2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10611456

Generalized Correspondence Matching via Flexible Hierarchical Refinement and Patch Descriptor Distillation

Yu Han, Ziwei Long, Yanting Zhang✉, Jin Wu, Zhijun Fang, Rui Fan✉

***Abstract*— Correspondence matching plays a crucial role in numerous robotics applications. In comparison to conventional hand-crafted methods and recent data-driven approaches, there is significant interest in plug-and-play algorithms that make full use of pre-trained backbone networks for multi-scale feature extraction and leverage hierarchical refinement strategies to generate matched correspondences. The primary focus of this paper is to address the limitations of deep feature matching (DFM), a state-of-the-art (SoTA) plug-and-play correspondence matching approach. First, we eliminate the pre-defined threshold employed in the hierarchical refinement process of DFM by leveraging a more flexible nearest neighbor search strategy, thereby preventing the exclusion of repetitive yet valid matches during the early stages. Our second technical contribution is the integration of a patch descriptor, which extends the applicability of DFM to accommodate a wide range of backbone networks pre-trained across diverse computer vision tasks, including image classification, semantic segmentation, and stereo matching. Taking into account the practical applicability of our method in real-world robotics applications, we also propose a novel patch descriptor distillation strategy to further reduce the computational complexity of correspondence matching. Extensive experiments conducted on three public datasets demonstrate the superior performance of our proposed method. Specifically, it achieves an overall performance in terms of mean matching accuracy of 0.68, 0.92, and 0.95 with respect to the tolerances of 1, 3, and 5 pixels, respectively, on the HPatches dataset, outperforming all other SoTA algorithms. Our source code, demo video, and supplement are publicly available at `mias.group/GCM`.**

I. Introduction

Correspondence matching between images is crucial for a wide range of computer vision and robotics applications, *e.g.*, simultaneous localization and mapping [1]–[3], 3D geometry reconstruction [4]–[6], and stereo matching [7]–[10]. Conventional hand-crafted approaches extract keypoints using human-designed local feature detectors and descriptors, such as the scale-invariant feature transform (SIFT) [11] and speeded up robust features (SURF) [12]. Correspondence pairs are then determined using the nearest neighbor search (NNS) algorithm [13]. With the recent advances in deep learning, data-driven approaches [14]–[20] have demonstrated compelling results.

Fig. 1. **Comparison between DFM and our proposed GCM on the Hpatches dataset.** Our approach produces more correspondences in the regions with low texture or repetitive patterns.

Conventional hand-crafted approaches generally leverage a sequential pipeline for keypoint detection, description, and matching [21], [22]. Their overall performance is often determined by the weakest component within this pipeline, akin to the "barrel effect" principle. Moreover, errors in the earlier stages can accumulate and propagate to the later stages, making it tricky to improve the overall performance [23]. While data-driven approaches have significantly outperformed hand-crafted methods, detector-based methods [14]–[17] may still struggle in texture-less regions, and detector-free approaches could face information loss due to manually selected scales [18]–[20]. Additionally, most data-driven approaches require a large amount of well-annotated data for model training, often resulting in unsatisfactory performance when applied to new scenarios [24].

To address these limitations, deep feature matching (DFM) [25], a plug-and-play approach built upon a hierarchical matching refinement paradigm is proposed. DFM utilizes a VGG [26] model pre-trained on the ImageNet [27] database to extract multi-scale features, with no need for additional training with well-annotated data. Furthermore, DFM leverages a coarse-to-fine strategy in conjunction with the NNS

This research was supported by the National Natural Science Foundation of China under Grants 62233013, 62206046, and U2033218, the Shanghai Sailing Program under Grant 21YF1401300, the Science and Technology Commission of Shanghai Municipal under Grant 22511104500, the Fundamental Research Funds for the Central Universities, and Xiaomi Young Talents Program. (*Y. Han and Z. Long contributed equally to this work*) (✉ *Corresponding authors: Y. Zhang and R. Fan*).

Y. Han, Y. Zhang, and Z. Fang are with the School of Computer Science and Technology, Donghua University, Shanghai 201620, P. R. China (e-mails: 2232816@mail.dhu.edu.cn, {ytzhang, zjfang}@dhu.edu.cn).

J. Wu is with the Department of Electronics and Computer Engineering, the Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong SAR, P. R. China (e-mail: jin_wu_uestc@hotmail.com).

Z. Long and R. Fan are with the Machine Intelligence & Autonomous Systems (MIAS) Group, the College of Electronics & Information Engineering, Shanghai Research Institute for Intelligent Autonomous Systems, the State Key Laboratory of Intelligent Autonomous Systems, and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, P. R. China (e-mails: {2053301, rfan}@tongji.edu.cn).

algorithm to perform hierarchical feature matching from the deepest layers to the shallowest ones. DFM significantly improves accuracy and robustness compared to conventional hand-crafted methods, outperforming even some approaches trained with correspondences [24].

