

Appendix

In this supplementary document, we provide results on three additional datasets: PF-PASCAL [1], TSS [6], and Freiburg cars [5]. We also visualize feature maps on SPair-71k [2] and show the results of keypoint matches.

A. Implementation details

For the results labeled as ‘Ours+SD’ (e.g. in Tab. 1 in the main paper), this model use our spherical mapper but at inference time also combines the DINOv2 and Stable Diffusion (SD) features, applying Eq. (7) with the fused DINO+SD features in place of ϕ . We also show architecture details in Fig. A1

B. Continuous surface embeddings







							
PCK	CSE	28.1	23.4	24.1	25.5	86.4	12.1
	Ours	80.0	80.2	66.7	71.2	63.9	68.6
	Ours + SD	82.4	82.0	67.3	73.9	60.0	69.8
KAP	CSE	36.7	30.2	32.8	32.8	72.2	27.1
	Ours	69.7	63.2	54.8	54.3	48.7	47.9
	Ours + SD	76.2	65.3	57.3	56.1	47.4	49.4

Table A1. Keypoint matching scores on SPair-71k evaluated using PCK@0.1 and KAP@0.1 for CSE [3] and our models.

Although not directly comparable, the spherical embedding learned by our model is similar to Continuous Surface Embeddings (CSE) [3], in that it aims to densely represent points on an object’s surface in a smooth way. However, CSE is learned in a fully supervised way, using images densely annotated with correspondences to 3D meshes of object categories. In comparison, we only rely on viewpoint and weak geometric priors. A consequence of this is that our sphere mappings do not necessarily converge to a unique solution, for instance, applying a random rotation to it would still satisfy the geometric constraints that we use during training. This makes it challenging to evaluate our spheres under CSE protocol, as recovering the transformation between them and ground truth meshes is non trivial. Still, CSE can easily be evaluated by keypoint metrics on SPair-71k for the categories on which it was trained by using the publicly released implementation. The PCK and KAP@0.1 are shown in Tab. A1. Despite being supervised, CSE only outperforms our approach on humans, most likely because it is the only category with enough high-quality annotations. Another interesting results is that CSE suffers a much smaller drop in performance when evaluated with KAP instead of PCK, as its fully supervised regime makes it more robust to object symmetries and repetitions, constituting a strong argument in favor of using KAP when evaluating correspondences.

DINOv2	0.1, 0.1, 0.1	0.3, 0.3, 0.3	1, 1, 0.3	0.1, 0.1, 0.03	0.3, 0.3, 0.1
56.2	62.2	62.5	61.4	62.0	63.6

Table A2. Average PCK@0.1 on SPair-71k for different values of geometric losses.

C. Additional ablation results

We investigate the balance between the different geometric losses to assess the sensitivity of our model. In Tab. A2, we show PCK@0.1 scores when training and testing on SPair-71k where we set the losses to the same weight, or globally increase or decrease the weight using the following color-coding: λ_{rd} , λ_o , and λ_{vp} . We observe that altering the losses balance slightly decreases performances, but all models consistently beat the DINOv2 backbone by a large margin.

D. Additional datasets

Freiburg cars. SPair-71k training data is relatively limited, i.e. it contains approximately only 50 training images per category. In order to explore the behavior of our model when more data is available, we trained it using images from the Freiburg cars dataset [5]. Freiburg cars contains 46 scenes each centered around a single car, and there is an average of 120 images sampled from 360° around each car. As it comes with precise viewpoint annotations, we can use it to study the sensitivity of our model to the granularity of the viewpoint supervision. We discretize the camera viewpoint supervision into different numbers of discrete bins (e.g. four bins would correspond to the camera viewing the car from the front, back, and two sides) and evaluate these models on SPair-71k car test pairs.

Our model trained and tested on SPair-71k from Tab. 1 in the main paper obtains at PCK@0.1 of 67.2 on cars. The results in Tab. A3 show that there is no significant benefit from having even finer-grained viewpoint supervision beyond a certain number of bins. The best performing model trained on Freiburg cars improves PCK@0.1 by 4.6 points compared to SPair-71k training. This illustrates the potential of adding additional training data even when the viewpoint supervision is coarse.

As Freiburg cars scenes are densely sampled, we can also use them to qualitatively assess the consistency of feature maps under viewpoint changes. Images in Fig. A2 show strong consistency of the maps across the whole viewpoint range, while maintaining semantic consistence between visually different instances for our sphere mapper. Note, our results in Fig. A2 are for our model trained on SPair-71k.

