Unsupervised Part Discovery from Contrastive Reconstruction

Subhabrata Choudhury; Iro Laina; Christian Rupprecht; Andrea Vedaldi

11.01.2023

Aasaipriya Chandran

Visual Geometry Group
University of Oxford
Oxford, UK

Outline

- Motivation
- Previous work
  - Unsupervised scene decomposition
  - Part discovery and segmentation
  - Self-supervised and contrastive learning
- Unsupervised part discovery
  - Part criteria and losses
  - Dataset details
  - Architecture and Implementation details
  - Evaluation metrics proposed
  - Comparison with the state of the art
- Experiments
- Limitations
- Conclusion

Motivation

- Major works: object or scene level segmentation
- Aim: Part segmentation of object
- Parts invariant to geometric and photometric changes
- Supervised learning requires manual annotations; infeasible. So unsupervised method.
- Unsupervised part discovery: Decompose an object into a collection of repeatable and informative parts without supervision.

Previous work

1. **Unsupervised scene decomposition:**

- Spatially decompose scene into objects - object centric representation
- Representative and discriminative approach
- Performance well on simple, synthetic scenes
- Limitation: Decomposing a scene into object is different from decomposing object into parts.

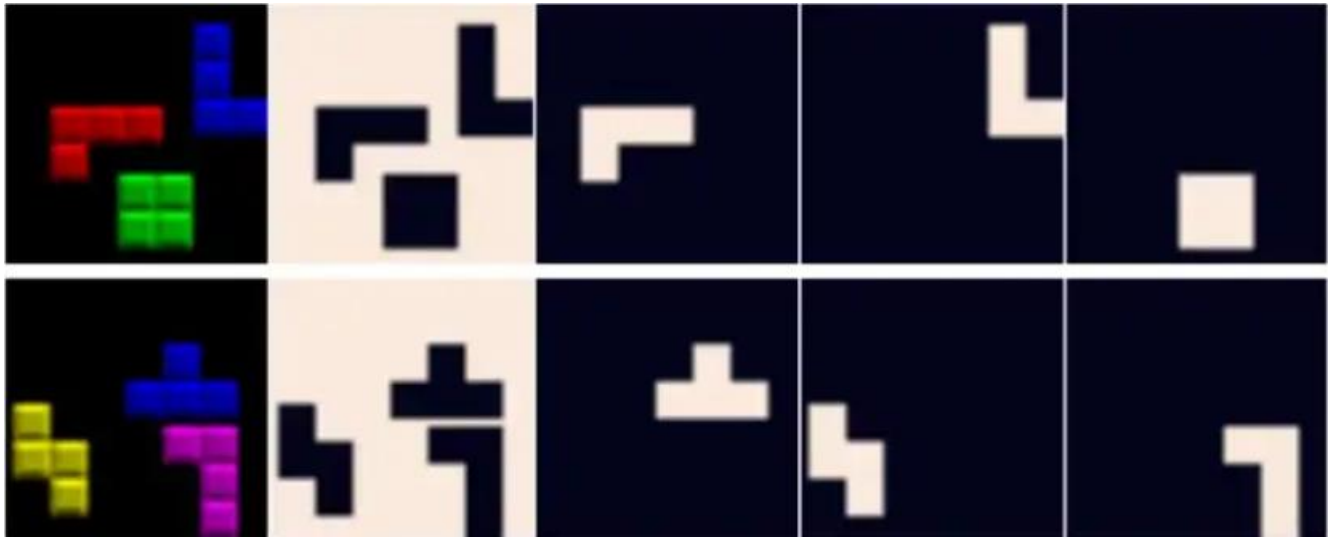


Fig: Introduction to Object Centric Learning by Michele De Vita (https://mik3dev.medium.com/)

Previous work

2. Part discovery and segmentation:

Part based models

- Various parts of the image are used separately
- Requires image labels for training
- Create attention maps to learn parts
- Intermediate step to discover important region to utilize for downstream task (fine-grained classification)

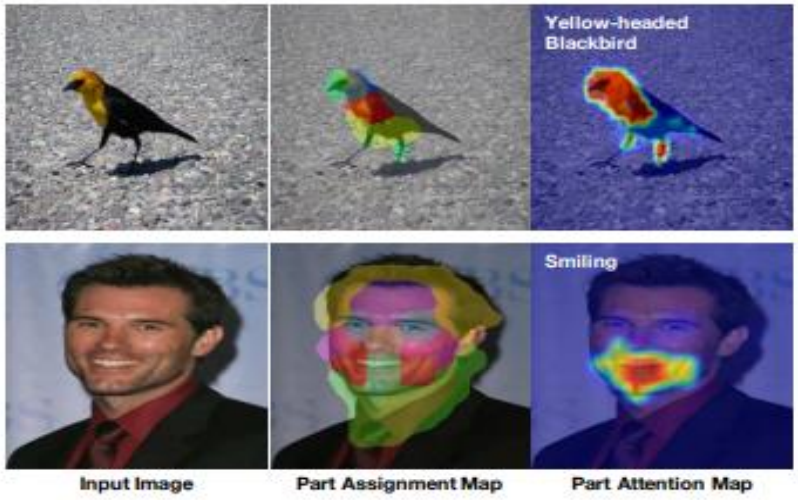


Fig: Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

Previous work

2. Part discovery and segmentation:

DFF (Deep Feature Factorization)

- Feature extraction from deep CNN
- Use NMF (non-negative matrix factorization) to get heat maps

