Improving Transformer-based Image Matching by Cascaded Capturing Spatially Informative Keypoints

Chenjie Cao, Yanwei Fu†
School of Data Science, Fudan University
{20110980001,yanweifu}@fudan.edu.cn

Abstract

Learning robust local image feature matching is a fundamental low-level vision task, which has been widely explored in the past few years. Recently, detector-free local feature matchers based on transformers have shown promising results, which largely outperform pure Convolutional Neural Network (CNN) based ones. But correlations produced by transformer-based methods are spatially limited to the center of source views' coarse patches, because of the costly attention learning. In this work, we rethink this issue and find that such matching formulation degrades pose estimation, especially for low-resolution images. So we propose a transformer-based cascade matching model – Cascade feature Matching TRansformer (CasMTR)§, to efficiently learn dense feature correlations, which allows us to choose more reliable matching pairs for the relative pose estimation. Instead of re-training a new detector, we use a simple yet effective Non-Maximum Suppression (NMS) post-process to filter keypoints through the confidence map, and largely improve the matching precision. CasMTR achieves state-of-the-art performance in indoor and outdoor pose estimation as well as visual localization. Moreover, thorough ablations show the efficacy of the proposed components and techniques.

1. Introduction

Image matching is an important vision problem that is widely employed for many downstream tasks like Structure-from-Motion [39], Simultaneous Localization and Mapping [31], and visual localization [28]. However, accurately matching two or more images remains difficult due to various factors, such as differences in viewpoints, illuminations, seasons, and surroundings. Classical approaches [26, 36] address it via the pipeline of *detection, description, and matching of features* by hand-crafted features. Recently, learning Convolutional Neural Network (CNN) based detectors [9, 33, 38, 49, 55] have been utilized to detect and describe keypoints, leading to significant improvements in this pipeline. But such detector-based CNNs suffer from limited receptive fields and search space, as noticed in [43].

Table 1: Summary of test image size, backbones, memory cost (GB), and inference speed (s/image) on MegaDepth [22] with AUC of different pose errors (%). Suffixes '-8c', '-4c', and '-2c' denote matching at 1/8, 1/4, and 1/2 of image size. Baseline: QuadTree [47] with the same backbone as ours. Directly implementing QuadTree-4c causes Out-of-memory (OOM) error in a 32GB GPU, so its inference speed is estimated in brackets.

Methods	Backbone	Size	Pose Est. AUC @5°	@10°	@20°	Mem.(G)	s/img
Baseline-8c	Twins+FPN	704	51.63	68.54	80.98	3.83	0.146
CasMTR-4c	Twins+FPN	704	52.59	69.78	82.31	3.99	0.212
CasMTR-2c	Twins+FPN	704	**54.91**	**71.27**	**83.01**	6.29	0.311
QuadTree-8c	FPN	832	52.87	69.24	81.32	5.72	0.203
Baseline-8c	Twins+FPN	832	52.90	69.78	82.05	4.91	0.207
QuadTree-4c	FPN	832	–	–	–	OOM	(0.602)
CasMTR-4c	Twins+FPN	832	53.63	70.34	82.55	4.91	0.304
CasMTR-2c	Twins+FPN	832	**55.61**	**71.96**	**83.52**	7.55	0.444
QuadTree-8c	FPN	1152	55.09	71.31	83.20	12.62	0.424
Baseline-8c	Twins+FPN	1152	55.77	72.01	83.64	13.33	0.423
QuadTree-4c	FPN	1152	–	–	–	OOM	(1.442)
CasMTR-4c	Twins+FPN	1152	56.34	72.11	83.55	12.40	0.649
CasMTR-2c	Twins+FPN	1152	**56.90**	**72.94**	**84.24**	14.36	0.887

To solve this issue, transformer-based detector-free methods have emerged as more robust alternatives, demonstrating impressive matching abilities in texture-less regions [43, 18, 47, 57, 4]. However, the high computational cost of attention limits transformer-based methods to '*semi-dense*' matching, where source matching points are spaced apart at intervals of coarse feature space, as shown in Fig.1(a,d). Such semi-dense matching leads to an issue that *keypoint locations are not informative enough*: the spatially restricted source points in coarse feature maps lack the necessary details to express structural information, making it difficult to accurately estimate pose. This problem is especially challenging for low-resolution images, as seen in our pilot study (Tab. 1). More experiments based on extreme resolutions are discussed in the supplementary. Furthermore, it remains unclear whether transformer-based methods

†Corresponding author.
§Code is available at https://github.com/ewrfcas/CasMTR

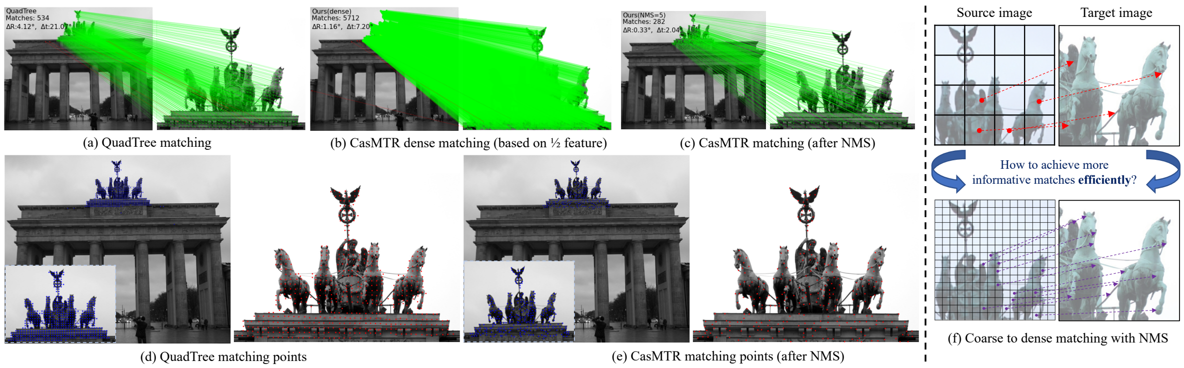


Figure 1: QuadTree [47] (a,d) vs our CasMTR (b,c,e). Our method achieves more fine-grained matching pairs for both source and target images (b). It is further improved by our NMS detection, which retains reliable matching results located in structural keypoints (c,e). We show an intuitive motivation for our spatially informative keypoints in (f). Best viewed in color.

