Telling Left from Right: Identifying Geometry-Aware Semantic Correspondence

Junyi Zhang[†] Charles Herrmann[‡] Junhwa Hur[‡] Eric Chen[§]
Varun Jampani[¶] Deqing Sun[‡ *] Ming-Hsuan Yang[‡,§ *]
[†]Shanghai Jiao Tong University [‡]Google Research [§]UIUC [¶]Stability AI [§]UC Merced

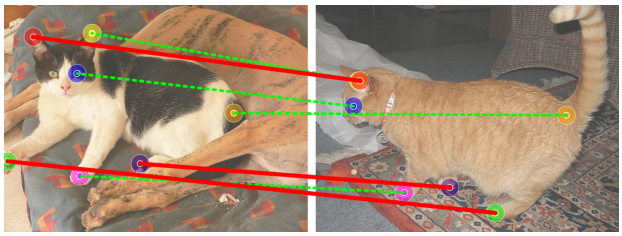
Abstract

*While pre-trained large-scale vision models have shown significant promise for semantic correspondence, their features often struggle to grasp the geometry and orientation of instances. This paper identifies the importance of being geometry-aware for semantic correspondence and reveals a limitation of the features of current foundation models under simple post-processing. We show that incorporating this information can markedly enhance semantic correspondence performance with simple but effective solutions in both zero-shot and supervised settings. We also construct a new challenging benchmark for semantic correspondence built from an existing animal pose estimation dataset, for both pre-training validating models. Our method achieves a PCK@0.10 score of **65.4** (zero-shot) and **85.6** (supervised) on the challenging SPair-71k dataset, surpassing the state of the art by 5.5p and 11.0p absolute gains, respectively. Our code and datasets are publicly available at: https://telling-left-from-right.github.io.*

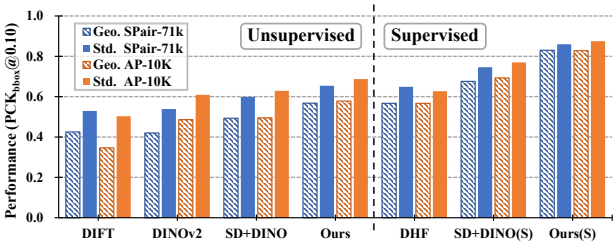
1. Introduction

Since the advent of high fidelity text-to-image (T2I) generative models [40, 41] and large vision foundation models [36], there has been significant interest in understanding both what these models are learning and what they are not. Numerous works show that these models have powerful feature embeddings that can be used for many computer vision tasks including depth estimation [36, 62], semantic segmentation [49, 56], and semantic correspondences [2, 10, 14, 18, 30, 35, 61]. While many works have shown their strengths, less analysis has been done on their weaknesses; in particular, what do these features struggle with?

We propose using semantic correspondence as a promising test bed. Semantic correspondence, the establishment of pixel-level matches between two images with semantically similar objects, is an important Computer Vision problem with a variety of downstream applications, *e.g.*, image editing [10, 34, 35, 61] and style transfer [9, 24]. It also has many difficult challenges, *e.g.*, from large intra-class variations to different backgrounds, lighting, or viewpoints.



(a) The state-of-the-art method [61] fails at matching keypoints with geometric ambiguity, or, "telling left from right" (red solid lines).



(b) **The performance gap between geometry-aware set (Geo.) and standard set (Std.) of state-of-the-art methods.** The geometry-aware set accounts for **59.6%** and **45.7%** of the total keypoint pairs on SPair-71k [32] and AP-10K [60], respectively.

Figure 1. Illustration of geometry-aware correspondence.

Despite these challenges, the large foundation model features currently achieve state-of-the-art performance [61]. However, a closer examination shows that this performance is inconsistent across all challenges. In particular, we find that these foundation model's features significantly underperform on "geometry-aware"[1] semantic correspondences: correspondences which share semantic properties but have different relations to the overall geometry of the object, *e.g.*, the "left" paw vs. the "right" paw as shown in Fig. 1a. Motivated by this, we conduct an in-depth analysis of these correspondences (Fig. 1b). We find that surprisingly, such cases account for a significant portion of the benchmark datasets (nearly 60% in SPair71k), and state-of-the-art methods with the deep features perform considerably worse on this challenging subset (up to 30% worse, Fig. 1b).

[*]Equal contribution.

[1]We use the term geometry in the loose sense and do not refer to 3D geometric properties, such as shape and surface normal.

Are these problems an innate failing of these features, or can they be alleviated through better post-processing? Based on the above observations, we propose several methods that resolve the geometric ambiguity during matching. First, we introduce a test-time viewpoint alignment strategy that approximately aligns viewpoints of instances to make the problem easier. Then we train a lightweight post-processing module that improves geometric awareness of features from visual foundation models [36, 40], by using a soft-argmax based dense training objective with given annotated sparse keypoints. We further introduce a pose-variant augmentation strategy as well as a window soft-argmax module. These not only significantly improve performance on standard benchmarks by 15% while costing only 0.32% of extra runtime.

For more advanced analysis, we create a new benchmark dataset using existing annotations from the AP-10K [60] animal pose estimation dataset. Compared to the largest existing benchmark [32], our new benchmark dataset includes 5 times more training pairs and also evaluates cross-species and cross-families semantic correspondence in addition to intra-species correspondence. We also demonstrate that this benchmark can serve as a valuable pre-training resource for improving geometry-aware semantic correspondence.

To summarize, we make the following contributions:

- We identify the problem of geometry-aware semantic correspondence and show that pre-trained features of foundation models (SD [40] and DINOv2 [36]) struggle with geometric information.
- We propose to improve geometric awareness of the features in both unsupervised and supervised manners.
- We introduce a large-scale and challenging benchmark, AP-10K, for both training and evaluation.
- Our method boosts the overall performance on multiple benchmark datasets, especially on the geometry-aware correspondence subset. It achieves an 85.6 PCK@0.10 score on SPair-71k, outperforming the state-of-the-art method by more than 15%.

2. Related Work

Semantic correspondence. Conventional approaches to semantic correspondence estimation follow a common pipeline that consists of i) feature extraction [1, 8, 13, 27, 29, 44]), ii) cost volume computation [6, 15, 23, 33], and iii) matching field regression [21, 50–53]. To handle challenging intra-class variations between images, previous work have presented various approaches such as matching uniqueness prior [26], parameterized spatial prior [16, 20, 38, 39, 42, 57], or end-to-end regression [6, 7, 16, 25, 51]. A few previous work have also explored semantic correspondence under unsupervised [10, 35, 43, 48] and weakly supervised [17, 37, 54] setting, by densely image aligning [10, 35, 37, 48] or automatic label generation [17, 43].

