OmniGlue: Generalizable Feature Matching with Foundation Model Guidance

Hanwen Jiang* Arjun Karpur† Bingyi Cao† Qixing Huang* André Araujo†

*University of Texas at Austin
(hwjiang,huangqx)@cs.utexas.edu

†Google Research
(arjunkarpur,bingyi,andrearaujo)@google.com

Abstract

The image matching field has been witnessing a continuous emergence of novel learnable feature matching techniques, with ever-improving performance on conventional benchmarks. However, our investigation shows that despite these gains, their potential for real-world applications is restricted by their limited generalization capabilities to novel image domains. In this paper, we introduce OmniGlue, the first learnable image matcher that is designed with generalization as a core principle. OmniGlue leverages broad knowledge from a vision foundation model to guide the feature matching process, boosting generalization to domains not seen at training time. Additionally, we propose a novel keypoint position-guided attention mechanism which disentangles spatial and appearance information, leading to enhanced matching descriptors. We perform comprehensive experiments on a suite of 7 datasets with varied image domains, including scene-level, object-centric and aerial images. OmniGlue's novel components lead to relative gains on unseen domains of 20.9% with respect to a directly comparable reference model, while also outperforming the recent LightGlue method by 9.5% relatively. Code and model can be found at https://hwjiang1510.github.io/OmniGlue.

1. Introduction

Local image feature matching techniques provide fine-grained visual correspondences between two images [30], which are critical for achieving accurate camera pose estimation [39, 41] and 3D reconstruction [4, 15, 19, 42]. The past decade has witnessed the evolution from hand-crafted [3, 29] to learning-based image features [10, 36, 38, 51, 55]. More recently, novel learnable image matchers have been proposed [12, 27, 41, 44, 47], demonstrating ever-improving performance on conventional benchmarks [1, 8, 25].

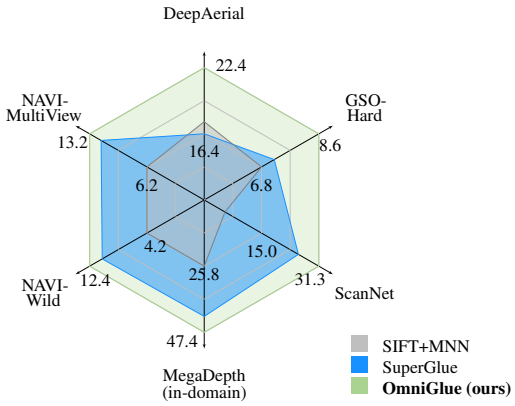

Figure 1. **OmniGlue is a generalizable learnable matcher.** Introducing foundation model guidance and an enhanced attention mechanism, OmniGlue learns effective image matching that transfers well to image domains not seen during training. We compare it against reference methods SIFT [29] and SuperGlue [41], with substantial improvements on a suite of diverse datasets: outdoor scenes (MegaDepth-1500 [25] pose AUC@5°), indoor scenes (ScanNet [8] pose accuracy @5°), aerial scenes (DeepAerial [35] PCK@1%) and object-centric images (GSO-Hard [11] and NAVI-MultiView / NAVI-Wild [18], pose accuracy @5°).

Despite substantial progress, these advancements overlook an essential aspect: the **generalization capability** of image matching models. Today, most local feature matching research [12, 27, 44] focuses on specific visual domains with abundant training data (*e.g.*, outdoor and indoor scenes), leading to models that are highly specialized for the training domain. Unfortunately, we observe that the performance of these methods usually drops dramatically on out-of-domain data (*e.g.*, object-centric or aerial captures), which may not even be significantly better than traditional approaches in some cases. For this reason, traditional domain-agnostic techniques, such as SIFT [29], are still widely used to obtain poses for downstream applications [2, 24, 31, 48]. Due to the cost of collecting high-quality correspondence annotations, we believe it is unrealistic to assume that abundant training data would be available for each image domain, like in some other vision tasks [9, 26]. Thus, the community should focus

*This work was completed while Hanwen was an intern at Google.

on developing architectural improvements to make learnable matching methods generalize.

Motivated by the above observations, we propose **OmniGlue**, the first learnable image matcher that is designed with generalization as a core principle. Building on top of domain-agnostic local features [10], we introduce novel techniques for improving the generalizability of matching layers: foundation model guidance and keypoint-position attention guidance. As shown in Fig. 1, with the introduced techniques, we enable OmniGlue to generalize better on out-of-distribution domains while maintaining quality performance on the source domain.

Firstly, we incorporate broad visual knowledge of a foundation model. By training on large-scale data, the foundation model, DINOv2 [34], performs well in diverse image domains on a variety of tasks, including robust region-level matching [21, 34, 56]. Even though the granularity of matching results yielded from foundational models is limited, these models provide generalizable guidance on potential matching regions when a specialized matcher cannot handle the domain shift. Thus, we use DINO to guide the inter-image feature propagation process, downgrading irrelevant keypoints and encouraging the model to fuse information from potentially matchable regions.

Secondly, we also guide the information propagation process with keypoint position information. We discover that previous positional encoding strategies [41] hurt performance when the model is applied to different domains – which motivates us to disentangle it from the matching descriptors used to estimate correspondence. We propose a novel keypoint-position guided attention mechanism designed to avoid specializing too strongly in the training distribution of keypoints and relative pose transformations.

Experimentally, we assess OmniGlue's generalization across diverse visual domains, spanning synthetic and real images, from scene-level to object-centric and aerial datasets, with small-baseline and wide-baseline cameras. We demonstrate significant improvements compared to previous work. In more detail, our contributions are as follows.

Contributions. **(1)** We introduce foundation model guidance to the learnable feature matching process, which leverages broad visual knowledge to enhance correspondences in domains that are not observed at training time, boosting pose estimation accuracy by up to **5.8%** (14.4% relatively). **(2)** A new strategy for leveraging positional encoding of keypoints, which avoids an overly reliant dependence on geometric priors from the training domain, boosting cross-domain transfer by up to **6.1%** (14.9% relatively). **(3)** We perform comprehensive experiments on 7 datasets from varied domains, demonstrating the limited generalizability of existing matching methods, and OmniGlue's strong improvements, with relative gains of **20.9%** on average in all novel domains. **(4)** By fine-tuning OmniGlue using limited amount of data from the target domain, we show that OmniGlue can be easily adapted with an improvement up to **8.1%** (94.2% relatively).

