D^{2}Feat: Dual Distillation for Semantic and Geometric Local Feature Learning

Yiyuan Wang[1,2], Xinze Li[2], Puzhen Wu[3], Junkai Zhang[4], Weifeng Su[2,5], Wentao Cheng[2,†]
[1]Department of Computer Science and Technology, Hong Kong Baptist University, Hong Kong, China
[2]Department of Computer Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, China
[3]Department of Orthopaedics and Traumatology, The University of Hong Kong, Hong Kong, China
[4]School of Law, Tsinghua University, Beijing, China
[5]Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science
{wangyiyuan, wfsu, wentaocheng}@bnbu.edu.cn, t330026083@mail.uic.edu.cn,
puzhenwu8@connect.hku.hk, jj-zhang25@mails.tsinghua.edu.cn

***Abstract*—Convolution-based feature descriptors remain widely adopted in many downstream tasks for their resource efficiency and ease of deployment. Recently, transformer-based feature matchers and their underlying self-supervised visual encoders have achieved remarkable progress in feature matching, redefining the frontier of visual correspondence learning. Motivated by these advances, we propose D^{2}Feat, a dual-distillation framework that injects the strengths of both self-supervised visual encoders and transformer-based matchers into a compact convolutional backbone. Specifically, we distill semantically rich but spatially coarse representations from self-supervised encoders at early stages, and later distill geometrically precise features from transformer-based feature matchers. Furthermore, we introduce a lightweight Fine-grained Perceiver Module (FPM) that integrates the distilled features in a bypass manner with minimal computational overhead. Our dual-stage distillation strategy ensures stable convergence and consistent optimization behavior across training stages. Extensive experiments on MegaDepth, ScanNet, and HPatches demonstrate that D^{2}Feat achieves outstanding performance. Code is available at https://github.com/YiyuanWang-001/D2Feat.**

***Index Terms*—image matching, knowledge distillation, Vision Foundation Model.**

I. Introduction

Local image feature extraction is a key component in many computer vision tasks, such as 3D reconstruction [1]–[3], visual localization [4]–[6], and augmented reality [7]–[9]. These features serve as the basis for accurate correspondence estimation, enabling robust geometric reasoning across multiple views. Due to their efficiency and light computational footprint, CNN-based models [10]–[13] remain a practical choice for local feature extraction, particularly on mobile and robotic platforms where memory and computational resources are limited. Although CNN-based models surpass traditional hand-crafted descriptors [14], [15] in accuracy and robustness, their local inductive bias limits the modeling of long-range dependencies and hampers generalization across diverse scenes. These limitations highlight the need for more discriminative yet efficient paradigms for robust local feature learning.

† Corresponding author

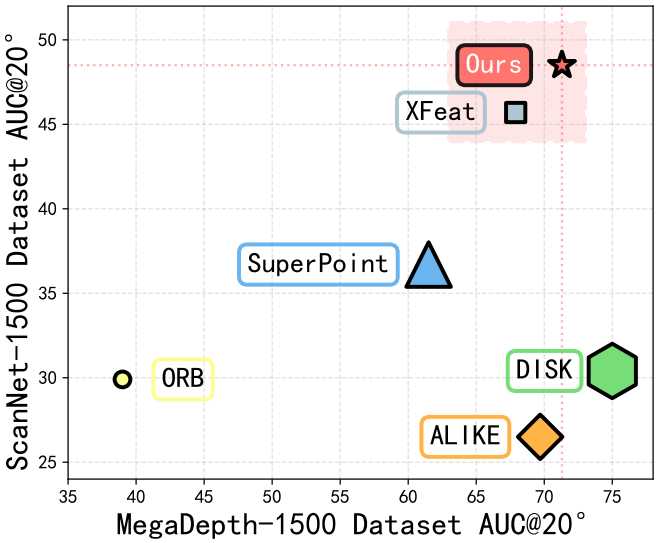

Fig. 1: Relative pose estimation results on both the MegaDepth-1500 [16] and untrained ScanNet-1500 [17] datasets. The smaller the marker, the higher the efficiency.