PF-PASCAL & TSS. We also evaluate our model on PF-PASCAL [1] and TSS [6]. As these sets exhibit less challenging pose variations compared with SPair-71k, the benefit of using spherical maps is more limited, as it can only

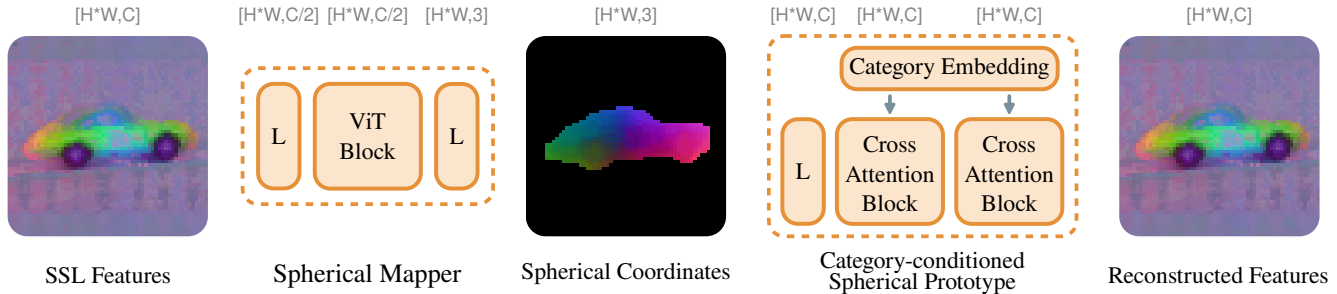


Figure A1. Architecture details for our sphere mapper and spherical prototype. C denotes the dimension of the SSL embedding, and block marked with L are simple linear layers used to change dimensionality.

# bins	4	8	16	32	64	128	360
PCK@0.1	60.1	71.8	71.1	71.2	69.2	68.1	71.0

Table A3. Impact of viewpoint supervision granularity. Here we train with coarse-to-finer discretized poses from Freiburg cars and evaluate on the car category in SPair-71k. Only when using very few bins (*i.e.* four) does the performance significantly drop. This indicates that our approach is capable of training on relatively weak pose supervision. For context, for the results in the main paper, we use the eight viewpoint bins provided by the SPair-71k annotations.

	PF-PASCAL, PCK@ κ			TSS, PCK@0.05			
	0.15	0.10	0.05	FG3DCar	JODS	Pascal	avg
DINOv2 [4]	61.1	77.3	83.3	82.8	73.9	53.9	72.0
SD [7]	61.0	83.3	86.3	93.9	69.4	57.7	77.7
DINOv2 + SD [7]	73.0	86.1	91.1	94.3	73.2	60.9	79.7
Ours	66.2	83.9	90.2	83.1	74.1	54.4	75.5
Ours + SD	74.0	88.4	92.6	95.3	78.7	64.2	82.3

Table A4. Scores for PF-PASCAL and TSS.

help separate repeated parts that appear in the same image and not issues due to large pose variation (as they are not present). Nonetheless the results in Tab. A4 indicate that our spherical maps yield consistent improvements.

E. Additional qualitative results and failure cases

In Fig. A3 we present qualitative results illustrating keypoint matching on some particularly hard SPair-71k evaluation pairs that exhibit large camera viewpoint differences. For each keypoint in a source image, we show where its matched nearest neighbor lies in the target image. These results show that our spherical maps make fewer mistakes on repeated parts, and are more likely to predict points on the correct side of objects in instances where there is visual ambiguity. It is particularly visible on the car example, where all models but ours map the left side of the source car to the right side of the target car, as they both appear on the same side of the image.

Our model still makes mistakes, though these are also present in other models. In particular, it struggles in the presence of large object scale variation (cow), confuses

quadruped legs (horse), and deals poorly with large intra-class shape variations (chair). A limitation of our model is the confusion it makes between legs of quadrupeds. However, these mistakes are also present in other models.