Nevertheless, DFM has three significant limitations. The most fatal drawback is its high demand on the backbone network, limiting compatibility to those capable of providing feature maps with the same size as the input image, such as VGG. Another drawback is the lack of extensive experimental evaluation of DFM in various computer vision and robotics tasks, except for image classification. This raises questions about its performance when using different backbone networks, especially those trained for dense correspondence matching, such as [28]–[30]. Finally, the hierarchical refinement strategy employed in DFM has the potential to propagate errors from deeper layers to shallower ones, resulting in lower density and quality of correspondence matching, as shown in Fig. 1.

To address the challenges mentioned above, this paper introduces **G**eneralized **C**orrespondence **M**atching (**GCM**) based on flexible hierarchical refinement and patch descriptor distillation. First, we omit the pre-defined threshold used in the hierarchical refinement process of DFM by leveraging a more flexible NNS strategy, thereby preventing the exclusion of repetitive yet valid matches in early stages. Furthermore, we expand the applicability of DFM to accommodate various types of backbones, pre-trained across diverse computer vision tasks, including image classification [26], [31]–[35], semantic segmentation [36]–[39], and stereo matching [28]–[30]. This is accomplished by incorporating a patch descriptor to function as the highest-resolution feature maps (with the same resolution as the input image) in DFM. Additionally, we propose a novel strategy for patch descriptor distillation, which further enhances the overall efficiency of correspondence matching. What surprises us is that several backbones demonstrate improved performance when the patch descriptor is distilled. Extensive experiments conducted on the HPatches dataset [40] demonstrate the superior mean matching accuracy (MMA) achieved by GCM. Moreover, as illustrated in Fig. 1, GCM is capable of producing denser and more accurate matched correspondences, particularly in repetitive or low-texture areas when compared to DFM.

II. Related Work

A. Correspondence Matching

Early approaches, as exemplified by SIFT [11], typically employ hand-crafted local visual features, *e.g.*, gradients, angles, and blobs, to detect distinctive interest points and generate descriptors. Correspondence pairs are subsequently determined via NNS [24]. However, these algorithms have limited adaptability to diverse and complex datasets owing to their reliance on manually designed features [16]. Moreover, they tend to be highly sensitive to variations in lighting, scale, and viewpoint, which constrains their robustness in real-world scenarios [14]. In addition, they often require significant domain expertise for feature design and may struggle to generalize effectively across different tasks or domains.

Over the last half-decade, data-driven methods [15]–[17] have demonstrated superior performance compared to conventional hand-crafted approaches. Among them, detector-based algorithms have remained the dominant choice for correspondence matching. As an example, SuperPoint [14] leverages homographic adaptation in conjunction with MagicPoint [41] to enhance detector performance and generate pseudo ground-truth interest points for unlabeled images in a self-supervised fashion. Another data-driven approach, repeatable and reliable detector and descriptor (R2D2) [16], is built upon a trainable convolutional neural network (CNN). R2D2 enhances the descriptor quality by placing a strong emphasis on the reliability of feature points. This is achieved in conjunction with a trainable CNN designed for both feature description and detection, known as D2-Net [15]. Such improvements greatly advance R2D2's feature description capabilities. In contrast, DFM [25], which relies on VGG-19 model, does not offer such precision in the description. Additionally, SuperGlue [17] leverages a graph neural network (GNN) equipped with an attention mechanism and a differentiable Sinkhorn algorithm [42] to compute matches between two sets of detected features and their corresponding descriptor vectors. SuperGlue achieves notably superior performance when compared to the NNS methods.

Other approaches, such as neighborhood consensus networks (NCNet) designed for image correspondence estimation [18] forgo the feature detection phase and instead directly match points distributed across a dense grid rather than sparse locations. NCNet constructs a 4D cost volume with neighborhood consensus to enumerate all possible matches between images. In contrast, efficient neighborhood consensus networks via submanifold sparse convolutions (SparseNCNet) [19] employs a more condensed form of the correlation tensor, storing a subset and substituting the dense 4D convolution with a sparser convolution technique to improve correspondence matching efficiency. Epipolar-guided pixel-level correspondences (Patch2pix) [43] utilizes a pre-trained backbone to extract patch-level matches and employs a two-stage regressor to refine these matches to pixel-level precision. On the other hand, detector-free local feature matching with Transformers (LoFTR) [20] achieves a high level of robustness but can be slower due to the large number of processes it involves. Although efficiency can be improved by reducing the resolution of the input image, this may affect matching accuracy to some extent. In this paper, our GCM is built upon the foundation of DFM. Our primary focus is to tackle the limitations of DFM, particularly the high demand for feature maps and the necessity of a predefined threshold for hierarchical refinement.

