PDiscoFormer

Relaxing Part Discovery Constraints with Vision Transformers

Ananthu Aniraj, Cassio F. Dantas, Dino Ienco, Diego Marcos

Code

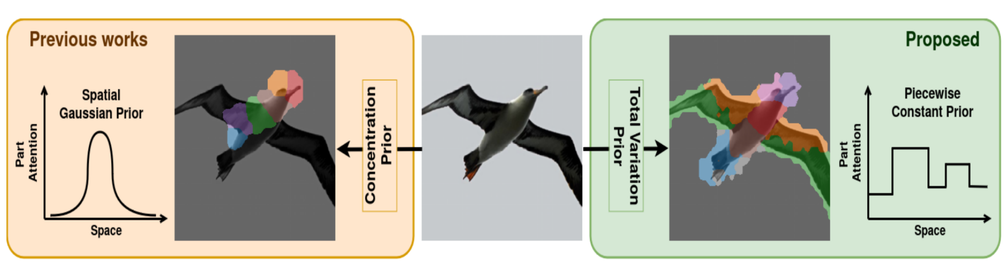
Motivation

Computer vision methods that explicitly detect object parts and reason on them are a step towards **inherently interpretable models**.

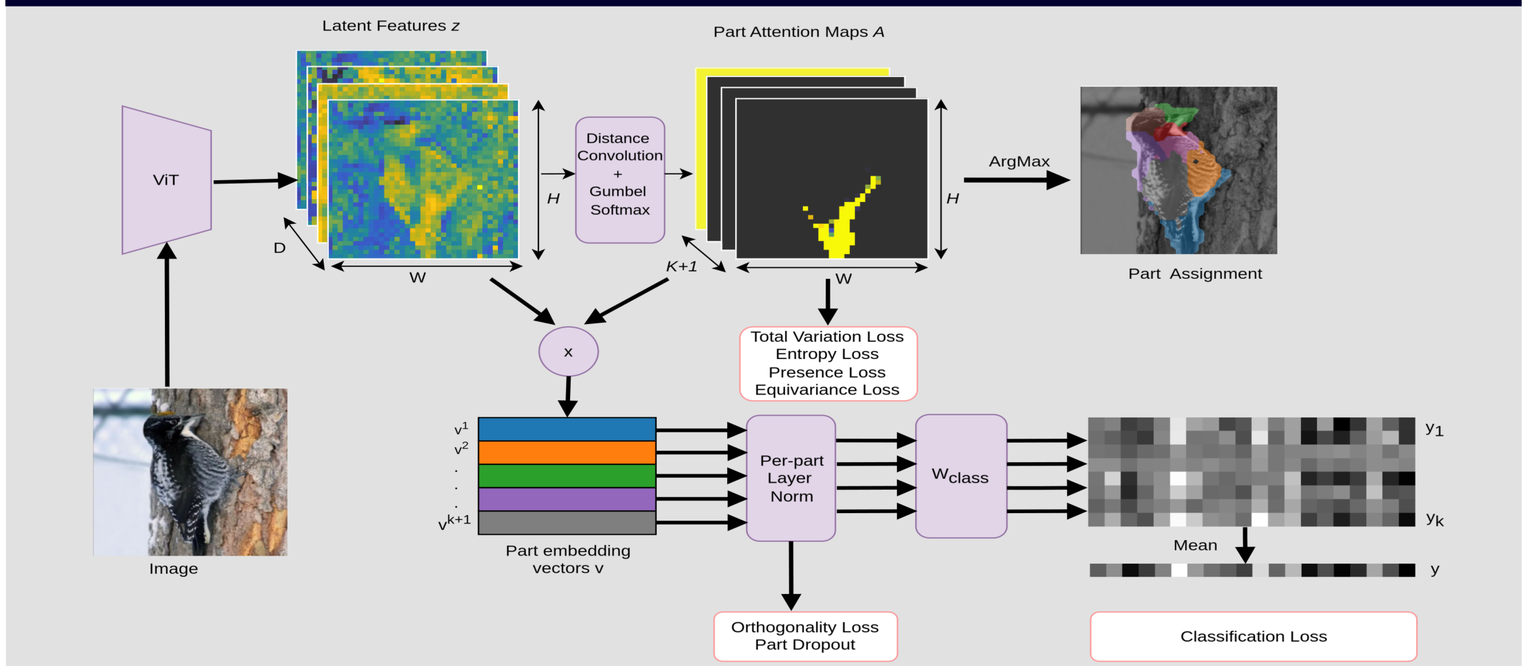
Existing approaches that perform part discovery make **restrictive assumptions** about the geometric properties of the discovered parts.

