Supplementary Materials for ECCV 2024 paper PDiscoFormer: Relaxing Part Discovery Constraints with Vision Transformers

Ananthu Aniraj[1], Cassio F. Dantas[2], Dino Ienco[2], and Diego Marcos[1]

[1] Inria, Univ. Montpellier, LIRMM, UMR TETIS, Montpellier, France
{ananthu.aniraj, diego.marcos}@inria.fr
[2] Inria, Inrae, Univ. Montpellier, UMR TETIS, Montpellier, France
{cassio.fraga-dantas, dino.ienco}@inrae.fr

A Training Settings

We trained all our models using the Adam optimizer [7]. The class token, position embedding, and register token of the ViT were kept unfrozen, while all other ViT layers were frozen during training. We used a starting learning rate of 10^{-6} for the ViT backbone fine-tuned tokens, 10^{-3} for the linear projection layer to form the part prototypes, and 10^{-2} for the modulation and the final linear layer used for classification. We used a variable batch size, with a minimum of 16, by adjusting the learning rate using the square root scaling rule [9]. Training lasted for a total of 28 epochs, and we employed a step learning rate schedule, reducing the learning rate by a factor of 0.5 every 4 epochs (as in [8]). Additionally, to regularize our training process, we applied gradient norm clipping [10] with a constant value of 2 for all experiments. In all our experiments, the loss weight of the background loss $\mathcal{L}_{\mathrm{p}_0}$ was set to 2, while all other loss weights were set to a value of 1. We used a constant part dropout value of 0.3.

B Compute Requirements and Model Performance

Compute Requirements (Training). We trained our models on a machine equipped with 4 NVIDIA GeForce RTX 2080 Ti GPUs. The training duration varied depending on the dataset and batch size, as detailed in Tab. A.
Inference Speed. Taking the models trained on the CUB dataset [11] with $K = 8$ as an example, the inference speed on a workstation with a single NVIDIA GeForce RTX 3090 GPU, averaged over 100 runs, is presented in Tab. B.

C Effect of backbone

In this experiment, we investigate the impact of different deep neural network backbones on the performance of our method. The results are summarized in Tab. C.

Table A: Training details for PDiscoFormer on different datasets.

Dataset	Batch Size per GPU	Training Time
CUB	8	4 hours
PartImageNet OOD	32	1 hour 10 minutes
Flowers	32	14 minutes

Table B: Inference speed comparison on the CUB dataset with $K = 8$.

Model	Inference Speed (images/second)
Huang [4]	139.31
PDiscoNet [8]	248.07
PDiscoFormer	**326.18**

Effect of Backbone Pre-Training for Part Discovery. From Tab. C, it is evident that our proposed part discovery priors perform optimally on the self-supervised DinoV2 ViT backbone. However, when compared to the PDiscoNet model with the same ResNet backbone, which utilizes the stricter concentration loss prior [5, 8], the resultant PDiscoFormer+R101 model exhibits inferior performance for both part discovery and classification. For instance, on the Oxford Flowers dataset, the PDiscoFormer+R101 model's training collapses for all tested values of K, resulting in classification accuracies of 8.4%, 7.1%, and 5%, respectively. In contrast, the PDiscoNet+R101 model achieves significantly higher accuracies of 77.5%, 83.1%, and 81% for the same values of K. Despite relatively better performance on the CUB and PartImageNet OOD datasets, the PDiscoFormer+R101 model still lags behind the self-supervised ViT and related methods from the literature. These results indicate that a strong part shape prior, such as the concentration loss, is indeed required to obtain consistent part maps for ImageNet-pretrained CNN backbones. Moreover, they highlight the crucial role of the strong inductive biases that the ViT model learns during the self-supervised pre-training stage, enabling the use of a more flexible geometric prior such as the total variation loss.