F. Supplementary video

Finally, our [project website](#) contains supplementary videos where we compare to different methods using held-out image sequences. While the results demonstrate predictions on images from short video sequences, the models do not use any temporal information, and in our case of our approach we are training on the held-out SPair-71k training set. It is very apparent in the video that the baseline methods confuse the different sides of the cars and horses, in addition to generating the same features for the different wheels of the car. This is evident by the fact that these distinct parts have the same color which is obtained by performing PCA to reduce the features to three dimensions. In contrast, our spherical-based approach attempts to map each point on the surface of the object to unique features. Note, that we show images from the same car sequences as in Fig. A2 where PCA is computed over images from those sequences, but in the video PCA is computed over five sequences.

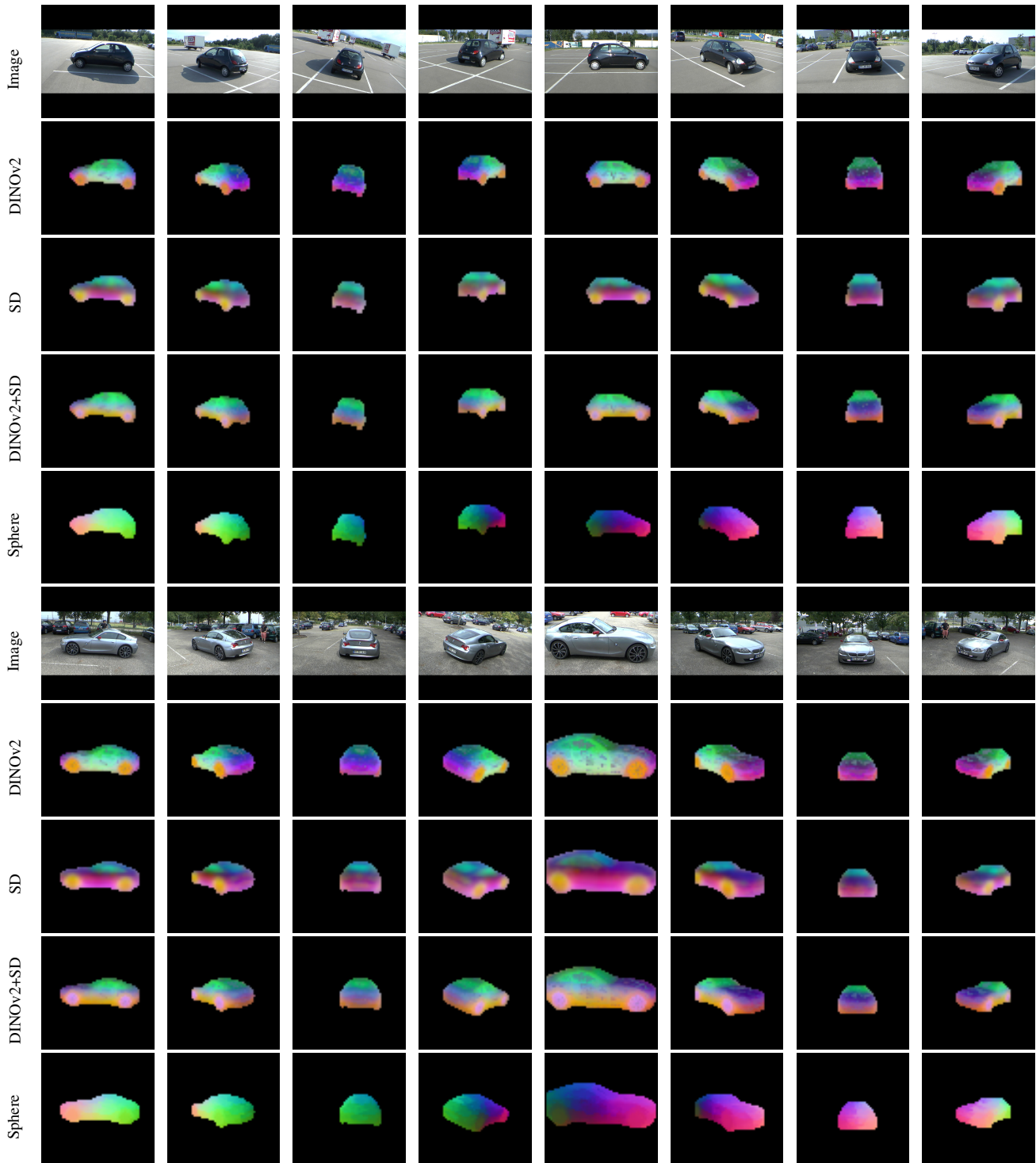


Figure A2. Here we illustrate the multi-view consistency of our approach at test time on two different car sequences from the Freiburg cars dataset [5]. For each sequence, we show input images from different view points, DINOv2, Stable Diffusion (SD), and DINOv2+SD PCA feature maps, and our predicted spherical maps. While other models capture semantic parts, in contrast to us, they fail to correctly disambiguate the two different sides of each car resulting in the same features for the left and right sides. They also fail to produce distinct features for individual car wheels. Note, these large viewpoint changes are typically not assessed in the Spair-71k [2] benchmark. Please see the supplementary video for 360° videos.

