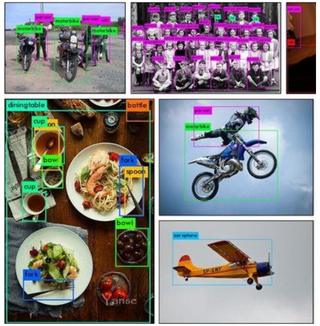
Object localization *(almost)* for **free** harnessing **self-supervised features**

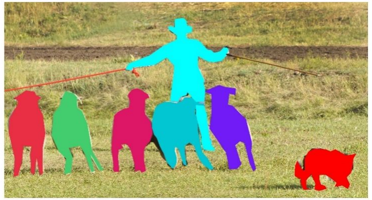
Oriane Siméoni
valeo.ai

Object localization

Classic benchmarks
Closed vocabulary setup



COCO [Lin et al. ECCV'14]



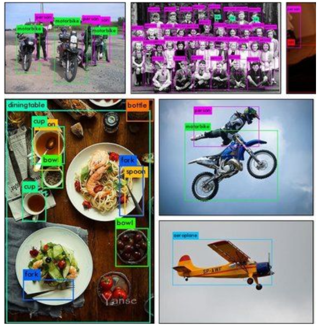
Object detection

Instance segmentation

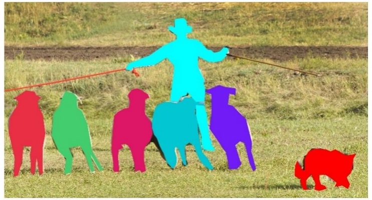
But, require
- the definition of a **finite** set of **classes**
  ➔ **limited** when we consider our world
- train a model in fully-supervised fashion
  ➔ a lot of **annotation** ✍️

Object localization

Classic benchmarks
Closed vocabulary setup



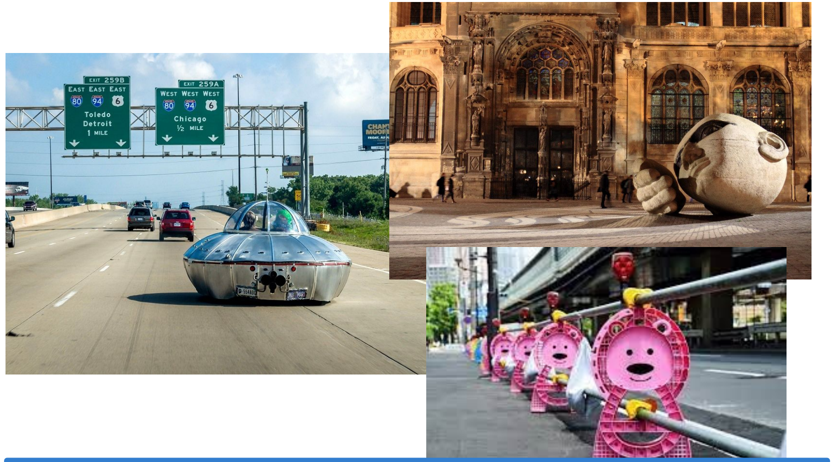
COCO [Lin et al. ECCV'14]



Object detection

Instance segmentation

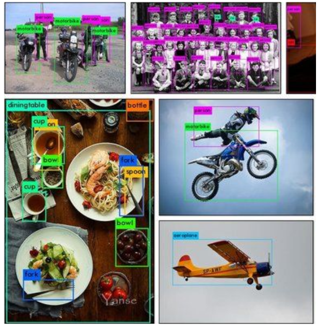
But, require
- the definition of a **finite** set of **classes**
  → **limited** when we consider our world
- train a model in fully-supervised fashion
  → a lot of **annotation** ✍️



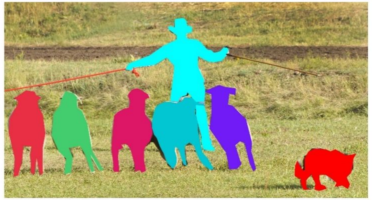
How to find **objects** *without knowing anything about them* ?

Object localization

Classic benchmarks
Closed vocabulary setup



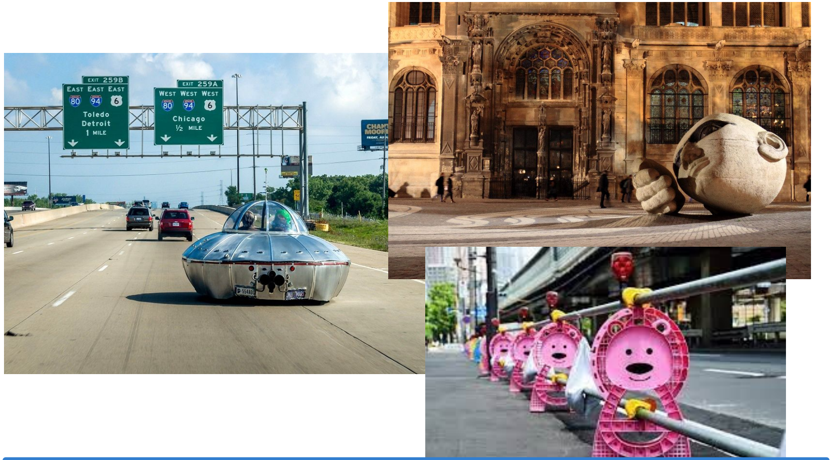
COCO [Lin et al. ECCV'14]



Object detection

Instance segmentation

But, require
- the definition of a **finite** set of **classes**
  → **limited** when we consider our world
- train a model in fully-supervised fashion
  → a lot of **annotation** ✍️



How to find **objects** *without knowing anything about them* ?

Segment anything [Kirillov et al., ICCV'23]

Without human-made supervision ?

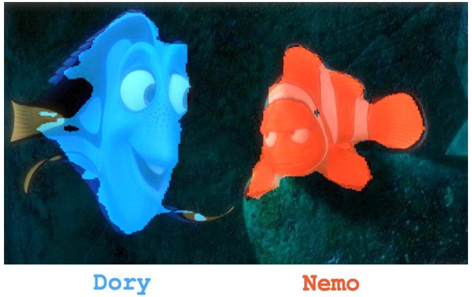
Unsupervised object localization

Goal

- Discovering objects in a 2d image
- No **information/supervision** about objects available

Unsupervised **object discovery**

Foreground/background segmentation



Zero-shot open-vocabulary semantic segmentation

Object localization *(almost)* for free harnessing **self-supervised features**

Object localization *(almost)* for free harnessing **self-supervised features**

Why self-supervised features ?

Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron[1,2] Hugo Touvron[1,3] Ishan Misra[1] Hervé Jegou[1]
Julien Mairal[2] Piotr Bojanowski[1] Armand Joulin[1]

[1] Facebook AI Research [2] Inria* [3] Sorbonne University

Figure 1: **Self-attention from a Vision Transformer with 8×8 patches trained with no supervision.** We look at the self-attention of the `[CLS]` token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

DINO [Caron et al. ICCV'21]

- **ViT models** pre-trained in a **self-supervised** manner have **good localization properties**
- Trained on **unlabelled** data with a **proxy task**

Are we done ?

Why self-supervised features ?