In recent years, transformer-based or attention-driven feature matching methods [18]–[22] have demonstrated remarkable generalization ability and strong robustness under severe appearance or viewpoint changes. Their success can be attributed both to the powerful representation capacity of large-scale visual encoders [23], [24] and to the geometric binding learned through multiple layers of self- and cross-attention that explicitly model relationships between image pairs. However, these transformer-based approaches are notoriously expensive: they rely on heavy backbones and dense attention operations that result in large parameter counts and computational overhead. Although several efforts have explored accelerating transformer-based matchers [25], [26], the underlying architectures remain unchanged, limiting their portability in resource-constrained environments. This trade-off between representational power and efficiency remains unresolved. One may

then ask: can the light of transformer-based methods extend to the CNN world, bringing their generalization and geometric reasoning capabilities to lightweight architectures?

To realize this idea, we design D^{2}Feat to mimic the learning dynamics of transformer-based feature matchers, where semantic understanding precedes geometric reasoning. Specifically, the framework adopts a dual-stage distillation paradigm. In the early stage, the student CNN learns from large visual encoders such as DINOv3 [24], acquiring broad and domain-general semantic priors. In the later stage, it distills fine-grained geometric cues from transformer-based feature matchers, whose attention mechanisms capture precise correspondences between images. Through this two-phase distillation, the CNN inherits semantic generalization and geometric reasoning capabilities from its teachers.

Heterogeneous distillation [27], [28] between transformers and CNNs, however, is unstable due to the large architectural and capacity gap. To address this, we treat distillation as an auxiliary bypass rather than a direct replacement of the backbone representation. The distilled features are injected into the network through a lightweight fusion pathway, allowing the main CNN to retain most of its representational capacity. To further integrate this bypass branch, we introduce the local Fine-grained Perceiver Module (FPM), which aggregates and refines multi-scale geometric features from both the backbone and the bypass stream. FPM enhances local geometric awareness and spatial consistency within the CNN representation.

In summary, this work has the following contributions:

- We propose D^{2}Feat, a dual distillation framework, which distills semantic and geometric understanding from teacher models in a staged manner.
- We introduce a Fine-grained Perceiver Module (FPM) to fuse geometric cues from transformer-based matcher.
- Extensive experiments on relative pose estimation, homography estimation, and visual localization demonstrate the superiority and robustness of D^{2}Feat.

II. METHOD

This section first presents our feature extraction network, followed by an explanation of how self-supervised encoders facilitate semantic generalization for potential matching regions. We then introduce the Fine-grained Perceiver Module to enhance geometric perception capability. Finally, we outline the training framework and loss function design.

A. Network Overview

As illustrated in Fig.2, given an input image $\mathbf{I} \in \mathbb{R}^{W\times H\times 3}$, D^{2}Feat employs a dual-branch architecture that balances semantic understanding with geometric precision. Following XFeat [10], we adopt blocks comprising 2D convolutions, ReLU activation, and BatchNorm for feature extraction. The first branch uses a feature pyramid network (FPN) to merge multi-scale intermediate representations via bilinear upsampling to $H/8 \times W/8 \times 64$, followed by element-wise summation. In parallel, the second branch serves as the student network, acquiring broad semantic knowledge from the frozen vision foundation model (VFM). Both branches are merged to produce a feature map of size $H/8 \times W/8 \times 64$. To enhance fine-grained geometric understanding, we leverage multi-scale geometric cues from the transformer-based matcher during training. The generated feature map is processed through specialized heads: a descriptor head produces initial dense feature $\mathbf{F}_{des} \in \mathbb{R}^{H/8\times W/8\times 64}$, a convolutional module predicts reliability map $\mathbf{R} \in \mathbb{R}^{H/8\times W/8}$ from updated $\mathbf{F}_{des}$ modeling match probability for each local feature, and a dedicated branch predicts keypoint map $\mathbf{K} \in \mathbb{R}^{H/8\times W/8\times (64+1)}$ encoding keypoint distribution. Both teacher models are discarded during inference. More details are provided in the appendix.

