Mind the Gap: Aligning Vision Foundation Models to Image Feature Matching

Supplementary Material

1. Discussion on feature behavior

Given the constraints imposed by limited resources (e.g., computational capacity, dataset availability, and time), our investigation is primarily centered on the empirical analysis of existing models. Unfortunately, conducting comprehensive ablation studies to isolate the factors influencing feature properties such as data composition, training methodologies, and architectural variations, exceeds our current resource capabilities. Nevertheless, we propose several conceptual hypotheses to account for the observed differences disparities between SD [26] and DINOv2 [23]:

Training Paradigms. The core distinction between diffusion models and the other models is that they are **generative**. The training objective of diffusion models, which involves a coarse-to-fine reconstruction loss, necessitates the generation of informative features for every object within the image. For instance, since the model is trained to generate diverse representations of objects (e.g., a big car and a small car running), it must inherently require the model contain representations of every car. This stands in contrast to non-generative models, which typically employ either contrastive learning objectives (e.g., CLIP [42]) or discriminative objectives (e.g., ResNet [13], ViT [8]). Such objectives often result in the loss of instance-specific and object-level details.

Architecture Differences. Conversely, the self-attention mechanism in SD's UNet architecture that governs the dependencies between different regions of the image. This mechanism also enables the generation of structurally coherent objects and facilitates the explicit representation of object instances.

Despite the absence of comprehensive ablation studies, several experimental observations provide valuable insights:

Model Capacity : In our experiments as shown in Figure 4 of main tex and Tabel 1, we evaluate reduced-capacity versions of both SD. The distilled variants of SD exhibit properties consistent with the base model, yield performance improvements on the IMIM benchmark. These findings suggest that model capacity alone does not fully account for the observed feature characteristics.

Training Protocols : As shown in Table 1, we examine models with identical architectures but differing training regimes. For instance, the SD-1-5 model undergoes finetuning with distinct step counts and datasets compared to SD-1-3 (195,000 steps on "laion-aesthetics" versus 595,000 steps on "laion-improved-aesthetics"). Similarly, the SD-2-1-base model is trained on the filtered LAION-5B dataset. Despite these variations, all three models demonstrate comparable performance, indicating robustness to moderate differences in training protocols and datasets.

Table 1. Ablation results of different variants of SD models.

Method	**MegaDepth dataset**			**IMIM**
	AUC@5°	AUC@10°	ACU@20°	
SD-tiny	58.5	73.2	83.8	84.1
SD-small	60.0	74.5	85.0	86.5
SD-1-3	60.5	75.3	85.7	88.0
SD-1-5	60.8	75.6	85.8	88.3
SD-2-1(ours)	**61.2**	**76.1**	**85.8**	**88.7**

In summary, we believe that trying to investigate the underlying causes and establishing empirical evidence represents a compelling research direction and a promising avenue for future work.

2. Perliminaries

2.1. Denoising UNet of Diffusion Model

Denoising architecture of a Text-to-Image diffusion model is elaborated starting with its various layers. In the Latent Diffusion Model [26], the diffusion process occurs within the latent space of a previously trained image autoencoder. This model utilizes a U-Net [27] structure, which is conditioned on a guiding text prompt T. The U-Net consists of multiple layers, each consists three distinct types of blocks: (1) a residual block, (2) a self-attention block, and (3) a cross-attention block, as depicted in Figure 1. During each step of the denoising sequence, the noisy latent code z_t serves as the input to the U-net. The residual block processes the image features z_t to generate intermediary features $\varphi(z_t)$. In the self-attention block, the features $\varphi(z_t)$ are transformed into "queries" Q, "keys" K, and "values" V. Each query vector $q_{i,j}$, which represents a specific patch at spatial location (i, j) in Q, produces a self-attention map.

The final block, the cross-attention block, promotes the interaction between the spatial image features from the self-attention block and the text prompt T's token embeddings. This mechanism is similar to the one in the self-attention layer, except that here, Q is sourced from the spatial features of the prior self-attention layer, whereas K and V are derived from the text prompt's token embeddings.