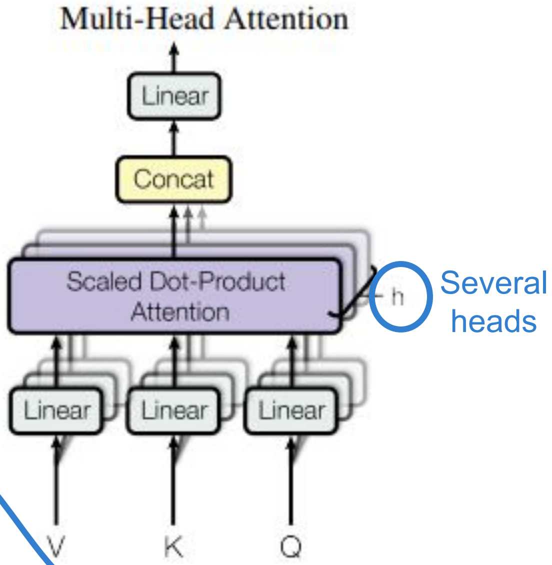
Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron[1,2] Hugo Touvron[1,3] Ishan Misra[1] Hervé Jegou[1]
Julien Mairal[2] Piotr Bojanowski[1] Armand Joulin[1]

[1] Facebook AI Research [2] Inria* [3] Sorbonne University

Figure 1: **Self-attention from a Vision Transformer with** 8×8 **patches trained with no supervision.** We look at the self-attention of the [CLS] token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

DINO [Caron et al. ICCV'21]



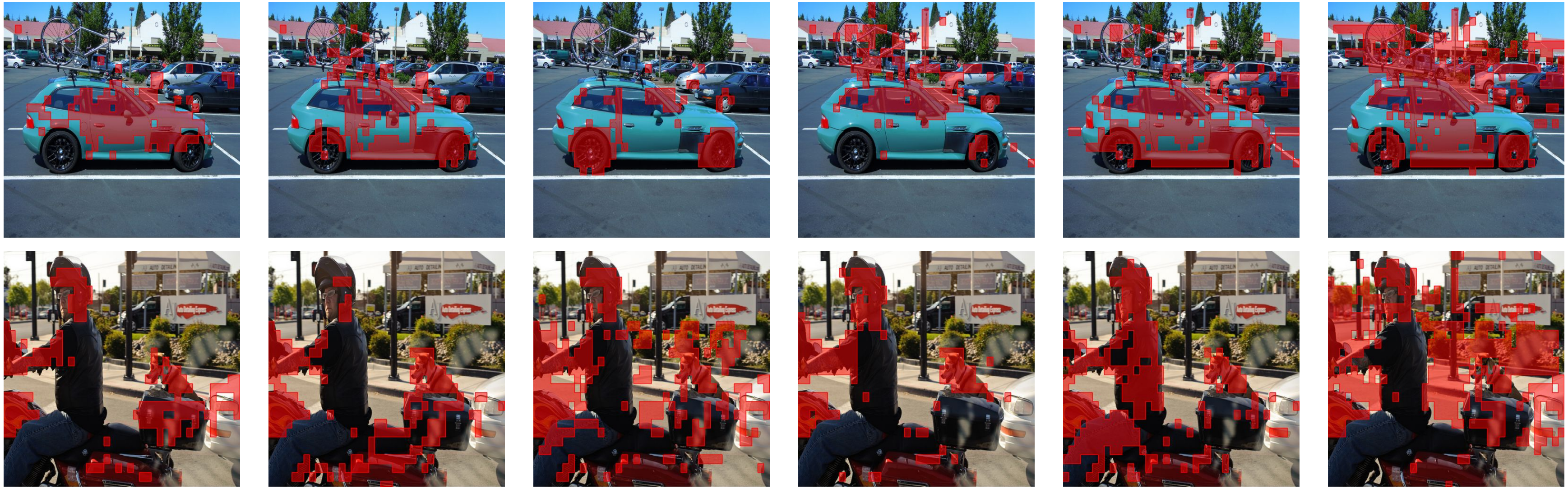
$$\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Attention is all you need [Vaswani et al. NeurIPS'17]

Self-attention maps

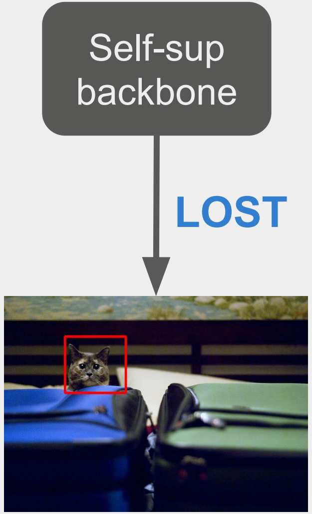
- The **6 heads** attend to **different parts** of an image
- Without supervision hard to distinguish **what is important** and is an object

[CLS] self-attention maps



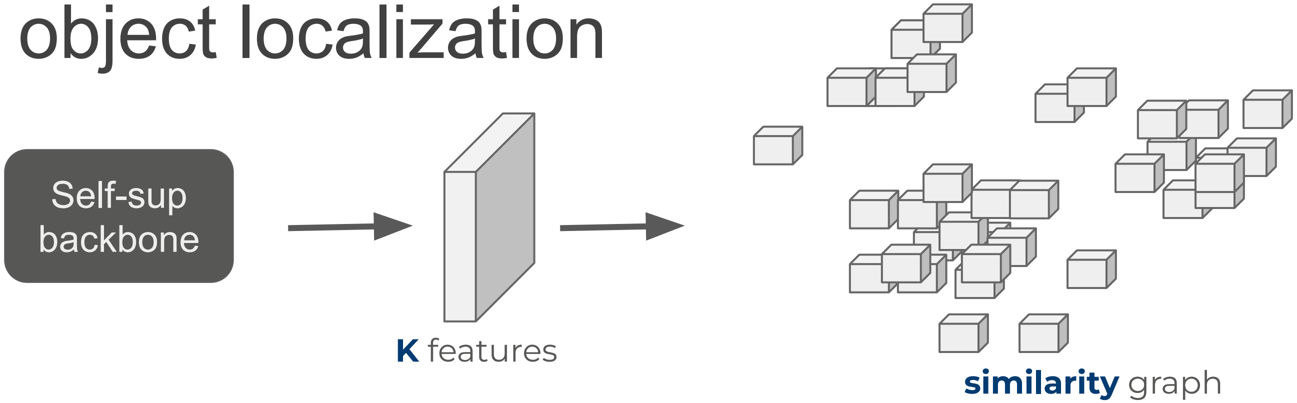
Head 1 Head 2 Head 3 Head 4 Head 5 Head 6

