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

Abstract. 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 driven by a fine-grained classification task make very restrictive assumptions on the geometric properties of the discovered parts; they should be small and compact. Although this prior is useful in some cases, in this paper we show that pre-trained transformer-based vision models, such as self-supervised DINOv2 ViT, enable the relaxation of these constraints. In particular, we find that a total variation (TV) prior, which allows for multiple connected components of any size, substantially outperforms previous work. We test our approach on three fine-grained classification benchmarks: CUB, PartImageNet and Oxford Flowers, and compare our results to previously published methods as well as a re-implementation of the state-of-the-art method PDiscoNet with a transformer-based backbone. We consistently obtain substantial improvements across the board, both on part discovery metrics and the downstream classification task, showing that the strong inductive biases in self-supervised ViT models require to rethink the geometric priors that can be used for unsupervised part discovery. Training code and pre-trained models are available at https://github.com/ananthu-aniraj/pdiscoformer.

1 Introduction

Deep neural networks, first in the form of convolutional neural networks (CNN) and, more recently, vision transformers (ViT), have revolutionized the field of
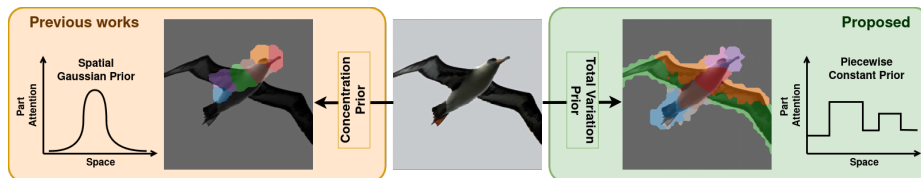

Fig. 1: Concentration *vs* total variation as part priors.

computer vision. However, their black-box nature often limits their application in domains where interpretability is crucial. To ameliorate the understanding of the internal behaviour of deep learning models trained for computer vision tasks, post-hoc interpretability methods [25,27,31] are commonly used. These methods generate saliency maps indicating regions in the image that contributed most to a classification decision. However, these are only approximations to the actual internal process of the model and may be misleading [28].

Recent efforts have been directed towards developing interpretable-by-design deep neural networks [4,13,14,17,21,22]. These approaches aim to identify semantically interpretable features in an image and utilize this information for solving a downstream task. Primarily tailored for fine-grained image classification, these methods seek to emulate human reasoning by subdividing the final task, such as classifying flower species or dog breeds, into multiple explicit intermediate stages, such as identifying lower-level semantic features like body parts, which in turn are used to provide a classification score.

However, improving interpretability while maintaining performance remains a challenge. Some models, such as those proposed in [13,14,17,18], rely on objective functions that impose priors on the shape and area of object parts such as compactness-inducing part shaping losses that, in practice, result in spatial Gaussian priors of fixed variance on the part maps. While effective for specific datasets, like the commonly used CUB [38] benchmark for bird species classification, where assumptions like parts occurring only once in an image and being compact are beneficial, these assumptions may not generalize well to other use cases. For instance, in plant species classification, the presence of multiple instances of object parts such as leaves or flowers, as well as irregularly-shaped parts like branches or stems, renders these assumptions inadequate.

As we show in this work, such strong part shape priors are required when using a CNN-based backbone pre-trained on ImageNet; in order to reach acceptable classification results, it is imperative to fine-tune the whole backbone. This means that the only inductive bias that is kept throughout the part discovery process is the one brought in by the convolutional nature of the CNN architecture. A strong part shape prior is required in order to obtain consistent part maps in this setting, leading to all recent methods for part discovery in the context of fine-grained classification to use a concentration loss that makes part maps small and compact [13,14,17]. Here, we propose a method that addresses this issue by refraining from assuming any priors on part shape or area. The only prior we assume is that the part maps should tend to be piece-wise constant and thus be formed of one or more spatially connected components. Our model is capable of localizing up to a fixed number of semantically consistent parts of an object without relying on part annotations, using only the classification labels as supervision signals. The lack of assumptions about part shape and size makes our approach suitable to various fine-grained image classification tasks, as demonstrated by our empirical results, generalizing beyond cases where compactness is a good part prior.

The **main contribution** of this paper is to build upon [17] and relax its constraining concentration-based part prior by a more flexible total variation prior. We investigate which additional elements are required to compensate for the flexibility of the total variation loss and find that a combination of an entropy loss and Gumbel-Softmax on the part maps allows for consistent and high quality part maps. Our approach makes better use of the representation of self-supervised pre-trained ViT models and allows to apply the model to tasks with diverse types of parts.

2   Related Work

Unsupervised Part Discovery. Machine learning methods designed for unsupervised part discovery aim to identify a small number of semantically consistent parts, typically around $K = 8$, shared among classes in a dataset. These methods visualize all discovered parts in an image through per-part saliency maps, facilitating easier interpretation. However, as this task is unsupervised, it often relies on assumptions about the discovered parts, such as their distribution and shape. For instance, Huang *et al.* [13] assume that the presence of each part across a mini-batch follows a beta distribution, while [17] assumes that discovered parts are present at least once in a mini-batch. Additionally, assumptions are often made about the shape of the parts. Methods such as [14, 17, 18, 40] impose a spatial Gaussian prior with a fixed variance on the part maps, resulting in compact parts appearing once in the image. Such assumptions on part shape can be highly restrictive and may not generalize well across different scenarios. To address this limitation, our proposed method refrains from assuming such priors on part shape and only imposes a spatial prior that discovered parts should form connected components.

