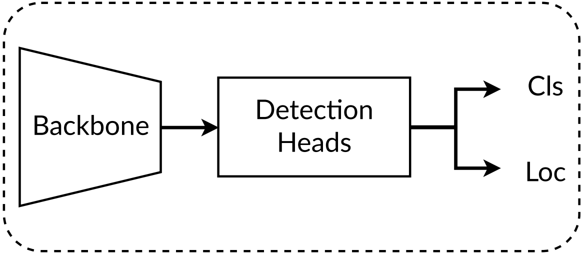
Proposal-Contrastive Pretraining for Object Detection from Fewer Data

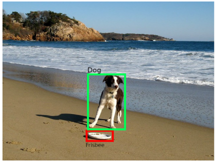
Quentin Bouniot [1,2] Romaric Audigier [1] Angelique Loesch [1] Amaury Habrard [2,3]

[1]CEA-List [2]Université Jean Monnet [3]Institut Universitaire de France

Object Detectors in a Nutshell

Object Detector

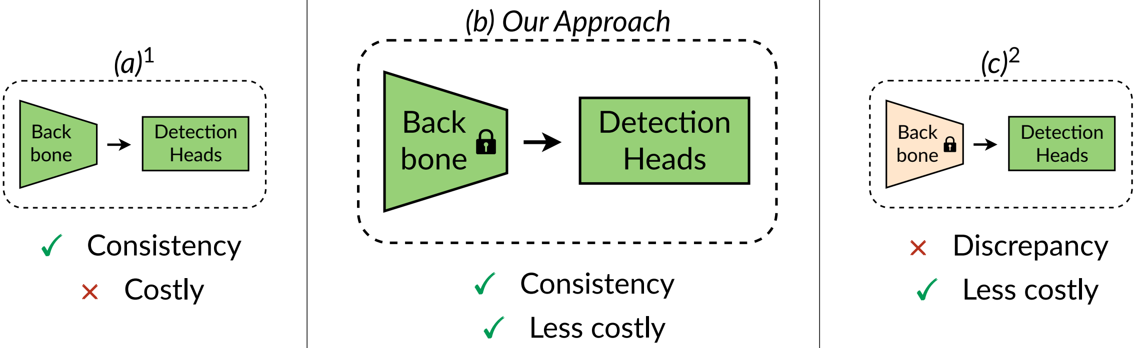




- Detectors composed of **backbone model** and **detection-specific heads.**
- Predict **class (Cls)** and **location (Loc)** for each objects in an image.

Pretraining in Object Detection

Overall Pretraining



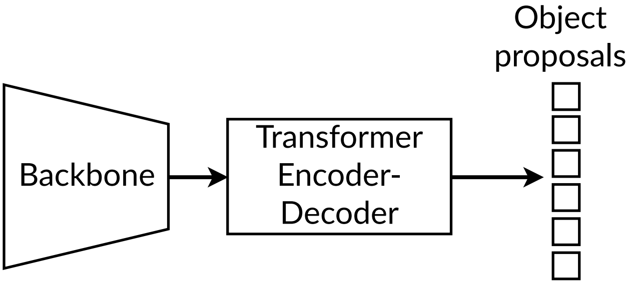
[1] Fangyun Wei et al. "Aligning pretraining for detection via object-level contrastive learning". In: *NeurIPS*. 2021

[2] Zhigang Dai et al. "Up-DETR: Unsupervised pre-training for object detection with transformers". In: *CVPR*. 2021; Amir Bar et al. "Detreg: Unsupervised pretraining with region priors for object detection". In: *CVPR*. 2022

Outline

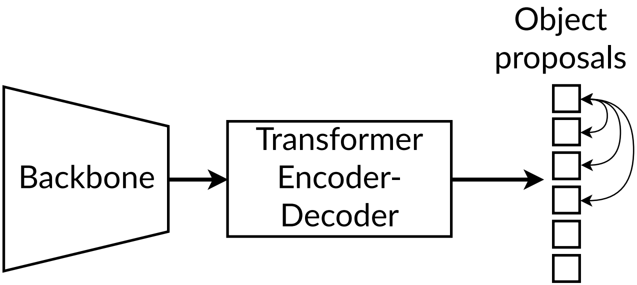
1. Context
2. Proposal Selection Contrast (ProSeCo)
   - Idea
   - Proposal-Contrastive Learning
   - Avoiding Collapse
3. Experimental Results
   - Comparison with state of the art
   - Ablation Studies
4. Conclusion

Transformer-based Detectors



- Transformer-based detectors generates N proposals $\gg k$ objects in images.

