Telling Left from Right: Identifying Geometry-Aware Semantic Correspondence

Supplementary Material

Contents

1. Introduction 1

2. Related Work 2

3. Geometric Awareness of Deep Features 2
   - 3.1. Geometry-Aware Semantic Correspondence . 3
   - 3.2. Evaluation on the Geometry-aware Subset . . 3
   - 3.3. Sensitivity to Pose Variation 3
   - 3.4. Global Pose Awareness of Deep Features . . 4

4. Improving Geo-Aware Correspondence 4
   - 4.1. Test-time Adaptive Pose Alignment 4
   - 4.2. Dense Training Objective 5
   - 4.3. Pose-variant Augmentation 5
   - 4.4. Window Soft Argmax 5

5. Experimental Results 6
   - 5.1. Quantitative Analysis 6
   - 5.2. Qualitative Analysis 8

6. Conclusion 8

A Further Implementation Details 11

B Benchmarking AP-10K Dataset for Semantic Correspondence 12

C Details on Geo-Aware Correspondence 12

D Additional Analysis 14
   - D.1. Detailed Performance on Geo-Aware Subset 14
   - D.2. Detailed Analysis on Window Soft Argmax . 14
   - D.3. Discussion on Generalizability 15
   - D.4. Additional Ablation Analysis under Supervised Setting 15
   - D.5. Ablation Study under Unsupervised Setting . 16

E. Additional Results 16
   - E.1. Alternative Metrics for Pose Alignment . . . 16
   - E.2. Qualitative Results on AP-10K 17
   - E.3. Additional Qualitative Results on SPair-71k . 17

A. Further Implementation Details

Feature extraction. The extraction of SD and DINOv2 features is conducted in a manner similar to that described in Zhang *et al*. [60]. Specifically, the SD features are extracted from SD-1-5's UNet decoder layer 2, 5, and 8 at timestep 50 with an implicit captioner, and the DINOv2 features are extracted from the token facet of the 11th layer.

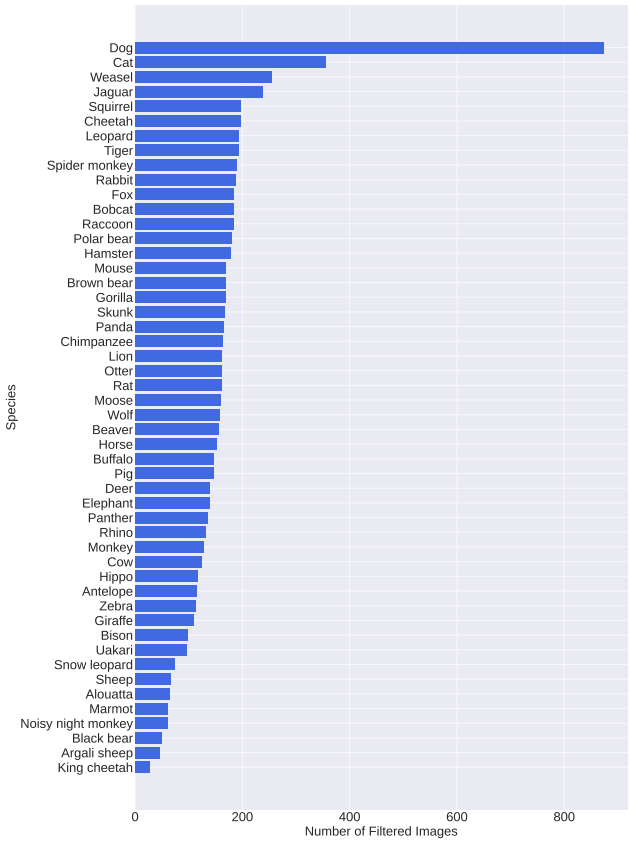


Figure 13. **Distribution of the filtered images across different species.** Note that only 50 species have annotated images.

Adaptive viewpoint alignment. For adaptive viewpoint (or pose) alignment in Sec. 4.1 in the main paper, we utilize segmentation masks from ODISE [55] to calculate the Instance Matching Distance (IMD). Considering the imbalanced viewpoint distribution in the images, "horizontal flip" is employed as the primary viewpoint augmentation for all categories. Specifically for the bottle category, to accommodate its unique viewpoint variations, we further apply rotations of +90°, 180°, and -90° as additional augmented viewpoints.

Pose-variant augmentation. In terms of pose-variant augmentation, we compute all the pair augmentations in a single batch and assign weights of 1 for both *single-flip* and *double-flip*, and a weight of 0.25 for the *self-flip*. Note that pose-variant augmentation is not applied during training on the PF-Pascal dataset due to all image pairs in this dataset

Figure 14. **Sample image pairs of AP-10K benchmark** including intra species, cross species, and cross family.

are of similar pose.

Training. Our model is trained for 100k steps (equivalent to 2 epochs) on the SPair-71k dataset, and 250k steps on AP-10K (equivalent to 1 epoch) and PF-Pascal (equivalent to 85 epochs), with a mini-batch size of 1. For a faster training, we pre-extract features from the visual foundation models offline and only train the post-processor online. This strategy significantly reduces the training duration, allowing it to be completed within just a few hours on a single GPU.

B. Benchmarking AP-10K Dataset for Semantic Correspondence

Image filtering. To start with, we exclude images with fewer than three visible keypoints or with multiple instances of the target category, to make the dataset less ambiguous for semantic matching.

Train/validation/test sets. After the filtering, there exists an imbalance in the number of images per species within the AP-10K dataset, as illustrated in Fig. 13. To ensure a balanced evaluation across different species, we uniformly sample an equivalent number of images for validation and test sets across all species — specifically, $N_{\text{val}} = 20$ for validation and $N_{\text{test}} = 30$ for testing, in line with the protocol established by SPair-71k [31]. The remaining images constitute the training set. It is important to note that for these three species, king cheetah, argali sheep, and black bear, whose numbers of images after the filtering are below 50, we earmark these as a hold-out set without including them in the training set. Thereby, it can also provide a measure for evaluating the generalization capability of semantic correspondence methods.

