3DG-STFM: 3D Geometric Guided Student-Teacher Feature Matching

Runyu Mao[1], Chen Bai[2], Yatong An[2], and Fengqing Zhu[1] and Cheng Lu[2]

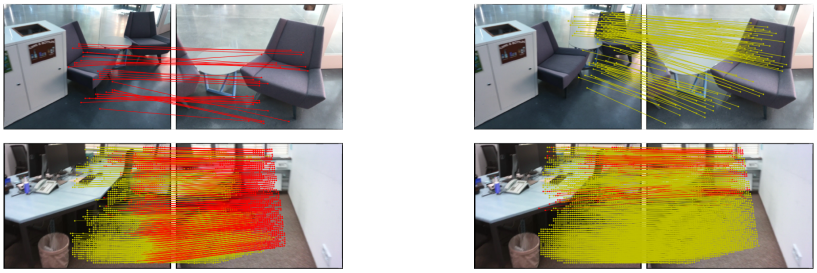
[1] Purdue University {mao111, zhu0}@purdue.edu
[2] XPeng Motors {chenbai, yatongan, luc}@xiaopeng.com

Abstract. We tackle the essential task of finding dense visual correspondences between a pair of images. This is a challenging problem due to various factors such as poor texture, repetitive patterns, illumination variation, and motion blur in practical scenarios. In contrast to methods that use dense correspondence ground-truths as direct supervision for local feature matching training, we train 3DG-STFM: a multi-modal matching model (Teacher) to enforce the depth consistency under 3D dense correspondence supervision and transfer the knowledge to 2D unimodal matching model (Student). Both teacher and student models consist of two transformer-based matching modules that obtain dense correspondences in a coarse-to-fine manner. The teacher model guides the student model to learn RGB-induced depth information for the matching purpose on both coarse and fine branches. We also evaluate 3DG-STFM on a model compression task. To the best of our knowledge, 3DG-STFM is the first student-teacher learning method for the local feature matching task. The experiments show that our method outperforms state-of-the-art methods on indoor and outdoor camera pose estimations, and homography estimation problems. Code is available at:https://github.com/Ryan-prime/3DG-STFM.

1 Introduction

Establishing correspondences between overlapped images is critical for many computer vision tasks including structure from motion (SfM), simultaneous localization and mapping (SLAM), visual localization, etc. Most existing methods that tackle this problem follow the classical tri-stage pipeline, i.e., feature detection [28,38], feature description [29,23,3,50,13,12], and feature matching [29,35,41]. To improve efficiency, HLoc [40] was proposed to incorporate these matching techniques for visual localization. Several recent works [35,36,24,44] attempted to avoid the detection step and established a dense matching by considering all points from a regular grid. These dense matching approaches aim to supply interest points in low-texture regions and provide sufficient candidates for the matching purpose.

To generate dense ground-truth correspondences as supervision, depth maps, camera intrinsic and extrinsic matrices are used for the calculation of point reprojections from one image to the other [41,44,24]. Although photometric objective,



(a) LoFTR (b) 3DG-STFM

Fig. 1: **Comparison between dense local feature matching method LoFTR [44] and the proposed method 3DG-STFM.** This example demonstrates that our approach, embedded the depth distribution via student-teacher learning, could find the correct correspondences under challenging scenario with repetitive patterns and low-texture regions. The red color indicates epipolar error beyond 5×10^{-4} (in the normalized image coordinates).

widely used in optical flow estimation [33,21,32], could provide dense correspondences, its constant brightness assumption is not allowed to be generalized for the geometric matching problem. One typical adversarial scenario is image pairs taken under radically different illumination. On the other hand, given a set of images with dense correspondences, triangulation could easily reconstruct the 3D scene and depth maps. Therefore, depth information is implicitly provided by dense correspondence supervision.

However, to the best of our knowledge, none of the existing methods explored the depth modality distribution during the training phase. Depth maps, unlike RGB images, provide 3D information, which depicts the geometry distribution in an explicit manner. We argue that the introduction of depth modality distribution can provide two-fold benefits. First, depth information, even if in lower quality or sparse, can remove lots of ambiguity in 2D image space and enforce geometric consistency for feature matching, which is very difficult using only RGB inputs. That is particularly true when there are multiple similar objects within the image pair. In that case, most of existing methods tend to find implausible matching candidates since they purely discriminate 2D descriptors without depth or size knowledge. An example is shown in the first row of Fig. 1, where the baseline method is confused by the similar 2D appearance and incorrectly matches the closer chair to the further one. Second, as the example shown in the second row of Fig. 1, low texture area of single object haunts 2D descriptor in terms of enforcing dense and consistent matching. That deficiency can also be nicely regularized by leveraging the discrimination of depth modality.

Despite the advantage of depth information, high quality RGB-D inputs can only be collected in well-controlled lab environment, and very few, especially low cost devices can capture similar well aligned RGB-D pairs in real world scenarios. Most imaging systems are only equipped with RGB sensors as input and cannot afford high computational cost stemmed from multi-modal inference. That makes naive multi-modal fusion of RGB and depth inputs during both inference and training a restrictive solution. Consequently, a good way of transferring expensive RGB-D knowledge into RGB modality inference is needed in practical scenarios, considering constraints from both hardware and computational load.

Motivated by these observations, we propose 3DG-STFM, a student-teacher learning framework, to transfer depth knowledge learned by a multi-modal teacher model to a unimodal student model to improve the local feature matching. To the best of our knowledge, 3DG-STFM is the first student-teacher learning architecture to transfer cross-modal knowledge on the image matching problem. The method aims to find the depth and RGB correlational distribution in RGB-D images and transfer the knowledge to the RGB student branch by maintaining such distribution. Therefore, depth modality is not explicitly required in the actual inference process (student branch).