B. Semantic Distillation

DINOv3 [24] is a vision foundation model pretrained via self-supervised learning, enabling exceptional generalization across diverse image domains. First, we employ knowledge distillation with DINOv3 as the teacher, enabling the student model to acquire comprehensive semantic knowledge and achieve strong generalization capability. Specifically, DINOv3 generates features $\mathbf{F}_{dino} \in \mathbb{R}^{P\times 768}$ (P is the number of DINOv3 patches), which are reshaped into the pseudo semantic label feature $\mathbf{F}_t \in \mathbb{R}^{H/8\times W/8\times 64}$ through convolution and bilinear interpolation. Meanwhile, the second branch of D^{2}Feat serves as the student network, with its output adopted as the student features $\mathbf{F}_s \in \mathbb{R}^{H/8\times W/8\times 64}$. To enable pixel-level alignment between the feature $\mathbf{F}_t$ and feature $\mathbf{F}_s$ in both magnitude and spatial directions, we introduce Mean Squared Error (MSE) loss to quantify the differences:

$$\mathcal{L}_{\text{MSE}-\text{SD}} = \frac{1}{N} \left\| \mathbf{F}_t - \mathbf{F}_s \right\|_2^2 \tag{1}$$

where N denotes the total number of elements in the feature maps. The MSE loss provides a stable and direct optimization signal for initial alignment. Meanwhile, to mitigate its sensitivity to scale and potential gradient issues, we incorporate the Kullback-Leibler (KL) divergence to enforce consistency in the feature distributions.

$$p_t = \text{Softmax}\left(\hat{\mathbf{F}}_t/\tau\right), \quad q_s = \text{Softmax}\left(\hat{\mathbf{F}}_s/\tau\right) \tag{2}$$

where $\hat{\mathbf{F}}_t$ and $\hat{\mathbf{F}}_s$ are the L2-normalized teacher and student feature maps respectively, and τ is temperature scaling factor. The resulting teacher probability p_t and student probability q_s are then used to compute the $\mathcal{L}_{\text{KL-DINO}}$, with the optimization objective of minimizing the divergence between the teacher and student distributions.

$$\mathcal{L}_{\text{KL-SD}} = \tau^2 \cdot \text{KL}(p_t \| q_s) \tag{3}$$

where KL divergence is averaged over all spatial locations. The final knowledge distillation loss is expressed as follows:

$$\mathcal{L}_{\text{KD-SD}} = \alpha \cdot \mathcal{L}_{\text{MSE-SD}} + \beta \cdot \mathcal{L}_{\text{KL-SD}} \tag{4}$$

where α, β are the weights. By integrating the hard constraints from $\mathcal{L}_{\text{MSE-SD}}$ with the soft guidance provided by $\mathcal{L}_{\text{KL-SD}}$, the student model effectively inherits comprehensive semantic