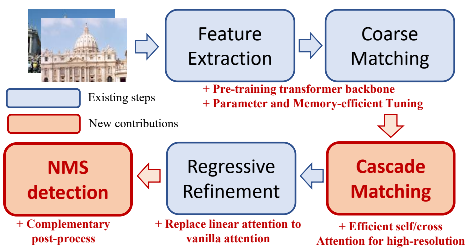


Figure 2: Illustration of CasMTR pipeline; and our novelties compared against the existing steps from detector-free matching methods [43, 47, 4] are highlighted in red.

can capture matching points in finer-grained image features rather than coarse ones (1/8) without a substantial increase in computational costs.

To address these challenges, we improve the existing transformer-based matching pipeline [43, 4] by efficiently capturing spatially informative keypoints in a cascaded manner. Particularly, our key idea is inspired by the coarse-to-fine Multi-View Stereo (MVS) [13]. We propose enhancing the transformer-based matching pipeline by adding the new stages of cascade matching and Non-Maximum Suppression (NMS) detection as summarized in Fig. 2. Such new stages increase and refine the matching candidates in source views. Thus, we can achieve *dense matching for both source and target views* as in Fig. 1(f), resulting in more precise matches focusing on more reliable positions with informative structures. Moreover, we elaborate on several novel techniques to support the newly incorporated stages in Fig. 2. Consequently, the proposed method can achieve dense and precise matches on 1/2 image size located in informative space.

Formally, we propose a transformer-based matching method called Cascade feature Matching TRansformer (CasMTR). It makes a significant contribution by enabling pure transformer-based models to conduct *dense* matching by cascaded capturing spatially informative keypoints without relying on the expensive learning of huge 4D correlations as merely extended from [43]. CasMTR develops several key components as follows. Firstly, inherited in MVS, coarse-to-fine cascade matching modules are repurposed with different efficient attention mechanisms [19, 58, 6, 47, 14, 67] to overcome the semi-dense matching in coarse features. We present the local non-overlapping [6] and overlapping [67] self-attention for outdoor and indoor cases respectively, due to different illuminations and surroundings. Secondly, CasMTR enjoys flexible training by a novel Parameter and Memory-efficient Tuning method (PMT), which is originally derived for NLP tasks [44]. Essentially, PMT can incrementally finetune CasMTR based on off-the-shelf matching models with reliable coarse matching initialization and fast convergence. Thirdly, we for the first time introduce the training-free NMS detection to complementarily filter precise matches based on dense matching confidence maps of CasMTR. Critically, NMS serves as a simple yet effective post-processing that preserves structurally meaningful keypoints rather than the coarse patch center as in Fig. 1(e). This improves the pose estimation as in Fig. 1(c) and has good generalization for various high-resolution matching tasks [22, 1, 45, 65]. Finally, in the development of our model, we have learned that the devil is in the details. Consequently, several non-trivial technical improvements have been implemented to our newly proposed matching pipeline (highlighted in Fig. 2), such as pre-training transformer backbones, improving efficient linear attention, and optimizing self and cross attention for high-resolution matching.

The proposed CasMTR is comprehensively evaluated in relative pose estimation [22, 7], homography estimation [1], and visual localization [45, 65], showing its state-of-the-art performance. Additionally, our exhaustive ablation studies show the effectiveness of all newly proposed components.

2. Related Work

Detector-based Matching. Detector-based matching methods following the process of feature detection, description, and matching have dominated this field for a long time. Traditional manners utilize heuristic hand-craft features [26, 36] for local feature matching, which enjoy great success and are still used in many 3D tasks nowadays. After the wave of deep learning, many learning-based methods [15, 62, 8, 11, 25, 27] were proposed based on the detector-dependent pipeline with better performance. SuperPoint [9] proposes to utilize the homographic adaptation for the self-supervised matching training. Then, SuperGlue [38] further improves the performance through the graph neural network. Moreover, DISK [55] leverages reinforcement learning to optimize the end-to-end detector-based pipeline. However, these methods still suffer from limited interest points in indistinctive regions [43]. On the other hand, D2D [49] proposes to describe first, and then detect based on deep descriptors [30, 50]. Compared with confidence-based NMS detection, feature-based D2D is complicated and slower. Besides, D2D is not compatible with the joint training model because it ignores the correlation between source and target views (Tab. 9).

Detector-free Matching. Detector-free methods enjoy an end-to-end pipeline to achieve the matching directly without an explicit keypoint detection phase [24, 5, 43]. Learning-based detector-free methods can be generally categorized into Convolutional Neural Network (CNN) based methods [35, 34, 20, 52, 54, 12] and transformer or attention-based ones [43, 18, 47, 57, 41, 4, 46]. CNN-based methods produce dense matching results through learning 4D cost volumes or warped features, which are limited by receptive fields. Some transformer-based manners [43, 47, 57, 4], led by LoFTR [43], largely enlarge the receptive fields with interlacing self/cross attention modules, and enjoy better performance in texture-less regions. On the other hand, COTR [18] jointly learns both matching images with self-attention together rather than modeling self/cross ones respectively in encoders. Then query points are decoded through cross-attention for the matching results. We focus on the former one in this paper. But matching density and accuracy of these approaches are insufficient to tackle many downstream tasks precisely, *e.g.*, pose estimation for low-resolution images. Moreover, it is non-trivial to extend these transformer-based matching solutions into dense and high-resolution cases because of the heavy computation of attention.

