MATCHA🍵: Towards Matching Anything

Fei Xue[1*] Sven Elflein[2,3,4] Laura Leal-Taixé[3] Qunjie Zhou[3†]
[1]University of Cambridge [2]University of Toronto [3]NVIDIA [4]Vector Institute

Abstract

Establishing correspondences across images is a fundamental challenge in computer vision, underpinning tasks like Structure-from-Motion, image editing, and point tracking. Traditional methods are often specialized for specific correspondence types, geometric, semantic, or temporal, whereas humans naturally identify alignments across these domains. Inspired by this flexibility, we propose MATCHA, a unified feature model designed to "rule them all", establishing robust correspondences across diverse matching tasks. Building on insights that diffusion model features can encode multiple correspondence types, MATCHA augments this capacity by dynamically fusing high-level semantic and low-level geometric features through an attention-based module, creating expressive, versatile, and robust features. Additionally, MATCHA integrates object-level features from DINOv2 to further boost generalization, enabling a single feature capable of matching anything. Extensive experiments validate that MATCHA consistently surpasses state-of-the-art methods across geometric, semantic, and temporal matching tasks, setting a new foundation for a unified approach for the fundamental correspondence problem in computer vision. To the best of our knowledge, MATCHA is the first approach that is able to effectively tackle diverse matching tasks with a single unified feature. Project page: https://github.com/feixue94/matcha.

1. Introduction

"In computer vision, there is only one problem: correspondence, correspondence, correspondence." –Takeo Kanade

Establishing correspondences between images is a fundamental problem in computer vision, integral to a variety of applications such as mapping and localization [44, 58], image editing [45], object pose estimation [70] and point tracking [13, 22]. Correspondence is typically categorized by type: geometric [11, 38, 49, 56], semantic [29, 73, 74] and temporal [7, 27, 55, 61] correspondences, as shown in Fig. 1. Geometric correspondences identify 2D points in images of static scenes that represent the same physical 3D point, with challenges in diverse illumination and viewpoint variations. They are typically used to eestimate geometric camera transformations *e.g.*, for structure-from-motion applications. Semantic correspondences connect similar object parts across distinct instances within a category, demanding high-level abstraction across different instances. Temporal correspondences, in contrast, match points of the same instance across video frames, require to handle both static and dynamic elements, occlusions, deformations and viewpoint changes stemming from complex motions.

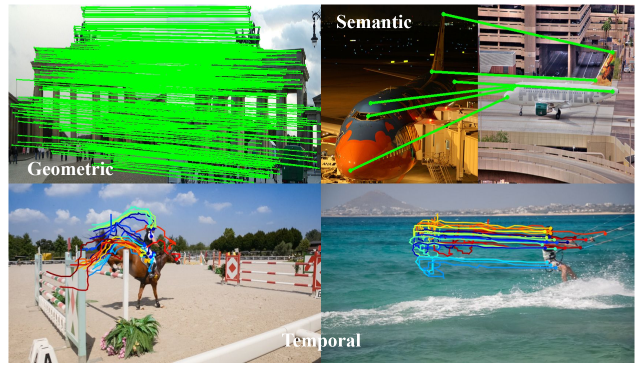

Figure 1. **MATCHA for *matching anything*.** We visualize geometric, semantic and temporal correspondences established by MATCHA, using a single feature descriptor.

Addressing these distinct challenges usually requires specialized models [11, 34, 40, 47, 64, 74]. However, humans can align points flexibly across different scenarios, *e.g.*, across static scenes, dynamic objects of different instances under various viewpoints, prompting the question: *Do we really need a separate feature for each type of correspondence problem?* DIFT [61] offers a step toward this, revealing that correspondence patterns can emerge naturally from diffusion models [12, 54]. However, DIFT still relies on distinct feature descriptors for different tasks, potentially limiting its utility when the matching type is unknown. More importantly, the unsupervised correspondences learned by DIFT fall short of fully supervised methods in matching accuracy (*cf*. Sec. 4.1 and Sec. 4.2).

In this work, we introduce MATCHA, a foundation fea-

*This work was done when Fei Xue was an intern at NVIDIA.
†Project leader.

ture model for *matching anything*. Unlike DIFT, MATCHA learns a single descriptor for geometric, semantic, and temporal matching, incorporating explicit supervision while leveraging the rich knowledge of foundation models. Our experiments confirm that combining the knowledge of foundation models with targeted supervision is key to accurate and generalized matching (*cf*. Tab. 3 and Tab. 4). While adding correspondence-level supervision is straightforward, annotated datasets for supervision are limited, compared to the scale of data a foundation model is usually trained on, especially for semantic and temporal matching tasks where human annotations are required for real-world data. Thus, the main challenge we need to address is to find a proper way to inject accurate correspondence supervision from only a limited amount of annotated data, without destroying the rich information and generalization of features learned by foundation models. To achieve this, we leverage an attention-based dynamic feature fusion module that learns to extract mutually supportive knowledge from two domains, *i.e*., semantic and geometric, to enhance themselves for improved matching performance. Guided by correspondence-level supervision, our attention-based fusion enhances diffusion features without losing generalization. Supported by the fusion process, we are able to combine the enhanced diffusion features with the complementary semantic knowledge from DINOv2, which captures robust, single-object correspondences (as shown in Fig. 2). The result is a unified, high-quality feature that achieves strong matching performance across different tasks.

We summarize the contributions of this work as follows: We (i) systematically analyze common feature models for matching, informing the design of MATCHA, a novel feature model that learns to dynamically fuse geometric and semantic information to improve representational robustness without loss of generality. MATCHA demonstrates that (ii) static fusion of features can offset the limitations of individual descriptors, enabling a single feature to address a range of correspondence tasks effectively. Comprehensive evaluations show (iii) MATCHA surpasses state-of-the-art on most benchmarks, significantly outperforming unsupervised methods in semantic and geometric matching, highlighting the importance of correspondence supervision for precision. For the first time, we show that (iv) a single feature is able to achieve the new state-of-the-art across all three types of common correspondence problems. (iiv) As a contribution to the community, we re-purpose the TAP-Vid point tracking benchmark [13] for temporal matching evaluation, establishing common feature baselines to support future research on unified feature learning for matching.

