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

SCAN ME

CVPR 2023

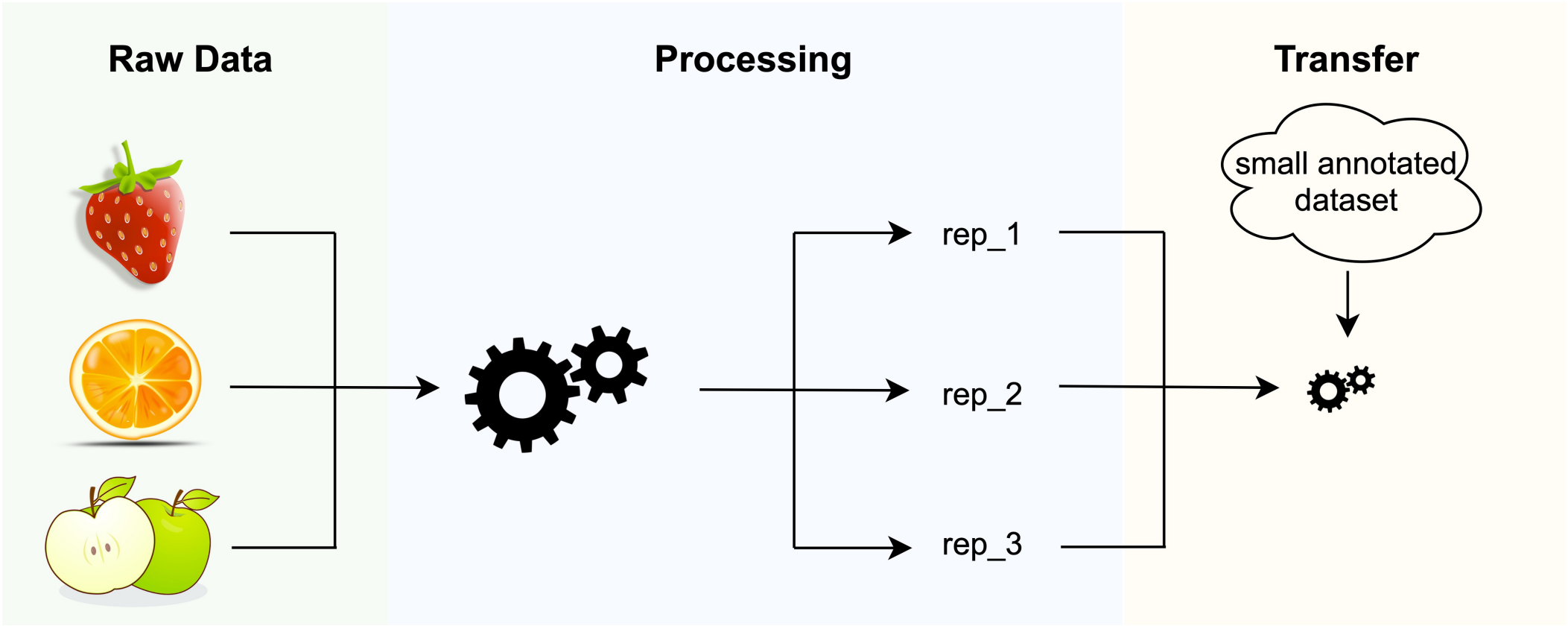
Thomas Stegmüller*, Tim Lebailly*, Behzad Bozorgtabar, Tinne Tuytelaars and Jean-Philippe Thiran