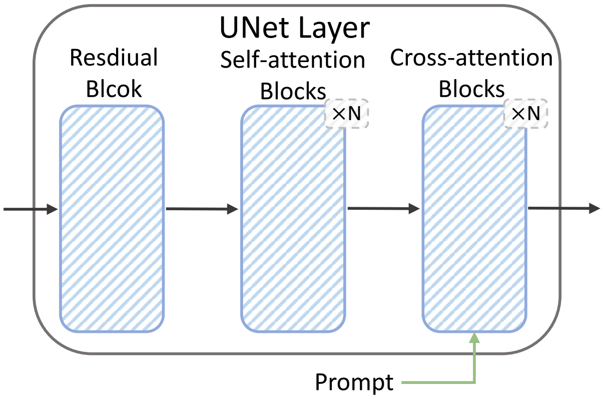

Figure 1. Block of Stable Diffusion UNet layer.

3. Implementation Details

3.1. Diffusion Models

The total number of time steps T for Stable Diffusion (SD) models is set to 1000. UNet architecture includes downsampling blocks, middle blocks, and upsampling blocks. We focus on extracting features from the upsampling blocks only. The UNet in SD consists of 4 upsampling blocks contained 15 layers. We utilize the feature maps from the n-th upsampling block as the final diffusion feature. The maximum prompt length supported by Stable Diffusion is 77 tokens, which includes two special tokens: SOS and EOS. Consequently, we configure the prompt length to 75 in CIPM thereby utilizing all 77 token positions.

3.2. Position Encoding

We employ the 2D extension of Rotary Position Encoding [30] to encode the relative positions between coarse features within self-attention modules. 2D RoPE enables the model to focus more on the interaction between features rather than their specific locations, which enhances the ability to capture the context of local features.

4. More Experiments Results

4.1. Visual Localization

Datasets and Evaluation Protocols. Visual localization is a critical task in image matching, aiming to determine the 6-DoF poses of query images based on a 3D scene model. The Aachen dataset comprises 6,697 daytime and 191 nighttime images, emphasizing the challenge of matching under significant illumination variations, particularly at night. We report the metrics separately for the daytime and nighttime subsets. Following the benchmark in [34], we compute query poses using the prescribed methodology. The candidate image pairs are identified using the pre-trained HLoc [28] system following [1].

Results. As shown in Table 2, regarding the Aachen V1.1 dataset, IMD yields strong competitive outcomes. Overall, the proposed method demonstrates strong performance and exhibits excellent generalization capabilities across a wide range of visual recognition tasks. These evaluations highlight the versatility of our approach in addressing diverse and complex problem scenarios.

Table 2. Results of Visual Localization on Aachen v1.1 Dataset.

Method	Day	Night
	(0.25m,2°)/(0.5m,5°)/(1.0m,10°)	
SP [7]+SG [29] CVPR'20	89.8 / 96.1 / **99.4**	77.0 / 90.6 / **100.0**
LoFTR [31] CVPR'21	88.7 / 95.6 / 99.0	**78.5** / 90.6 / 99.0
CasMTR [2] ICCV'23	**90.4** / **96.2** / 99.3	**78.5** / **91.6** / 99.5
AspanFormer [3] ECCV'22	89.4 / 95.6 / 99.0	77.5 / **91.6** / 99.5
PRISM [1] ACMMM'24	89.4 / **96.2** / 99.3	**78.5** / 91.1 / 99.5
IMD (**Ours**)	90.2 / **96.2** / 99.0	77.5 / 91.1 / 99.5

Table 3. Results of Homography Estimation on Hpatches Dataset.