Part Prototype Networks. Chen *et al.* [4] introduced the Prototypical Part Network (ProtoPNet), an inherently interpretable method featuring a fixed number of learned prototypes per class. These prototypes represent visual patterns or features learned by the network during training. The model computes the similarity between each learned prototype and image patches and uses a linear combination of these similarity scores for classification. While explanations provided by ProtoPNet visualize all prototypes along with their weighted similarity scores, the sheer number of prototypes, often proportional to the number of classes, can lead to redundancy and interpretability challenges [10, 21]. Subsequent works in this line of research have sought to address these challenges. Some approaches focus on reducing the number of prototypes shared per class [30, 37], while others explore sharing prototypes between classes [20, 22]. Despite these efforts, recent works like [22] still utilize a large number ($K > 500$) of prototypes, which can pose significant challenges for end-user interpretation.

Regularization in Image Segmentation. Early works on image segmentation made direct use of low-level image features, chiefly color, along with geometric priors, in order to find coherent segments. The Mumford-Shah model [19], originally devised for piece-wise image reconstruction, was applied to find segments
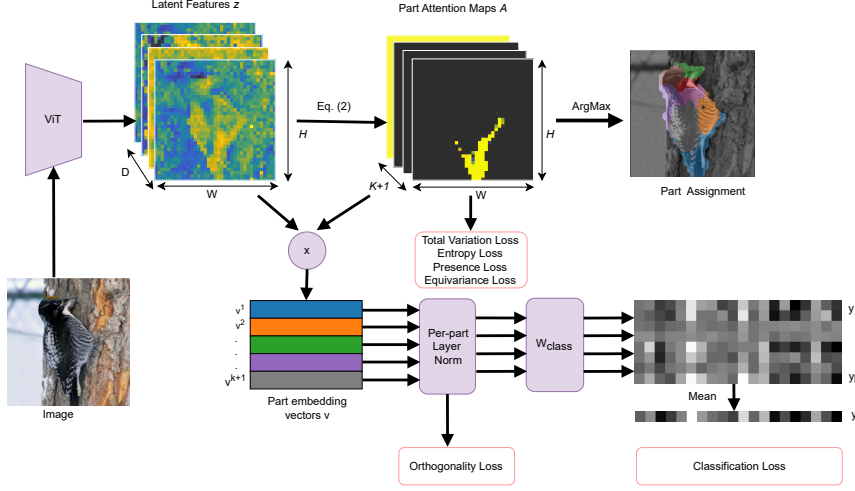
Fig. 2: The **PDiscoFormer** architecture for part discovery.

that are as homogeneous as possible while trying to minimize the length of their boundary as a regularization [3]. Total variation (TV) [29] has been shown to be a powerful and efficient regularization to approximate this boundary length prior, and has been extensively used for piece-wise signal and image reconstruction [7,35], as well as unsupervised image segmentation [8,36]. With the advent of deep learning, TV regularization has proven valuable for style transfer [16] and edge-preserving smoothing [41]. In this work we explore the use of TV regularization for weakly-supervised part discovery in order to alleviate the issues brought by previously used priors.

Emergent grouping of ViT representations. Some recent research [26,39] has studied the emergence of object segmentation by supervising with image-text pairs. These methods learn to group similar patches in a ViT and assign them to textual concepts. They aim to identify the pixels in an input image corresponding to a predefined text prompt, focusing on segmenting whole objects rather than decomposing them into parts, as our model does with part discovery, and are limited to concepts that are represented in the text annotations.

3   Methodology

Here we introduce an end-to-end trainable method (see Fig. 2) capable of automatically identifying K discriminative parts (semantically interpretable image regions) useful for solving a fine-grained image classification task. In contrast to earlier methods [13,17], our approach avoids any assumptions related to part shape, size or number of part instances. As in [13,17], our model is weakly supervised and relies solely on image-level class labels, thus uncovering discriminative parts without requiring any additional supervision. Our formulation is based

on [17], from which we take take most elements of the architecture, except for the Gumbel-Softmax and the layer normalization-based part modulation. The total variation, entropy and background losses have not been used in previous part discovery works and are shown to be important in our results.

3.1   Model architecture

Let $\mathbf{x} \in \mathbb{R}^{3 \times M \times N}$ be an image in the dataset, and let $y \in \{1, 2, ..., C\}$ represent the corresponding classification label. Given a backbone model $h_\theta(\cdot)$, we extract a feature tensor:

$$\mathbf{z} = h_\theta(\mathbf{x}) \in \mathbb{R}^{D \times H \times W} \tag{1}$$

We drop the class and register tokens of the ViT and reshape the patch tokens to obtain $\mathbf{z}$. Following prior research [13, 14, 17], we compute attention maps $\mathbf{A} \in [0, 1]^{(K+1) \times H \times W}$, where the last channel in the first dimension is used to represent the background region in the image. These attention maps are computed using the negative squared Euclidean distance function between patch token features $\mathbf{z}_{ij} \in \mathbb{R}^D$, $i \in \{1, ..., H\}$, $j \in \{1, ..., W\}$ and each of the learnable prototypes $\mathbf{p}^k \in \mathbb{R}^D$, with $k \in \{1, \ldots, K+1\}$, followed by a Gumbel-Softmax [15] across the $K + 1$ channels:

$$a_{ij}^k = \frac{\exp\left(-\|\mathbf{z}_{ij} - \mathbf{p}^k\|^2 + \gamma_k\right)}{\sum_{l=1}^{K+1} \exp\left(-\|\mathbf{z}_{ij} - \mathbf{p}^l\|^2 + \gamma_l\right)} \tag{2}$$

where γ_k and γ_l are independently drawn samples from the `Gumbel(0,1)` distribution, and a_{ij}^k represents an element of $\mathbf{A}$.