Intra-species image pair sampling. For each species, we construct all possible image matching pairs within each validation and test set (*i.e.*, $\binom{N_{\text{val}}}{2}$ and $\binom{N_{\text{test}}}{2}$) that are established in the previous step. On the other hand, the training set exhibits a more significant variance in the number of images; to circumvent the unbalanced distribution that arises from quadratic pairing growth, we limit the pairing to a maximum of either $50 \times N_{\text{train}}$ or $\binom{N_{\text{train}}}{2}$ pairs, whichever is fewer. Considering that the AP-10K dataset was not initially curated for the task of semantic correspondence, we apply an additional filtration criterion to the image pairs, retaining only those with a minimum of three mutual visible keypoints. This results in a total number of 260,950 training, 8816 validation, and 20,630 testing image pairs.

Cross-species and cross-family image pair sampling. We also include correspondence matching pairs across different species and families. For all 11 families with multiple species, we sample $\binom{N_{\text{val}}}{1} \cdot \binom{N_{\text{val}}}{1}$ validation pairs and $\binom{N_{\text{test}}}{1} \cdot \binom{N_{\text{test}}}{1}$ testing pairs for each family. For the cross-family setting, among all the $\binom{21}{2}$ combination of the total of 21 families, we only sample N_{val} validation and N_{test} testing pairs to save compute. A filtering process based on the mutually visible keypoints is also applied, yielding a total number of 4300 and 4200 validation pairs, alongside 9619 and 6300 testing pairs for cross-species and cross-family correspondence, respectively. Please refer to Fig. 14 for sample image pairs.

C. Details on Geo-Aware Correspondence

Keypoint subgroups. We list the keypoint subgroups of each category in Tab. 6. We exclude very few parts (nostril, eyes, *etc*.) that are close to each other and thus cannot be easily distinguished by existing metrics. We suggest that an improved metric (*e.g.*, a keypoint can be only regarded as a prediction to its nearest ground truth point) can make up this issue.

Per-category proportion. We show the average proportion of the geometry-aware subset with respect to both image pairs and keypoint pairs for each category in Fig. 15. For most of the categories, the geometry-aware subset accounts for a considerable fraction of all pairs.

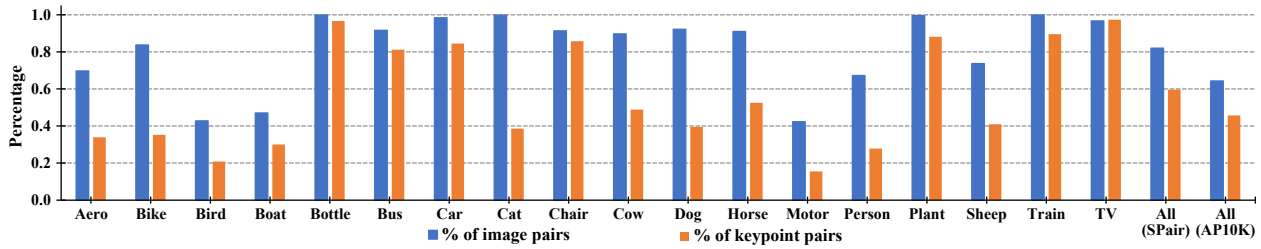


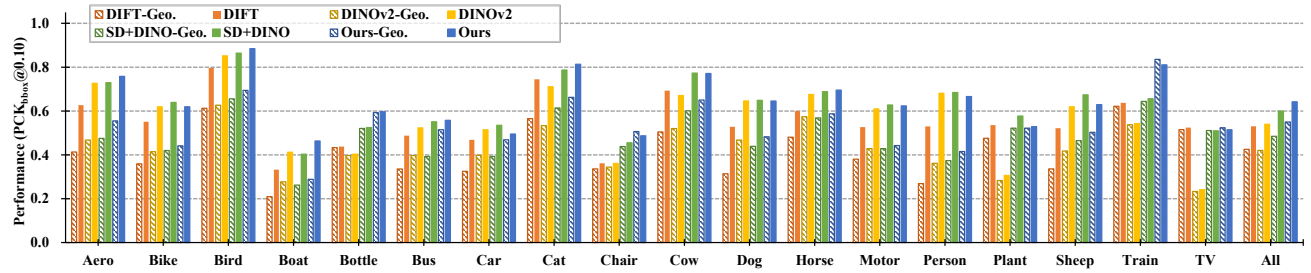
Figure 15. **Proportion of the geometry-aware subset with respect to image pair and keypoint pair.** We show the per-category results of SPair-71k as well as the average results of SPair-71k and AP-10K intra-species set.

Table 6. **Semantically similar keypoint subgroups.** We list the keypoint subgroups for categories from both SPair-71k and AP-10K. The number in the bracket indicates the number of keypoints in each subgroup. The annotation in the index version will also be released.