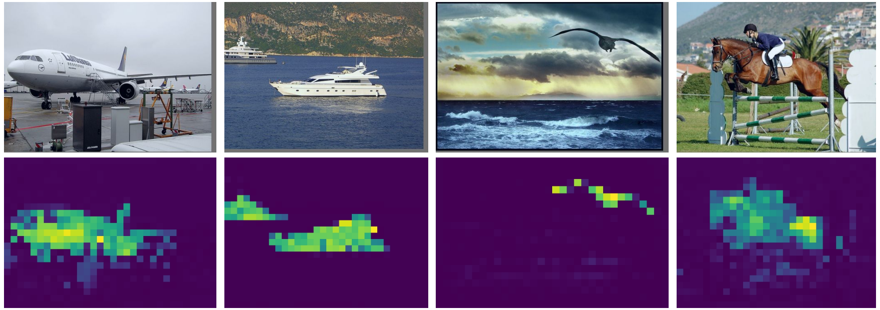
Unsupervised object localization



Single object
localization

Single object localization





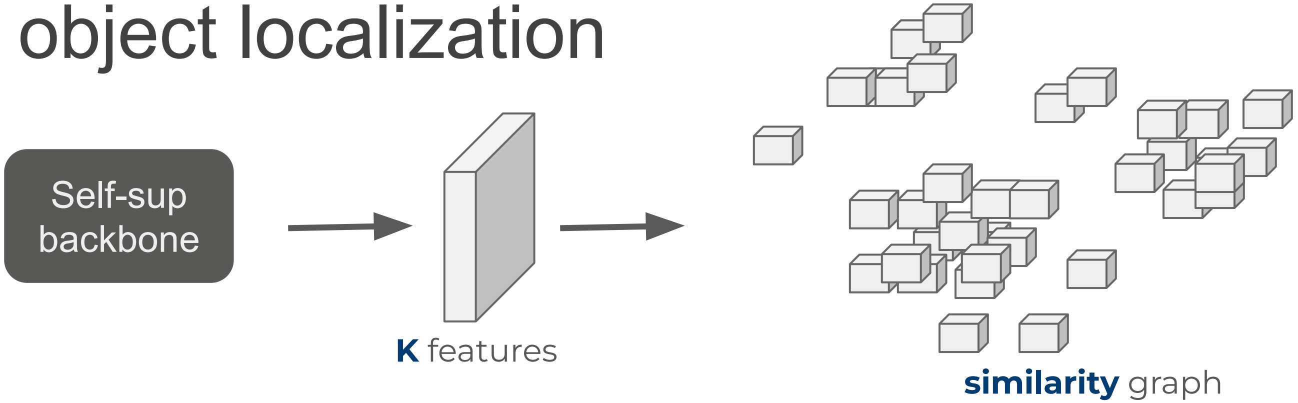
Patch **degrees**
Low to **high**

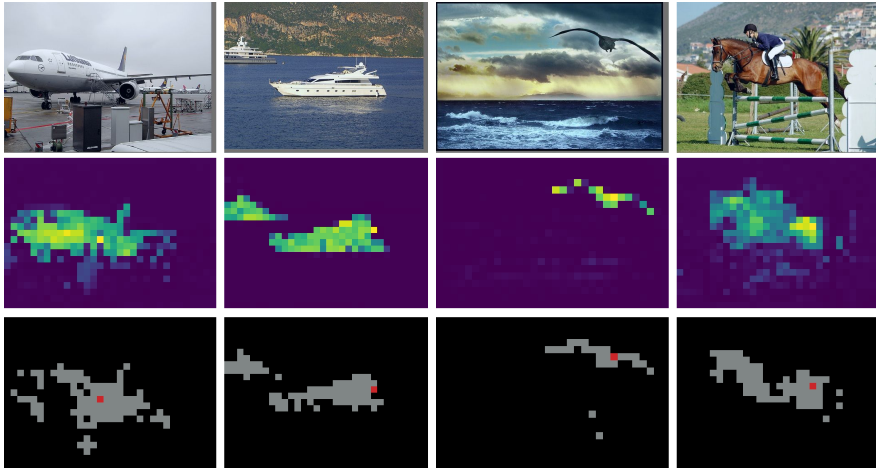
LOST [Siméoni et al. BMVC'21]

- Patches of **foreground** are **less** correlated than those of background

Siméoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC'21

Single object localization





Patch **degrees**
Low to **high**

initial **seed**

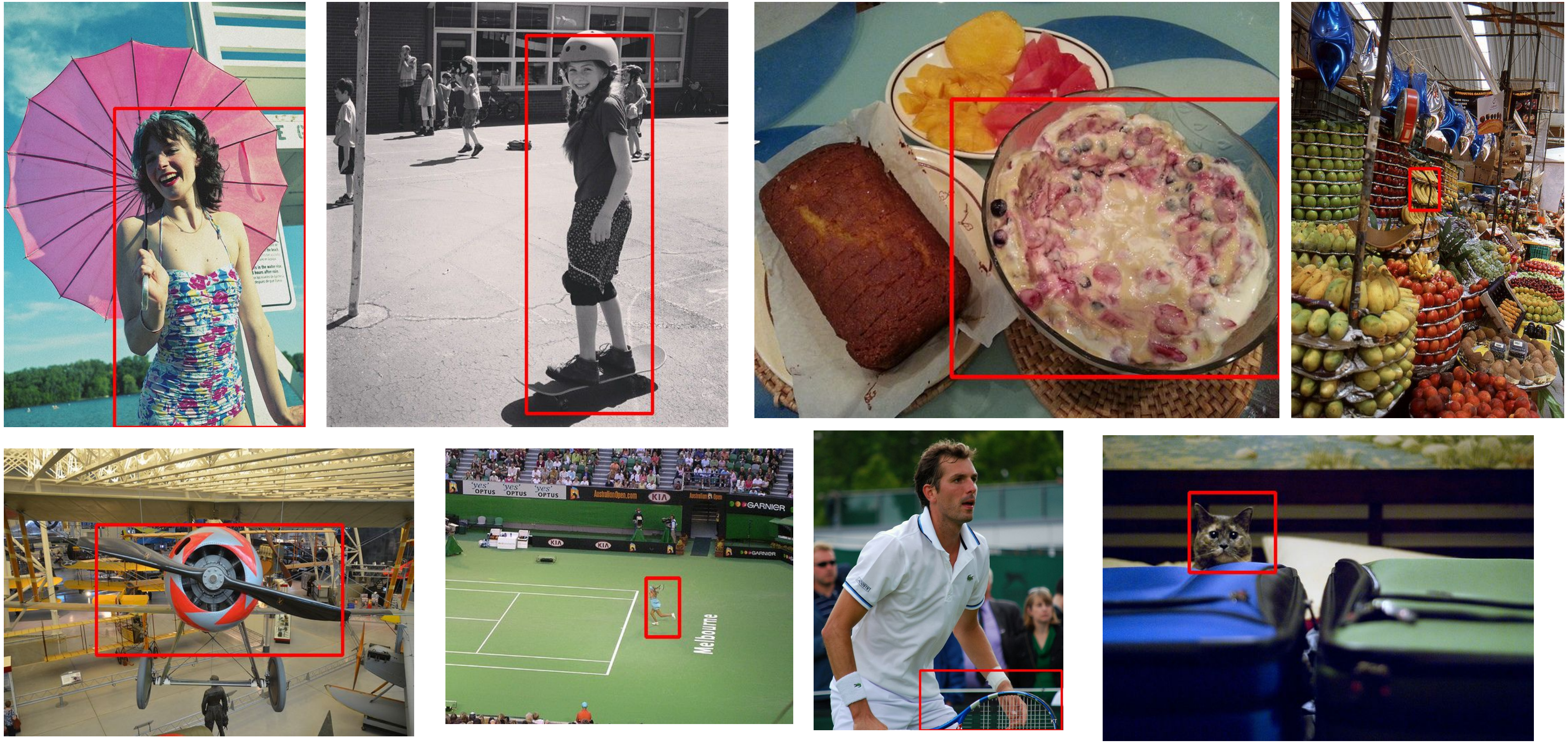
similar patches

LOST [Siméoni et al. BMVC'21]

- Patches of **foreground** are **less** correlated than those of background
- **Object =** patch with the **lowest degree** & connected **correlated patches**
- Additional expansion step

Siméoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC'21

Qualitative results



Siméoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC'21

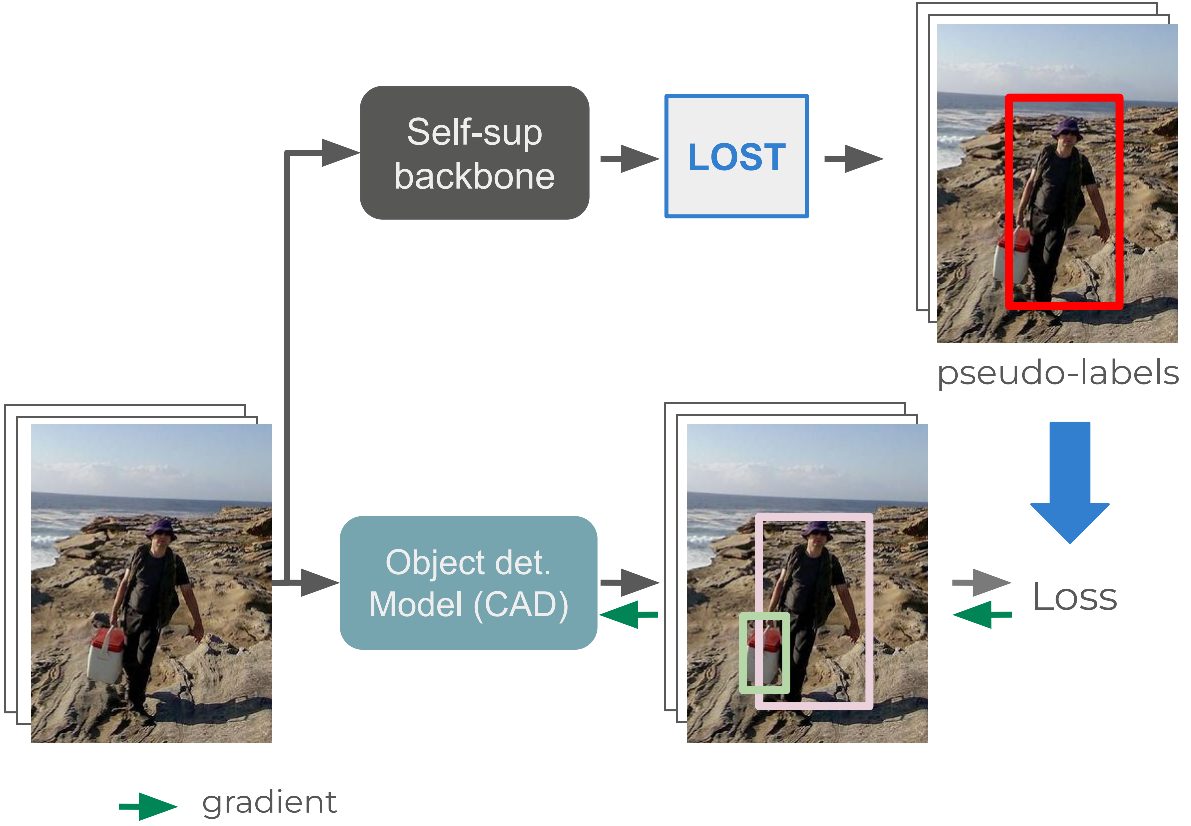
Qualitative results

Method	VOC07_trainval	VOC12_trainval	COCO_20k
Selective Search [65]	18.8	20.9	16.0
EdgeBoxes [84]	31.1	31.6	28.8
Kim *et al*. [38]	43.9	46.4	35.1
Zhang *et al*. [80]	46.2	50.5	34.8
DDT+ [72]	50.2	53.1	38.2
rOSD [68]	54.5	55.3	48.5
LOD [69]	53.6	55.1	48.5
DINO-seg (w. ViT-S/16)	45.8	46.2	42.1
LOST (ours)	**61.9**	**64.0**	**50.7**
	+ 7.4	**+ 8.7**	**+ 2.2**

Corloc metric = % of correct boxes
→ a predicted box is correct if has **IoU > 0.5 with one of gt boxes**

Siméoni et al., Localizing Objects with Self-Supervised Transformers and no Labels, BMVC’21

Improving results through learning



LOST+CAD [Siméoni et al. BMVC'21]

- Train a **class-agnostic** object **detector** (eg Faster R-CNN)
- Use **LOST** predictions as **pseudo ground-truth**

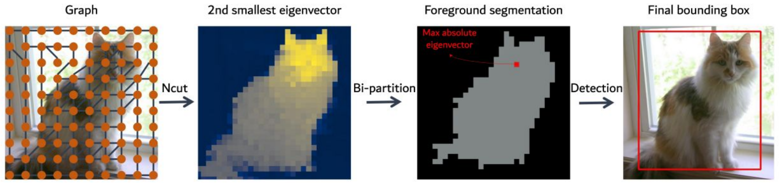
→ **Regularization** & predicts several boxes

- **+7pts** corloc
- more than one prediction per image

More powerful algorithms

TokenCut [Wang et al. CVPR'22], **Deep Spectral Methods** [Melas-Kyriazi et al. CVPR'22], **SelfMask** [Shi et al. CVPRW'22]

- Same features, *similar graph*
- Solve a normalized **graph-cut** problem with **spectral clustering** → improved localization

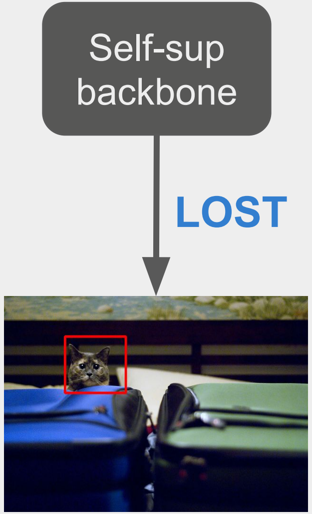


CutLer [Wang et al. CVPR'23]

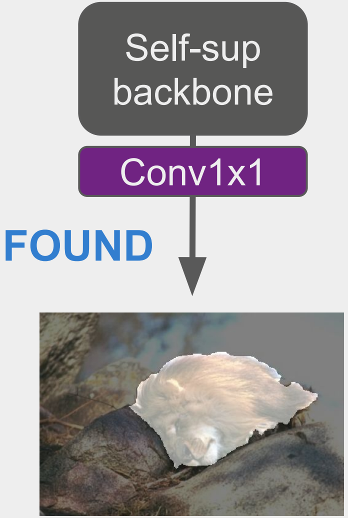
- Detect several objects
- Remove **already** discovered nodes from the graph and **repeat** the operation
- Also propose an **improved training** scheme (propose to repeat **3x** a training → increase number of detected boxes)

More details/discussion in our recent **survey**:
Unsupervised Object Localization in the Era of Self-Supervised ViTs: A Survey, Siméoni et al., arxiv'23

Unsupervised object localization

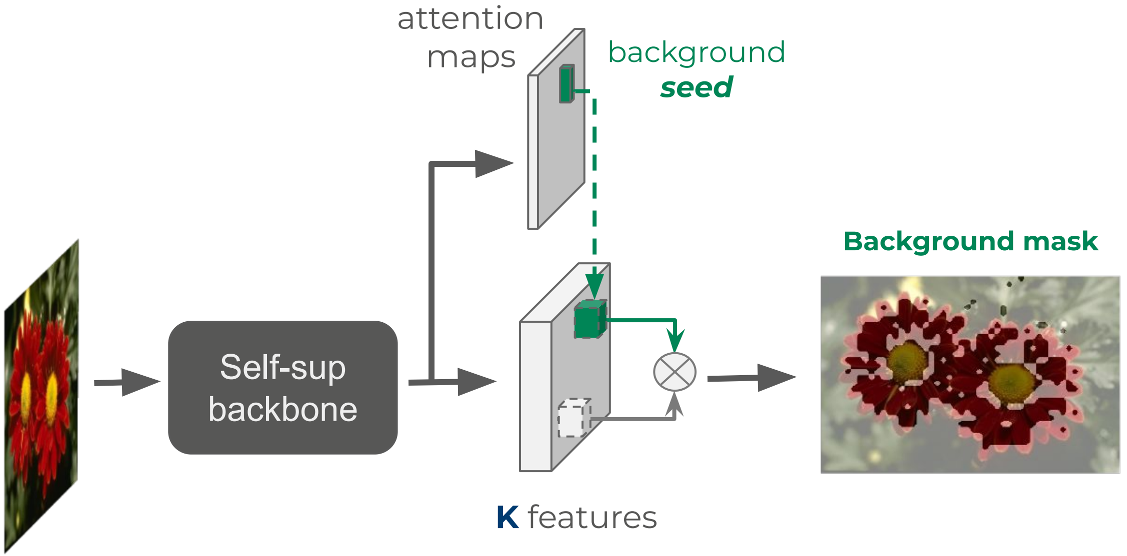


Single object
localization



Foreground/background
segmentation

Discovering the background to highlight objects



FOUND [Siméoni et al. CVPR'23]