Frozen vs partially fine-tuned ViT. We observe from Tab. C that our method performs best when we fine-tune the position embeddings, class, and register tokens along with our additional layers while keeping the rest of the ViT frozen. Although our losses can still operate on a completely frozen ViT, as shown in Tab. C, the performance is generally a bit lower, particularly for higher values of K. For instance, on the CUB dataset with 16 parts, our method with the partially fine-tuned ViT achieved 55.8% ARI, 73.4% NMI, and 88.7% classification accuracy, compared to 50% ARI, 69.5% NMI, and 85.1% classification accuracy for our method with the fully frozen backbone. Similarly, on the PartImageNet OOD dataset with 50 parts, our method with the partially fine-tuned ViT achieved 62.2% ARI, 46.3% NMI, and 91% classification accuracy, compared to 57.9% ARI, 44.5% NMI, and 90.6% classification accuracy for our method with the fully frozen backbone. These results indicate that some

Table C: Performance of our method and the state-of-the-art method from the literature [8] under different backbone configurations.

Method	CUB (%)					PartImageNet OOD (%)				Flowers (%)		
	K	Kp ↓	NMI ↑	ARI ↑	Top-1 Acc. ↑	K	NMI ↑	ARI ↑	Top-1 Acc. ↑	K	Fg. mIoU ↑	Top-1 Acc. ↑
PDiscoNet [8] + R101	4	9.12	37.82	15.26	86.17	8	27.13	8.76	88.58	2	19.04	77.51
	8	8.52	50.08	26.96	86.72	25	32.41	10.69	89.00	4	34.76	83.05
	16	7.67	56.87	38.05	87.49	50	41.49	14.17	86.06	8	49.10	81.04
PDiscoFormer + R101	4	11.07	34.32	16.94	82.59	8	11.44	29.64	87.33	2	0.89	8.41
	8	8.27	44.59	25.63	84.25	25	13.86	27.55	88.42	4	0.08	7.12
	16	9.53	35.64	17.61	83.84	50	7.91	19.64	88.78	8	0.00	4.96
PDiscoNet + ViT-B	4	7.70	52.59	26.66	88.61	8	19.28	34.72	90.95	2	4.92	92.75
	8	6.34	65.01	37.90	86.95	25	28.23	50.35	90.29	4	1.95	95.48
	16	5.95	68.63	43.41	84.04	50	29.48	27.80	89.69	8	13.18	97.40
PDiscoFormer + frozen ViT-B	4	8.19	52.88	23.22	88.87	8	28.84	55.66	90.35	2	67.28	99.41
	8	6.23	67.59	41.35	88.56	25	43.36	62.82	90.47	4	58.71	99.43
	16	6.44	69.54	49.99	85.10	50	44.48	57.91	90.59	8	72.27	99.53
PDiscoFormer + partially fine-tuned ViT-B	4	7.41	58.13	25.11	89.06	8	29.00	52.40	89.75	2	**73.62**	99.61
	8	6.54	69.87	42.76	**89.41**	25	44.71	59.27	90.77	4	73.32	99.54
	16	**5.74**	**73.38**	**55.83**	88.72	50	**46.29**	**62.21**	**91.01**	8	69.59	**99.64**

level of fine-tuning, combined with our proposed training objective function, is beneficial for discovering consistent parts from the self-supervised ViT.

Qualitative Analysis. For our qualitative analysis, we examine the results obtained for the CUB ($K = 8$ parts), Oxford Flowers ($K = 2$ parts), and PartImageNet OOD ($K = 8$ parts) datasets, as depicted in Fig. A, Fig. B, and Fig. C, respectively. In Fig. A, our model with the partially fine-tuned self-supervised ViT (last row) demonstrates the most consistent results for part discovery and exhibits superior segmentation of discovered parts, such as the bird wings. The frozen ViT model (second last row) generally identifies consistent parts but occasionally misassigns background regions, such as the sky, as foreground parts. Notably, models with the self-supervised ViT backbone tend to more accurately segment foreground image regions and bird wings, indicating the effectiveness of representations learned during self-supervised pre-training for part discovery. Turning to Fig. B, only our models with the partially fine-tuned and frozen ViT (last two rows) successfully identify consistent parts. Among these, the partially fine-tuned model achieves better segmentation of the flowers and produces more semantically interpretable part assignments, with one part clearly corresponding to the flower calyx and the other to the corolla. Finally, in Fig. C, our model with the partially fine-tuned ViT (last row) demonstrates the most consistent and semantically interpretable parts. For instance, the blue part corresponds to the upper part of the animal's face, the green part to the snout, and the orange part to the mouth. The frozen ViT model (second last row) generally localizes the object of interest well but produces parts that are qualitatively more challenging to interpret. For example, it assigns the same part (red color) to the face of a bear and an alligator, although this part corresponds to the snout for dogs and