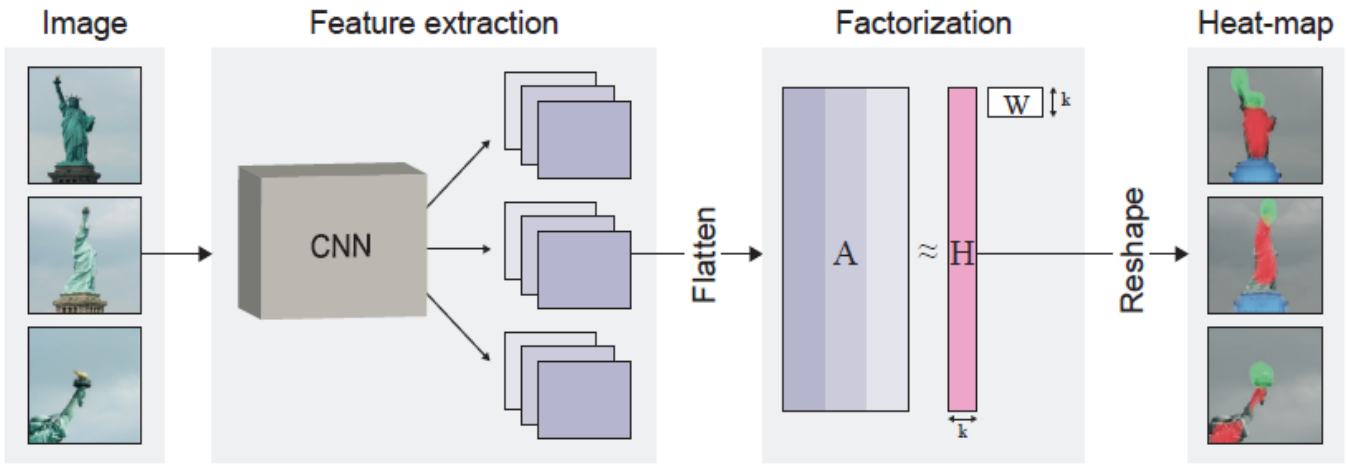


Fig: Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In Proceedings of the European Conference on Computer Vision (ECCV)

Previous work

2. Part discovery and segmentation:

ULD (Unsupervised Landmark Detection)

- Learn landmarks without supervision
- Rely on geometric constrains landmarks equivariance to transformations

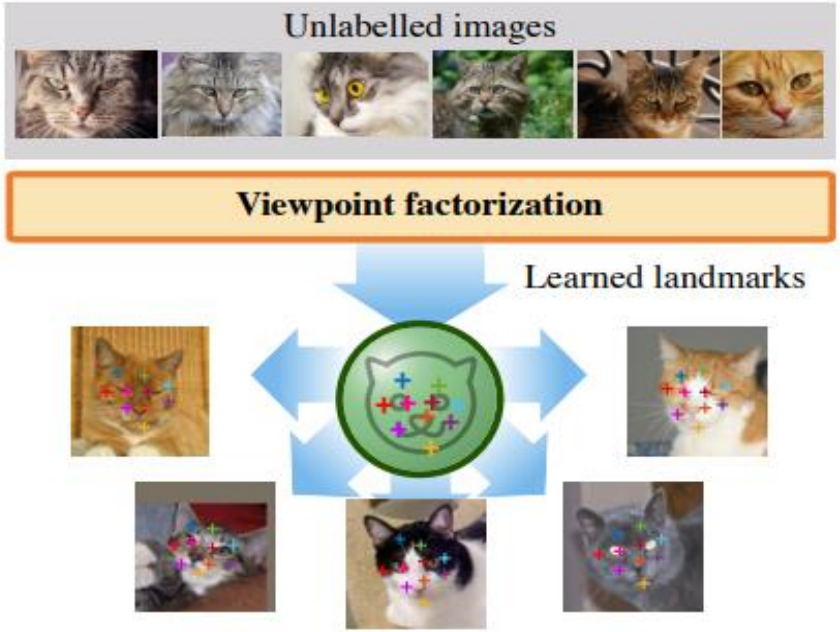


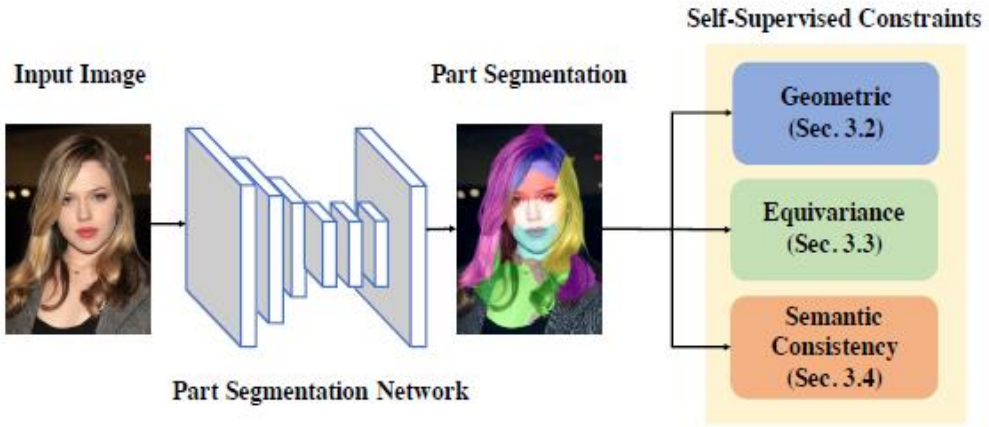
Fig: James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In Proceedings of the IEEE international conference on computer vision

Previous work

2. Part discovery and segmentation:

SCOPS (Self-supervised Co-Part Segmentation)

- Related to proposed paper – loss functions and backbone architecture
- Use features from convolutional layers for pretraining
- Losses – geometric concentration, equivariance, semantic consistency and objects as the union of parts



Other

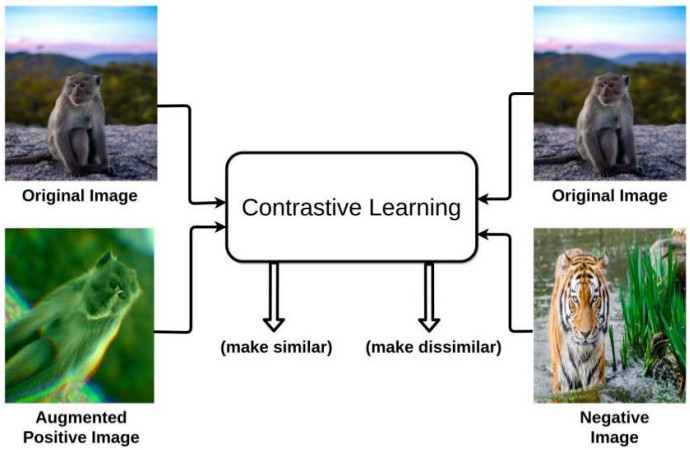
- Generative adversarial methods – no supervision. Use motion in videos
- Probabilistic generative model

Fig: Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Previous work

3. Self-supervised with contrastive learning (Related to current method):

- Pretext/Proxy using contrastive learning
- Key idea of contrastive learning: Encode two similar data points with similar embeddings; pushing the embeddings of dissimilar data apart
- No labels; Use data augmentations to create positive pair
- Learning utilized for downstream task