Categeory	**Method**	**Homography est. AUC**			
		@3px	@5px	10px	mAUC
Zero-shot	CLIP [24] ICML'21	46.6	60.9	73.0	60.2
	DINOv2 [23] Arxiv'23	48.6	62.7	74.9	62.1
	DIFT [32] NeurIPS'23	54.1	64.5	76.3	65.0
Sparse	R2D2 [25]+NN NeurIPS'19	50.6	63.9	76.8	63.8
	D2Net [9]+NN CVPR'19	23.2	35.9	53.6	37.6
	DISK [37]+NN NeurIPS'20	52.3	64.9	78.9	65.4
	OmniGlue [14] CVPR'24	55.3	69.0	82.5	68.9
	SP [7]+SG [29] CVPR'20	53.9	68.3	81.7	68.0
Semi-Dense	LoFTR [31] CVPR'21	65.9	75.6	84.6	75.4
	Quadtree [33] ICLR'22	66.3	76.5	84.9	75.8
	ASpanFormer [3] ECCV'22	67.4	76.9	85.6	76.6
	Effcient LoFTR [39] CVPR'24	66.5	76.4	85.5	76.1
	EcoMatcher [4] ECCV'24	68.0	77.8	86.4	77.4
	JamMa [21] CVPR'25	68.1	77.0	85.4	76.8
	HomoMatcher [38] AAAI'25	70.2	79.6	87.8	79.2
	SRMatcher [18] ACMMM'24	71.2	79.3	87.0	79.2
	ASTR [40] CVPR'23	71.7	80.3	88.0	80.0
	CasMTR [2] ICCV'23	71.4	80.2	87.9	79.8
	PRISM [1] ACMMM'24	71.9	80.4	88.3	80.2
	IMD (Ours)	**73.9**	**82.0**	88.6	**81.5**
Dense	DKM [10] CVPR'23	71.3	80.6	88.5	80.1
	PMatch [43] CVPR'23	71.9	80.7	88.5	80.4
	RoMa [11] CVPR'24	72.2	81.2	**89.1**	80.8

4.2. The extend version of Homography Estimation

Table 3 shows the extend version of performance of various methods for homography estimation on HPatches. It is evident that, even compared to dense methods, our approach achieves the highest accuracy across multiple thresholds and in terms of mean average precision, demonstrates the effectiveness of our approach.

4.3. The extend version of Pose Estimation

Table 4 shows the extend version of results of multi instances evaluation on IMIM and two-view pose estimation on the MegaDepth, ScanNet datasets. Our proposed IMD outperforms across all evaluation metrics compared with spare and semi-dense methods on both benchmarks signif-

Table 4. **Results of multi instances evaluation on IMIM and two-view pose estimation on the MegaDepth [16], ScanNet [6] datasets.** Methods are grouped into 3 groups: 1) methods that are zero-shot and not fine-tuned on the training data, 2) sparse methods, 3) semi-dense methods. All the IMD result have gray background for easy lookup and annotate **best** results. The extend version are provided in the Suppl. B, G denote the model's size.

Category	**Method**	**MegaDepth**			**ScanNet**			**IMIM**
		AUC@5°	AUC@10°	AUC@20°	AUC@5°	AUC@10°	AUC@20°	
Zero-Shot	CLIP [24] ICML'21	30.8	48.1	63.2	10.1	20.6	31.3	54.4
	DINOv2 [23] Arxiv'23	32.5	50.8	65.3	13.0	28.5	40.8	57.9
	DIFT [32] NeurIPS'23	38.4	55.9	70.5	15.7	32.0	45.1	61.2
Sparse	SP [7]+NN CVPRW'23	31.7	46.8	60.1	7.5	18.6	32.1	55.9
	OmniGlue [14] CVPR'24	47.4	65.0	77.8	**31.3**	**50.2**	**65.0**	**77.6**
	SP [7]+LG [17] ICCV'23	**49.9**	**67.0**	**80.1**	14.8	30.8	47.5	60.5
Semi-Dense	LoFTR [31] CVPR'21	52.8	69.2	81.2	16.9	33.6	50.6	68.9
	RCM [20] ECCV'24	53.2	69.4	81.5	17.3	34.6	52.1	-
	Effcient LoFTR [39] CVPR'24	56.4	72.2	83.5	19.2	37.0	53.6	70.6
	EcoMatcher [4] ECCV'24	56.5	72.0	83.4	-	-	-	-
	HomoMatcher [38] AAAI'25	57.8	73.5	84.4	22.1	40.9	57.5	-
	TopicFM [12] AAAI'23	58.2	72.8	83.2	17.3	34.5	50.9	76.5
	MESA_ASpan [41] CVPR'24	58.4	74.1	84.8	-	-	-	-
	CasMTR [2] ICCV'23	59.1	74.3	84.8	22.6	40.7	58.0	79.2
	PRISM [1] ACMMM'24	60.0	74.9	85.1	23.9	41.8	58.9	-
	IMD (**Ours**)	**61.2**	**76.0**	**85.8**	**29.8**	**48.3**	**64.2**	**88.7**
Dense	DKM [10] CVPR'23	60.4	74.9	85.1	26.6	47.1	64.2	75.9
	RoMa [11] CVPR'24	62.6	76.7	86.3	28.9	50.4	68.3	79.7