The attention maps are used to compute part embedding vectors $\mathbf{v}^k \in \mathbb{R}^D$ by using the attention values to calculate a weighted average over the feature tensor $\mathbf{z}$:

$$\mathbf{v}^k = \frac{\sum_i \sum_j a_{ij}^k \mathbf{z}_{ij}}{HW} \tag{3}$$

These embedding vectors are then modulated using a layer normalization [2] operation, with normalization applied over the feature and prototype dimensions. This operation helps to make each embedding vector distinct for the classifier:

$$\mathbf{v}_m^k = \frac{(\mathbf{v}^k - \mathrm{E}[\mathbf{v}^k])}{\sqrt{\mathrm{Var}[\mathbf{v}^k] + \epsilon}} \odot \mathbf{w}_m^k + \mathbf{b}_m^k \tag{4}$$

with $\mathbf{w}_m^k, \mathbf{b}_m^k \in \mathbb{R}^D$ being the modulation weights and biases per part. These modulated embedding vectors $\mathbf{v}_m^k$ for $k \in \{1, \ldots, K\}$ (we omit the background embedding) are then linearly projected to obtain a vector of class scores conditioned on the part embedding:

$$\mathbf{y}^k = W_c \cdot \mathbf{v}_m^k \tag{5}$$

The same linear classifier $W_c \in \mathbb{R}^{C \times D}$ is applied to each of the modulated embedding vectors, i.e., for each of the K discovered foreground parts. We then take the mean of the per-part class scores to obtain the classification scores:

$$\mathbf{y} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{y}^k \qquad (6)$$

Finally, a softmax function is employed to obtain the classification probabilities.
Part dropout. We adopt dropout [34] on the modulated embedding vectors $\mathbf{v}_m^k$, as proposed in [17], during model training. This drops entire part embedding vectors at a time and ensures that all discovered parts are discriminative, since discovering a single discriminative part would not be enough to solve the downstream classification task.

3.2 Loss Functions

Our main contribution consists in proposing a set of part-shaping loss functions that are able to make the most out of the inductive biases learned by self-supervised ViT models.
Classification Loss ($\mathcal{L}_{\mathbf{c}}$). Our model is mainly driven by the fine-grained classification objective during training. We use the standard negative log likelihood or multi-class cross-entropy for this purpose. While the classification loss helps ensure the discriminative nature of the discovered parts, it may not directly enforce the learned attention maps to represent salient parts of the object in the image. To address this, we incorporate additional objective functions to enforce priors on the discovered parts. These objectives effectively guide the learning process to focus on relevant regions of the image.
Orthogonality Loss ($\mathcal{L}_{\perp}$). We aim to discover parts that are semantically distinct. To achieve this, we apply the orthogonality loss, which encourages the learned part embedding vectors to be decorrelated from one another. This is accomplished by minimizing the cosine similarity between the embedding vectors. We apply this loss on the modulated embedding vectors $\mathbf{v}_m^k$.

$$\mathcal{L}_{\perp} = \sum_{k=1}^{K+1} \sum_{l \neq k} \frac{\mathbf{v}_m^k \cdot \mathbf{v}_m^l}{\|\mathbf{v}_m^k\| \cdot \|\mathbf{v}_m^l\|} \qquad (7)$$

Equivariance Loss ($\mathcal{L}_{\mathbf{eq}}$). We aim to detect the same parts even if the image is translated, rotated, or scaled. To achieve this, we formulate the equivariance loss, which encourages the learned part attention maps to be equivariant to rigid transformations. Specifically, given an input image, we use an affine transform T with parameters randomly sampled from a pre-defined range and pass both the original image and the transformed image through our model. Then, we invert the transformation on the part attention maps of the transformed image. Finally, the equivariance loss is formulated as the cosine distance between the part attention maps of the original image and those of the corresponding transformed image.

Let $A^k(\mathbf{x})$ be a function that returns the kth attention map given an input image $\mathbf{x}$. Then:

$$\mathcal{L}_{\text{eq}} = 1 - \frac{1}{K} \sum_k \frac{\left\| A^k(\mathbf{x}) \cdot T^{-1}\left(A^k\left(T(\mathbf{x})\right)\right)\right\|}{\|A^k(\mathbf{x})\| \cdot \|A^k\left(T(\mathbf{x})\right)\|} \tag{8}$$

Presence Loss ($\mathcal{L}_{\mathbf{p}}$). We assume that all discovered foreground parts are present in at least some images of the training dataset, while the background, the $(K+1)^{th}$ part, is expected to be present in all images in the dataset. To enforce this, we formulate the presence loss, which consists of two parts. Firstly, we define a loss applied only to the discovered foreground parts, ensuring that each part is present at least once in a mini-batch. Given a batch $\{\mathbf{x}_1, ..., \mathbf{x}_B\}$, the presence loss is calculated as:

$$\mathcal{L}_{\text{P}_1} = 1 - \frac{1}{K} \sum_k \max_{b,i,j} \bar{a}_{ij}^k(\mathbf{x}_b) \tag{9}$$

where $\bar{A}^k(\mathbf{x}_b) = \text{avgpool}(A^k(\mathbf{x}_b))$ is the output of a 2D average pooling operator with a small kernel size employed to prevent single-pixel attention maps.

