ReCo: Retrieve and Co-Segment

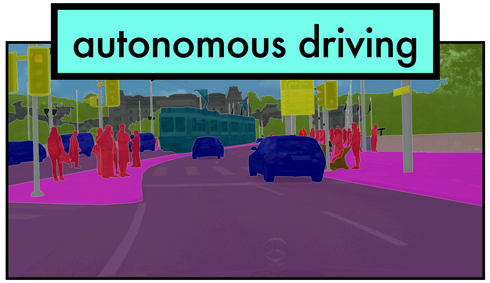
Why it is useful | How it works

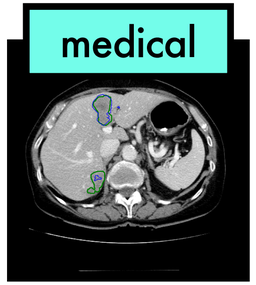
Paper: G. Shin, W. Xie, S. Albanie, "ReCo: Retrieve and Co-segment for Zero-shot Transfer", NeurIPS (2022)

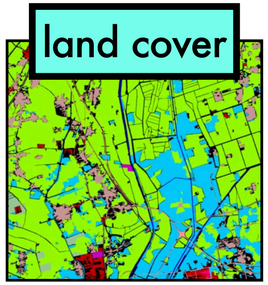
Motivation

Task: semantic segmentation (label the pixels)

Applications of semantic segmentation:







Four key challenges for existing approaches:

Cost	Flexibility	Deployment	Access
manual annotation	limited categories	need labelled examples	need target distribution

ReCo targets these challenges based on two findings:

❶ Modern DNNs learn object extent/correspondences without pixel-level supervision — DINO | DeiT-S-SIN | STEGO

❷ Language & vision pretraining produces models with large vocabulary/zero-shot transfer — CLIP

ReCo: Curate archives | Co-segment concepts | Construct segmenter

(optional) **ReCo+:** target domain adaptation via pseudolabelling

Example predictions (COCO Stuff)

References/image credits:
M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding", CVPR (2016)
P. Bilic et al., "The liver tumor segmentation benchmark (lits)", arxiv (2019)
X-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models", RSE (2020)
(DINO) M. Caron et al., "Emerging properties in self-supervised vision transformers", ICCV (2021)
(DeiT-S-SIN) M. Naseer et al., "Intriguing properties of vision transformers", NeurIPS (2021)
(STEGO) M. Hamilton et al., "Unsupervised Semantic Segmentation by Distilling Feature Correspondences", ICLR (2022)
(CLIP) A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)
(ReCo) G. Shin et al., "ReCo: Retrieve and Co-segment for Zero-shot Transfer", NeurIPS (2022)
(COCO Stuff) H. Caesar, "Coco-stuff: Thing and stuff classes in context", CVPR (2018)

ReCo

image collection

archive

CLIP

Image encoder

F_1
F_2
$\vdots$
F_k

Hypothesis 1: modern image datasets contain concepts of interest

A photo of a lionfish

Dense CLIP

(1) Retrieval

(2) Co-segmentation

$F_1 \otimes F_1$ $F_1 \otimes F_2$ $\cdots$ $F_1 \otimes F_k$
$F_2 \otimes F_1$ $F_2 \otimes F_2$ $\cdots$ $F_2 \otimes F_k$
$\vdots$ $\vdots$ $\ddots$ $\vdots$
$F_k \otimes F_1$ $F_k \otimes F_2$ $\cdots$ $F_k \otimes F_k$