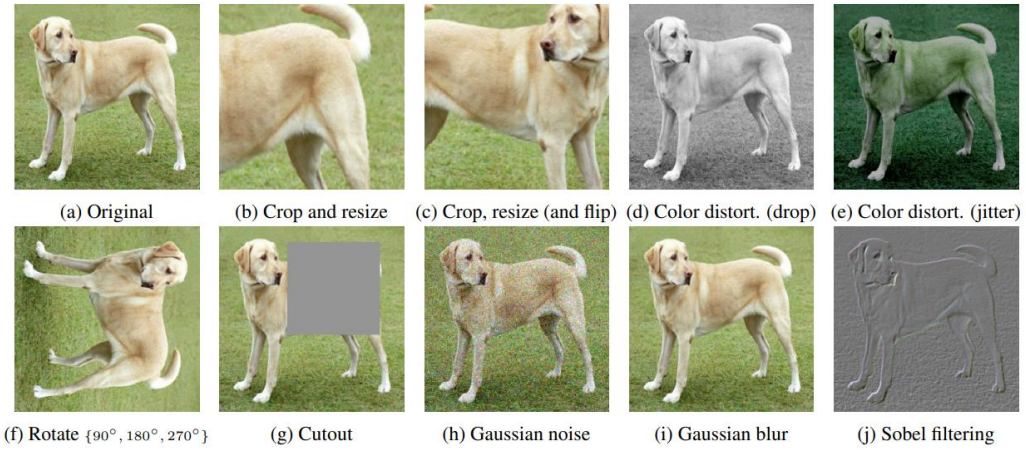


Fig: Advancing Self-Supervised and Semi-Supervised Learning with SimCLR by Ting Chen and Geoffrey Hinton

Unsupervised part discovery

Defining Part: Two main ideas to define parts in unsupervised part segmentation

1. Motion based approach:

"What moves together belongs together"

Learns to group pixels using motion as cue

Limitation: Segmentation not possible when all parts move together

Fig: Sara Sabour, Andrea Tagliasacchi, Soroosh Yazdani, Geoffrey Hinton, and David J Fleet. Unsupervised part representation by flow capsules. In International Conference on Machine Learningz; "Unsupervised Discovery of Parts, Structure, and Dynamics" ICLR 2021

Unsupervised part discovery

Defining Part:

2. Semantic correspondence based:

Learning based on semantic correspondence across collection of images

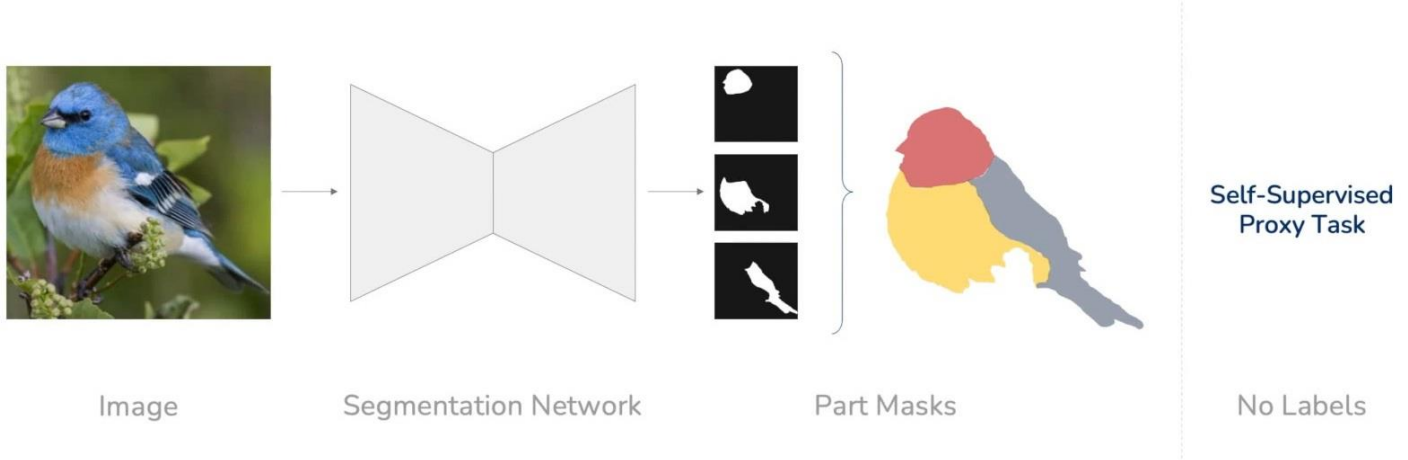
Challenge: Reidentify the same parts across different instances

Proposed method based on this approach

Fig: Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): "Unsupervised Part Discovery by Unsupervised Disentaglement", CGPR 2020

Unsupervised part discovery

- Automatically learn a part detector.
- Assign each pixel to one of the K semantic parts
- No supervision, requires Proxy task.
- Part segmenter is a function $f : I \mapsto M$
- Predict mask $M \in \{0,1\}^{K \times H \times W}$ on image $I \in \mathbb{R}^{3 \times H \times W}$ where $\sum_{k=1}^{K} M_u = 1$



Unsupervised part discovery

Part: Criteria and its loss

•Parts should have uniform feature information – Feature loss

•Parts should be consistent across images and distinct from other parts – Contrastive loss

•Parts should be visually consistent – Visual consistency loss

•Part should be invariant to geometric and photometric transformations – Equivariance loss