Secondly, to ensure the presence of the background, we introduce a stricter presence loss, which is satisfied if the background is present in all images of the mini-batch. This loss is formulated as follows:

$$\mathcal{L}_{\text{P}_0} = -\frac{1}{B} \sum_b \log \left(\max_{i,j} \; m_{ij} \bar{a}_{ij}^{K+1}(\mathbf{x}_b) \right) \tag{10}$$

where $M = [m_{ij}]^{H \times W}$ is a soft mask with $m_{ij} \in [0,1]$ that privileges entries placed farther from the image center (towards the borders):

$$m_{ij} = 2 \left(\frac{i-1}{H-1} - \frac{1}{2} \right)^2 + 2 \left(\frac{j-1}{W-1} - \frac{1}{2} \right)^2 \tag{11}$$

This loss captures the notion that the background is at the same time expected in every image and is more likely to appear near the boundaries of the image.

Entropy Loss ($\mathcal{L}_{\mathbf{ent}}$). We aim to ensure that each patch token is assigned to a unique part. To achieve this, we formulate an entropy [32] loss applied to the part attention maps:

$$\mathcal{L}_{\text{ent}} = \frac{-1}{K+1} \sum_{k=1}^{K+1} \sum_{ij} a_{ij}^k \log \left(a_{ij}^k \right) \tag{12}$$

Total Variation Loss ($\mathcal{L}_{\mathbf{tv}}$). We do not want to make any assumptions about the part shape. However, we would still like the discovered parts to be composed of one or, at most, a few connected components. To achieve this, we make use of a total variation loss [29] applied to the part attention maps:

$$\mathcal{L}_{\text{tv}} = \frac{1}{HW} \sum_{k=1}^{K+1} \sum_{ij} \left| \nabla a_{ij}^k \right| \tag{13}$$

where ∇a_{ij}^k is the spatial image gradient of part map A^k in location ij.

4   Experimental Setup

We aim to perform part discovery using only image-level class labels as supervision signals, assuming that the discovered parts are shared among classes, which is common in fine-grained image classification settings. Moreover, we aim to compare our model with similar methods in the literature in scenarios where they have been empirically shown to work well, as well as in new and more challenging scenarios where underlying data characteristics may require more flexible geometric modeling.

All models were implemented in PyTorch. We used the DinoV2 [24] ViT [9] model with register tokens [6] as the starting weights for all experiments, utilizing the Base (ViT-B) variant. The training settings and compute requirements of our model are detailed in the supplementary material.

4.1   Datasets

CUB Dataset. The CUB dataset [38] consists of images depicting 200 bird species, with 5,994 images in the training set and 5,794 images in the testing set. Keypoint annottions for 15 different bird body parts are available for each image. Bird parts are typically compact, although their presence may vary in images due to changes in pose and occlusion. Most images contain only a single instance of a bird. For this dataset, we evaluate the results using keypoint regression (Kp) using ground truth part keypoint annotations to compare them with centroids discovered by the model. Furthermore, we compute the Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) to quantify the clustering quality of the discovered parts, as in related literature [5].

PartImageNet Dataset. The PartImageNet dataset [11] comprises 158 classes distributed among 11 super-classes, such as airplanes, birds, and snakes. Each super-class is associated to between two and five possible parts, for a total of 41 part classes. Additionally, test images may contain multiple instances of objects, sometimes from different super-classes, while the classification label provides information about a single object. Notably, several super-classes, including snakes, airplanes, and reptiles, often exhibit irregular and non-compact part shapes. For our experiments, we use the `PartImageNet OOD` variant. Although originally intended for out-of-distribution detection, this variant is commonly used in related literature. We train our models on 14,865 training images and test on 1,658 images, respectively. For evaluation, we derive part centroids from ground truth part masks and calculate NMI and ARI scores to assess clustering quality. Additionally, we include an experiment on the newly released `PartImageNet Seg` variant in the supplementary material.

Table 1: Quantitative comparison of the part discovery and classification capabilities of our model with related approaches from the literature. In this table, K refers to the number of parts. The method with the label ** does not use any class supervision.

Method	K	CUB(%) Kp ↓	NMI ↑	ARI ↑	Top-1 Acc. ↑	K	PartImageNet OOD(%) NMI ↑	ARI ↑	Top-1 Acc. ↑	K	Flowers(%) Fg. mIoU ↑	Top-1 Acc. ↑
Dino [1]**	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 [13]	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 [17]	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 [6]	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**

Oxford 102 Flower Dataset. The Oxford 102 Flower Dataset [23] comprises 8,189 images divided into 102 flower classes. The training and validation sets contain 10 images per class, totaling 1,020 images each, while the test set consists of 6,149 images, with a minimum of 20 images per class. Since no part annotations are provided for this dataset, we evaluate the quality of the discovered parts by combining them and comparing the result to the provided foreground/background masks. We evaluate model performance by calculating the mean Intersection over Union (mIoU) for the foreground mask. This dataset often includes multiple instances of parts in a single image, such as flowers, as well as parts with irregular shapes, such as flower petals.