We propose attention mechanisms to guide the student model to study the teacher model's matching distribution and learning priority. Therefore, with RGB images as input, the student unimodal model could explore RGB-induced depth information and learn multi-modal matching strategies. The main contributions of this paper are summarized as follows:

- We propose the first student-teacher learning architecture on the local feature matching problem that learns the induced depth distribution distilled from dense RGB-D correspondence supervision.
- We propose attentive knowledge transfer strategies to help the student model understand the matching distribution and learning priority during the training instead of learning point-to-point matching.
- We show that the proposed model produces high-quality dense correspondences on a range of matching tasks and achieves state-of-the-art results on both camera pose and homography estimation tasks.

2 Related Work

2.1 Learning-based Dense Local Feature Matching

In the past decades, many groups made great efforts to improve the local feature matching pipeline, i.e., feature detection, feature description and feature matching, and achieved promising performance by leveraging learning-based techniques. DeTone et al. proposed Superpoint [13], a self-supervised learnable interest point detector and descriptor. ViewSynth [31] designed a depth map keypoint detection method without using RGB domain information. Instead of learning better task-agnostic local features, SuperGlue [41] built a densely connected graph between two sets of keypoints by leveraging a Graph Neural Network (GNN). Geometric correlation of the keypoints and their visual features are integrated and exchanged within the GNN using the self and cross attention mechanism. However, those detector-based local feature matching algorithms only produced sparse keypoints, especially in low-texture regions.

To address the above problem, detector-free methods [36,24,44,22] proposed pixel-wise dense matching methods. [10] and [42] used contrastive loss to learn dense feature descriptors and were followed by the nearest neighbor search for the matching purpose. NCNet [35] proposed an end-to-end approach by directly learning the dense correspondences. It enumerated all possible matches between two images and constructed a 4D correlation tensor map. The 4D neighborhood consensus networks learned to identify reliable matching pairs and filtered out unreliable matches accordingly. Based on this concept, Sparse NCNet [36] improved NCNet's efficiency and performance by processing the 4D correlation map with submanifold sparse convolutions [17]. And DRC-Net [24] proposed a coarse-to-fine approach to generate higher accuracy dense correspondences. Recently, LoFTR [44] was proposed to learn global consensus between image correspondences by leveraging Transformers. Inspired by [41], the attention mechanism was used to learn the mutual relationship among features. For memory efficiency, the coarse matching features were first predicted and then fed to a small transformer to produce the final fine-level matches. Benefiting from the global receptive field of Transformers, LoFTR improved the matching performance by a large margin. All above mentioned dense local feature matching approaches needed dense ground-truth correspondences as supervision. None of the dense matching methods has explored any modality beyond 2D image space in which the feature ambiguities often exist due to the missing information in depth.

2.2 Student-Teacher Learning

Student-teacher learning has been actively studied in knowledge transfer context including model compression [1,19], acceleration [53,9], and cross-modal knowledge transfer [18,16]. Given a well-trained teacher model with large weight, the goal of the student-teacher learning is to distill and compress the knowledge from the teacher, and guides the lightweight student model for better performance. On the other hand, data with multiple modalities commonly provides more valuable supervisions than single modality data and could benefit model performance. However, due to the lack of data or labels for some modalities during training or testing, it is important to transfer knowledge between different modalities.

Due to different network architectures, many different knowledge transfer approaches have been proposed. The most popular response-based knowledge for image classification was Knowledge Distillation (KD) loss proposed by [19]. In this method, KD loss employed the distribution of neural response of the last output layer, logits layer, of the teacher model and guided the student to learn the distribution. Besides the output, the intermediate layer's feature representation was also used to train the student model [37]. Zagoruyko et al. [54] proposed a method to transfer the attention instead of the feature representations to achieve a better distillation performance. And NST [20] provided a method to learn a similar activation of the neurons. Moreover, there are many other related approaches [51,27,49,47,7,45]. However, none of them provided a knowledge transfer solution for correspondence matching problems, which need to consider all mutual relationships among local features of different images.

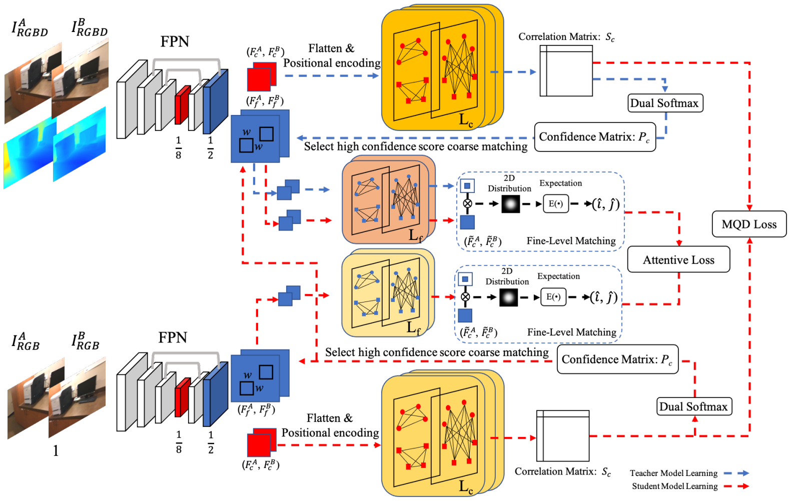


Fig. 2: **Overview of 3DG-STFM.** For each of student or teacher branch, Feature Pyramid Networks (FPN) [26] are used to extract coarse-level local features (F^A_c, F^B_c) and fine-level features (F^A_f, F^B_f) with $\frac{1}{8}$ and $\frac{1}{2}$ of the original image resolution. The coarse level transformer consisting of $\mathrm{L_c}$ attention layers finds coarse pairs and their matching scores. Matches with high confidence scores will be selected and mapped to a fine-level feature map. Surrounding features on F^A_f, F^B_f are collected by the $w \times w$ size window and fed to a fine-level transformer with $\mathrm{L_f}$ attention layers. The fine-level matching module is applied to predict correspondences $(\hat{i}, \hat{j})$ on subpixel-level. The teacher model is first trained under direct supervision. During student training, it will be frozen and provide additional supervision via attentive loss and Mutual Query Divergence (MQD) loss.