Contrastive feature discovery

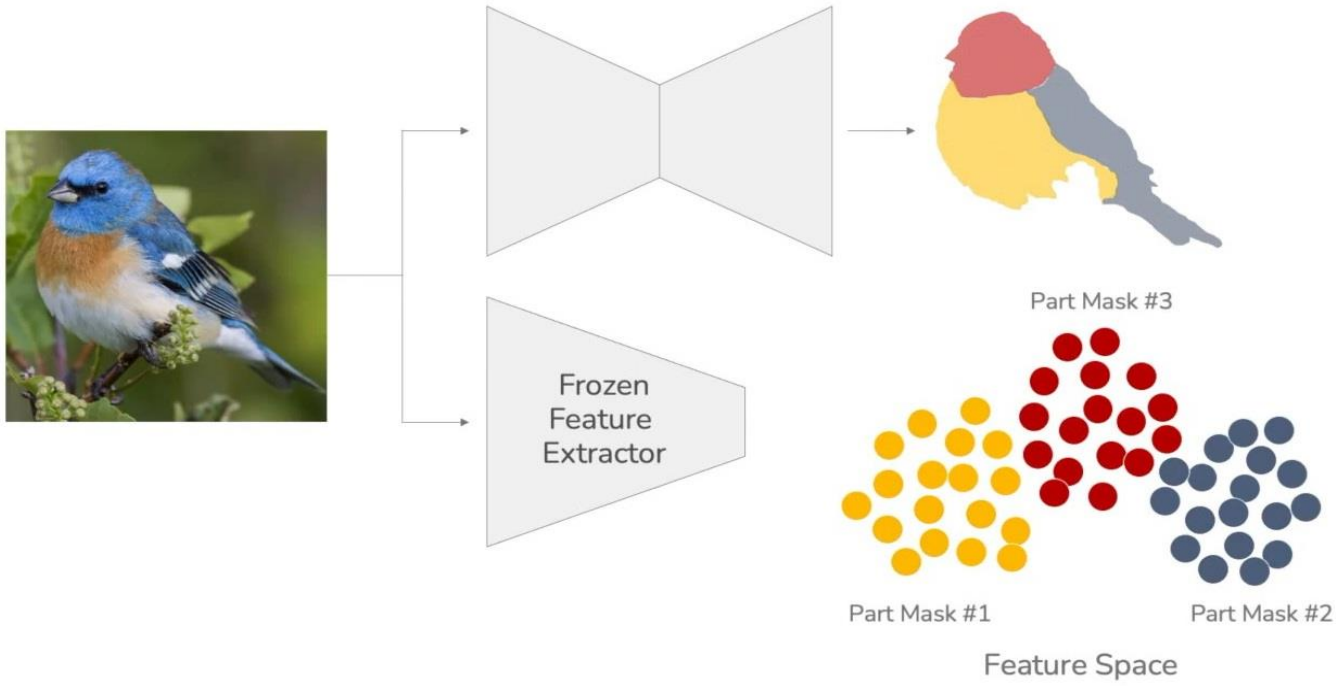
Feature belonging to same part type are similar. Parts should have uniform information

- Average part descriptor

$$z_k(I) = \frac{1}{|M_k|} \sum_{u \in \Omega} M_{ku} \, [\phi(I)]_u, \quad |M_k| = \sum_{u \in \Omega} M_{ku}$$

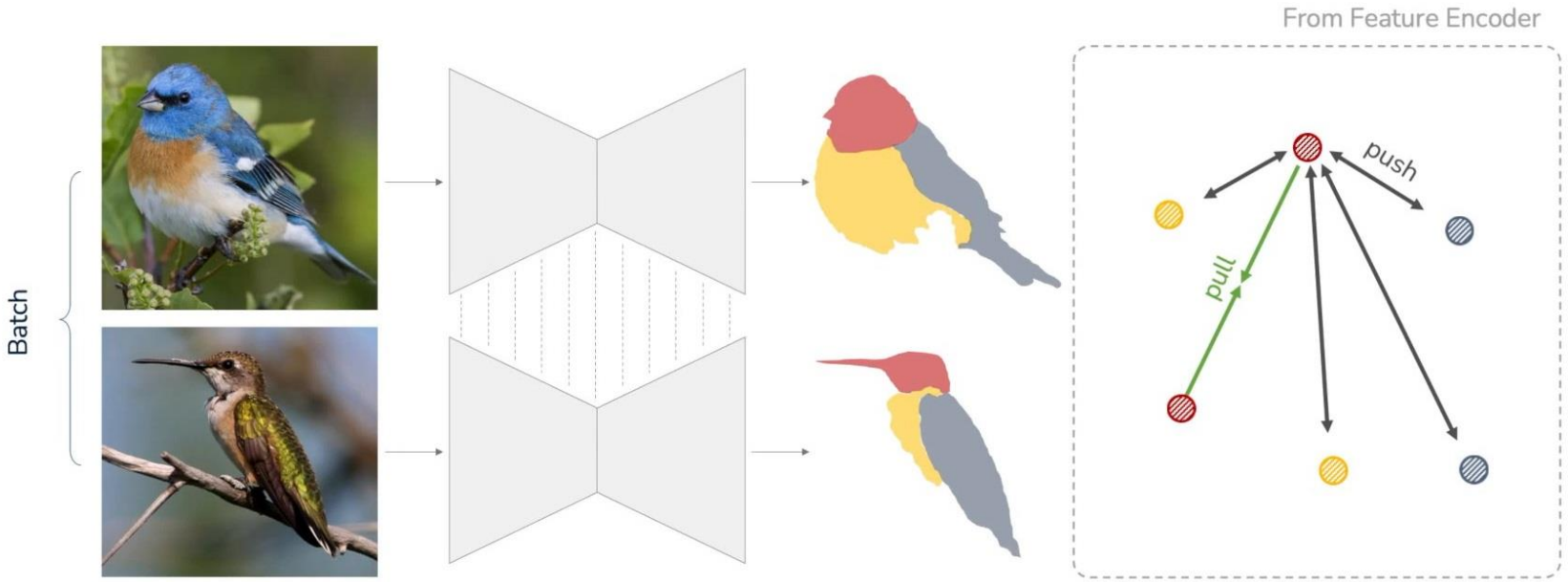
- Feature Loss

$$\mathcal{L}_f(M) = \sum_{k=1}^{K} \sum_{u \in \Omega} M_{ku} \, \|z_k(I) - [\phi(I)]_u\|_2^2.$$