Coarse-to-fine Learning. The efficient coarse-to-fine manner plays an important role in learning-based stereo matching [51, 59, 63, 13], MVS [13, 64, 56, 29], and optical flow [32, 42, 60, 66]. CasMVSNet [13] builds coarse cost volume at early stages with large depth ranges and makes later stages refine details. On the other hand, PWC-Net [42] warps pyramid features into cost volumes to further refine the coarse-to-fine flow estimation. Different from the depth prediction and the optical flow, learning geometric image matching with coarse-to-fine manners is more solid to tackle the error propagation [48]. Because the geometric image matching is usually based on static landmarks with consistent displacements. Thus the coarse matching in low-resolution will not inevitably mislead local details. Patch2pix [68] proposed a coarse-to-fine refinement for pixel-level matching just for CNNs. COTR [18] needs to recursively crop finer patches for more precise matching results, which is very time-consuming. ECO-TR [46] proposes to crop coarse-to-fine feature patches and train them end-to-end to improve efficiency. But the feature cropping of ECO-TR still limits the receptive fields across different patches. Hence learning a coarse-to-fine transformer-based matching model with global receptive fields is still challenging.

3. Method

Preliminary and Overview. We briefly review the transformer-based matching baseline in the example of LoFTR [43]. LoFTR uses a local feature CNN to extract coarse (1/8) and fine (1/2) feature maps from image pairs. Then interlaced self/cross-attention modules are leveraged to learn coarse-level matching predictions. Additionally, LoFTR utilizes a refinement module to model sub-pixel match prediction in fine-level features. However, the source point of each matched pair is still restricted at the coarse level (1/8), which limits the performance. Some follow-ups [47, 4] improve the linear attention [19] of LoFTR while retaining the whole pipeline unchanged. Inherited from the LoFTR, we develop a novel coarse-to-fine CasMTR as in Fig. 3. Given the matching image pair $\mathbf{I}_A, \mathbf{I}_B$, we first extract their multi-scale features by a feature encoder. Then the self and cross QuadTree attention [47] based coarse matching is performed in coarse-level features (Sec. 3.1). According to the coarse matches, a couple of local attention-based cascade matching modules are proposed to refine the matching pairs (Sec. 3.2). After that, a sub-pixel refinement leverages the spatial expectation to predict exact matching results (Sec. 3.3). Finally, the NMS post-process detects local keypoints based on confidence maps, which largely improves the pose estimation in outdoor scenes (Sec. 3.4).

3.1. Feature Extraction and Coarse Matching

Feature Encoder. We first follow [43] and use FPN to produce coarse-to-fine features $\mathbf{F}^s_A, \mathbf{F}^s_B$ for the image pair $\mathbf{I}_A, \mathbf{I}_B$, where $s \in \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$ indicate the image scale. Inspired by [17], we try to replace $\{\frac{1}{4}, \frac{1}{8}\}$ layers with partial pre-trained Twins [6] layers. To balance the computation, the feature encoder's channels are reduced in our CasMTR, which also benefits the efficiency of subsequent cascade modules. The pre-trained attention-based encoder strengthens the matching learning as evaluated in Tab. 1.

Parameter and Memory-efficient Tuning (PMT). Since we pay more attention to the coarse-to-fine matching, our coarse matching is simply based on the state-of-the-art QuadTree attention [47]. Essentially, the proposed model

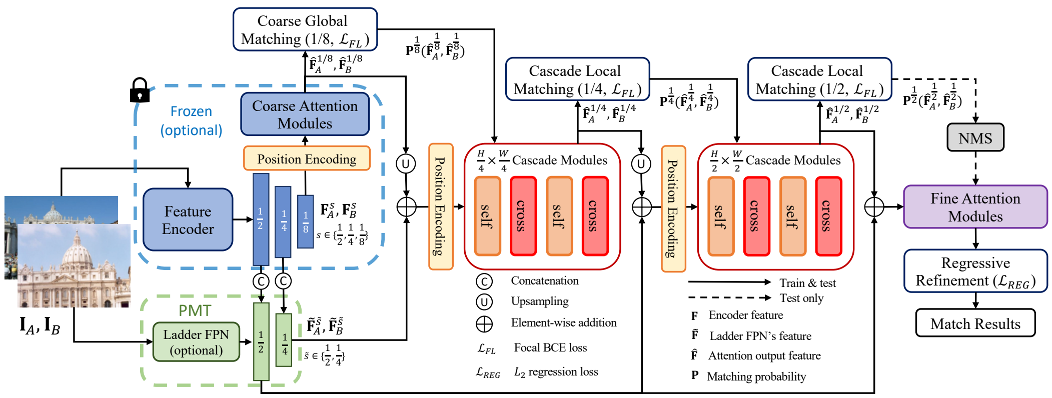


Figure 3: Overview of CasMTR. Optionally, our model can work as an incremental refinement. Particularly, we could freeze feature encoder and coarse attention modules during training with a lightweight trainable ladder FPN to save the computation and memory footprint. Matching scales and loss functions are denoted in the bracket of each matching module, while feature scales are shown in superscripts. Softmax matching probabilities $\mathbf{P}(\hat{\mathbf{F}}_A, \hat{\mathbf{F}}_B)$ got from global (1/8) and local (1/4, 1/2 detailed in Eq. 1) dot products are utilized to decide the next local matching candidates and NMS (test only).

is learned independently from the coarse matching, *i.e.*, we can freeze the feature encoder and coarse matching module, and incrementally finetune the cascade matching with the coarse matching initialization. To improve the representation of high-level features, we introduce PMT to incrementally finetune the matching model as shown in Fig. 3. Specifically, a lightweight trainable ladder side FPN is utilized to receive and concatenate features from the frozen feature encoder as $\tilde{\mathbf{F}}_A^{\tilde{s}}, \tilde{\mathbf{F}}_B^{\tilde{s}}$, where $\tilde{s} \in \{\frac{1}{2}, \frac{1}{4}\}$. Different from other tuning techniques [16, 21, 10], PMT is not only parameter-efficient but also memory-efficient. Because the FPN of PMT could be well-updated by fine-grained features without any gradients propagated back from frozen models. In practice, we leverage the PMT to finetune our cascade modules based on the off-the-shelf QuadTree matching on the large ScanNet dataset [7]. Our algorithm can be converged in about two epochs and achieve appreciable improvements as in Tab. 4.