Figure 2. Generated samples from SD-2-1 with the prompt (left) "A cat holding up its *left front paw*" and (right) "A car with the *right front door* open". SD has difficulty generating images that require understanding the intrinsic geometry of instances.

However, due to the limited capacity of their features or the usage of strong spatial prior, they still exhibit difficulties handling challenging intra-class variations such as large pose changes or non-rigid deformation.

Recently, visual foundation models (*e.g.* DINO [5, 36] and SD [40]) demonstrate that their pretrained features, learned by self-supervised learning or generative tasks [2, 14, 30, 46, 61], can serve as powerful descriptors for semantic matching by surpassing prior arts specifically designed for semantic matching. Yet, we reveal that such models still show limitations [10] in comprehending the intrinsic geometry of instances (*e.g.*, Fig. 2) and formally investigate this issue, termed "geometry-aware" semantic correspondence.

Benchmark datasets. Recent advances in semantic correspondence have continuously revealed limitations of existing benchmark datasets. For example, widely used datasets (PF-Pascal [11], PF-Willow [12], CUB-200-2011 [55], and TSS [47]) provide image pairs with only limited viewpoints or pose variations, making it hard to evaluate methods on handling large object viewpoint changes. The CUB dataset [55] provides images of a single object class, bird, only. SPair-71k [32] introduces a more challenging benchmark dataset that consists of 1,800 images across 18 object categories with substantial intra-class variations. Recently, Aygün and Mac Aodha [3] leveraged an animal pose dataset, Awa-Pose [4], to create 10k image pairs for evaluating the inter-class semantic correspondence. While existing methods [3, 7, 30, 61] have low performance on the SPair-71k and Awa-Pose, these benchmarks are still small-scale. To address these shortcomings, we introduce a new, large-scale, and challenging benchmark using the animal pose estimation dataset, AP-10K [60]. This new benchmark further facilitates comprehensive evaluations of geometric awareness and provides annotations for training models.

3. Geometric Awareness of Deep Features

In this section, we first provide the clear problem definition of "geometry-aware semantic correspondence" as challenging cases of semantic correspondence, which requires an understanding of relations of similar semantic parts. Then we provide comprehensive analyses on the performance of pretrained features of foundation models on the problem and what geometric information those features possess.

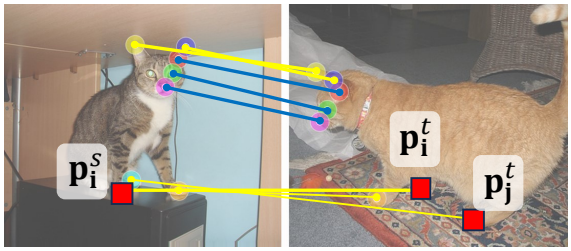


Figure 3. Annotations of geometry-aware semantic correspondence (yellow) and standard semantic correspondence (blue).

3.1. Geometry-Aware Semantic Correspondence

We define geometry-aware semantic correspondence as a challenging case of semantic correspondence, where there exist geometry-ambiguous matching cases, and thus it requires an understanding of instances' orientations or geometry. Fig. 1a illustrates the exemplar cases that require proper understanding of both semantic parts (*i.e.* paws) and their associations (*i.e.* left paw or right paw) with the orientation of instances.

As a formal definition, for each instance category, we first cluster keypoints into subgroups $\mathcal{G}_{parts}$ by their semantic parts. Each subgroup $\mathcal{G}_{parts}$ consists of a set of keypoints $\mathbf{p}_{(parts,\ index)}$ that fall into the same subgroup but position in different part locations according to their orientations. For the *cat* category as an example, the subgroups are $\mathcal{G}_{parts}$ with $parts$ = {ears, paws, ... }, and $\mathcal{G}_{\text{paws}} = \{\mathbf{p}_{(\text{paws, front left})}, \mathbf{p}_{(\text{paws, front right})}, \mathbf{p}_{(\text{paws, rear left})}, \mathbf{p}_{(\text{paws, rear right})}\}$.

Then, give a source $\mathbf{I}^s$ and a target image $\mathbf{I}^t$ that contains the same/similar instance category with their keypoint correspondence annotations, the correspondence $\langle \mathbf{p}_\mathbf{i}^s, \mathbf{p}_\mathbf{i}^t \rangle$ is considered as a "geometry-aware" correspondence if the two conditions are met. First, two keypoints $\mathbf{p}_\mathbf{i}^s$ and $\mathbf{p}_\mathbf{i}^t$ belong to the same subgroup, $\mathbf{p}_\mathbf{i}^s \in \mathcal{G}_{part}^s$ and $\mathbf{p}_\mathbf{i}^t \in \mathcal{G}_{part}^t$. Second, there are other visible keypoint(s) belonging to same subgroup in the target image, $\exists\, \mathbf{j} \neq \mathbf{i}$ s.t. $\mathbf{p}_\mathbf{j}^t \in \mathcal{G}_{part}^t$. As illustrated in Fig. 3, the front right paw ($\mathbf{p}_\mathbf{i}^s$) of the cat in the source image has several semantically similar correspondences, such as ($\mathbf{p}_\mathbf{j}^t$) and ($\mathbf{p}_\mathbf{i}^t$), which requires proper understanding of geometry to find the correct match.

3.2. Evaluation on the Geometry-aware Subset

We evaluate the state-of-the-art methods on geometry-aware semantic correspondence to see if their features are geometry-aware and how well they perform on this challenging task. From the challenging SPair-71k [32] datasets, we first cluster keypoint subgroups $\mathcal{G}_{parts}$ for each category and gather geometry-aware correspondence cases as the "geometry-aware subset". Surprisingly such cases account for a substantial portion, 82.4% of total image pairs and 59.6% of matching keypoints, of the dataset. (Please refer to Supp. C for more details.)

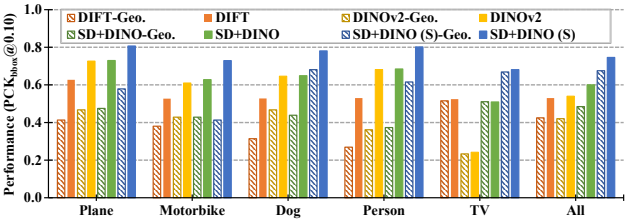


Figure 4. **Per-category evaluation of state-of-the-art methods on SPair-71k geometry-aware subset (Geo.) and standard set.** While the geometry-aware subset accounts for 60% of the total matching keypoints, we observe a substantial performance gap between the two sets for all the methods.