icantly. However, it still falls short of the current state-of-the-art dense matching method RoMa. Our approach prioritizes overcoming misalignment in the adaptation of the misalignment model to the matching task, rather than driving final performance.

Table 5. Comparison with the state-of-the-art visual representations. B, L denote the model's size.

Model	Training Data	MegaDepth dataset			IMIM dataset
		AUC@5°	AUC@10°	ACU@20°	
Pre-trained with discriminative objectives					
DeiT-v3-B[35]	IN-21k	57.9	73.5	83.8	76.4
Swin-B[19]	IN-22k	57.5	73.2	83.6	74.0
Twins-SVT-L[5]	IN-1k	56.9	72.8	83.5	75.5
Pre-trained with contrastive learning objectives					
CLIP-L[24]	WIT	57.6	73.1	83.5	76.3
DINOv2-B[23]	IN-22k	57.8	73.5	83.7	75.5
Pre-trained with text					
SLIP-B[22]	YFCC15M	58.0	73.3	83.6	77.3
IMD	LAION	**61.2**	**76.0**	**85.8**	**88.7**

4.4. Ablations about Visual Representations

We compare the internal representations of text-to-image diffusion models with those of other state-of-the-art pre-trained models. Specifically, we evaluate a range of pre-trained models trained under different paradigms, including discriminative objectives, contrastive learning objectives, and pre-trained with text supervision. In all experiments, the weights of the pre-trained models are kept frozen, and we employ identical training hyperparameters as those used in our proposed method. For each category, we select the best-performing and largest publicly available models to ensure a comprehensive comparison. Table 5 below shows that IMD outperforms all pre-trained models by a large margin on both datasets, especially on IMIM. This highlights that the internal representation of the diffusion model is significantly more effective for feature matching and it can handle multi-instance scenarios well.

4.5. Ablations about Diffusion Time Steps

We further conduct more experiments about which diffusion step(s) are most effective for feature extraction. The level of noise distortion introduced to the input image escalates as the value of t increases. In the context of stable diffusion [26], the process comprises a total of 1000 time steps. As illustrated in Table 6, all evaluation metrics exhibit a decline as t increases, with the optimal performance observed at $t = 0$, which aligns with our final selected value. Combining features from three time steps, 0, 100, and 200, results in accuracy comparable to using only $t = 0$, but at the cost of being three times slower. This phenomenon also confirms the findings of DIFT [32], that larger values of t tend to produce features that are more semantically meaningful, whereas smaller values of t emphasize low-level details. The optimal selection of t depends on the specific requirements of the correspondence task, as different tasks may necessitate distinct balances between semantic and low-level features. For instance, tasks involving semantic correspondence are likely to benefit more from features with strong semantic representations as the setting of $t = 261$ in SD4Match [15]. However, feature matching tasks achieve better performance with features that capture low-level details with smaller t.

Table 6. Ablation results of different diffusion time steps.

Method	MegaDepth dataset			IMIM
	AUC@5°	AUC@10°	ACU@20°	
0	61.2	**76.0**	85.8	**88.7**
100	60.9	75.7	85.8	88.1
200	60.1	74.9	85.7	87.2
500	59.4	74.2	85.0	86.3
0+100+200	**61.4**	75.8	**86.0**	88.6

4.6. Ablations about Index of UNet Blocks

The index n of the U-Net upsampling block to extract the feature map, choose from [0, 1, 2, 3]. If $n = 0$, the output size would be 1/32 of image size; if $n = 1$, it would be 1/16; if $n = 2$ or 3, it would be 1/8. It is worth noting that our IMD employ the coasre-to-fine strategy following previous semi-dense methods [3, 31]. Large a scale difference may lead to a failure of refinement in the second stage, so we conduct experiments about $n = 1, 2, 3$. As shown in Table 7, when we extend our model with a more coarse level (1/16), IMD fails to produce accurate matching results, these matching errors inevitably propagate into subsequent learning stages.