3 Method

Our proposed system, 3DG-STFM, is to train a unimodal local feature matching model (Student) by leveraging the knowledge from a well-trained multi-modal model (Teacher). As shown in Fig. 2, the RGBD image pairs (I^A_{RGBD}, I^B_{RGBD}) and RGB image pairs (I^A_{RGB}, I^B_{RGB}) are fed to teacher and student branches separately. The labels of dense correspondences provide direct supervision during the teacher or student training. Once we reach a well-trained multi-modal teacher model, two strategies are proposed for cross-modal knowledge transfer: (1) Using the Mutual Query Divergence (MQD) loss guides the student model to learn the coarse-level matching distributions embedded in the teacher model's correlation matrix S_c. (2) Using the attentive loss guides the student at the fine-level module to pay more attention to the teacher's confident predictions and learn the matching distribution with priority.

Our method is based on the matching strategies mentioned in LoFTR [44] due to their high performances. In this section, we will first introduce the

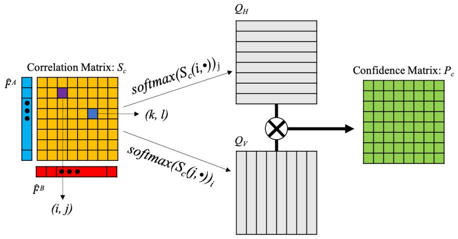

Q_H
Correlation Matrix: S_c
$\hat{F}^A$
softmax$(S_c(i,\bullet))_j$
(k, l)
Confidence Matrix: P_c
Q_V
softmax$(S_c(j,\bullet))_i$
$\hat{F}^B$
(i, j)

Fig. 3: **Coarse-level differentiable matching mechanism.** Transformer outputs are correlated to generate correlation matrix S_c. Dual-softmax [46,35] operation is applied on two dimensions to obtain the matching probability.

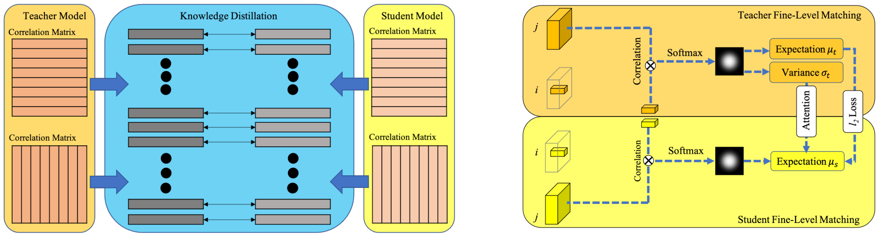
transformer-based model in Section 3.1. Section 3.2 and 3.3 will describe our knowledge transfer strategies over both coarse and fine levels.

3.1 Transformer-based Local Feature Matching

As shown in Fig. 2, two transformer-based matching modules, inspired by [44], are adopted in both teacher and student branches of our 3DG-STFM system.
Coarse-level Matching. Given the coarse-level feature map in dimension $h \times w \times c$, we flatten them into $hw \times c$ and do the positional encoding [5]. The encoded local feature vector will be fed to a coarse-level matching transformer. Unlike classical vision transformer [14,6,48] focusing on self-attention, the matching transformer adds a cross-attention layer to consider the relations between pixels from different images. We interleave the self-attention and cross-attention layers in matching transformer modules by $\mathrm{L_c}$ times. As shown in Fig. 3, the output of the coarse matching transformer $\{\hat{F}^A, \hat{F}^B\}$ corresponding to two different images $\{I^A, I^B\}$ will be used to calculate correlation matrix S_c by $S_c(i,j) = Corr(\hat{F}_i^A, \hat{F}_j^B)$, in which $\hat{F}_i^A$ and $\hat{F}_j^B$ indicate local feature at position i of I^A and local feature at position j of I^B. The dual-softmax [46,35], two softmax operations with temperature $\tau = 0.1$ in horizontal and vertical directions, is applied on the correlation matrix to calculate forward and backward matching probability: $P_{A\rightarrow B}(i,j) = softmax(\frac{1}{\tau}S_c(i,\cdot))_j$ and $P_{A\leftarrow B}(i,j) = softmax(\frac{1}{\tau}$ $S_c(\cdot,j))_i$. The confidence matrix P_c with the final matching probabilities has same dimension as S_c and is calculated by: $P_c(i,j) = P_{A\rightarrow B}(i,j) \cdot P_{A\leftarrow B}(i,j)$. We call the output of horizontal and vertical softmax as query matrix $\{Q_H, Q_V\}$ since they depict query results of each feature from one image to another and vice versa. Given the ground-truth matrix derived from correspondence labels, we calculate the cross-entropy loss:

$$\mathcal{L}_c = -\frac{1}{|\mathcal{M}_{gt}|} \sum_{(i,j)\in\mathcal{M}_c^{gt}} FL(P_c(i,j)) \log P_c(i,j) \tag{1}$$

$$FL(p) = \alpha(1-\hat{p})^{\gamma}, \hat{p} = \begin{cases} p \text{ if y=1} \\ 1-p \text{ otherwise} \end{cases} \tag{2}$$



(a) Coarse-level knowledge transfer. (b) Fine-level attentive knowledge transfer

Fig. 4: **Knowledge transfer on coarse-level and fine-level.** (a) The correlation matrix is decomposed into multiple independent distributions depicting mutual query processes for the student learning.(b) One center point of a fine-level feature is selected and correlates with all points of the other feature map for heatmap distribution generation. Both expectation and variance of teacher branch's heatmap are used for fine-level knowledge transfer.

in which P_c is the confidence matrix and $\mathcal{M}_{gt}$ is the correspondence set generated by ground-truth labels. We follow [44] to set a focal loss term, FL with predicted probability p, to address the imbalance between matching and non-matching pairs.