2. Related Work

Generalizable Local Feature Matching. Prior to the deep learning era, researchers focused on developing generalizable local feature models. For example, SIFT [29], SURF [3] and ORB [40] have been widely used for image matching tasks across diverse image domains. Still today, many computer vision systems ignore recent advances in learnable local features and rely on hand-crafted methods, for example, to obtain poses for downstream applications [2, 24, 31, 48]. One of the main reasons for such old hand-crafted methods to continue being adopted is that most of the recent learning-based methods [13, 32, 33, 38, 49] are specialized to domains with abundant training data, such as outdoor building scenes, and do not generalize well to other domains. Recently, the community shifted the main focus to develop learnable image matchers, which associate local features produced by off-the-shelf methods [10] or jointly learn feature description and association [44]. While they demonstrate better performance compared with hand-crafted matching systems, they make the entire image matching pipeline even more domain-specific. Our experiments show that learnable matchers specialize strongly in the training domain, with limited generalization. Our proposed OmniGlue improves the generalization capability of existing learnable matchers by introducing guidance from foundation models and improved positional encoding.

Sparse Learnable Matching. Sparse learnable image matching methods [6, 27, 41] associate sparse keypoints, produced by keypoint detectors. For example, SuperGlue [41] uses SuperPoint [10] for keypoint detection and leverages the attention mechanism [50] to perform intra- and inter-image keypoint feature propagation. However, SuperGlues shows limited generalization capability. One reason is that it entangles the local descriptors and positional information of the keypoints, making the matching process overly dependent on learned positional patterns. It hinders the generalizability to data with different position-related matching patterns. To solve this problem, OmniGlue proposes to disentangle them during the feature propagation, releasing the reliance on positional patterns and improving the generalization capability to images from diverse domains.

(Semi-)Dense Learnable Matching. Dense image matching methods jointly learn the image descriptors and the matching module, performing pixel-wise matching on the entire input images [7, 12, 44, 46, 52]. They benefit from the end-to-end learning pipeline and demonstrate better performance in the training domain. For example, the semi-dense method LoFTR introduces a coarse-to-fine correspondence prediction paradigm [44]. Another line of work di-

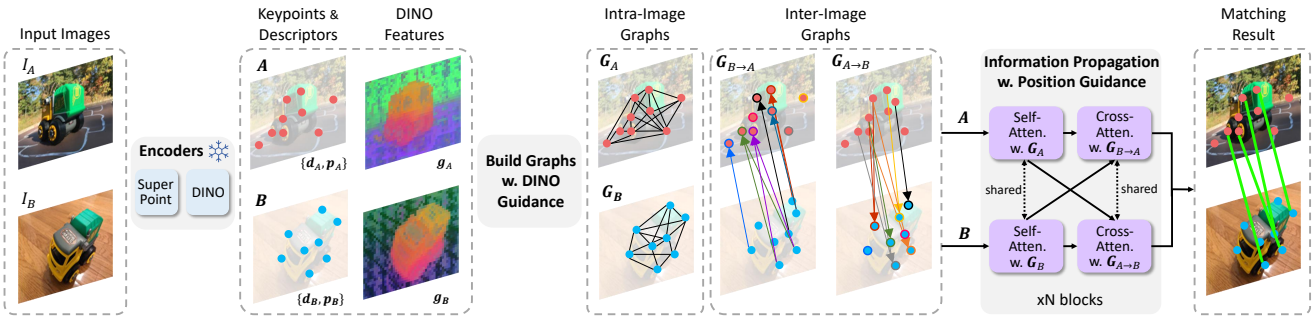

Figure 2. **OmniGlue overview.** We use frozen DINO and SuperPoint to detect keypoints and extract features. Then, we build densely connected intra-image keypoint graphs and leverage DINO features to build inter-image graphs. We refine the keypoint features based on the constructed graphs, performing information propagation. In this process, we use keypoint positions solely for guidance, disentangling them from the keypoint local descriptors. Finally, the matching results are produced based on the updated keypoint local descriptors.

rectly predicts the matching results as a 4D correlation volume [12, 46]. However, we notice that some of them generalize worse on new domains compared with sparse methods. Thus, OmniGlue chooses to focus on sparse methods, which can have better potential to be generalizable due to the use of domain-agnostic local descriptors.

Matching with Additional Image Representations. Leveraging robust image representations is a promising avenue toward generalizable image matching. One line of work uses geometric image representations, e.g., depth map [53] and NOCS map [23], to augment the image matching process. However, they are dependent on a highly accurate monocular estimation of these geometric representations. Differently, SFD2 [54] uses semantic segmentation results to reject indistinguishable keypoints in background regions. Nevertheless, the semantic segmentation model has to be trained on each specific target domain. Recently, large vision models, e.g., self-supervised vision backbones [5, 16, 34] and Diffusion models [17, 45, 56] demonstrate robust semantic understanding properties. By training on large data, these models showcase strong generalization capability across diverse domains [20, 21, 28], which enables them to obtain coarse patch-level matching results. However, performing matching using image features extracted by these models demonstrates limited performance on regions/keypoints without strong semantic information and the accuracy is limited [22, 56]. Instead of directly incorporating these coarse signals into the keypoint features and using them to perform matching, OmniGlue uses DINOv2 features to identify potentially related regions and guide the attention-based feature refinement process. Thanks to the wide domain knowledge encoded in this model, OmniGlue can boost the generalization ability of our method to diverse domains.

3. OmniGlue

We first introduce the overview and technical details of our method OmniGlue. Then we compare OmniGlue with SuperGlue and LightGlue for clarifying their differences.

3.1. Model Overview

Fig. 2 presents a high-level overview of our OmniGlue method, with four main stages. First, image features are extracted using two complementary types of encoders: SuperPoint [10], focusing on generic fine-grained matching; and DINOv2 [34], an image foundation model which encodes coarse but broad visual knowledge. Second, we build keypoint association graphs using these features, both intra and inter-image. In contrast to previous work, our inter-image graph leverages DINOv2 guidance, which provides a coarse signal capturing general similarity between SuperPoint keypoints. Third, we propagate information among the keypoints in both images based on the built graphs, using self and cross-attention layers for intra and inter-image communication, respectively. Crucially, we disentangle positional and appearance signals at this stage, different from other models that overlook this aspect. This design enables feature refinement to be guided by both keypoint spatial arrangement and their feature similarities, but without contaminating the final descriptors with positional information, which hinders generalizability. Finally, once the refined descriptors are obtained, optimal matching layers are applied to produce a mapping between the keypoints in the two images. These stages are described in more detail in the following section.