Table 7. Ablation results of different index of UNet blocks.

Method	MegaDepth dataset			IMIM
	AUC@5°	AUC@10°	ACU@20°	
$n = 1(1/16)$	59.3	74.7	85.0	86.1
$n = 2(1/8)$	**61.2**	76.0	**85.8**	**88.7**
$n = 3(1/8)$	61.3	**76.1**	85.8	88.5

4.7. More Qualitative Comparisons

More qualitative results on the MegaDepth dataset and IMIM dataset are shown in Figure 2 and Figure 3, showing the effective of our method.

5. Time Cost

Our IMD requires only a single inference step for diffusion extraction, while $t = 0$ mitigating the inversion steps. As a result, the computational overhead is comparable to that of competing contrastive-learning based feature extraction methods when processing images of identical resolution. We reported the running time of matching each image pair in the ScanNet datasets for comprehensively understanding. All results are based on a singe NVIDIA 3090 GPU: IMD takes 165 ms vs. RoMa's 303 ms.

6. Limitation

The incorporation of the Stable Diffusion architecture, despite requiring merely a single forward pass, substantially escalates the computational overhead associated with correspondence estimation when compared to the methods that do not use the foundation models. Regarding practical efficiency, it is noteworthy that IMD demonstrates comparable computational speed to other widely-used contrastive-learning based feature extraction methods. Specifically, as detailed in Sec 5. We believe with the integration of advanced feature matching techniques which focus on the efficiency [17, 39], could potentially enhance both the performance and computational efficiency of IMD.

7. Future Work

The Scalability of IMD. Our investigation primarily focuses on UNet-based diffusion models due to their current prevalence as state-of-the-art text-to-image generation systems. However, we posit that analogous feature representations may exist in alternative diffusion architectures, supported by the following evidence: (1) Recent investigations [36] have demonstrated the presence of structural and appearance features in vision transformer-based architectures; (2) Empirical observations have consistently revealed instance-level features across multiple UNet diffusion variants. Based on these findings, we hypothesize that other diffusion model implementations (e.g., DiTs) may exhibit similar or potentially superior instance feature characteristics. This is a compelling direction for our future work.

Aggregation of Multi-layer Diffusion Features. All experimental results presented in this work utilize single-layer diffusion features exclusively. Preliminary investigations in Sec 4.5 suggest that feature aggregation across multiple time steps could yield performance improvements. However, such an approach introduces numerous design considerations (e.g., aggregation strategies) and hyperparameters (e.g., layer selection, timestep choices, and weighting schemes) that may require task-specific optimization. While parameter tuning could potentially enhance performance, it risks conflating feature quality with optimization efficacy. Given our primary objective of overcoming the misalignment between foundation models and feature matching, we employ the simplest configuration using raw, single timestep/layer features. The development of multi-layer/timestep diffusion feature pyramids remains an important direction for future research.