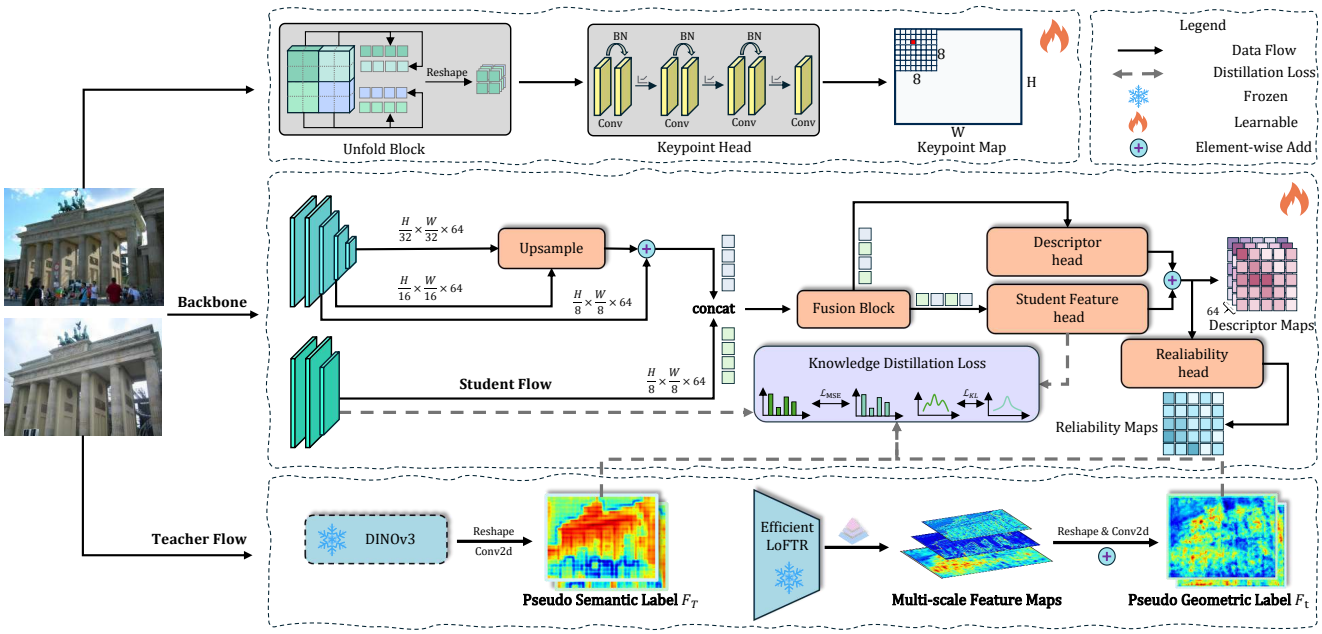

Fig. 2: Overview of the proposed D^{2}Feat. The model extracts a dense descriptor map $\mathbf{F}_{des}$, a reliability heatmap $\mathbf{R}$, and a keypoint heatmap $\mathbf{K}$. During the training phase, we obtain semantically rich pseudo-label features from the frozen VFM DINOv3 to assist the student network in learning a broad range of visual knowledge. Additionally, we leverage multi-scale geometric cues captured from the transformer-based matcher Efficient-LoFTR to enhance fine-grained perceptual abilities. It is important to note that the pre-trained teacher model will be discarded during the inference phase.

knowledge from the teacher, while adapting to domain-specific noise and distribution shifts.

C. Geometric Distillation

While DINOv3 provides broad semantic knowledge, its features are coarse and lack fine-grained detail [20] [29]. To enhance fine-grained perception, we propose a local Fine-grained Perceiver Module (FPM) to incorporate knowledge from Efficient-LoFTR [26], a transformer-based matcher that generates multi-scale features at {2×, 4×, 8×} downsampling with {64, 128, 256} channels. These features contain rich geometric information, improving robustness to illumination, viewpoint, and scale variations. The multi-scale feature maps are fused via element-wise summation to produce the pseudo geometric label feature $\mathbf{F}'_t \in \mathbb{R}^{H/8\times W/8\times 64}$ as follows:

$$\mathbf{F}'_t = \text{Conv2d}\left(\mathbf{F}'_1 + \mathbf{F}'_2 + \mathbf{F}'_3\right) \tag{5}$$

where $\mathbf{F}'_1$, $\mathbf{F}'_2$ and $\mathbf{F}'_3 \in \mathbb{R}^{H/8\times W/8\times 64}$ are derived from decoupled feature maps through convolutional operations and bilinear interpolation. As described in Section II-A, the output feature $\mathbf{F}$ from the dual-branch network is fed into the descriptor head to obtain the initial feature descriptor $\mathbf{F}_{des}$. Additionally, we design a parallel simple transformation head to convert $\mathbf{F}$ into the student feature $\mathbf{F}'_s \in \mathbb{R}^{H/8\times W/8\times 64}$, which is aligned with the pseudo label feature $\mathbf{F}'_t$ generated by Efficient-LoFTR to facilitate knowledge distillation.

Similarly, employing the MSE loss and KL divergence yields the optimized joint loss function as follows:

$$\mathcal{L}_{\text{KD-GD}} = \alpha \cdot \mathcal{L}_{\text{MSE}}(\mathbf{F}'_t, \mathbf{F}'_s) + \beta \cdot \mathcal{L}_{\text{KL}}(\mathbf{F}'_t, \mathbf{F}'_s) \tag{6}$$

where α, β are the weights. This paradigm effectively integrates the geometric perception capability from transformer-based matcher into the lightweight student model. The final feature descriptor is updated and obtained as shown below:

$$\mathbf{F'}_{des} = \text{Conv2d}\left(\mathbf{F}_{des} + \mathbf{F}'_s\right) \tag{7}$$