3.2. Cascade Matching Modules

Following the coarse matching results, we additionally propose multi-stage cascade modules to further refine more detailed matching results for both source and target images. For each stage, we first add sinusoidal position encoding as other methods [43, 47, 4]. We normalize the position encoding as [4] during the inference, which makes CasMTR robust to various test sizes. Then, self and cross-attention layers are interleaved in the cascade module for better local feature learning. Different from 1D-cascade architectures [13], extending the cascade mechanism to 2D is non-trivial. The main concern about cascade matching learning is the computation for high-resolution features. To address this, we thoroughly compare various efficient self and cross-attention mechanisms and choose the best combination among them.

Self-Attention. Global self-attention suffers from quadratic spatial complexity, especially for high-resolution features.

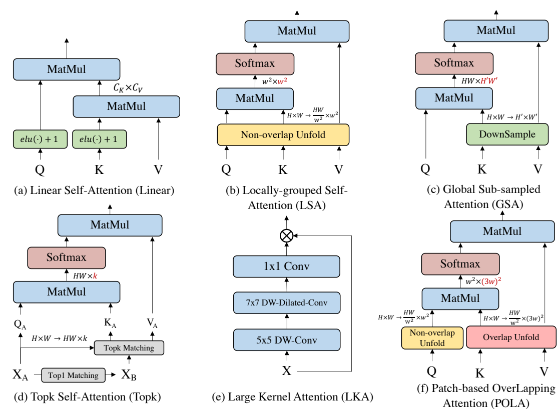


Figure 4: Six self-attention modules explored in CasMTR.

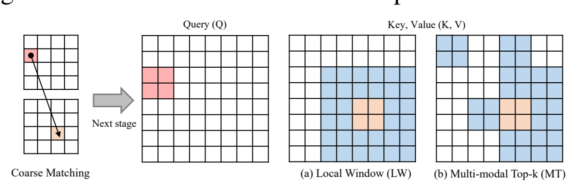


Figure 5: Two cross-attentions explored in cascade modules.

But without the self-attention, the pure cross-attention model performs not well as in Tab. 2. To balance the computation and the performance, we explore six efficient self-attention mechanisms shown in Fig. 4 and verified in Tab. 2, including Linear attention [19], Locally-grouped Self-Attention (LSA) [6], Global Sub-sampled Attention (GSA) [58], simplified top-k attention, Large Kernel Attention (LKA) [14], and Patch-based OverLapping Attention (POLA) [67]. We list more details of these manners in the supplementary.

Cross-Attention. The cross-attention plays an important role in cascade matching. Two types of attention modules, designed in Fig. 5 are Local Window (LW) and Multi-modal Top-k (MT) cross-attention respectively. Given the coarse matching result, each query patch in LW intuitively selects

Table 2: Pilot study of AUC and FLOPs about different attention mechanisms based on 1/4 cascade model (Ours-4c) on MegaDepth. All FLOPs of cascade modules are based on 1152×1152 test images. LSA+LW is adopted on MegaDepth, while POLA+LW is adopted on ScanNet.

Cascade (layers)		MegaDepth			ScanNet			FLOPs(G)
self	cross	@5°	@10°	@20°	@5°	@10°	@20°	
Linear(2)	LW(2)	56.01	72.03	83.43	–	–	–	129.21
LSA(2)	LW(2)	56.34	72.11	83.55	26.24	46.45	63.94	142.32
LSA+GSA(2)	LW(2)	55.71	71.60	83.17	–	–	–	202.56
LSA(2)	MT(2)	55.60	71.92	83.27	–	–	–	143.51
LKA(2)	LW(2)	55.75	72.02	83.20	–	–	–	136.53
–	LW(4)	55.16	71.48	83.01	–	–	–	143.28
Top-k(2)	LW(2)	55.47	71.28	83.02	–	–	–	141.76
LKA(2)	MT(2)	**56.99**	**72.56**	**83.89**	25.79	45.87	63.50	137.72
POLA(2)	LW(2)	56.31	72.35	83.51	**27.08**	**47.02**	**64.44**	223.42

neighbor patches around the top-1 matching target from another image as key and value patches. In Fig. 5(a), the LW example is based on window size 6 and 36 neighbors in all. LW dramatically reduces the sequence length of keys and values, which makes the cascade matching for high-resolution features possible. Furthermore, LW is more capable of learning detailed feature correlations. On the other hand, to alleviate the intractable error propagation caused by the coarse stage [61], we propose MT to model multi-modal distribution for cascade matching. In particular, MT holds top-k coarse matching patches as candidates. Then they are upsampled as key and value patches for the cross-attention. As in Fig. 5(b), the MT example is on top-36: top-9 in the coarse stage, and each coarse block can be further divided into 4 patches in the cascade stage. MT can potentially address the mismatch from the coarse stage, as long as the top-k candidates can cover the ground truth. Because of the scale upsampling, both source and target features are quadrupled as shown in Fig. 5, which influence the practical kernel size and top-k in LW and MT respectively. Technically, we implement both LW and MT by CUDA to improve efficiency, and their speed is almost the same in practice.

Analysis of Attention Modules. We conduct pilot study to verify the results and FLOPs of all attention combinations in Tab. 2. Such a pilot thus guides how we design the model. Specifically, 'LSA+LW' enjoys a good trade-off between performance and efficiency in the outdoor MegaDepth. Furthermore, the extended 'LSA+GSA' fails to achieve better results with more computation. We think that local feature learning is more important than global one in our cascade modules. Besides, for the indoor scenes from ScanNet with more challenging texture-less matching instances, 'POLA+LW' outperforms 'LSA+LW' with larger receptive fields. Therefore, 'LSA+LW' and 'POLA+LW' are used to comprise our cascade modules. Note that 'LKA+MT' can produce superior results in MegaDepth. But we did not choose 'LKA+MT' for two reasons. First, the depthwise convolutions used in LKA are not well optimized in PyTorch, which largely slows down the training. Second, 'LKA+MT' is not stable enough, as it fails to achieve reliable results in ScanNet.