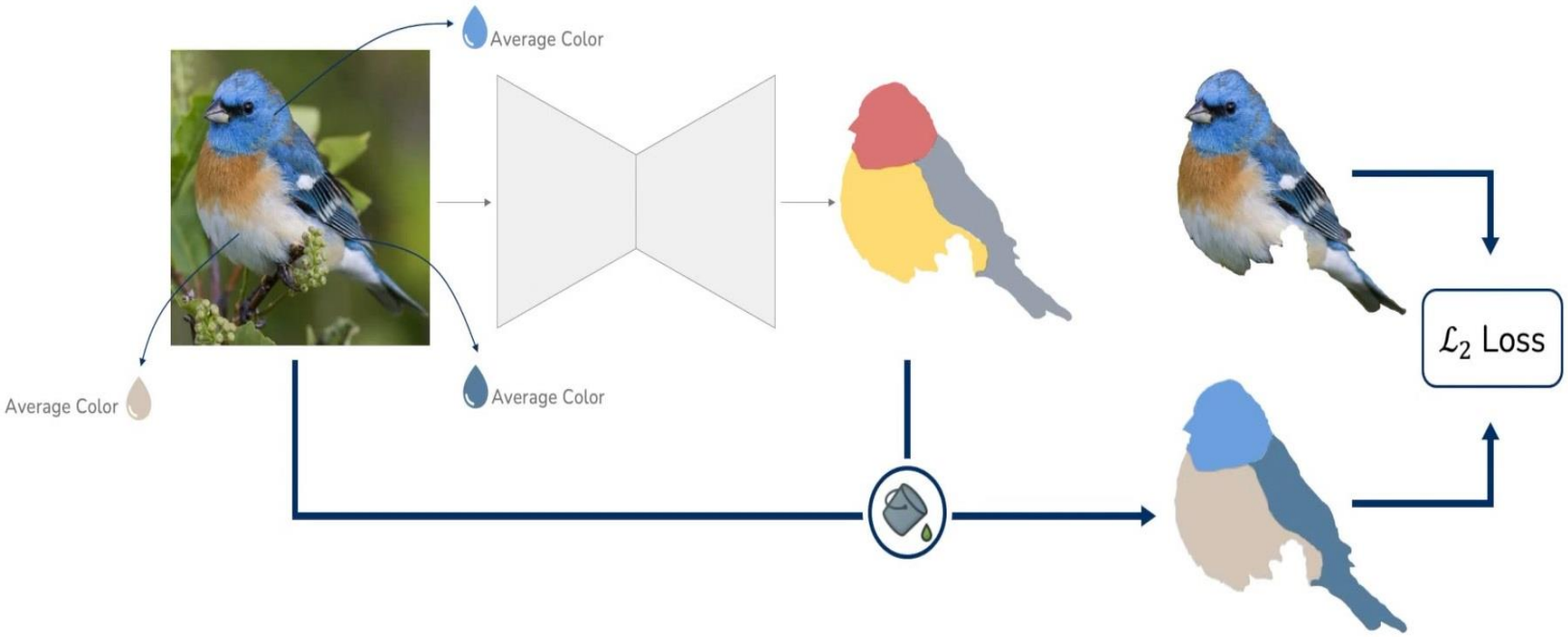
Contrastive feature discovery

- Parts should be distinct across images and distinct from other parts
- Maximize the semantic similarity of same part across images
- Minimize the semantic similarity between all other parts in the same and other images
- Contrastive loss $$\mathcal{L}_c = -\sum_{n=1}^{N}\sum_{k=1}^{K} \log \frac{\exp(z_k^{(n)} \cdot \hat{z}_k^{(n)}/\tau)}{\exp(z_k^{(n)} \cdot \hat{z}_k^{(n)}/\tau) + \sum_{j\neq k}\sum_{i\neq n} \exp(z_k^{(n)} \cdot z_j^{(i)}/\tau)}$$



Visual Consistency

- Parts should be visually consistent. They are roughly uniformly colored.
- Visual Consistency Loss: $\mathcal{L}_v(M) = \sum_{k=1}^{K} \sum_{u \in \Omega} M_{ku} \left\| I_u - \frac{1}{|M_k|} \sum_{v \in \Omega} M_{kv} I_v \right\|_2^2$.



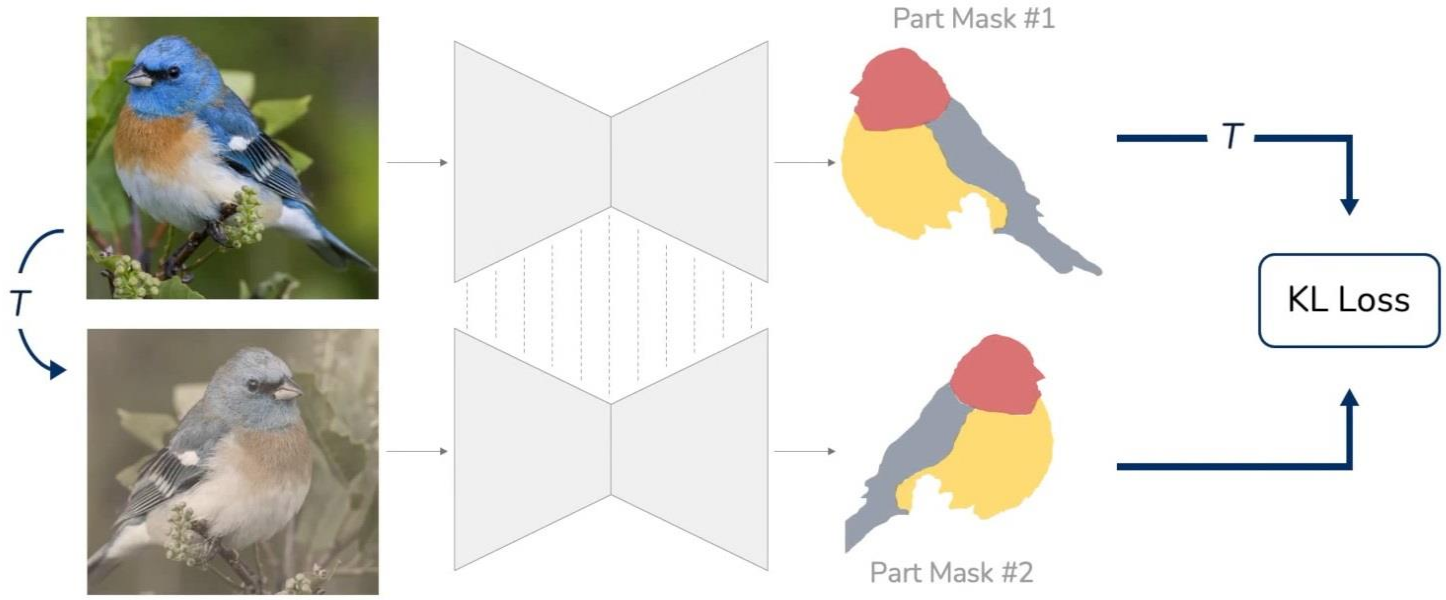
Transformation Equivarience

- Following transformations are applied. color jitter, brightness (±30%), contrast (±30%), saturation (±30%), hue (±30%), random rotations (±60◦) and translations (±10%)

- Commutativity of the function: $T(f(I)) = f(T(I))$

- Equivariance Loss:

$$\mathcal{L}_e(I, T(I)) = \sum_{u \in \Omega} \mathcal{KL}\left(T_u(f(I)), f_u(T(I))\right) + \mathcal{KL}\left(f_u(T(I)), T_u(f(I))\right)$$