3.2. OmniGlue Details

Feature Extraction. The inputs are two images with shared content, denoted as I_A and I_B. We denote the SuperPoint keypoint sets of the two images as $\mathbf{A} := \{A_1, ..., A_N\}$ and $\mathbf{B} := \{B_1, ..., B_M\}$. Note that N and M are the number of identified keypoints of I_A and I_B, respectively. Each keypoint is associated with its SuperPoint local descriptor $\mathbf{d} \in \mathbb{R}^C$. Additionally, normalized keypoint locations are encoded with positional embeddings, and we further refine them using MLP layers. We denote the resulting positional features of a keypoint as $\mathbf{p} \in \mathbb{R}^C$. Furthermore, we extract dense DINOv2 feature maps of the two images. We interpolate the feature maps using the location of SuperPoint

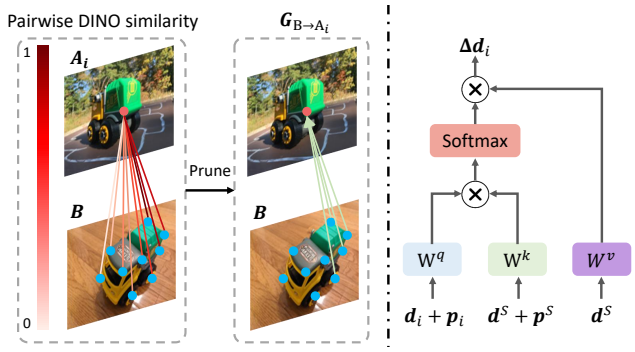

Figure 3. **(Left) Building inter-image graph.** We prune the dense pairwise graph based on the DINO feature similarity. **(Right) Position-guided attention.** The keypoint position is involved in computing attention weights, while the output attention update is only composed of local descriptor components.

keypoints to obtain DINOv2 descriptors for each keypoint, denoted as $\mathbf{g} \in \mathbb{R}^{C'}$. For clarity, we denote the three features of the i^{th} keypoint in set $\mathbf{A}$ as $\mathbf{d}_i^A$, $\mathbf{p}_i^A$ and $\mathbf{g}_i^A$. The features of the keypoints in set $\mathbf{B}$ are denoted accordingly. The goal of our OmniGlue model is to estimate correspondences between the two keypoint sets.

Graph Building Leveraging DINOv2. We build four keypoint association graphs: two inter-image graphs and two intra-image graphs. The two inter-image graphs represent the connectivity between the keypoints of the two images, from I_A to I_B and vice versa. We denote them as $\mathbf{G}_{A\to B}$ and $\mathbf{G}_{B\to A}$, respectively. The two inter-image graphs are directed, where information is propagated from the source node to the target node.

We leverage DINOv2 features to guide the building of the inter-image graphs. As depicted in Fig. 3 (left), we take $\mathbf{G}_{B\to A_i}$ as an example. For each keypoint A_i in keypoint set $\mathbf{A}$, we compute its DINOv2 feature similarities with all keypoints in set $\mathbf{B}$. Note that we perform channel-wise normalization on the DINOv2 features $\mathbf{g}_i^A$ and $\mathbf{g}^B$ before computing the similarities. We select the top half of keypoints in set $\mathbf{B}$ with the largest DINOv2 similarities to connect with A_i, which prunes the densely-connected pairwise graph between the keypoints of the two images. We perform the same operation on all keypoints in A to obtain $\mathbf{G}_{B\to A}$, and the graph $\mathbf{G}_{A\to B}$ is built in a similar manner.

Similarly, the intra-image graphs represent the connectivity between keypoints belonging to the same image. We denote them as $\mathbf{G}_A$ and $\mathbf{G}_B$, which are undirected – information is propagated bi-directionally between connected keypoints. Each keypoint is densely connected with all other keypoints within the same image.

Information Propagation with Novel Guidance. We perform information propagation based on the keypoint graphs. This module contains multiple blocks, where each block has two attention layers. The first one updates keypoints based on the intra-image graphs, performing self-attention; The second updates keypoints based on the inter-image graphs, performing cross-attention. In particular, this stage introduces two novel elements compared to previous work, which we show are critical towards generalizable matching: suitable guidance from DINOv2 and from keypoint positions.

First, DINOv2 guidance: during cross-attention, for keypoint A_i, it only aggregates information from the DINOv2-pruned potential matching set selected from $\mathbf{B}$, instead of all its keypoints. This is particularly helpful for generalized image matching, where DINO's broad knowledge may guide the feature matching process in a domain that the model has not seen at training time. In this manner, information from irrelevant keypoints will not be fused into the query keypoint features. This process also encourages the cross-attention module to focus on distinguishing the matching point in the smaller potential matching set. Note, however, that we do not forcibly limit the matching space to the potential matching sets, as DINO may also be incorrect in some cases.

Second, we introduce refined keypoint guidance. We observe that prior methods entangle keypoint positional features and local descriptors during feature propagation [41], which makes the model overly dependent on learned position-related priors – our ablation experiments in Section 4 highlight this issue. The learned priors are vulnerable under image pairs with matching patterns that were not seen at training time, limiting the generalization capability. To deal with this issue, we propose a novel position-guided attention, which disentangles the keypoint positional features $\mathbf{p}$ and the local descriptors $\mathbf{d}$. The positional information is used as spatial context in this module and is not incorporated in the final local descriptor representation used for matching.