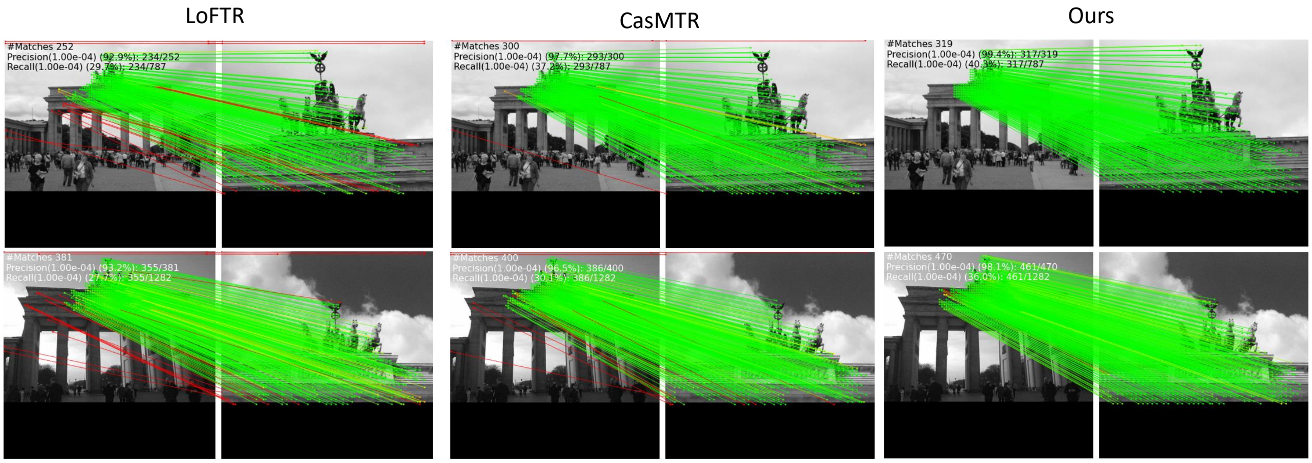
LoFTR
CasMTR
Ours
#Matches 252
Precision(1.00e-04) (92.9%): 234/252
Recall(1.00e-04) (29.7%): 234/787
#Matches 300
Precision(1.00e-04) (97.7%): 293/300
Recall(1.00e-04) (37.2%): 293/787
#Matches 319
Precision(1.00e-04) (99.4%): 317/319
Recall(1.00e-04) (40.3%): 317/787
#Matches 381
Precision(1.00e-04) (93.2%): 355/381
Recall(1.00e-04) (27.7%): 355/1282
#Matches 400
Precision(1.00e-04) (96.5%): 386/400
Recall(1.00e-04) (30.1%): 386/1282
#Matches 470
Precision(1.00e-04) (98.1%): 461/470
Recall(1.00e-04) (36.0%): 461/1282

Figure 2. Qualitative outdoor matching results compared with LoFTR [31], CasMTR [2] and ours IMD. The red color indicates epiploar error beyond 1×10^{-4} (in the normalized image coordinates).