Fine-level Matching. Based on the confidence matrix P_c, matching pairs with probability scores higher than a threshold θ_c are selected and refined by a fine-level matching module. The selected coarse-level features are upsampled and concatenated to fine-level features cropped by $w \times w$ size windows before passing to the fine-level matching transformer.

The fine-level matching transformer is a lightweight transformer containing L_f attention layers. It aggregates the contextual information to generate features $\{\tilde{F}_f^A, \tilde{F}_f^B\}$ and passes them to a differentiable matching module. Instead of generating a confidence matrix, the fine-level matching module selects the center feature of $\tilde{F}_f^A$ and correlates with all features in $\tilde{F}_f^B$. The similarity distribution is generated and the expectation μ is treated as the prediction. The final loss based on direct supervision is calculated by:

$$\mathcal{L}_f = \frac{1}{|\mathcal{M}_f|} \sum_{(\hat{i},\hat{j}) \in \mathcal{M}_f} \frac{1}{\sigma^2(\hat{i})} ||\mu(\hat{i}) - \hat{j}_{gt}||_2^2 \tag{3}$$

where $\hat{j}_{gt}$ is the ground-truth position we wrap from image solution to fine-level heatmap scale. $\mu(\hat{i})$ is the prediction associated to coarse position $\hat{i}$ and $\sigma^2(\hat{i})$ is the total variance of heatmap distribution. $\mathcal{M}_f$ is the set of fine matches predicted by module. The total variance of the similarity distribution is treated as uncertainty to assign a weight to each fine-level match. The larger total variance indicates it is an uncertain prediction and associate with low weights.

3.2 Coarse-level Knowledge Distillation

A response-based knowledge distillation strategy is applied to help the student learn from the teacher on a coarse level. This method distills the logits layer's distribution and guides the student to learn. As aforementioned, the logits layer's output in our case is correlation matrix S_c with size $hw \times hw$. Each row or column depicts the relation between one pixel and each pixel of the other image. The dual softmax operation could be treated as a query process in two directions. Local features at position $i \in F_c^A$ retrieve the closest feature from all positions $j \in F_c^B$ and vice versa. Many existing response-based knowledge [19,18] distillation methods treat the logits layer's output as a single distribution. However, as shown in Fig. 3, exploring the relationship between $S_c(i, j)$ and $S_c(k, l)$ in different row and column is meaningless for matching purpose. These uninterpretable relations could produce extra loss that confuses the knowledge transfer process.

Instead of learning a single distribution from the correlation matrix S_c in teacher, we split the distribution into two matching query matrices, as shown in Fig. 4a, to avoid the unpredictable correlations between them. Based on Knowledge Distillation (KD) loss [19], we propose Mutual Query Divergence loss $\mathcal{L}_{MQD}$ that employs all $2 \times hw$ mutual query distributions:

$$\mathcal{L}_{MQD} = \frac{1}{n}\big(-\sum_{i=1}^{n} FL(p_S^{(i)})\hat{p}_S^{(i)}\log(\hat{p}_T^{(i)})\big) \tag{4}$$

$$p_S^{(l)} = \frac{\exp(o_S^{(k)})}{\sum_{k=1}^{L}\exp(o_S^{k})}, \hat{p}_S^{(l)} = \frac{\exp(\frac{o_S^{(k)}}{T})}{\sum_{k=1}^{L}\exp(\frac{o_S^{k}}{T})}, \hat{p}_T^{(l)} = \frac{\exp(\frac{o_T^{(k)}}{T})}{\sum_{k=1}^{L}\exp(\frac{o_T^{k}}{T})} \tag{5}$$

in which $\hat{p}_S^l$ and $\hat{p}_T^l$ are student and teacher's query distributions distilled from their logits layer's outputs o_S and o_T at temperature T. Additional focal loss weight FL (Equation 2) is added to balance the matching/unmatching ground-truth pairs. p_S^l is the standard confidence score predicted by student model expressed in Equation 5. The $\mathcal{L}_{MQD}$ on coarse level is the mean of KD loss of all n distribution, where n is equal to $2 \times hw$ in our case. Based on this loss, the coarse level matching module pays attention to the distributions benefit matching and ignores noisy information across distributions.

3.3 Fine-level Attentive Knowledge Transfer

After transferring the coarse level matching knowledge from the teacher model with the mutual query distribution distillation, an attentive loss ($\mathcal{L}_{att}$) is proposed for the student's fine-level matching module. Instead of learning point-to-point matching under the supervision of ground-truth, $\mathcal{L}_{att}$ explores the matching distribution and learning priority of the teacher model.