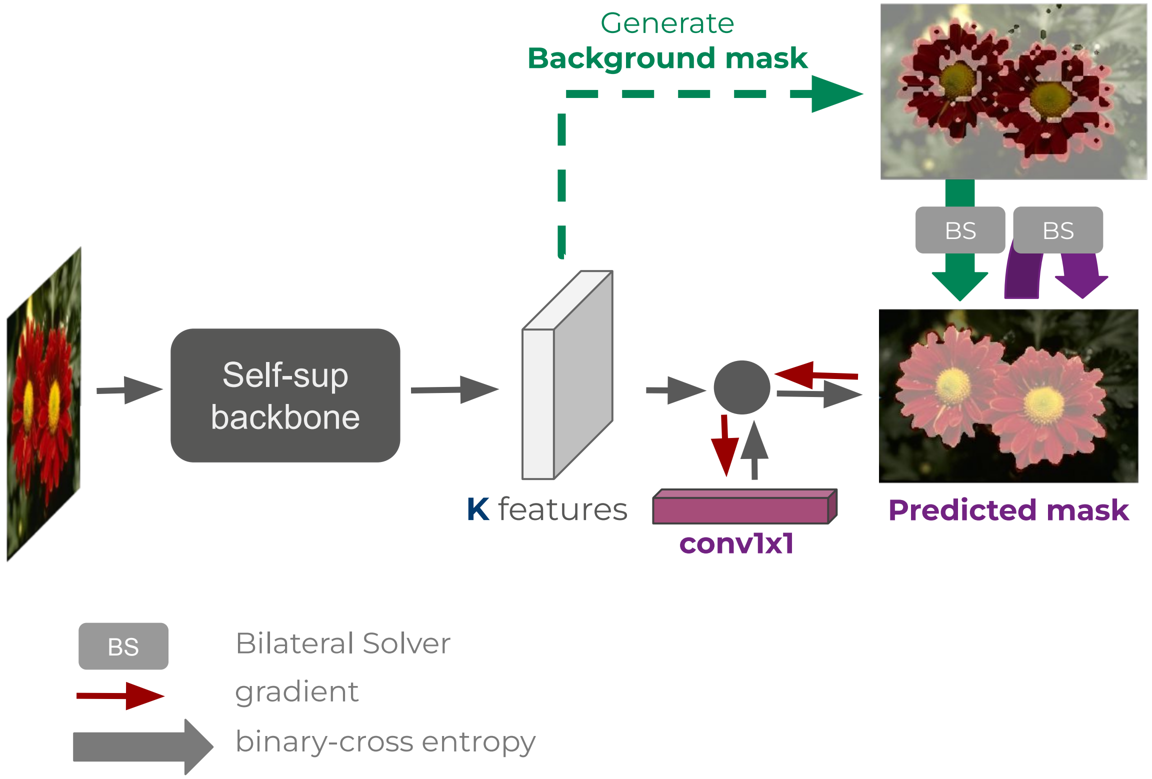
- Look for the **background** instead of objects
- No hypotheses about objects

Background mask

- Seed = patch receiving **least attention**
- Mask = **correlated** patches to seed

Siméoni et al., Unsupervised Object Localization: Observing the Background to Discover Objects, CVPR'23

Self-supervised refinement



FOUND [Siméoni et al. CVPR'23]

- Look for the **background** instead of objects
- No hypotheses about objects

FOUND = a single conv 1x1

- Trained using background masks as **pseudo-labels**
- **Bilateral Solver** used to refine masks along pixel edges

Siméoni et al., Unsupervised Object Localization: Observing the Background to Discover Objects, CVPR'23

Out-of-domain predictions (*no post-processing)*

FOUND [Siméoni et al. CVPR'23]

- **Single conv 1x1** layer trained with pseudo-labels
- Trained for 500 it. on DUTS-TR (10k images) [Wang et al, CVPR17] ~ **2h** with a **single GPU**
- Inference at **80 FPS** 🚀 on a V100

Siméoni et al., Unsupervised Object Localization: Observing the Background to Discover Objects, CVPR'23

Quantitative results

		DUT-OMRON [65]			DUTS-TE [55]			ECSSD [43]		
Method	Learning	Acc	IoU	max F_β	Acc	IoU	max F_β	Acc	IoU	max F_β
— ***Without post-processing bilateral solver*** —										
HS [63]		.843	.433	.561	.826	.369	.504	.847	.508	.673
wCtr [73]		838	.416	.541	.835	.392	.522	.862	.517	.684
WSC [28]		.865	.387	.523	.862	.384	.528	.852	.498	.683
DeepUSPS [36]		.779	.305	.414	.773	.305	.425	.795	.440	.584
BigBiGAN [54]		.856	.453	.549	.878	.498	.608	.899	.672	.782
E-BigBiGAN [54]		.860	.464	.563	.882	.511	.624	.906	.684	.797
Melas-Kyriazi et al. [33]		.883	.509	—	.893	.528	-	.915	.713	—
LOST [45] ViT-S/16 [6]		.797	.410	.473	.871	.518	.611	.895	.654	.758
DSS [34] [59]		—	.567	—	—	.514	—	—	.733	—
TokenCut [59] ViT-S/16 [6]		.880	.533	.600	.903	.576	.672	.918	.712	.803
SelfMask [44]	✓	.901	.582	—	.923	.626	—	.944	.781	—
FOUND — single ViT-S/8 [6]	✓	**.920**	**.586**	**.683**	**.939**	.637	**.733**	.912	.793	.946
FOUND — multi ViT-S/8 [6]	✓	.912	.578	.663	.938	**.645**	.715	**.949**	**.807**	**.955**

- **80 FPS** vs
  60 FPS (LOST)
  13 FPS (SelfMask, FreeSolo)
- **<1000** learned **parameters**

Unsupervised object localization

Self-sup backbone

LOST

Single object localization

Self-sup backbone

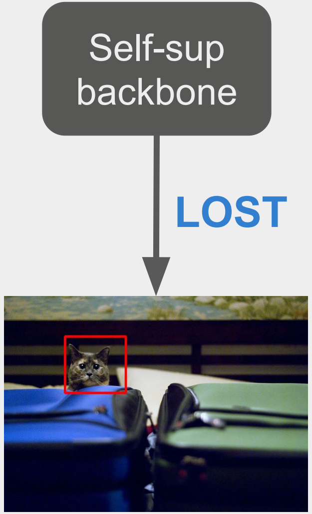
Conv1x1

FOUND

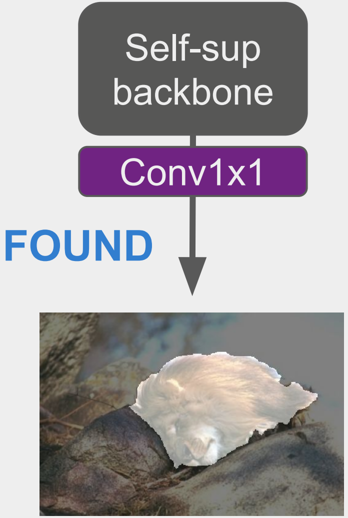
Foreground/background segmentation

What about classes ?