B. Knowledge Distillation

Knowledge distillation is a commonly used technique for model compression, initially introduced for image classification. Unlike pruning and quantization techniques [44] used in model compression, knowledge distillation techniques focus

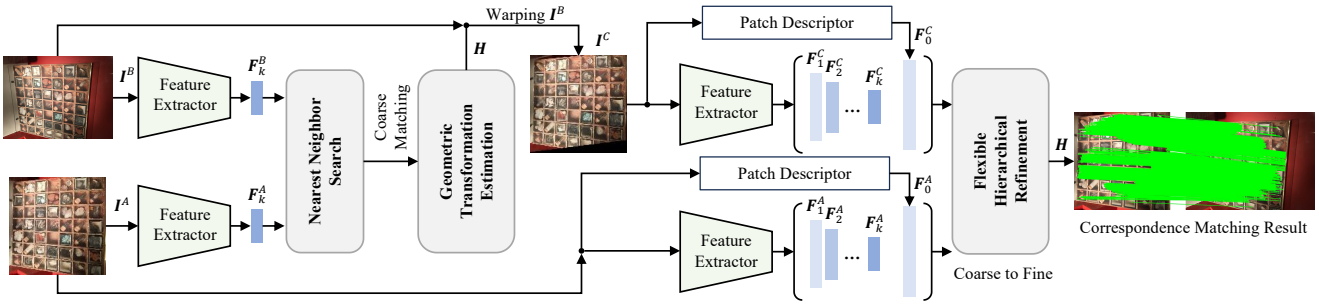

Fig. 2. **The architecture of our proposed GCM.**

on training a smaller, more lightweight model [45]. This is achieved by exploiting supervised information from larger, high-performance models, ultimately reducing both time and space complexity. Due to its success across various tasks, knowledge distillation is regarded as an effective multi-tasking approach, applicable to classification [46], semantic segmentation [47], and object detection [48], [49]. In this paper, we utilize the knowledge distillation technique to further reduce the complexity of the patch descriptor.

III. Methodology

A. Architecture Overview

The architecture of our proposed GCM is shown in Fig. 2. GCM follows a two-step correspondence matching strategy, similar to DFM [25]. Initially, we employ a pre-trained backbone network as the deep feature extractor to obtain the feature maps $\mathcal{F} = \{\boldsymbol{F}_1, ..., \boldsymbol{F}_k\}$. Here, $\boldsymbol{F}_l \in \mathbb{R}^{\frac{H}{2^l} \times \frac{W}{2^l} \times C_l}$ represents the l-th layer of feature maps, with H and W denoting the height and width of the input RGB image $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$, respectively. Next, we perform NNS on the last layers of feature maps, denoted as $\boldsymbol{F}_k^A$ and $\boldsymbol{F}_k^B$, to obtain coarse matches $\mathcal{M}_k^{A,B}$. Here, the superscripts A and B represent two images of $\boldsymbol{I}^A$ and $\boldsymbol{I}^B$. Based on these matches, we estimate the homography matrix $\boldsymbol{H}$, which is further utilized to generate a warped image $\boldsymbol{I}^C$ from $\boldsymbol{I}^B$.

In the second step, we perform feature extraction followed by a flexible hierarchical refinement (FHR) module, which will be detailed in Section III-D, to obtain the matched correspondences. In specific, we construct hierarchical feature map layers $\{\boldsymbol{F}_1^A, \boldsymbol{F}_2^A, ..., \boldsymbol{F}_k^A, \boldsymbol{F}_0^A\}$ for $\boldsymbol{I}^A$, and $\{\boldsymbol{F}_1^B, \boldsymbol{F}_2^B, ..., \boldsymbol{F}_k^B, \boldsymbol{F}_0^B\}$ for $\boldsymbol{I}^B$. Note that, $\boldsymbol{F}_0 \in \mathbb{R}^{H \times W \times C_0}$ is obtained through a patch descriptor described in Section III-C, which provides feature maps with the same size as the original images. The FHR module operates in a coarse-to-fine manner, processing from the deepest layer $\boldsymbol{F}_k$ to shallow layers down to $\boldsymbol{F}_1$, and finally $\boldsymbol{F}_0$, where we obtain the final matches between $\boldsymbol{I}^A$ and the warped image $\boldsymbol{I}^C$. Later, taking advantage of the estimated $\boldsymbol{H}$ in the first step, we can trace back to the matched correspondences between $\boldsymbol{I}^A$ and the original image of $\boldsymbol{I}^B$.

Our method generalizes the baseline DFM approach and is compatible with various types of backbone networks. These networks can be pre-trained for a diverse range of computer vision tasks, regardless of whether they can produce feature maps with the same resolution as the original images.

B. Nearest Neighbor Search

Since the dense NNS utilized in DFM is a local matching approach that necessitates a manually defined threshold and tends to discard correct matches in cases of repetitiveness, we adopt a flexible, parameter-free NNS strategy.

Given the feature maps $\boldsymbol{F}^A$ and $\boldsymbol{F}^B$ extracted from images $\boldsymbol{I}^A$ and $\boldsymbol{I}^B$, we identify potential matches by determining the nearest neighbors based on the feature distance between p^A and p^B (with $\boldsymbol{f}^A$ and $\boldsymbol{f}^B$ denoting the features at the point p^A and p^B in $\boldsymbol{F}^A$ and $\boldsymbol{F}^B$, respectively):

$$d(p^A, p^B) = 1 - \phi(\boldsymbol{f}^A, \boldsymbol{f}^B), \tag{1}$$

where $\phi(\cdot, \cdot)$ represents the cosine similarity as follows:

$$\phi(\boldsymbol{f}^A, \boldsymbol{f}^B) = \frac{\boldsymbol{f}^A \cdot \boldsymbol{f}^B}{\|\boldsymbol{f}^A\|_2 \, \|\boldsymbol{f}^B\|_2}. \tag{2}$$

For a specific point p^A within the feature map $\boldsymbol{F}^A$, it is matched to p^B if the distance to p^B is minimal. A match (p^A, p^B) is confirmed only if it is mutual, meaning that p^A and p^B are recognized as a matched pair only if p^B is also matched with p^A.