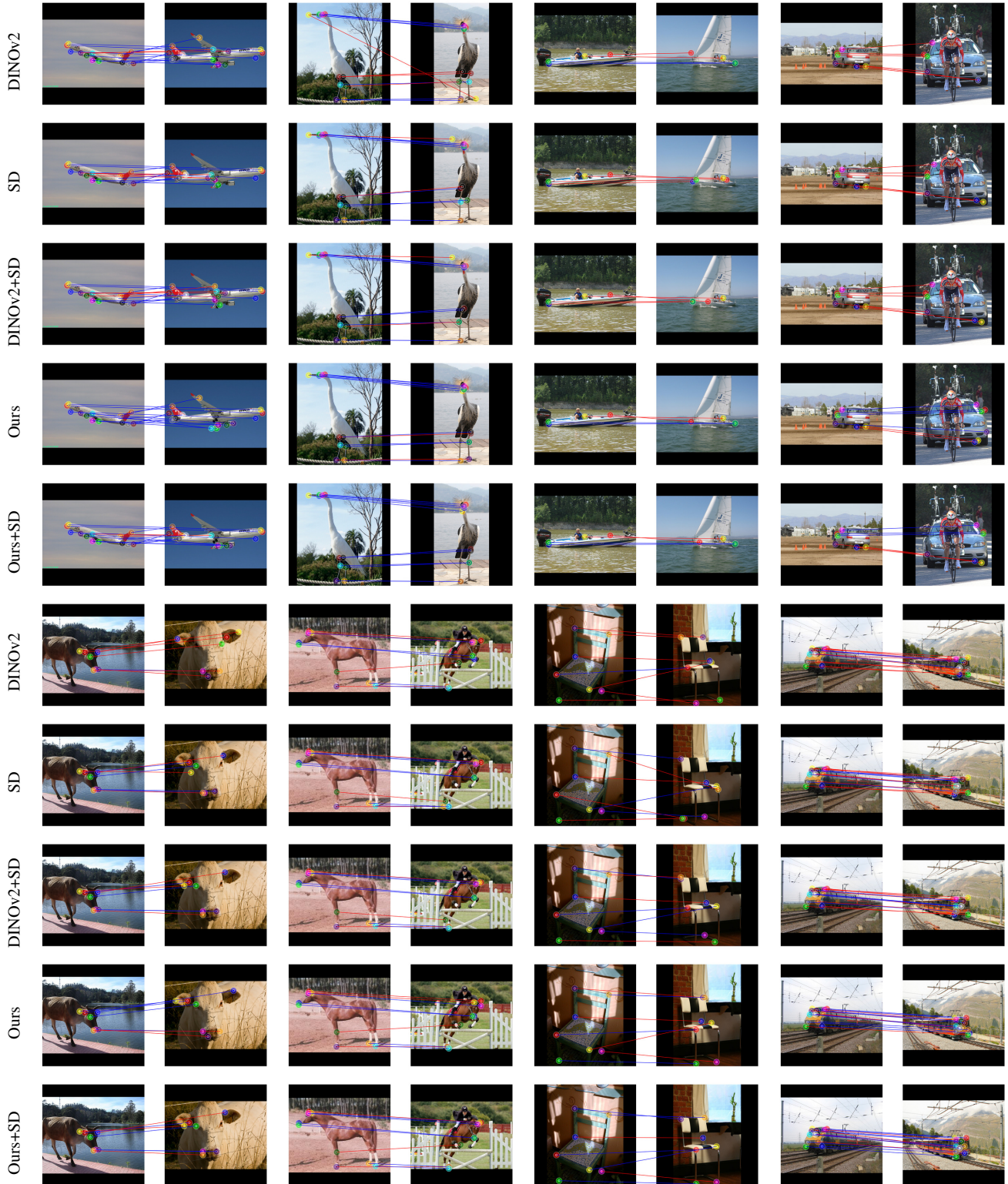


Figure A3. Keypoint matching on SPair-71k [2]. In each pair, the left image acts as the source, from which keypoint features are sampled, then the nearest neighbor of each feature is computed on the target to its right. Blue lines indicate correct matches, *i.e.* within a $\tau = 0.1$ threshold of the ground truth, while red lines indicate incorrect matches.