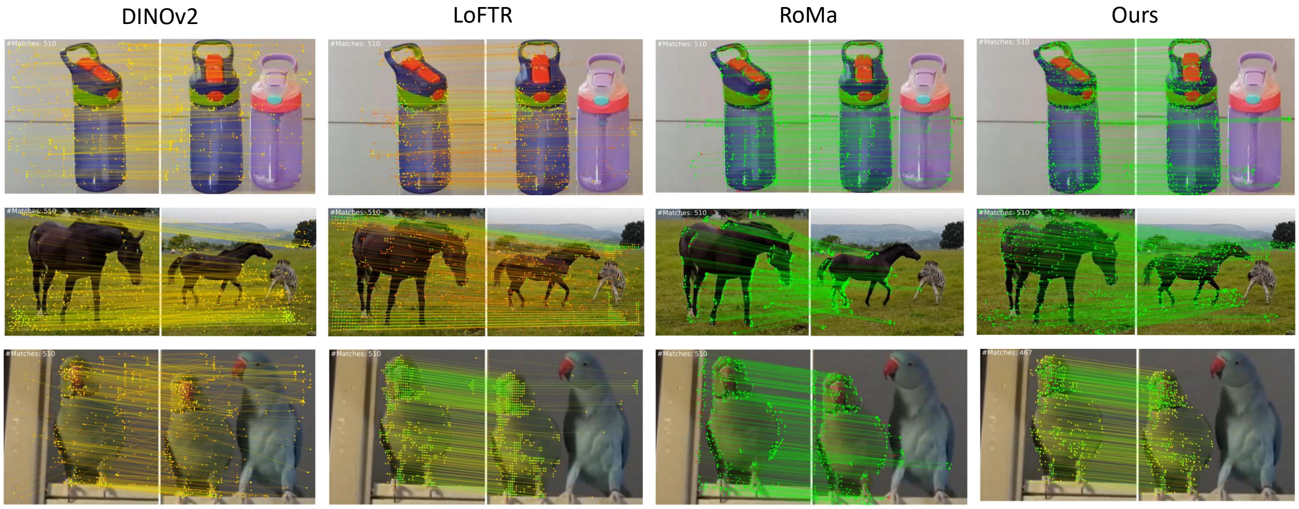

Figure 3. Qualitative IMIM matching results compared with DINOv2 [23], LoFTR[31], RoMa [11] and ours IMD. Green indicate the correct matching result. The results show scenarios where existing similar objects in the scene, whereas our proposed PDM effectively distinguishes and match the target object with high precision despite significant variations (view angle, pose).

References

[1] Xudong Cai, Yongcai Wang, Lun Luo, Minhang Wang, Deying Li, Jintao Xu, Weihao Gu, and Rui Ai. Prism: Progressive dependency maximization for scale-invariant image matching. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5250–5259, 2024. 2, 3

[2] Chenjie Cao and Yanwei Fu. Improving transformer-based image matching by cascaded capturing spatially informative keypoints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12129–12139, 2023. 2, 3, 5

[3] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 2, 4

[4] Peiqi Chen, Lei Yu, Yi Wan, Yongjun Zhang, Jian Wang, Liheng Zhong, Jingdong Chen, and Ming Yang. Ecomatcher: Efficient clustering oriented matcher for detector-free image matching. In *European Conference on Computer Vision*, pages 344–360. Springer, 2024. 2, 3

[5] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in neural information processing systems*, 34:9355–9366, 2021. 3

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

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2, 3

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[9] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8092–8101, 2019. 2

[10] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 2, 3

[11] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 2, 3, 5

[12] Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm: Robust and interpretable topic-assisted feature matching. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2447–2455, 2023. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[14] Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. Omniglue: Generalizable feature matching with foundation model guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19865–19875, 2024. 2, 3

[15] Xinghui Li, Jingyi Lu, Kai Han, and Victor Adrian Prisacariu. Sd4match: Learning to prompt stable diffusion model for semantic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27558–27568, 2024. 4

[16] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 3

[17] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 3, 4

[18] Yuhan Liu, Qianxin Huang, Siqi Hui, Jingwen Fu, Sanping Zhou, Kangyi Wu, Pengna Li, and Jinjun Wang. Semantic-aware representation learning for homography estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2506–2514, 2024. 2

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[20] Xiaoyong Lu and Songlin Du. Raising the ceiling: Conflict-free local feature matching with dynamic view switching. In *European Conference on Computer Vision*, pages 256–273. Springer, 2024. 3

[21] Xiaoyong Lu and Songlin Du. Jamma: Ultra-lightweight local feature matching with joint mamba. *arXiv preprint arXiv:2503.03437*, 2025. 2

[22] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022. 3

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 3, 5

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[25] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 2

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[28] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12716–12725, 2019. 2

[29] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2

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

[32] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 2, 3

[33] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. *arXiv preprint arXiv:2201.02767*, 2022. 2

[34] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, et al. Long-term visual localization revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2074–2088, 2020. 2

[35] Hugo Touvron, Matthieu Cord, and Herve Jegou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 3

[36] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 4

[37] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 2

[38] Xiaolong Wang, Lei Yu, Yingying Zhang, Jiangwei Lao, Lixiang Ru, Liheng Zhong, Jingdong Chen, Yu Zhang, and Ming Yang. Homomatcher: Dense feature matching results with semi-dense efficiency by homography estimation. *arXiv preprint arXiv:2411.06700*, 2024. 2, 3

[39] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei 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*, pages 21666–21675, 2024. 2, 3, 4

[40] Jiahuan Yu, Jiahao Chang, Jianfeng He, Tianzhu Zhang, Jiyang Yu, and Feng Wu. Adaptive spot-guided transformer for consistent local feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21898–21908, 2023. 2

[41] Yesheng Zhang and Xu Zhao. Mesa: Matching everything by segmenting anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20217–20226, 2024. 3

[42] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1

[43] Shengjie Zhu and Xiaoming Liu. Pmatch: Paired masked image modeling for dense geometric matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21918, 2023. 2