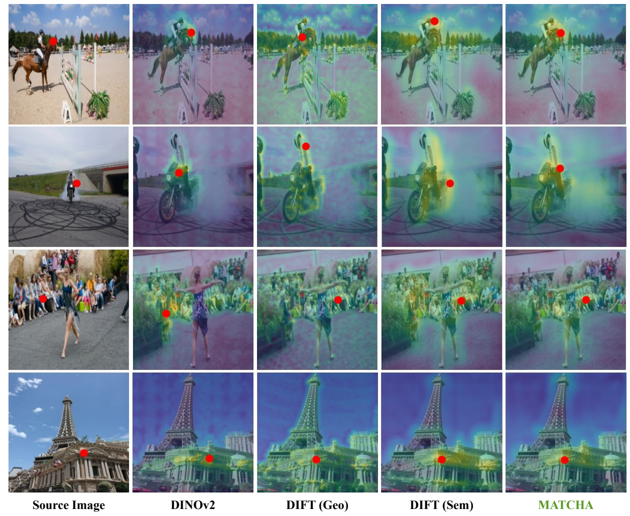

Figure 2. **Heatmap of features from DINOv2, DIFT, and MATCHA.** Given a query point from the source image (1st column), DINOv2 features give more accurate correspondences on single object (1st and 2nd row) but struggle when multiple instances of the same class (3rd row) or similar structures (4th row) exist. Both geometric and semantic features of DIFT perform reversely. By unifying knowledge in the three foundation features, MATCHA produces more accurate and reliable correspondences.

2. Related Work

Geometric Correspondence. Geometric matching refers to searching physically correspondent point pairs between two images captured in the same scene. Geometric correspondences are commonly established by detecting, describing, and matching local features. An abundant of local feature detection and description methods have been developed starting from hand-crafted local features such as SURF [2] and SIFT [38] and then evolving towards learned ones [11, 15, 41, 49, 62, 64, 71, 72]. Benefiting from massive training data, the learned features show better discriminative ability to viewpoint and illumination changes than handcrafted ones. However, as most of these learned features are trained with purely geometric ground-truth correspondences mainly from static objects, despite their promising accuracy on geometric matching, they have poor performance especially on semantic matching (*cf*. Sec. 4.1 and Sec. 4.2). Geometric correspondences can be obtained with nearest neighbor matching [43] based on descriptor distance. Although more powerful learned sparse [26, 36, 56] and dense [4, 16, 18, 19, 52, 60, 75] matchers are proposed, in this paper, we focus mainly on the feature itself and use nearest neighbor matching to find correspondences.

Semantic Correspondence. Semantic matching aims to match points with similar semantic meaning across different instances of the same category, *e.g*., matching the eyes of a cat in one image to another cat in the other image. Semantic matching methods focus on extracting feature descriptors [8, 29] to capture semantic information. Recent works [23, 34, 40, 61, 73, 74] leverage features extracted from foundation models [3, 12, 46, 48, 54] due to their rich semantic knowledge which is hard to learn from a limited amount of annotated semantic matching training data. These methods, *e.g.*, DIFT [61] and SD+DINO [74] use the foundation model features directly for semantic matching. However, their performance is not comparable to those finetuned with supervision, *e.g.*, DHF [40] and SD4Match [34]. Some works also build semantic matchers for matching from the perspective of customized matching functions [28, 32, 37], correspondence networks [50, 51] or semantic flow [5, 24, 30, 32, 52, 63]. These methods require paired images as input rather than single images.

Temporal Correspondence. Temporal matching targets at establishing correspondences of the same object across video frames. It generalizes the geometric matching task from static scenes to general natural scenes that contain both static and dynamic content. Recently, temporal correspondence has been largely investigated in its downstream application task, *i.e.*, tracking any point (TAP) [13]. The point tracking work [6, 7, 13, 14, 22, 27, 55] focuses on occlusion handling and exploring temporal priors, *e.g.*, long-term consistency, motion constraints, as well as leveraging 3D reconstruction [39, 59, 66, 67, 69]. Compared to these works, we are interested in the general problem of establishing pair-wise correspondences of any two frames from a video without leveraging any temporal constraints.

Vision Foundation Model. Modern vision foundation models, *e.g.*, DINO [3, 46], CLIP [25, 48], and diffusion models [12, 54], exhibit strong generalization performance across a variety of tasks or domains. Excitingly, their features show promising accuracy for both geometric [18, 26] and semantic [23, 34, 40, 50] correspondences, or even directly delivering emergent correspondences without an explicit supervision [61, 74]. DIFT [61] demonstrates that rich semantic and geometric features have been learned by image diffusion models and can be utilized to directly establish semantic, geometric and temporal correspondences without further supervision. SD+DINO [74] reveals that features from different foundation models have different properties and demonstrates that the combination of SD feature and DINO feature gives better semantic accuracy than either of them. Inspired by these two works, we leverage SD model and DINOv2 as our backbones to provide raw features. However, essentially different with these two works, we focus on how to obtain a single feature for all three types of matching by involving the supervision signals.

3. MATCHA

In this section, we present MATCHA, a novel feature model that unifies knowledge from multiple foundation models [3, 12, 46, 54] and enhances features for accurate correspondences through precise supervision, enabling a single feature descriptor for correspondence problems across different domains, reaching the state-of-the-art performance.

3.1. Preliminary

Our method is inspired by previous work DIFT that extracts features from a diffusion model for unsupervised matching. We also build on top of DINOv2 [46], a powerful self-supervised foundation model.

DIFT [61]. DIFT demonstrates that diffusion models trained for image generation implicitly learn correspondences. By extracting features from specific layers and timestamps, DIFT identifies effective feature descriptors for geometric, semantic, and temporal matching tasks. Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$, DIFT extracts a semantic descriptor $F_h \in \mathbb{R}^{H/16\times W/16\times 1280}$ and a geometric descriptor $F_l \in \mathbb{R}^{H/8\times W/8\times 640}$ from a pre-trained stable diffusion model [53]. While F_h is used for semantic matching and F_l for geometric and matching, DIFT requires manual selection of descriptors per task, which limits flexibility and generalization. Our approach eliminates the need for task-specific descriptors, achieving greater accuracy across tasks while maintaining a single, unified descriptor.