With these novel elements, our attention layer, illustrated in Fig. 3 (right), is defined as follows, where we take the example of keypoint A_i:

$$\mathbf{d}_i^A \leftarrow \mathbf{d}_i^A + \text{MLP}([\mathbf{d}_i^A | \Delta\mathbf{d}_i^A]), \text{where} \quad (1)$$

$$\Delta\mathbf{d}_i^A = \text{Softmax}(\frac{\mathbf{q}_i^A(\mathbf{k}^S)^T}{\sqrt{C}}) \cdot \mathbf{v}^S \in \mathbb{R}^C, \text{and} \quad (2)$$

$$\mathbf{q}_i^A = \mathbf{W}^q(\mathbf{d}_i^A + \mathbf{p}_i^A) + \mathbf{b}^q \in \mathbb{R}^C, \quad (3)$$

$$\mathbf{k}^S = \mathbf{W}^k(\mathbf{d}^S + \mathbf{p}^S) + \mathbf{b}^k \in \mathbb{R}^{K\times C}, \quad (4)$$

$$\mathbf{v}^S = \mathbf{W}^v(\mathbf{d}^S) + \mathbf{b}^v \in \mathbb{R}^{K\times C}. \quad (5)$$

As described in Eq. 1, the attention has a residual connection, which integrates the attention update value $\Delta\mathbf{d}_i^A$. The notation $\leftarrow$ is the updating operation and $[\cdot|\cdot]$ is the channel-wise concatenation. To compute the attention update value, as described in Eq. 2, we compute the feature similarity between the keypoint A_i and its source connected keypoints in a graph, which is denoted as S containing K keypoints. The query, key and value of the attention are $\mathbf{q}_i^A$, $\mathbf{k}^S$, and $\mathbf{v}^S$, respectively. Specifically, as shown in Eq. 3-5, the query and key are computed by fusing both local descriptors and

positional features. The value, however, is transformed from only the local descriptors. We note that the weights ($\mathbf{W}$) and bias ($\mathbf{b}$), which map features into query, key and value tokens in attention, are not shared across different attention layers. In self-attention ($\mathbf{G}_A$ and $\mathbf{G}_B$), S is composed by all keypoints; in cross-attention ($\mathbf{G}_{A\to B}$ and $\mathbf{G}_{B\to A}$), S contains the keypoints identified by DINO.

Intuitively, the query and key compute the attention weights, where both feature affinity and spatial correlations are considered. However, the attention update value, $\Delta\mathbf{d}_i^A$, is composed of local descriptor components only. This design allows the model to reason about spatial correlation between keypoints using their positional features while avoiding an over-reliance on it.

Matching Layer and Loss Function. We use the refined keypoint representations to produce a pairwise similarity matrix $\mathbf{S} \in \mathbb{R}^{N\times M}$, where $\mathbf{S}_{i,j} = \mathbf{d}_i^A \cdot (\mathbf{d}_j^B)^T$. Then we use the Sinkhorn algorithm [43] to refine the similarities, which produces the matching matrix $\mathbf{M} \in [0,1]^{N\times M}$, where $\mathbf{M}_{i,j}$ represents the matching probability between keypoint A_i and B_j. To train OmniGlue, we minimize the negative log-likelihood of the matching matrix with ground truth [41, 44].

3.3. Comparison Against SuperGlue and LightGlue

It is important to highlight differences between our model and reference sparse learnable feature matching methods, SuperGlue [41] and LightGlue [27]. While neither of these is designed to target generalizability to multiple domains, there are common elements in the model structure, so we would like to emphasize our novelty.

Both works use attention layers for information propagation. Differently, OmniGlue leverages a foundation model to guide this process, which significantly helps with transferring to image domains that are not observed during training.

In terms of local descriptor refinement, OmniGlue departs from SuperGlue to disentangle positional and appearance features. For reference, SuperGlue represents keypoint with entangling the two features as $\mathbf{d} + \mathbf{p}$, where positional features are also used to produce matching results. Similar to our design, LightGlue removes the dependency of the updated descriptors on the positional features. However, it proposes a very specific positional encoding formulation, based on rotary encodings, only in self-attention layers.

Overall, SuperGlue is the closest model to OmniGlue, serving as a directly comparable reference where our contributions can be clearly ablated. For this reason, in the following section, we use SuperGlue as the main reference comparison for experimental validation.

4. Experiments

We first introduce the experiment setup and then present our results as well as ablation studies.

	(1) **Type**	(2) **Scene**	(3) **Real Img.**	(4) **Syn. Trans.**	(5) **Mask**	(6) **Cam. Bl.**	(7) **Diff. Bg.**	(8) **Task**
MegaDepth	Scene	Outdoor	✓	✗	✗	Large	✗	Corr. & Pose Est.
GSO-Hard	Object	None	✗	✗	✗	Large	✗	Pose Est.
GSO-Easy	Object	None	✗	✗	✗	Small	✗	Pose Est.
NAVI-MV	Object	In & Outdoor	✓	✗	✓	Large	✗	Pose Est.
NAVI-Wild	Object	In & Outdoor	✓	✗	✓	Large	✓	Pose Est.
ScanNet	Scene	Indoor	✓	✗	✗	Large	✗	Pose Est.
SH	Scene	Outdoor	✓	✓	✗	Small	✗	Corr. Est.
DeepAerial	Scene	Aerial	✓	✓	✗	N/A	✓	Image Reg.

Table 1. Dataset and task comparisons on: (1) The general type; (2) The background scene type; (3) Use of real (✓) or rendered (✗) images; (4) Whether the pose transformation is synthetic; (5) Whether foreground masks are used to filter correspondence predictions; (6) The camera baseline type; (7) Whether two input images have different backgrounds; (8) Evaluated tasks: Correspondence Estimation, Pose Estimation or Image Registration.

4.1. Experimental Setup

We list the datasets and tasks used for evaluating OmniGlue in Table 1. We include details of **datasets** as follows:

- **Synthetic Homography (SH)** contains images from the Oxford and Paris dataset [37]. We generate random crops and homography transformations to sample image patch pairs, similar to [41]. Two subsets are generated, SH100 and SH200, wherein the perturbations of the image corners for homography generation are within 100 and 200 pixels, respectively. For each subset, we generate roughly 9 million training pairs and 10K test pairs.
- **MegaDepth (MD)** [25] is a large-scale outdoor image dataset. The ground-truth matches are computed using SfM [42]. We follow the train/test split of prior works [44], with roughly 625K training pairs and 1500 test pairs.
- **Google Scanned Objects (GSO)** [11] comprises 1400 daily object model scans of 17 categories. We render synthetic images with large (60°- 90°) rotation (Hard subset) and small (15°- 45°) rotation (Easy subset) camera baselines, intentionally distinct from the training distribution. We produce 50 image pairs for each object model, resulting in around 140K test cases.
- **NAVI** [18] focuses on objects and encompasses a variety of both indoor and outdoor images. It is divided into two subsets: the multiview subset (25K image pairs), featuring input images captured in the same environment; and the wild subset (36K image pairs), where the two input images are taken in different environments with distinct backgrounds, lighting conditions and camera models.
- **ScanNet** [8] collects indoor images. We follow the split of prior works [44] with 1500 evaluation pairs.
- **DeepAerialMatching** [35] provides aligned pairs of satellite images under varying conditions (i.e. different seasons, weather, time-of-day). We introduce random 2D rotations and crop 520×520 image patches to produce image pairs with known affine transformations (500 in total).