C. Patch Descriptor Distillation

To reduce time and space complexity, we develop an additional distilled patch descriptor as a more lightweight feature description network alternative. Given that descriptors of R2D2 [16] are sourced from an L2-normalized feature map, our strategy is to train the final layer of the student model's backbone using the final feature layer of the teacher model's backbone. As shown in Fig. 4, the student model reduces the number of intermediate layers and directly assimilates the final feature map from the teacher model. $\boldsymbol{X}^T \in \mathbb{R}^{H \times W \times 128}$ denotes the output of the last layer of the teacher model backbone network, and $\boldsymbol{X}^S \in \mathbb{R}^{H \times W \times 128}$ denotes the last layer output of the student model backbone network. The loss function is defined as follows:

$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} (1 - |\phi(\boldsymbol{X}_{ij}^T, \boldsymbol{X}_{ij}^S)|), \tag{3}$$

where the point set $\mathcal{P}$ contains all points in the image, represented by coordinates (i, j) for each pixel. The function ϕ is defined in Equation (2). This lightweight patch descriptor enables the faster generation of $\boldsymbol{F}_0 \in \mathbb{R}^{H \times W \times 128}$ for facilitating flexible hierarchical refinement.

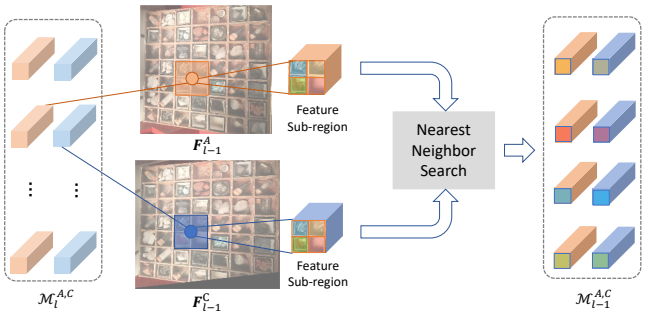

Fig. 3. **An illustration of one iteration within the hierarchical refinement module.**

D. Flexible Hierarchical Refinement

The flexible hierarchical refinement (FHR) module consists of multiple iterations, with a single iteration illustrated in Fig. 3. Given the matched features $\mathcal{M}_l^{A,C}$ in the l-th layer, where every match $(p_l^A, p_l^C) \in \mathcal{M}_l^{A,C}$ is associated with feature $(\boldsymbol{f}_l^A, \boldsymbol{f}_l^C)$, this module returns matched correspondences $\mathcal{M}_{l-1}^{A,C}$ in the $(l-1)$-th layer. During this step, it allows the matched features in $\boldsymbol{F}_l^A$ and $\boldsymbol{F}_l^C$ to identify the corresponding sub-region pairs in $\boldsymbol{F}_{l-1}^A$ and $\boldsymbol{F}_{l-1}^C$. In specific, every point in $\boldsymbol{F}_l$ is mapped to $\boldsymbol{F}_{l-1}$ and corresponds to a 2×2 patch on $\boldsymbol{F}_{l-1}$. By applying NNS to these sub-regions, matched correspondences $\mathcal{M}_{l-1}^{A,C}$ between $\boldsymbol{F}_{l-1}^A$ and $\boldsymbol{F}_{l-1}^C$ can be obtained.

There are $(k+1)$ iterations in total, consistent with the number of hierarchical feature layers from $\boldsymbol{F}_k$ to $\boldsymbol{F}_0$. As for the initial input of FHR, we employ the NNS on the k-th layer's $\boldsymbol{F}_k$ from the two images to obtain an initial correspondence. Subsequently, we map these points to $\boldsymbol{F}_{k-1}$ and establish paired sub-regions between $\boldsymbol{I}^A$ and $\boldsymbol{I}^C$. Using this correspondence, we employ the NNS to establish refined correspondences within each group of patches. This process is iteratively applied until the first layer of $\boldsymbol{F}_1$ is reached, resulting in correspondences at half size of the input images. Finally, we project the matched points onto the descriptor layer $\boldsymbol{F}_0$ and perform correspondence matching at this layer using NNS between every paired patch associated with $\boldsymbol{I}^A$ and $\boldsymbol{I}^C$. Due to potential accumulated errors from the previous stages, we introduce a ratio test as in [11], [15] at the final iteration to filter out inferior matches.

It is worth noting that this hierarchical refinement paradigm is adaptable to a wide range of backbone networks.

IV. EXPERIMENTS

A. Datasets and Evaluation Protocol

Here is a summary of the datasets and evaluation setups used in our experiments:

① HPatches: We conduct experiments on the HPatches dataset, utilizing mean matching accuracy (MMA) and homography estimation accuracy as metrics. For both image matching and homography estimation, we follow the evaluation setup detailed in [25].

② MegaDepth: We evaluate the outdoor pose estimation accuracy of our proposed method on the MegaDepth dataset [50]. We quantify the pose error by computing the area under the receiver operating characteristic (AUC), following the setup detailed in [51] and [20]. Additionally, we also compute match precision following the setup in [17].