DINOv2 [46]. DINOv2 is trained on millions of images for object- and patch-level discrimination, allowing its features to capture rich semantic information for establishing object-level correspondences, as shown in recent work [74]. In our experiments, DINOv2 also exhibits robust handling of extreme viewpoint changes and scale variations for individual objects, excelling in temporal matching tasks (*cf*. Sec. 4.3). While DINOv2 and DIFT (F_h) both provide semantic descriptors, our results show complementary strengths between the two that enhance general matching capability (*cf*. Tab. 3). However, DINOv2's lack of spatial detail limits its geometric matching performance. Our approach integrates knowledge from stable diffusion and DINOv2, unifying them into a powerful, single representation for matching across diverse tasks. We denote the feature extracted from DINOv2 as $F_d \in \mathbb{R}^{H/14\times W/14\times 1024}$.

3.2. Architecture

Overview. As shown in Fig. 3, given an RGB image I as input, MATCHA outputs a single feature descriptor $F_m \in R^{H/8\times W/8\times D_m}$ (D_m is its channel size) for matching, including geometric, semantic and temporal matching tasks. (i) First, we build on top of the foundation feature models, DIFT and DINOv2, by obtaining two semantic feature descriptors F_h and F_d and a geometric descriptor F_l (*cf*. Sec. 3.1). (ii) We next enhance the DIFT geometric and semantic features F_l and F_h by learning to extract supportive information from the other domain's descriptor. Such dynamic fusion is learned via correspondence-level joint supervision on semantic and geometric matching. We show in our later ablations that such a learned dynamic fusion is critical for a successful and balanced merging stage where each descriptor can build on top of each other. (iii) Finally, we directly merge the two enhance features and the DINOv2

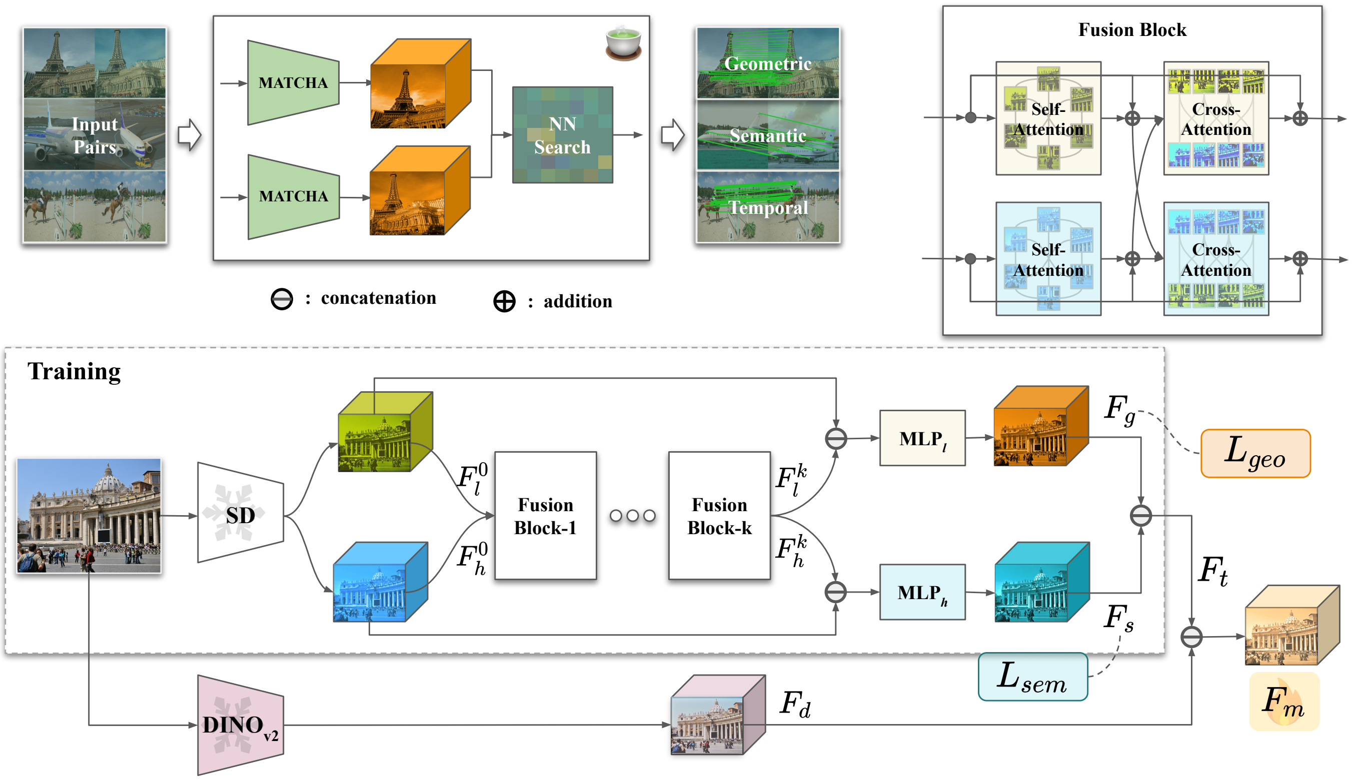

Figure 3. **Architecture of MATCHA.** Given an RGB image, MATCHA produces a single feature for geometric, semantic and temporal matching with nearest neighbor searching. MATCHA is built on top of stable diffusion (SD) models [54] and DINOv2 [46]. Specifically, original geometric and semantic features extracted from SD are first fused dynamically with a transformer [65] consists of self and cross attention blocks. In this dynamic fusion process, both geometric and semantic features are augmented with each other which are supervised with corresponding ground-truth signals in the training process. Then, augmented geometric and semantics features along with DINOv2 feature are unified statically via concatenations into a single feature for *matching anything*.

semantic feature F_d into a single unified feature F_m. We describe the detail of each step in the followings.