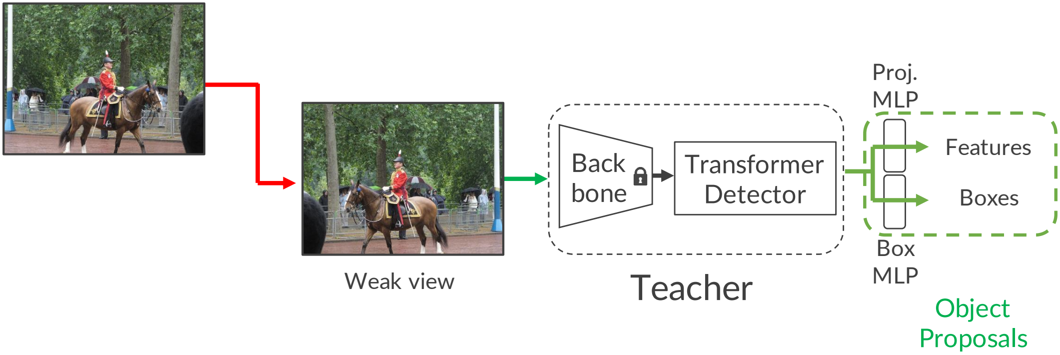
Transformer-based Detectors



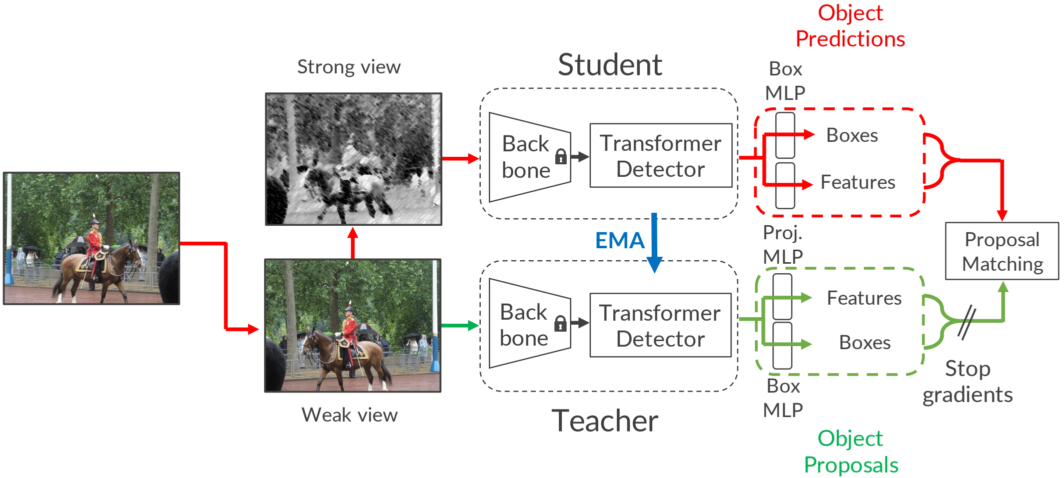
- Transformer-based detectors generates N proposals $\gg k$ objects in images.

Contribution: Contrastive learning **between** proposals.

Proposal-Contrastive Learning



Proposal-Contrastive Learning



- Object Proposals from **Teacher** are matched with Predictions from **Student**.

Proposal-Contrastive Learning

Unsupervised Proposal Matching

Object Proposals

$$\hat{\sigma}_i^{\text{prop}} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{j=1}^{N} \mathcal{L}_{\text{prop\_match}}\left(\mathbf{y}_{(i,j)}, \hat{\mathbf{y}}_{(i,\sigma(j))}\right)$$

Permutations of N elements

Object Predictions

- Proposal j found by the **teacher** associated to prediction $\hat{\sigma}_i^{\text{prop}}(j)$ of the **student.**

Proposal-Contrastive Learning

Unsupervised Proposal Matching

Object Proposals

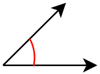
$$\hat{\sigma}_i^{\text{prop}} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{j=1}^{N} \mathcal{L}_{\text{prop\_match}}\left(\mathbf{y}_{(i,j)}, \hat{\mathbf{y}}_{(i,\sigma(j))}\right)$$

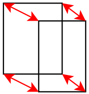
Permutations of N elements Object Predictions

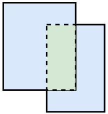
- Proposal j found by the **teacher** associated to prediction $\hat{\sigma}_i^{\text{prop}}(j)$ of the **student.**