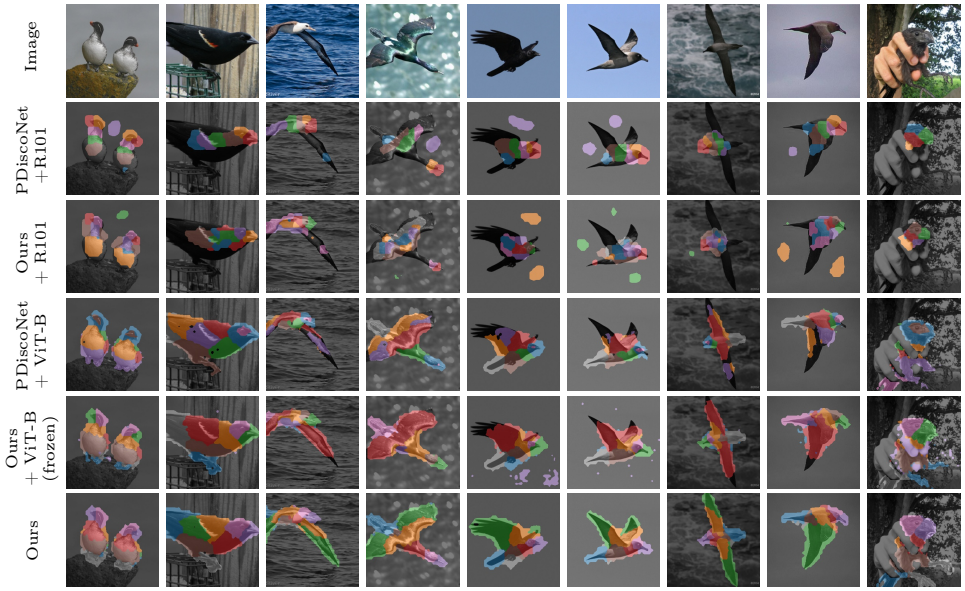

Fig. A: Qualitative results on CUB for $K = 8$.

leopards. Additionally, the PDiscoNet+ViT-B model consistently discovers the mouth of the animal as a part (blue color) and assigns the upper part of the face to the same part (orange color) in most cases. However, some of the other discovered parts (red and green colors) are less consistent and harder to interpret. Furthermore, this model achieves lower-quality segmentation of foreground parts due to the compactness prior of the concentration loss. In comparison, the PDiscoNet model and our model with the ResNet backbone find parts that are even more inconsistent, making them harder to interpret. Additionally, these models demonstrate poorer segmentation of the salient object in the image compared to our models with the frozen and partially fine-tuned ViT.

D Entropy Analysis of Part Attention Maps

We conducted an analysis of the average entropy (see Eq. (12) in the main paper) in the part attention maps of our model ablations in the CUB ($K = 16$) and PartImageNet OOD ($K = 25$) datasets. The results are summarized in Fig. D. From Fig. D, it is evident that the Gumbel-Softmax mechanism [6], followed by the total variation loss ($\mathcal{L}_{tv}$) and entropy loss ($\mathcal{L}_{ent}$), significantly contribute to reducing the overall entropy in our part attention maps for both the CUB and PartImageNet OOD datasets. This underscores their crucial role in minimizing ambiguity in our part assignments and highlights the necessity of considering multiple components for effective entropy reduction, rather than solely focusing on minimizing the entropy loss ($\mathcal{L}_{ent}$). By reducing information leakage between