Dynamic feature fusion. We adopt the transformer [65] with self and cross attention mechanism for fusion. This strategy allows our model to dynamically gather complementary information from the geometric and semantic descriptors and supervise them jointly in the training process. We patchify both features F_h and F_l with a patch size of p and project their feature dimension to a common feature dimension D_h with a linear layer, which produces the input semantic feature $F_h^0 \in R^{N \times D_h}$ and geometric feature $F_l^0 \in R^{N \times D_h}$ for the fusion stage, where $N = \frac{H}{p*8} \times \frac{W}{p*8}$ is the number of patchified features. The fusion module consists of k self- and cross-attention blocks. For i-th block with $i \in \{1, ..., k\}$, the updating process is as follows:

$$F_{hs}^i = F_h^{i-1} + \mathtt{self}_h^i(F_h^{i-1}), \quad (1)$$

$$F_{ls}^i = F_l^{i-1} + \mathtt{self}_l^i(F_l^{i-1}), \quad (2)$$

$$F_h^i = F_h^{i-1} + \mathtt{cross}_h^i(F_{hs}^i, F_{ls}^i), \quad (3)$$

$$F_l^i = F_l^{i-1} + \mathtt{cross}_l^i(F_{ls}^i, F_{hs}^i), \quad (4)$$

where $\mathtt{self}_h^i$ and $\mathtt{self}_l^i$ are i-th self-attention blocks for F_h and F_l, respectively. $\mathtt{cross}_h^i$ and $\mathtt{cross}_l^i$ are i-th cross-attention blocks for F_h and F_l, respectively. We use the same multi-head attention architecture for each feature branch with non-sharing parameters. Finally, we concatenate the original input features and the fused features along the channel dimension and feed them into a two-layer MLP to output the final semantic feature F_s and geometric feature F_g, defined as:

$$F_s = \mathtt{MLP}_h([F_h^0||F_h^k]), F_g = \mathtt{MLP}_l([F_l^0||F_l^k]), \quad (5)$$

where $[.||.]$ denotes channel-wise concatenation. F_g and F_s are augmented geometric and descriptors and can be used directly for geometric and semantic matching, respectively.

Feature Merging. With the previous preparation of the fusion, we are able to smoothly merge the three features to unify their knowledge. We start by concatenating the enhanced semantic and geometric features, F_s and F_g, to form F_t, which effectively captures both semantic and geometric information within the image. As shown in Tab. 3, this explicit merging, built upon the dynamic fusion process, results in a single feature that significantly outperforms the direct merging of raw DIFT features without fusion enhancement. This demonstrates its superior ability to handle both semantic and geometric information simultaneously. To further boost its matching ability, we equip F_t with the strong semantic cues of DINO-v2 in F_d by another concatenation to obtain the final unified matching feature F_m. Specifically, the two concatenations are defined by:

$$F_t = (F_g || F_s(..., :: d_s)), F_m = (F_t || F_d(..., :: d_t)), \quad (6)$$

where $d_s = \frac{D_s}{D_g}$ and $d_t = \frac{D_d}{D_t}$ are strides adopted to downsample F_s and F_d along the channel dimension. In our experiments, F_g, F_s, and F_d have dimensionalities of 256, 768, and 1024, respectively, leading to the dimensionality of F_m being 1024.

3.3. Supervision

Instead of providing supervision on the final unified feature, we choose to only provide supervision signals to the dynamic fusion enhancement. Ideally, we want to introduce precise signals on each of the tasks directly to our unified feature, which usually requires large-scale annotated data for balanced training across different tasks. However, it is highly expensive to obtain large-scale and accurate correspondence annotations, especially for semantic matching and temporal matching. Therefore, with the limited amount of supervision, we choose to customize the DIFT feature for semantic matching, and support the general semantic understanding from DINOv2 descriptor without further tuning it. Specifically, we apply semantic matching supervision to F_s using CLIP contrastive loss [48] combined with a dense semantic flow loss [32] and geometric matching supervision to F_g using the dual softmax loss function [47]. We provide more information about our supervision losses and training details in the supplementary material.

4. Experiments

We evaluate MATCHA on three matching tasks. We also test a variant of our method, denoted as MATCHA-Light, which evaluates the individual performance of F_s, F_g and F_t (*cf*. Sec. 3.2) on semantic, geometric, and temporal matching tasks. This model is lighter due to no fusion from DINOv2, and follows DIFT by tackling different matching tasks using different descriptors. As we aim to provide a universal descriptor for different matching tasks, we mainly compare MATCHA with previous descriptors in our experiments. More experiments are provided in the supplementary material.

4.1. Semantic Matching

Datasets. Following [5, 34, 73], we use three widely used datasets. SPair-71k [42] contains 12,234 testing pairs split from 70,958 annotated pairs across 18 classes, with diverse scenes and significant viewpoint and scale variation. PF-PASCAL [21] includes 299 testing pairs split from 3547 annotated pairs with similar viewpoints and instance pose. PF-WILLOW [20] contains 900 testing pairs across 4 categories and is used to verify the generalization capability. We evaluate all datasets at an image resolution of 512 × 512.

Method	SM. Sup.	SPair-71k [42] PCK$_{@0.01/0.05/0.1}$(↑)	PF-Pascal [21] PCK$_{@0.05/0.1/0.15}$(↑)	PF-Willow [20] PCK$_{@0.05/0.1/0.15}$(↑)
DINOv2 [46]	✗	6.3 / 38.4 / 53.9	63.0 / 79.2 / 85.1	43.8 / 75.4 / 86.1
⋆DIFT [61]	✗	7.2 / 39.7 / 52.9	66.0 / 81.1 / 87.2	58.1 / 81.2 / -
DIFT	✗	3.1 / 37.9 / 54.3	58.7 / 81.8 / 87.8	55.7 / 85.1 / 92.9
DIFT.Uni + DINO [46, 61]	✗	8.0 / 40.0 / 52.4	61.8 / 78.2 / 85.2	58.7 / 82.9 / 90.7
USC [23]	✗	- / 28.9 / 45.4	-	53.0 / 84.3 / -
SD+DINO [74]	✗	7.9 / 44.7 / 59.9	71.5 / 85.8 / 90.6	-
†GeoASM [73]	✗	9.9 / 49.1 / 65.4	74.0 / 86.2 / 90.7	-
DHF [40]	✓	8.7 / 50.2 / 64.9	78.0 / 90.4 / 94.1	-
*SCorrSAN [24]	✓	3.6 / 36.3 / 55.3	81.5 / 93.3 / 96.6	54.1 / 80.0 / 89.8
*CATs++ [5]	✓	4.3 / 40.7 / 59.8	84.9 / 93.8 / 96.8	56.7 / - / 81.2
*SD4Match [34]	✓	- / 59.5 / 75.5	84.4 / 95.2 / 97.5	56.7 / 80.9 / 91.6
*SD+DINO [74]	✓	9.6 / 57.7 / 74.6	80.9 / 93.6 / 96.9	-
*†GeoASM [73]	✓	**22.0 / 75.3 / 85.6**	**85.9 / 95.7 / 98.0**	-
MATCHA-Light	✓	10.4 / 65.5 / 78.9	82.3 / 93.5 / 96.6	69.0 / 90.1 / 96.2
MATCHA	✓	12.2 / 67.1 / 79.6	79.5 / 93.0 / 96.8	**70.2 / 91.3 / 97.0**