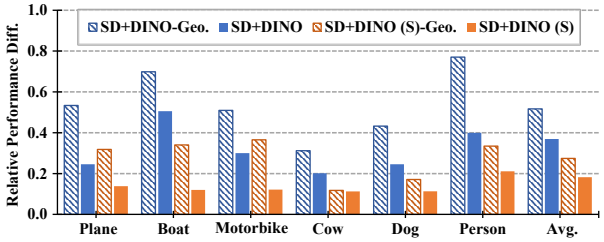


Figure 5. **Evaluation of the sensitivity to pose variations.** The y-axis shows the normalized difference between the best and the worst performance among 5 different azimuth-variation subsets. We report the results of the unsupervised and supervised methods on both the geometry-aware (Geo.) and standard set. The larger the value, the more sensitive the performance is to pose variation.

Fig. 4 shows the performance of the state-of-the-arts on the subset (in both zero-shot and supervised (S)). For all methods, there exists a substantial performance gap between the geometry-aware subset and the standard set, around 20% for zero-shot methods and still 10% for supervised methods. This reveals the weakness of current methods in matching keypoints where the geometry ambiguity is involved and the limitation on geometric awareness.

3.3. Sensitivity to Pose Variation

For certain categories, however, where the pose variation of the pair images is small (*e.g.*, potted plant and TV in SPair-71k), performance gaps on both the standard and geometry-aware sets are nearly marginal. This suggests that the pose variation is one of the key factors that affect the accuracy of geometry-aware correspondence. To delve deeper into it, we divide image pairs in SPair-71k into 5 subsets, based on their annotated azimuth differences, ranging from 0 (identical poses) to 4 (completely opposing directions). For each category, we then again evaluate the performance on these 5 subsets, $\mathcal{A} = \{\mathbf{a_0}, \mathbf{a_1}, \ldots, \mathbf{a_4}\}$ and define the normalized relative difference, $\mathbf{d} = \frac{\max(\mathcal{A}) - \min(\mathcal{A})}{\max(\mathcal{A})}$, which measures the sensitivity to pose variations. As shown in Fig. 5, the performance on the geometry-aware subset is more sensitive to the pose variation than the standard set across all categories, indicating that the pose variation affects the performance on geometry-aware semantic correspondence.

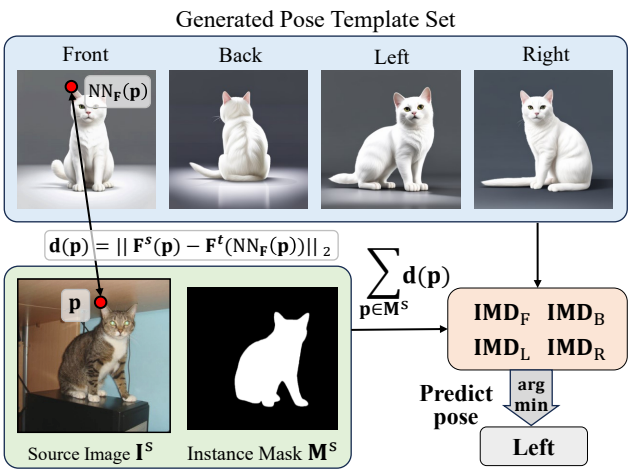


Figure 6. **Rough pose prediction with feature distance.** By computing the instance matching distance (IMD) of the source image to the generated pose templates, we can utilize the feature maps to predict rough pose and evaluate the global pose awareness of current deep features. We only show one template set for brevity.

3.4. Global Pose Awareness of Deep Features

We further analyze if deep features are aware of high-level pose (or viewpoint) information of an instance in an image. We explore this pose awareness by a template-matching approach in the feature space.

Instance matching distance (IMD). We introduce this metric to examine pose prediction accuracy. Given a source $\mathbf{I}^s$ and target image $\mathbf{I}^t$, their normalized feature maps $\mathbf{F}^s$ and $\mathbf{F}^t$, and a source instance mask $\mathbf{M}^s$, we define the IMD metric as:

$$\text{IMD}(\mathbf{I}^s, \mathbf{I}^t, \mathbf{M}^s) = \sum_{\mathbf{p} \in \mathbf{M}^s} \|\mathbf{F}^s(\mathbf{p}) - \text{NN}(\mathbf{F}^s(\mathbf{p}), \mathbf{F}^t)\|_2, \quad (1)$$

where $\mathbf{p}$ denotes a pixel within the source instance mask, $\mathbf{F}^s(\mathbf{p})$ is the feature vector at $\mathbf{p}$, and $\text{NN}(\mathbf{F}^s(\mathbf{p}), \mathbf{F}^t)$ represents the nearest-neighboring feature vector in the target feature map. IMD measures the similarity of two images via the average feature distance of corresponding pixels.

Pose prediction via IMD. With the IMD metric, we can evaluate the global pose awareness of features from existing methods via pose prediction. We start by generating multiple pose template sets (in Fig. 6). We then compute the IMD between the input and template images for each set and predict the pose whose IMD is the smallest. A collective vote across all sets determines the final pose estimate.

We manually annotated 100 cat images from SPair-71k with pose labels {left, right, front, and back} and evaluate the pose prediction performance of the following deep features: DINOv2 [36], SD [40], and fused SD+DINO [61]. Due to some ambiguous cases for annotations, we also report the performance of binary classification into {left, right} or {front, back}. As in Tab. 1, DINOv2 struggles with left/right (L/R) distinction [10] but excels in front/back (F/B) prediction; SD performs well in both distinguishing L/R and F/B; and SD+DINO surpass both in all cases, achieving near-perfect results. This suggests that the deep features are aware of global pose information.

Table 1. **Zero-shot rough pose prediction result with IDM (Eq. (1)).** We report the accuracy of predicting left or right (L/R), front or back (F/B), the former two cases (L/R or F/B), and one of the four directions (L/R/F/B).

Feature	L/R	F/B	L/R or F/B	L/R/F/B
DINOv2	63.8	100.0	75.0	51.0
SD	95.7	96.8	96.0	78.0
SD+DINO	98.6	100.0	99.0	84.0

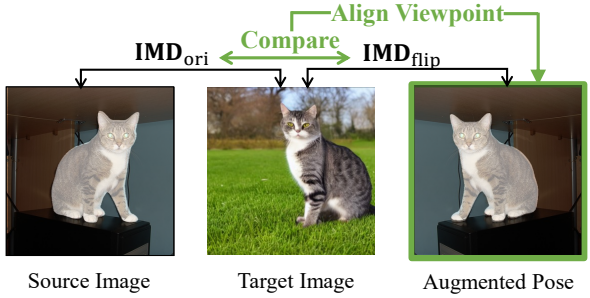


Figure 7. **Adaptive pose alignment.** By comparing the matching distance between the target image and augments of the source image, we can reduce the pose variation of pair images at test time for better correspondence.