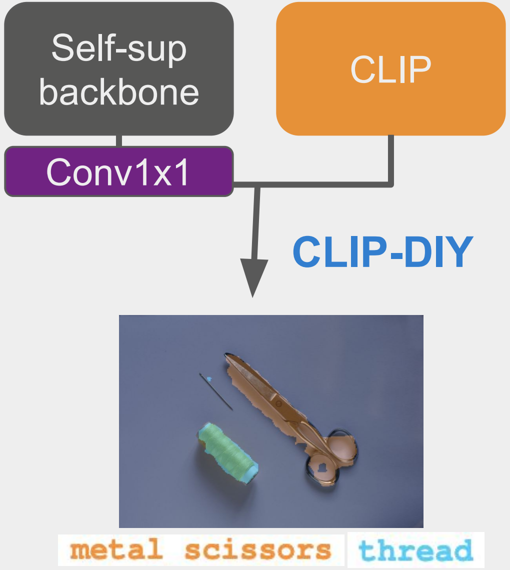
Unsupervised object localization



Single object
localization

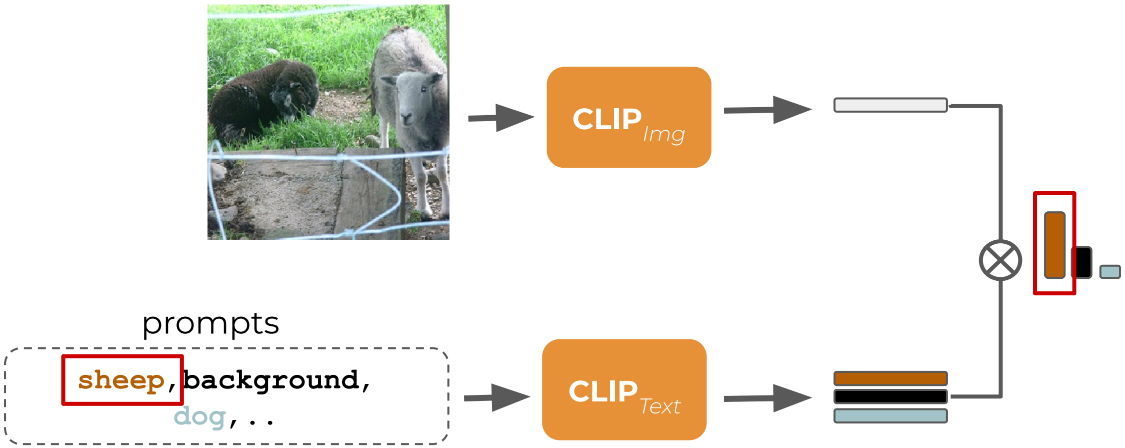


Foreground/background
segmentation



Open-vocabulary
segmentation

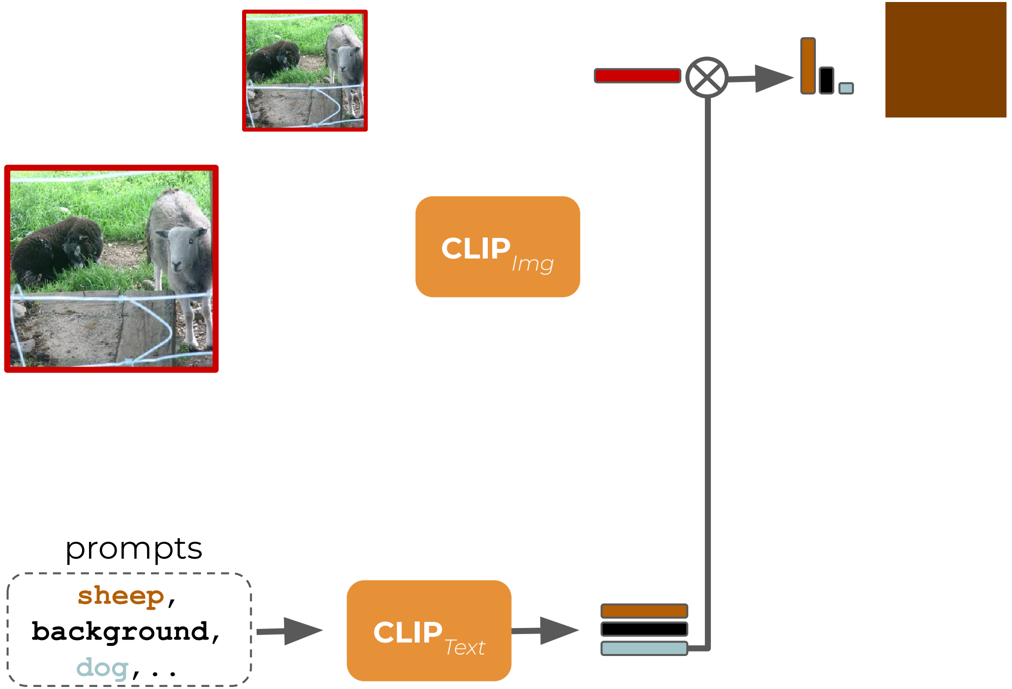
Open-vocabulary text/**global** image alignment



- Powerful VLMs which **align text and images**
- **CLIP** [Ilharco et al. 21] trained with a **global** objective to **align** *text to images*
  → good zero-shot classification
- **Densifying** CLIP is a hard task: require training (**TCL** [Cha et al. CVPR'23], **CLIPpy** [Ranasinghe et al. ICCV'23]), very noisy (**MaskCLIP** [Zhou et al. ECCV'22]), extra annotation, etc..

Wysoczanska et al., CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free, WACV'24

CLIP densification

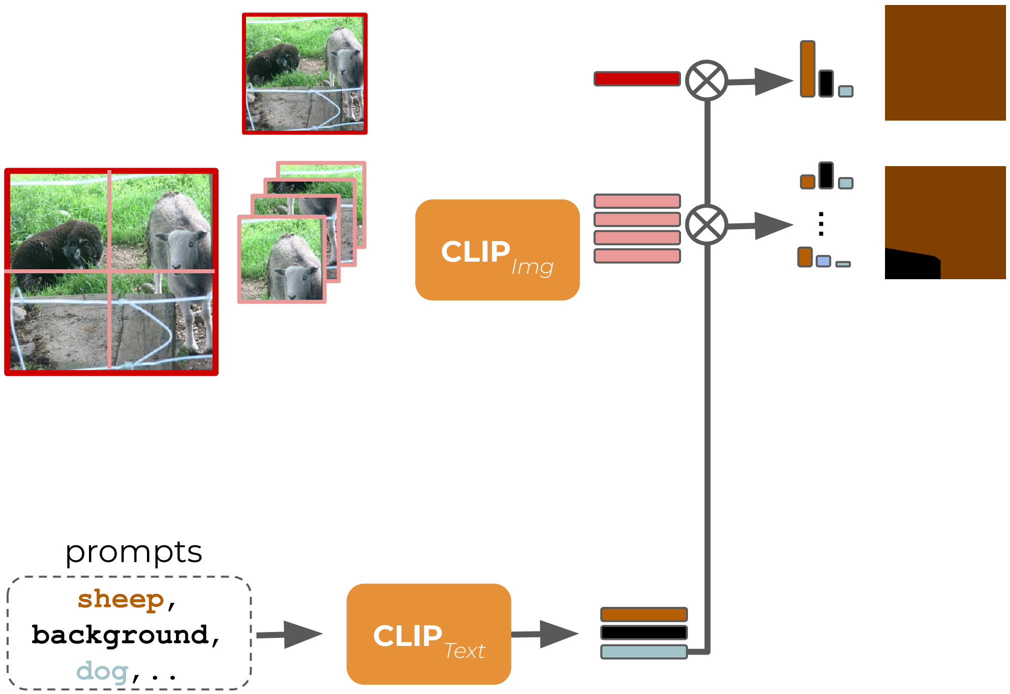


CLIP-DIY [Wysoczanska et al. WACV'24]

- **Idea:** leverage CLIP good **global** properties
- Perform prompt assignment is a **sliding window** fashion

Wysoczanska et al., CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free, WACV'24