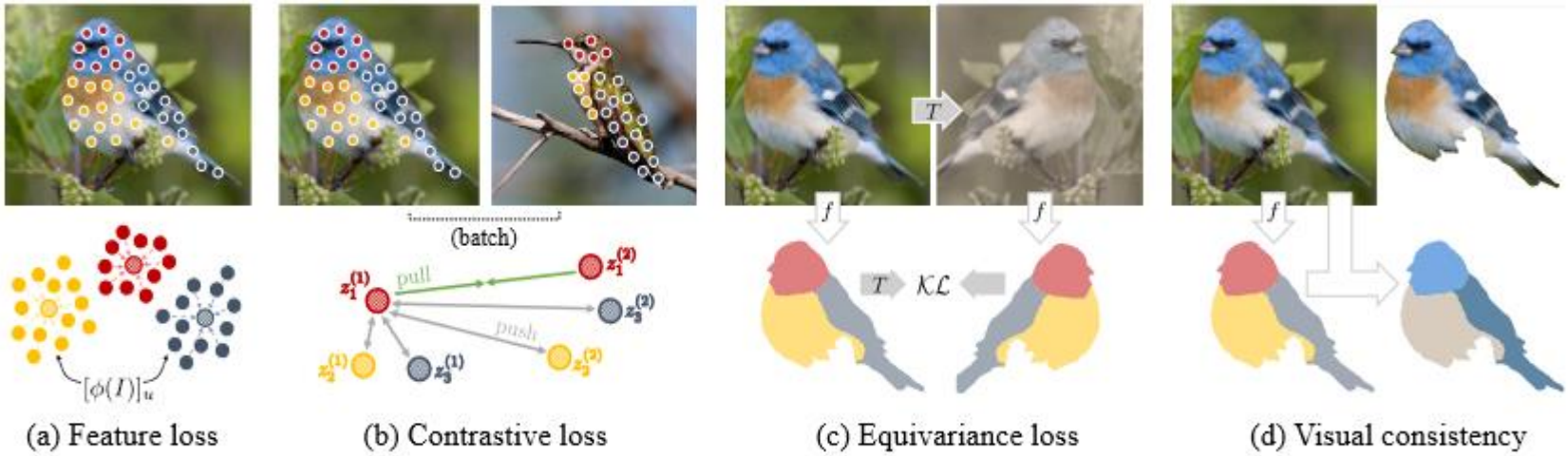


Unsupervised part discovery

Objective:

Learn the function f, by minimizing the weighted sum of the prior losses

$$\lambda_f \mathcal{L}_f + \lambda_c \mathcal{L}_c + \lambda_v \mathcal{L}_v + \lambda_e \mathcal{L}_e$$



(a) Feature loss (b) Contrastive loss (c) Equivariance loss (d) Visual consistency

Dataset

- **The Caltech-UCSD Birds-200 dataset (CUB-200-2011)** - dataset for fine-grained recognition, comprising of 11,788 images of 200 bird species with annotations for 15 part locations

Fig: https://www.researchgate.net/figure/Examples-of-images-in-the-Caltech-UCSD-Birds-200-2011-Dataset-Corresponding-categories_fig6_318204948

Dataset

- **The large-scale fashion database (DeepFashion)** – fashion dataset containing 52,712 densely labelled images of people in different clothing items. The labels include 15 categories and a background class

Fig: https://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html

Dataset

- **PASCAL-Part** - Extension of the PASCAL VOC 2010 dataset; contains 10,103 training and validation images and 9,637 testing images with part level annotations for the 20 categories.
- Current model is trained for following 10 categories – sheep, horse, cow, motorbike, plane, bus, car, bike, dog, cat.

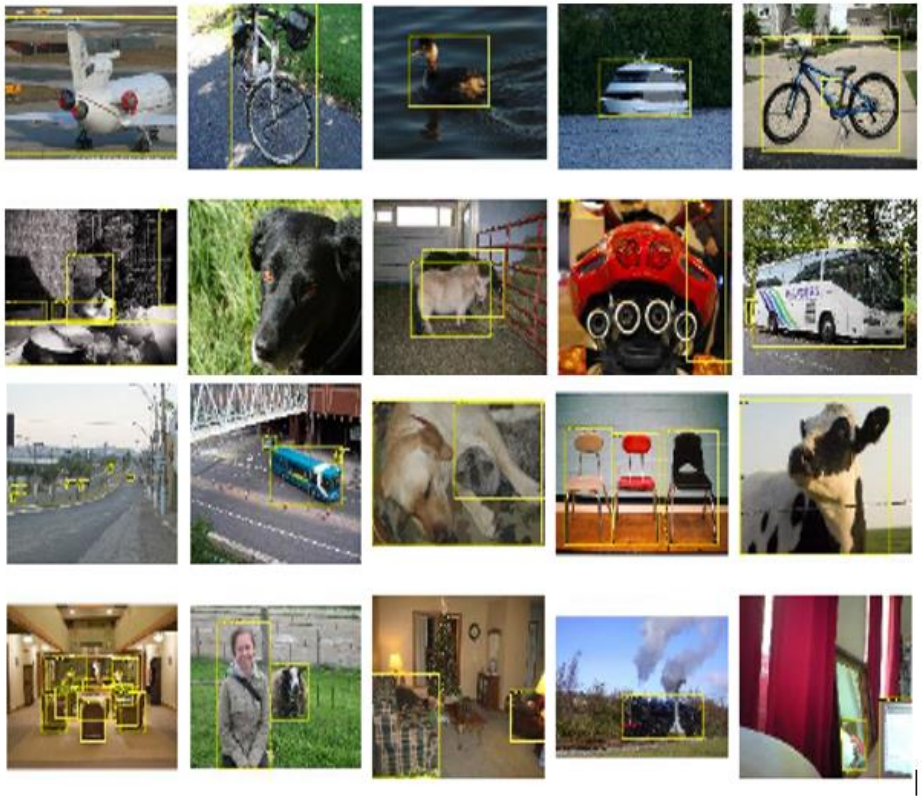


Fig: http://host.robots.ox.ac.uk/pascal/VOC/voc2010/

Architecture and Implementation details

- Model f as deep neural network - **DeepLab-v2 with ResNet-50**. Pretrained on ImageNet.
- Perceptual Network Φ – **VGG19**. For contrastive and feature objectives (L_f *and* L_c). Pretrained on ImageNet. Frozen.
- CUB-200-2011 and DeepFashion: $\lambda_f = 5$, $\lambda_c = 2.3 \cdot 10^3$, $\lambda_v = 30$, $\lambda_e = 5.7 \cdot 10^3$
- PASCAL-Part: $\lambda_f = 5$, $\lambda_c = 2.3 \cdot 10^3$, $\lambda_v = 30$, $\lambda_e = 5.7 \cdot 10^4$, i.e. higher equivariance
- SGD using a learning rate of 10^{-5} ; Weight decay of $5 \cdot 10^{-4}$; Batch size of 6; Image size of 256 × 256
- Trained on foreground pixels only

Evaluation metrics proposed

Previous works' evaluation metrics: Landmark/Keypoint regression error

- Convert part segmentation into Landmark (part center) and evaluate against ground truth
- Use Linear regressor to fit detected landmark against ground truth landmark for training
- If model predicts one single keypoint; then error is low. Does not correlate well with segmentation performance.