Tasks and metrics. We assess the models across three

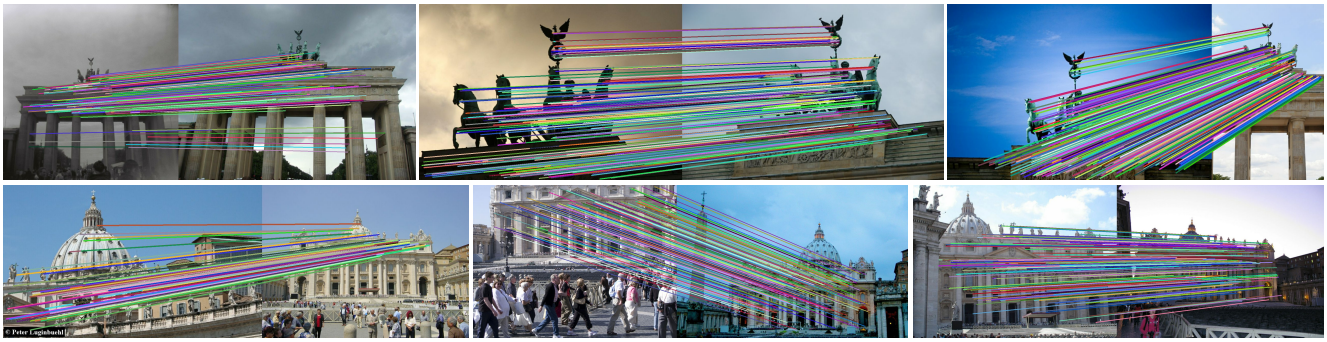

Figure 4. Visualization of correspondences predicted by OmniGlue on the MegaDepth-1500 benchmark. We distinguish the matches by different colors. We show results for scene "0022" and "0015" on the top and bottom rows, respectively.

tasks: (1) *Correspondence estimation*, evaluated with correspondence-level precision and recall (for sparse methods only). Following SuperGlue [41], we employ thresholds of $< 3px$ and $> 5px$ to label a correspondence as correct and incorrect, respectively. (2) *Camera pose estimation*, evaluated with pose accuracy (% of correct poses within $\{5°, 10°, 20°\}$ of error) and AUC, with accuracy being used by default unless otherwise specified. The poses are derived from the estimated correspondences using RANSAC [14], and we use Rodrigues' formula to calculate relative rotation error between the predicted/ground truth rotation matrices; (3) *Aerial image registration*, evaluated with percentage of correct keypoints (PCK). We use RANSAC-based affine estimation from the estimated correspondences, and apply the predicted/ground truth affine transformations to 20 test keypoints with fixed positions to calculate the PCK within $\tau \cdot max(h, w)$ pixels of error, for $\tau \in \{0.01, 0.03, 0.05\}$.

Baselines. We compare OmniGlue against:

- **SIFT** [29] and **SuperPoint** [10] provide domain-agnostic local visual descriptors for keypoints. We generate matching results using both nearest neighbor + ratio test (NN/ratio) and mutual nearest neighbor (MNN), with the best outcomes being reported.
- **Sparse matchers: SuperGlue** [41] employs attention layers for intra- and inter-image keypoint information aggregation, using descriptors derived from SuperPoint [10]. It is the closest reference of OmniGlue. **LightGlue** [27] improves SuperGlue [41] with better performance and speed. Besides, we also test with **DINOv2** [34]+SuperGlue, by substituting SuperPoint descriptors with DINO features.
- **(Semi-)Dense matchers: LoFTR** [44] and **PDCNet** [46] are used as reference dense matching techniques, to contextualize our sparse matching performance with respect to other types of approaches.

Implementation details. In line with SuperGlue [41], we implement 9 contextual reasoning blocks, each comprising an intra-image aggregation layer (self-attention) and an inter-image aggregation layer (cross-attention). This configuration results in a total of 18 attentional layers. Across all sparse methods, we use 1024 keypoints and 256-dimensional descriptors. See more training details in supplementary.

Setting →	**Test Performance (in-domain)**	
	SH100	*SH200*
DINOv2 [34]+SG [41]	87.6 / 88.4	79.8 / 80.2
SP[10]+SG [41]	**99.2** / 99.4	95.4 / 96.0
OmniGlue (ours)	**99.2 / 99.5**	**96.4 / 98.0**
Setting →	**Test Generalization** (*src* → *trg*)	
	SH100 → *SH200*	*SH200* → *MD*
DINOv2 [34]+SG [41]	72.6 / 77.3	19.2 / 18.8
SP[10]+SG [41]	78.3 / 75.6	34.9 / 39.0
OmniGlue (ours)	**90.0 / 89.6**	**36.0 / 54.7**
relative gain (%)	+14.9 / 18.5	+4.3 / +40.3

Table 2. Results for in-domain (top) and zero-shot generalization to out-of-domain datasets (bottom), for models trained on Synthetic Homography (SH) datasets. We measure precision / recall at the correspondence level.

4.2. Results

Following SuperGlue and LightGlue, we first initialize OmniGlue by training it on SH100. Then we further pretrain OmniGlue on SH200, and finally train OmniGlue on MegaDepth (MD). We evaluate OmniGlue and all baseline methods on the test splits of each training domain, and test their generalization to both subsequent training datasets or out-of-domain test datasets. Finally, we experiment with adapting OmniGlue to out-of-domain images with limited target domain training data.