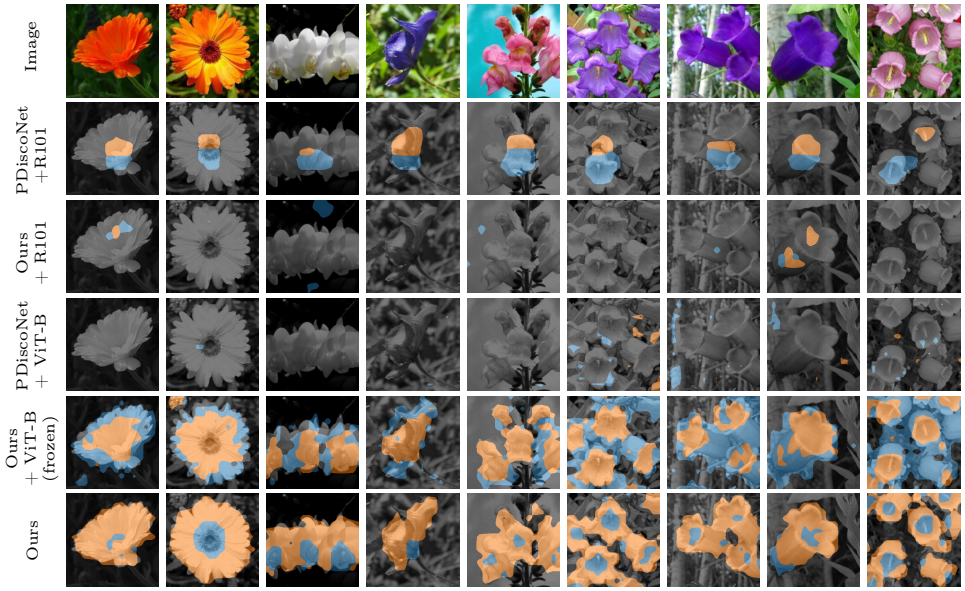

Fig. B: Qualitative results on Flowers for $K = 2$.

background and foreground assignments, lower entropy could help prevent the model from learning spurious correlations from background image regions [1, 2, 12], thereby enhancing model robustness across diverse imaging environments. Additionally, lower entropy between foreground part assignments would ensure that each discovered foreground part is a unique, independent feature. We believe this mechanism allows our model, as demonstrated in our empirical results, to effectively scale for different values of K without any hyper-parameter tuning.

E Experiment on the PartImageNet Seg Dataset

We conducted an experiment on the `PartImageNet Seg` dataset, a recent addition to the PartImageNet family specifically tailored for image classification tasks [3]. This dataset comprises 158 classes organized into 11 super-classes, with a total of 41 part classes, maintaining consistency with the `PartImageNet OOD` version. With a training set comprising 21,662 images and a test set containing 2,405 images, assessing the part discovery and classification capabilities of our method on this dataset can provide further insights into its overall generalization. Given the absence of comparable results in the existing literature for unsupervised part discovery on this new dataset, we conducted our experiments from scratch for all models presented in Tab. D. This approach ensures that our evaluation is thorough and offers valuable insights into the effectiveness of our method in this evaluation setting.