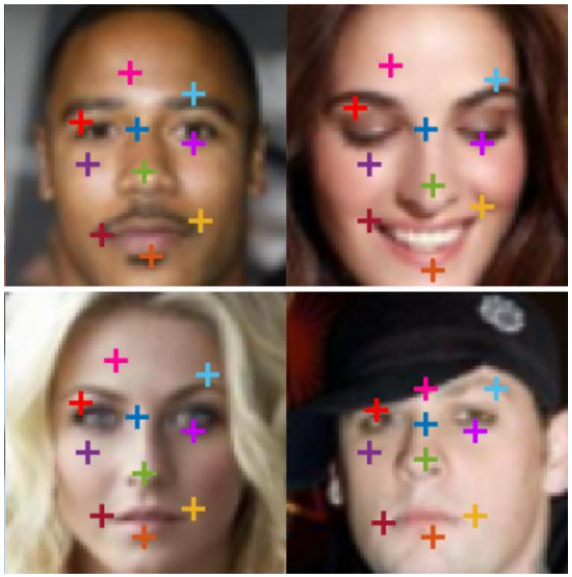


Fig: James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In Proceedings of the IEEE international conference on computer vision

Evaluation metrics proposed

Proposed: Adjusted Rand Index(ARI) to measure the information overlap between predicted and ground truth

$$RI = \frac{\text{Number of Agreeing Pairs}}{\text{Number of Pairs}}$$

$$ARI = \frac{\text{RI - Expected RI}}{\text{Max(RI) - Expected RI}}$$

Evaluation metrics proposed

Proposed: Normalized Mutual Information(NMI) to measure the information overlap between predicted and ground truth

$$NMI(Y, C) = \frac{2 \times I(Y; C)}{[H(Y) + H(C)]}$$

Stricter NMI and ARI using only foreground information: FG-NMI, FG-ARI

Advantages of NMI and ARI: Comparing to Intersection-over-Union (IoU);

- Do not require the ground truth annotation to align exactly
- Do not impose a constraint in the value of number of parts (K)

Comparison with the state of the art

CUB-200

Method	Keypoint Regression Error ↓				FG-NMI↑	FG-ARI↑	NMI↑	ARI↑
	CUB-001	CUB-002	CUB-003	CUB-all				
Image midpoint	27.3	26.7	27.2	23.5	0.0	0.0	0.0	0.0
GT keypoint avg	20.9	22.4	19.9	17.9	0.0	0.0	0.0	0.0
"throat" kpt only	16.4	14.9	15.2	12.1	11.6	-16.2	4.6	-8.3
ULD [68, 86]	30.1	29.4	28.2	-	-	-	-	-
DFF [14]	22.4	21.6	22.0	-	32.4	14.3	25.9	12.4
SCOPS [37] (paper)	18.5	18.8	21.1	-	-	-	-	-
SCOPS [37] (model)	18.3	17.7	17.0	12.6	39.1	17.9	24.4	7.1
Huang and Li [36]†	15.1	17.1	15.7	11.6	-	-	26.1	13.2
Ours	**11.3**	**15.0**	**10.6**	**9.2**	**46.0**	**21.0**	**43.5**	**19.6**