seeds

reference embedding

lionfish

Hypothesis 2: CLIP curates high purity archives

context elimination
suppress background

Language-guided co-segmentation | enhance foreground

unseen image

A photo of a lionfish

Image encoder

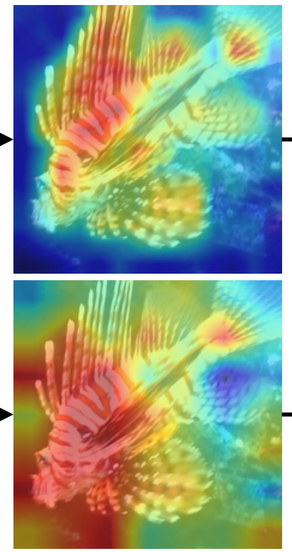
F_{new}

$\otimes$

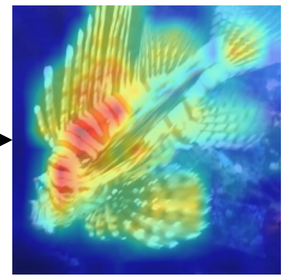
lionfish

reference embedding

Dense CLIP



(3) Inference



CRF | postprocess

ReCo+ | Train DeepLabv3+ on ReCo pseudolabels

References/image credits:
(ReCo) G. Shin et al., "ReCo: Retrieve and Co-segment for Zero-shot Transfer", NeurIPS (2022)
(CLIP) A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)
(DenseCLIP) C. Zhou et al., "Extract free dense labels from clip", arxiv (2021) - note: there is an updated version of this work called MaskCLIP (ReCo builds on the original DenseCLIP)

Data Flows

Pretraining data

CLIP	WebImageText ?
400M (image, alt-text) pairs	

Co-segmentation backbone

DeiT-S-SIN	ImageNet (SIN) 
1.2M (class) labelled images	

DeepLabv3+ (in ReCo+)

ResNet-101	ImageNet 1K 
1.2M (class) labelled images	

Archive/co-segmentation data

ImageNet 1K 

1.2M unlabelled images

LAION-5B (filtered) 

Filtered subsets from 5B images

Ablation studies

Pascal Context (val set) 

5K segmented images with 59 categories

Test data

Cityscapes (val set) 

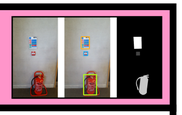
500 segmented images, 27 categories

KITTI-STEP (val set) 

3K segmented images, 19 categories

COCO Stuff (val set) 

4K segmented images, 27 mid-level categories

Firenet 

1.5K segmented images (*fire extinguisher* class)

ReCo+

Cityscapes (train)	no labels
KITTI-STEP (train)	COCO Stuff10K (train)

References:
(WebImageText) A. Radford et al., "Learning transferable visual models from natural language supervision", ICML (2021)
(ImageNet SIN) R. Geirhos et al., "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness", ICLR (2018)
(DeepLabv3+) L-C. Chen et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation", ECCV (2018)
(ImageNet) J. Deng et al., "Imagenet: A large-scale hierarchical image database", CVPR (2009)
(Pascal Context) R. Mottaghi et al., "The role of context for object detection and semantic segmentation in the wild", CVPR (2014)

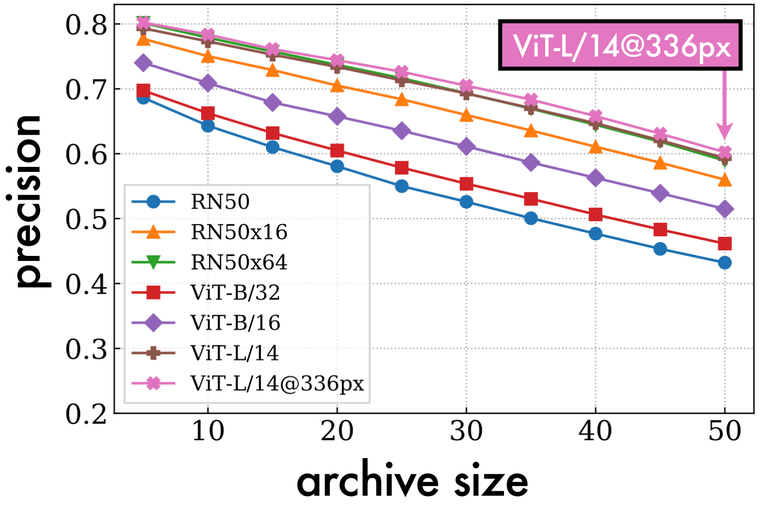
M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding", CVPR (2016)
M. Weber et al., "STEP: Segmenting and Tracking Every Pixel", NeurIPS Data Track (2021)
F. Panella et al., "Firenet dataset", https://www.firenet.xyz/