5 Experiments

5.1 Comparison with State-of-the-Art

We conduct a quantitative analysis of our model's part discovery and classification capabilities and compare it with those of related methods from the literature. Additionally, we compare our model against a re-implementation of the state-of-the-art PDiscoNet [17], where we replaced the backbone with the variant used in this paper. The results of this comparison are summarized in Tab. 1. From the results, our method consistently outperforms existing approaches for both part discovery and image classification across all datasets. We attribute this to the synergistic effects of the strong inductive biases ingrained in the self-supervised ViT backbone and the enhanced flexibility of the proposed combina-
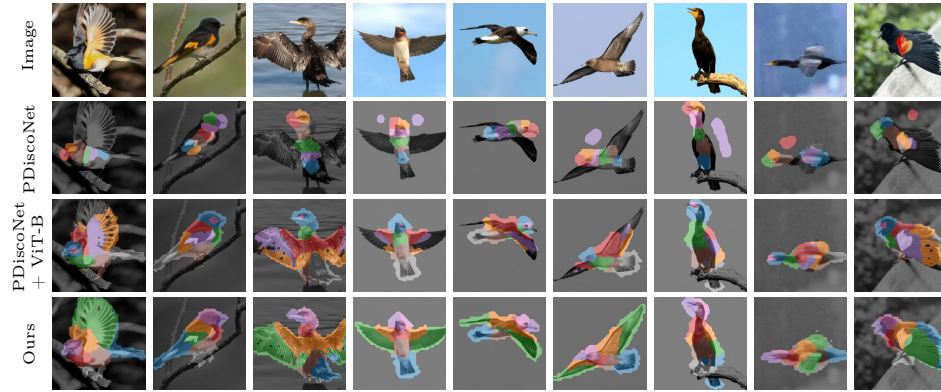
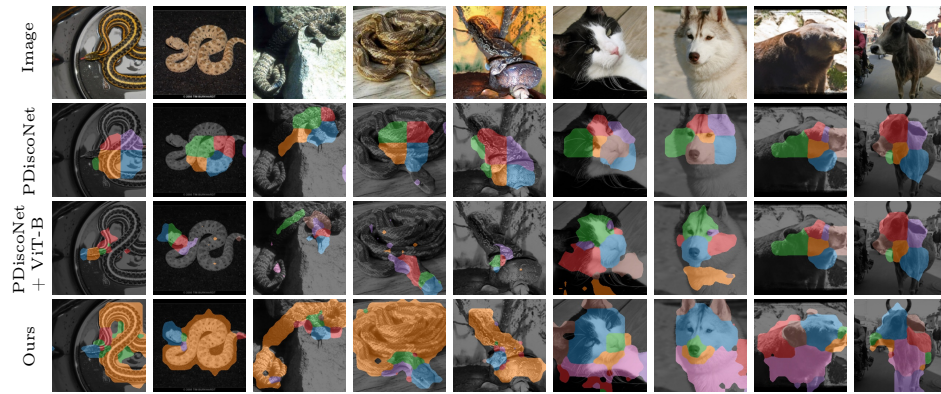**Fig. 3:** Qualitative results on CUB for $K = 8$.

Fig. 4: Qualitative results on PartImageNet for $K = 8$.