③ ScanNet: We evaluate the indoor pose estimation accuracy on the ScanNet dataset [52]. Please refer to [17] for detailed setups.

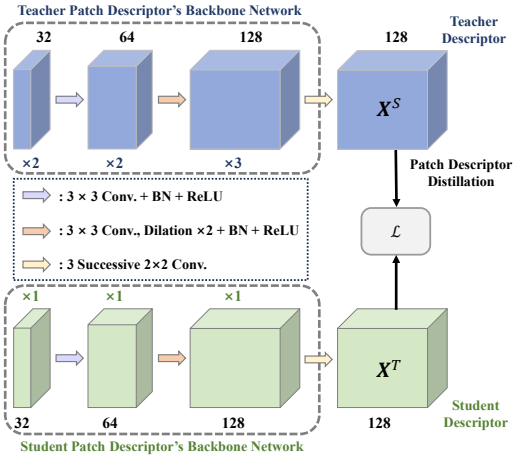

Fig. 4. **Our proposed patch descriptor distillation strategy**. We employ the R2D2 backbone as the teacher descriptor to generate a more lightweight student descriptor. BN refers to batch normalization.

B. Implementation Details

In our experiments, we employ ResNet18, pre-trained on the ImageNet dataset, as the default deep feature extractor. For the feature description, we use the patch descriptor from the distilled R2D2. We only incorporate a ratio test at the descriptor layer to correct errors from previous stages. Similar to the setup used in [25], two ratio tests (thresholds: 0.95 and 0.60) are applied. Our model is trained on the same datasets used to train R2D2. During the training process, we utilize the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 2. The training process converges after 10 epochs on an NVIDIA GTX1650 GPU. The student model contains 292K parameters, compared to the teacher model, which has 486K parameters. This reduction in parameters demonstrates the efficiency gains achieved through our patch descriptor distillation strategy.

We incorporate a patch descriptor that allows the proposed method to be compatible with the most popular backbone networks. It is worth noting that our method is not directly compatible with the Swin Transformer architecture [39], which initially applies a $4\times$ downsampling, resulting in a feature map with a maximum resolution equivalent to only a quarter of the input image. To ensure compatibility with Swin Transformer, we up-sample the points directly to match the required resolution.

C. Image Matching Performance Evaluation

Table I demonstrates that our proposed GCM outperforms all other existing algorithms in terms of overall and viewpoint MMA on the HPatches dataset. It achieves results comparable to LoFTR and better results than R2D2.

TABLE I

MMA COMPARISON ON THE HPATCHES DATASET. "r" DENOTES RATIO TEST THRESHOLD.

Category	Method	Overall			Illumination			Viewpoint			Matches
		@1px	@3px	@5px	@1px	@3px	@5px	@1px	@3px	@5px	
Hand-Crafted	SIFT+NNS [11]	0.35	0.50	0.54	0.37	0.49	0.52	0.33	0.52	0.55	0.4K
Fully Supervised	R2D2+NNS [16]	0.33	0.76	0.84	0.38	0.81	0.90	0.29	0.71	0.79	1.6k
	LoFTR [20]	0.63	0.91	0.93	0.68	**0.95**	0.96	0.59	0.86	0.91	2.6K
Plug-and-Play	DFM (r=0.90) [25]	0.51	0.85	0.93	0.63	0.91	0.97	0.42	0.81	0.89	7.0K
	DFM (r=0.60) [25]	0.61	0.88	0.94	**0.77**	0.93	**0.98**	0.47	0.84	0.90	1.0K
	Ours (r=0.95)	0.53	0.85	0.92	0.56	0.87	0.94	0.50	0.84	0.90	**14.4K**
	Ours (r=0.60)	**0.68**	**0.92**	**0.95**	0.74	0.93	0.97	**0.63**	**0.91**	**0.94**	3.1K

TABLE II

MMA COMPARISON ON THE HPATCHES DATASET AMONG BACKBONE NETWORKS PRE-TRAINED FOR A VARIETY OF COMPUTER VISION TASKS. "r" DENOTES RATIO TEST THRESHOLD. "$*$" DENOTES THE BACKBONE USED IN R2D2 [16].