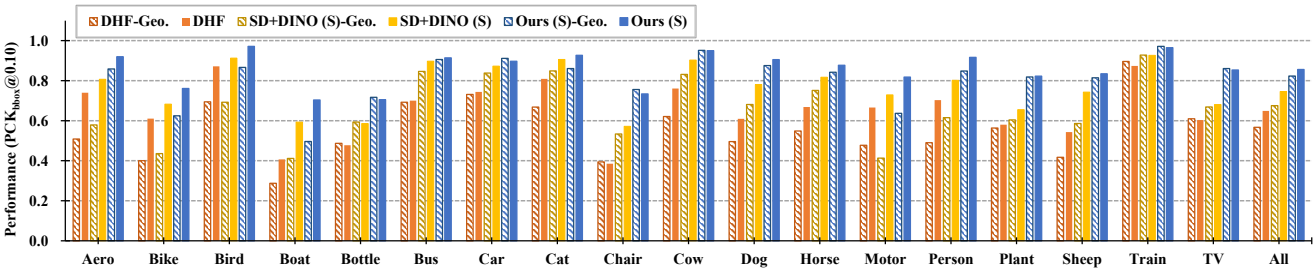
Dataset	Category	Subgroups
SPair-71k	Aeroplane	$\mathcal{G}_{\text{landing\_gear}}$ (2), $\mathcal{G}_{\text{engine\_front}}$ (2), $\mathcal{G}_{\text{wing\_end}}$ (2), $\mathcal{G}_{\text{engine\_back}}$ (2), $\mathcal{G}_{\text{wing\_foot\_front}}$ (2), $\mathcal{G}_{\text{wing\_foot\_back}}$ (2), $\mathcal{G}_{\text{tailplane\_end}}$ (2), $\mathcal{G}_{\text{tailplane\_foot\_front}}$ (2), $\mathcal{G}_{\text{tailplane\_foot\_back}}$ (2)
	Bicycle	$\mathcal{G}_{\text{handle}}$ (2), $\mathcal{G}_{\text{seat\_back\_end}}$ (2), $\mathcal{G}_{\text{pedal}}$ (2)
	Bird	$\mathcal{G}_{\text{wing\_end}}$ (2), $\mathcal{G}_{\text{foot}}$ (2), $\mathcal{G}_{\text{knee}}$ (2), $\mathcal{G}_{\text{hip}}$ (2)
	Boat	$\mathcal{G}_{\text{upper\_front}}$ (2), $\mathcal{G}_{\text{upper\_side}}$ (2), $\mathcal{G}_{\text{upper\_back}}$ (2), $\mathcal{G}_{\text{lower\_front}}$ (2), $\mathcal{G}_{\text{lower\_side}}$ (2), $\mathcal{G}_{\text{lower\_back}}$ (2)
	Bottle	$\mathcal{G}_{\text{cap}}$ (2), $\mathcal{G}_{\text{neck}}$ (2), $\mathcal{G}_{\text{shoulder}}$ (2), $\mathcal{G}_{\text{body}}$ (2), $\mathcal{G}_{\text{base}}$ (2)
	Bus	$\mathcal{G}_{\text{rearview\_mirror}}$ (2), $\mathcal{G}_{\text{light}}$ (2), $\mathcal{G}_{\text{licence\_plate}}$ (2), $\mathcal{G}_{\text{front\_fender}}$ (4), $\mathcal{G}_{\text{wheel}}$ (4), $\mathcal{G}_{\text{rear\_fender}}$ (4), $\mathcal{G}_{\text{window\_top\_corner}}$ (4), $\mathcal{G}_{\text{window\_bottom\_corner}}$ (4)
	Car	$\mathcal{G}_{\text{rearview\_mirror}}$ (2), $\mathcal{G}_{\text{light}}$ (2), $\mathcal{G}_{\text{licence\_plate}}$ (2), $\mathcal{G}_{\text{brand\_logo}}$ (2), $\mathcal{G}_{\text{rear\_fender}}$ (4), $\mathcal{G}_{\text{wheel}}$ (4), $\mathcal{G}_{\text{front\_fender}}$ (4), $\mathcal{G}_{\text{window\_bottom\_corner}}$ (4), $\mathcal{G}_{\text{window\_top\_corner}}$ (4)
	Cat	$\mathcal{G}_{\text{ear}}$ (2), $\mathcal{G}_{\text{paw}}$ (4)
	Chair	$\mathcal{G}_{\text{cushion\_front}}$ (2), $\mathcal{G}_{\text{cushion\_back}}$ (2), $\mathcal{G}_{\text{leg}}$ (4), $\mathcal{G}_{\text{backrest\_top}}$ (2), $\mathcal{G}_{\text{armrest\_front}}$ (2), $\mathcal{G}_{\text{armrest\_back}}$ (2)
	Cow	$\mathcal{G}_{\text{ear}}$ (2), $\mathcal{G}_{\text{hoof}}$ (4), $\mathcal{G}_{\text{knee}}$ (4), $\mathcal{G}_{\text{horn}}$ (2)
	Dog	$\mathcal{G}_{\text{ear}}$ (2), $\mathcal{G}_{\text{paw}}$ (4)
	Horse	$\mathcal{G}_{\text{ear}}$ (2), $\mathcal{G}_{\text{hoof}}$ (4), $\mathcal{G}_{\text{knee}}$ (4)
	Motorbike	$\mathcal{G}_{\text{rearview\_mirror}}$ (2), $\mathcal{G}_{\text{handle}}$ (2)
	Person	$\mathcal{G}_{\text{shoulder}}$ (2), $\mathcal{G}_{\text{elbow}}$ (2), $\mathcal{G}_{\text{wrist}}$ (2), $\mathcal{G}_{\text{knee}}$ (2), $\mathcal{G}_{\text{ankle}}$ (2), $\mathcal{G}_{\text{foot}}$ (2)
	Pottedplant	$\mathcal{G}_{\text{top}}$ (4), $\mathcal{G}_{\text{side\_wall}}$ (2), $\mathcal{G}_{\text{bottom}}$ (2)
	Sheep	$\mathcal{G}_{\text{ear}}$ (2), $\mathcal{G}_{\text{hoof}}$ (4), $\mathcal{G}_{\text{knee}}$ (4), $\mathcal{G}_{\text{horn}}$ (2)
	Train	$\mathcal{G}_{\text{front\_top}}$ (2), $\mathcal{G}_{\text{front\_bottom}}$ (2), $\mathcal{G}_{\text{back\_top}}$ (2), $\mathcal{G}_{\text{back\_bottom}}$ (2), $\mathcal{G}_{\text{window\_top\_outer\_corner}}$ (2), $\mathcal{G}_{\text{window\_bottom\_outer\_corner}}$ (2), $\mathcal{G}_{\text{window\_top\_inner\_corner}}$ (2), $\mathcal{G}_{\text{window\_bottom\_inner\_corner}}$ (2), $\mathcal{G}_{\text{front\_light}}$ (2)
	Tvmonitor	$\mathcal{G}_{\text{outer\_corner}}$ (4), $\mathcal{G}_{\text{outer\_side}}$ (4), $\mathcal{G}_{\text{inner\_corner}}$ (4), $\mathcal{G}_{\text{inner\_side}}$ (4)
AP-10K	All	$\mathcal{G}_{\text{shoulder}}$ (2), $\mathcal{G}_{\text{foot}}$ (4), $\mathcal{G}_{\text{knee}}$ (4), $\mathcal{G}_{\text{hip}}$ (2)