D. Network Training

Descriptor Learning. To guide the learning of local feature embeddings, we utilize a Negative Log-Likelihood (NLL) loss. Descriptor sets $\mathbf{F}_1$ and $\mathbf{F}_2 \in \mathbb{R}^{N\times 64}$ are sampled from dense maps $\mathbf{F}'_{des}$, and then a similarity matrix $\mathbf{S} = \mathbf{F}_1\mathbf{F}_2^\top \in \mathbb{R}^{N\times N}$ is defined to evaluate similarity. For matching, we adopt an approach based on LoFTR [19], resulting in the dual-softmax loss $\mathcal{L}_{\text{des}}$. The similarity measure of corresponding features lies along the main diagonal $\mathbf{S}_{ii}$ of $\mathbf{S}$, with Softmax$_r$ applied row-wise as described below:

TABLE I: **Relative pose estimation on MegaDepth-1500.** The best method is highlighted in bold, the second best is underlined, and the methods are separated by class (standard/fast), with * denoting 10k keypoints (Tab. I, II, III, IV, V, VI) .

	Method	AUC@5°	AUC@10°	AUC@20°	Acc@10°	MIR	#inliers
Standard	SuperPoint [12], CVPRW2018	37.3	50.1	61.5	67.4	0.35	495
	DISK [13], NeurIPS2020	53.8	65.9	75.0	**81.3**	**0.72**	1231
	DISK* [13], NeurIPS2020	**55.2**	**66.8**	**75.3**	**81.3**	0.71	**1997**
	SiLK [30], ICCV2023	14.7	21.5	29.3	31.9	0.17	235
	SiLK* [30], ICCV2023	16.2	23.2	31.8	34.7	0.14	478
Fast	ORB [15], ICCV2021	17.9	27.6	39.0	43.1	0.25	288
	ZippyPoint [31], CVPR2023	23.6	34.9	46.3	51.8	0.23	192
	ALIKE [11], TIM2023	**47.6**	59.8	69.7	75.9	0.53	625
	XFeat [10], CVPR2024	42.8	56.7	67.9	74.9	0.56	914
	D^{2}Feat (Ours)	46.5	**60.1**	**71.3**	**78.8**	**0.62**	**1050**

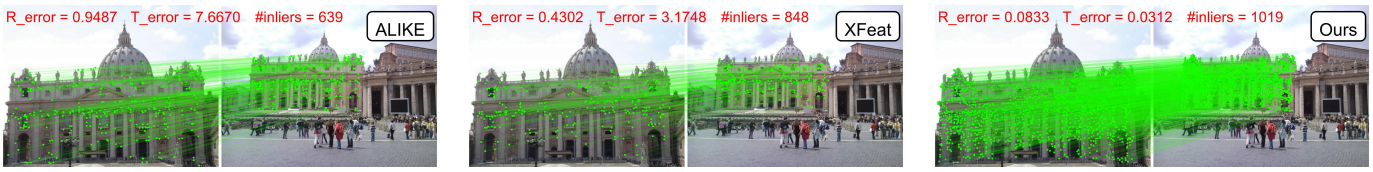

Fig. 3: **Qualitative results.** More visualizations are provided in appendix.

$$\mathcal{L}_{\text{des}} = -\sum_i \log\left(\text{Softmax}_r(\mathbf{S})_{ii}\right) - \sum_i \log\left(\text{Softmax}_r\left(\mathbf{S}^\top\right)_{ii}\right) \tag{8}$$

Reliability Learning. Following XFeat [10], we oversee the reliability map during the training process by treating the dual-softmax probability as a confidence indicator, represented as $\bar{\mathbf{R}} \in \mathbb{R}^N$. Similarly, the vectors $\bar{\mathbf{R}}_1$ and $\bar{\mathbf{R}}_2$ are derived by applying the dual-softmax method to match $\mathbf{F}_1$ and $\mathbf{F}_2$: $\bar{\mathbf{R}}_1 = \max_r(\text{Softmax}_r(\mathbf{S}))$ and $\bar{\mathbf{R}}_2 = \max_r(\text{Softmax}_r(\mathbf{S}^\top))$. Finally, we supervise the reliability map using the L1 loss:

$$\mathcal{L}_{\text{rel}} = |\sigma(\mathbf{R}_1) - \bar{\mathbf{R}}_1 \odot \bar{\mathbf{R}}_2| + |\sigma(\mathbf{R}_2) - \bar{\mathbf{R}}_1 \odot \bar{\mathbf{R}}_2| \tag{9}$$

where σ denotes the sigmoid activation function and $\odot$ represents the Hadamard product.