From Synthetic Homography to MegaDepth. As depicted in Table 2, in comparison to the base method SuperGlue, OmniGlue not only exhibits superior performance on the in-domain data but also demonstrates robust generalization. Even with a minimal data distribution shift from SH100 to SH200, SuperGlue experiences substantial drops in performance with a 20% reduction in precision and recall. This result implies that SuperGlue is overly dependent on learned position-related patterns and is unable to handle further image warping distortion. In contrast, OmniGlue showcases strong generalization capability, surpassing Su-

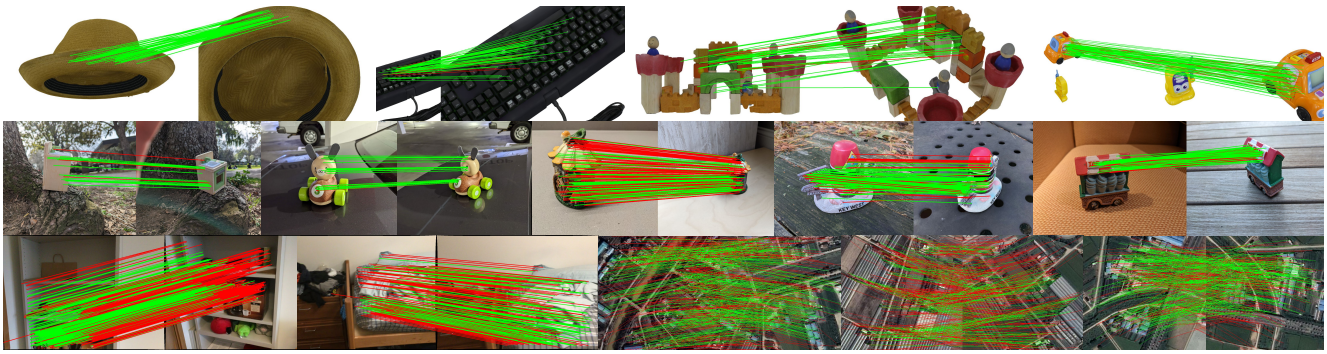

Figure 5. Zero-shot generalization to novel domains. The top and middle row show results on GSO and NAVI, the last row shows results on ScanNet and DeepAerial. We draw the correct and incorrect estimated correspondences as green and red, respectively.

	In-domain	**Out-of-domain** (Zero-shot Generalization)					
	MegaDepth-1500	*Google Scanned Object*		*NAVI*		*ScanNet*	*DeepAerial*
		Hard (60-90 deg.)	Easy (15-45 deg.)	Multiview	Wild		
Method	AUC@5°/ 10°/ 20°	Acc@5°/ 10°/ 20°	Acc@5°/ 10°/ 20°	Acc@5°/ 10°/ 20°	Acc@5°/ 10°/ 20°	Acc@5°/ 10°/ 20°	PCK@1%/3%/5%
DENSE AND SEMI-DENSE METHODS							
PDCNet [46]	51.5 / 67.5 / 78.2	5.1 / 8.9 / 14.9	24.8 / 36.7 / 49.3	3.9 / 7.1 / 11.6	6.6 / 11.6 / 17.0	**38.6 / 60.0** / 71.3	14.0 / 20.9 / 22.6
LoFTR [44]	**52.8 / 69.2 / 81.2**	7.6 / 14.0 / 22.9	38.2 / 54.1 / 67.5	12.5 / 22.7 / 34.2	9.8 / 18.4 / 29.8	36.2 / 56.1 / 68.6	17.8 / 23.7 / 25.0
DESCRIPTOR+HAND-CRAFTED RULES							
SIFT [29]+MNN	25.8 / 41.5 / 54.2	6.8 / 12.1 / 20.3	32.5 / 46.2 / 60.3	6.2 / 11.9 / 22.7	4.2 / 8.1 / 23.1	4.6 / 10.6 / 20.2	17.5 / 25.9 / 32.2
SuperPoint [10]+MNN	**31.7 / 46.8 / 60.1**	5.4 / 10.5 / 18.8	28.9 / 43.4 / 58.0	10.0 / 19.2 / 31.6	8.2 / 16.0 / 28.0	18.8 / 35.2 / 49.6	16.0 / 24.3 / 31.9
SPARSE METHODS							
DINOv2 [34]+SG [41]	31.5 / 40.8 / 45.3	3.6 / 7.3 / 15.1	12.0 / 22.7 / 38.7	7.3 / 15.6 / 28.3	8.4 / 17.2 / 30.6	9.7 / 26.7 / 41.5	11.4 / 18.2 / 23.1
SuperGlue [41]	42.2 / 61.2 / 76.0	7.2 / 13.2 / 21.6	32.3 / 48.4 / 62.9	11.8 / 21.9 / 34.4	10.6 / 19.8 / 31.8	25.5 / 43.4 / 57.3	16.4 / 26.2 / 28.8
LightGlue [27]	**47.6** / 64.8 / **77.9**	7.5 / 13.8 / 21.7	36.4 / 53.2 / 66.9	**13.2** / 24.0 / 34.8	9.7 / 17.6 / 25.9	36.7 / 59.4 / **71.6**	18.1 / 25.8 / 27.3
OmniGlue (ours)	47.4 / **65.0** / 77.8	**8.6 / 15.3 / 25.0**	**38.4 / 54.8 / 68.8**	**13.2 / 24.8 / 37.7**	**12.4 / 22.8 / 35.0**	31.3 / 50.2 / 65.0	**22.4 / 33.5 / 36.6**
rel. gain (%) over [41]	+12.3 / +6.2 / +2.4	+19.4 / +15.9 / +15.7	+18.9 / +13.2 / +9.4	+11.9 / +13.4 / +9.6	+16.7 / +15.2 / +10.1	+22.0 / +15.7 / +13.4	+36.6 / +27.9 / +27.0

Table 3. Results for in-domain (left, measured with AUC) and zero-shot generalization to out-of-domain datasets (right, measured with pose accuracy / PCK), for models trained on the MegaDepth dataset. We highlight the best results on out-of-domain data and show our relative improvement against our base method SuperGlue. All sparse methods use 1024 keypoints.

perGlue with a 12% improvement in precision and a 14% boost in recall. Similarly, during the transfer from SH200 to Megadepth, OmniGlue outperforms SuperGlue with a drastic 15% improvement in recall.

From MegaDepth to other Domains. As shown in Table 3, OmniGlue not only achieves comparable performance on MegaDepth-1500 with the state-of-the-art sparse matcher LightGlue, but also demonstrates better generalization capability on 5 out of 6 novel domains, when compared to all other methods. In detail, on MegaDepth-1500, OmniGlue showcases 12.3% relative gain (pose AUC @5°) over the base method SuperGlue. On the 6 novel domains, OmniGlue shows 20.9% and 9.5% averaged relative gains (for pose and registration accuracy at the tightest thresholds) over SuperGlue and LightGlue, respectively. Moreover, OmniGlue demonstrates larger performance gains on harder novel domains against LightGlue, i.e., on GSO-Hard, NAVI-Wild, and DeepAerial. We show visualization in Fig. 5 and Fig 4 for zero-shot generalization on novel domains and its performance on the source domain.