tion of geometric priors. Particularly noteworthy are the significant performance gains observed on the Flowers and PartImageNet datasets, which contain objects with irregular shapes occurring in multiple instances within the same image, underscoring the importance of our geometric part prior that favors piece-wise constant and spatially connected components. Even on CUB, where most parts fit well into the compactness prior, we see a consistent improvements of results with respect to PDiscoNet, even when using the same ViT backbone. With 16 parts, our model achieves 55.8% ARI and 73.4% NMI, compared to 43.4% ARI and 68.6% NMI of the best competing method, PDiscoNet with a ViT backbone. Despite competitive performance in classification tasks on CUB and PartImageNet, the PdiscoNet+ViT-B model tends to degrade in accuracy as part complexity increases, particularly in CUB, unlike our model, which maintains robust accuracy for any number of parts. In the Flowers dataset, all competing methods result in subpar performance, while our approach excels at both detecting the foreground and classification no matter the number of parts. In this
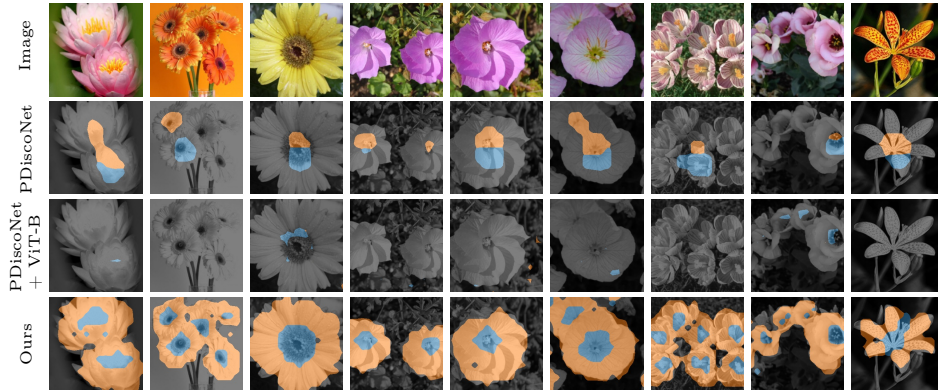
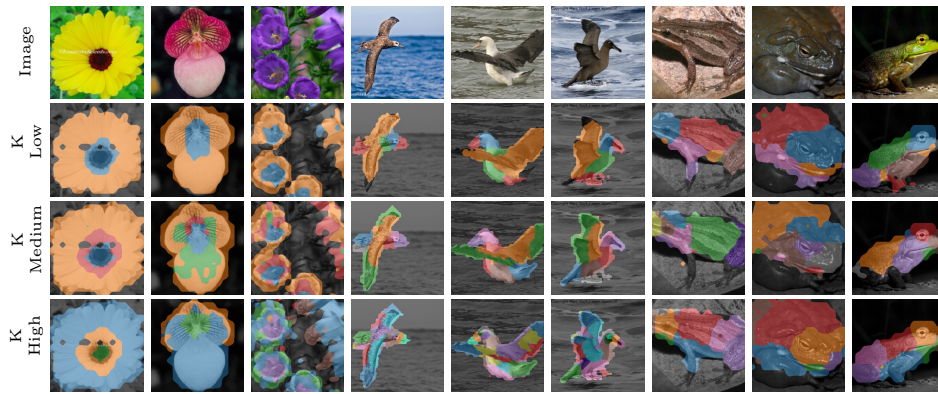**Fig. 5:** Qualitative results on Flowers for $K = 2$.

Fig. 6: Effect of K on PDiscoFormer. The left three images correspond to Flowers (from top to bottom, $K = 2, 4, 8$). The three middle ones are CUB, with K being 4, 8 and 16. The three on the right belong to PartImageNet, with $K = 8, 25, 50$.

dataset, PDiscoNet+ViT-B results in the second best classification accuracy, 97.4%, after the 99.6% obtained by our method, but completely fails to locate the foreground, with an mIoU of 13.2%, against 73.6% for PDiscoFormer. Additionally, we extend our comparisons by applying our proposed losses to a ResNet backbone and a fully frozen ViT architecture, showing that a pre-trained ViT is required to fully profit from the proposed set of losses and that fine-tuning the position, class and register tokens, while keeping the rest of the ViT frozen, results in better performances than leaving the whole ViT frozen. The results of these additional comparisons are presented in the supplementary material.

Qualitative Analysis. A qualitative analysis, as illustrated in Fig. 3, Fig. 4, Fig. 5 and Fig. 6 corroborates our quantitative findings, showcasing our model's ability to accurately discover irregularly shaped body parts such as bird wings (CUB), flower petals (Flowers) and the body of a snake (PartImageNet). At the

Table 2: Results of the ablation study in PartImageNet and CUB.

	PartImageNet OOD (K=25)			CUB (K=16)			
	NMI ↑	ARI ↑	Top-1 Acc. ↑	Kp ↓	NMI ↑	ARI ↑	Top-1 Acc. ↑
Full Model	**44.71**	**59.27**	90.77	5.74	**73.38**	**55.83**	88.72
No $\mathcal{L}_{\mathrm{p_0}}$	41.05	49.08	89.81	5.85	68.66	43.09	83.64
No $\mathcal{L}_{ent}$	39.16	54.46	**91.19**	**5.57**	66.88	45.68	88.21
No $\mathcal{L}_{eq}$	43.38	54.19	90.47	9.05	56.85	30.63	83.71
No $\mathcal{L}_{\perp}$	42.44	58.86	91.01	5.79	70.55	50.90	88.63
No $\mathcal{L}_{\mathrm{p_1}}$	33.98	58.31	91.07	7.14	61.04	34.28	**89.04**
No $\mathcal{L}_{tv}$	33.73	25.35	90.47	6.01	70.78	51.83	80.22
No Per-Part LayerNorm	31.14	40.73	89.69	5.74	70.70	50.62	69.73
No Part Dropout	33.08	43.98	89.99	6.08	70.28	41.69	87.14
No Gumbel	32.98	54.68	90.71	6.44	68.92	43.61	86.43

same time, PDiscoNet [17] fails to discover these parts altogether, even when retrained on a ViT-B backbone. This indicates the usefulness of the flexible geometric priors enabled by our total variation loss. Compared to PDiscoNet [17], not only do we achieve a much tighter coverage of the object of interest, with our PDiscoFormer more closely following the object boundaries. This is particularly the case in parts that are relatively large, such as open bird wings and flower corollas. In addition to that, we can observe a substantial improvement in the consistency and semantic interpretability of parts. For instance, the parts discovered by our model for the four quadruped images displayed in Fig. 4 can immediately be interpreted (the green part corresponds to the snout while the brown one to the ears), while those from PDiscoNet are not clear. In Fig. 6 we vary the number of parts, K. On Flowers, using two parts leads to a clear assignment of one part to the flower calyx and one to the corolla, while more parts tends to exhibit oversegmentation. CUB, on the other hand, is able to absorb a larger number of parts. With $K = 4$, beak and legs are covered by the same part, while a fine-grained and consistent distinction of the different wing parts is achieved with $K = 16$. Interestingly, in the case of PartImageNet, the effective number of parts being used by a specific super-class remains almost constant when increasing K.