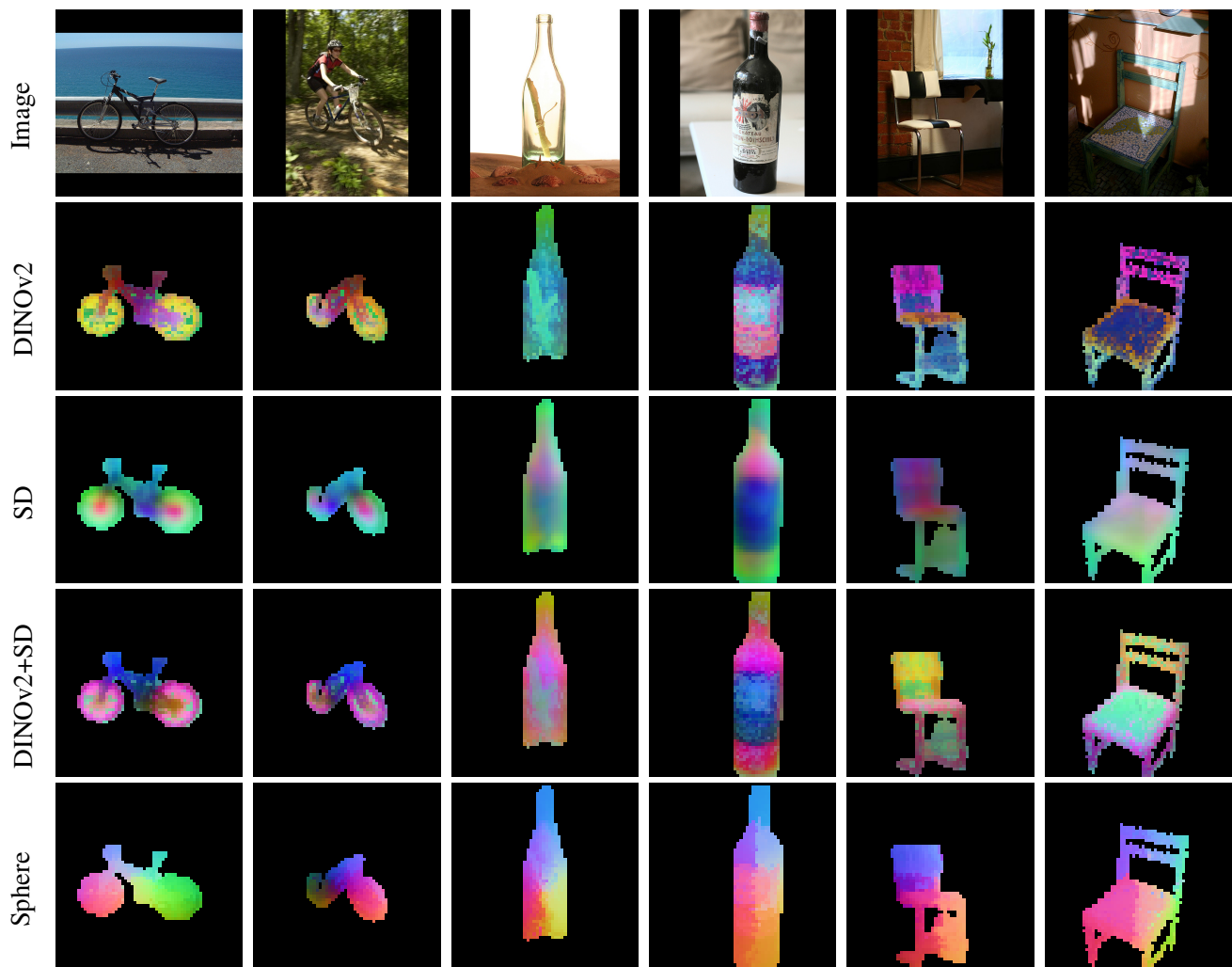


Figure A4. Example dense correspondence maps for categories from the SPair-71k [2] dataset.