Matching and Loss. Given $\hat{\mathbf{F}}^{\tilde{s}}_A, \hat{\mathbf{F}}^{\tilde{s}}_B, \tilde{s} \in \{\frac{1}{2}, \frac{1}{4}\}$ after the interlaced attention learning of cascade modules, we use the same key candidates from the cross-attention (*i.e.*, LW or MT) for the dot product similarity matrix as

$$\mathbf{S}(i,j) = \frac{1}{\tau} \cdot \left\langle \hat{\mathbf{F}}^{\tilde{s}}_A(i), \hat{\mathbf{F}}^{\tilde{s}}_B(j) \right\rangle \in \mathbb{R}^{H^{\tilde{s}}W^{\tilde{s}} \times k} \quad (1)$$

where $\tau = 0.1$ is a scale parameter; the key length k is 100 and 128 for LW and MT respectively. Note that Eq. 1 presents a local correlation with k candidates for each feature point rather than the full correlation in the coarse level. Softmax is used to normalize Eq. 1 into local matching probability $\mathbf{P}^{\tilde{s}}(i,j)$. We also adopt cycle-consistent matching to enforce that two features in different images are matched each other. Following [43], the Focal binary cross-entropy Loss (FL) [23] is used to optimize the cascade matching as

$$\mathcal{L}^{\tilde{s}}_{FL} = -\mathbb{E}_{\mathcal{M}^{\tilde{s}}}[(1-\mathbf{P}^{\tilde{s}})^{\gamma} \log(\mathbf{P}^{\tilde{s}})], \quad (2)$$

where $\gamma = 2$; $\mathcal{M}^{\tilde{s}}$ indicates matching queries which satisfy the cycle-consistent and have one ground truth target in k candidates. In cascade stages, the classification loss enjoys the priority because we have to learn proper confidence [3] for the detection (Sec. 3.4). We also tried the vanilla cross-entropy, but it performed slightly worse than FL.

Discussions. Since cascade matching facilitates pose estimation in limited resolution (Tab. 1), our method achieves prominent improvement on 480×640 ScanNet [7] without any post-processing (Tab. 4). When input images become larger, the coarse matching pairs also become dense gradually to alleviate the pose estimation error. Moreover, for large image scales, we find that our cascade matching can be strengthened by a simple yet effective NMS post-processing with negligible cost. Besides, our CasMTR is efficient enough compared with the trivial extension (QuadTree-4c).

3.3. Local Regressive Refinement

The patch-wise refinement module in LoFTR [43] is also incorporated in our work for sub-pixel matching. The refinement module first unfolds all features into 5×5 patches. Different from the one in LoFTR, we use the standard attention instead of the linear one in both self and cross attention. Because the refinement module only calculates the attention map with a sequence length $5 \times 5 = 25$. Therefore, standard attention even enjoys less computation compared with linear attention and performs better. The refinement module utilizes soft-argmax to regress the residual matching flow. One may ask whether such local refinement can replace cascade modules for dense matching. We should clarify that the patch-wise refinement is extremely limited by the matching range and receptive fields, which discourages the results. We tried the trivial solution to densely match through the refinement module in Tab. 8, but it worked worse than the baseline. Even NMS failed to make its results competitive.

3.4. Confidence based Detection with NMS

Different from detector-based methods [11, 9, 38], latest attention based methods [43, 47, 4, 57] achieve good results even without detector. These detector-free methods only use a confidence threshold to filter unconvinced matching. Moreover, the sparse matching (1/8) is not ready for further keypoint detection. Except for the confidence threshold, we propose to use the simple NMS to detect local keypoints through the cascade confidence maps as shown in Fig. 1(c). Specifically, we apply the overlapping max-pooling on the confidence map. Then if the local maximal confidence locates in the center of the pooling kernel, we retain this matching pair. So the minimum interval in feature space of two keypoints is equivalent to half of NMS's kernel size. The main difference between the NMS and the threshold refusing is that NMS detects local keypoints through confidence rather than global filtering. Thanks to the dense matching from cascade modules, NMS can shift the matching prediction to some structural keypoints with relatively higher confidence. Thus NMS is complementary to CasMTR. We find that the simple NMS outperforms other traditional detectors [26], and feature-based detector [11, 49]. Further, we train CasMTR with trainable detectors, working worse than NMS as in Tab. 9, which is discussed in Sec. 4.4.

4. Experiments

Datasets. CasMTRs are trained on outdoor MegaDepth [22] and indoor ScanNet [7] to verify the relative pose estimation. MegaDepth comprises 196 scene reconstructions with 1M Internet images. Ground-truth matching pairs are from COLMAP [40] computed depth maps, Following [43], for one epoch, we randomly sample 200 pairs from each scene for the training, and 1500 pairs from independent two scenes are selected as the test set. For the ScanNet, there are 1613 monocular sequences with 230M and 1500 pairs for training and testing respectively. For one epoch, 100 images are sampled for training on each scene.
Implementation. We extend CasMTR into $\{\frac{1}{4},\frac{1}{2}\}$ resolutions with cascade modules, *i.e.*, CasMTR-4c and CasMTR-2c. The NMS kernel is fixed in 5 for the pose estimation. For MegaDepth, CasMTR is trained progressively in 704×704 and tested in 1152×1152. In particular, we first train CasMTR in the coarse stage with $\frac{1}{8}$ matching for 8 epochs. Then CasMTR-4c and CasMTR-2c are further finetuned with 16 and 8 epochs respectively. CasMTR-2c converges faster than CasMTR-4c because more supervised matching pairs are learned in the high-resolution learning for each epoch. For ScanNet, both training and testing image size is 480×640. To tackle the mega data scale, we use PMT to incrementally finetune CasMTR-4c based on the off-the-shelf QuadTree [47] weights. PMT-CasMTR-4c can converge in only 2 epochs. CasMTR shares a 0.2 threshold in all stages.

Table 3: Pose estimation on outdoor MegaDepth with AUC of different pose errors (%).

