**87.1**	**95.4**	**98.8**
Night V1.0	D2-Net	19.3K	512	80.6	87.8	96.9
	R2D2	10K	128	79.6	87.8	95.9
	ASLFeat	10K	128	**82.7**	87.8	95.9
	Ours	10K	128	78.8	**89.3**	**99.0**
Night V1.1	D2-Net	19.3K	512	68.1	85.9	97.9
	R2D2	10K	128	69.6	84.3	97.9
	ASLFeat	10K	128	**72.8**	85.3	96.9
	Ours	10K	128	72.5	**87.3**	**97.9**

Table 4: **Performance on Aachen Day-Night dataset for visual localization.** The benchmark website updates the evaluation metrics this year. The results are derived from authors' public implementations with default parameters.

Results. Our SPTD2 is compared with the typical joint detector and descriptor methods D2-Net, R2D2 and ASLFeat. Note that there exist some greater scores in the benchmark website, while they use greater matching strategy, which is unfair to evaluate. Here all methods are evaluated with the default matching strategy to compare fairly. As shown in Table 4, our SPTD2 performs surprisingly well under challenging illumination changes especially for strict accuracy metrics for the estimated pose. While in the night environment setting, the cross-scale attention modules bring some noises from the non-discriminative dark background, which hinders our performance. On the other hand, methods in Table 4, build image pyramid (MS) in inference to improve the localization performance, while making low running speed. We employ the multi-scale detection and description with the multi-level detector and descriptor in decoder, which is over 2 times quicker than MS operation. With $2^{1/4}$ scaling-factor MS, we improve the localization accuracy with $\{$+1.7%,+2.3%,+1.8%$\}$ for (0.25m, 2°).

5 Conclusions

In this paper, we propose a novel transformer-based architecture to jointly learn the local features descriptor and detector. The novel soft point-wise transformer simultaneously mines the long-range intrinsic and cross-scale dependencies of local features. The cross-scale attention module and multi-level decoder can guarantee the keypoints localization accuracy and discriminative descriptions especially in repetitive regions. Compared to other attention optimization methods, the soft point-wise attention remarkably decreases the computation and memory complexity. Experiments show SPTD2 significantly outperforms prior state-of-the-art methods.

Acknowledgments

This research is supported by Research Project of Intelligent sense technology of swarming drone under Grant 301021304.

References

[Balntas *et al.*, 2017] 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.

[Barroso-Laguna *et al.*, 2019] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *ICCV*, 2019.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.

[DeTone *et al.*, 2018] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR*, 2018.

[Dosovitskiy *et al.*, 2020] 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. *arXiv preprint arXiv:2010.11929*, 2020.

[Dusmanu *et al.*, 2019] 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.

[Dusmanu *et al.*, 2020] Mihai Dusmanu, Johannes L Schönberger, and Marc Pollefeys. Multi-view optimization of local feature geometry. In *ECCV*, 2020.

[Fu *et al.*, 2019] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[He *et al.*, 2018] Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *CVPR*, 2018.

[Huang *et al.*, 2019a] Lang Huang, Yuhui Yuan, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Interlaced sparse self-attention for semantic segmentation. In *ICCV*, 2019.

[Huang *et al.*, 2019b] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

[Li *et al.*, 2020] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. *AAAI*, 2020.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[Luo *et al.*, 2020] 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.

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

[Noh *et al.*, 2017] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *ICCV*, pages 3456–3465, 2017.

[Pautrat *et al.*, 2020] Rémi Pautrat, Viktor Larsson, Martin R. Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *ECCV*, 2020.

[Radenović *et al.*, 2018] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, pages 5706–5715, 2018.

[Revaud *et al.*, 2019] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2d2: Repeatable and reliable detector and descriptor. In *NeurIPS*, 2019.

[Sattler *et al.*, 2018] 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*, 2018.

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

[Sun *et al.*, 2020] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, 2020.

[Tian *et al.*, 2017] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *CVPR*, 2017.

[Tyszkiewicz *et al.*, 2020] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *NeurIPS*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[Wang *et al.*, 2018] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[Zhang *et al.*, 2020] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *ECCV*, 2020.

[Zhang, 2000] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE TPAMI*, 2000.

[Zheng *et al.*, 2020] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020.