Comparison with the state of the art

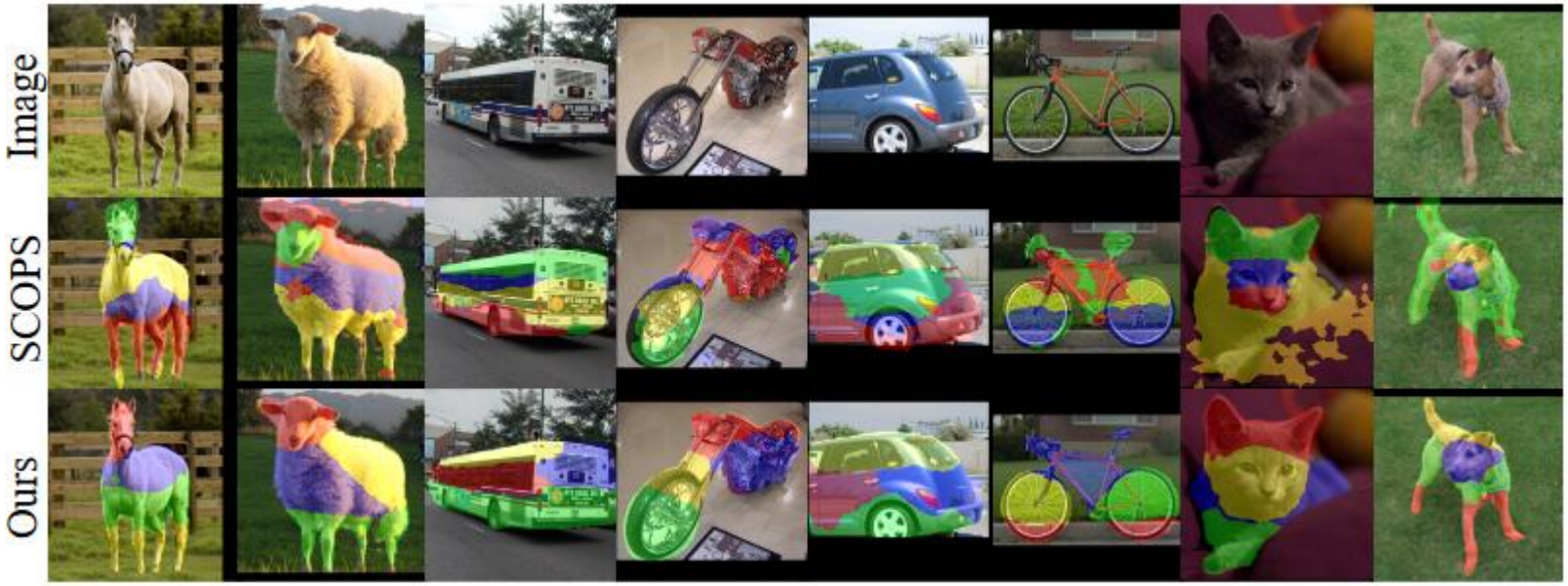
DeepFashion

	FG-NMI	FG-ARI	NMI	ARI
SCOPS [37]	30.7	27.6	56.6	81.4
Ours	**44.8**	**46.6**	**68.1**	**90.6**

Comparison with the state of the art

PASCAL-Part

	NMI										ARI									
	sheep	horse	cow	mbike	plane	bus	car	bike	dog	cat	sheep	horse	cow	mbike	plane	bus	car	bike	dog	cat
DFF [14]	12.2	14.4	12.7	19.1	16.4	13.5	9.0	17.8	14.8	18.0	21.6	32.3	23.3	37.2	38.3	28.5	24.1	39.1	32.3	37.5
SCOPS [37]	26.5	29.4	28.8	35.4	35.1	35.7	33.6	28.9	30.1	33.7	46.3	55.7	51.2	59.2	68.0	66.0	67.1	52.4	52.2	46.6
K-means	34.5	33.3	33.0	38.9	42.8	37.5	**38.4**	35.2	40.4	44.2	58.3	66.8	59.0	63.1	76.8	66.4	70.6	63.2	70.2	71.9
Ours	**35.0**	**37.4**	**35.3**	**40.5**	**45.1**	**38.8**	36.8	**34.8**	**46.6**	**47.9**	**59.8**	**68.9**	**59.7**	**64.7**	**79.6**	**67.6**	**72.7**	**64.7**	**73.6**	**75.4**



Experiments

Ablation experiment:

- Baseline: Clustering the perceptual features of concatenated layers relu5_2 and relu5_4 from VGG19 with K-means
- Remove various parts of the model and measure the decrease in performance.
- L2 instead of contrastive – Replace contrastive loss with simple L2 loss
- Lc w/ different views - use parts in differently augmented versions

Variant		CUB-200-2011 (kp)		DeepFashion (fg)	
		FG-NMI	FG-ARI	FG-NMI	FG-ARI
k-means cluster (VGG19)	[relu5_2, relu5_4]	34.9	14.7	30.3	21.4
w/o consistency within parts	$(\lambda_f = 0)$	29.7	11.7	40.3	40.0
w/o consistency across parts	$(\lambda_c = 0)$	41.3	19.0	39.0	40.1
w/o visual consistency	$(\lambda_v = 0)$	38.5	17.9	31.3	25.2
w/o equivariance	$(\lambda_e = 0)$	29.3	11.2	41.5	42.7
$\mathcal{L}_2$ instead of contrastive	$\mathcal{L}_c = \|z_k^{(n)} - \hat{z}_k^{(n)}\|_2^2$	34.0	13.4	36.7	32.0
$\mathcal{L}_c$ w/ different views		44.4	20.2	36.4	33.4
Ours	(full model)	**46.0**	**21.0**	**44.8**	**46.6**

Experiments

Eliminating Supervision

- Method still rely on backbones pre-trained with ImageNet supervision (IN-1lk)
- Removing these supervised components and replace with unsupervised models
- Below results is determined on CUB-200-2011 dataset.

Backbone of f	Perceptual Network ϕ	FG Mask	FG-NMI	FG-ARI
ResNet50 (IN-1k supervised)	VGG19 (IN-1k supervised)	GT	46.0	21.0
ResNet50 (IN-1k supervised)	VGG16 (IN-1k supervised)	GT	39.7	19.1
ResNet50 (SwAV[7])	VGG16 (IN-1k supervised)	GT	35.4	16.4
ResNet50 (SwAV[7])	VGG16 (DeepCluster-v1 [6])	GT	32.3	14.0
ResNet50 (SwAV[7])	VGG16 (DeepCluster-v1 [6])	[56]	31.9	14.9
ResNet50 (IN-1k supervised)	ViT (DiNO[8])	GT	43.9	19.7
ResNet50 (SwAV[7])	ViT (DINO [8])	[56]	42.7	20.0

Additional Results

Evaluation of different number of parts:

	CUB-200-2011 (kp)				DeepFashion (fg)			
Variant	FG-NMI	FG-ARI	NMI	ARI	FG-NMI	FG-ARI	NMI	ARI
$K = 4$	46.0	21.0	43.5	19.6	44.8	46.6	68.1	90.6
$K = 6$	47.2	23.0	44.4	20.7	43.5	42.2	66.2	91.0
$K = 8$	58.2	34.0	51.5	28.3	39.2	30.7	62.4	90.6

Variability in results: Trained the model with K = 4 (with 5 different random seeds) and below are mean ± standard deviation of NMI and ARI

Dataset	FG-NMI	FG-ARI	NMI	ARI
CUB	45.3 ± 2.8	20.5 ± 1.5	42.8 ± 1.7	19.2 ± 0.5
DeepFashion	44.6 ± 0.4	46.1 ± 0.6	68.2 ± 0.2	90.7 ± 0.1

Additional Results

- CUB-200 with K=6 and K=8

(a) CUB, K=6

(b) CUB, K=8

Additional Results

- DeepFashion with K=6 and K=8

(c) DeepFashion, K=6

(d) DeepFashion, K=8

Limitations

- Enforced visual consistency objective (e.g., the wheels of a car or a striped garment)
- Parts discovered in a self-supervised manner might not necessarily agree with human intuition (e.g., for humans one could segment arms, legs, torso, head or decompose arms into hands, fingers, etc)
- Failing to separate the foreground from the background

Conclusion

Proposed method expands on prior work by introducing constraints on contrastive formulation, equivariance and visual consistency in segmenting the object parts.

Few opinions for open discussion:

- Is using ‘supervised features’ reasonable for the model’s motivation of being unsupervised task?
- K=4 looks more visually better in part detection compared to other K values.

References

- James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In Proceedings of the IEEE international conference on computer vision, pages 5916–5925, 2017.
- Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2694–2703, 2018.
- Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In Proceedings of the European Conference on Computer Vision (ECCV), pages 336–352, 2018.
- Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8662–8672, 2020.
- Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
- Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.
- David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6541–6549, 2017.

Discussion/Questions

Thank you !!!

Discussion/Questions?