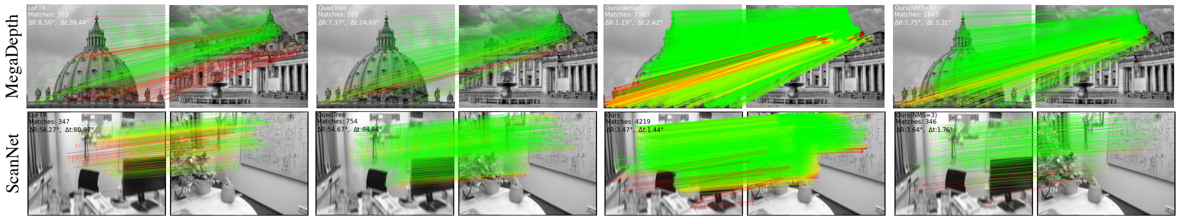
Methods	Pose Estimation AUC ↑		
	@5°	@10°	@20°
SP [9]+SuperGlue [38]	42.2	61.2	76.0
PDCNet+(H) [53]	43.1	61.9	76.1
LoFTR [43]	52.8	69.2	81.2
QuadTree [47]	54.6	70.5	82.2
MatchFormer [57]	53.3	69.7	81.8
DKM [12]	54.5	70.7	82.3
ASpanFormer [4]	55.3	71.5	83.1
CasMTR-4c	58.0	73.6	84.6
CasMTR-2c	**59.1**	**74.3**	**84.8**

4.1. Relative Pose Estimation

As in [38, 43], we evaluate the relative pose estimation with AUC of pose errors at thresholds $(5^\circ, 10^\circ, 20^\circ)$, while the pose error is defined as the maximum angular error of rotation and translation. The essential matrix is optimized by OpenCV RANSAC with model-predicted matching pairs.
Outdoor MegaDepth. We show MegaDepth results in Tab. 3. CasMTR can outperform all competitors especially in AUC5° and AUC10°, which include both transformer-based [43, 47, 57, 4] and CNN-based [12] manners. Moreover, our CasMTR-2c with denser feature matching capability can further improve the performance. Besides, our NMS detection is effective for outdoor scenes with large displacements and appearance transformations as verified in Tab. 9. Qualitative results are compared in Fig. 6. Our CasMTR achieves denser and more exact matching results.
Indoor ScanNet. ScanNet results are in Tab. 4. CasMTR-4c achieves the best result among all competitors. Note that our PMT-enhanced method only needs to be finetuned with 2 epochs, which is flexible and efficient for the practice. CasMTR-2c did not obtain more improvement compared with CasMTR-4c in ScanNet. We think that texture-less regions of indoor scenes with motion blur and inferior annotations are too challenging for local attention learning in the 1/2 resolution. Since the resolution of ScanNet is much lower than MegaDepth, NMS is not applied to CasMTR to remain dense enough matching pairs, which results in more precise pose estimation as qualitatively compared in Fig. 6.

4.2. Homography Estimation

CasMTR is also evaluated in on HPatches dataset [1] for the homography estimation. HPatches contains 116 planar scenes with viewpoint or illumination changes, which is widely used to evaluate the low-level matching performance. Following [38, 43], we report the AUC of corner error up to thresholds 3, 5, and 10 pixels in Tab. 5. RANSAC is adopted to get the robust homography matrix. To ensure fairness, we



(a) LoFTR (b) QuadTree (c) CasMTR (dense) (d) CasMTR (NMS)

Figure 6: Qualitative outdoor and indoor matching results compared with LoFTR [43], QuadTree [47], CasMTR-4c (ScanNet), CasMTR-2c (MegaDepth), and our NMS detected results.

Table 4: Pose estimation on indoor ScanNet [7] with AUC of different pose errors (%).

Methods	Pose Estimation AUC ↑		
	@5°	@10°	@20°
SP [9]+SuperGlue [38]	16.2	33.8	51.8
PDCNet+(H) [53]	20.2	39.4	57.1
LoFTR [43]	22.0	40.8	57.6
QuadTree [47]	24.9	44.7	61.8
MatchFormer [57]	24.3	43.9	61.4
DKM [12]	24.8	44.4	61.9
ASpanFormer [4]	25.6	46.0	63.3
CasMTR-4c	**27.1**	**47.0**	**64.4**

Table 5: Homography estimation on HPatches [1] with AUC of different corner errors (%).

Methods	Pose Estimation AUC ↑			matches
	@3px	@5px	@10px	
DISK [55]+NN	52.3	64.9	78.9	1.1k
SP [9]+SuperGlue [38]	53.9	68.3	81.7	0.6k
LoFTR [43]	64.6	74.8	84.2	2.6k
QuadTree [47]	66.3	76.2	84.9	2.7k
CasMTR-4c	67.5	77.1	86.3	11.4k
CasMTR-2c	69.6	78.9	87.1	44.7k
CasMTR-4c (NMS=5)	69.7	78.8	87.0	0.4k
CasMTR-2c (NMS=9)	**71.4**	**80.2**	**87.9**	0.5k

resize the short side of each image to 480 as LoFTR. From Tab. 5, CasMTRs trained on MegaDepth outperform other methods with denser matching results. But we should clarify that dense matching is not the key factor to improve the homography estimation. After the NMS detection, results from CasMTR are further improved with even fewer matches than LoFTR or QuadTree. Therefore, experiments on HPatches sufficiently show the effectiveness of the proposed method. More details are discussed in the supplementary.

4.3. Visual Localization

We also evaluate CasMTR on the InLoc [45] and Aachen Day-Night v1.1 [65] benchmarks of visual localization to further validate the robustness of our model. Following the pipeline of HLoc [37], we replace the matching stage with compared methods for getting matching pairs between query and database images. Since no official codes are provided from [43], we re-implement the visual localization and report results in Tab. 6 and Tab. 7. Our baseline

Table 6: Visual localization on InLoc [45]. * means our implementation of LoFTR; note that our re-implementations are better on DUC1 and worse on DUC2 compared with [43].