4. Improving Geo-Aware Correspondence

We propose several techniques that improve geometric awareness during matching, in both zero-shot and supervised settings. We first introduce an adaptive pose alignment strategy that runs at test time without any training involved. Then, we further introduce a post-processing module with various training strategies that can improve the geometry awareness of deep features.

4.1. Test-time Adaptive Pose Alignment

In Sec. 3.3, we find that pose variations can largely affect the performance of geometry-aware semantic correspondence. To address this, we introduce a very simple test-time pose alignment strategy that utilizes the global pose information inherent in deep features (Sec. 3.4) and improves correspondence accuracy.

As in Fig. 7, we first augment the source image by using a set of pose-variant augmentations (*e.g.*, flip, rotations *etc.*), calculate the IMD (Eq. (1)) between the augmented source images and the target image, and choose the optimal pose with the minimum IMD distance.[2] As in Fig. 9, this simple pose alignment can drastically improve the correspondence accuracy in a test-time, unsupervised manner.

[2]Refer to Supp. E.1 for alternative metrics that does not require mask.

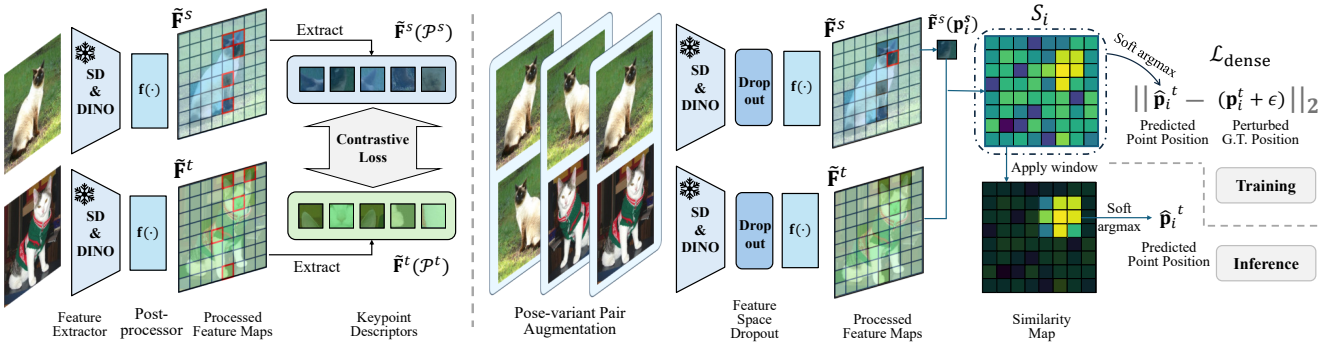


Figure 8. (Left) previous supervised methods [30, 61] with a sparse training objective. (Right) an overview of our supervised method. Only the lightweight post-processor is updated during training. Both the pair augmentation and feature space Dropout are for training only.

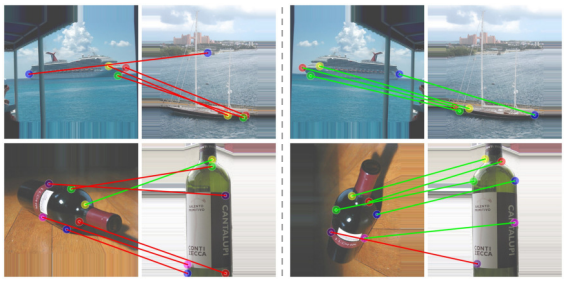


Figure 9. (Left) original image pairs. (Right) image pairs with the test-time aligned pose. The reduced pose variation improves the correspondence accuracy.

4.2. Dense Training Objective

Let $\mathbf{F}$ represent the raw feature map and $\mathrm{f}(\cdot)$ be the post-processing model that outputs the refined feature map $\tilde{\mathbf{F}} = \mathrm{f}(\mathbf{F})$, illustrated in Fig. 8. Given a set of annotated keypoint pairs from source images $\mathcal{P}^s = \{\mathbf{p}_1^s, \mathbf{p}_2^s, \ldots, \mathbf{p}_n^s\}$ and target images $\mathcal{P}^t = \{\mathbf{p}_1^t, \mathbf{p}_2^t, \ldots, \mathbf{p}_n^t\}$, previous works with pretrained foundation model features [30, 61] adopt a CLIP-style symmetric contrastive loss $\mathrm{CL}(\cdot, \cdot)$ to train the post-processing model:

$$\mathcal{L}_{\text{sparse}} = \mathrm{CL}(\tilde{\mathbf{F}}^s(\mathcal{P}^s), \tilde{\mathbf{F}}^t(\mathcal{P}^t)), \tag{2}$$

where $\tilde{\mathbf{F}}^s$ and $\tilde{\mathbf{F}}^t$ are the post-processed source and target features. However, the loss is applied only to features with sparsely annotated keypoints, which potentially neglects additional informative features.

Instead, we employ the soft-argmax operator [19, 23, 58] so that gradients calculated from sparse annotations can be back-propagated to all spatial locations. Specifically, we compute the similarity map $S_i = \tilde{\mathbf{F}}^s(\mathbf{p}_i^s)^T \tilde{\mathbf{F}}^t$ between the normalized query keypoint descriptor $\tilde{\mathbf{F}}^s(\mathbf{p}_i^s)$ and the target feature map $\tilde{\mathbf{F}}^t$. Then, we take soft-argmax over the similarity map to get the predicted position $\hat{\mathbf{p}}_i^t = \mathrm{SoftArgmax}(S_i)$. The L2 norm penalizes the distance between the predicted position and the target position $\hat{\mathbf{p}}_i^t$:

$$\mathcal{L}_{\text{dense}} = \sum_i \|\hat{\mathbf{p}}_i^t - (\mathbf{p}_i^t + \epsilon)\|_2, \tag{3}$$

To prevent overfitting, we also apply Dropout at the input feature map $\mathbf{F}$ and Gaussian noise ϵ that perturbs the ground truth keypoint positions $\mathbf{p}_i^t$. We empirically find that combining the two objectives achieves better performance; thus our final training objective is $\mathcal{L} = \mathcal{L}_{\text{dense}} + \mathcal{L}_{\text{sparse}}$.

