Do It Yourself: Learning Semantic Correspondence from Pseudo-Labels

Supplementary Material

A. Details on ImageNet-3D Trained Model

In this subsection, we present more details about the model that was trained on ImageNet-3D, including more details about the improved spherical mapper.

Pose Conversion in ImageNet-3D. We reformulate the loss objective for taking into account viewpoint information as presented in Eq. (10). We acquire the needed labels in the following way: Given the rotation matrix R presenting the 3D pose in the ImageNet-3D dataset [34], we compute the corresponding coordinate on the 2-sphere $\psi = [\theta, \phi] \in \mathcal{S}^2$ as follows:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = R \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

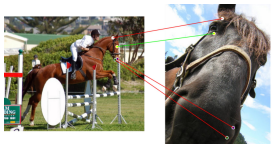
$$\theta = \arccos(z), \quad \phi = \operatorname{atan2}(y, x)$$

Experimental Details For training the spherical mapper, we use the same hyperparameter as Mariotti et al. [35] and we train for 200 epochs on the ImageNet-3D dataset. The generation of the pseudo-labels and training of the adapter follow the same hyperparameters as our presented model for SPair-71k. We train on this larger dataset for 400k steps.

B. Discussion of Failure Cases

While we show significant improvements for most object categories, our method does not improve results for all classes compared to the SOTA. Our approach performs worse when objects are vertically flipped, *e.g.*, for the airplane category or bicycles, as presented in the challenging example in Tab. 6a. In such cases, a limiting factor is that the polar angle is not available for 3D-aware sampling and training of the spherical mapper. Our method also fails for heavy perspective changes Tab. 6b.

(a) Our model can fail for objects that are up-side down.



(b) Matching objects under heavy scale changes is challenging.

Table 6. **Examples of failure cases.**

C. Training with Less Supervision

To explore the scalability of our proposed strategy to larger datasets without 3D pose annotations or masks, we study how the performance deteriorates for SPair-71k when not having access to the viewpoint annotation. For this, we extract pose information using Orient-Anything [54] and extract object masks using `rembg` for unsupervised foreground extraction. Using SAM masks, which are of slightly higher quality but not completely unsupervised (although feasible through, e.g., using GroundingSAM), and the automatically extracted poses, the PCK@0.1 *per-img* of our method drops to 69.6% compared to when using GT pose annotations (71.6%). Using `rembg` masks, it drops further to 68.0%, which is expected but still around 2.7p better than the previous best weakly supervised method [9]. While [47] uses dataset-specific information about the keypoint label convention, our approach only requires class labels.We report the results in Tab. 7.

3D pose label	mask label	PCK@0.1
Ground Truth	SAM	71.6
Orient-Anything	SAM	69.6
Orient-Anything	`rembg`	68.0

Table 7. **PCK@0.1 *per-image* on SPair-71k without ground-truth viewpoint annotations.**

D. Pre-Training with Pseudo-Labels

We explore whether pre-training with our pseudo-labels also improves the supervised performance. For this purpose, we fine-tune the adapter with ground truth labels of the SPair-71k dataset, which improves the supervised performance from 82.9% [63] to 83.5% (PCK@0.1 *per-img*).

E. Pseudo-Label Generation without SD

When training a refiner of DINOv2 features with pseudo-labels that are acquired only from DINOv2 features, i.e., not from SD+DINOv2 as in the main paper, the performance drops to 67.17% (*per-img*) and to 70.29% (*per-kpt*), which is still on par with recent SOTA models.

Channels	PCK@0.1	PCK@0.05	PCK@0.01
128	74.08	56.28	11.26
384	74.39	56.87	11.53
768	74.43	56.76	11.22
1536	74.63	57.13	11.60

Table 8. **PCK metrics for varying numbers of feature channels.**

F. Ablation of Number of Feature Channels

Learning a refining module allows reducing the number of channels of the features used for nearest neighbor computation. The performance does not heavily drop, which might be a valuable option for memory-constrained applications.

G. More Qualitative Results

We show detected correspondences for uncurated image pairs from the SPair-71k test dataset for DistillDIFT, TLR, SphMap, and our method in Fig. 5 and Fig. 6.

Figure 5. Uncurated image pairs of SPair-71K dataset of three SOTA models and ours.

Figure 6. Uncurated image pairs of SPair-71K dataset of three SOTA models and ours.