Notably, the reference dense matchers, which achieve better performance on the in-domain MegaDepth dataset, generalize worse. Their performances are close, or even worse, to SuperGlue, which has 10% lower in-domain AUC@5°. We conjecture this may be due to the joint learning of visual descriptors and the matching module, making them easier to specialize strongly to the training domain.

Low-Shot Fine-tuning on Target Domain. In certain real-world scenarios, a limited set of target domain data may be available for fine-tuning. To test this scenario, we fine-tune OmniGlue on the target domain (object-centric GSO dataset), comparing its performance with the base model, SuperGlue. We create small training subsets by utilizing only a few dozen object scans. Notably, these small training sets consist of instances from the sneaker object category only, covering a significantly minor subset of the testing object category distribution.

As depicted in Table 4, OmniGlue is more readily adapted to the target domain. In detail, when scaling from 0 to 30 instances for training, OmniGlue consistently exhibits enhanced performance for both test subsets. With just 10 instances for training, OmniGlue improves pose estimation accuracy by 5.3% and 4.0% on the two subsets. Expanding the training sets by incorporating 10 more objects leads to

#Train Inst.	Model	Hard (60-90 deg.) @5°/ 10°/ 20°	Easy (15-45 deg.) @5°/ 10°/ 20°
0	SG	7.2 / 13.2 / 21.6	32.3 / 48.4 / 62.9
	OG	**8.6 / 15.3 / 25.0**	**38.4 / 54.8 / 68.8**
10	SG	11.6 / 20.8 / 31.7	38.9 / 55.7 / 68.6
	OG	**13.9 / 24.6 / 36.8**	**42.4 / 60.1 / 74.0**
	rel. gain (%)	**+61.6 / +60.8 / +47.2**	**+10.4 / +9.7 / +7.6**
20	SG	13.0 / 22.9 / 35.2	40.3 / 57.0 / 70.5
	OG	**15.3 / 27.0 / 39.7**	**44.1 / 61.5 / 75.0**
	rel. gain (%)	**+77.9 / +76.5 / +58.8**	**+14.8 / +12.2 / +9.0**
30	SG	14.6 / 25.2 / 37.9	42.0 / 59.2 / 71.2
	OG	**16.7 / 29.1 / 42.3**	**45.8 / 62.5 / 76.0**
	rel. gain (%)	**+94.2 / +90.2 / +69.2**	**+19.3 / +14.1 / +10.5**

Table 4. Fine-tuning results of SuperGlue [41] (SG) and our method OmniGlue (OG) on Google Scanned Object (GSO) dataset. We use dozens of sneaker object instances to generate training data and test on all 17 GSO categories. We also show a relative gain compared with the zero-shot performance.

a further performance improvement of 2%. Furthermore, OmniGlue consistently surpasses SuperGlue, achieving a relative gain of approximately 10% across all experiments. The results collectively demonstrate the applicability of OmniGlue in real-world scenarios as a versatile and generalizable method.

4.3. Ablation Study and Insights

We conduct a comprehensive ablation study on each proposed module, as detailed in Table 5. Please note that the numbers reported on the GSO dataset are based on a subset, encompassing half of all test cases, for rapid evaluation.

The effectiveness of each proposed technique. The results in Table 5 (1) highlight the effectiveness of our foundation model guidance, which enhances the generalization capability on out-of-domain data. Additionally, the third row of Table 5 (2) illustrates the impact of the position-guided attention, showcasing improvement in both in-domain and out-of-domain data. Furthermore, we conduct ablations with different approaches to disentangling keypoint positional features. The first two rows of Table 5 (2) demonstrate that performance degrades when either not using any positional features or applying the position-guidance only on self-attention (without positional guidance on cross-attention). This emphasizes the effectiveness of our position-guided attention in facilitating information propagation within both intra- and inter-image contexts. Besides, after removing the positional embeddings, the model shows better generalization even though the in-domain performance drops. This result implies that the inappropriate way that SuperGlue uses positional information limits its generalization.

The ways of incorporating DINO features. As shown in Table 5 (3), we explore different methods of incorporating DINOv2. The first involves merging DINO features and SuperPoint local descriptors. This integration is performed before the information propagation module using an MLP.

		In-domain *MegaDepth* P / R	**Out-of-domain** *Google Scanned Object* Hard @5°/ 10°/ 20°	Easy @5°/ 10°/ 20°
(0)	SuperGlue [41]	67.2 / 68.3	9.0 / 16.9 / 27.3	40.4 / 60.5 / 76.6
(1)	only DINO-guide	66.6 / 68.0	10.0 / 18.7 / 29.6	46.2 / 65.4 / 79.5
(2)	only no pos. emb. - all	60.5 / 58.1	9.1 / 17.2 / 27.7	43.5 / 63.2 / 78.2
	only no pos. emb. - cross	63.3 / 62.1	9.3 / 17.0 / 28.0	44.8 / 64.1 / 79.4
	only pos. guidance	**69.2** / 73.9	9.8 / 18.0 / 28.6	46.4 / 66.6 / 80.2
(3)	(2) + DINO-SP-merge	62.6 / 65.6	7.8 / 14.9 / 24.9	42.5 / 61.3 / 75.4
	(2) + DINO-guide-intra+inter	66.4 / 72.2	10.5 / 19.4 / 30.5	47.1 / 66.8 / 80.8
(4)	(2) + DINO-guide-0.3	66.8 / 73.3	10.3 / 19.3 / 30.8	47.3 / 67.1 / 81.0
	(2) + DINO-guide-0.4	66.8 / 73.1	10.2 / 18.9 / 30.4	47.2 / 66.9 / 80.8
	(2) + DINO-guide-0.6	66.7 / 74.1	10.2 / 19.1 / 30.3	47.7 / 67.4 / 81.1
(5)	(2) + DINO-guide-0.5 (full)	66.2 / 74.1	**11.0 / 20.4 / 32.0**	**48.7 / 68.4 / 82.3**

Table 5. Ablation study on (1) only with DINO guidance, (2) only with the disentangled keypoint representation variants, (3) DINO guidance variants analysis (based on (2) with position guidance), (4) DINO guidance threshold analysis, and (5) full model OmniGlue.

The experiment reveals a decline in performance, suggesting that the two features are not compatible, likely due to the coarse granularity of DINO. The manner in which these features can be effectively merged remains an open problem.