We find that **such restrictive priors are not required** to detect consistent parts using a pre-trained ViT such as DinoV2.

We use the **total variation (TV)** prior which enforces that the parts form spatially connected components.



Method



Part losses

We encourage attention maps to behave like parts: they should form connected components, be present in at least some images, distinct, equivariant to rigid transforms and should have low entropy.

Total Variation loss: parts form spatially connected components

$$\mathcal{L}_{\text{tv}} = \frac{1}{HW}\sum_{k=1}^{K+1}\sum_{ij}\left|\nabla a_{ij}^{k}\right|$$

Presence loss: stimulate part presence per batch

$$\mathcal{L}_{\text{p}_1} = 1 - \frac{1}{K}\sum_{k}\max_{b,i,j}\bar{a}_{ij}^{k}(\mathbf{x}_b)$$

$$\mathcal{L}_{\text{p}_0} = -\frac{1}{B}\sum_{b}\log\left(\max_{i,j}\; m_{ij}\bar{a}_{ij}^{K+1}(\mathbf{x}_b)\right)$$

Orthogonality loss: stimulate part distinctness

$$\mathcal{L}_{\perp} = \sum_{k=1}^{K+1}\sum_{l\neq k}\frac{\mathbf{v}_m^k \cdot \mathbf{v}_m^l}{\|\mathbf{v}_m^k\| \cdot \|\mathbf{v}_m^l\|}$$

Equivariance loss: stimulate equivariance to rigid transforms T

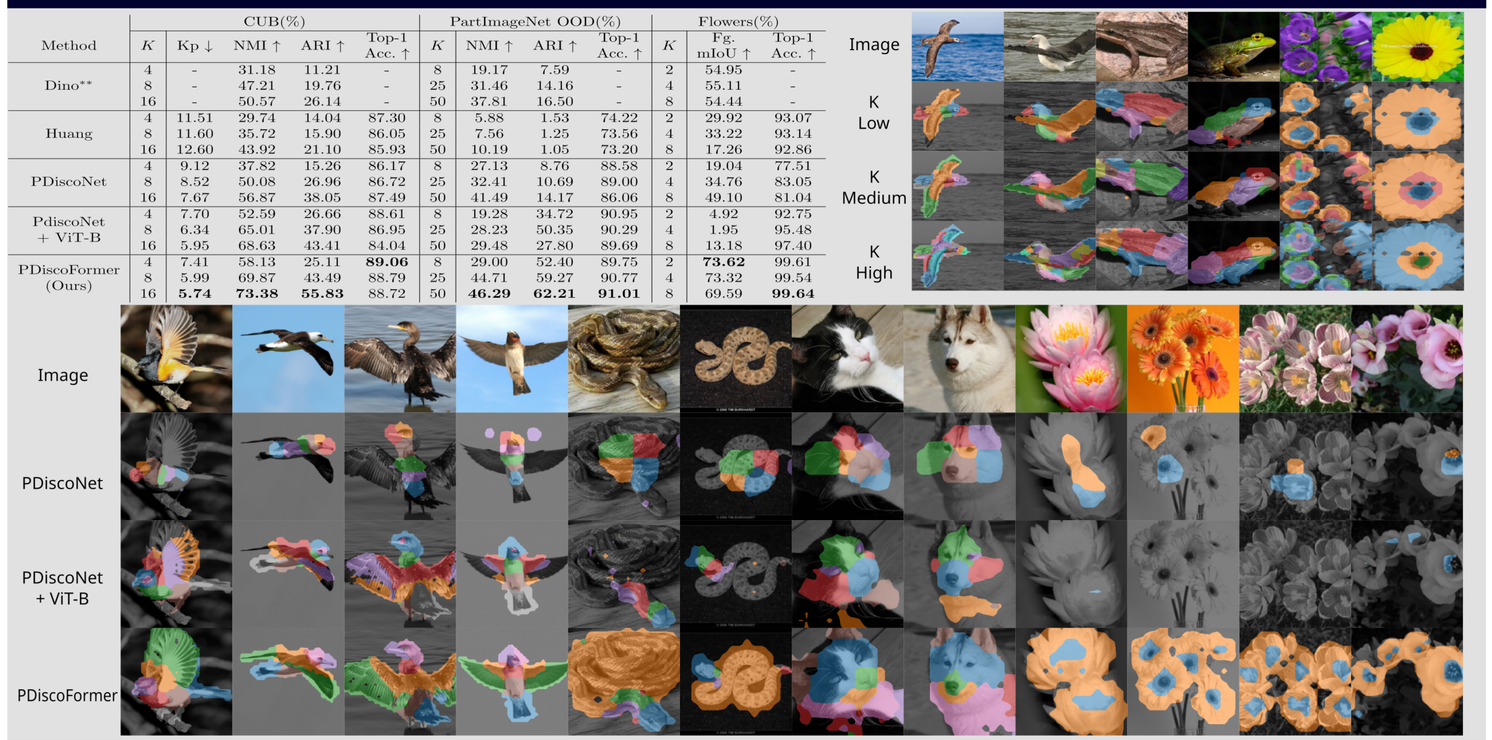
$$\mathcal{L}_{\text{equiv}} = 1 - \frac{1}{K}\sum_{k}\frac{||A^k(\mathbf{X}) \odot T^{-1}(A^k(T(\mathbf{X})))||}{||A^k(\mathbf{X})|| \cdot ||A^k(T(\mathbf{X}))||}$$

