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

CVPR 2024

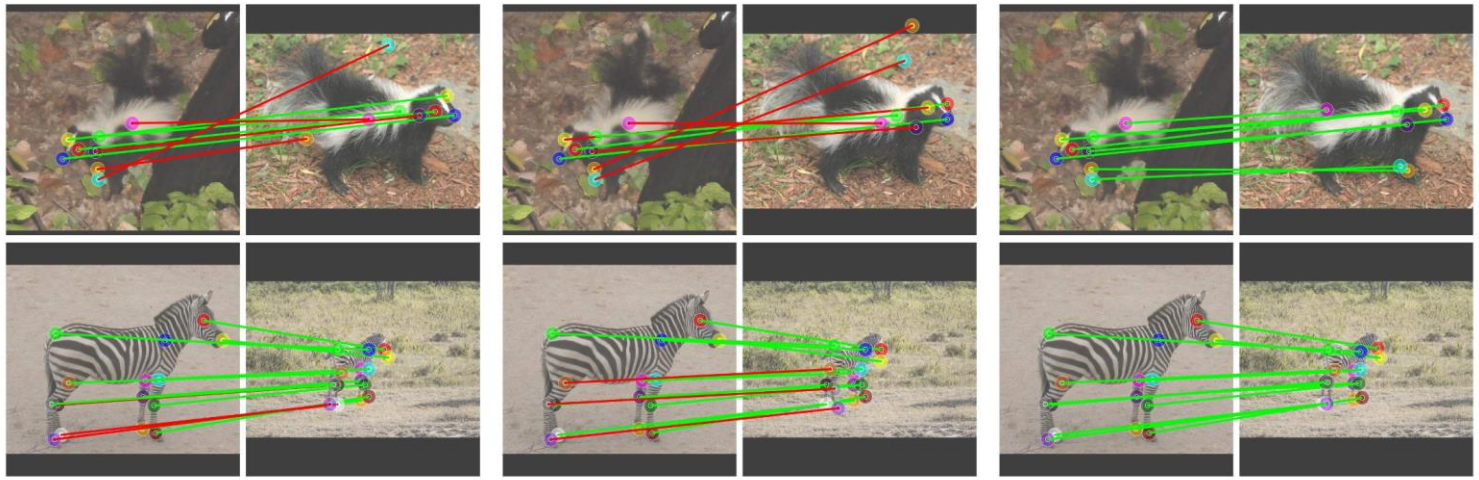
Contributions

•We identify **the problem of geometry-aware semantic correspondence** and show that pre-trained features of foundation models (SD and DINOv2) struggle with geometric information.