As shown in Fig. 4b, the fine-level local feature matching is based on the differentiable matching approach that could produce a heatmap that represents the matching probability of each pixel in the neighborhood of j with i. By computing expectation μ over the probability distribution, we get the final position $\hat{j}$ with sub-pixel accuracy on I^B. The uncertainty of the prediction is also measured by the total variance of the correlation distribution. During the student-teacher learning process, both branches could generate heatmaps. We treat heatmaps of teacher model and student model as gaussian distributions $\mathcal{N}_t(\mu_t, \sigma_t^2)$ and $\mathcal{N}_s(\mu_s, \sigma_s^2)$. The Kullback–Leibler (KL) divergence loss ($\mathcal{L}_{KL}$) is applied to help the student learn the distribution from the teacher. The KL divergence of two gaussian distributions could be written as:

$$\mathcal{L}_{KL}(\mathcal{N}_s, \mathcal{N}_t) = \log(\frac{\sigma_t}{\sigma_s}) + \frac{\sigma_s^2 + (\mu_s - \mu_t)^2}{2\sigma_t^2} - \frac{1}{2} \tag{6}$$

Although total variance $\{\sigma_t, \sigma_s\}$, and $\sigma(\hat{i})$ (Equation 3) are included in the loss, the optimizer would decrease loss by increase the total variance. To avoid the incorrect loss, the gradient is not backpropagated through σ_s, σ_t, and $\sigma(\hat{i})$. Therefore, we generate $\mathcal{L}_{att}$ by removing those constant variable of $\mathcal{L}_{KL}$:

$$\mathcal{L}_{att} = \frac{1}{|\mathcal{M}_f|} \sum_{(\hat{i},\hat{j}) \in \mathcal{M}_f} \frac{(\mu_s^{(\hat{i})} - \mu_t^{(\hat{i})})^2}{2\sigma_t^{(\hat{i})2}} \tag{7}$$

where the $\mu_s^{(\hat{i})}$ and the $\mu_t^{(\hat{i})}$ are the expectations of student's and teacher's output distributions which corresponding to match $(\hat{i}, \hat{j})$ in fine-level correspondence set $\mathcal{M}_f$. Therefore, the total loss is a mean of the weighted sum of all the fine-level pairs' l_2 loss in matching set $\mathcal{M}_f$. We call this attentive loss since it could be treated as a l_2 distance loss that pays more attention to the prediction associated with large attention weight $\frac{1}{2\sigma_t^2}$. The total variance is commonly treated as a metric for certainty measure. The teacher prediction with a small total variance indicates the teacher is quite certain about the location of the correspondence. In this case, the loss is assigned with a large weight to guide the student model to learn those certain predictions from the teacher in priority.

3.4 Supervision

Both teacher and student training processes are under the direct supervision provided by correspondence ground-truths. The teacher model provides extra supervision during the student model training. For direct supervision, we follow the same procedure mentioned in [41,35,44] that uses the camera intrinsic, extrinsic matrices, and depth maps to compute the dense correspondences. To supervise coarse-level matching training, mutual nearest neighbors of the two sets of $\frac{1}{8}$-resolution grids are selected as ground-truth $\mathcal{M}_c^{gt}$. The pixel-level matching positions could be used for l_2 loss and supervise the fine-level matching learning. The final loss for the teacher and student model is:

$$\mathcal{L}_{teacher} = \lambda_0 \mathcal{L}_c + \lambda_1 \mathcal{L}_f \tag{8}$$

$$\mathcal{L}_{student} = \lambda_0 \mathcal{L}_c + \lambda_1 \mathcal{L}_f + \lambda_2 \mathcal{L}_{MQD} + \lambda_3 \mathcal{L}_{att} \tag{9}$$

in which $\mathcal{L}_c$ and $\mathcal{L}_f$ are coarse-level and fine-level loss under direct supervision described in Equation 1 and Equation 3. The student model is also guided by the teacher model via Mutual Query Divergence loss $\mathcal{L}_{MQD}$ and attentive loss $\mathcal{L}_{att}$ for the coarse and the fine level knowledge transfer.

3.5 Implementation Details

We train the indoor model of 3DG-STFM on the ScanNet [11] dataset and the outdoor model on the MegaDepth [25] dataset. The coarse-level transformer contains 4 attention layers, and the fine-level transformer has 1 attention layer. Each attention layer consists of a self-attention and a cross-attention layer with 8 heads. The focal loss parameters $\{\alpha, \gamma\}$ are set as $\{0.25, 2.0\}$. The confidence score threshold θ_c is set to 0.2 to remove unreliable correspondences. The window size w is 5. For indoor dataset ScanNet, the models are trained using AdamW with an initial learning rate of 6×10^{-3} on 32 2080Ti GPUs. All images are resized to 640×480. The weights of losses $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}$ are set as $\{0.25, 0.25, 4.0, 0.25\}$. The outdoor models for Megadepth are trained using AdamW with an initial learning rate of 8×10^{-3} on 16 P100 GPUs. The weights of losses $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}$ are set as $\{0.25, 0.25, 1.0, 0.25\}$. It is worth mentioning that our method is based on LoFTR [44], which provides two version implementations for the outdoor dataset in their official code. One is for 840×840 resolution image and consumes 24 GB RAM during the training. The other is training on 640×640 image pairs and feasible for 16 GB RAM GPUs. In this work, we treat the latter one as the baseline for the outdoor pose estimation and homography estimation tasks, and our 3DG-STFM is also trained on images resized to 640×640 with padding. We normalize depth maps in both ScanNet and Megadepth in the training process. The depth maps of ScanNet are in the range of 0 to 10 meters. We normalize it to $[0, 1]$ and concatenate it to RGB images for multi-modal training. On the other hand, Megadepth's depth maps are relative estimations that come from COLMAP [43] reconstructions and have a pretty large range. We normalized them to $[0, 1]$ for each pair of images for teacher model training.

4 Experiments

4.1 Indoor Pose Estimation

Dataset. We use ScanNet [11], a large-scale indoor scene dataset composed of 1613 monocular sequences with depth maps and camera poses. This dataset is quite challenging due to extensive texture-less regions and repetitive patterns. Following the [41,44], we sample 230M image pairs with overlap scores between 0.4 and 0.8 for training and the student model is evaluated on the 1500 testing pairs. The images are resized to 640×480 to fit the depth map's dimension.
Evaluation Protocol. Following [44], we report the AUC of the pose error at thresholds $(5°, 10°, 20°)$. The pose error is defined as the maximum of angular error in rotation and translation. The predicted matches are used to solve the essential matrix with RANSAC.