keypoints Learning. The keypoint detection branch is supervised using keypoints generated by ALIKE [11]. Considering the keypoint map $\mathbf{K} \in \mathbb{R}^{H/8 \times W/8 \times (64+1)}$, the coordinates of keypoints derived from the teacher network (t_x, t_y) within each cell $\mathbf{K}_{i,j} \in \mathbb{R}^{65}$ are transformed into a linear index $t_{idx} = (t_x + t_y \times 8)$, where t_{idx} ranges from 0 to 63. To oversee the dustbin category, t_{idx} is set to 64 when no keypoint is detected in cell $\mathbf{K}_{i,j}$. Then the keypoint loss $\mathcal{L}_{\text{key}}$ is computed using the NLL loss as follows:

$$\mathcal{L}_{\text{key}} = -\sum_k \log(\text{Softmax}(\mathbf{K}_{i,j}))_{t_{idx}} \tag{10}$$

Knowledge Distillation. The total knowledge distillation loss function is formulated as:

$$\mathcal{L}_{\text{kd}} = \mathcal{L}_{\text{KD-SD}} + \mathcal{L}_{\text{KD-GD}} \tag{11}$$

Finally, combining all individual loss components, the overall loss $\mathcal{L}$ is computed as described below:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{des}} + \lambda_2 \mathcal{L}_{\text{rel}} + \lambda_3 \mathcal{L}_{\text{key}} + \lambda_4 \mathcal{L}_{\text{kd}} \tag{12}$$

where λ_1, λ_2, λ_3 and λ_4 are weights for balancing.

III. Experiments

We evaluated D^{2}Feat on three widely adopted tasks: relative pose estimation, homography estimation, and visual localization. Additionally, we conducted ablation studies and performed overhead comparisons.

Implementation Details. To assess generalization and robustness, the model was trained only on MegaDepth [16] and synthetically warped COCO [32] dataset. More information and experimental details can be found in the appendix.

Comparative methods. Since our research primarily focuses on lightweight image matching frameworks and deployment in resource constrained environments, we selected several standard and fast baseline models for comparison, including DISK [13], ORB [15], SuperPoint [12], ALIKE [11], SilK [30], ZippyPoint [31], and XFeat [10]. Specifically, for SilK and ALIKE, we chose the VGGnp-μ and ALIKE-Tiny versions respectively. All baseline models were evaluated using the top 4096 detected keypoints, with the Mutual Nearest Neighbor (MNN) algorithm during the matching stage.

A. Relative Pose Estimation

Datasets. MegaDepth [16] is an outdoor multi-scene dataset, while ScanNet [17] is an indoor RGB-D dataset containing 2.5 million views in more than 1500 scans. Both feature complex scenes with significant viewpoint and illumination variations. We select the MegaDepth-1500 and ScanNet-1500 subsets as our test sets. Following XFeat [10], we use the LO-RANSAC [33] estimate the essential matrix.

Metrics. We report AUC at thresholds of {5°, 10°, 20°}, Acc@10° (fraction of poses with max angular error <10°), MIR (mean ratio of post-RANSAC inlier matches), and #inliers (number of inlier matches). In qualitative results (Fig.3), T_error (translation error) and R_error (rotation error), computed using ground truth labels, are also reported.

Results. As shown in Tab.I and Tab.II, we present the pose estimation results for outdoor and indoor scenes. Our method exhibits comparable computational overhead and inference speed (Tab.VI) to XFeat, while achieving outstanding performance on the MegaDepth-1500 dataset. In particular, on the unseen ScanNet-1500 dataset, our model demonstrates strong generalization capability, achieving accuracy improvements of 9.1%, 16%, and 22% over ALIKE [11], and 1.2%, 1.9%, and 2.8% over XFeat on AUC@{5°, 10°, 20°} respectively.