Task	Method	Overall			Illumination			Viewpoint			Matches
		@1px	@3px	@5px	@1px	@3px	@5px	@1px	@3px	@5px	
Image Classification	ResNet18 [31]	0.53	0.85	0.92	0.56	0.87	0.94	0.50	0.84	0.90	14.4K
	ResNet18 (r=0.6) [31]	0.68	**0.92**	**0.95**	0.74	0.93	0.97	0.63	0.91	**0.94**	3.1K
	ResNet18* (r=0.6) [16]	**0.69**	0.91	0.95	0.74	0.91	0.96	**0.64**	**0.92**	0.94	2.0K
	VGG19 [26]	0.56	0.86	0.92	0.60	0.89	0.95	0.51	0.84	0.89	13.3K
	ResNet50 [31]	0.52	0.85	0.92	0.57	0.87	0.94	0.47	0.83	0.90	13.7K
	ResNeXt50 [32]	0.50	0.85	0.92	0.55	0.88	0.94	0.46	0.83	0.90	12.4K
	MobileV3 [33]	0.50	0.82	0.89	0.55	0.84	0.90	0.45	0.81	0.88	12.8K
	EfficientNetV2 [34]	0.50	0.84	0.92	0.56	0.87	0.93	0.44	0.82	0.90	13.9K
Semantic Segmentation	FCN [38]	0.46	0.78	0.85	0.49	0.78	0.85	0.44	0.77	0.85	10.9K
	DeepLabV3 [37]	0.37	0.62	0.70	0.47	0.76	0.84	0.26	0.50	0.57	4.4K
	Swin Tranformer [39]	0.53	0.80	0.88	0.60	0.81	0.89	0.46	0.79	0.87	4.7K
Stereo Matching	CREStereo [29]	0.51	0.73	0.75	0.57	0.78	0.81	0.46	0.68	0.70	**23.3K**
	PSMNet [30]	0.31	0.51	0.56	0.38	0.58	0.63	0.24	0.44	0.49	4.8K
	DeepPruner [28]	0.31	0.50	0.55	0.37	0.57	0.62	0.24	0.44	0.48	5.2K
-	Average Pooling	0.60	0.85	0.88	0.66	0.88	0.92	0.53	0.81	0.84	17.4K

TABLE III

HOMOGRAPHY ESTIMATION RESULTS ON THE HPATCHES DATASET.

Method	Overall			Illumination			Viewpoint		
	@1px	@3px	@5px	@1px	@3px	@5px	@1px	@3px	@5px
SIFT [11]+NNS	0.42	0.74	0.85	0.54	0.86	0.93	0.30	0.54	0.71
SuperPoint [14]+NNS	0.46	0.78	0.85	0.57	0.92	0.97	0.35	0.65	0.74
D2Net [15]+NNS	0.38	0.71	0.82	0.66	0.95	0.98	0.12	0.49	0.67
R2D2 [16]+NNS	0.47	0.77	0.82	0.63	0.93	0.98	0.32	0.64	0.70
SuperPoint [14]+SuperGlue [17]	0.51	0.82	0.89	0.60	0.92	0.98	0.42	0.71	0.81
SuperPoint [14]+CAPS [53]	0.49	0.79	0.86	0.62	0.93	0.98	0.36	0.65	0.75
SuperPoint [14]+ClusterGNN [54]	0.52	0.84	0.90	0.61	0.93	0.98	0.44	0.74	0.81
SIFT+CAPS [53]	0.36	0.77	0.85	0.48	0.89	0.95	0.26	0.65	0.76
Patch2Pix [43]	0.50	0.79	0.87	0.71	0.95	0.98	0.30	0.64	0.76
MatchFormer	0.55	0.81	0.87	**0.75**	0.95	0.98	0.37	0.68	0.78
LoFTR	**0.63**	**0.91**	**0.93**	0.68	0.95	0.96	**0.59**	**0.86**	**0.91**
DFM [25]	0.41	0.74	0.85	0.63	0.91	0.97	0.21	0.59	0.74
Ours	0.55	0.84	0.90	0.73	**0.95**	**0.98**	0.38	0.73	0.81

Furthermore, the results presented in Table II confirm the compatibility of our algorithm with various backbone networks that are pre-trained for a wide range of computer vision tasks, including image classification, semantic segmentation, and stereo matching. Additionally, we propose an alternative approach that does not rely on a visual backbone network, by substituting the feature maps in the hierarchical refinement strategy with descriptors average pooling to different scales. This method is represented as Average Pooling in Table. II. The unsatisfactory results achieved with stereo-matching backbone networks are somewhat unexpected, and it is possible that these networks, primarily designed for 1D dense search, may not be well-suited for solving 2D search problems that are more relevant to correspondence matching tasks. This analysis provides valuable insights into the performance variation observed with different types of backbone networks and highlights the need for specialized feature extraction and matching strategies in correspondence matching applications. Considering the speed advantage of the ResNet18 model, we choose ResNet18 as the backbone for experiments. It is also worth noting that the difference in performance between using the original R2D2 backbone and a lightweight distilled R2D2 backbone is minimal. This demonstrates the effectiveness of the distillation strategy in reducing computational complexity without sacrificing performance.

TABLE IV

POSE ESTIMATION PERFORMANCE EVALUATION ON THE MEGADEPTH DATASET. THESE RESULTS DENOTE THE PERCENTAGES OF CORRECTLY ESTIMATED POSES WITH POSE ERRORS BELOW 5/10/20°, RESPECTIVELY.

Method	Pose Estimation AUC			Precision
	@5°	@10°	@20°	
SIFT+NNS	16.70	28.15	41.84	35.53
D2-Net	4.44	8.27	14.17	48.58
SuperPoint+NNS	29.01	44.74	58.92	57.43
R2D2+NNS	41.15	58.88	72.84	81.51
SuperPoint+SuperGlue	46.10	63.82	77.68	**99.66**
Patch2Pix	39.70	55.06	67.77	80.17
LoFTR	**52.42**	**69.26**	**81.41**	96.78
DFM	25.91	41.70	56.19	91.92
Ours	34.61	52.24	67.10	87.41

TABLE V

POSE ESTIMATION PERFORMANCE EVALUATION ON THE SCANNET DATASET. "†" DENOTES THE MODELS TRAINED ON THE OUTDOOR DATASETS AND EVALUATED ON THE SCANNET DATASET (INDOOR).

