Tech Details for LoFTR in the IMW Challenge

Xingyi He[1,2*] Yuang Wang[1,2*] Jiaming Sun[1,2*] Zehong Shen[1] Hujun Bao[1] Xiaowei Zhou[1†]
[1]Zhejiang University [2]SenseTime Research

Abstract

Technical details for the submission `29f8cfce` (named LoFTR_v4) of the 2021 Image Matching Challenge.

1. Method and Technical Details

Our method is based on LoFTR [6], a detector-free local feature matching method with Transformers [8]. Given two images to be matched, LoFTR directly produces a set of semi-dense matches conditioned on both images. Similar to other detector-free feature matching or correspondences estimation methods such as NCNet [5], [7], LoFTR does not depend on pre-extracted local features, thus lacks consistent features within a single image. However, the IMC benchmark requires the submission of a fixed set of keypoints for each image and based on which the matches of all possible image pairs. This submission format is specifically designed for traditional image matching pipelines based on local features, which conflicts with detector-free image matching pipelines. Therefore, we accompany LoFTR with a set of pre-extracted local features to strengthen the consistency of features within one view. Moreover, we design a postprocessing pipeline upon LoFTR's semi-dense matches to further merging features to satisfy the restriction on the number of features.

1.1. LoFTR and LoFTR-SPP

LoFTR builds semi-dense matches of two image pairs in a coarse-to-fine pipeline. First, coarse matches are built based on regular grids of coarse-level feature maps. Then, the coordinates of right matches are refined to a sub-pixel level using fine-level feature maps and a coarse-to-fine module. Though the right matches are refined, LoFTR always uses evenly distributed grid points as left matches, which is acceptable for estimating relative poses but hinders the performance of SfM because of inconsistent features within a single view and unrepeatable features among multiple views. We mitigate the above problems by pairing LoFTR with pre-extracted features (SuperPoint [2] is used in our submission) in the left view. After the establishment of coarse-level matches, we replace a left grid point with a pre-extracted keypoint if there exists a keypoint in the window centering around the grid point. Then, the coarse-to-fine module will refine the right match, orienting toward the matching of the left keypoint, instead of the grid point. We still keep a left grid point without any keypoint in its support region. This strategy keeps the semi-dense property of LoFTR, which is vital for relative pose estimation. We name this modified pipeline LoFTR-SPP. Note that this pipeline is different from guided-matching since the right matches are always refined.

1.2. Greedy Points Merging

After the bidirectional exhaustive matching of image pairs of a scene, we collect all points of each image from its matches with other images. The points consist of both integer-valued grid points when the image is acting as the left image and floating-valued sub-pixel refined points when it is acting as a right image. In general, an image generally contains twice more than the 8k points restriction; thus further point merging is required. We use a straightforward greedy strategy as shown in Fig. 1. First, we run non-maximum suppression upon all points within an image. The suppression radius of each image is determined adaptively through a bisection search, such that a maximum number of points satisfying the 8k restriction are kept. The sum of matching scores of each point among all of its matches produced in the exhaustive matching are used in the nms. We also add a small bonus to the scores of pre-extracted features in LoFTR-SPP for a biased selection of those points. We call the points surviving the nms process "pillar points." Then, we traverse through all points of an image except for those pillar points. If a point p_i is near to (upon a distance threshold) the pillar point ${p_p}^{(1)}$ it was suppressed by, we would merge it to ${p_p}^{(1)}$. However, there might be another pillar point ${p_p}^{(2)}$ nearer to p_i (with a smaller score for sure), we could optionally merge p_i to ${p_p}^{(2)}$ instead for a smaller localization error. Finally, we update the matches of those merged points for an acceptable IMC submission. Note that there might be conflicting one-to-many matches produced by our points merging pipeline, which could be further filtered and selected but is actually not handled by our submissions.

1.3. Implementation Details

We use custom geometric verification with DEGENSAC [1] upon our matches, with a maximum iterations of 10000 and a confidence threshold of 0.99999.

*Equal contribution.

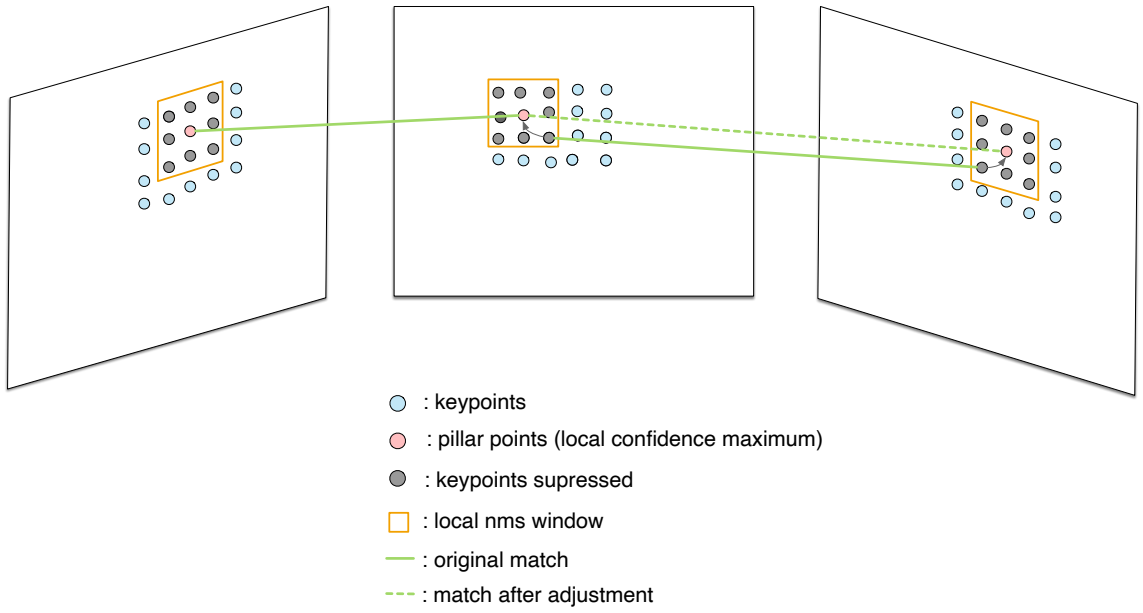

Figure 1. **Greedy points merging.** We extract pillar points using non-maximum suppression with bisection-searched suppression radiuses. Suppressed keypoints within a local window of their corresponding pillar points are merged to the pillar point. Matches of the suppressed keypoints are updated accordingly.

2. Dataset and Pre-trained Models

We use the MegaDepth [4] dataset to train our models, following the same setup as in [6]. We remove scenes used as the test and validation set in IMC, as well as scenes with low-quality depth maps pointed out by [3]. Among scenes kept, we enumerate all image pairs with covisible scores in a range of $[0.1, 0.7]$ and further split each scene into sub-scenes with covisible scores in ranges $[0.1, 0.3], [0.3, 0.5], [0.5, 0.7]$ respectively. These sub-scenes are used for training, leading to a training set composed of 368 sub-scenes in total. All models are trained from scratch, with no pre-trained model.

References

[1] Ondrej Chum, Tomas Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. In *CVPR*, 2005.

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

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

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

[5] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *NeurIPS*, 2018.

[6] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, 2021.

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

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