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

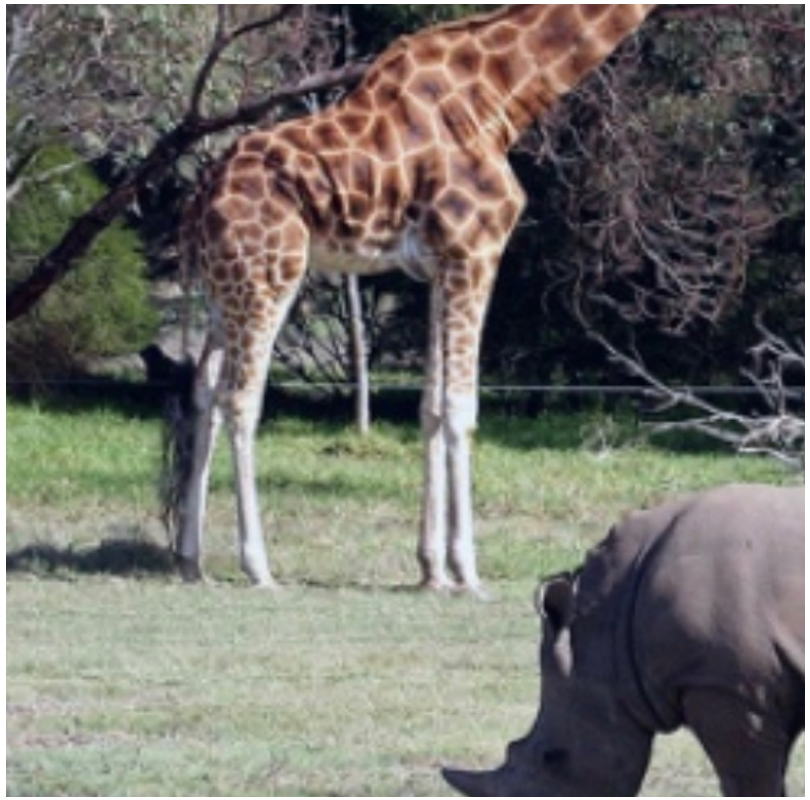
EPFL

KU LEUVEN

PROBLEM SETTING



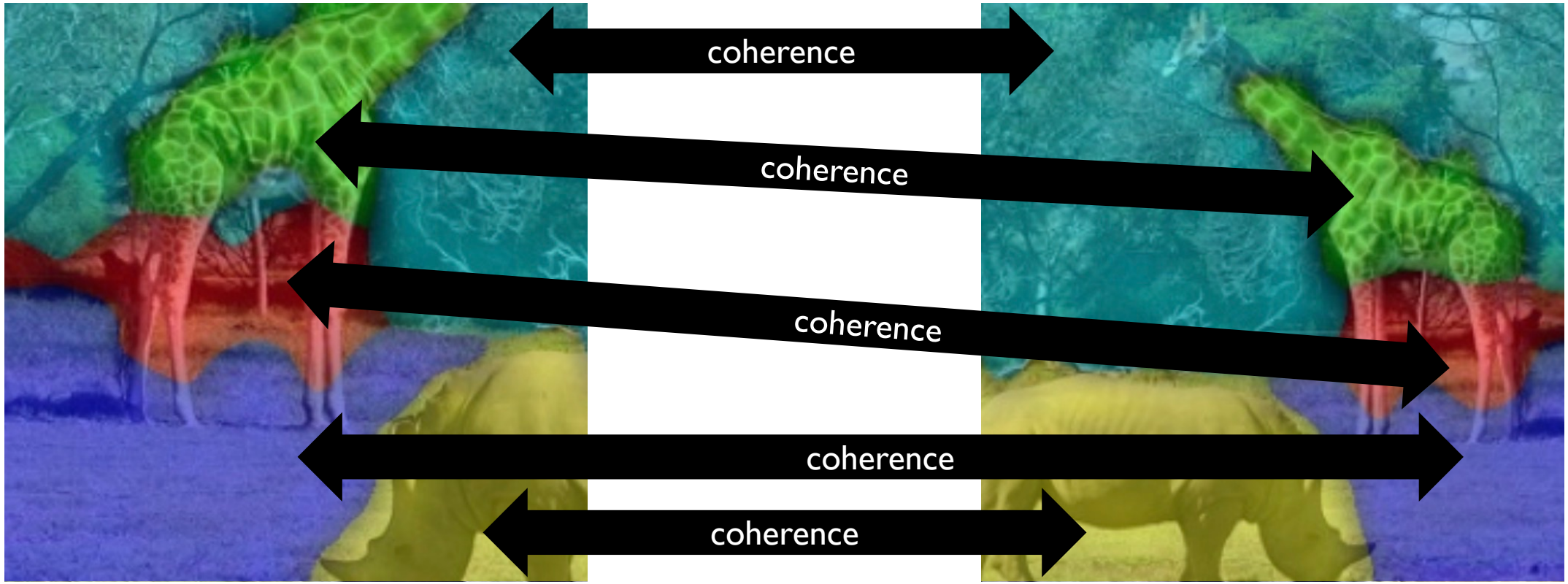
PROBLEM SETTING



coherence

PROBLEM SETTING

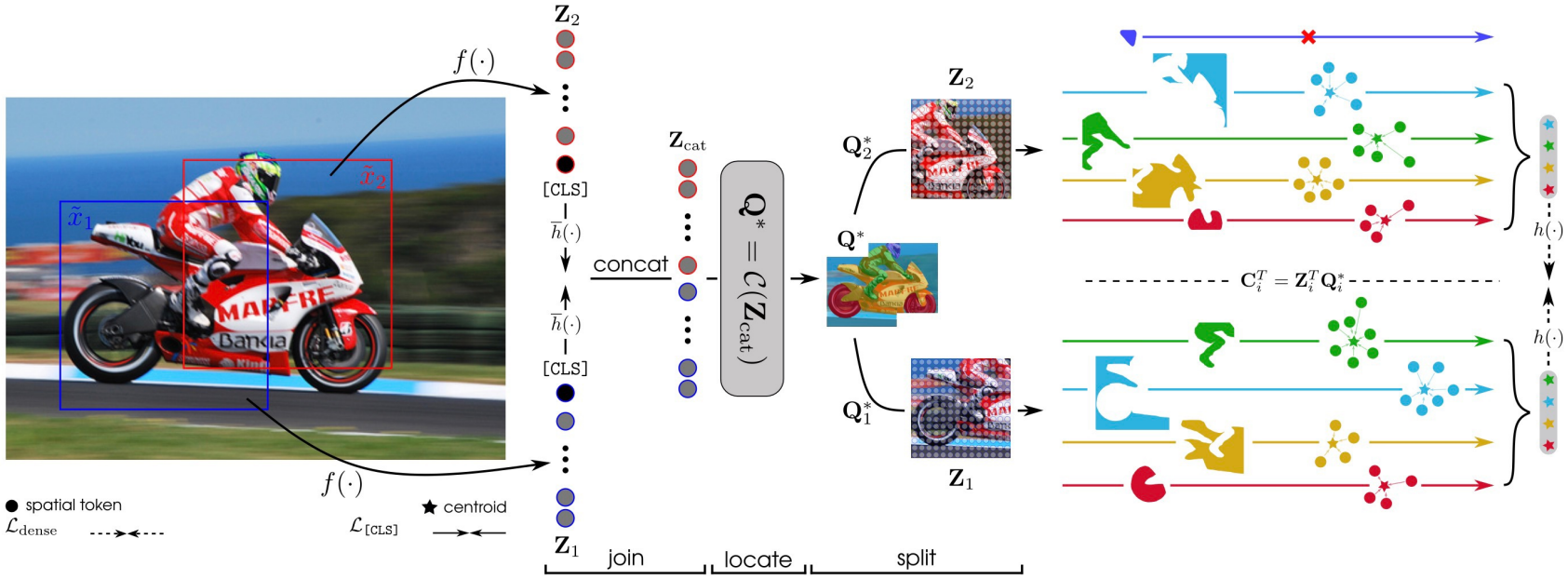
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA



OBJECT-LEVEL SELF-DISTILLATION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Where are the objects ?
- How to match objects ?
- Efficiency ?



1. IMAGE-LEVEL SELF-DISTILLATION
2. OBJECT-LEVEL SELF-DISTILLATION
3. RESULTS

IMAGE-LEVEL SELF-DISTILLATION

EPFL KU LEUVEN

IMAGE-LEVEL SELF-DISTILLATION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Encoder:
  f
- Augmented view:
  $\tilde{\boldsymbol{x}}_1$
- Dense representation:
  $\mathbf{Z}_1 \in \mathbb{R}^{N \times d}$
- Image-level representation:
  $\bar{\boldsymbol{z}}_1 = \mathbf{avg}(\mathbf{Z}_1) \in \mathbb{R}^d$