Matching Cost $\mathcal{L}_{\text{prop\_match}}$ depends on:

- features similarity
- L_1 loss of box coordinates
- generalized IoU loss

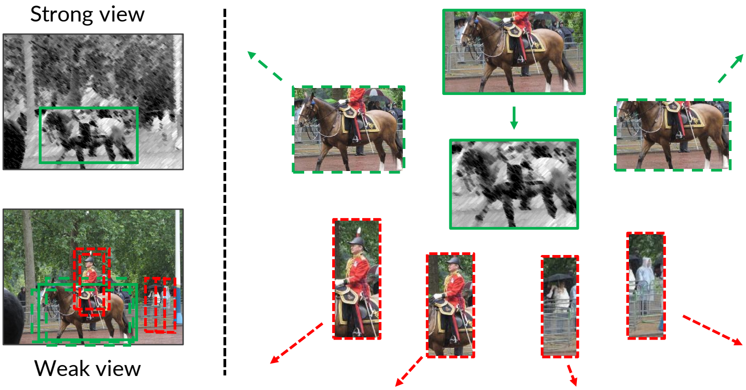






Proposal-Contrastive Learning

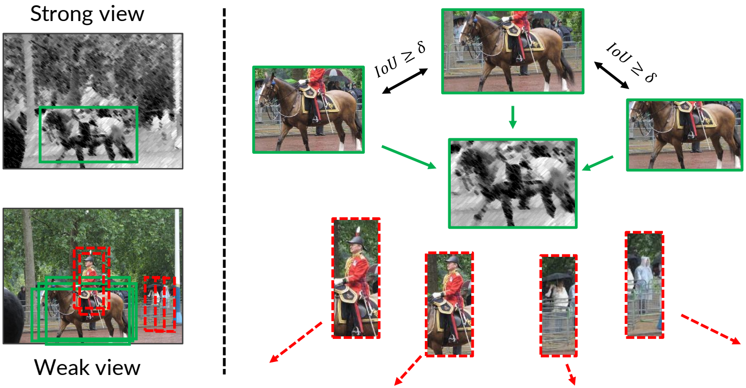
Naive way



✗ **Close** proposals considered as **negative** examples.

Proposal-Contrastive Learning

Localization-aware Contrastive loss



✓ **Overlapping** proposals are considered as **positive** examples.

Proposal-Contrastive Learning

Soft Contrastive Estimation (SCE) loss function[3]

$$\underbrace{\boxed{p'_{(in,jm)}}}_{\text{Relations between proposals}} = \frac{\mathbb{1}_{i \neq n} \mathbb{1}_{j \neq m} \exp(\mathbf{z}_{(i,j)} \cdot \mathbf{z}_{(n,m)} / \overbrace{\tau_t}^{\text{Temperature}})}{\sum_{k=1}^{N_b} \sum_{l=1}^{N} \mathbb{1}_{i \neq k} \mathbb{1}_{j \neq l} \exp(\mathbf{z}_{(i,j)} \cdot \underbrace{\mathbf{z}_{(k,l)}}_{\text{Features of Object Proposals}} / \tau_t)}$$

[3] Julien Denize et al. "Similarity contrastive estimation for self-supervised soft contrastive learning". In: *WACV*. 2023.

Proposal-Contrastive Learning

Soft Contrastive Estimation (SCE) loss function[3]

Relations between proposals

Temperature

$$\boxed{p'_{(in,jm)}} = \frac{\mathbb{1}_{i\neq n}\mathbb{1}_{j\neq m}\exp(\mathbf{z}_{(i,j)} \cdot \mathbf{z}_{(n,m)}/\tau_t)}{\sum_{k=1}^{N_b}\sum_{l=1}^{N}\mathbb{1}_{i\neq k}\mathbb{1}_{j\neq l}\exp(\mathbf{z}_{(i,j)} \cdot \mathbf{z}_{(k,l)}/\tau_t)}$$

Features of Object Proposals

Features of Object Predictions

$$\boxed{p''_{(in,jm)}} = \frac{\exp(\mathbf{z}_{(i,j)} \cdot \hat{\mathbf{z}}_{(n,m)}/\tau)}{\sum_{k=1}^{N_b}\sum_{l=1}^{N}\exp(\mathbf{z}_{(i,j)} \cdot \hat{\mathbf{z}}_{(k,l)}/\tau)}$$

Contrastive aspect between predictions and proposals

[3] Julien Denize et al. “Similarity contrastive estimation for self-supervised soft contrastive learning”. In: *WACV*. 2023.