Notably, due to the unbalanced pose distribution exhibited in specific categories of the SPair-71k (*e.g.* bottles, potted plants, TVs, and trains) where image pairs often share similar poses, almost all keypoint subgroups in these categories are mutually visible, which results in proportions to be near 100%. In contrast, the AP-10K dataset, comprised solely of animal images, does not exhibit this imbalance.

Per-category performance. In Fig. 16a and Fig. 16b, we provide detailed per-category performance for both unsupervised and supervised state-of-the-art methods on the geometry-aware subset and the standard set. These figures provide an expanded view of Fig. 4 from the main paper. Regardless of the method or category, performance on the geometry-aware subset consistently lags behind that of the



(a) **Performance of the unsupervised methods.**



(b) **Performance of the supervised methods.**

Figure 16. **Per-category performance of the state-of-the-art methods and ours (blue).** We report both the geometry-aware subset (Geo.) and the standard set on SPair-71k. Our methods consistently outperform previous arts across all categories.

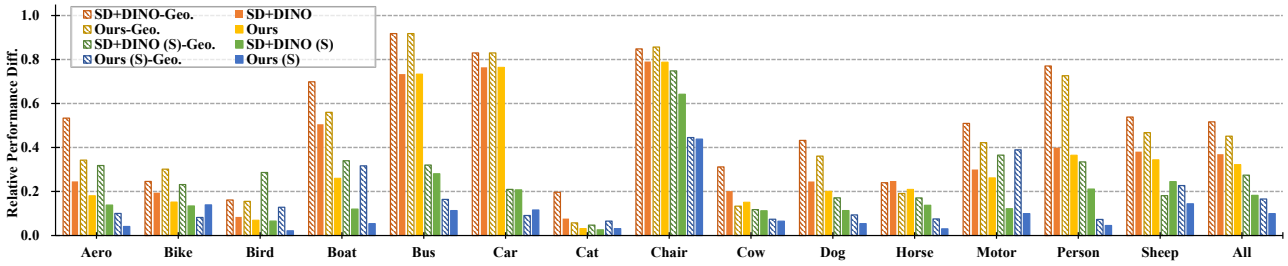


Figure 17. **Per-category evaluation of the sensitivity to pose variations.** Both our zero-shot (yellow) and supervised methods (blue) considerably improve the robustness to pose variations on both the geometry-aware set (Geo., hashed bar) and the standard set (solid bar) compared to the state-of-the-art methods [60]. We exclude categories that only have one azimuth-variation subset.

standard set.

Additionally, in Fig. 17, we offer a per-category analysis of pose variation sensitivity. The results for both unsupervised and supervised variants of SD+DINO [60] are presented, comparing their performance on both the geometry-aware and standard sets. This analysis serves as an extended version of Fig. 5 from the main paper. The findings clearly show that sensitivity to pose variation is considerably higher in the geometry-aware subset across all categories and methodologies.

D. Additional Analysis

D.1. Detailed Performance on Geo-Aware Subset

We provide the per-category performance on the geometry-aware subset in Fig. 16 as well as the pose-sensitivity analysis of our methods in Fig. 17.

D.2. Detailed Analysis on Window Soft Argmax

Performance in accordance with window size. We evaluate the effect of soft-argmax's window size on the performance at different PCK thresholds. As depicted in Fig. 18, the performance across all PCK levels initially improves and then declines as the window size increases from 0 (hard argmax) to 60 (soft argmax). Notably, the peak PCK values for 0.01, 0.05, and 0.1 are observed at window sizes of 5, 11, and 17, respectively. We opt for a window size of 15 to achieve an optimal balance in performance.

Comparison with Gaussian kernel soft argmax. Previous work [23] also explored a trade-off solution between hard and soft argmax by applying a Gaussian kernel on the feature map, centered at the hard argmax position.

We also search different σ values for the Gaussian kernel to achieve the best performance across different PCK levels. We then compare our window soft argmax with the kernel

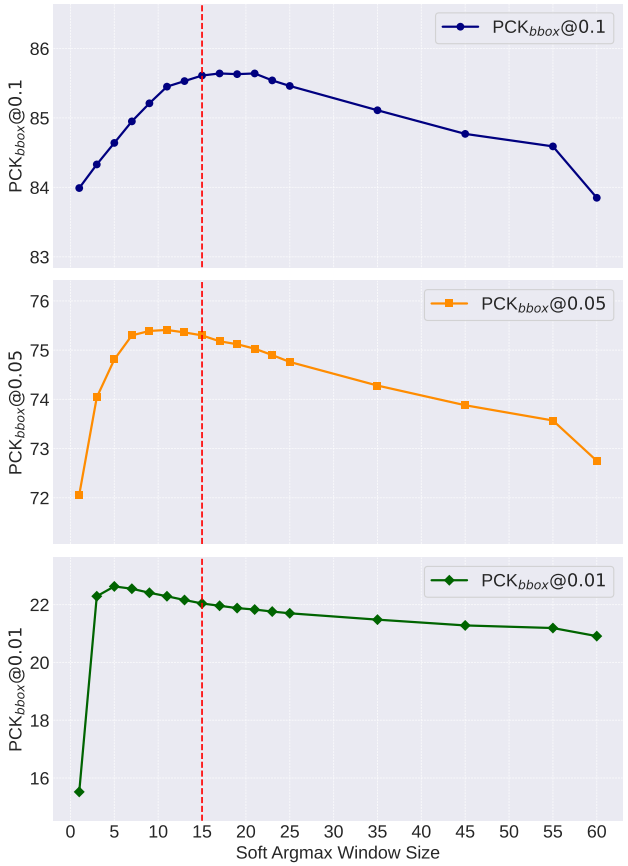


Figure 18. **Performance of different PCK levels vs. soft argmax window size.** We test the performance on the SPair-71k dataset and set the window size as 15 for optimal balance.