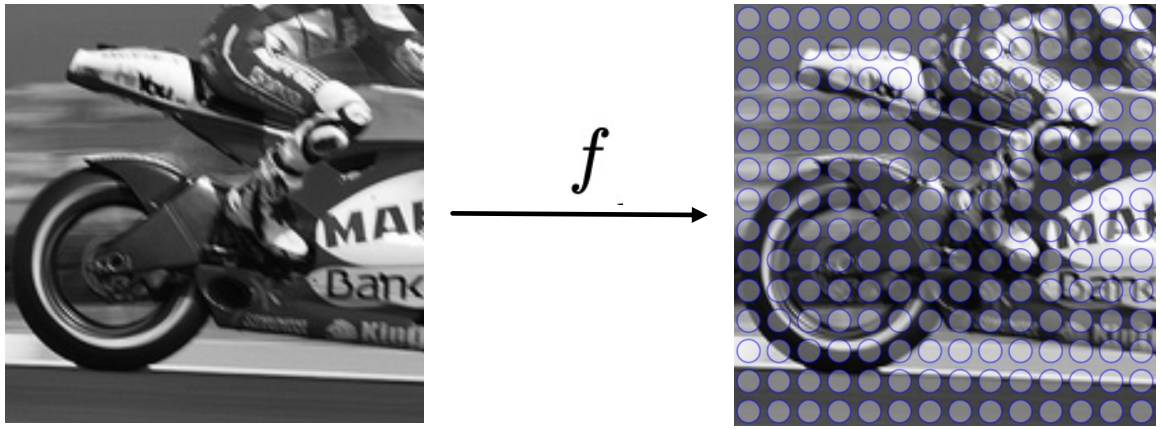


IMAGE-LEVEL SELF-DISTILLATION

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

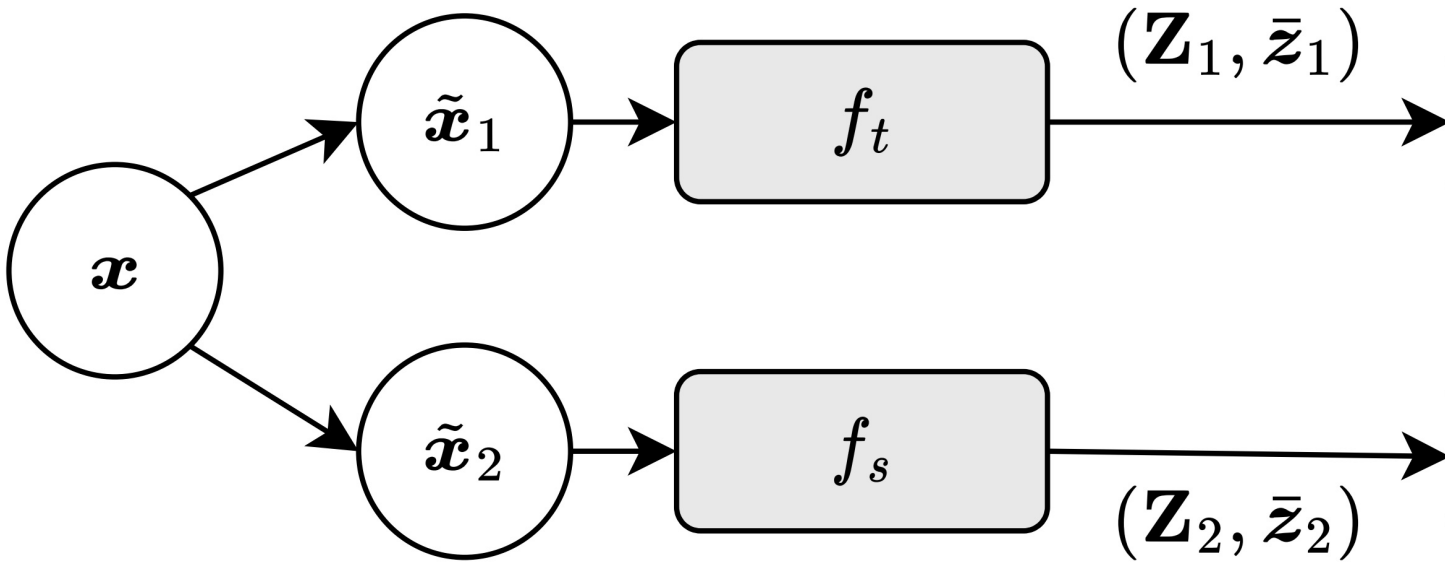


IMAGE-LEVEL SELF-DISTILLATION

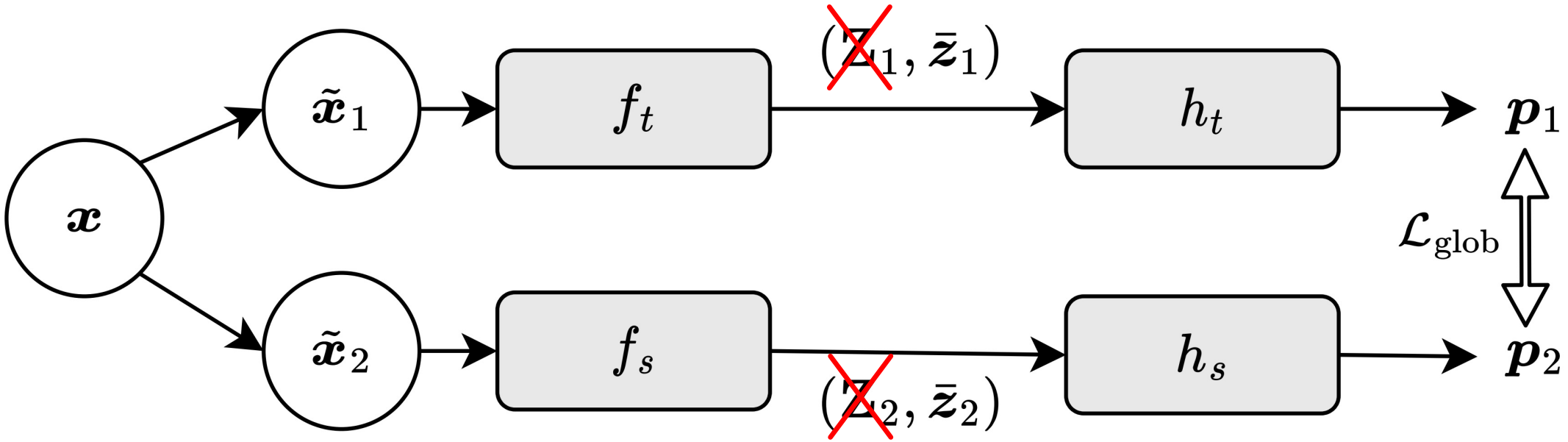


IMAGE-LEVEL SELF-DISTILLATION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Loss function :

$$\boldsymbol{p}_{t,\{1,2\}} = \underset{\bar{L}}{\texttt{softmax}}\left(\bar{h}_t(\bar{\boldsymbol{z}}_{t,\{1,2\}}/\bar{\tau}_t)\right)$$

$$\boldsymbol{p}_{s,\{1,2\}} = \underset{\bar{L}}{\texttt{softmax}}\left(\bar{h}_s(\bar{\boldsymbol{z}}_{s,\{1,2\}}/\bar{\tau}_s)\right)$$