CLIP densification

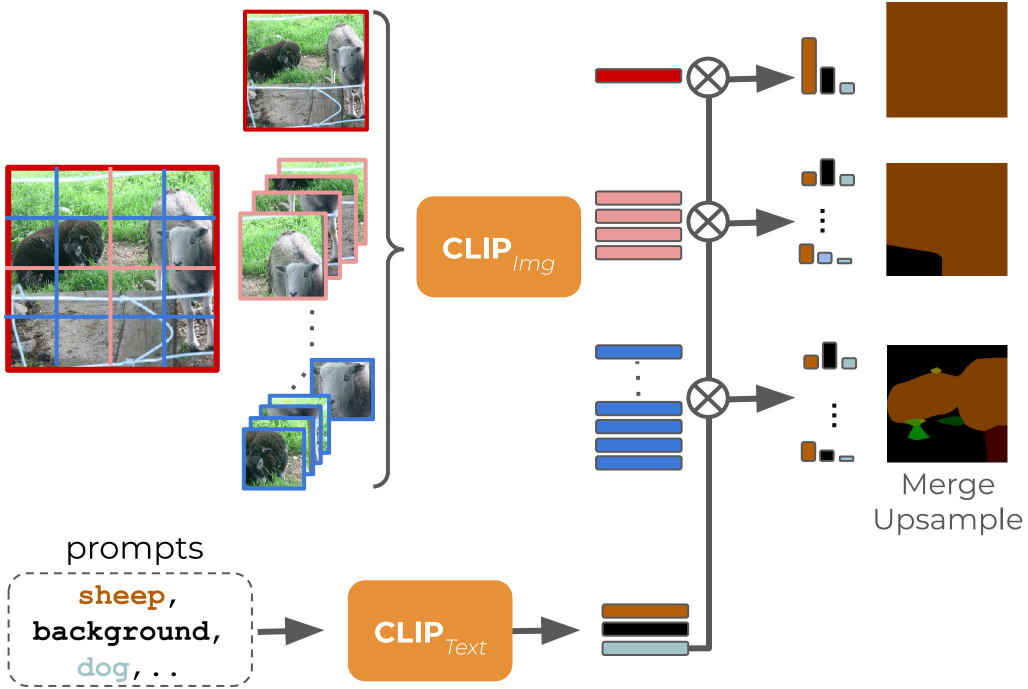


CLIP-DIY [Wysoczanska et al. WACV'24]

- **Idea:** leverage CLIP good **global** properties
- Perform prompt assignment is a **sliding window** fashion

Wysoczanska et al., CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free, WACV'24

CLIP densification

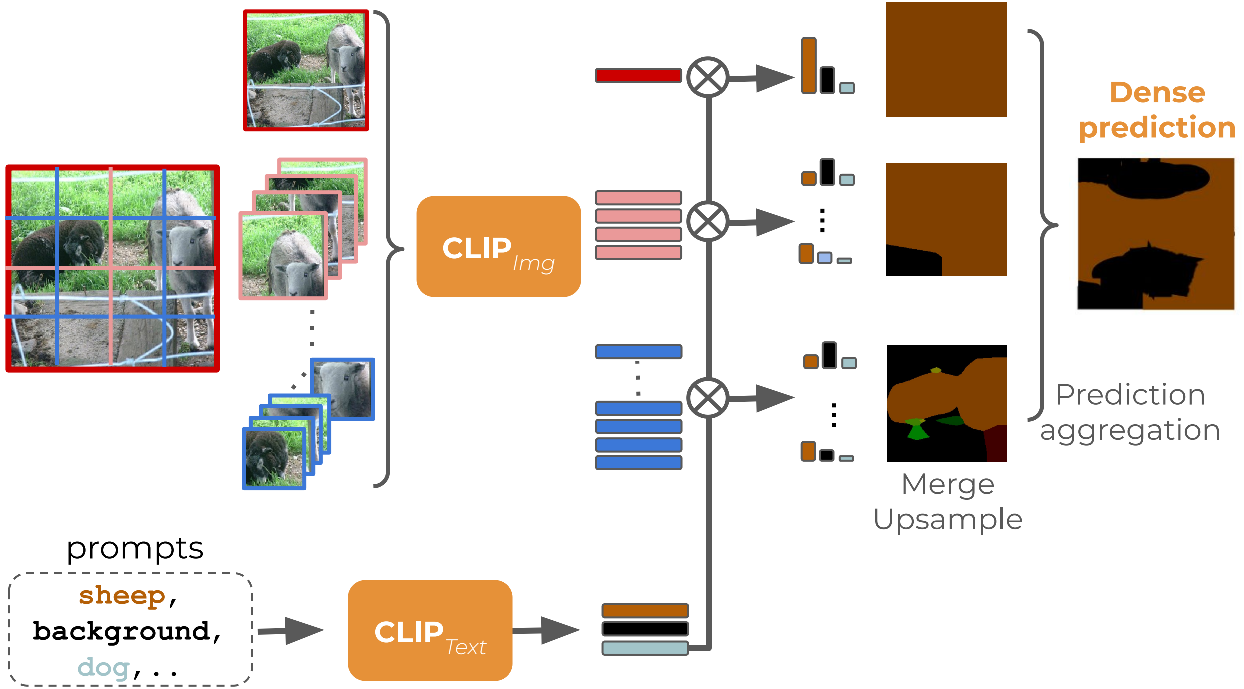


CLIP-DIY [Wysoczanska et al. WACV'24]

- **Idea:** leverage CLIP good **global** properties
- Perform prompt assignment is a **sliding window** fashion

Wysoczanska et al., CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free, WACV'24

CLIP densification

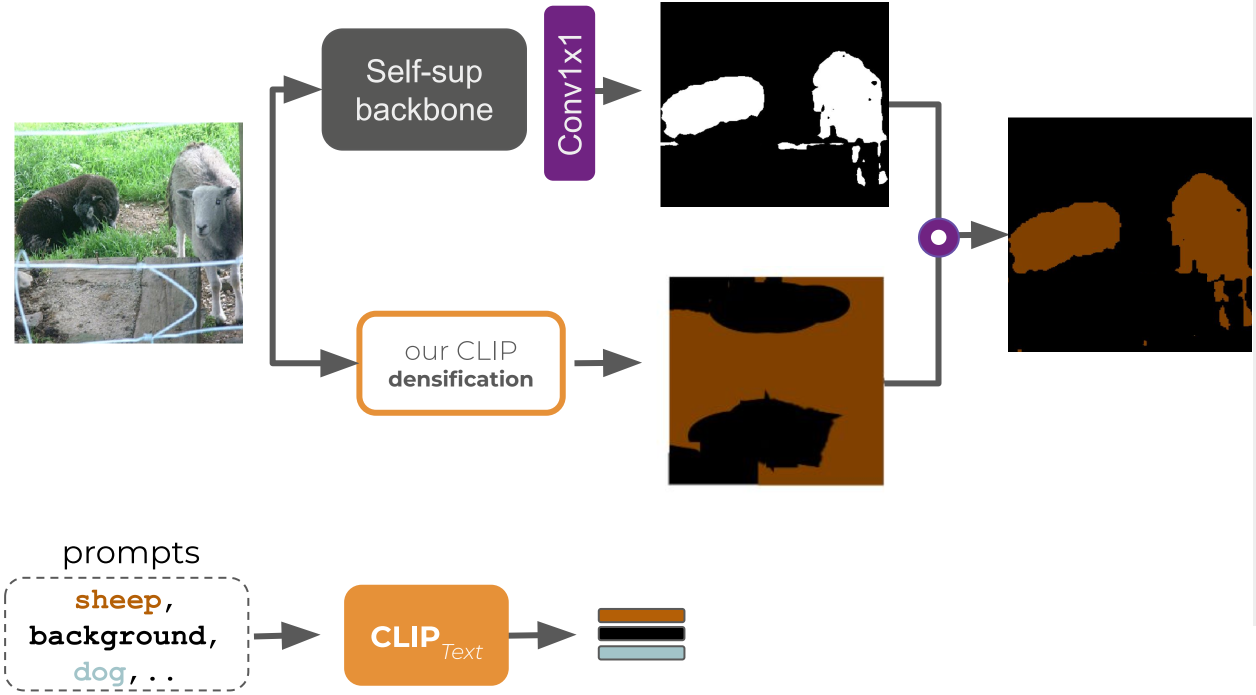


CLIP-DIY [Wysoczanska et al. WACV’24]