soft argmax in Tab. 7 on different peak PCK levels and the default value as reported in [23]. Our window soft argmax consistently outperforms kernel argmax across all settings, suggesting the superiority of our approach. We hypothesize that this is because when using the argmax-centered Gaussian kernel to scale the similarity map, it makes the similarity map biased to argmax locations, while our method treats the window region with the same scale.

Training with window soft argmax. We also experiment if applying the window soft argmax during training is beneficial. As shown in Tab. 8, applying window soft argmax during training hurts the PCK performances with the loose thresholds, while helping the stricter threshold (*i.e.*, PCK@0.01). Our hypothesis is that applying windows during training helps the model focus on the local region but overlook global information.

D.3. Discussion on Generalizability

As shown in the main paper, we validate the generalizability of our method by training on AP-10K intra-species set and testing on cross-species and cross-family subsets. Here, we extend this analysis with additional tests:

Table 7. **Comparison with Gaussian kernel soft argmax on SPair-71k.** Default and peak values for each PCK level are reported for both methods, with the best results **bolded**.

Setting	Method	PCK@0.01	PCK@0.05	PCK@0.10
Default	Kernel	19.7	73.5	84.3
	Window	22.0	75.3	85.6
Best PCK@0.01	Kernel	22.4	75.0	84.9
	Window	**22.6**	75.3	85.0
Best PCK@0.05	Kernel	22.3	75.3	85.3
	Window	22.3	**75.4**	85.5
Best PCK@0.10	Kernel	21.9	75.1	85.5
	Window	22.0	75.2	**85.7**

Table 8. **Effect of applying window soft argmax during training.** We train all the post-processors on SPair-71k for one epoch and from scratch. The best results are **bolded**.

Setting	PCK@0.01	PCK@0.05	PCK@0.10
Window Soft Argmax (7)	21.4	69.0	79.0
Window Soft Argmax (15)	21.8	69.5	80.1
Window Soft Argmax (22)	**22.1**	70.3	81.2
Soft Argmax	20.4	**70.8**	**82.1**

Training on PF-PASCAL and testing on other datasets. We evaluate the generalizability of our method by training it on PF-PASCAL and then testing it on SPair-71k and AP-10K intra-species test sets (see Tab. 10). While previous studies [7, 16] have noted a potential performance decrease due to models' overfitting to the limited distribution of pose variation in PF-PASCAL, our method consistently outperforms across different datasets and PCK thresholds, demonstrating its robustness.

Training on SPair-71k and testing on AP-10K and PF-PASCAL. In a similar vein, we trained our model on the SPair-71k dataset and evaluated its performance on PF-PASCAL and AP-10K intra-species test sets (see Tab. 11). The findings mirrored those from Tab. 10, with our approach achieving the best results across all datasets and PCK metrics, confirming its generalizability again.

D.4. Additional Ablation Analysis under Supervised Setting

To further evaluate the effect of each component on improving semantic correspondence, we conduct a leave-one-out ablation analysis. For an in-depth understanding of the specific improvements, we incorporate the breakdown analysis protocol from "Demystifying" [3] into our ablation study. This analysis introduces four metrics, as delineated in [3]: 1) *Jitter*: the ratio of matches near their correct locations; 2) *Miss*: the ratio of points incorrectly matched to the back-

Table 9. **Leave-one-out ablation study on SPair-71k.** We report the *per image* results and four metrics introduced in [3] (*i.e.*, Jitter, Miss, Swap, and PCK†) for a detailed analysis of the effect of each module. The best results are **bold**.

Variations	Jitter↓	Miss↓	Swap↓	Swap^{LR}↓	PCK†@0.1↑	PCK@0.01↑	PCK@0.05↑	PCK@0.1↑
SD+DINO (S) **(Baseline)**	9.7	13.7	15.8	9.4	70.5	9.6	57.7	74.6
w/o Dense Training Objective	8.3	11.8	13.7	8.5	74.5	15.2	64.5	78.3
w/o Pose-variant Augmentation	7.4	10.0	13.9	8.7	76.1	19.0	70.3	81.5
w/o Perturbation & Dropout	6.9	9.9	12.3	7.2	77.8	20.3	71.8	82.3
w/o Window Soft Argmax	8.1	9.8	14.1	8.7	76.1	15.1	69.3	81.3
Ours	6.9	9.3	12.0	7.0	78.7	21.6	72.6	82.9
Ours w/ AP-10k Pretraining	**6.1**	**8.7**	**10.4**	**5.6**	**80.9**	**22.0**	**75.3**	**85.6**

Table 10. **Generalizability test with training on PF-PASCAL.** We test the generalizability of our method by training the model on the PF-PASCAL dataset and testing on the SPair-71k and AP-10K intra-species (I.S.) test set. The best results are **bold**.

	SPair-71k			AP-10K-I.S.		
Method	0.01	0.05	0.10	0.01	0.05	0.10
SCorrSAN [16]	1.5	18.4	32.7	-	-	-
CATs++ [7]	2.1	19.7	32.0	-	-	-
DHF [30]	4.6	30.1	41.8	7.3	37.0	49.1
SD+DINO (S) [60]	**5.3**	34.1	46.9	8.2	43.4	59.2
Ours	**5.3**	**37.1**	**54.3**	**10.1**	**44.0**	**62.5**

Table 11. **Generalizability test with training on SPair-71k.** We test the generalizability of our method by training the model on the SPair-71k dataset and testing on the PF-PASCAL and AP-10K intra-species (I.S.) test set. The best results are **bold**.