Table 1. **Evaluation on Semantic Matching.** We report PCK under different thresholds. * denotes methods with dataset-specific models and † denotes semantic masks being required. Red indicates methods that compute correlation volume from image pairs while others produce separate descriptors for matching. Both results of DIFT from its original paper [61] (⋆DIFT) and our implementation (DIFT) are included.

Baselines. The baseline methods include those without supervision, *e.g.*, DIFT [61], USC [23], DINOv2 [46] as well as those supervised with GT semantic correspondences, *e.g.*, SD4Match [34] and DHF [40]. We also show results of SD+DINO [74] and GeoASM [73] which provide models with and without supervision. Besides, numbers of two semantic matchers SCorrSAN [24] and CATs++ [5] are also included as a reference.

Metrics. We adopt the standard metric of Percentage of Correct Keypoints (PCK) under different thresholds (0.01/0.05/0.1 for SPair and 0.05/0.1/0.15 for others).

Results. As shown in Tab. 1, both MATCHA and MATCHA-Light surpass all other semantic features except for GeoASM [73] which requires dataset-specific trained models for evaluation and applies task-specific augmentation on top of its baseline SD+DINO [74]. Such test-time augmentation requires masks of the dominant object to flip test images and is not applicable for geometric matching and temporal matching. In contrast, we pursue general improvement in feature representation to better handle matching across various situations using a single feature model (*cf*. Tab. 4). We show that our models stand out on PF-Willow, indicating strong generalization capability.

4.2. Geometric Matching

Datasets. Following prior works [15, 47, 49, 64], HPatches [1] is used to test feature matching performance. We also utilize testing splits [60] of ScanNet [10] and

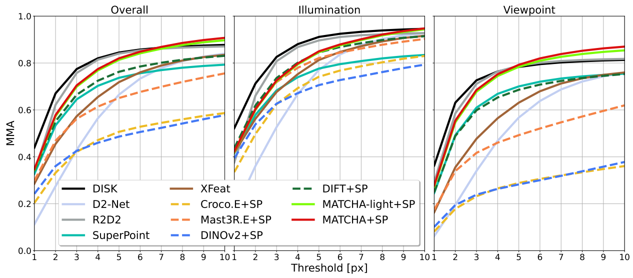

Figure 4. **Geometric Matching on HPatches**. We report Mean Matching Accuracy (MMA) at error thresholds ranging from 1-10 pixel. Concrete and dash lines denote methods with and without supervision, respectively.

Method	GM Sup.	MegaDepth [35]	ScanNet [10] AUC$_{@5/10/20}$(↑)	Aachen [57]
Croco.E [68] + SP	✗	8.0 / 14.7 / 24.2	1.8 / 4.2 / 8.4	11.4 / 18.2 / 26.3
DINOv2 [46] + SP	✗	24.6 / 37.4 / 50.9	2.3 / 5.9 / 12.3	17.2 / 26.1 / 36.4
DIFT [61] + SP	✗	49.7 / 62.8 / 72.8	9.3 / 18.7 / 29.4	43.7 / 53.1 / 61.3
DIFT.Uni [61] + DINOv2 [46] + SP	✗	50.9 / 63.9 / 74.5	9.5 / 19.4 / 30.6	41.9 / 51.3 / 60.0
SP [11]	✓	47.2 / 60.0 / 69.9	6.8 / 14.9 / 24.7	41.6 / 50.2 / 58.1
XFeat [47]	✓	45.4 / 58.9 / 69.3	12.3 / 25.9 / 40.6	36.1 / 45.9 / 55.1
DISK [64]	✓	55.4 / 67.7 / 76.7	6.8 / 14.9 / 24.7	48.9 / 57.5 / 64.6
R2D2 [49]	✓	39.6 / 54.3 / 66.2	5.4 / 11.3 / 19.3	27.6 / 36.4 / 44.1
D2Net [15]	✓	32.5 / 47.7 / 61.4	10.6 / 22.9 / 37.3	30.3 / 41.8 / 52.5
MASt3R.E [33] + SP	✓	37.8 / 51.6 / 63.6	7.4 / 16.8 / 28.5	31.2 / 41.3 / 51.3
MATCHA-Light + SP	✓	**57.1 / 70.9 / 81.2**	**13.0 / 26.6 / 41.8**	<u>51.4 / 60.1 / 67.1</u>
MATCHA + SP	✓	<u>55.8 / 69.3 / 80.0</u>	<u>12.7 / 26.1 / 40.8</u>	**51.7 / 61.0 / 68.5**

Table 2. **Evaluation on Relative Pose Estimation.** We report the AUC values at error thresholds of $5°/10°/20°$ on all datasets.

Megadepth [35] to evaluate the relative pose estimation. We further create randomly selected 1500 pairs of images with large viewpoint and appearance changes from the database of Aachen (Day&Night) v1.0 [57] to validate the generalization ability. These four datasets cover geometric correspondences induced by homography and perspective transformations under indoor, outdoor and planar scenes with moderate to strong viewpoint and illumination changes.

Baselines. Our baselines include local features (*e.g.*, SuperPoint (SP) [11], DISK [64], *etc.*), foundation models (*e.g.*, DINOv2 [46], DIFT [61]) as well as encoders of recent popular geometric foundation models (*e.g.*, Croco.E [68] and MASt3R.E [33]). Following DIFT [61], we use SP to provide keypoints for DINOv2, Croco.E, MASt3R.E and our method MATCHA and MATCHA-Light. We run all methods in the same setting on the original image resolution. We compute matches using nearest neighbour matching with mutual check and estimate relative pose using Poselib [31] with LO-RANSAC [9] as XFeat [47].