Method	Pose Estimation AUC			Precision
	@5°	@10°	@20°	
SIFT+NNS	5.83	13.06	22.47	40.30
SuperPoint+NNS	9.43	21.53	36.40	50.40
Patch2Pix†	9.57	21.22	34.56	50.59
R2D2†	6.82	16.37	28.02	46.51
SuperPoint+SuperGlue	16.24	34.13	52.88	**84.33**
SuperPoint+SuperGlue†	15.68	32.66	49.87	80.45
LoFTR	**20.27**	**39.63**	**57.47**	83.92
LoFTR†	17.57	34.46	51.88	70.16
DFM	3.73	9.76	18.94	74.77
Ours†	11.03	23.73	38.93	67.63

D. Homography Estimation Performance Evaluation

In Table III, the accuracy metrics reported for overall, illumination, and viewpoint matching at various thresholds provide a comprehensive performance evaluation of homography estimation for both our method and the compared algorithms. Notably, while GCM achieves the second-best performance across all these categories, it outperforms the majority of data-driven approaches that are trained using fully supervised correspondence ground truth.

E. Pose Estimation Performance Evaluation

Tables IV and V provide the performance evaluations for pose estimation in outdoor and indoor environments, respectively. The MegaDepth dataset presents a significant challenge due to the necessity of matching under extreme viewpoint changes and addressing issues related to high texture repetition. Despite utilizing a non-specialized network for deep feature extraction, our method consistently demonstrates impressive performance on demanding test data. While it may not reach the state-of-the-art performance level achieved by fully supervised methods in pose estimation, our approach represents a notable improvement over the baseline plug-and-play algorithm DFM. Additionally, the experimental results on the ScanNet dataset demonstrate that our proposed GCM achieves performance levels comparable to fully supervised methods and outperforms the baseline algorithm DFM across most threshold values.

TABLE VI

ABLATION STUDY OF EACH COMPONENT ON HPATCHES DATASET.

FHR	Descriptor	Distillation	Homography Estimation Accuracy		
			@1px	@3px	@5px
			0.37	0.68	0.80
	✓		0.41	0.72	0.82
	✓	✓	0.42	0.72	0.83
✓			0.38	0.71	0.83
✓	✓		**0.45**	0.75	0.84
✓	✓	✓	0.44	**0.76**	**0.85**

F. Abaltion Studies

We conduct an ablation study to determine the individual and combined effects of FHR, patch descriptor, and descriptor distillation on the homography estimation accuracy across the overall Hpatches dataset, with the setup as referenced in Sec. IV-C. The results are presented in Table VI. VGG19 is utilized as the backbone network and the baseline model (the first row) is designated as DFM, with random sample consensus (RANSAC) for homography matrix estimation to ensure experimental consistency. Our key insights:

- Patch descriptors play a crucial role in improving accuracy, in the configurations where it is employed (the second and fifth rows). These results underscore its essential contribution to baseline performance enhancement and integration with FHR for further improvements.
- Descriptor distillation significantly refines patch descriptor quality, leading to more accurate homography estimation. This is evident in the configurations incorporating descriptor distillation (the third and sixth rows).
- FHR contributes to a notable increase in accuracy (from the fourth to the sixth rows). When applied in combination with patch descriptor and distillation, FHR significantly elevates the model's overall performance.

V. CONCLUSION

In summary, this paper introduced three significant technical contributions to address the limitations present in DFM: (1) a flexible hierarchical refinement strategy that eliminates the need for a pre-defined threshold, initially utilized in the hierarchical refinement process of DFM; (2) the incorporation of a patch descriptor that extends the applicability of DFM to accommodate a wide range of backbone networks, pre-trained across diverse robot perception tasks, such as semantic segmentation and stereo matching; (3) a novel patch descriptor distillation strategy, which further reduces the computational complexity of correspondence matching and enhances the practical applicability of our method in real-world robotics applications. Through performance evaluations including image matching, homography estimation, and pose estimation, the effectiveness of our proposed algorithm is validated. Particularly noteworthy is that our method achieves state-of-the-art overall mean matching accuracy, outperforming both conventional handcrafted approaches and data-driven methods trained via fully supervised learning. We are confident that our method can be readily integrated into a variety of downstream real-world robotics applications.

REFERENCES

[1] M. U. M. Bhutta *et al.*, "Loop-box: Multiagent direct slam triggered by single loop closure for large-scale mapping," *IEEE Transactions on Cybernetics*, vol. 52, no. 6, pp. 5088–5097, 2020.

[2] R. Mur-Artal *et al.*, "ORB-SLAM: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[3] M. U. M. Bhutta and M. Liu, "Pcr-pro: 3d sparse and different scale point clouds registration and robust estimation of information matrix for pose graph slam," in *2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 354–359, IEEE, 2018.