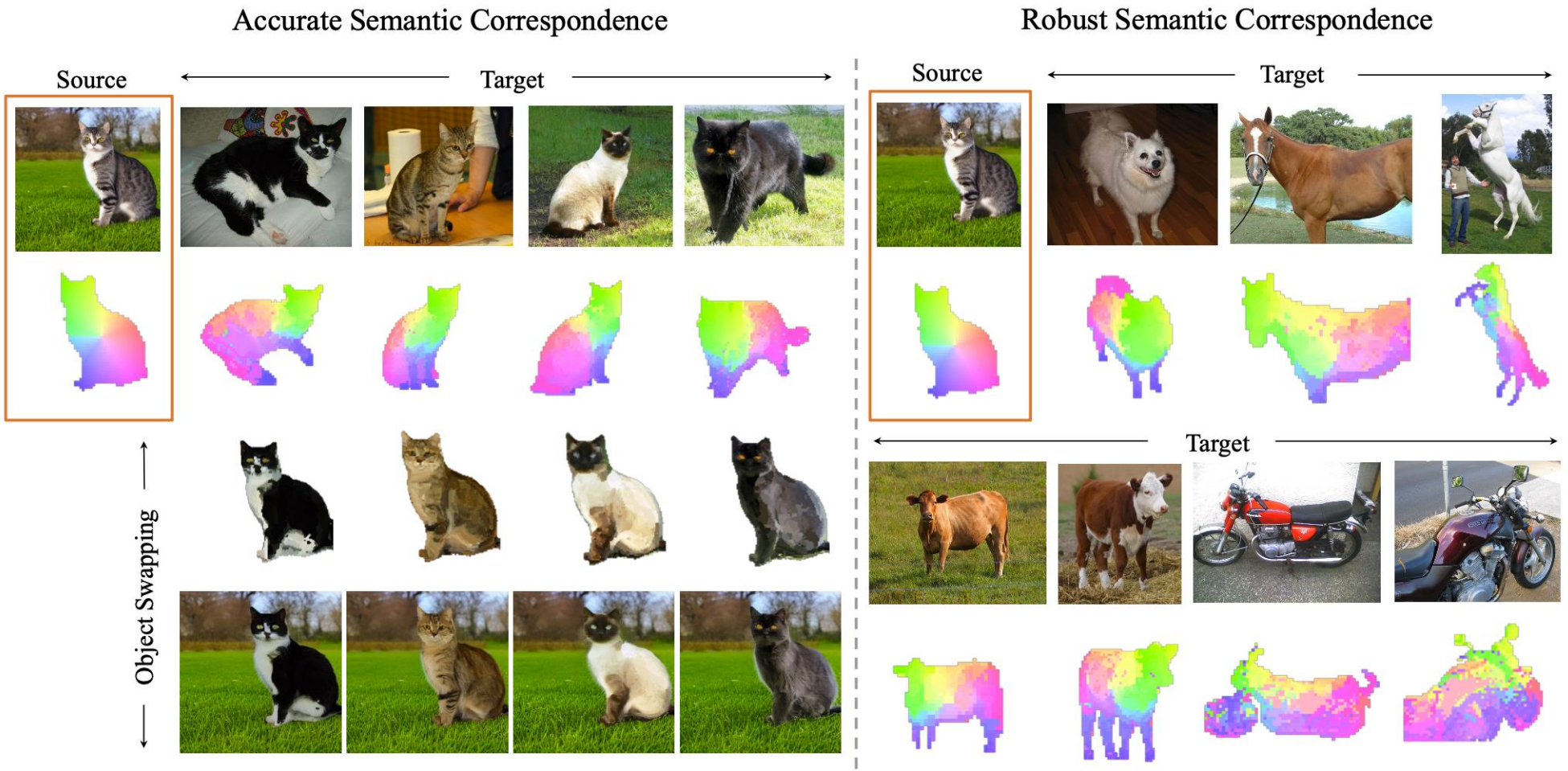
•We propose to improve geometric awareness of the features in **both unsupervised and supervised manners**.

•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%.



Preliminary: Semantic correspondence using SD + DINOv2 feature



[1] Zhang, Junyi, et al. "A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence." *Advances in Neural Information Processing Systems* 36 (2024).