- **Idea:** leverage CLIP good **global** properties
- Perform prompt assignment is a **sliding window** fashion
- Aggregate **predictions**

Wysoczanska et al., CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free, WACV’24

Objectness guided fusion



CLIP-DIY [Wysoczanska et al. WACV’24]

- **Idea:** leverage CLIP good **global** properties
- Perform prompt assignment is a **sliding window** fashion
- Aggregate **predictions**

Objectness guided fusion

- Assign text prompts to **FOUND foreground** pixels
- Leverage CLIP at best: in it **global** ability

Wysoczanska et al., CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free, WACV’24

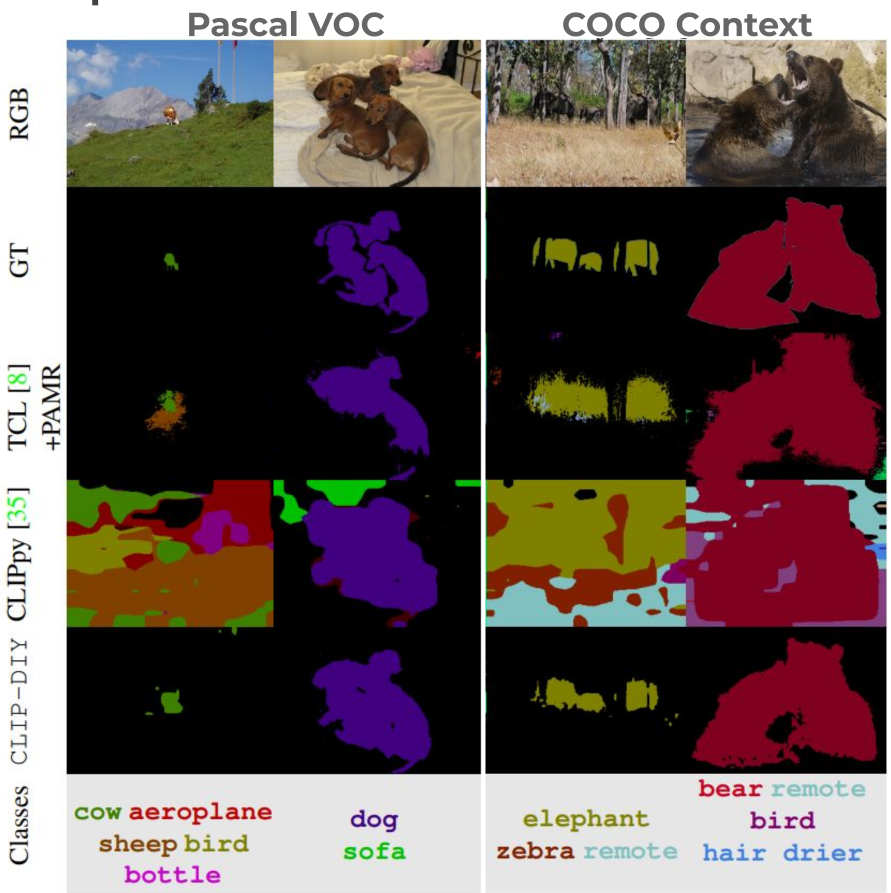
Qualitative results

Pascal VOC

COCO

Wysoczanska et al., CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free, WACV'24

Comparison to SOTA



CLIP-DIY [Wysoczanska et al. WACV'24]

- Use **CLIP** as is designed
- **Training-free**
- No post-processing

Comparison to SOTA

Method	extra training ?	Backbones Visual	Backbones Text	PASCAL VOC	COCO Object
ReCo† [41]	✓	ViT-L/14*	CLIP-ViT-L/14*	25.1	15.7
ViL-Seg [26]	✓	ViT-B/16		37.3	-
MaskCLIP+† [58]	✓	ResNet101 [19]		38.8	20.6
CLIPpy [35]	✓	ViT-B/16	T-5 [34]	52.2	**32.0**
GroupViT [53]	✓	ViT-S/16	12T	52.3	-
ViewCo [37]	✓	ViT-S/16	12T	52.4	23.5
SegCLIP [27]	✓	ViT-B/16	CLIP-ViT-B/16	52.6	26.5
OVSegmentor [54]	✓	ViT-B/16	BERT-ViT-B/16	53.8	25.1
TCL [8] + PAMR [2]	✓	ViT-B/16	CLIP-ViT-B/16	55.0	31.6
CLIP-DIY (ours)		ViT-B/16	CLIP-ViT-B/16	59.0	30.4
CLIP-DIY (ours)		ViT-B/32	CLIP-ViT-B/32	**59.9**	31.0

CLIP-DIY [Wysoczanska et al. WACV'24]

- Use **CLIP** as is designed
- **Training-free**
- No post-processing

Wysoczanska et al., CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free, WACV'24

CLIP-DIY: In the wild

Wysoczanska et al., CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free, WACV'24

Sneak peek to our recent work

https://arxiv.org/abs/2312.12359

CLIP-DINOiser: Teaching CLIP a few DINO tricks

Monika Wysoczańska[1] Oriane Siméoni[2] Michaël Ramamonjisoa[3] Andrei Bursuc[2]
Tomasz Trzcinski[1,4,5] Patrick Pérez[2]
[1]Warsaw University of Technology, [2]Valeo.ai, [3]Meta AI, [4]Tooploox, [5]IDEAS NCBR

Related works

Unsupervised object localization

- **LOD:** Large-Scale Unsupervised Object Discovery. *Vo et al.* NeurIPS'21
- **TokenCut:** Self-supervised transformers for unsupervised object discovery using normalized cut. *Wang et al.* CVPR'22
- **Deep Spectral Methods:** A surprisingly strong baseline for unsupervised semantic segmentation and localization. *Melas-Kyriazi et al.* CVPR'22
- **SelfMask:** Unsupervised salient object detection with spectral cluster voting. *Shi et al.* CVPRW'22
- **CutLER:** Cut and Learn for Unsupervised Object Detection and Instance Segmentation. *Wang et al.* CVPR'23

Zero-shot semantic segmentation

- **CLIP:** Openclip. Ilharco et al. 2021
- **MaskCLIP:** Extract free dense labels from clip. *Zhou et al.* ECCV'22
- **TCL:** Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. *Cha et al.* CVPR'23
- **CLIPpy:** Perceptual grouping in contrastive vision-language models. *Ranasinghe et al.* ICCV'23

References

Unsupervised object localization

- Localizing Objects with Self-Supervised Transformers and no Labels, *Siméoni et al.*, BMVC'21
- Unsupervised Object Localization: Observing the Background to Discover Objects, *Siméoni et al.*, CVPR'23
- Unsupervised Object Localization in the Era of Self-Supervised ViTs: A Survey, *Siméoni et al.*, arxiv'23

Open-vocabulary zero-shot semantic segmentation

- CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free, *Wysoczanska et al.*, WACV'24
- CLIP-DINOiser: Teaching CLIP a few DINO tricks, *Wysoczanska et al.*, arxiv'23

Collaborators


Gilles Puy


Eloi Zablocki


Patrick Pérez


Monika Wysoczańska


Spyros Gidaris


Michaël Ramamonjisoa


Antonin Vobecky


Renaud Marlet


Chloé Sekkat


Huy V. Vo


Jean Ponce


Andrei Bursuc


Simon Roburin

Conclusion

- We can find objects **without knowing anything** about them
- **Self-supervised features** are powerful and contain good localization properties without any human made annotation
- We can easily extract **one object** or localize *all* by looking for the **background**
- We can leverage **open-vocabulary** features to **densely** assign prompts to pixels

Perspective

- The definition of object is **ill-defined**, we might want to handle **different level of granularity**
- SSL correlation do not allow to separate similar objects → leverage more type of features ?
- What about features learnt on **non object-centric/curated data**?