Bridging Viewpoint Gaps: Geometric Reasoning Boosts Semantic Correspondence

Supplementary Material

1. Additional Quantitative Experiments

Evaluation on PF-Pascal. PF-Pascal [1] comprises 2,941 training pairs, 308 validation pairs, and 299 testing pairs across 20 categories of objects. Unlike SPair-71k, the image pairs in PF-Pascal exhibit similar viewpoints and instance poses, focusing on scenarios where objects are more consistently aligned.

We also focus our evaluation on the **rigid body categories** in the PF-Pascal dataset. Specifically, 13 categories in PF-Pascal.

Similarly to the evaluation on SPair-71k dataset, we use **Percentage of Correct Keypoints (PCK)** metric. The PCK measures the proportion of predicted keypoints that are correctly matched to the ground truth within a specified threshold.

A predicted keypoint is considered correct if it lies within a threshold of $\alpha \times \max(h, w)$ pixels from the ground truth keypoint, where $\alpha \in [0, 1]$ (e.g., $\alpha = 0.10$), and h and w represent height and width, respectively. For the **PF-Pascal** dataset, h and w are the dimensions of the entire *image* (α_{img}). In addition, we follow the previous method [5] to evaluate on PF-Pascal using different PCK levels at $\alpha = 0.05, 0.10, 0.15$.

Table 1. **Evaluation on PF-Pascal at Different PCK Levels:** We present the performance on PF-Pascal test split (rigid-body categories only) using PCK *per image* criteria at different thresholds. The highest PCK values are highlighted in **bold** font, and the second-highest values are underlined. $*$: Finetuned on SPair-71k training dataset.

	PF-Pascal		
Methods	0.05	0.10	0.15
DINOv2+NN [3, 4]	61.3	77.1	83.5
SD+DINO [4]	74.8	86.4	90.8
Spherical Maps* [2]	64.9	82.3	89.1
Spherical Maps + SD* [2]	74.1	87.0	91.4
Telling Left from Right [5]	72.4	85.7	90.2
Ours (Geo only)	86.2	91.9	93.8
Ours (Geo + Semantic (DINOv2 ViT-S))	<u>87.0</u>	**92.5**	<u>94.4</u>
Ours (Geo + Semantic (DINOv2 + SD))	**87.1**	<u>92.3</u>	**94.4**

Quantitative result on PF-Pascal. We present our performance in the PF-Pascal test split in Table 1 . Similar to our results on SPair-71k shown in Table **??** , our method **Ours (Geo only)** without using any semantic features achieves improvements of 2.4 percentage points in PCK@0.15, 4.9 percentage points in PCK@0.10, and 11.4 percentage points in PCK@0.05, respectively. When incorporating semantic matching using DINOv2 + SD **Ours (Geo + Semantic (DINOv2 + SD)**, these improvements are further increased to 3.0 percentage points at PCK@0.15, 5.3 percentage points at PCK@0.10 and 12.3 percentage points at PCK@0.05. In particular, unlike the ablation in Table **??**, which demonstrates that in SPair-71k, our method gains more performance with the help of semantic features at stricter thresholds such as PCK@0.05, in PF-Pascal the effectiveness of using additional semantic matching provides a consistent improvement of around 0.7 percentage points across all PCK levels. This observation suggests that in a relatively simple problem setting where viewpoints are similar, our method is effective enough with geometric matching alone. In conclusion, these results indicate that our method not only improves upon other methods by addressing the large viewpoint issue in semantic correspondence, but also achieves significant enhancements on datasets with similar viewpoint image pairs. This highlights that our method provides general improvements over other methods in all circumstances.

2. Ambiguity in Semantic Correspondence

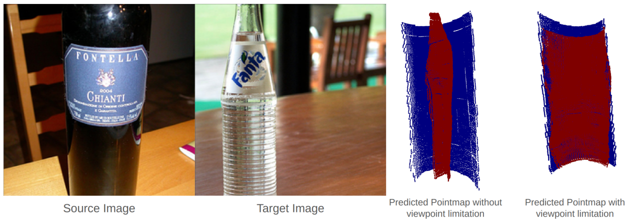
Source Image
Target Image
Predicted Pointmap without viewpoint limitation
Predicted Pointmap with viewpoint limitation

Figure 1. Visualization of effectiveness on viewpoint limitation for symmetric object, blue points represent predicted pointmap of source image and red points represent deformed pointmap of target image.