4.3. Pose-variant Augmentation

Standard data augmentation schemes (*e.g.*, random cropping, color jittering, *etc.*) have been generally used to augment the limited number of annotated data. However, such standard augmentations show two shortcomings in naively adopting them to our approach. Diverse augmentations on input images require our model to process the feature map of each augmented image using visual foundation models, which linearly increases the computational cost along with the number of augmentation schemes used. Besides, such augmentations (*e.g.*, cropping, scaling, or photometric augmentations) do not augment images with different poses or viewpoints, which might not bring additional effective supervision signals for geometric awareness.

Instead, we introduce a set of pose-varying augmentation schemes tailored to our approach, which needs to process only one feature map from a single additional augmented image (horizontal flipped) yet can utilize the feature in multiple ways. The motivation is that the deep features are aware of the global pose; thus, the processed feature map of the flipped image can add an additional signal; compared to simply flipping the feature map. We introduce the following three augmentation settings: 1) *double flip*: flipped source image and flipped target image; 2) *single flip*: flipped source image and original target image; and 3) *self flip*: source image and flipped source image. For setting 2 and 3, keypoint annotations are correspondingly flipped to preserve the inherent geometric concept, *e.g.*, the left paw in the flipped image should be the right paw of the original image. The keypoint flipping in *self flip* also ensures that the model learns to discern concepts rather than simply matching keypoints based on appearances.

4.4. Window Soft Argmax

At test time, current methods [30, 61] use the argmax operation on the similarity map to infer correspondence. How-

Table 2. **Evaluation on SPair-71k.** Per-class and average PCK@0.10 on test split. The methods are categorized into two types: supervised (S) and unsupervised (U). †: index is used to flip source keypoints at test time. ∗: fine-tuned backbone. We report *per point* PCK result for the (U) methods, following [10, 35], and *per image* result for the (S) methods, following [7, 16, 25, 26]. The highest PCK are highlighted in **bold**, while the second highest are underlined. Both our zero-shot and supervised methods outperform prior arts across all categories.

	Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	**All**
U	ASIC [10]	57.9	25.2	68.1	24.7	35.4	28.4	30.9	54.8	21.6	45.0	47.2	39.9	26.2	48.8	14.5	24.5	49.0	24.6	36.9
	DINOv2+NN [36, 61]	72.7	62.0	85.2	41.3	40.4	52.3	51.5	71.1	36.2	67.1	64.6	67.6	61.0	68.2	30.7	62.0	54.3	24.2	55.6
	DIFT [46]	63.5	54.5	80.8	34.5	46.2	52.7	48.3	77.7	39.0	76.0	54.9	61.3	53.3	46.0	57.8	57.1	71.1	63.4	57.7
	SD+DINO [61]	73.0	64.1	86.4	40.7	52.9	55.0	53.8	78.6	45.5	77.3	64.7	69.7	63.3	69.2	58.4	67.6	66.2	53.5	64.0
U†	NeuCongeal† [35]	-	29.1	-	-	-	-	-	53.3	-	-	35.2	-	-	-	-	-	-	-	-
	Ours-Zero-Shot†	**78.0**	**66.4**	**90.2**	**44.5**	**60.1**	**66.6**	**60.8**	**82.7**	**53.2**	**82.3**	**69.5**	**75.1**	**66.1**	**71.7**	**58.9**	**71.6**	**83.8**	**55.5**	**69.6**
S	SCOT [26]	34.9	20.7	63.8	21.1	43.5	27.3	21.3	63.1	20.0	42.9	42.5	31.1	29.8	35.0	27.7	24.4	48.4	40.8	35.6
	PMNC* [25]	54.1	35.9	74.9	36.5	42.1	48.8	40.0	72.6	21.1	67.6	58.1	50.5	40.1	54.1	43.3	35.7	74.5	59.9	50.4
	SCorrSAN* [16]	57.1	40.3	78.3	38.1	51.8	57.8	47.1	67.9	25.2	71.3	63.9	49.3	45.3	49.8	48.8	40.3	77.7	69.7	55.3
	CATs++* [7]	60.6	46.9	82.5	41.6	56.8	64.9	50.4	72.8	29.2	75.8	65.4	62.5	50.9	56.1	54.8	48.2	80.9	74.9	59.8
	DHF [30]	74.0	61.0	87.2	40.7	47.8	70.0	74.4	80.9	38.5	76.1	60.9	66.8	66.6	70.3	58.0	54.3	87.4	60.3	64.9
	SD+DINO (S) [61]	81.2	66.9	91.6	61.4	57.4	85.3	83.1	90.8	54.5	88.5	75.1	80.2	71.9	77.9	60.7	68.9	92.4	65.8	74.6
	Ours	87.0	73.7	95.4	69.0	66.1	91.6	86.9	90.7	68.6	93.6	85.2	84.6	78.7	86.9	79.7	79.0	**96.9**	84.3	82.9
	Ours (Adapt. Pose)†	87.6	74.1	95.5	70.1	66.7	**92.0**	87.4	91.4	68.0	93.2	85.5	84.7	79.9	87.8	79.9	78.9	**96.9**	84.8	83.2
	Ours (AP-10K P.T.)	**92.0**	**76.1**	**97.2**	**70.4**	**70.5**	91.4	**89.7**	**92.7**	**73.4**	**95.0**	**90.5**	**87.7**	**81.8**	**91.6**	**82.3**	**83.4**	96.5	**85.3**	**85.6**

ever, it shows two major limitations: argmax is limited to discrete pixel coordinates without sub-pixel reasoning, and it does not incorporate any spatial context with neighboring pixels when determining correspondence. One could use soft-argmax at time time too, but our study shows in Tab. 5 that it does not improve the performance on all metrics, probably due to its nature of incorporating similarities from all pixels with possible noisy response.

To complement the weaknesses of both, we propose a *window soft argmax* technique for both supervised and unsupervised settings. First, we determine the target center location using the argmax operation and apply soft-argmax on the pre-defined window, as illustrated in Fig. 8. This hybrid approach naturally enables sub-pixel reasoning but also prevents it from being affected by any noisy response in the similarity map. Tab. 5 shows that the usage of window soft argmax substantially improves the correspondence performance on all metrics.

5. Experimental Results

Implementation details. We follow [61] to resize the input image to 960^2 and 840^2 to extract the SD and DINOv2 features, respectively, yielding a feature map at a resolution of 60×60. The post-processor on top of the fused features is four bottleneck layers [13] with 5M parameters in total. The model is trained with the AdamW optimizer [28] of weight decay rate 0.001 and the one-cycle scheduler [45] of 1.25×10^{-3} learning rate and 0.3 percentage for the increasing cycle. We train all our models on one NVIDIA RTX3090 GPU. Refer to Supp. A for more details.