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

- Self-distillation:

$$\boldsymbol{\theta}_t \leftarrow \lambda\boldsymbol{\theta}_t + (1-\lambda)\boldsymbol{\theta}_s$$

OBJECT-LEVEL SELF-DISTILLATION

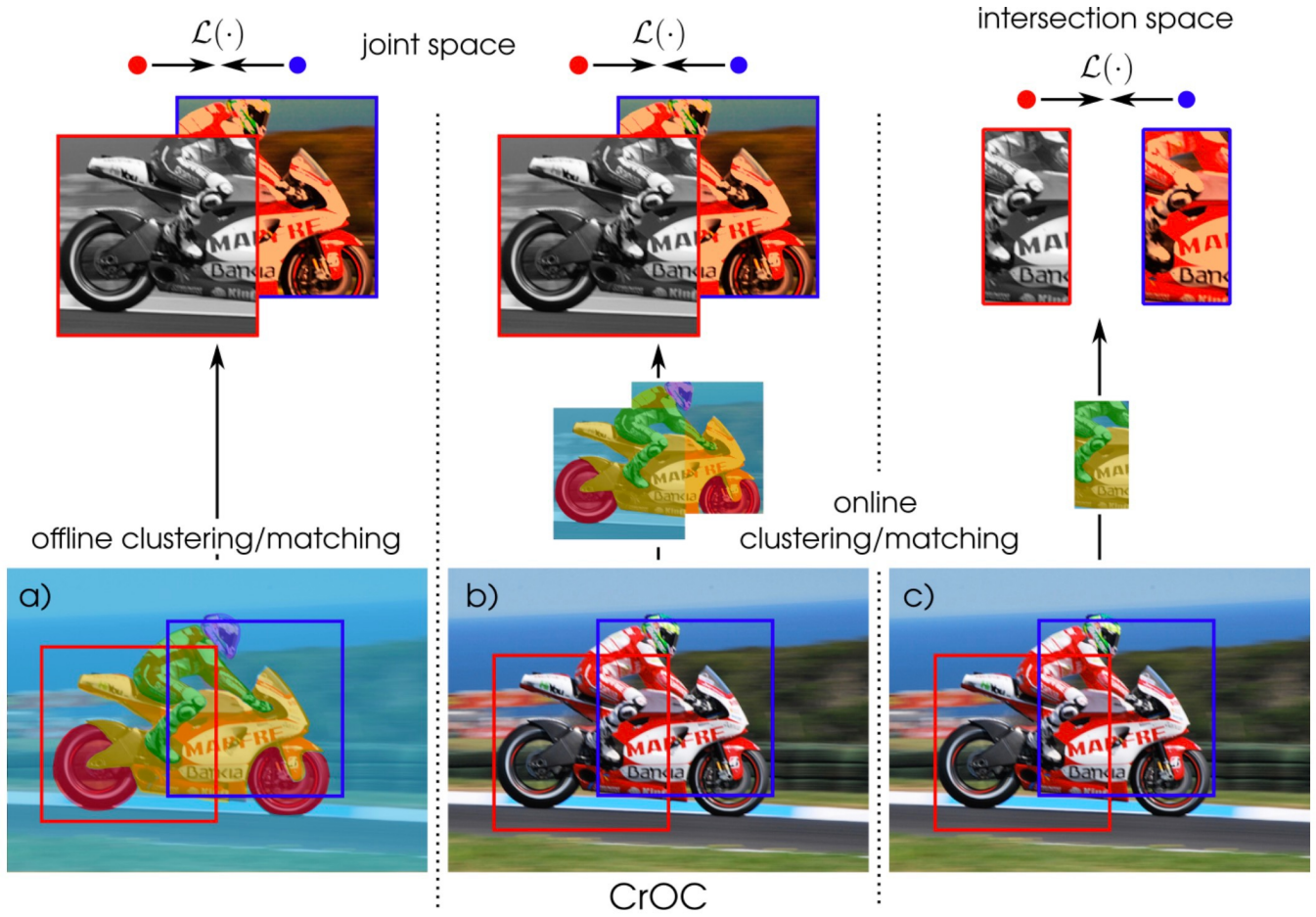
EPFL KU LEUVEN

OBJECT-LEVEL SELF-DISTILLATION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Goal:
  1. Self-distill object representations cross-view.
- Requirements:
  1. Object representations in each view.
  2. A matching between representations from each view.