Metrics. Following [15, 75], we report the mean matching accuracy (MMA) under 1-10 pixel error thresholds on HPatches and report the area under the curve (AUC) of poses accuracy at error thresholds of $5/10/20$ degrees for relative pose estimation.

Feature matching results. As shown in Fig. 4, both our models have rather close performance on planar scenes, achieving overall the best matching accuracy at bigger thresholds, *e.g.*, above 7px errors. At smaller thresholds, we are only less accurate than DISK and R2D2 which benefit from feature maps at the original image resolution. Note that all other methods including our models use downscaled feature maps ($8\times$ downsampling), but our models give the best accuracy among them.

We observe that supervised methods, *e.g.*, DISK, are much better than methods without supervision, *e.g.*, DIFT, at handling viewpoint variations. This strongly suggests that accurate geometric matching against viewpoints is rather hard to learn from large data without precise correspondence supervisions.

Relative pose estimation results. As shown in Tab. 2, our models achieve the best performance on both indoor and outdoor datasets. While DIFT is the most superior unsupervised method, we are able to drastically increase its AUC score by 6.1-8.4 point and 5.8-7.1 point on outdoor MegaDepth and Aachen, and by 3.7-11.4 point on indoor ScanNet. Additionally, even if trained with a huge number of 3D correspondences, the encoder of MASt3R is significantly less accurate than other feature models that are supervised with much less yet explicit geometric correspondences, *e.g.*, DISK, SP and both of our models. Those observations further validate the importance of a precise supervision direct on descriptors for robust and accurate geometric correspondences.

4.3. Zero-shot Temporal Matching

We further evaluate the zero-shot performance of our models on the challenging temporal matching task.

Datasets. We re-purpose the existing TAPVid dataset [13] to benchmark feature models for temporal matching. TAPVid dataset consists of 30 highly varying real-world video sequences with unknown camera poses, among which some contain highly dynamic objects and extreme camera motions. We perform matching between the first frame and all following frames in each sequence to test the ability of features on handling temporal challenges.

Baselines. We compare our models to previous state-of-the-art geometric (*e.g.*, DISK [64]) and semantic (*e.g.*, DIFT [61]) matching baselines. Rather than using DIFT original feature for temporal matching as in their paper, we instead follow MATCHA-Light to use its concatenated geometric and semantic feature, which leads to better performance. We further consider a hybrid version of DIFT, DIFT.Uni+DINOv2, which combines geometric and semantic DIFT features as well as DINOv2 descriptors as in MATCHA and can be considered as an unsupervised version of MATCHA.

Metrics. As TAPVid provides sparse query points for images, we report the same PCK metric at thresholds of $0.05/0.1/0.15$ as in semantic matching (*cf*. Sec. 4.1).

Results. As shown in Tab. 4, among supervised methods,

Figure 5. **Visualization of temporal matches on TapVID-Davis [13].** We visualize several challenging cases for establishing temporal correspondences, where MATCHA generally achieves the best performance in handling extreme scale and viewpoint changes, as well as scenes with multiple similar instances. DIFT* is the adapted DIFT where we use its concatenated semantic and geometric feature for temporal matching for better performance.

the geometric-matching-only models are generally better on temporal matching than the semantic-matching-only models. However, among unsupervised methods, DINOv2 despite of its poor geometric matching performance (Sec. 4.2) and moderate semantic matching capability (Sec. 4.1), achieves surprisingly superior temporal matching ability. While those two observations seem contradictory, our hypothesis is that DINOv2, benefiting from its large-scale learning on single object-centric data, is able to well handle large viewpoint and scale changes especially when there is a single dominant object in the scene. However, it is poor at handling repetitive structures, and therefore it fails to achieve good geometric matching as well as temporal matching when many similar instances exist. We provide a visual example in Fig. 5 that supports our hypothesis.

Furthermore, MATCHA and DIFT.Uni+DINOv2 standing on top of DINOv2 are significantly better than the other baseline models, which infers that part of the semantic knowledge required for tackling temporal matching is uniquely supported by DINOv2. Finally, MATCHA outperforms DIFT.Uni+DINOv2, indicating that the accurate correspondence supervision signals from semantic and geometric matching provide additional help to improve temporal matching accuracy as well.

Baseline	with DINOv2	Feat. Fusion	Corres. Sup	Desc. Type	Aachen AUC$_{@5/10/20}$($\uparrow$)	PF-Willow PCK$_{@0.05/0.1/0.15}$($\uparrow$)
DIFT	✗	✗	✗	SM	25.6 / 35.6 / 46.3	55.7 / 85.1 / 92.9
DIFT.S	✗	✗	✓	SM	11.5 / 18.6 / 27.7	63.6 / 88.4 / 95.7
MATCHA-Light	✗	✓	✓	SM	21.9 / 31.4 / 41.3	69.0 / 90.6 / 96.2
M1	✓	✓	✓	SM	29.2 / 39.5 / 49.7	**70.3 / 92.4 / 97.6**
DIFT	✗	✗	✗	GM	43.7 / 53.1 / 61.3	26.4 / 40.4 / 50.6
DIFT.S	✗	✗	✓	GM	50.4 / 58.7 / 65.7	32.7 / 46.4 / 55.6
MATCHA-Light	✗	✓	✓	GM	51.4 / 60.1 / 67.1	33.2 / 49.4 / 59.1
M1	✓	✓	✓	GM	**54.0 / 62.7 / 69.8**	53.1 / 76.8 / 85.5
DIFT.Uni	✗	✗	✗	Uni	43.6 / 52.7 / 60.8	26.4 / 40.4 / 50.6
DIFT.Uni + DINO	✓	✗	✗	Uni	41.9 / 51.3 / 60.0	58.7 / 82.9 / 90.7
M2	✗	✗	✓	Uni	50.5 / 58.9 / 65.9	31.8 / 45.6 / 55.4
M3	✗	✓	✓	Uni	50.0 / 59.0 / 66.5	60.8 / 82.8 / 90.4
M4	✓	✗	✓	Uni	53.0 / 61.8 / 69.0	53.9 / 78.1 / 88.2
MATCHA	✓	✓	✓	Uni	51.7 / 61.0 / 68.5	70.2 / 91.3 / 97.0

Table 3. **MATCHA Ablation Study.** We ablate different components of proposed model on Aachen [57] for geometric matching and PF-Willow [20] for semantic matching using the same metrics defined in the previous sections. We denote their descriptor types using **SM/GM/Uni** that stand for semantic/geometric/unified features. We use green cells for evaluations on a supervised matching task and gray on zero-shot matching tasks.