5.2   Ablation Study

We conducted an ablation study by systematically removing one loss function or architectural element at a time and by replacing the Gumbel-Softmax with a regular Softmax. This was performed on the PartImageNet OOD dataset with $K = 25$ parts and the CUB dataset with $K = 16$ parts. The quantitative results of the ablation study are shown in Tab. 2 and some qualitative results can be seen in Fig. 7. From the results, we can observe the following: **Total Variation** $(\mathcal{L}_{tv})$ prevents spurious part activations in the background, possibly reducing
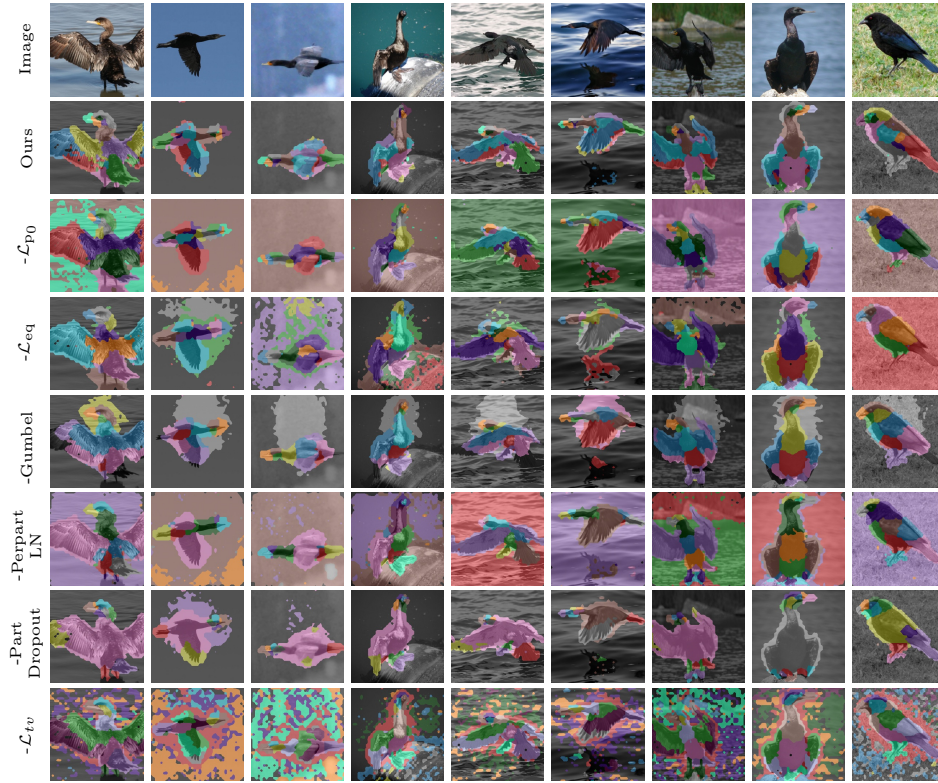
Fig. 7: Qualitative results from the ablation study.

information leakage between foreground parts and the background. It has the largest impact on PartImageNet in terms of ARI. The **absence of the Background Loss ($\mathcal{L}_{\mathbf{P}_0}$)** results in the background part not being activated, causing the model to discover the background as a foreground part. **Presence Loss ($\mathcal{L}_{\mathbf{p}}$) and Part Dropout** promote the utilization of all parts, preventing domination by a few parts which can lead to multiple unused parts (Fig. 7). **Removing the Per-Part Layer Norm** leads to inconsistent part discovery (Fig. 7) indicating that it contributes to the creation of discriminative part-specific features driven by the classification loss. **Equivariance Loss ($\mathcal{L}_{eq}$)** enhances robustness to affine geometric transformations, supporting stable part discovery across diverse inputs (Fig. 7). **Gumbel-Softmax and Entropy Loss ($\mathcal{L}_{\mathbf{ent}}$)** ensure unique part assignments for each patch token, further reducing information leakage, as shown in our supplementary material. As shown in Tab. 2, the complete model, incorporating all loss functions and architectural choices, generally performs best for both part discovery and classification, demonstrating the efficacy of our approach.

6   Limitations and Future Work

Our study focused on datasets with part annotations or FG-BG masks available for the test set. This allowed for quantitative comparison with state-of-the-art methods. However, exploring the training of our model on larger and more diverse datasets could provide valuable insights and further validate its performance in real-world scenarios. For example, larger datasets may introduce more variability in object appearances and poses, challenging the model to generalize better. Methods to enable the evaluation of part discovery in such settings would allow to take steps in this direction. The main hyper-parameter K, the number of parts to discover, has a large impact on the granularity of the discovered parts (Fig. 6). Exploring ways of automatically estimating this value while preserving interpretability in the discovered semantic parts would be a valuable contribution. This could enhance the adaptability of the model to various datasets and streamline the training process by removing the need for manual selection of K. Additionally, while our approach can detect parts of various shapes due to the total variation prior, it cannot ascertain if they belong to the same object. Integrating our method with unsupervised object discovery research [12,33] could enable part discovery within each detected object, improving capabilities in complex scenes. These methods aim to identify and localize individual objects without explicit supervision, complementing our approach for a more comprehensive scene understanding. These considerations suggest promising directions for extending and refining our method.