Datasets. We evaluate our methods on two widely-used benchmarks, namely PF-Pascal and SPair-71k, and our new proposed benchmark. *PF-Pascal* [11] consists of 2941 training, 308 validation, and 299 testing image pairs with similar viewpoints and instance pose. The images span across 20 categories of objects. *SPair-71k* [32] is a more challenging and larger-scale dataset with $53,340$ training pairs, $5,384$ validation pairs, and $12,234$ testing pairs across 18 categories, with large intra-class variation.

AP-10K benchmark. To further validate and improve our method in an in-the-wild setting, we build a new large-scale, challenging semantic correspondence benchmark with an existing animal pose estimation dataset. The AP-10K dataset [60] consists of 10,015 images across 23 families and 54 species. All the images share the same keypoint annotation of 17 keypoints. After manually filtering images with multiple instances and images with less than three visible keypoints, we construct a benchmark with 261k training, 17k validation, and 36k testing image pairs. The validation and testing image pairs span three settings: the main intra-species set, the cross-species set, and the cross-family set. It is 5× larger than the largest existing benchmark [32]. Please refer to Supp. B for more details.

Evaluation metrics. We follow the common practice and use the Percentage of Correct Keypoints (PCK) [59] to evaluate the correspondence accuracy. The PCK is computed within a threshold of $\alpha \cdot max(h, w)$ where α is a positive decimal (*e.g.*, 0.10) and (h, w) denotes the dimensions of the bounding box of an instance in SPair-71k and AP-10K, and the dimensions of the images in PF-Pascal, respectively.

5.1. Quantitative Analysis

Overall semantic correspondence. The per-category evaluation results, presented in Tab. 2, demonstrate the efficacy

Table 3. **Evaluation on SPair-71k, AP-10K, and PF-Pascal datasets at different PCK levels.** We report the performance of the AP-10K intra-species (I.S.), cross-species (C.S.), and cross-family (C.F.) test sets. †: index is used to flip source keypoints at test time. ∗: fine-tuned backbone. We report the *per image* PCK results (hence the (U) results are different from Tab. 2). The highest and second PCK among each category is **bold** and underlined, respectively. Both our zero-shot and supervised methods outperform all previous methods significantly.

		SPair-71k			AP-10K-I.S.			AP-10K-C.S.			AP-10K-C.F.			PF-Pascal		
	Method	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10	0.05	0.10	0.15
U	DINOv2+NN [36, 61]	6.3	38.4	53.9	6.4	41.0	60.9	5.3	37.0	57.3	4.4	29.4	47.4	63.0	79.2	85.1
	DIFT [46]	7.2	39.7	52.9	6.2	34.8	50.3	5.1	30.8	46.0	3.7	22.4	35.0	66.0	81.1	87.2
	SD+DINO [61]	7.9	44.7	59.9	7.6	43.5	62.9	6.4	39.7	59.3	5.2	30.8	48.3	71.5	85.8	90.6
U†	**Ours-Zero-Shot†**	**9.9**	**49.1**	**65.4**	**11.3**	**49.8**	**68.7**	**9.3**	**44.9**	**64.6**	**7.4**	**34.9**	**52.7**	**74.0**	**86.2**	**90.7**
S	SCorrSAN* [16]	3.6	36.3	55.3	-	-	-	-	-	-	-	-	-	81.5	93.3	96.6
	CATs++* [7]	4.3	40.7	59.8	-	-	-	-	-	-	-	-	-	84.9	93.8	96.8
	DHF [30]	8.7	50.2	64.9	8.0	45.8	62.7	6.8	42.4	60.0	5.0	32.7	47.8	78.0	90.4	94.1
	SD+DINO (S) [61]	9.6	57.7	74.6	9.9	57.0	77.0	8.8	53.9	74.0	6.9	46.2	65.8	80.9	93.6	96.9
	Ours	21.6	72.6	82.9	23.1	73.0	87.5	**21.7**	70.2	85.8	**18.4**	63.1	78.4	85.5	95.1	97.4
	Ours (Adapt. Pose)†	21.7	72.8	83.2	**23.2**	**73.2**	**87.7**	**21.7**	**70.3**	**85.9**	18.3	**63.2**	**78.5**	85.3	95.0	97.4
	Ours (AP-10K P.T.)	**22.0**	**75.3**	**85.6**	-	-	-	-	-	-	-	-	-	**85.9**	**95.7**	**98.0**

Table 4. **Evaluation on the geometry-aware subset.** We report the results on both SPair-71k and AP-10K intra-species test sets across three PCK levels. The best performances are **bold**.

		SPair-71k			AP-10K-I.S.		
	Method	0.01	0.05	0.10	0.01	0.05	0.10
U	DINOv2+NN [36, 61]	3.4	28.2	42.0	2.1	26.8	48.6
	DIFT [46]	4.6	30.0	42.5	1.8	18.9	34.6
	SD+DINO [61]	5.3	34.5	49.3	2.5	28.0	49.5
U†	**Ours-Zero-Shot†**	**6.9**	**39.5**	**56.8**	**3.5**	**35.9**	**57.8**
S	SCorrSAN* [16]	2.8	30.0	49.4	-	-	-
	CATs++* [7]	3.2	33.1	53.0	-	-	-
	DHF [30]	6.8	42.1	56.7	2.5	30.0	50.7
	SD+DINO (S) [61]	7.5	50.3	67.6	4.0	43.7	69.3
	Ours	18.2	66.0	77.4	10.4	64.8	82.8
	Ours (Adapt. Pose)†	18.3	66.3	78.0	**10.5**	**65.0**	**83.2**
	Ours (AP-10K P.T.)	**20.1**	**71.0**	**82.3**	-	-	-

of our methods. Our *zero-shot* approach, despite its simplicity, achieves considerable gains over SD+DINO, highlighting the significance of pose alignment in semantic correspondence. In the *supervised* category, our methods outperform existing works across all 18 categories, registering a substantial improvement of **11.0p** (from 74.6 to 85.6). Notably, pre-training on the AP-10K dataset contributes a gain of 2.7p, underscoring the untapped potential of animal pose datasets in this domain.

Further comparisons across different datasets and three PCK levels are in Tab. 3. Our methods exhibit significant improvements across most metrics, particularly with notable gains in the more strict thresholds (e.g., PCK@0.05 and PCK@0.01), especially considering that SD+DINO uses the same raw feature maps as our model. Despite the methods being trained only on AP-10K intra-species sets, the robust performance on cross-species and cross-family test sets showcases the generalizability of our approach.