4.4. Ablations

We perform ablation studies on Aachen [57] and PF-Willow [20] for geometric and semantic matching, respectively. In Tab. 3, we present intermediate variants that evolve from DIFT baseline towards our final MATCHA. We focus on studying the impact of four design choices: (i) correspondence supervision, (ii) feature fusion between semantic and geometric features, (iii) leveraging DINOv2 and (iv) using separate semantic (SM) and geometric (GM) descriptors versus a unified (Uni) feature for both tasks. The baselines include M1 (DIFT + DINOv2 + feature fusion + supervision), M2 (DIFT + supervision + unified descriptor), M3 (DIFT + feature fusion + unified descriptor), and M4 (DIFT + DINOv2 + supervision + unified descriptor)

Impact of correspondence supervisions. We add the same number of self-attention layers (as in MATCHA) to process the original DIFT semantic and geometric descriptors and supervise them accordingly using the same semantic and geometric supervisions individually. We name this variant DIFT.S. As shown in Tab. 3, the geometric supervision leads to improved performance both on geometric and semantic matching, verifying that a general improvement in matching capability was gained with geometric supervision. While supervised semantic DIFT descriptor also shows clear improvement on semantic matching, it leads to worse geometric matching performance, indicating the loss of generalization capability in its feature potentially due to the limited semantic matching data.

Impact of dynamic feature fusion. After turning on our proposed fusion module (*cf*. Sec. 3.2), MATCHA-Light is able to further improve the accuracy on top of DIFT.S when being evaluated on both supervised and unsupervised semantic and geometric matching tasks. While semantic and geometric features contain information to support each other, it is not trivial to extract and fuse them to realize the

Method	Single Desc	Corres. Sup.	*Geometric* Aachen AUC$_{@5/10/20}$(↑)	Avg(↑)	*Semantic* PF-Willow PCK$_{@0.05/0.1/0.15}$(↑)	Avg(↑)	*Temporal* TapVid-Davis PCK$_{@0.05/0.1/0.15}$(↑)	Avg(↑)	Average Score(↑)
DISK [64]	✓	GM	48.9 / 57.5 / 64.6	57.0	10.2 / 17.0 / 23.1	16.8	57.0 / 61.7 / 65.0	61.2	45.0
DeDoDe-G [17]	✓	GM	62.4 / 70.6 / 76.8	69.9	27.1 / 42.2 / 51.7	40.3	80.6 / 85.8 / 88.4	84.9	65.1
XFeat [47]	✓	GM	36.1 / 45.9 / 55.1	45.7	25.7 / 40.0 / 48.8	38.2	63.3 / 71.4 / 77.1	70.6	51.5
MASt3R.E [33]	✓	GM	31.2 / 41.3 / 51.3	41.3	24.0 / 42.1 / 54.7	40.3	75.2 / 83.8 / 87.9	82.3	54.6
DIFT [61]	✗	✗	43.7 / 53.1 / 61.3	52.7	55.7 / 85.1 / 92.9	77.9	79.7 / 86.7 / 90.5	85.6	72.1
MATCHA-Light	✗	GM+SM	51.4 / 60.1 / 67.1	<u>59.5</u>	69.0 / 90.6 / 96.2	<u>85.3</u>	78.7 / 86.3 / 90.2	85.1	<u>76.6</u>
DINOv2 [46]	✓	✗	17.2 / 26.1 / 36.4	26.6	43.8 / 75.4 / 86.1	68.4	83.2 / 89.7 / 92.0	88.3	61.1
DIFT.Uni +DINOv2	✓	✗	41.9 / 51.3 / 60.0	51.1	58.7 / 82.9 / 90.7	77.4	86.4 / 91.6 / 93.5	<u>90.5</u>	73.0
MATCHA	✓	GM+SM	**51.7 / 61.0 / 68.5**	**60.4**	**70.2 / 91.3 / 97.0**	**86.2**	**87.8 / 93.5 / 95.5**	**92.3**	**79.6**

Table 4. **Towards Matching Anything with A Unified Feature.** We compare ourselves to various feature models across geometric, semantic and temporal matching and compute the ranking of each method for each task and averaged over tasks. We show that MATCHA achieves the topk averaged ranking among all types of methods using a single feature for matching anything.

mutual helping goal. For example, naively concatenating DIFT semantic and geometric features as in DIFT.Uni, or DIFT supervised features as in M2, both lead to a big drop in semantic matching performance compared to using those feature individually. In contrast, we show that with the help of feature fusion, semantic and geometric features not only improve themselves as in MATCHA-Light, but also become more cooperative and consistent with each other when being concatenated as in M3. The above experiments fully demonstrate that our proposed feature fusion module enables effective extraction and fusion of helpful information from the semantic and geometric features into each other, leading to enhanced feature matching accuracy.

Role of DINOv2. As shown in Tab. 3, M1, DIFT.Uni+DINO, MATCHA building on top of DINOv2, achieve constant improvement on both geometric and semantic matching performance compared to their baselines MATCHA-Light, DIFT.Uni and M3. Such conclusion is consistent with our discussion in Sec. 4.3, showing that DINOv2 provides interesting complementary knowledge to DIFT as well as our supervised MATCHA-Light, to significantly boost their general matching capabilities.

A unified feature. As shown in the upper two parts of Tab. 3, using only the semantic or geometric descriptor, it is hard to achieve a good performance on both tasks. Among those, M1 geometric descriptor is the most promising feature that achieves the best geometric matching performance with proper generalization on semantic matching. However, unifying the semantic and geometric feature of M1 into one as in MATCHA largely improves its performance on semantic matching, with a slight drop at geometric matching accuracy, achieving the best balance between the two matching tasks. We further evaluate MATCHA in the next section towards our end goal.

