Appendix

In this supplementary material, we provide additional details about our approach, experiments and results.

A. Discussion of the Architecture

Here we further discuss the architecture of our model.

Feature Extractor To evaluate the importance of using ViT-DINO features, we compare a number of different pre-trained self-supervised ViTs in Tab. 1. We find that our method works best with ViT-DINO features, followed closely by ViT-MSN [1]. MAE does not seem to learn features with strong semantic correspondence properties, possibly due to its image reconstruction objective.

Method	Backbone	**CUB**	**Spair71k**	**PF-Pascal**
MAE [4]	ViT-16	34.6	18.4	54.7
MSN [1]	ViT-16	62.1	29.5	64.8
DINO [3]	ViT-16	64.9	–	–
DINO [3]	ViT-8	**66.2**	**36.4**	**73.4**

Table 1. **Comparison of different backbones.** We compare the PCK of a model trained on CUB with different self-supervised ViTB-16 backbones (we use ViTS-8 in our all other experiments, but ViT-8 weights were not available for some of the backbones we evaluate here).

Fine-tuning the Backbone. We found that keeping the ViT-DINO backbone frozen to be essential for our method to work well. If the pre-trained ViT-DINO is fine-tuned together with the transformer adapter, it returns features Φ that minimize the contrastive loss in Eq. 2 but have lost their semantic power. We thus keep the teacher frozen in all our experiments. Moreover, keeping the features used in the distillation loss frozen, while fine-tuning the features that go into the transformer adapter, does not lead to any significant gains in performance.

Additional ablations We ablate the values of the temperature parameters τ_H and τ_d on CUB in Table Tab. 2. We see that performance is highly dependent on both parameters.

Value	Ablated param. τ_H	Ablated param. τ_d
1.0	32.4	38.8
0.1	65.3	**66.8**
0.05	**66.8**	66.7
0.01	64.1	66.4

Table 2. **Ablation of τ_H and τ_d .** When ablating τ_H, we keep $\tau_d = 0.1$ and when ablating τ_d, we keep $\tau_H = 0.05$. We ablate on CUB.

B. Implementation details

In the contrastive distillation loss in Eq. 2, the denominator contains negative features from the source and target image. In practice, we sample 200 features from both images due to memory constraints. We found that the model converges faster as more negative features are used, but got diminishing returns when using more than 200.

When selecting pseudo pairs in training, we discard images whose salient region is smaller than 20% of the area of the image. We found that many of the unsupervised masks of [6] were very small and not representing any object, and we use this step to filter such masks.

While during training we keep features at their original 28×28 spatial resolution, during test-time in the final matching step we bilinearly upsample the target features to 128×128, and do this for all baselines and comparisons, too. We found this improves matching performance for small objects across the board.

The adapter has 14M parameters and the FLOPs for a single pass are 15B. We train on a single 48G GPU with batch size 32 for less than 2 days.

C. Qualitative Evaluations

Our model can track keypoints on an object as it undergoes significant appearance changes, as shown in Fig. 1.

In Figs. 2 to 5 we qualitatively compare our method to the baseline ViT-DINO features and the methods from [2, 5] on random examples from PF-Pascal, Spair-71k, CUB, and SDogs. We use green squares for correct matches and red circles for incorrect ones.

We see that as discussed in the paper, PMD performs very well when the transformations between the objects can be approximated by a simple warp, as in the PF-Pascal dataset in Fig. 2. In the evaluations of PMD on Spair-71k, CUB and Sdogs, we use their weakly-supervised SPair-71k pre-trained model. We see that in all these cases the model fails to establish semantic correspondences.

Finally, we note a problem of the SDogs dataset that can be seen in the last column of Fig. 5. In a large number of the pairs, there is more than one dog in the target image. This leads to ambiguous annotated keypoints, where there should be more than one correct correspondence in the target image. Upon closer inspection, our model (and other models as well) predicts correct keypoints, but not on the arbitrarily selected dog with the ground-truth correspondences.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. *arXiv preprint arXiv:2204.07141*, 2022. 1

Figure 1. **Semantic Correspondences.** Given annotations in one frame, our model can find correspondences when the object undergoes significant appearance and pose changes (transformation to a wolf) in the video.

Figure 2. **Qualitative Evaluation on PF-Pascal.** We see that PMD manages to find correct correspondences when the source and target images are aligned.