Methods	DUC1	DUC2
	(0.25m,2°)/(0.5m,5°)/(1m,10°)	
HLoc [37]+LoFTR [43]*	49.5/73.7/82.8	**51.9**/69.5/80.9
HLoc [37]+Baseline	47.5/71.7/83.8	48.1/**70.2**/79.4
HLoc [37]+CasMTR	**53.5/76.8/85.4**	**51.9/70.2/83.2**

Table 7: Visual localization on Aachen Day-Night [65].

Methods	Day	Night
	(0.25m,2°) / (0.5m,5°) / (1m,10°)	
HLoc [37]+LoFTR [43]	88.7/95.6/99.0	**78.5**/90.6/99.0
HLoc [37]+Aspanformer [4]	89.4/95.6/99.0	77.5/**91.6/99.5**
HLoc [37]+CasMTR	**90.4/96.2/99.3**	**78.5/91.6/99.5**

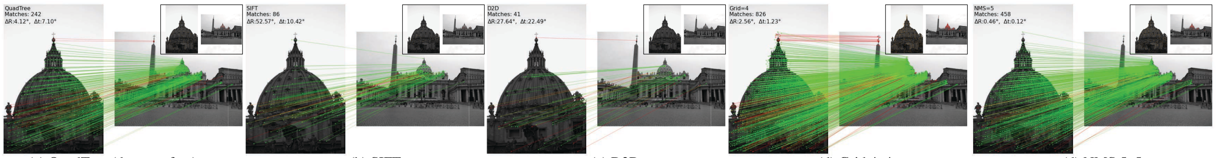
Table 8: Comparison between our CasMTR and the trivial refinement extension of baseline (Baseline-Tri) on MegaDepth.

Methods	Pose Estimation AUC ↑		
	@5°	@10°	@20°
Baseline	55.77	72.01	83.64
Baseline-Tri	47.09	64.76	78.13
Baseline-Tri (NMS=5)	51.19	67.62	80.00
Ours-4c (NMS=5)	57.99	72.42	84.58
Ours-2c (NMS=5)	**59.08**	**74.33**	**84.80**

is based on QuadTree with Twins backbone. Considering high-resolution inputs and large-scale images, we use the MegaDepth pre-trained CasMTR-4c enhanced with NMS kernel size 5 to evaluate both benchmarks. As shown in Tab. 6 and Tab. 7, CasMTR outperforms other competitors.

4.4. Ablations

Cascade Matching vs Dense Refinement. As mentioned in Sec. 3.3, the straightforward way to achieve dense matching with detector-free methods [43, 47] is to make all patch-wise features in the refinement module produce matching results as much as possible. Theoretically, a such trivial extension can get as many matching pairs as CasMTR-2c. But as verified in Tab. 8, such trivial extension (Baseline-Tri) fails to get good results. Because receptive fields of the patch-wise refinement are limited, while the self-attention only considers features in the same patch. Moreover, the



(a) QuadTree (detector-free) (b) SIFT (c) D2D (d) Grid 4x4 (d) NMS 5x5

Figure 7: Qualitative comparisons among different detection methods. Detected points are shown in the right-up corner..

Table 9: Ablation studies about various detection methods based on CasMTR-4c trained on MegaDepth. Thr.>0.5 means that increasing the confidence threshold from 0.2 to 0.5. Trainable* and † indicate that finetuning the baseline with extra trainable detectors. * only optimizes detected points while † learns detected points with higher weights (3 times). Grid 4×4 means that selecting top-1 from non-overlapping 4×4 confidence windows, while NMS is max-pooled on 5×5 overlapping ones.

Detector	Pose Estimation AUC ↑		
	@5°	@10°	@20°
–	56.34	72.11	83.55
Thr.> 0.5	53.81	70.46	82.54
Trainable*	56.06	72.05	83.26
Trainable†	57.21	73.05	84.03
SIFT	51.84	68.78	81.39
D2D	53.92	70.50	82.58
Grid 4×4	56.54	72.42	83.98
NMS 5×5	**57.99**	**73.36**	**84.58**

cross-attention is also corrupted by non-overlapping cross windows. Although the NMS can improve the Baseline-Tri a little, it still has a large gap compared with CasMTR.

Detection Methods. We evaluate different strategies in Tab. 9, while qualitative comparisons are shown in Fig. 7. Simply increasing the threshold is not effective for the pose estimation without considering the local relation. For trainable versions, we finetune CasMTR with another trainable detector jointly through the straight-through estimation [2] with grid size 4×4 and tested with NMS kernel 5 as [55]. But the trainable detector even reduces the performance. We think that such a jointly detect-and-describe pipeline is unsuitable for supervised image matching learning. Because the detector devotes to searching keypoints which are *easy* to be matched, while the descriptor becomes lazy to ignore hard matching cases. Thus, we train another version by simply increasing the weights of detected points instead. This slightly improves the performance, but still has a gap from NMS. Note that NMS is much more efficient because it is training-free. Moreover, both traditional SIFT [26] and feature-based D2D [49] post-processing fail to improve the matching performance. So these detectors are incompatible with the transformer-based matching, *e.g.*, they cannot ensure that the detected keypoints enjoy confident model probabilities. From Tab. 9, the overlapping maxpool filtering (NMS) outperforms the non-overlapping one (Grid).

Table 10: Ablation studies about NMS kernel size in post-processing (Post.) on MegaDepth.

Cascade		Post.	Pose Estimation AUC ↑		
4c	2c		@5°	@10°	@20°
		–	55.77	72.01	83.64
✓		–	56.34	72.11	83.55
✓	✓	–	56.90	72.94	84.24
		NMS 3×3	56.23	72.17	83.37
		NMS 5×5	55.41	71.10	82.67
		NMS 7×7	55.75	71.09	82.26
✓		NMS 3×3	56.95	73.02	84.36
✓		NMS 5×5	57.99	73.36	84.58
✓		NMS 7×7	56.99	72.78	84.11
✓	✓	NMS 3×3	57.56	73.40	84.60
✓	✓	NMS 5×5	**59.08**	**74.33**	**84.80**
✓	✓	NMS 7×7	57.64	73.44	84.38

Kernel Size of NMS. We evaluate the NMS detection with different kernel sizes in Tab. 10 on MegaDepth 1152×1152 image pairs. NMS can not improve the pose estimation of the baseline method without cascade dense matching. Because the semi-dense solution does not contain sufficient matching pairs to be detected as local keypoints. Besides, both CasMTR-4c and CasMTR-2c can achieve the best results with NMS kernel size 5. More experiments in HPatches (Tab. 5) and visual localization (Tab. 6, Tab. 7) denote that our NMS detection can work robustly with CasMTR.