[4] R. Fan *et al.*, "Road surface 3D reconstruction based on dense subpixel disparity map estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.

[5] J. L. Schonberger *et al.*, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] R. Fan *et al.*, "Rethinking road surface 3-D reconstruction and pothole detection: From perspective transformation to disparity map segmentation," *IEEE Transactions on Cybernetics*, vol. 52, no. 7, pp. 5799–5808, 2022.

[7] Z. Wu *et al.*, "S^3M-Net: Joint learning of semantic segmentation and stereo matching for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2024. DOI: 10.1109/TIV.2024.3357056.

[8] H. Zhao *et al.*, "Dive deeper into rectifying homography for stereo camera online self-calibration," in *2024 International Conference on Robotics and Automation (ICRA)*, 2024. in press.

[9] H. Wang *et al.*, "PVStereo: Pyramid voting module for end-to-end self-supervised stereo matching," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4353–4360, 2021.

[10] Z. Wu *et al.*, "SG-RoadSeg: End-to-end collision-free space detection sharing encoder representations jointly learned via unsupervised deep stereo," in *2024 International Conference on Robotics and Automation (ICRA)*, 2024. in press.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[12] H. Bay *et al.*, "SURF: Speeded up robust features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 404–417, 2006.

[13] M. Muja *et al.*, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2227–2240, 2014.

[14] D. DeTone *et al.*, "SuperPoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 224–236, 2018.

[15] M. Dusmanu *et al.*, "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 (CVPR)*, pp. 8092–8101, 2019.

[16] J. Revaud *et al.*, "R2D2: Reliable and repeatable detector and descriptor," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[17] P.-E. Sarlin *et al.*, "SuperGlue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4938–4947, 2020.

[18] I. Rocco *et al.*, "NCNet: Neighbourhood consensus networks for estimating image correspondences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1020–1034, 2020.

[19] I. Rocco *et al.*, "Efficient neighbourhood consensus networks via submanifold sparse convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 605–621, 2020.

[20] J. Sun *et al.*, "LoFTR: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8922–8931, 2021.

[21] M. U. M. Bhutta, *Towards a Swift Multiagent Slam System for Large-Scale Robotics Applications*. Hong Kong University of Science and Technology (Hong Kong), 2021.

[22] J. Ma *et al.*, "Image matching from handcrafted to deep features: A survey," *International Journal of Computer Vision*, vol. 129, pp. 23–79, 2021.

[23] U. Efe *et al.*, "Effect of parameter optimization on classical and learning-based image matching methods," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2506–2513, 2021.

[24] C. Zhou *et al.*, "E3CM: Epipolar-constrained cascade correspondence matching," *Neurocomputing*, vol. 559, p. 126788, 2023.

[25] U. Efe *et al.*, "DFM: A performance baseline for deep feature matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4284–4293, 2021.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*, Computational and Biological Learning Society, 2015.

[27] A. Krizhevsky *et al.*, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, Curran Associates, Inc., 2012.

[28] S. Duggal *et al.*, "DeepPruner: Learning efficient stereo matching via differentiable patchmatch," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4384–4393, 2019.

[29] J. Li *et al.*, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16263–16272, 2022.

[30] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5410–5418, 2018.

[31] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[32] S. Xie *et al.*, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1492–1500, 2017.

[33] A. Howard *et al.*, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.

[34] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *International Conference on Machine Learning (ICML)*, pp. 10096–10106, PMLR, 2021.

[35] N. Ma *et al.*, "ShuffleNet V2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, 2018.

[36] O. Ronneberger *et al.*, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, Springer, 2015.

[37] L.-C. Chen *et al.*, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[38] J. Long *et al.*, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[39] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

[40] V. Balntas *et al.*, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5173–5182, 2017.

[41] D. DeTone *et al.*, "Toward geometric deep SLAM," *arXiv preprint arXiv:1707.07410*, 2017.

[42] R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *The Annals of Mathematical Statistics*, vol. 35, no. 2, pp. 876–879, 1964.

[43] Q. Zhou *et al.*, "Patch2Pix: Epipolar-guided pixel-level correspondences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4669–4678, 2021.

[44] J.-H. Park *et al.*, "Quantized sparse training: A unified trainable framework for joint pruning and quantization in dnns," *ACM Transactions on Embedded Computing Systems*, vol. 21, no. 5, pp. 1–22, 2022.

[45] B. Zhao *et al.*, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11953–11962, 2022.

[46] G. Hinton *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[47] T. He *et al.*, "Knowledge adaptation for efficient semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 578–587, 2019.

[48] Q. Li *et al.*, "Mimicking very efficient network for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6356–6364, 2017.
[49] G. Chen *et al.*, "Learning efficient object detection models with knowledge distillation," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 742–751, 2017.
[50] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2041–2050, 2018.
[51] M. Tyszkiewicz *et al.*, "Disk: Learning local features with policy gradient," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 14254–14265, 2020.
[52] A. Dai *et al.*, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5828–5839, 2017.
[53] Q. Wang *et al.*, "Learning feature descriptors using camera pose supervision," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 757–774, 2020.
[54] Y. Shi *et al.*, "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 (CVPR)*, pp. 12517–12526, 2022.