	PF-PASCAL			AP-10K-I.S.		
Method	0.05	0.10	0.15	0.01	0.05	0.10
SCorrSAN [16]	54.5	71.2	78.8	-	-	-
CATs++ [7]	54.8	68.7	76.1	-	-	-
DHF [30]	64.2	77.8	84.0	9.3	42.0	55.2
SD+DINO (S) [60]	68.9	81.7	87.2	9.7	50.4	65.9
Ours	**74.0**	**85.3**	**89.7**	**16.5**	**56.7**	**70.2**

ground; 3) *Swap*: the ratio of matches that are in the correct area but nearer to a different semantic part; 4) *PCK*†: the PCK metric adjusted to exclude *Swap* errors. For comprehensive details, please see Sec. 4.1 of [3]. To advance our evaluation of geometry-aware correspondence further, we introduce an additional metric, *Swap*^{LR}, for geometric confusion (left/right) cases.

As shown in Tab. 9, our method significantly improves *Jitter*, *Miss*, *Swap*, *Swap*^{LR} by 37.1%, 36.5%, 36.0%, and 40.4%, respectively. Specifically, the integration of spatial context through our proposed dense training objective and the window soft argmax technique notably boosts the performance for *Jitter*, *Swap*, and *Swap*^{LR}, which relies on detailed spatial understanding. Besides, the dense training objective also contributes largely in overcoming the *Miss* error, we hypothesize that the soft argmax operator in dense training objective can effectively suppress the background noise. Moreover, by encouraging the pose-awareness, the proposed pose-variant pair augmentation notably reduces both the *Swap* errors, and especially the geomety-aware *Swap*^{LR} error.

In summary, the improvement in *Swap*^{LR} metric further validates the effectiveness of our designs in improving the geometric-awareness of the pretrained features, while the gain in *Miss* showcases that our method also reduces mismatches to the image background.

D.5. Ablation Study under Unsupervised Setting

In the main paper, our zero shot method consists of two techniques: adaptive pose alignment and window soft argmax. In this section, we further ablate different techniques to evaluate the effectiveness of each module under the unsupervised setting.

As shown in Tab. 12, our adaptive pose alignment technique consistently improves the performance across all five inference settings. Additionally, under both with or without adaptive pose alignment settings, our window soft argmax method consistently boosts the performance on both the geometry-aware subset and standard set, outperforming either the argmax or soft argmax. This further demonstrates the effectiveness of our method.

E. Additional Results

E.1. Alternative Metrics for Pose Alignment

In our adaptive pose alignment method, we leverage the mask of the source image, obtained through an off-the-shelf segmentation method, ODISE [55], to calculate the matching distance. While this mask is solely used for pose alignment and does not restrict the solution space for the target image, we propose more flexible approaches to calculate the metric for pose alignment.

Table 12. **Ablation study under unsupervised setting.** We report the PCK@α_{bbox} results on both the standard set (Std.) and geometry-aware set (Geo.) of SPair-71k. The best performances are **bold**.

Method Variants	Inference Strategy	SPair-71k (Std.)			SPair-71k (Geo.)		
		0.01	0.05	0.10	0.01	0.05	0.10
SD+DINO [60]	Argmax Inference (Default)	7.9	44.7	59.9	5.3	34.5	49.3
	Soft Argmax Inference	6.4	36.5	53.7	6.4	36.5	53.7
	Window Soft Argmax (3)	**10.0**	45.9	60.1	**6.7**	35.5	49.6
	Window Soft Argmax (5)	9.9	**46.3**	60.5	6.6	**35.8**	50.1
	Window Soft Argmax (11)	8.7	45.3	**61.3**	5.5	34.3	**51.1**
SD+DINO [60] **w/ Adapt. Pose**	Argmax Inference (Default)	8.9	48.7	64.2	6.3	39.6	55.0
	Soft Argmax Inference	7.6	40.7	58.4	4.1	29.0	48.2
	Window Soft Argmax (3)	**11.2**	49.7	64.3	**8.3**	40.8	55.4
	Window Soft Argmax (5)	11.1	**50.1**	64.8	8.1	**41.1**	56.0
	Window Soft Argmax (11)	9.9	49.1	**65.4**	6.9	39.5	**56.8**

Table 13. **Effect of different adaptive pose alignment metric.** Alternative approaches with relaxed conditions can achieve very competitive results that are much better than the baseline.

Setting	PCK@0.01	PCK@0.05	PCK@0.10
None **(Baseline)**	7.9	44.7	59.9
Mutual-NN	8.5	47.8	63.1
SAM-mask [22]	8.6	48.5	64.0
ODISE-mask [55] **(Default)**	8.9	48.7	64.2

Firstly, as an alternative to generating masks based on object categories (as in ODISE), we can employ a query-point-based segmentation method, *e.g.*, SAM [22], to obtain the instance mask. Such setting has a more relaxed condition because the semantic correspondence task naturally provides query keypoints of the instance in the source image. Furthermore, we can eliminate the need for masks at all by using the average distance of mutual nearest-neighbor pixels as the alignment metric. As shown in the Tab. 13, both alternative metrics yield highly competitive results, significantly surpassing our baseline.

E.2. Qualitative Results on AP-10K

We show the qualitative comparison of our supervised methods with both unsupervised and supervised versions of SD+DINO [60] on AP-10K intra-species (Fig. 19), cross-species (Fig. 20), and cross-family (Fig. 21) subset.

E.3. Additional Qualitative Results on SPair-71k

In Fig. 22 and Fig. 23, we show the qualitative comparison of our supervised methods with both the unsupervised and supervised versions of SD+DINO [60] on SPair-71k dataset. Our method establishes correct correspondence for challenging cases that previous works cannot handle.