Limitations: geometric ambiguity

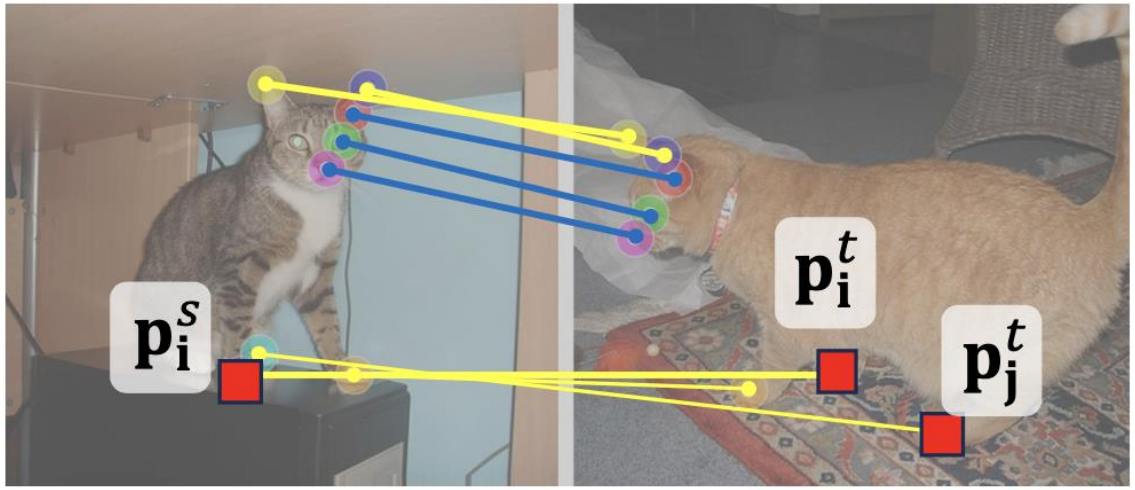
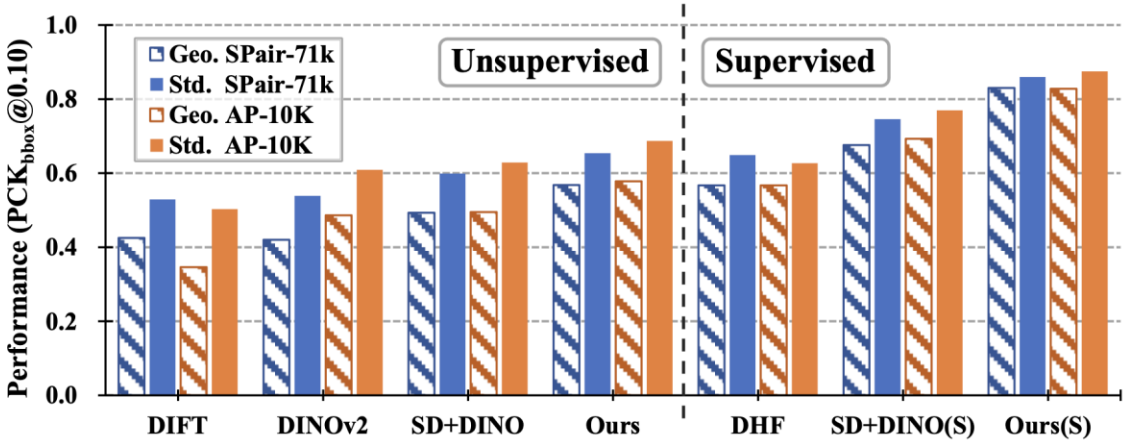


Figure 3. Annotations of geometry-aware semantic correspondence (yellow) and standard semantic correspondence (blue).



(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.

Geometry-ambiguous matching cases require an understanding of instances' orientations or geometry

Gather geometry-aware subset:

(a) A keypoint from the source image and a keypoint from the target image belong to the same semantic subgroup (eyes, paws…).

(b) There are other visible keypoint(s) belonging to the same subgroup in the target image.

(c) Such cases account for 82.4% of total image pairs and 59.6% of matching keypoints.