Geometry-aware semantic correspondence. Our methods achieve even more significant improvements in the geometry-aware subset, as reported in Tab. 4. We reduce the relative gap from 9.38% (SD+DINO (S)) to 3.86% on the SPair-71k PCK@0.10 metric. Notably, the proposed adaptive viewpoint alignment brings more substantial gain on the geometry-aware subset for both zero-shot and supervised settings, suggesting its effectiveness in improving the geometric ambiguity by mitigating the pose variation. Besides, pre-training on the AP-10K dataset brings even a gain of 4.3p on the geo-aware subset.

Table 5. **Ablation study on SPair-71k.** We report the PCK@α_{bbox} results for both standard set (Std.) and geometry-aware set (Geo.). The best performances are **bold**. Our default method is underlined.

		SPair-71k (Std.)			SPair-71k (Geo.)		
	Model Variants	0.01	0.05	0.10	0.01	0.05	0.10
	Baseline	9.6	57.7	74.6	7.5	50.3	67.6
+	Dense Training Objective	13.0	65.2	78.3	11.1	58.8	71.9
+	Pose-variant Augmentation	13.8	66.7	80.0	11.4	60.5	73.9
+	Perturbation & Dropout	15.1	69.3	81.3	13.5	63.3	75.4
	Soft Argmax Inference	20.5	69.6	81.0	16.9	61.9	75.0
+	Window Soft Argmax (5)	**22.3**	72.1	82.0	**19.8**	66.0	76.5
+	Window Soft Argmax (9)	22.0	**72.7**	82.5	19.2	**66.3**	77.1
	Window Soft Argmax (15)	21.6	72.6	**82.9**	18.2	66.0	**77.4**

Ablation studies. Further ablation studies are in Tab. 5. Each element of our designs brings about moderate improvements. The dense training objective, pose-variant augmentation, and window soft argmax notably enhance results in the geometry-aware subset, while ground truth perturbation and feature map Dropout improve the overall correspondence (as shown in the similar gain on both sets). Regarding the window soft argmax, varying window sizes have different effects across three thresholds. We set the window size to 15×15 and 11×11 for the supervised and unsupervised setting respectively, for the optimal balance.

In the Supp. D.4, we also provide a leave-one-out ablation study with the breakdown evaluation protocol introduced in [3], to evaluate the detailed effect of each of our proposed module. In short, all our designs expect pertuba-

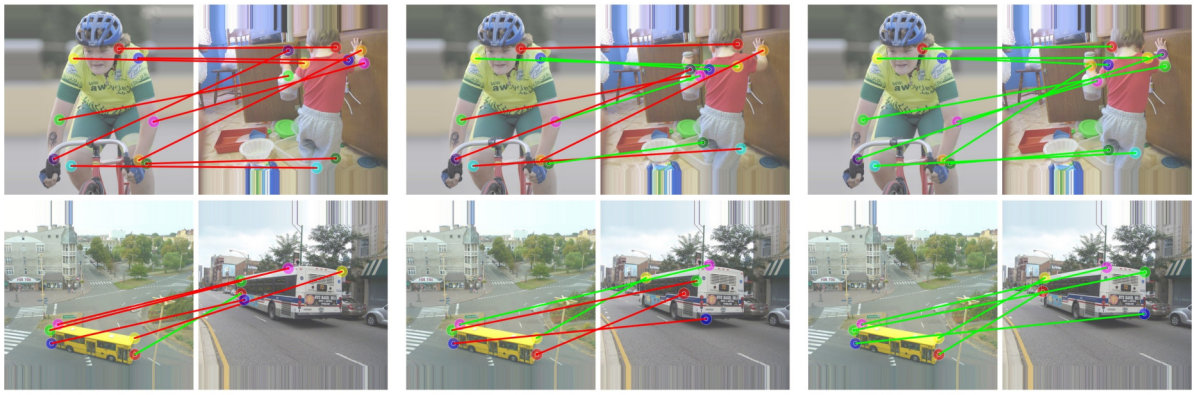


Figure 10. **Qualitative comparison.** Green lines indicate correct matches and red incorrect. Our method can build geometrically correct semantic correspondence even at extreme view variation, while both versions of SD+DINO struggle with geometric ambiguity (*e.g.*, ear and hands in the person example, corners in the bus example). Please refer to Supp. E.2 and E.3 for more results.

Figure 11. **Visualization of the similarity map.** For the red query point, SD+DINO matches appearance-similar points (wooden desk, floor); SD+DINO (S) returns a noisy similarity map due to the query point being out of supervision. Our method locates both semantically and geometrically correct points. The keypoint supervision of "chair" category is in blue, though these images are not in the training set.

tion&dropout notably improves the geometry-aware (*e.g.*, left/right) confusion, while the dense training objective also reduces mismatches to the image background.

5.2. Qualitative Analysis

We qualitatively compare our methods against both zero-shot and supervised versions of SD+DINO. As shown in Fig. 10, our approach significantly enhances semantic correspondence under the extreme view-variation cases. While additional supervision in SD+DINO does aid in keypoint localization to some extent, both versions of SD+DINO struggle with geometric ambiguity.

We further investigate cases where the query point lacks meaning and without direct supervision. As the visualization of the similarity map shown in the Fig. 11, SD+DINO highlights the regions with similar appearance (wooden materials) but fails to locate the chair; SD+DINO (S) generates noisy similarity maps (all regions are highlighted) when the query point is out of supervision, due to the sparse training objective; Our method locates the points both semantically and geometrically correct. Despite all methods sharing the same raw feature maps and our approach using the same feature post-processor as SD+DINO (S), the improvements in our method underscore the effectiveness of our design.

Limitations. As shown in Fig. 12 (top), small instances may be challenging for our method due to the resolution limits of raw feature maps. Our method may fail on extreme pose variations with severe deformation (see Fig. 12, bottom). Future work may address these complex scenarios by advanced reasoning mechanisms or geometry prior, such as the spherical constraint proposed in concurrent work [31].

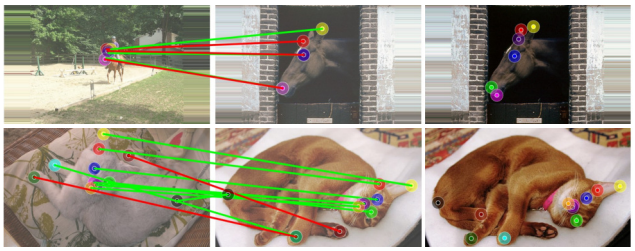


Figure 12. **Limitations.** Top: small instance. Bottom: scenarios combining both large pose variation and severe deformation.