Figure 19. **Qualitative comparison on the AP-10K intra-species set.**

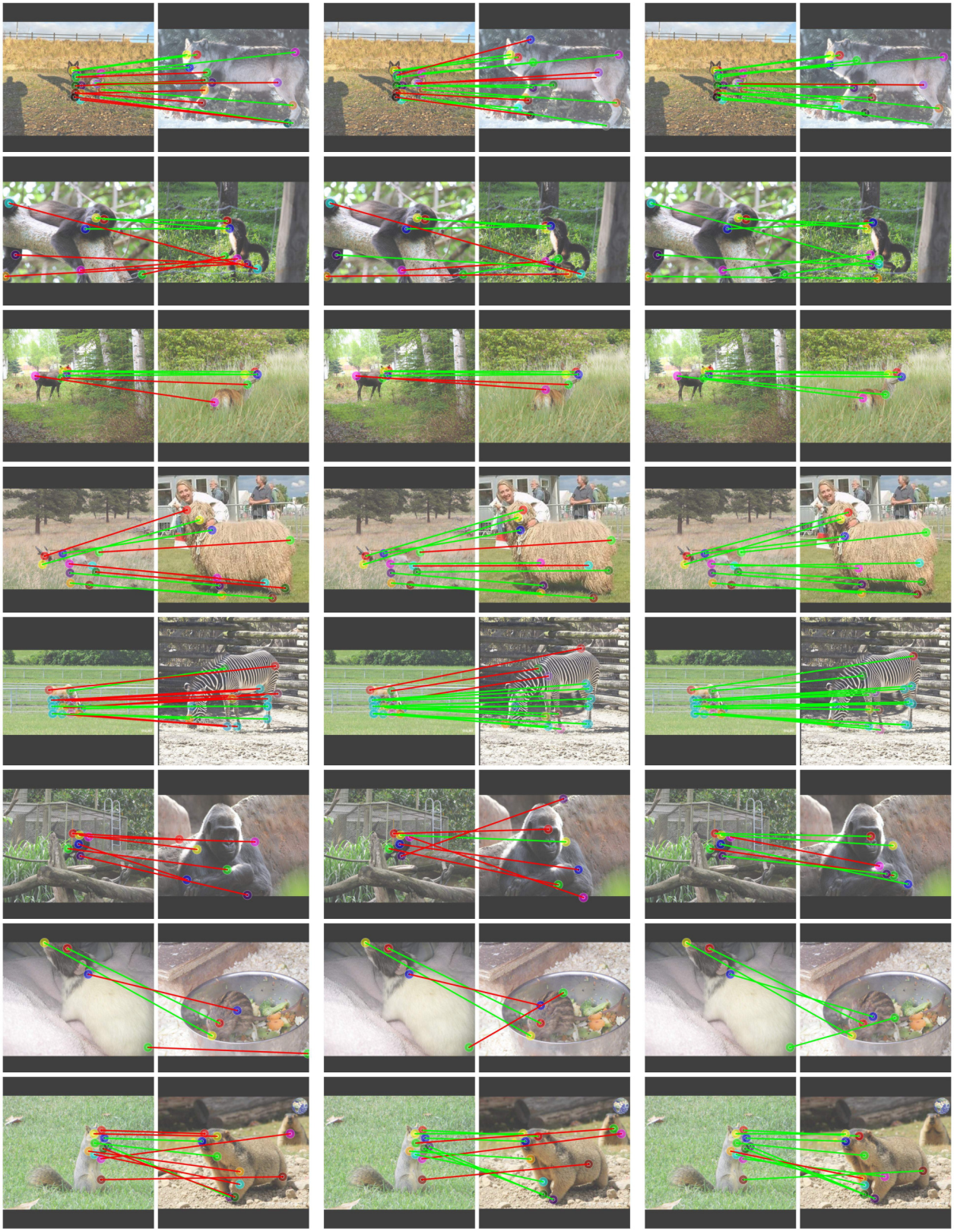


Figure 20. **Qualitative comparison on the AP-10K cross-species set.**

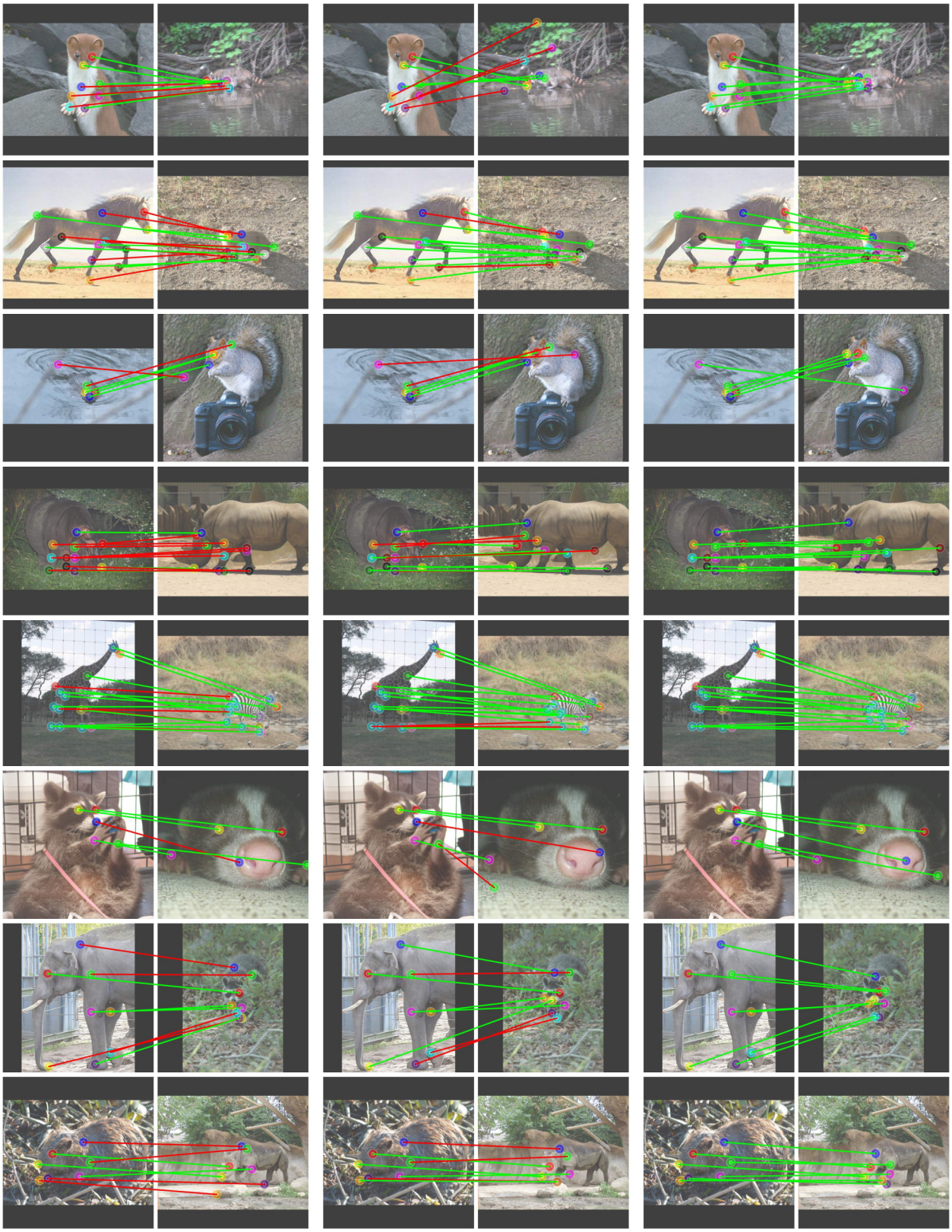


Figure 21. **Qualitative comparison on the AP-10K cross-family set.**