Unlike visual correspondence, where textures and illumination can help in finding correspondences for geometrically symmetric objects, the ambiguity problem in semantic correspondence for symmetric objects (e.g., tables, bottles) is different. For example, even humans struggle to distinguish between the front and back of a table when given a front view of another table. Similarly, during training, when provided with two arbitrary views I^1 and I^2 of two symmetric objects O^1 and O^2 to regress pointmaps into unified 3D coordinates X^1 and X^2, the network may encounter ambiguity. Given a set of views I' with their groundtruth pointmaps X'^{gt}, representing all views similar to I^2 that cause ambiguity, the network tends to predict the average of X'^{gt} to minimize the L_2 loss as visualized in Figure 1. Although some symmetry-aware loss functions have been designed to resolve this issue, we observe that within the same category where symmetric objects exist, there is still a considerable number of objects without symmetry. Therefore, for categories with potentially symmetric objects, we limit the viewpoint change between pairs of images. For example, in the bottle category, we only use image pairs with changes in altitude angle while keeping the azimuth angle the same.

3. Additional Qualitative Experiments

In Figures 2 and 3, we show qualitative comparisons between our method and the unsupervised version of telling left-from-right [5].

References

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

[2] Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. Improving semantic correspondence with viewpoint-guided spherical maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19521–19530, 2024.

[3] 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.

[4] 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.

