SemAlign3D: Semantic Correspondence between RGB-Images through Aligning 3D Object-Class Representations

Supplementary Material

A. Additional Experimental Results

We qualitatively compare our method with DINO+SD [31] and GeoAware [32] on all SPair-71k [22] categories in Fig. 9 and Fig. 10. The results demonstrate that our method is applicable to a wide range of object-classes and can improve performance even if the 3D object-class representation is very coarse.

B. Hyper-Parameters

We list the hyper-parameters in Tab. 5. The hyper-parameters σ_{dense}, σ_{sparse}, and $\sigma_{\text{sparse}}^{\text{inference}}$ correspond to the σ in Eq. (5) for C_{dense} and C_{sparse}, respectively. The hyper-parameter $\sigma_{\text{sparse}}^{\text{inference}}$ is used for sparse semantic correspondence in Eq. (14). As discussed in the paper, it is crucial to start with a high σ and decrease its value over time to obtain a good alignment. Furthermore, we generally decrease σ_{dense} faster than σ_{sparse} to avoid local minima. Additionally, we choose $\sigma_{\text{sparse}}^{\text{inference}}$ large for categories where our spatial prior is rather imprecise. However, as seen in Tab. 1, the imprecise spatial prior can still improve performance.

The weights w_{dense} and w_{sparse} are from Eq. (9) and w_{geom}, $w_{\text{background}}$, and w_{depth} are from Eq. (13). We start with $w_{\text{sparse}} = 0$ and increase its value over time to avoid local minima. Similarly, we also start with a small value for $w_{\text{background}}$ as this term can lead to divergence if the representation does not sufficiently overlap with the object instance in the image, which is the case at the beginning of the optimization.

Group	**Parameter**	**Value**
Optimizer	Type	AdamW
	lr	5e-3
	n_{steps}	1000
Sigmas	σ_{dense}	Timesteps: [0, 300, 500] Values: [1.0, 0.1, 0.03]
	σ_{sparse}	Timesteps: [0, 300, 500, 700] Values (Bottle): [1.0, 0.3, 0.3, 0.05] Values (Other): [1.0, 0.1, 0.1, 0.05]
	$\sigma_{\text{sparse}}^{\text{inference}}$	Chair, TV: 0.01; Airplane, Bicycle, Bottle: 0.03; Other: 1.0
Weights	w_{dense}	1.0
	w_{sparse}	Timesteps: [0, 500, 700, 1000] Values (Bottle): [0, 0, 1, 10] Values (Other): [0, 0, 10, 1]
	w_{geom}	0.5
	$w_{\text{background}}$	Bottle: Timesteps: [0, 700, 900, 1000] Values: [0, 0, 1, 20] Other: Timesteps: [0, 300, 500, 700] Values: [1, 10, 100, 10]
	w_{depth}	10.0

Table 5. Hyper-Parameters. *Values* are linearly interpolated according to *Timesteps*.