7   Conclusion

In this paper, we presented PDiscoFormer, a new approach for unsupervised part discovery. Unlike previous methods that imposed strict geometric priors on discovered object parts, our model, based on self-supervised vision transformers, demonstrated the capability to relax these constraints. By leveraging a total variation (TV) prior, our model achieved substantial improvements in both part discovery and downstream classification accuracy when compared to the state-of-the-art across multiple datasets, specifically CUB, PartImageNet, and Oxford Flowers. Our findings highlight the importance of rethinking geometric priors in unsupervised part discovery tasks, especially in scenarios where objects exhibit complex shapes or occur multiple times within the same image. By refraining from assuming specific shape or size priors, our model demonstrated enhanced generalization capabilities, making it applicable to a broader range of fine-grained image classification tasks. Overall, our study shows that the strong inductive biases learned by vision transformers during the self-supervised pre-training stage enable the use of flexible shape priors such as total variation, leading to robust performance on part discovery and the downstream classification task on all of the tested datasets.

References

1. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. arXiv preprint arXiv:2112.05814 **2**(3), 4 (2021)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Transactions on image processing **10**(2), 266–277 (2001)
4. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems **32** (2019)
5. Choudhury, S., Laina, I., Rupprecht, C., Vedaldi, A.: Unsupervised part discovery from contrastive reconstruction. Advances in Neural Information Processing Systems **34**, 28104–28118 (2021)
6. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2023)
7. Dobson, D.C., Santosa, F.: Recovery of blocky images from noisy and blurred data. SIAM Journal on Applied Mathematics **56**(4), 1181–1198 (1996)
8. Donoser, M., Urschler, M., Hirzer, M., Bischof, H.: Saliency driven total variation segmentation. In: ICCV. pp. 817–824. IEEE (2009)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
10. Gautam, S., Höhne, M.M.C., Hansen, S., Jenssen, R., Kampffmeyer, M.: This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. Pattern Recognition **136**, 109172 (2023)
11. 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)
12. Hénaff, O.J., Koppula, S., Shelhamer, E., Zoran, D., Jaegle, A., Zisserman, A., Carreira, J., Arandjelović, R.: Object discovery and representation networks. In: European Conference on Computer Vision. pp. 123–143. Springer (2022)
13. 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)
14. 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)
15. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: International Conference on Learning Representations (2017)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
17. 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)

18. Marcos, D., Kierdorf, J., Cheeseman, T., Tuia, D., Roscher, R.: A whale's tail-finding the right whale in an uncertain world. In: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. pp. 297–313. Springer (2020)
19. Mumford, D.B., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Communications on pure and applied mathematics (1989)
20. Nauta, M., van Bree, R., Seifert, C.: Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14933–14943 (6 2021)
21. Nauta, M., Jutte, A., Provoost, J., Seifert, C.: This looks like that, because... explaining prototypes for interpretable image recognition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 441–456. Springer (2021)
22. Nauta, M., Schlötterer, J., van Keulen, M., Seifert, C.: PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
23. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics and Image Processing (Dec 2008)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. arXiv preprint arXiv:2304.07193 (2023), `http://arxiv.org/abs/2304.07193`
25. Ramaswamy, H.G., et al.: Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In: proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 983–991 (2020)
26. Ranasinghe, K., McKinzie, B., Ravi, S., Yang, Y., Toshev, A., Shlens, J.: Perceptual grouping in contrastive vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5571–5584 (2023)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144 (2016)
28. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
29. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena **60**(1-4), 259–268 (1992)
30. Rymarczyk, D., Struski, Å., Tabor, J., Zieliński, B.: Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 1420–1430 (2021)
31. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

32. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal **27**(3), 379–423 (1948). `https://doi.org/10.1002/j.1538-7305.1948.tb01338.x`
33. Siméoni, O., Puy, G., Vo, H.V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J.: Localizing objects with self-supervised transformers and no labels. arXiv preprint arXiv:2109.14279 (2021)
34. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
35. Strong, D., Chan, T.: Edge-preserving and scale-dependent properties of total variation regularization. Inverse problems **19**(6), S165 (2003)
36. Unger, M., Pock, T., Trobin, W., Cremers, D., Bischof, H.: Tvseg-interactive total variation based image segmentation. In: BMVC. vol. 31, pp. 44–46. Citeseer (2008)
37. Wang, J., Liu, H., Wang, X., Jing, L.: Interpretable image recognition by constructing transparent embedding space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 895–904 (2021)
38. 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)
39. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18134–18144 (2022)
40. Xu-Darme, R., Quénot, G., Chihani, Z., Rousset, M.C.: Particul: part identification with confidence measure using unsupervised learning. In: International Conference on Pattern Recognition. pp. 173–187. Springer (2022)
41. Yeh, R.A., Hu, Y.T., Ren, Z., Schwing, A.G.: Total variation optimization layers for computer vision. In: CVPR. pp. 711–721 (2022)