Table 1: **Evaluation on ScanNet [11] for indoor pose estimation.** The AUC of the pose error in percentage is reported.

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Multi-Modal	3DG-STFM Teacher	27.93	47.11	63.74
Detector-based	ORB [39]+GMS [4]	5.21	13.65	25.36
	D2-Net [15]+NN	5.25	14.53	27.96
	ContextDesc [30]+Ratio Test [29]	6.64	15.01	25.75
	SP [13]+NN	9.43	21.53	36.40
	SP [13]+PointCN [52]	11.40	25.47	41.41
	SP [13]+OANet [55]	11.76	26.90	43.85
	SP [13]+SGMNet [8]	15.40	32.06	48.32
	SP [13]+SuperGlue [41]	16.16	33.81	51.84
Detector free	LoFTR [44]	22.06	40.80	57.62
	3DG-STFM Student	**23.58**	**43.60**	**61.17**

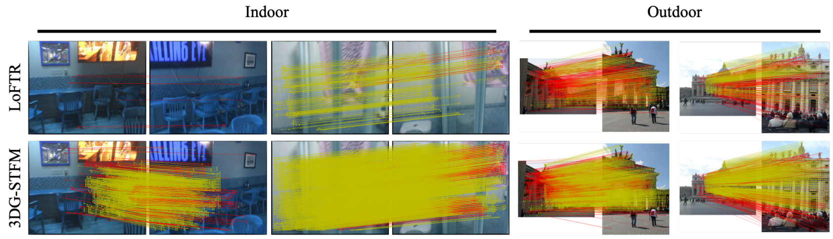


Fig. 5: **Qualitative results.** Our student model is compared to LoFTR [44] in indoor and outdoor scenes. Our method performs better in challenge scenarios with repetitive pattern and low texture region. The red color indicates epipolar error beyond 5×10^{-4} for indoor scenes and 1×10^{-4} for outdoor scenes (in the normalized image coordinates). More qualitative results in the supplementary.

Results. Since the released DRC-Net is trained on MegaDepth and LoFTR is proved to have better performance, we only consider LoFTR as the state-of-the-art for comparison. The results in Table 1 show that our student model learns from the teacher model and outperforms all unimodal competitors. For detector free methods, our student model outperforms LoFTR by $\sim 3\%$ at AUC@10°.

For the visual comparison in Fig. 5, our student model shows denser and more reliable correspondences than LoFTR does, especially in regions with repetitive patterns. In addition, our model provides more robust correspondences in the low-texture region, which also benefits the pose estimation. On average, our student model detects **1192.84** inlier (epipolar error less than 5×10^{-4} in the normalized image coordinates) correspondences on each pair of indoor images, which is much higher than **887.04** inlier correspondences of LoFTR. Both numeric and qualitative results demonstrate the effectiveness of our student model that learns the RGB-induced depth distribution from the teacher model.

Table 2: **Evaluation on MegaDepth [25] for outdoor pose estimation.** The AUC of the pose error in percentage is reported.

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Multi-Modal	3DG-STFM Teacher	53.43	69.81	81.79
Detector-based	SP [13]+SuperGlue [41]	42.18	61.16	75.96
Detector free	DRC-Net [24]	27.01	42.96	58.31
	LoFTR [44]	51.38	67.11	79.29
	3DG-STFM Student	**52.58**	**68.46**	**80.04**

4.2 Outdoor Pose Estimation

Dataset. We use MegaDepth [25], a dataset consisting of 1M internet images of 196 different outdoor scenes, for outdoor pose estimation evaluation. We follow DISK [46] to select 1500 pairs for validation.

Results. We resize the images with the long side to 1200 during the inference and follow the same evaluation protocol as indoor pose estimation. As shown in Fig. 5, since the outdoor images contain less low texture regions and repetitive patterns, the unimodal model baseline (LoFTR) could also predict many correct correspondences for robust camera pose estimations. However, the results in Table 2 indicate that our 3DG-STFM teacher model achieves better performance by leveraging the relative depth. The student model learned from the teacher could also outperform LoFTR, the state-of-art unimodal competitor. We find our student model averagely detects **1864.63** inlier (epipolar error less than 1×10^{-4} in the normalized image coordinates) correspondences on each outdoor image pair, which is also higher than LoFTR's **1694.60** inlier detections.

4.3 Homography Estimation

We also evaluate our student model for homography estimation on HPatches dataset [2]. Following previous work [41,44], we select 108 image sequences under large illumination changes or significant viewpoint variations for evaluation. Every test image sequence contains one reference image and five pairing images.

Evaluation Protocol. We resize the original images with shorter dimensions equal to 480 and find the top 1K correspondences for each pair for detector free methods. Our 3DG-STFM student model is trained on Megadepth [25] mentioned in Section 4.2. All baseline results are reported using their original default implementation hyperparameters. Homography estimation is performed by the OpenCV RANSAC implementation. Following [13], we compute the reprojected mean error of the four corners of the image and report the area under the cumulative curve (AUC) up to three values: 3, 5, and 10 pixels in Table 3.

Results. Our 3DG-STFM student model is generalized well on the homography estimation task and achieves best performance compare with detector free

Table 3: **Homography estimation on HPatches [2].** The AUC of the corner error in percentage is reported.