4.5. Towards Matching Anything

Keeping multiple versions of descriptors for an image is not effective in general. Therefore, we aim at pursuing a foundation feature model that produces a single descriptor that is designed for *matching anything*. In this section, we thoroughly evaluate the state-of-the-art feature models across the three matching tasks, *i.e.*, geometric, semantic, and temporal matching. As shown in Tab. 4, geometric features are not able to perform semantic matching well and have limited generalization ability on temporal matching. While the unsupervised foundation feature DIFT shows promising matching capability generalizing across three tasks, it requires different descriptors to handle different tasks and has a clear gap compared to task-specific best performing models. MATCHA, building on top of the feature knowledge learned in DIFT and DINOv2, further enhanced with precise correspondence supervision and supported by a careful fusion mechanism, for the first time, outperforms all other methods across all tasks, using only a single feature.

5. Conclusion

In this work, we introduce a new vision challenge: achieving *match-anything* capability with a single, unified feature representation. We propose MATCHA, a novel feature model that harnesses existing correspondence supervision resources to narrow the accuracy gap between foundational features and task-specific supervised methods, while preserving generalization across diverse correspondence tasks. By incorporating limited, high-quality supervision, we take a significant step toward eliminating the need for task-specific feature descriptors, moving closer to universal matching features. This approach has direct implications for applications relying on robust correspondence, including 3D reconstruction, tracking and localization, image retrieval, and image editing.

Limitations. Our experiments reveal that while features derived from foundation models capture rich information, they still face challenges in resolution precision for fine-grained geometric matching and are often not optimized for runtime efficiency. We encourage future work to address these limitations for broader applicability.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017. 5

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3

[4] 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 Nature Switzerland Cham, 2022. 2

[5] Seokju Cho, Sunghwan Hong, and Seungryong Kim. Cats++: Boosting cost aggregation with convolutions and transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7174–7194, 2022. 3, 5

[6] Seokju Cho, Jiahui Huang, Seungryong Kim, and Joon-Young Lee. Flowtrack: Revisiting optical flow for long-range dense tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19268–19277, 2024. 3

[7] Seokju Cho, Jiahui Huang, Jisu Nam, Honggyu An, Seungryong Kim, and Joon-Young Lee. Local all-pair correspondence for point tracking. *arXiv preprint arXiv:2407.15420*, 2024. 1, 3

[8] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *Advances in neural information processing systems*, 29, 2016. 2

[9] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. In *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*, pages 236–243. Springer, 2003. 6

[10] 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. 5, 6

[11] 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, 6

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2, 3

[13] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022. 1, 2, 3, 6, 7

[14] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023. 3

[15] 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, 5, 6

[16] 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

[17] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Dedode: Detect, don't describe—describe, don't detect for local feature matching. In *2024 International Conference on 3D Vision (3DV)*, pages 148–157. IEEE, 2024. 8

[18] 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

[19] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2dnet: Learning accurate correspondences for sparse-to-dense feature matching. *arXiv preprint arXiv:2004.01673*, 2020. 2

[20] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3475–3484, 2016. 5, 7

[21] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1711–1725, 2017. 5

[22] Adam W. Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *ECCV*, 2022. 1, 3

[23] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 5

[24] Shuaiyi Huang, Luyu Yang, Bo He, Songyang Zhang, Xuming He, and Abhinav Shrivastava. Learning semantic correspondence with sparse annotations. In *European Conference on Computer Vision*, pages 267–284. Springer, 2022. 3, 5

[25] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, et al. Openclip, 2021. 3

[26] 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

[27] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 1, 3

[28] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2307–2314, 2013. 3

[29] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6560–6569, 2017. 1, 2

[30] Seungryong Kim, Dongbo Min, Somi Jeong, Sunok Kim, Sangryul Jeon, and Kwanghoon Sohn. Semantic attribute matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12339–12348, 2019. 3

[31] Viktor Larsson. Poselib-minimal solvers for camera pose estimation, 2020. 6

[32] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. Sfnet: Learning object-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2278–2287, 2019. 3, 5

[33] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *European Conference on Computer Vision*, 2024. 6, 8

[34] 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. 1, 2, 3, 5

[35] 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. 6

[36] 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. 2

[37] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010. 3

[38] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 1, 2

[39] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 3

[40] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 5

[41] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Contextdesc: Local descriptor augmentation with cross-modality context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2527–2536, 2019. 2

[42] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019. 5

[43] Marius Muja and David G Lowe. Scalable nearest neighbor algorithms for high dimensional data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2227–2240, 2014. 2

[44] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 1

[45] Dolev Ofri-Amar, Michal Geyer, Yoni Kasten, and Tali Dekel. Neural congealing: Aligning images to a joint semantic atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19403–19412, 2023. 1

[46] 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. 2, 3, 4, 5, 6, 8

[47] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024. 1, 5, 6, 8

[48] 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, 5

[49] 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. 1, 2, 5, 6

[50] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017. 3

[51] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6917–6925, 2018. 3

[52] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31, 2018. 2, 3

[53] 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. 3

[54] 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, 2, 3, 4

[55] Peter Sand and Seth Teller. Particle video: Long-range motion estimation using point trajectories. *International journal of computer vision*, 80:72–91, 2008. 1, 3

[56] 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

[57] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8601–8610, 2018. 6, 7

[58] 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

[59] Jenny Seidenschwarz, Qunjie Zhou, Bardienus Duisterhof, Deva Ramanan, and Laura Leal-Taixé. Dynomo: Online point tracking by dynamic online monocular gaussian reconstruction. *arXiv preprint arXiv:2409.02104*, 2024. 3

[60] 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, 5

[61] 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. 1, 2, 3, 5, 6, 8

[62] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11016–11025, 2019. 2

[63] Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6258–6268, 2020. 3

[64] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 1, 2, 5, 6, 8

[65] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[66] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *International Conference on Computer Vision*, 2023. 3

[67] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *arXiv preprint arXiv:2407.13764*, 2024. 3

[68] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 6

[69] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *CVPR*, 2024. 3

[70] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14880–14890, 2022. 1

[71] 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. 2

[72] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 467–483. Springer International Publishing, 2016. 2

[73] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3076–3085, 2024. 1, 2, 5

[74] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 5

[75] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4669–4678, 2021. 2, 6