TABLE II: **Relative pose estimation on ScanNet-1500.**

	Method	**AUC**@5°	**AUC**@10°	**AUC**@20°
Standard	SuperPoint	12.5	24.4	36.7
	DISK / DISK*	9.6 / 11.3	19.3 / 22.3	30.4 / 33.9
Fast	ORB	9.0	18.5	29.9
	ALIKE	8.3	16.9	26.5
	XFeat	16.2	31.0	45.7
	D^{2}Feat	**17.4**	**32.9**	**48.5**

TABLE III: **Homography estimation on HPatches.**

		Illumination			Viewpoint		
	Method	@3	@5	@7	@3	@5	@7
Standard	SiLK	78.5	82.3	83.8	48.6	59.6	62.5
	SuperPoint	**94.6**	98.5	98.8	**71.1**	**79.6**	**83.9**
	DISK	**94.6**	**98.8**	**99.6**	66.4	77.5	81.8
Fast	ORB	74.6	84.6	85.4	63.2	71.4	78.6
	ZippyPoint	94.2	96.9	98.5	66.1	76.8	80.7
	ALIKE	**94.6**	98.5	**99.6**	67.9	78.2	82.9
	XFeat	94.2	97.7	98.9	68.2	81.1	85.7
	D^{2}Feat	94.2	**99.2**	99.2	**70.4**	**82.5**	**87.5**

B. Homography Estimation

Datasets. We utilized the widely adopted HPatches [34] dataset, comprising image sequences featuring diverse illumination conditions and viewpoint variations.

Metrics. Following ALIKE [11], we evaluate performance using the Mean Homography Accuracy (MHA). Here, we use the thresholds of 3, 5, and 7 pixels.

Results. Tab. III shows that our model outperforms XFeat [10] and ALIKE [11], as XFeat struggles with severe image variations and ALIKE is less effective under viewpoint changes.

C. Visual Localization

Datasets. We adopt the hierarchical localization pipeline HLoc [5] to localize images in Aachen Day-Night dataset [35].

Metrics. We adopt the standard HLoc metric, which evaluates the accuracy of correctly localized camera poses using position error thresholds of {0.25m, 0.5m, 5m} and corresponding rotation error thresholds of {2°, 5°, 10°}.

Results. Tab. IV show that our model achieves outstanding performance in visual localization compared to other lightweight models. Whether in daytime or night scenes, D^{2}Feat achieves accurate localization. Notably, at thresholds of (0.5m/5°) and (5m/10°), we achieve accuracy improvements of 3.1% and 1.6%, respectively, in night scenes.

TABLE IV: **Visual localization on Aachen Day-Night.**

		Day			Night		
	Method	0.25m 2°	0.5m 5°	5m 10°	0.25m 2°	0.5m 5°	5m 10°
Standard	SuperPoint	**87.4**	93.2	97.0	77.6	85.7	95.9
	DISK	86.9	**95.1**	**97.8**	**83.7**	**89.8**	**99.0**
Fast	ORB	66.9	76.1	81.7	10.2	12.2	19.4
	ZippyPoint	80.7	88.6	93.7	61.2	70.4	79.6
	ALIKE	**85.4**	91.5	95.0	**68.6**	83.8	92.1
	XFeat	82.3	89.6	**96.4**	65.4	83.8	95.8
	D^{2}Feat	84.2	**91.7**	**96.4**	**68.6**	**86.9**	**97.4**

TABLE V: **Ablation study on ScanNet-1500.**

Method	**AUC**@5°	**AUC**@10°	**AUC**@20°
Default	15.9	30.8	46.0
+ CNN	16.1	31.2	45.8
+ Single LoFTR	16.6	31.7	47.1
+ Single DINOv3	16.8	32.1	47.8
+ LoFTR & DINOv3	**17.4**	**32.9**	**48.5**

D. Ablation Study

We conduct ablation experiments to evaluate enhancement modules in Tab. V. Our baseline uses a single FPN backbone. Adding a second CNN branch as the student network increases parameters but yields no significant gains on the challenging ScanNet-1500 dataset. Distilling either DINOv3 [24] alone or Efficient-LoFTR [26] improves performance, showing effective learning of visual knowledge. Finally, joint distillation of DINOv3 and Efficient-LoFTR delivers the best results, with 1.5%, 2.1%, and 2.5% improvements on respective metrics, highlighting the benefits of fusing semantic and geometric cues from diverse pre-trained models for image matching.