Category	Method	Homography est. AUC			#matches
		@3px	@5px	@10px	
Detector-based	D2Net [15]+NN	23.2	35.9	53.6	0.2k
	R2D2 [34]+NN	50.6	63.9	76.8	0.5k
	DISK [46]+NN	52.3	64.9	78.9	1.1k
	SP [13]+SuperGlue [41]	53.9	68.3	**81.7**	0.6k
Detector free	Sparse-NCNet [36]	48.9	54.2	67.1	1.0k
	DRC-Net [24]	50.6	56.2	68.3	1.0k
	LoFTR [44]	63.4	71.9	79.9	1.0k
	3DG-STFM	**64.7**	**73.1**	81.0	1.0k

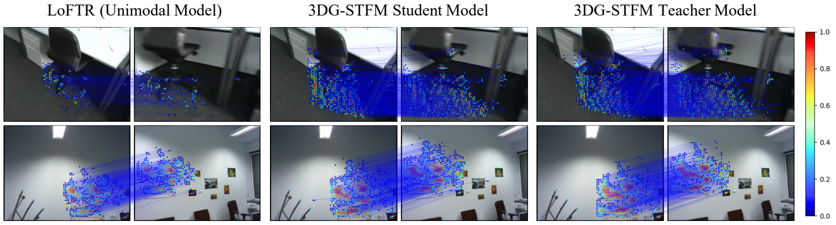


Fig. 6: **Visualization of matching distribution change for better understand student-teacher learning.** The color of correspondence scatter is determined by the confidence score predictions of each model. The teacher model not only guide the student model to find more correspondences, but also teaches the confidence score distribution to the student model.

methods as shown in Table 3. The method based on Superpoint [13] and Superglue [41] get better performance at AUC@10px than our approach. However, the 3DG-STFM student model shows more accurate performances under the other two strict metrics. We provide more details in the supplementary material.

4.4 Student-Teacher Learning Understanding

Visualizing Knowledge Transfer. To understand how our teacher model transfers knowledge to the student, we visualize the matching details to compare our student model and teacher model on ScanNet. We remove the teacher branch and training student branch solely based on direct supervision and treat it as the vanilla unimodal model for comparisons. Since we adopt the LoFTR's matching strategy, this vanilla unimodal model is the same as LoFTR. In Fig. 6, we plot all the predicted matches of models with a confidence score higher than 0.2. We show in the first row of Fig. 6 that both the teacher and student model find much more correspondences around low-texture regions than state-of-the-art. The teacher model explores the depth modality and then guides the student model to learn the RGB-induced depth information to increase the discriminant

Table 4: **Ablation study.**

Method	Pose estimation AUC		
	@5°	@10°	@20°
Multi-model Teacher	18.41	36.53	54.07
Unimodal	14.78	31.47	48.44
Unimodal+MQD	16.46	33.62	51.70
Unimodal+MQD+Att	**17.05**	**34.77**	**52.26**

Table 5: **Model compression study.**

Method	L_c	L_f	Pose estimation AUC		
			@5°	@10°	@20°
Teacher Model	4	1	18.41	36.53	54.07
Full-Size Student Model	4	1	17.05	34.77	52.26
Full-Size Model	4	1	**14.78**	31.47	48.44
Slim Model	2	1	14.18	29.68	46.45
Slim Student Model	2	1	14.49	**31.76**	**49.51**

features in areas with low texture but depth variations. We also show that the student follows the teacher's confidence score pattern, while both have different patterns compared to LoFTR, shown in the second row of Fig. 6. The confidence scores are indicated by color, high in red, low in blue. This knowledge transfer is achieved by proposed MQD loss and attentive loss for coarse-level and fine-level matching for our student-teacher architecture.

Ablation Study. To better understand the contribution in each module, we randomly select 150 scenes from ScanNet as a mini version dataset and test different variants of our model. The test set is the same as the original ScanNet. As shown in Table 4, the teacher model achieves the best performance and is used to teach the two models, i.e., Unimodal+MQD and Unimodal+MQD+Att. The unimodal model is trained under direct supervision provided by dense correspondences labels based on Equation 8. Compared with the unimodal model, both MQD and attentive loss help the knowledge transfer from teacher to student.

Model Compression Performance. Our architecture can also be generalized to model compression tasks. We implement the model compression experiments on the mini version of ScanNet. The results shown in Table 5 indicate our slim model has the competitive performance with the uncompressed model. In Table 5, the slim matching model is proposed with half attention layers on the coarse-level transformer L_c. The Slim Student Model is trained with our student-teacher architecture, while both the Full-Size Model and the Slim Model are trained under the direct supervision provided by ground-truth. By learning knowledge from the Teacher Model, the Slim Student Model improves 2% compared to the Slim Model and also shows better performance than the Full-Size Model at AUC@10° and @20°. More details are provided in the supplementary.

5 Conclusion

In this paper, we propose 3DG-STFM: a novel student-teacher learning framework for the dense local feature matching problem. Our proposed framework mines depth knowledge from one multi-modal teacher model to guide the student model to learn the hidden depth information embedded in the RGB domain. Two attentive mechanisms, i.e., MQD loss and attentive loss, are proposed to help the knowledge transfer. Our student model is evaluated on several image matching and camera pose estimation tasks on indoor and outdoor datasets and achieves state-of-the-art performances. Our 3DG-STFM also shows generalization ability on model compression tasks.