RELATED WORKS



OBJECT-LEVEL SELF-DISTILLATION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

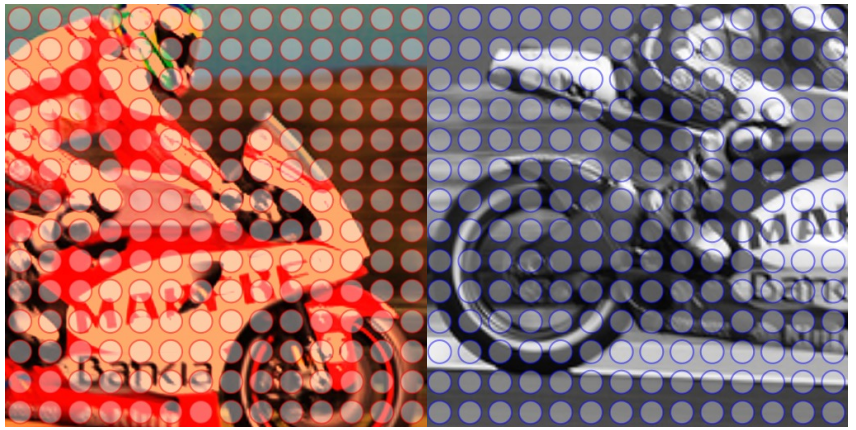
- Discover objects:

$$\mathbf{Q}^* = \mathcal{C}(\mathbf{Z}_{\text{cat}}) \in \mathbb{R}^{2N \times K}$$

- Split assignment view-wise:

$$\mathbf{Q}^*_{\{1,2\}} \in \mathbb{R}^{N \times K}$$

→ free cross-view matching !



$$\mathbf{Z}_{\text{cat}} \in \mathbb{R}^{2N \times d}$$

- Compute object-level representations:

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

OBJECT-LEVEL SELF-DISTILLATION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Object-level loss:

$$\mathbf{P}_{t,\{1,2\}} = \underset{L}{\texttt{softmax}}\left(h_t(\mathbf{C}_{t,\{1,2\}})/\tau_t\right)$$

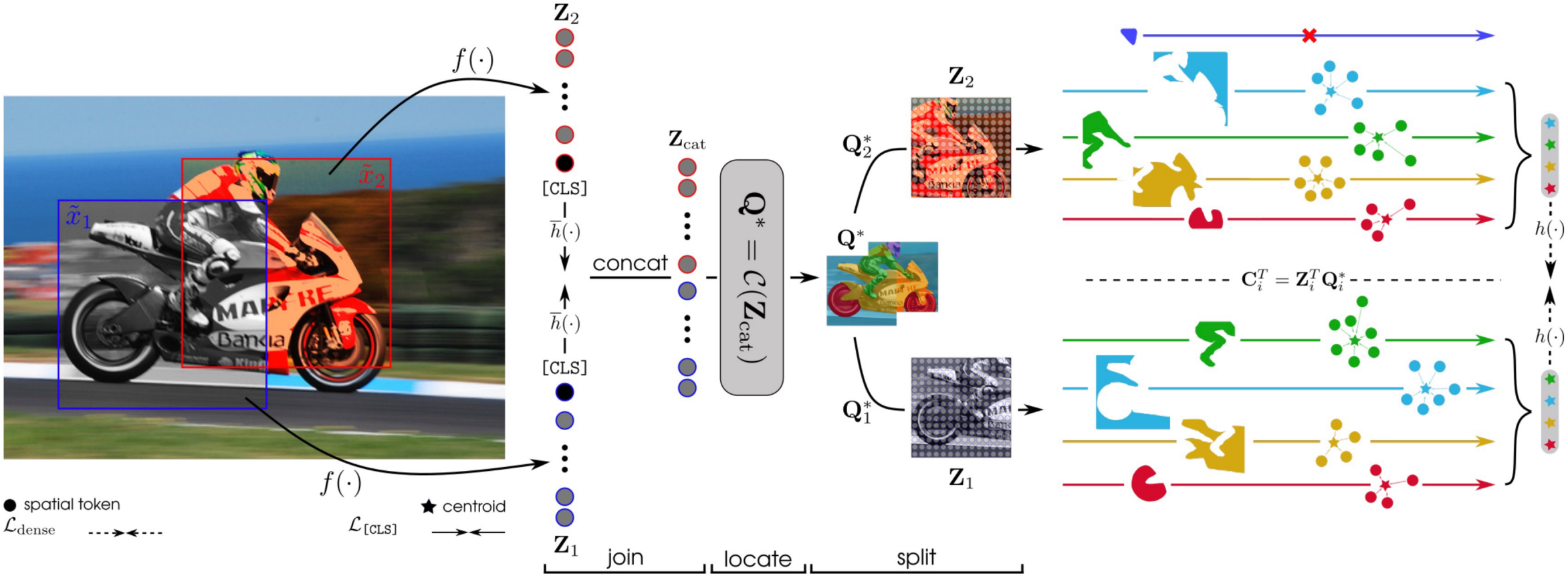
$$\mathbf{P}_{s,\{1,2\}} = \underset{L}{\texttt{softmax}}\left(h_s(\mathbf{C}_{s,\{1,2\}})/\tau_s\right)$$