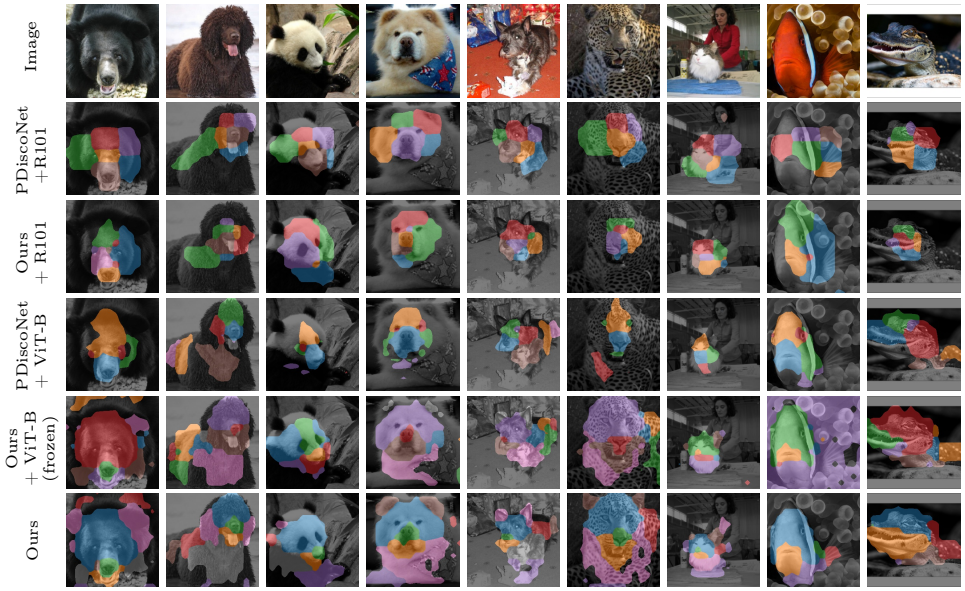

Fig. C: Qualitative results on PartImageNet OOD for $K = 8$.

Quantitative Results. We observe from Tab. D that our method with the self-supervised ViT backbone consistently outperforms the state-of-the-art method in terms of consistency in part discovery, as indicated by the NMI and ARI scores. For instance, for $K = 41$ parts, our method achieves NMI and ARI scores of 44.9% and 60%, respectively, compared to 27.9% and 40% by the PDiscoNet+ViT-B method and 24% and 40.3% by the original PDiscoNet. Additionally, similar to our previous experiments, we observe that the classification accuracy for the PDiscoNet method (also with the self-supervised ViT backbone) reduces with an increase in the number of parts to be discovered (K). Specifically, from Tab. D, our model's classification accuracy increases from 87.7% to 88.7%, while it reduces from 88.4% to 87.7% for PDiscoNet+ViT-B and from 85% to 84.3% for the original PDiscoNet. These results further indicate the generalization capability of our model for datasets containing objects with diverse part shapes.

Qualitative Results. Qualitative results for the PartImageNet Seg dataset with $K = 25$ parts are shown in Fig. E, featuring three images each for the super-classes *Car*, *Quadruped*, and *Snake*. From Fig. E, it is evident that our model is capable of consistently identifying parts with diverse shapes, such as the vehicle bumper, the dog's snout and ears, and the snake's body. The parts discovered by our model are also, on average, easier to interpret semantically. In contrast, the compared methods struggle to localize parts with irregular shapes, such as the snake body, and are generally less consistent, making them harder to interpret.

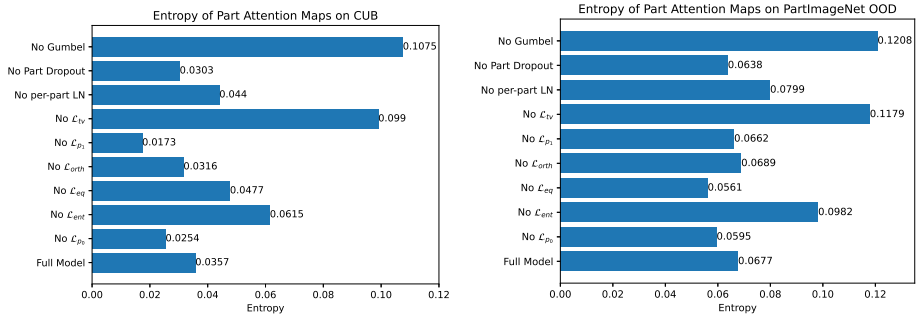

Fig. D: Entropy of part attention maps of the different model ablations on the CUB ($K = 16$) and PartImageNet OOD ($K = 25$) test sets.