References

1. Ba, L.J., Caruana, R.: Do deep nets really need to be deep? arXiv preprint arXiv:1312.6184 (2013) 4
2. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 5173–5182 (2017) 12, 13
3. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. European conference on computer vision pp. 404–417 (2006) 1
4. Bian, J., Lin, W.Y., Matsushita, Y., Yeung, S.K., Nguyen, T.D., Cheng, M.M.: Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 4181–4190 (2017) 11
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. European Conference on Computer Vision pp. 213–229 (2020) 6
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. European Conference on Computer Vision pp. 213–229 (2020) 6
7. Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., Tian, Q.: Data-free learning of student networks. Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 3514–3522 (2019) 4
8. Chen, H., Luo, Z., Zhang, J., Zhou, L., Bai, X., Hu, Z., Tai, C.L., Quan, L.: Learning to match features with seeded graph matching network. Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 6301–6310 (2021) 11
9. Chen, T., Goodfellow, I., Shlens, J.: Net2net: Accelerating learning via knowledge transfer. arXiv preprint arXiv:1511.05641 (2015) 4
10. Choy, C.B., Gwak, J., Savarese, S., Chandraker, M.: Universal correspondence network. Advances in Neural Information Processing Systems (2016) 4
11. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 5828–5839 (2017) 10, 11
12. DeTone, D., Malisiewicz, T., Rabinovich, A.: Toward geometric deep slam. arXiv preprint arXiv:1707.07410 (2017) 1
13. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. Proceedings of the IEEE conference on computer vision and pattern recognition workshops pp. 224–236 (2018) 1, 3, 11, 12, 13
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 6
15. Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T.: D2-net: A trainable cnn for joint description and detection of local features. Proceedings of the ieee/cvf conference on computer vision and pattern recognition pp. 8092–8101 (2019) 11, 13
16. Garcia, N.C., Morerio, P., Murino, V.: Modality distillation with multiple stream networks for action recognition. Proceedings of the European Conference on Computer Vision (ECCV) pp. 103–118 (2018) 4

17. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 9224–9232 (2018) 4
18. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer pp. 2827–2836 (2016) 4, 8
19. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 4, 8
20. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219 (2017) 4
21. Jason, J.Y., Harley, A.W., Derpanis, K.G.: Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. European Conference on Computer Vision pp. 3–10 (2016) 2
22. Jiang, W., Trulls, E., Hosang, J., Tagliasacchi, A., Yi, K.M.: Cotr: Correspondence transformer for matching across images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6207–6217 (October 2021) 4
23. Leutenegger, S., Chli, M., Siegwart, R.Y.: Brisk: Binary robust invariant scalable keypoints. 2011 International conference on computer vision pp. 2548–2555 (2011) 1
24. Li, X., Han, K., Li, S., Prisacariu, V.: Dual-resolution correspondence networks. Advances in Neural Information Processing Systems **33** (2020) 1, 4, 12, 13
25. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 2041–2050 (2018) 10, 12
26. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 2117–2125 (2017) 5
27. Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7096–7104 (2019) 4
28. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004) 1
29. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004) 1, 11
30. Luo, Z., Shen, T., Zhou, L., Zhang, J., Yao, Y., Li, S., Fang, T., Quan, L.: Contextdesc: Local descriptor augmentation with cross-modality context. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 2527–2536 (2019) 11
31. Mahmud, J., Singh, R.V., Akiva, P., Kundu, S., Peng, K., Frahm, J.: Viewsynth: Learning local features from depth using view synthesis. 31st British Machine Vision Conference (2020) 3
32. Meister, S., Hur, J., Roth, S.: Unflow: Unsupervised learning of optical flow with a bidirectional census loss. Thirty-Second AAAI Conference on Artificial Intelligence (2018) 2
33. Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., Zha, H.: Unsupervised deep learning for optical flow estimation. Thirty-First AAAI Conference on Artificial Intelligence (2017) 2
34. Revaud, J., De Souza, C., Humenberger, M., Weinzaepfel, P.: R2d2: Reliable and repeatable detector and descriptor. Advances in neural information processing systems **32**, 12405–12415 (2019) 13

35. Rocco, I., Cimpoi, M., Arandjelović, R., Torii, A., Pajdla, T., Sivic, J.: Neighbourhood consensus networks. Proceedings of the 32nd Conference on Neural Information Processing Systems (2018) 1, 4, 6, 9
36. Rocco, I., Arandjelović, R., Sivic, J.: Efficient neighbourhood consensus networks via submanifold sparse convolutions. European Conference on Computer Vision pp. 605–621 (2020) 1, 4, 13
37. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014) 4
38. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. European conference on computer vision pp. 430–443 (2006) 1
39. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. 2011 International conference on computer vision pp. 2564–2571 (2011) 11
40. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 12716–12725 (2019) 1
41. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition pp. 4938–4947 (2020) 1, 3, 4, 9, 10, 11, 12, 13
42. Schmidt, T., Newcombe, R., Fox, D.: Self-supervised visual descriptor learning for dense correspondence. IEEE Robotics and Automation Letters **2**(2), 420–427 (2016) 4
43. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 10
44. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 8922–8931 (2021) 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 13
45. Tarvainen, A., Valpola, H.: Weight-averaged consistency targets improve semi-supervised deep learning results. corr abs/1703.01780. arXiv preprint arXiv:1703.01780 (2017) 4
46. Tyszkiewicz, M.J., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. arXiv preprint arXiv:2006.13566 (2020) 6, 12, 13
47. Wang, H., Lian, D., Ge, Y.: Binarized collaborative filtering with distilling graph convolutional networks. arXiv preprint arXiv:1906.01829 (2019) 4
48. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. European Conference on Computer Vision pp. 108–126 (2020) 6
49. Wang, Z., Deng, Z., Wang, S.: Accelerating convolutional neural networks with dominant convolutional kernel and knowledge pre-regression. European Conference on Computer Vision pp. 533–548 (2016) 4
50. Yang, T.Y., Hsu, J.H., Lin, Y.Y., Chuang, Y.Y.: Deepcd: Learning deep complementary descriptors for patch representations. Proceedings of the IEEE International Conference on Computer Vision pp. 3314–3322 (2017) 1
51. Yang, Y., Qiu, J., Song, M., Tao, D., Wang, X.: Distilling knowledge from graph convolutional networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7074–7083 (2020) 4
52. Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P.: Learning to find good correspondences. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 2666–2674 (2018) 11

53. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning pp. 4133–4141 (2017) 4
54. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016) 4
55. Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., Chen, Y., Quan, L., Liao, H.: Learning two-view correspondences and geometry using order-aware network. Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 5845–5854 (2019) 11