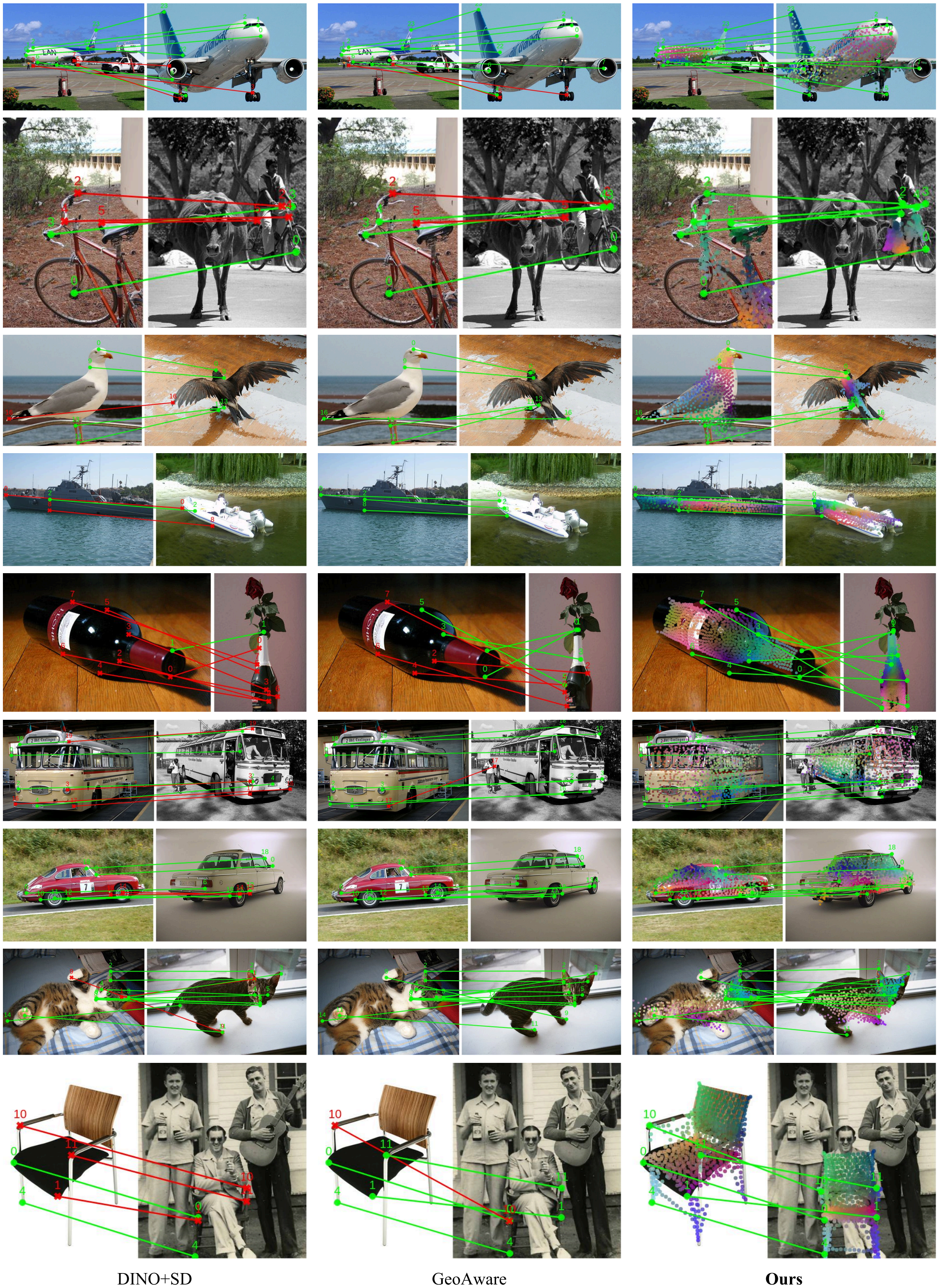

Figure 9. Qualitative comparison on SPair-71k [22] categories for sparse correspondence between DINO+SD [31], GeoAware (AP-10K P.T.) [32], and our method. Green lines (o-o) denote correct matches and red lines depict wrong matches (x-x).

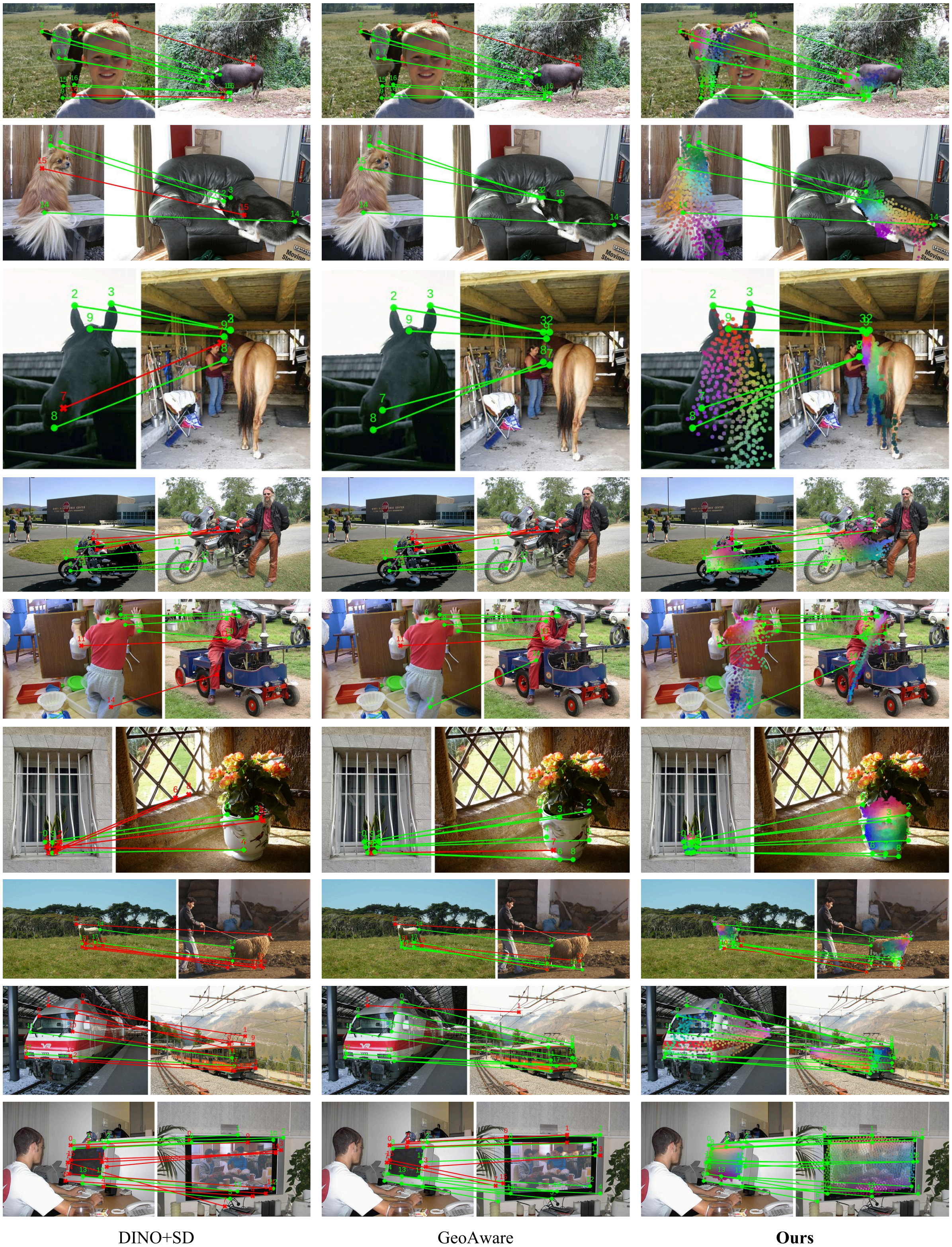

Figure 10. Qualitative comparison on SPair-71k [22] categories for sparse correspondence between DINO+SD [31], GeoAware (AP-10K P.T.) [32], and our method. Green lines (o-o) denote correct matches and red lines depict wrong matches (x-x).