CrOC: Cross-View Online Clustering for Dense Visual Representation Learning

Thomas Stegmüller*, Tim Lebailly*, Behzad Bozorgtabar, Tinne Tuytelaars, Jean-Philippe Thiran

EPFL - KU LEUVEN - CHUV

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

SCAN ME

MOTIVATION

Objective

- Self-supervised pre-training aligned with dense downstream tasks.

Challenges

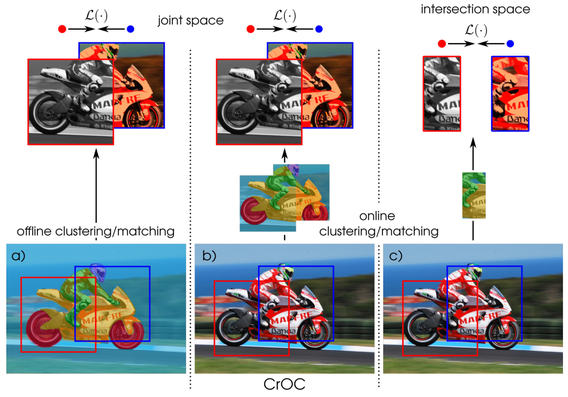
- Where are the objects in the image?
- How to match objects from both views?
- Compatibility with scene-centric images?

Ideas

- Online clustering of the features.
- Cluster the views conjointly.
- Cluster pruning.

1

JOINT SPACE CLUSTERING



Object clustering

Clustering of the joint views driven by semantics:

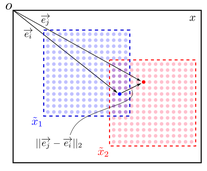
$$\mathbf{T}^{(\text{sem})} = -\mathbf{Z}_{\text{cat}}\mathbf{C}^\top \qquad \mathbf{Q}^* = \underset{\mathbf{Q}\in\mathcal{U}(\mathbf{r},\mathbf{c})}{\arg\min} < \mathbf{Q}, \mathbf{T}^{(\text{sem})} > -\frac{1}{\lambda}H(\mathbf{Q})$$

$$\mathcal{U}(\mathbf{r},\mathbf{c}) = \{\mathbf{Q} \in \mathbb{R}_+^{2N\times K} \mid \mathbf{Q}\mathbf{1}_K = \mathbf{r}, \mathbf{Q}^\top\mathbf{1}_{2N} = \mathbf{c}\}$$

Positional prior

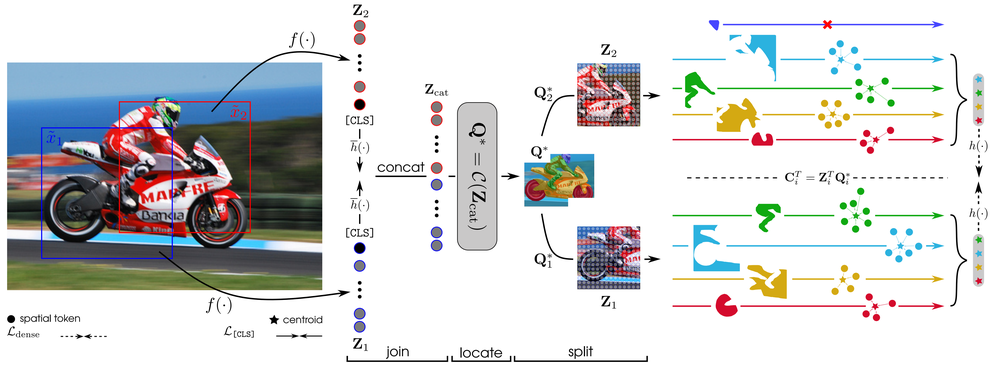
$$\mathbf{T}^{(\text{pos})}_{ij} = \frac{1}{S}||\mathbf{e}_i^{(\text{cat})} - \mathbf{e}_j^{(\text{cen})}||_2$$

$$\mathbf{T}^{(\text{tot})} = \mathbf{T}^{(\text{sem})} + \lambda_{\text{pos}}\mathbf{T}^{(\text{pos})}$$



2

CENTROID-LEVEL SELF-DISTILLATION



Centroids representation

The 2N concatenated spatial tokens from both views are clustered:

$$\mathbf{Q}^* = \mathcal{C}(\mathbf{Z}_{\text{cat}})$$

Centroids from each views are computed by splitting the assignments view-wise:

$$\mathbf{C}^\top_{\{t,s\},\{1,2\}} = \mathbf{Z}^\top_{\{t,s\},\{1,2\}}\mathbf{Q}^*_{\{1,2\}}$$

Self-supervised loss

Consistency is enforced between valid pairs of centroids...

$$\mathcal{L}_{\text{dense}} = \frac{1}{2}\left(H(\mathbf{P}_{t,1}, \mathbf{P}_{s,2}) + H(\mathbf{P}_{t,2}, \mathbf{P}_{s,1})\right)$$

... but not only:

$$\mathcal{L}_{\text{glob}} = \frac{1}{2}\left(H(\boldsymbol{p}_{t,1}, \boldsymbol{p}_{s,2}) + H(\boldsymbol{p}_{t,2}, \boldsymbol{p}_{s,1})\right)$$

3

RESULTS

Linear segmentation + frozen backbone

Method	Model / Dataset	PVOC12	COCO-Thing	COCO-Stuff	Avg.
Global features					
BYOL	ResNet50 / COCO+	38.7	50.4	39.8	43.0
DINO	ViT-S/16 / COCO	47.2	47.1	46.2	46.8
Local features					
ORL	ResNet50 / COCO+	45.2	55.6	45.6	48.8
DenseCL	ResNet50 / ImageNet	57.9	60.4	47.5	55.3
SoCo	ResNet50 / ImageNet	54.0	56.8	44.2	51.7
ReSim	ResNet50 / ImageNet	55.1	57.7	46.5	53.1
PixPro	ResNet50 / ImageNet	57.1	54.7	45.9	52.6
VICRegL	ResNet50 / ImageNet	58.9	58.7	48.2	55.3
MAE	ViT-S/16 / COCO	31.7	35.1	39.6	35.5
CP^2	ViT-S/16 / ImageNet+PVOC12	63.1	59.4	46.5	56.3
Ours					
CrOC	ViT-S/16 / COCO	54.5	55.6	49.7	53.3
CrOC	ViT-S/16 / COCO+	60.6	62.7	51.7	58.3
CrOC	ViT-S/16 / ImageNet	**70.6**	**66.1**	**52.6**	**63.1**

Online joint clusters visualization

4