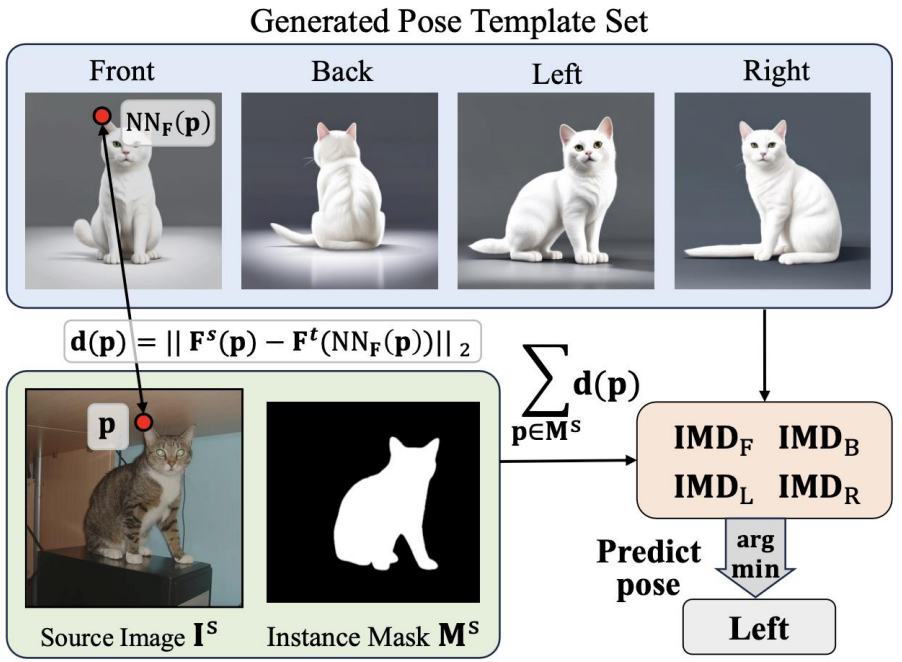
Global Pose Awareness of Deep Features

Analyze if deep features are aware of high-level pose (or viewpoint) information of an instance in an image.

Instance matching distance (IMD).

Nearest neighboring

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



Pose prediction via IMD

- Generate multiple pose template sets.
- Compute the IMD between the input and template images for each set.
- Predict the pose whose IMD is the smallest by a collective vote across all sets.

The deep features are aware of global pose information.

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

Improving Geo-Aware Correspondence

Test-time Adaptive Pose Alignment (zero-shot setting)

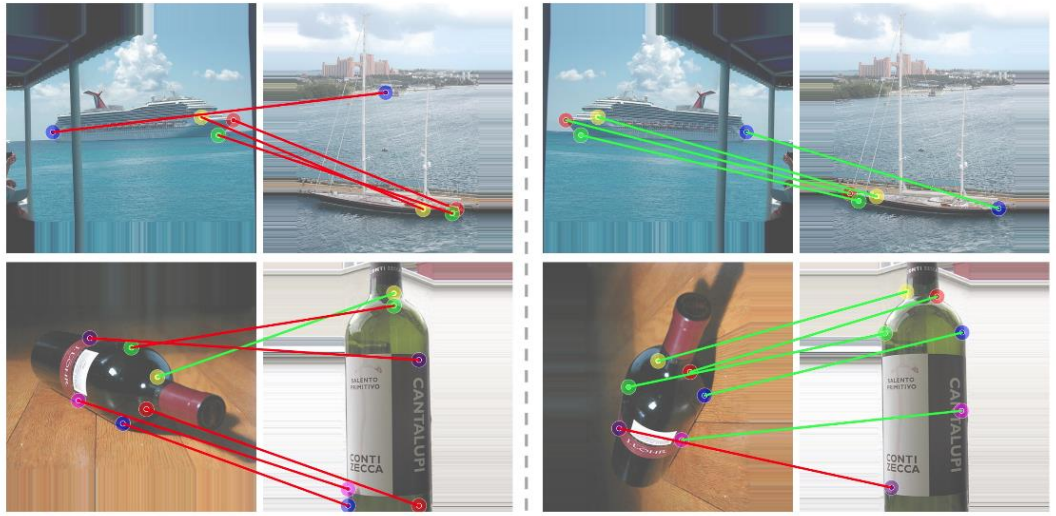


Figure 9. (Left) original image pairs. (Right) image pairs with the test-time aligned pose. The reduced pose variation improves the correspondence accuracy.

- Augment the source image by using a set of pose-variant augmentations (e.g. flip, rotations etc.).
- Calculate the IMD between the augmented source images and the target image.
- Choose the optimal pose with the minimum IMD distance.

⟸ This simple pose alignment can drastically improve the correspondence accuracy in a test-time, unsupervised manner.