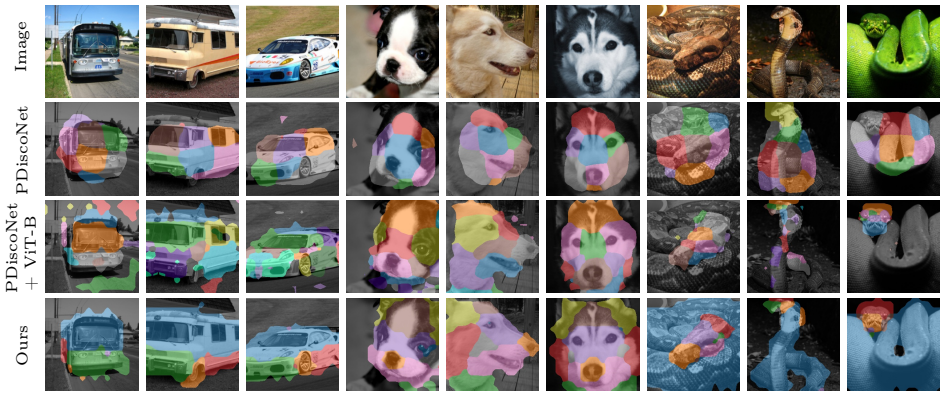

Fig. E: Qualitative results on PartImageNet Seg for $K = 25$ parts. The first 3 images belong to the super-class *Car*, the next 3 to the super-class *Quadruped* and the final 3 images belong to the super-class *Snake*.

References

1. Aniraj, A., Dantas, C.F., Ienco, D., Marcos, D.: Masking strategies for background bias removal in computer vision models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 4397–4405 (October 2023)
2. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018)
3. He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.N., Liu, S., Yang, C., Yu, Q., Yuille, A.: Partimagenet: A large, high-quality dataset of parts. In: European Conference on Computer Vision. pp. 128–145. Springer (2022)
4. Huang, Z., Li, Y.: Interpretable and accurate fine-grained recognition via region grouping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8662–8672 (2020)

Table D: Quantitative results on the PartImageNet Seg dataset

Method	PartImageNet Seg (%)			
	K	NMI↑	ARI↑	Top-1 Acc.↑
PDiscoNet [8]	8	10.69	32.43	85.03
	16	16.91	37.41	84.49
	25	18.44	40.70	84.12
	41	23.99	40.27	84.28
	50	21.98	39.40	84.32
PDiscoNet + ViT-B	8	16.68	32.01	88.36
	16	26.24	48.18	88.81
	25	25.93	46.81	**88.90**
	41	27.93	39.96	87.86
	50	28.68	34.13	87.73
PDiscoFormer (Ours)	8	20.29	38.90	87.65
	16	39.07	56.95	87.98
	25	43.18	**64.61**	88.15
	41	**44.85**	59.95	88.57
	50	44.06	60.10	88.65

5. Hung, W.C., Jampani, V., Liu, S., Molchanov, P., Yang, M.H., Kautz, J.: SCOPS: Self-Supervised Co-Part Segmentation. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
6. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: International Conference on Learning Representations (2017)
7. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2015), https://arxiv.org/abs/1412.6980
8. van der Klis, R., Alaniz, S., Mancini, M., Dantas, C.F., Ienco, D., Akata, Z., Marcos, D.: PDiscoNet: Semantically consistent part discovery for fine-grained recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1866–1876 (2023)
9. Krizhevsky, A.: One weird trick for parallelizing convolutional neural networks. arXiv preprint arXiv:1404.5997 (2014), http://arxiv.org/abs/1404.5997
10. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: International conference on machine learning. pp. 1310–1318. Pmlr (2013)
11. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology (2010)
12. Xiao, K.Y., Engstrom, L., Ilyas, A., Madry, A.: Noise or signal: The role of image backgrounds in object recognition. In: International Conference on Learning Representations (2020)