Proposal-Contrastive Learning

Localization-aware similarity distribution

$$w^{\text{Loc}}_{(in,jm)} = \lambda_{\text{SCE}} \cdot \mathbb{1}_{i=n}\mathbb{1}_{IoU_i(j,m)\geq\delta} + (1-\lambda_{\text{SCE}}) \cdot p'_{(in,jm)}$$

IoU between proposals in same image above threshold

Proposal-Contrastive Learning

Localization-aware similarity distribution

$$w^{\text{Loc}}_{(in,jm)} = \lambda_{\text{SCE}} \cdot \mathbb{1}_{i=n} \mathbb{1}_{IoU_i(j,m) \geq \delta} + (1 - \lambda_{\text{SCE}}) \cdot p'_{(in,jm)}$$

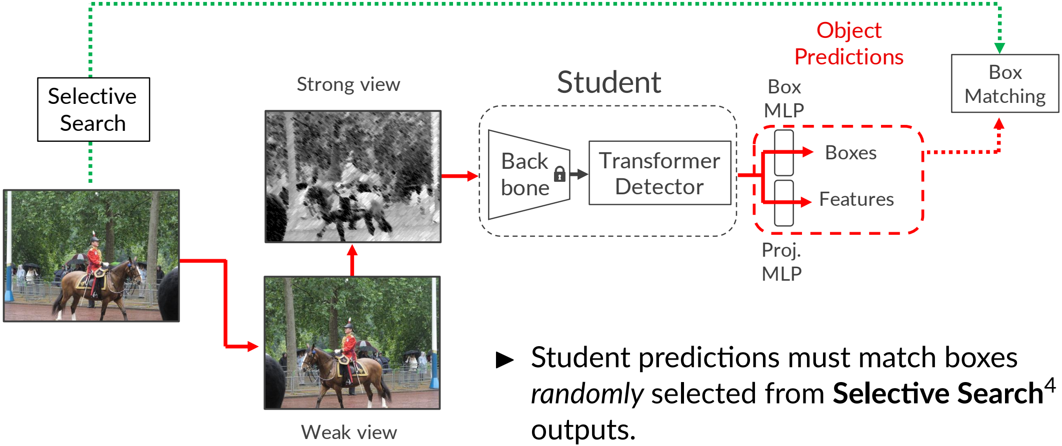
IoU between proposals in same image above threshold

Localized SCE (LocSCE) function

$$\mathcal{L}_{\text{LocSCE}}(\mathbf{y}, \hat{\mathbf{y}}, \hat{\sigma}^{\text{prop}}) = -\frac{1}{N_b N} \sum_{i=1}^{N_b} \sum_{n=1}^{N_b} \sum_{j=1}^{N} \sum_{m=1}^{N} w^{\text{Loc}}_{(in,jm)} \log(p''_{(in,j\hat{\sigma}^{\text{prop}}_n(m))})$$

Effective batch size

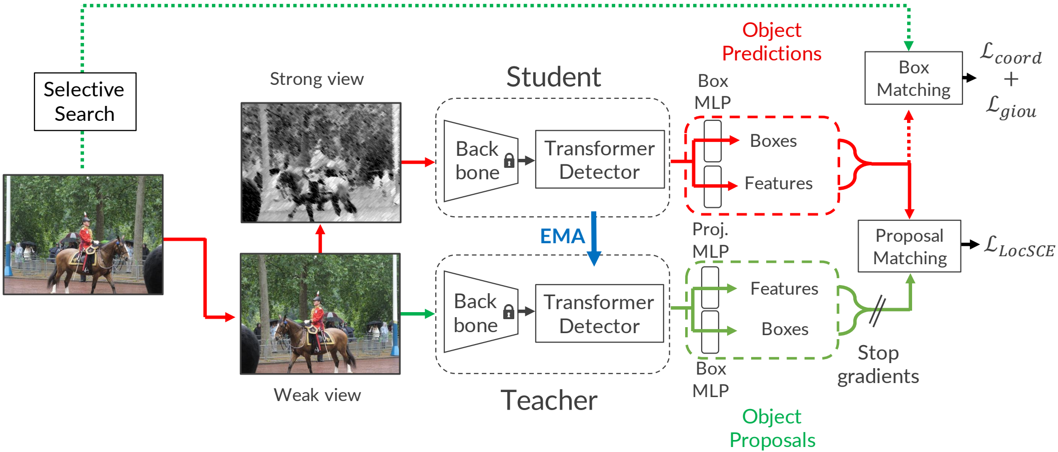
Avoiding Collapse



- Student predictions must match boxes *randomly* selected from **Selective Search**[4] outputs.