Improving Geo-Aware Correspondence

Dense Training Objective (supervised setting)

$$\mathcal{L} = \mathcal{L}_{\text{dense}} + \mathcal{L}_{\text{sparse}}.$$

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

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

CL: Contrastive loss

$\tilde{F} = f(F)$: the refined feature map

$f(\cdot)$: a trainable lightweight post-processor

F: the raw feature map

P: Annotated keypoint pairs set

$$\hat{\mathbf{p}}_i^t = \text{SoftArgmax}(S_i)$$

$$S_i = \tilde{\mathbf{F}}^s(\mathbf{p}_i^s)^T \tilde{\mathbf{F}}^t$$

Add dropout at the input feature map F and Gaussian noise ϵ that perturbs the GT to prevent overfitting

Improving Geo-Aware Correspondence

Pose-variant Augmentation

A set of pose-varying augmentation schemes

1) *double flip*: flipped source image and flipped target image;
2) *single flip*: flipped source image and original target image;
3) *self flip*: source image and flipped source image.

} keypoint annotations are correspondingly flipped to preserve the inherent geometric concept (e.g. the left paw should be the right paw after flip).

Window Soft Argmax

1) we determine the target center location using the argmax operation.

2) Apply soft-argmax on the pre-defined window.

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**

Improving Geo-Aware Correspondence

Dense Training Objective (supervised setting)

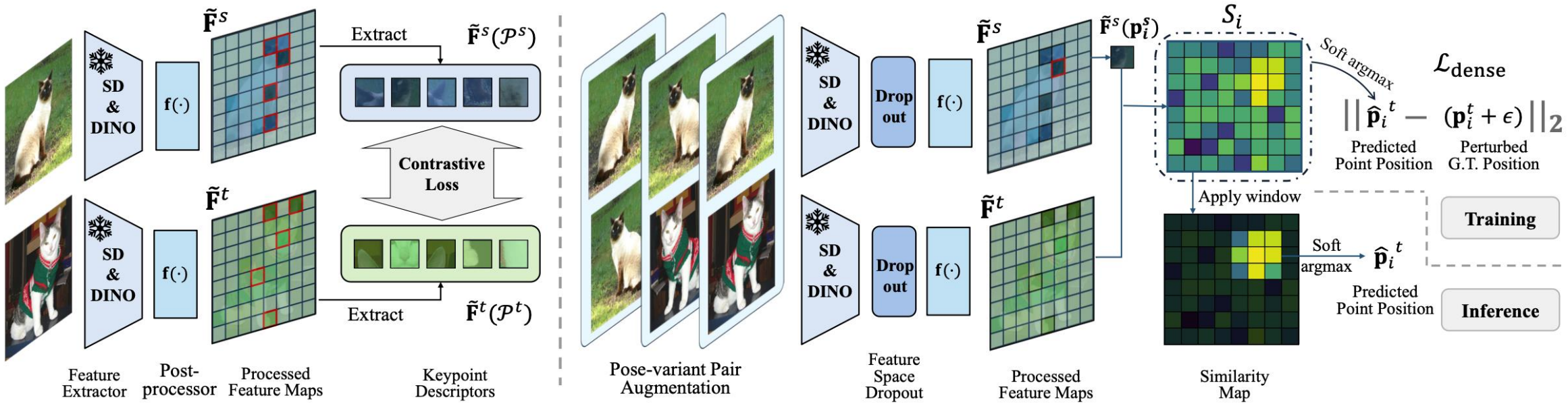


Figure 8. (Left) previous supervised methods [30, 61] with a sparse training objective. (Right) an overview of our supervised method. Only the lightweight post-processor is updated during training. Both the pair augmentation and feature space Dropout are for training only.

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.

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.

Datasets

Two widely-used benchmark

PF-Pascal and Spair-71k

Propose a new large-scale benchmark with AP-10K dataset

AP-10K: an existing animal pose estimation dataset consists of 10015 images across 23 families and 54 species. All images share the same keypoint annotation of 17 keypoints.