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

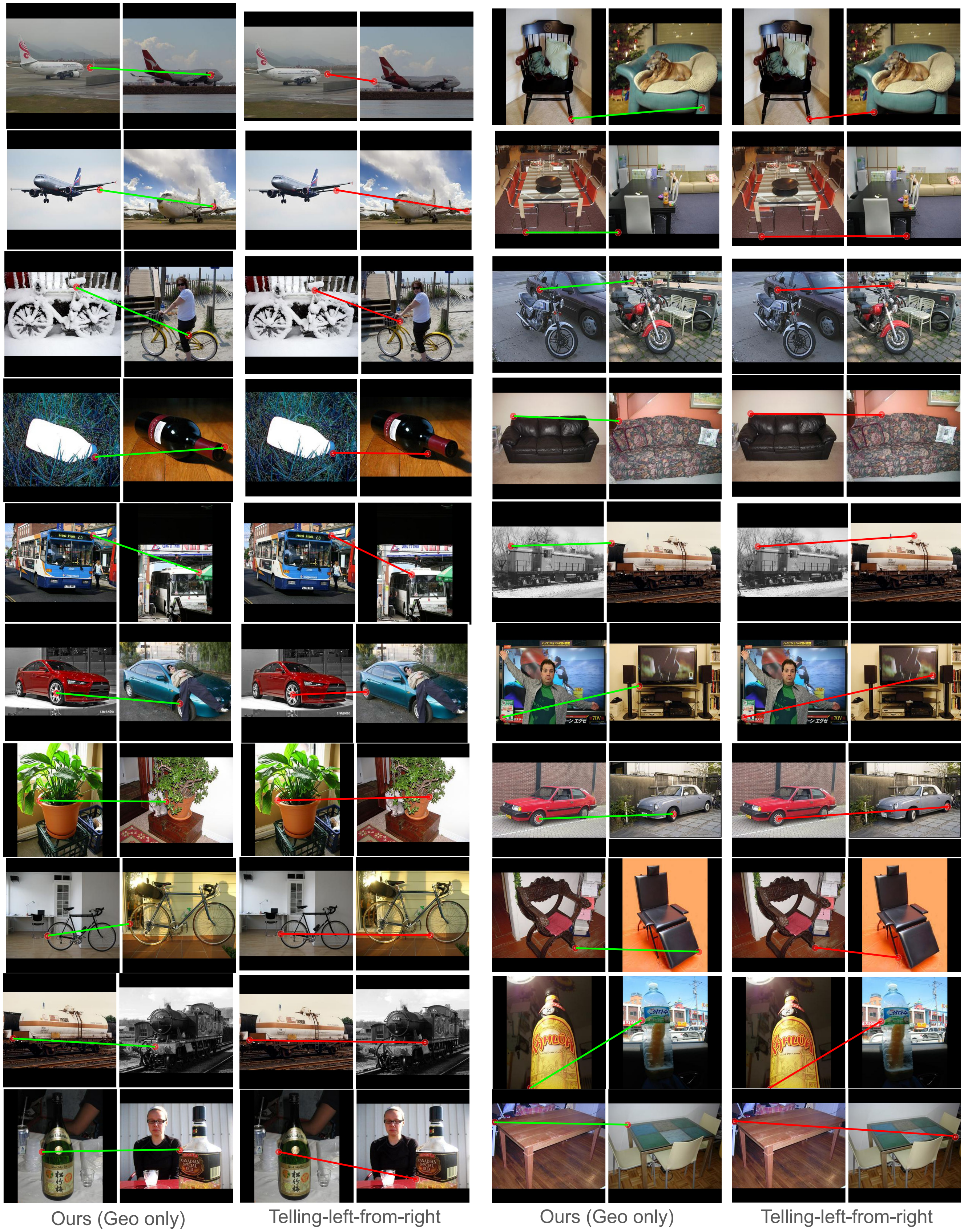

Figure 2. Qualitative Comparison on PF-Pascal

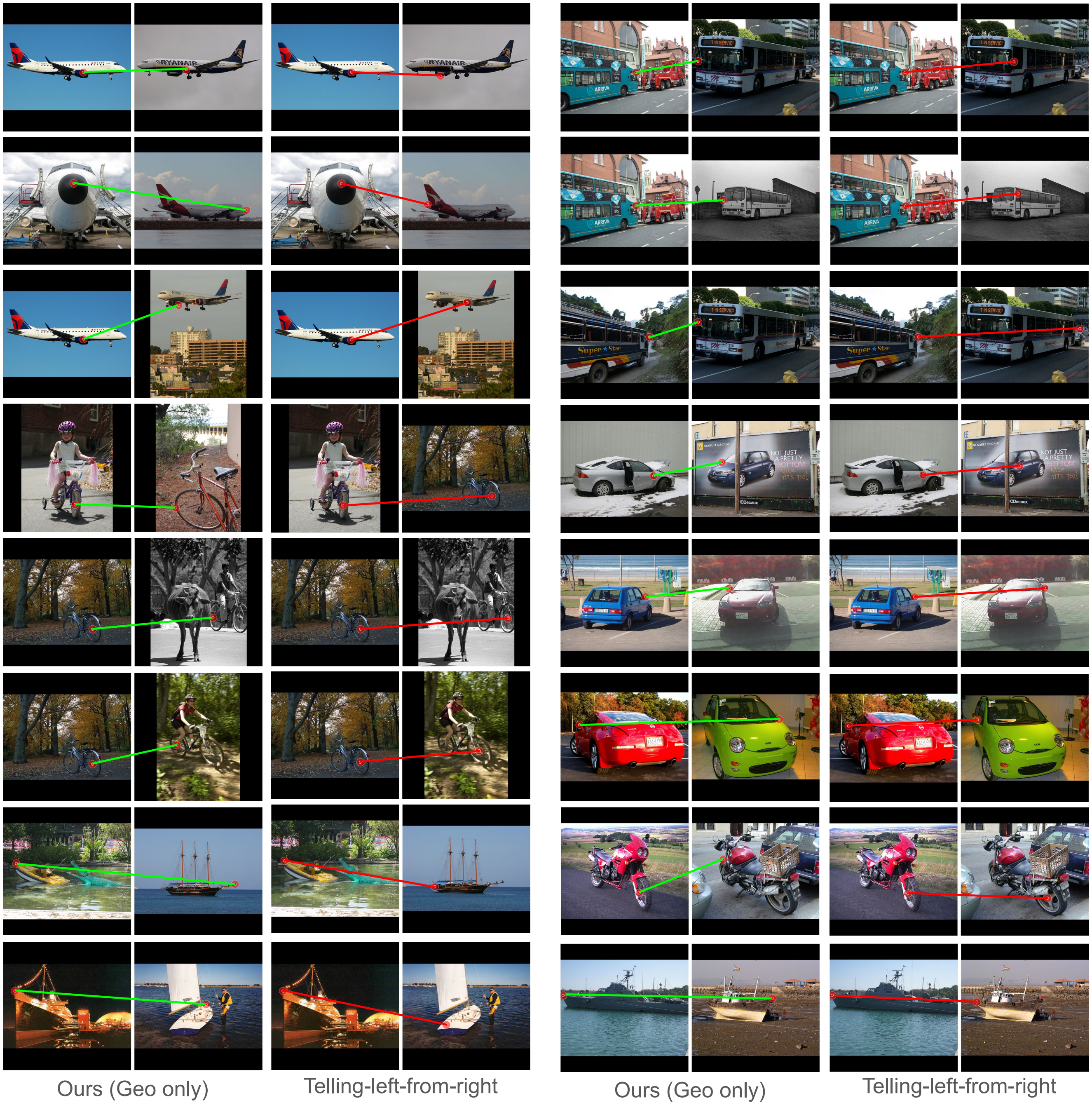

Figure 3. Qualitative Comparison on SPair-71k