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

$$\text{where } H(\mathbf{A}, \mathbf{B}) = -\frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{L}\mathbf{A}_{kl}\log(\mathbf{B}_{kl})$$

OBJECT-LEVEL SELF-DISTILLATION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA



JOINT SPACE CLUSTERING

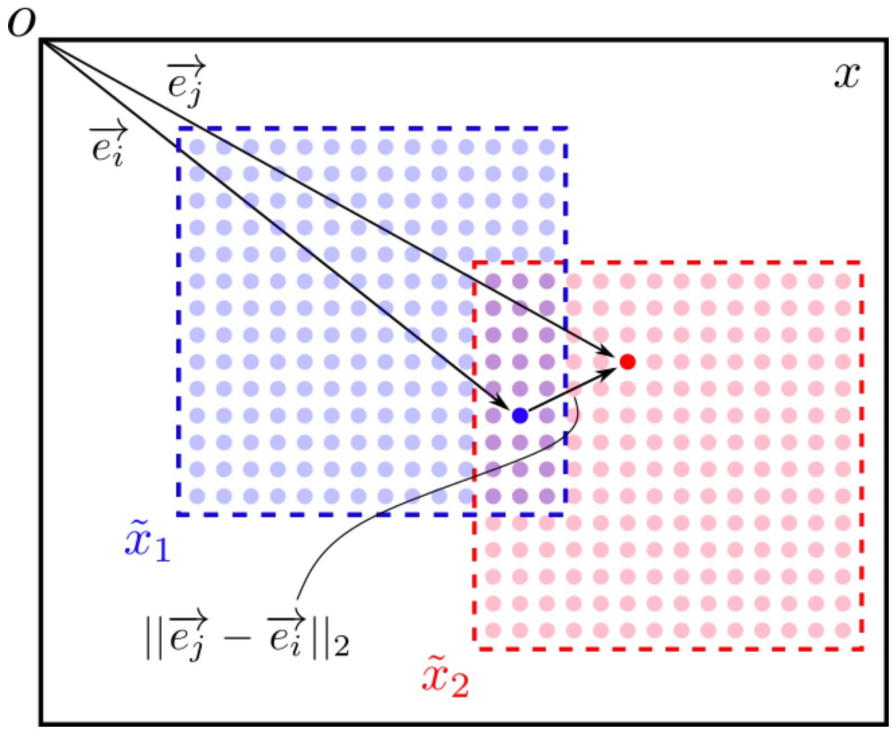
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Sinkhorn Knopp algorithm:
  - Constrained (avoid collapse).
  - Fast.
- Semantic cost

$$\mathbf{T}^{(\text{sem})} = -\mathbf{Z}_{\text{cat}}\mathbf{C}^{\top}$$

- Positional cost

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



JOINT SPACE CLUSTERING

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Total cost:

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

- Optimization:

$$\mathbf{Q}^* = \underset{\mathbf{Q} \in \mathcal{U}(\mathbf{r},\mathbf{c})}{\arg\min} < \mathbf{Q}, \mathbf{T}^{(\text{tot})} > -\frac{1}{\lambda} H(\mathbf{Q})$$

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

RESULTS

EPFL KU LEUVEN

ONLINE CLUSTER VISUALIZATION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

ONLINE CLUSTER VISUALIZATION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Spans a single view only