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

Figure 22. **Qualitative comparison on the SPair-71k.** Our method shines even in cases with large viewpoint variations.

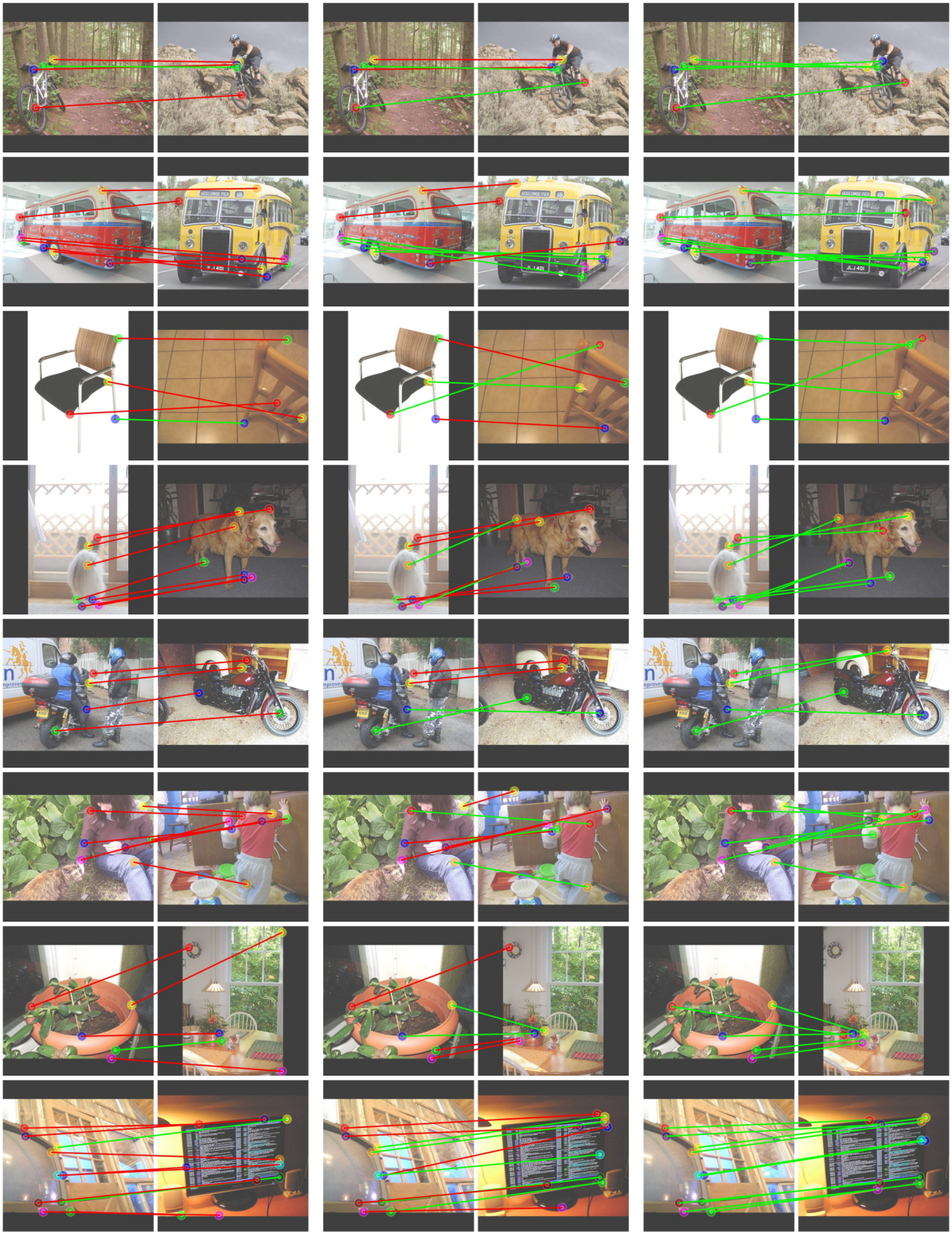


Figure 23. **Qualitative comparison on the SPair-71k.** Our method shines even in cases with large viewpoint variations.