Entropy loss: each patch is uniquely assigned to a part

$$\mathcal{L}_{\text{ent}} = \frac{-1}{K+1}\sum_{k=1}^{K+1}\sum_{ij}a_{ij}^{k}\log\left(a_{ij}^{k}\right)$$

Results

Method	K	CUB(%) Kp ↓	CUB(%) NMI ↑	CUB(%) ARI ↑	CUB(%) Top-1 Acc. ↑	K	PartImageNet OOD(%) NMI ↑	PartImageNet OOD(%) ARI ↑	PartImageNet OOD(%) Top-1 Acc. ↑	K	Flowers(%) Fg. mIoU ↑	Flowers(%) Top-1 Acc. ↑
Dino**	4	-	31.18	11.21	-	8	19.17	7.59	-	2	54.95	-
	8	-	47.21	19.76	-	25	31.46	14.16	-	4	55.11	-
	16	-	50.57	26.14	-	50	37.81	16.50	-	8	54.44	-
Huang	4	11.51	29.74	14.04	87.30	8	5.88	1.53	74.22	2	29.92	93.07
	8	11.60	35.72	15.90	86.05	25	7.56	1.25	73.56	4	33.22	93.14
	16	12.60	43.92	21.10	85.93	50	10.19	1.05	73.20	8	17.26	92.86
PDiscoNet	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
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 (Ours)	4	7.41	58.13	25.11	**89.06**	8	29.00	52.40	89.75	2	**73.62**	99.61
	8	5.99	69.87	43.49	88.79	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**



Conclusion

We propose a training objective that enables consistent discovery of parts using a pre-trained self-supervised Vision Transformer.

We do not impose any constraints on part shape and only check if they form spatially connected components.

PDiscoFormer discovers semantically meaningful parts with only image class labels as supervision.

References

[1] Zixuan Huang and Yin Li, Interpretable and Accurate Fine-grained Recognition via Region Grouping. CVPR, 2020.
[2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT Features as Dense Visual Descriptors. arXiv, 2022.
[3] van der Klis, R., Alaniz, S., Mancini, M., Dantas, C. F., Ienco, D., Akata, Z., & Marcos, D. (2023). PDiscoNet: Semantically consistent part discovery for fine-grained recognition. ICCV, 2023.