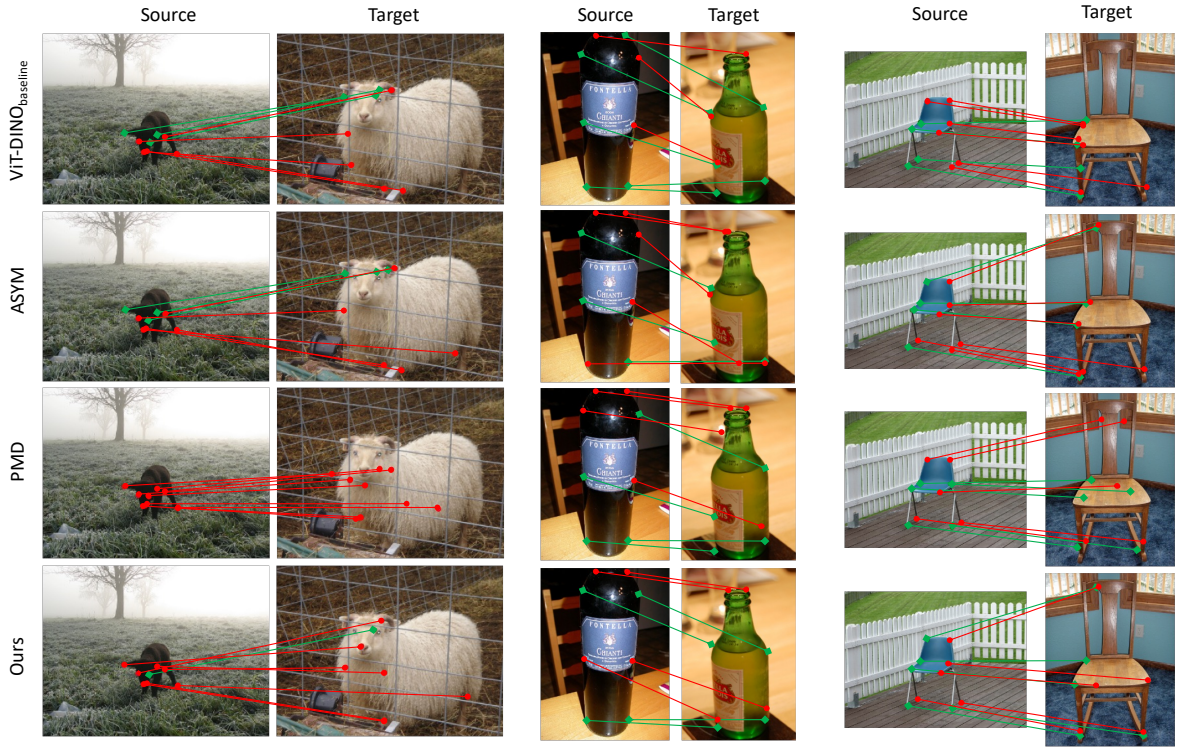

Figure 3. **Qualitative Evaluation on SPair-71k.** We see this is a much more difficult dataset than PF-Pascal.

[2] Mehmet Aygun and Oisin Mac Aodha. Demystifying unsupervised semantic correspondence estimation. In *European*