261k training / 17k validation / 35k testing pairs.

3 setting for validation and testing:
intra-species / cross-species / cross-family

Experimental Results

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**

Experimental Results

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**

Experimental Results

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**	-	-	-

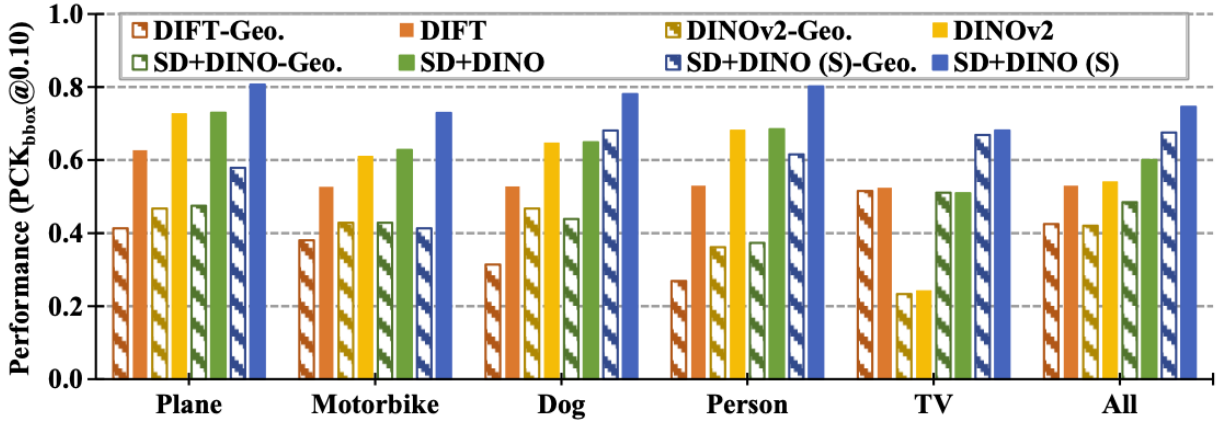
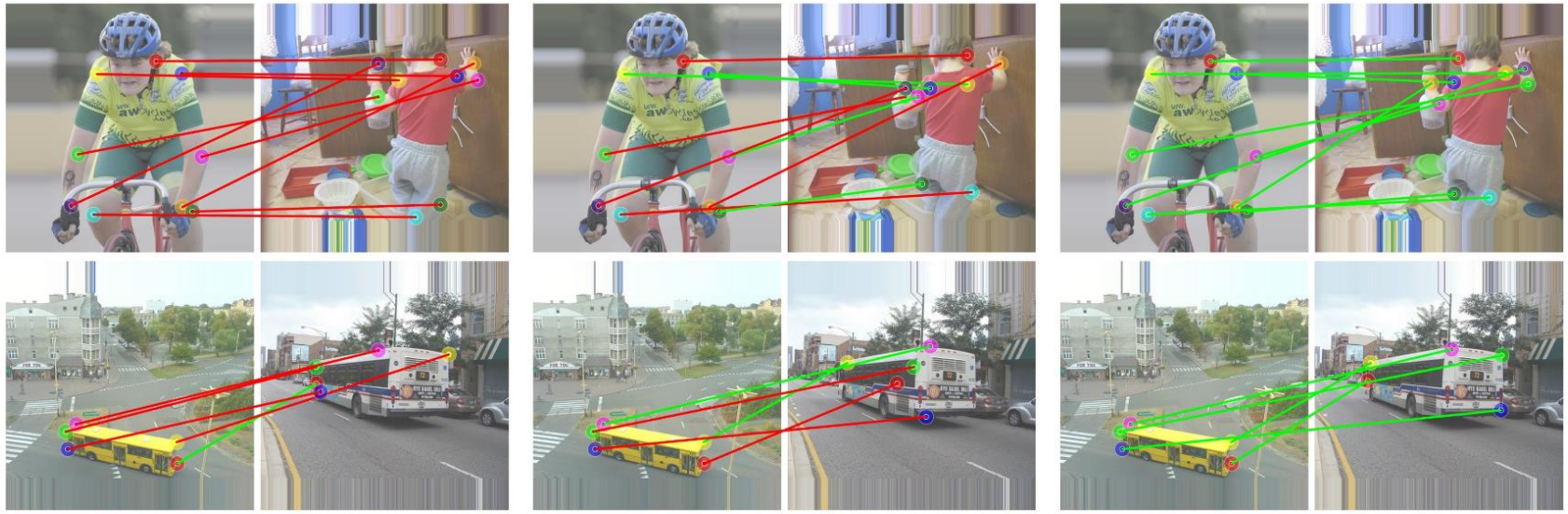