6. Conclusion

We identified the problem of geometry ambiguity in semantic correspondence and introduced simple and effective techniques to improve current methods. We also developed a new benchmark to train and validate existing methods. Extensive experiments demonstrate that our method not only significantly improves the overall semantic correspondence but also narrows the gap between the geometry-aware sub-set and the standard set, thereby benefiting various downstream tasks and providing another angle to understand the internal representation of foundation models.

References

[1] Kfir Aberman, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Neural best-buddies: Sparse cross-domain correspondence. *ACM TOG*, 37(4):1–14, 2018. 2

[2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 1, 2

[3] Mehmet Aygün and Oisin Mac Aodha. Demystifying unsupervised semantic correspondence estimation. In *ECCV*, pages 125–142. Springer, 2022. 2, 7, 15, 16

[4] Prianka Banik, Lin Li, and Xishuang Dong. A novel dataset for keypoint detection of quadruped animals from images. *arXiv preprint arXiv:2108.13958*, 2021. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[6] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. Cats: Cost aggregation transformers for visual correspondence. In *NeurIPS*, pages 9011–9023, 2021. 2

[7] Seokju Cho, Sunghwan Hong, and Seungryong Kim. Cats++: Boosting cost aggregation with convolutions and transformers. *IEEE TPAMI*, 2022. 2, 6, 7, 15, 16

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893. Ieee, 2005. 2

[9] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 1

[10] Kamal Gupta, Varun Jampani, Carlos Esteves, Abhinav Shrivastava, Ameesh Makadia, Noah Snavely, and Abhishek Kar. Asic: Aligning sparse in-the-wild image collections. In *ICCV*, 2023. 1, 2, 4, 6

[11] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *CVPR*, pages 3475–3484, 2016. 2, 6

[12] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE TPAMI*, 40(7):1711–1725, 2017. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 6

[14] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. In *NeurIPS*, 2023. 1, 2

[15] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4D convolutional swin transformer for few-shot segmentation. In *ECCV*, pages 108–126. Springer, 2022. 2

[16] Shuaiyi Huang, Luyu Yang, Bo He, Songyang Zhang, Xuming He, and Abhinav Shrivastava. Learning semantic correspondence with sparse annotations. In *ECCV*, pages 267–284. Springer, 2022. 2, 6, 7, 15, 16

[17] Yiwen Huang, Yixuan Sun, Chenghang Lai, Qing Xu, Xiaomei Wang, Xuli Shen, and Weifeng Ge. Weakly supervised learning of semantic correspondence through cascaded online correspondence refinement. In *ICCV*, pages 16254–16263, 2023. 2

[18] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *CVPR*, pages 869–878, 2019. 1

[19] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, pages 66–75, 2017. 5

[20] Seungryong Kim, Stephen Lin, Sang Ryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *NeurIPS*, 2018. 2

[21] Seungryong Kim, Dongbo Min, Somi Jeong, Sunok Kim, Sangryul Jeon, and Kwanghoon Sohn. Semantic attribute matching networks. In *CVPR*, pages 12339–12348, 2019. 2

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 17

[23] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsub Ham. SFNet: Learning object-aware semantic correspondence. In *CVPR*, pages 2278–2287, 2019. 2, 5, 14, 15

[24] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *CVPR*, pages 5801–5810, 2020. 1

[25] Jae Yong Lee, Joseph DeGol, Victor Fragoso, and Sudipta N. Sinha. Patchmatch-based neighborhood consensus for semantic correspondence. In *CVPR*, pages 13153–13163, 2021. 2, 6

[26] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *CVPR*, pages 4463–4472, 2020. 2, 6

[27] Jonathan L. Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *NeurIPS*, 2014. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[29] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157. Ieee, 1999. 2

[30] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *NeurIPS*, 2023. 1, 2, 5, 6, 7, 16

[31] Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. Improving semantic correspondence with viewpoint-guided spherical maps. *arXiv preprint arXiv:2312.13216*, 2023. 8

[32] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv prepreint arXiv:1908.10543*, 2019. 1, 2, 3, 6, 12

[33] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Learning to compose hypercolumns for visual correspondence. In *ECCV*, pages 346–363. Springer, 2020. 2

[34] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. DragonDiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 1

[35] Dolev Ofri-Amar, Michal Geyer, Yoni Kasten, and Tali Dekel. Neural congealing: Aligning images to a joint semantic atlas. In *CVPR*, pages 19403–19412, 2023. 1, 2, 6

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 4, 6, 7

[37] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A. Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *CVPR*, pages 13470–13481, 2022. 2

[38] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, pages 6148–6157, 2017. 2

[39] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *CVPR*, pages 6917–6925, 2018. 2

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 4

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022. 1

[42] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *ECCV*, pages 349–364, 2018. 2

[43] Aleksandar Shtedritski, Andrea Vedaldi, and Christian Rupprecht. Learning universal semantic correspondences with no supervision and automatic data curation. In *ICCV Workshops*, pages 933–943, 2023. 2

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2

[45] Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. 6

[46] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, 2023. 2, 6, 7

[47] Tatsunori Taniai, Sudipta N. Sinha, and Yoichi Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *CVPR*, pages 4246–4255, 2016. 2

[48] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. 30, 2017. 2

[49] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv preprint arXiv:2308.12469*, 2023. 1

[50] Prune Truong, Martin Danelljan, Luc V. Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. In *NeurIPS*, pages 14278–14290, 2020. 2

[51] Prune Truong, Martin Danelljan, and Radu Timofte. Glu-net: Global-local universal network for dense flow and correspondences. In *CVPR*, pages 6258–6268, 2020. 2

[52] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *CVPR*, pages 5714–5724, 2021.

[53] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Warp consistency for unsupervised learning of dense correspondences. In *ICCV*, pages 10346–10356, 2021. 2

[54] Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *CVPR*, pages 8708–8718, 2022. 2

[55] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. 2011. 2

[56] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, pages 2955–2966, 2023. 1, 11, 16, 17

[57] Fan Yang, Xin Li, Hong Cheng, Jianping Li, and Leiting Chen. Object-aware dense semantic correspondence. In *CVPR*, pages 2777–2785, 2017. 2

[58] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. *NeurIPS*, 32, 2019. 5

[59] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE TPAMI*, 35(12):2878–2890, 2012. 6

[60] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *NeurIPS*, 2021. 1, 2, 6

[61] 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. In *NeurIPS*, 2023. 1, 2, 4, 5, 6, 7, 11, 14, 16, 17

[62] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023. 1