[4] Jasper RR Uijlings et al. “Selective search for object recognition”. In: *IJCV*. 2013.

Proposal Selection Contrast (ProSeCo)



- Full pretraining procedure with both **contrastive** and **localization** learning.

Pretraining on ImageNet, finetuning on Mini-COCO

Pretraining	Detector	Mini-COCO 1% (1.2k)	Mini-COCO 5% (5.9k)	Mini-COCO 10% (11.8k)
Supervised	Def. DETR	13.0	23.6	28.6
SwAV[5]	Def. DETR	13.3	24.5	29.5
SCRL[6]	Def. DETR	16.4	26.2	30.6
DETReg[7]	Def. DETR	15.9	26.1	30.9
Supervised	Mask R-CNN	–	19.4	24.7
SoCo*[8]	Mask R-CNN	–	26.8	31.1
ProSeCo (Ours)	Def. DETR	**18.0**	**28.8**	**32.8**

[5] Mathilde Caron et al. "Unsupervised learning of visual features by contrasting cluster assignments". In: *NeurIPS*. 2020.

[6] Byungseok Roh et al. "Spatially consistent representation learning". In: *CVPR*. 2021.

[7] Amir Bar et al. "Detreg: Unsupervised pretraining with region priors for object detection". In: *CVPR*. 2022.

[8] Fangyun Wei et al. "Aligning pretraining for detection via object-level contrastive learning". In: *NeurIPS*. 2021.

Finetuning on other datasets

Method	FSOD-test	FSOD-train	PASCAL VOC	Mini-VOC	
	100% (11k)	100% (42k)	100% (16k)	5% (0.8k)	10% (1.6k)
Supervised	39.3	42.6	59.5	33.9	40.8
DETReg[9]	43.2	43.3	63.5	43.1	48.2
ProSeCo (Ours)	**46.6**	**47.2**	**65.1**	**46.1**	**51.3**

✓ **ProSeCo** improves over SOTA on all datasets considered, with **various amount** of labeled data.

[9] Amir Bar et al. “Detreg: Unsupervised pretraining with region priors for object detection”. In: *CVPR*. 2022.

Ablation Studies

Pretraining	Dataset	mAP
ProSeCo w/ SwAV	COCO	27.4
ProSeCo w/ SwAV	IN	27.8
DETReg w/ SCRL	IN	28.0
ProSeCo w/ SCRL	IN	**28.8**

Loss	δ	mAP
SCE	1.0	26.1
LocSCE (Ours)	0.2	27.0
LocSCE (Ours)	0.7	27.1
LocSCE (Ours)	0.5	**27.8**

- ▶ **Dataset diversity** more important than closeness to downstream task
- ✓ **Consistency** in the features improves performance
- ✓ **Location of proposals** helps for introducing **easy positives** for contrastive learning

Take Home Message

We propose ProSeCo, a Proposal-Contrastive Pretraining strategy for Object Detection with Transformers.

- ✓ Leverage high number of Object Proposals for **Proposal-Contrastive Learning.**
- ✓ Our **ProSeCo improves performance** when training with limited labeled data.
- ✓ **Consistency** with object-level features is important for Object Detection.
- ✓ **Location information** helps for Proposal-Contrastive learning.

Thank You !

Do not hesitate to contact us for question !

Bouniot et al., “Proposal-Contrastive Pretraining for Object Detection from Fewer Data”

References I

- Fangyun Wei et al. “Aligning pretraining for detection via object-level contrastive learning”. In: *NeurIPS*. 2021.
- Zhigang Dai et al. “Up-DETR: Unsupervised pre-training for object detection with transformers”. In: *CVPR*. 2021.
- Amir Bar et al. “Detreg: Unsupervised pretraining with region priors for object detection”. In: *CVPR*. 2022.
- Julien Denize et al. “Similarity contrastive estimation for self-supervised soft contrastive learning”. In: *WACV*. 2023.
- Jasper RR Uijlings et al. “Selective search for object recognition”. In: *IJCV*. 2013.
- Mathilde Caron et al. “Unsupervised learning of visual features by contrasting cluster assignments”. In: *NeurIPS*. 2020.

References II

- Byungseok Roh et al. “Spatially consistent representation learning”. In: *CVPR*. 2021.
- Quentin Bouniot et al. “Proposal-Contrastive Pretraining for Object Detection from Fewer Data”. In: *ICLR*. 2023.