Figure 4. **Per-category evaluation of state-of-the-art methods on SPair-71k geometry-aware subset (Geo.) and standard set.** While the geometry-aware subset accounts for 60% of the total matching keypoints, we observe a substantial performance gap between the two sets for all the methods.

Experimental Results



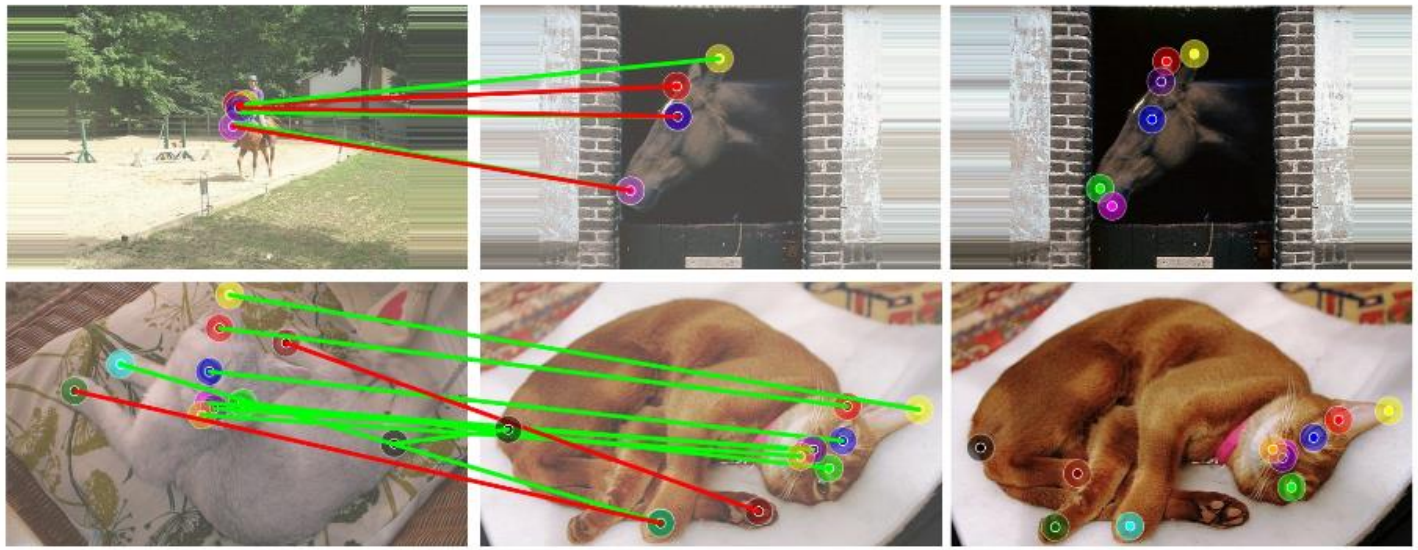
SD+DINO | SD+DINO (S) | **Ours**

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.

Source Image | SD+DINO | SD+DINO (S) | **Ours**

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.

Limitations



Source Image Predicted Points G.T. Points

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