TABLE VI: **Comparison of computation resources.**

	Method	**FLOPs (G)**	**GPU (FPS)**	**CPU (FPS)**	**Desc**
Fast	ALIKE	2.11	89.89	0.67	64
	XFeat	**1.33**	**99.01**	**2.27**	64
	D^{2}Feat	2.99	87.43	1.85	64

E. Performance Overhead

We compare the floating-point operations (FLOPs) and frames per second (FPS) of different lightweight models. The results in Tab. VI indicate that our model, compared to the state-of-the-art XFeat [10] and ALIKE [11] models, does not introduce excessive overhead but maintains a similar order of magnitude. While GPU inference speeds are comparable, our model achieves a 2.8× speedup over ALIKE on CPU.

IV. Conclusion

In this work, we introduced D^{2}Feat, a dual-distillation framework that unifies the strengths of self-supervised visual encoder and transformer-based feature matcher within a compact convolutional backbone. Through semantic and geometric distillation, complemented by the local Fine-grained Perceiver Module, our approach achieves efficient image matching.

V. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 62306154).

References

[1] C. Wu, "Towards linear-time incremental structure from motion," in *2013 International Conference on 3D Vision-3DV 2013*. IEEE, 2013, pp. 127–134.

[2] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[3] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE transactions on robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[4] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1744–1756, 2016.

[5] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 716–12 725.

[6] W. Cheng, W. Lin, K. Chen, and X. Zhang, "Cascaded parallel filtering for memory-efficient image-based localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1032–1041.

[7] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou, "Onepose: One-shot object pose estimation without cad models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6825–6834.

[8] X. He, J. Sun, Y. Wang, D. Huang, H. Bao, and X. Zhou, "Onepose++: Keypoint-free one-shot object pose estimation without cad models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 103–35 115, 2022.

[9] P.-E. Sarlin, M. Dusmanu, J. L. Schönberger, P. Speciale, L. Gruber, V. Larsson, O. Miksik, and M. Pollefeys, "Lamar: Benchmarking localization and mapping for augmented reality," in *European Conference on Computer Vision*. Springer, 2022, pp. 686–704.

[10] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento, "Xfeat: Accelerated features for lightweight image matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2682–2691.

[11] X. Zhao, X. Wu, J. Miao, W. Chen, P. C. Chen, and Z. Li, "Alike: Accurate and lightweight keypoint detection and descriptor extraction," *IEEE Transactions on Multimedia*, vol. 25, pp. 3101–3112, 2022.

[12] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[13] M. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," *Advances in neural information processing systems*, vol. 33, pp. 14 254–14 265, 2020.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[15] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[16] 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*, 2018, pp. 2041–2050.

[17] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[18] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[19] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[20] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "Roma: Robust dense feature matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 790–19 800.

[21] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," in *European Conference on Computer Vision*. Springer, 2024, pp. 71–91.

[22] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.

[23] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[24] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[25] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 17 627–17 638.

[26] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou, "Efficient loftr: Semi-dense local feature matching with sparse-like speed," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21 666–21 675.

[27] W. Zhang, Y. Liu, W. Ran, and C. Ma, "Cross-architecture distillation made simple with redundancy suppression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 23 256–23 266.

[28] Z. Hao, J. Guo, K. Han, Y. Tang, H. Hu, Y. Wang, and C. Xu, "One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 79 570–79 582, 2023.

[29] H. Jiang, A. Karpur, B. Cao, Q. Huang, and A. Araujo, "Omniglue: Generalizable feature matching with foundation model guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 865–19 875.

[30] P. Gleize, W. Wang, and M. Feiszli, "Silk: Simple learned keypoints," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 22 499–22 508.

[31] M. Kanakis, S. Maurer, M. Spallanzani, A. Chhatkuli, and L. Van Gool, "Zippypoint: Fast interest point detection, description, and matching through mixed precision discretization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6114–6123.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[33] V. Larsson *et al.*, "Poselib-minimal solvers for camera pose estimation," *PoseLib—Minimal Solvers for Camera Pose Estimation*, 2020.

[34] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5173–5182.

[35] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic *et al.*, "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8601–8610.