5. Conclusion

We rethink the transformer-based image matching pipeline and find that locating spatially informative source points is critical. So we propose a transformer-based cascade matching model called CasMTR, which can produce denser matches compared with previous transformer-based methods through its coarse-to-fine cascade modules. Benefiting from the thorough investigation of efficient attention, CasMTR enjoys a good balance between performance and efficiency. Further, CasMTR can be finetuned based on off-the-shelf matching models through the PMT. The newly repurposed NMS further detects more precise matching pairs in informative keypoints, improving the pose estimation. CasMTR enjoys state-of-the-art results in relative pose estimation, homography estimation, and visual localization.

Acknowledge. Dr. Fu is also with Shanghai Key Lab of Intelligent Information Processing, Fudan University, and Fudan ISTBI—ZJNU Algorithm Centre for Brain-inspired Intelligence, Zhejiang Normal University, Jinhua, China.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017. 2, 6, 7

[2] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 8

[3] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Learning robust image representations via transformers and temperature-based depth for multi-view stereo. *arXiv preprint arXiv:2208.02541*, 2022. 5

[4] 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. *arXiv preprint arXiv:2208.14201*, 2022. 1, 2, 3, 4, 6, 7

[5] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *Advances in neural information processing systems*, 29, 2016. 3

[6] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34:9355–9366, 2021. 2, 3, 4

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 4, 5, 6, 7

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Toward geometric deep slam. *arXiv preprint arXiv:1707.07410*, 2017. 3

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 1, 3, 6, 7

[10] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11368, 2022. 4

[11] 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 *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 3, 6

[12] Johan Edstedt, Mårten Wadenbäck, and Michael Felsberg. Deep kernelized dense geometric matching. *arXiv preprint arXiv:2202.00667*, 2022. 3, 6, 7

[13] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 2, 3, 4

[14] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *arXiv preprint arXiv:2202.09741*, 2022. 2, 4

[15] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3279–3286, 2015. 3

[16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 4

[17] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. *arXiv preprint arXiv:2203.16194*, 2022. 3

[18] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 1, 3

[19] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020. 2, 3, 4

[20] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *Advances in Neural Information Processing Systems*, 33:17346–17357, 2020. 3

[21] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 4

[22] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1, 2, 6

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[24] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010. 3

[25] Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. Gift: Learning transformation-invariant dense visual descriptors via group cnns. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[26] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1, 3, 6, 8

[27] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6589–6598, 2020. 3

[28] Simon Lynen, Bernhard Zeisl, Dror Aiger, Michael Bosse, Joel Hesch, Marc Pollefeys, Roland Siegwart, and Torsten Sattler. Large-scale, real-time visual–inertial localization revisited. *The International Journal of Robotics Research*, 39(9):1061–1084, 2020. 1

[29] Zhenxing Mi, Chang Di, and Dan Xu. Generalized binary search network for highly-efficient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12991–13000, 2022. 3

[30] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. *Advances in neural information processing systems*, 30, 2017. 3

[31] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 1

[32] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 3

[33] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2d2: repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*, 2019. 1

[34] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *European conference on computer vision*, pages 605–621. Springer, 2020. 3

[35] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31, 2018. 3

[36] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 1, 3

[37] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 7

[38] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 3, 6, 7

[39] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[40] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 6

[41] Yan Shi, Jun-Xiong Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12517–12526, 2022. 3

[42] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 3

[43] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 1, 2, 3, 4, 5, 6, 7

[44] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *arXiv preprint arXiv:2206.06522*, 2022. 2

[45] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 2, 7

[46] Dongli Tan, Jiang-Jiang Liu, Xingyu Chen, Chao Chen, Ruixin Zhang, Yunhang Shen, Shouhong Ding, and Rongrong Ji. Eco-tr: Efficient correspondences finding via coarse-to-fine refinement. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 317–334. Springer, 2022. 3

[47] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. *arXiv preprint arXiv:2201.02767*, 2022. 1, 2, 3, 4, 6, 7

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 3

[49] Yurun Tian, Vassileios Balntas, Tony Ng, Axel Barroso-Laguna, Yiannis Demiris, and Krystian Mikolajczyk. D2d: Keypoint extraction with describe to detect approach. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 3, 6, 8

[50] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11016–11025, 2019. 3

[51] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–204, 2019. 3

[52] Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 3

[53] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *arXiv preprint arXiv:2109.13912*, 2021. 6, 7

[54] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2021. 3

[55] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 1, 3, 7, 8

[56] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021. 3

[57] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. *arXiv preprint arXiv:2203.09645*, 2022. 1, 3, 6, 7

[58] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 2, 4

[59] Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens Van Der Maaten, Mark Campbell, and Kilian Q Weinberger. Anytime stereo image depth estimation on mobile devices. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5893–5900. IEEE, 2019. 3

[60] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. *Advances in neural information processing systems*, 32, 2019. 3

[61] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. Non-parametric depth distribution modelling based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8626–8634, 2022. 5

[62] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016. 3

[63] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6044–6053, 2019. 3

[64] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *arXiv preprint arXiv:2008.07928*, 2020. 3

[65] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *International Journal of Computer Vision*, 129:821–844, 2021. 2, 7

[66] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6278–6287, 2020. 3

[67] Shiyu Zhao, Long Zhao, Zhixing Zhang, Enyu Zhou, and Dimitris Metaxas. Global matching with overlapping attention for optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17592–17601, 2022. 2, 4

[68] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4669–4678, 2021. 3