The second method entails applying DINOv2 guidance for constructing both intra and inter-image graphs, demonstrating diminished performance compared to (5). We hypothesize that the reason lies in the fact that intra-image information propagation (self-attention) requires a global context, particularly for distinguishing all keypoints in the feature space. Reducing connectivity on the intra-image graph adversely affects the global context, aligning with findings in the study of attention span in SuperGlue.

Details of foundation model guidance. We ablate the hyperparameter used to determine the number of source keypoint in a graph, as presented in Table 5 (4). The results indicate that selecting the top half of keypoints in the other image for building inter-image graphs is the optimal choice.

5. Conclusions and Future Work

We propose OmniGlue, the first learnable image matcher that is designed with generalization as a core principle. We introduce the broad visual knowledge of a foundation model, which guides the graph-building process. We identify the limitation of the previous descriptor-position entangled representation and present a novel attention module to deal with it. We demonstrate that OmniGlue outperforms prior work with better cross-domain generalization. Moreover, OmniGlue can also be easily adapted to a target domain with a limited amount of data collected for fine-tuning. For future work, it is also worth exploring how to leverage unannotated data in target domains to improve generalization. Both of better architectural designs and better data strategies can pave the way for a foundational matching model.

Acknowledgement. Qixing Huang is supported by NSF IIS-2047677, HDR-1934932, CCF-2019844, and IARPA WRIVA program.

References

[1] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk. Hpatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. In *Proc. CVPR*, 2017. 1

[2] J. Barron, B. Mildenhall, D. Verbin, P. Srinivasan, and P. Hedman. Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. In *Proc. ICCV*, 2023. 1, 2

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, 2006. 1, 2

[4] César Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian D. Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32:1309–1332, 2016. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 3

[6] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6281–6290, 2021. 2

[7] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David N. R. McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, 2022. 2

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 5

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[10] 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. 1, 2, 3, 6, 7

[11] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Michael Hickman, Krista Reymann, Thomas Barlow McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560, 2022. 1, 5

[12] 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. 1, 2, 3

[13] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Dedode: Detect, don't describe - describe, don't detect for local feature matching. *ArXiv*, abs/2308.08479, 2023. 2

[14] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, 1981. 6

[15] Michael Goesele, Brian Curless, and Steven M Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2402–2409. IEEE, 2006. 1

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019. 3

[17] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. 2023. 3

[18] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. 1, 5

[19] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. *arXiv preprint arXiv:2212.04492*, 2022. 1

[20] Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. Leap: Liberate sparse-view 3d modeling from camera poses. *arXiv preprint arXiv:2310.01410*, 2023. 3

[21] Hanwen Jiang, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Single-stage visual query localization in egocentric videos. *arXiv preprint arXiv:2306.09324*, 2023. 2, 3

[22] Zhenyu Jiang, Hanwen Jiang, and Yuke Zhu. Doduo: Learning dense visual correspondence from unsupervised semantic-aware flow. *arXiv preprint arXiv:2309.15110*, 2023. 3

[23] Arjun Karpur, Guilherme Perrotta, Ricardo Martin-Brualla, Howard Zhou, and Andre Araujo. Lfm-3d: Learnable feature matching across wide baselines using 3d signals. In *Proc. 3DV*, 2024. 3

[24] K. Li, M. Runz, M. Tang, L. Ma, C. Kong, T. Schmidt, I. Reid, L. Agapito, J. Straub, S. Lovegrove, and R. Newcombe. FroDO: From Detections to 3D Objects. In *Proc. CVPR*, 2020. 1, 2

[25] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 1, 5

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[27] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *Proc. ICCV*, 2023. 1, 2, 5, 6, 7

[28] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 3

[29] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 1, 2, 6, 7

[30] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129:23–79, 2020. 1

[31] B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proc. ECCV*, 2020. 1, 2

[32] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 2

[33] Yuki Ono, Eduard Trulls, Pascal V. Fua, and Kwang Moo Yi. Lf-net: Learning local features from images. In *Neural Information Processing Systems*, 2018. 2

[34] Maxime Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 2, 3, 6, 7

[35] Jae-Hyun Park, Woo-Jeoung Nam, and Seong-Whan Lee. A two-stream symmetric network with bidirectional ensemble for aerial image matching. *Remote Sensing*, 12(3):465, 2020. 1, 5

[36] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. Nascimento. Enhancing Deformable Local Features by Jointly Learning to Detect and Describe Keypoints. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1

[37] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5706–5715, 2018. 5

[38] Jérôme Revaud, Philippe Weinzaepfel, César Roberto de Souza, No'e Pion, Gabriela Csurka, Yohann Cabon, and M. Humenberger. R2d2: Repeatable and reliable detector and descriptor. *ArXiv*, abs/1906.06195, 2019. 1, 2

[39] Barbara Roessle and Matthias Nießner. End2end multi-view feature matching with differentiable pose optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 477–487, 2023. 1

[40] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. Orb: An efficient alternative to sift or surf. *2011 International Conference on Computer Vision*, pages 2564–2571, 2011. 2

[41] 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. 1, 2, 4, 5, 6, 7, 8

[42] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 5

[43] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 5

[44] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8918–8927, 2021. 1, 2, 5, 6, 7

[45] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 3

[46] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5710–5720, 2021. 2, 3, 6, 7

[47] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. 2023. 1

[48] M. Tyszkiewicz, K.-K. Maninis, S. Popov, and V. Ferrari. RayTran: 3D pose estimation and shape reconstruction of multiple objects from videos with ray-traced transformers. In *Proc. ECCV*, 2022. 1, 2

[49] Michal J. Tyszkiewicz, P. Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *ArXiv*, abs/2006.13566, 2020. 2

[50] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. 2

[51] Yannick Verdie, Kwang Moo Yi, Pascal Fua, and Vincent Lepetit. TILDE: A Temporally Invariant Learned Detector. In *Proc. CVPR*, 2015. 1

[52] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Asian Conference on Computer Vision*, 2022. 2

[53] Shuzhe Wang, Juho Kannala, Marc Pollefeys, and Daniel Barath. Guiding local feature matching with surface curvature. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17981–17991, 2023. 3

[54] Fei Xue, Ignas Budvytis, and Roberto Cipolla. Sfd2: Semantic-guided feature detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5216, 2023. 3

[55] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned Invariant Feature Transform. In *Proc. ECCV*, 2016. 1

[56] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming Yang. A tale of two features: Stable diffusion complements dino for

zero-shot semantic correspondence. *ArXiv*, abs/2305.15347, 2023. 2, 3