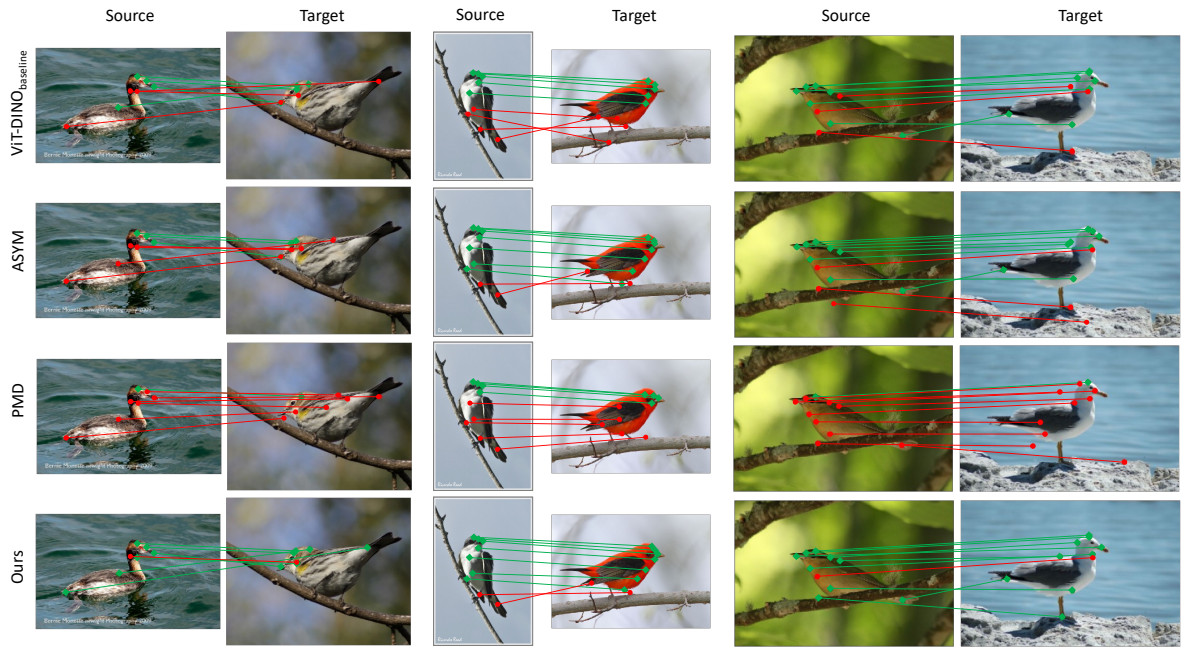

Figure 4. **Qualitative Evaluation on CUB.** Our method improves on the baseline features in almost all example points. We see that PMD fails to establish correspondences when there are large pose changes between the source and target images.

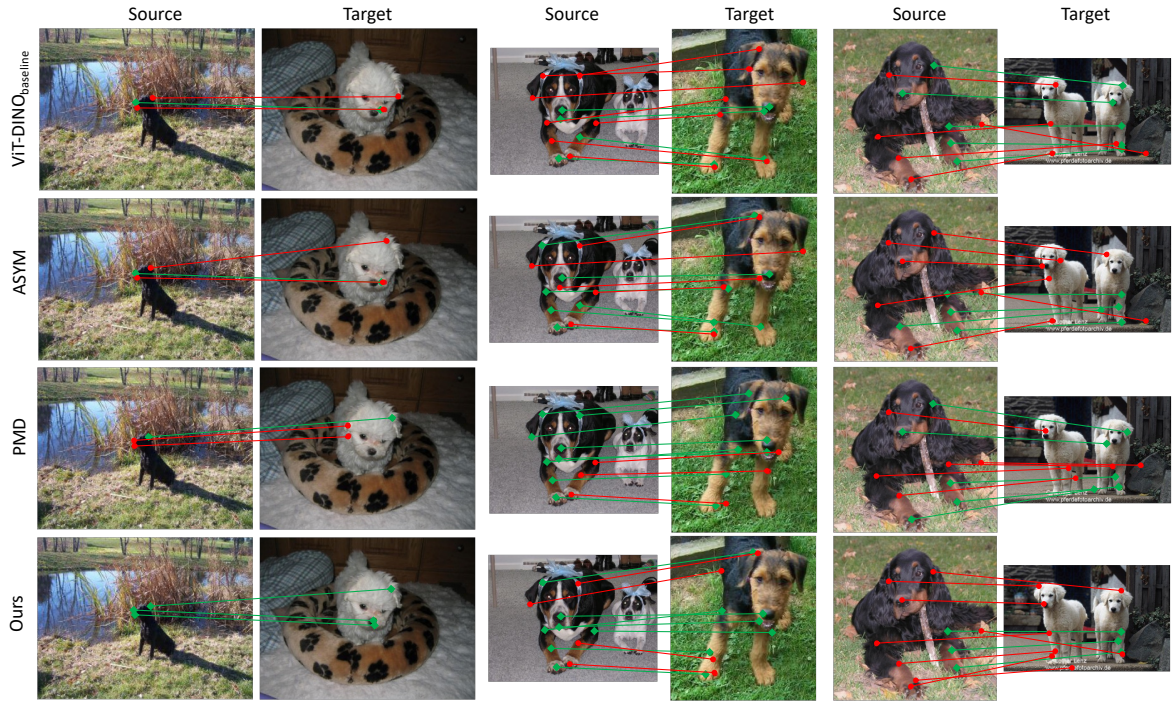

Figure 5. **Qualitative Evaluation on SDogs.** Our method generally improves on the baseline features. We emphasize the last column, where the dog on the right is arbitrarily annotated as the correct match.

Conference on Computer Vision (ECCV), 2022. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. ICCV*, 2021. 1

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *Proc. CVPR*, volume abs/2111.06377,

2021. 1

[5] Xin Li, Deng-Ping Fan, Fan Yang, Ao Luo, Hong Cheng, and Zicheng Liu. Probabilistic model distillation for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7505–7514, 2021. 1

[6] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 1