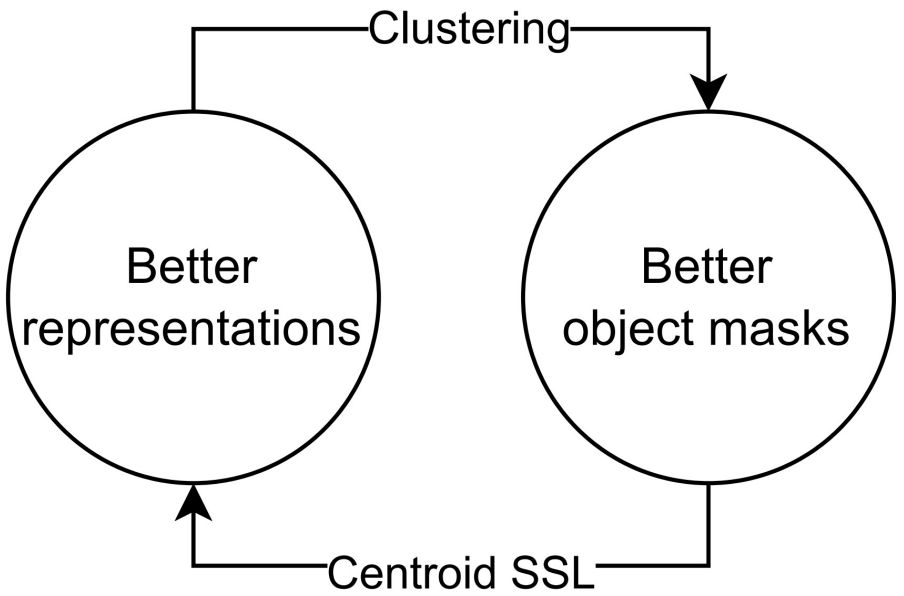
DOWNSTREAM LINEAR SEGMENTATION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Method	Model / Dataset	PVOC12	CC-Th.	CC-St.	Avg.
Global features					
BYOL [15]	ResNet50 / CC+	38.7	50.4	39.8	43.0
DINO [6]	ViT-S/16 / CC	47.2	47.1	46.2	46.8
Local features					
ORL [47]	ResNet50 / CC+	45.2	55.6	45.6	48.8
DenseCL [41]	ResNet50 / IN	57.9	60.4	47.5	55.3
SoCo [43]	ResNet50 / IN	54.0	56.8	44.2	51.7
ReSim [45]	ResNet50 / IN	55.1	57.7	46.5	53.1
PixPro [48]	ResNet50 / IN	57.1	54.7	45.9	52.6
VICRegL [2]	ResNet50 / IN	58.9	58.7	48.2	55.3
MAE [18]	ViT-S/16 / CC	31.7	35.1	39.6	35.5
CP^2 [39]	ViT-S/16 / IN+PVOC12	<u>63.1</u>	59.4	46.5	56.3
Ours					
CrOC	ViT-S/16 / CC	54.5	55.6	49.7	53.3
CrOC	ViT-S/16 / CC+	60.6	<u>62.7</u>	<u>51.7</u>	<u>58.3</u>
CrOC	ViT-S/16 / IN	**70.6**	**66.1**	**52.6**	**63.1**

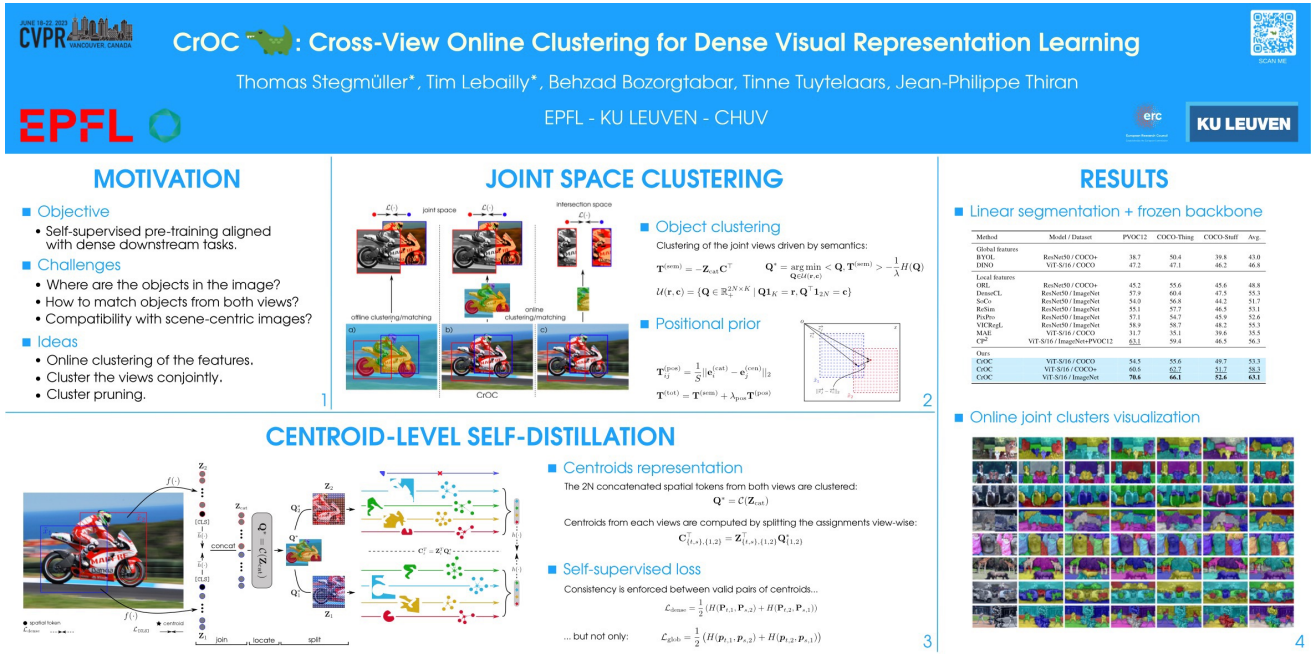
CONCLUSION

- Novel online centroid-level self-distillation algorithm.
- Clustering/matching in single step:
  - Efficient.
  - Free cross-view matching.
  - Object present only in single view ?
- Excellent performance on dense downstream tasks.



SEE YOU AT THE POSTER!

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA



Strategic Focus Area
Personalized Health and Related Technologies

European Research Council
Established by the European Commission

SCAN ME

EPFL KU LEUVEN