Ablation Studies

Archive curation

Assess CLIP archive curation

ImageNet-1K retrieval

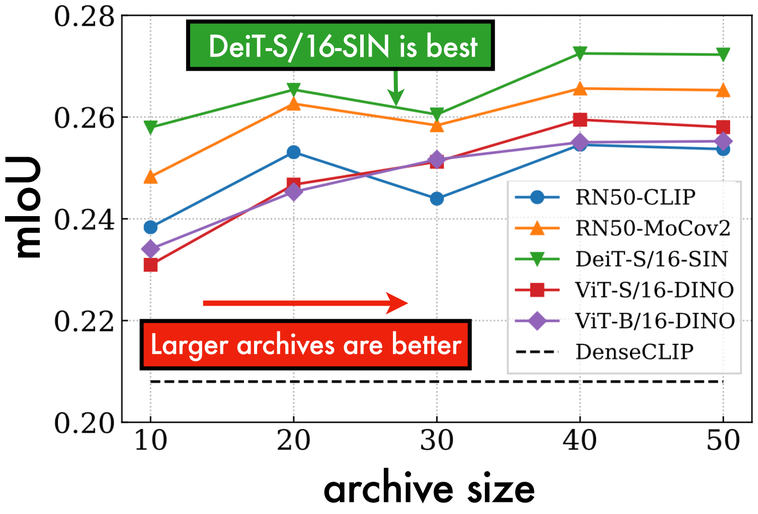


Takeaway: CLIP can curate archives of high purity

Co-segmentation

Influence of archive size/encoder

PASCAL-Context segmentation



Takeaways: DeiT-S/16-SIN best; larger archives work better

ReCo Components

Framework contributions

PASCAL-Context segmentation

DenseCLIP	LGC	CE	CRF	mIoU
✗	✗	✗	✗	5.7
✓	✗	✗	✗	21.8
✓	✓	✗	✗	23.1
✓	✗	✓	✗	26.0
✓	✓	✓	✗	26.6
✓	✓	✓	✓	**27.2**

Takeaways:

- DenseCLIP - major gain
- CRF - minor gain
- All components help

References/image credits:
(ReCo) G. Shin et al., "ReCo: Retrieve and Co-segment for Zero-shot Transfer", NeurIPS (2022)
(ImageNet) J. Deng et al., "Imagenet: A large-scale hierarchical image database", CVPR (2009)
(Pascal Context) R. Mottaghi et al., "The role of context for object detection and semantic segmentation in the wild", CVPR (2014)
(DenseCLIP) C. Zhou et al., "Extract free dense labels from clip", arxiv (2021)

Comparisons with existing approaches

COCO Stuff

Model	Acc.	mIoU
Zero-shot transfer		
DenseCLIP‡	32.2	19.6
ReCo‡	**46.1**	**26.3**
Unsupervised adaptation		
IIC	21.8	6.7
MDC	32.2	9.8
PiCIE	48.1	13.8
PiCIE + H	50.0	14.4
STEGO	**56.9**	28.2
ReCo+‡	54.1	**32.6**

Cityscapes

Model	Acc.	mIoU
Zero-shot transfer		
DenseCLIP⋆‡	35.9	10.0
MDC⋆†	-	7.0
PiCIE⋆†	-	9.7
D&S⋆†	-	16.2
ReCo⋆‡	**65.4**	**22.0**
Unsupervised adaptation		
IIC	47.9	6.4
MDC	40.7	7.1
PiCIE	65.5	12.3
STEGO	73.2	21.0
ReCo+‡	**83.7**	**24.2**

KITTI-STEP

Model	Acc.	mIoU
Zero-shot transfer		
DenseCLIP‡	34.1	15.3
ReCo‡	**70.6**	**29.8**
Unsupervised adaptation		
SegSort	69.8	19.2
HSG	73.8	21.7
ReCo+‡	**75.3**	**31.9**

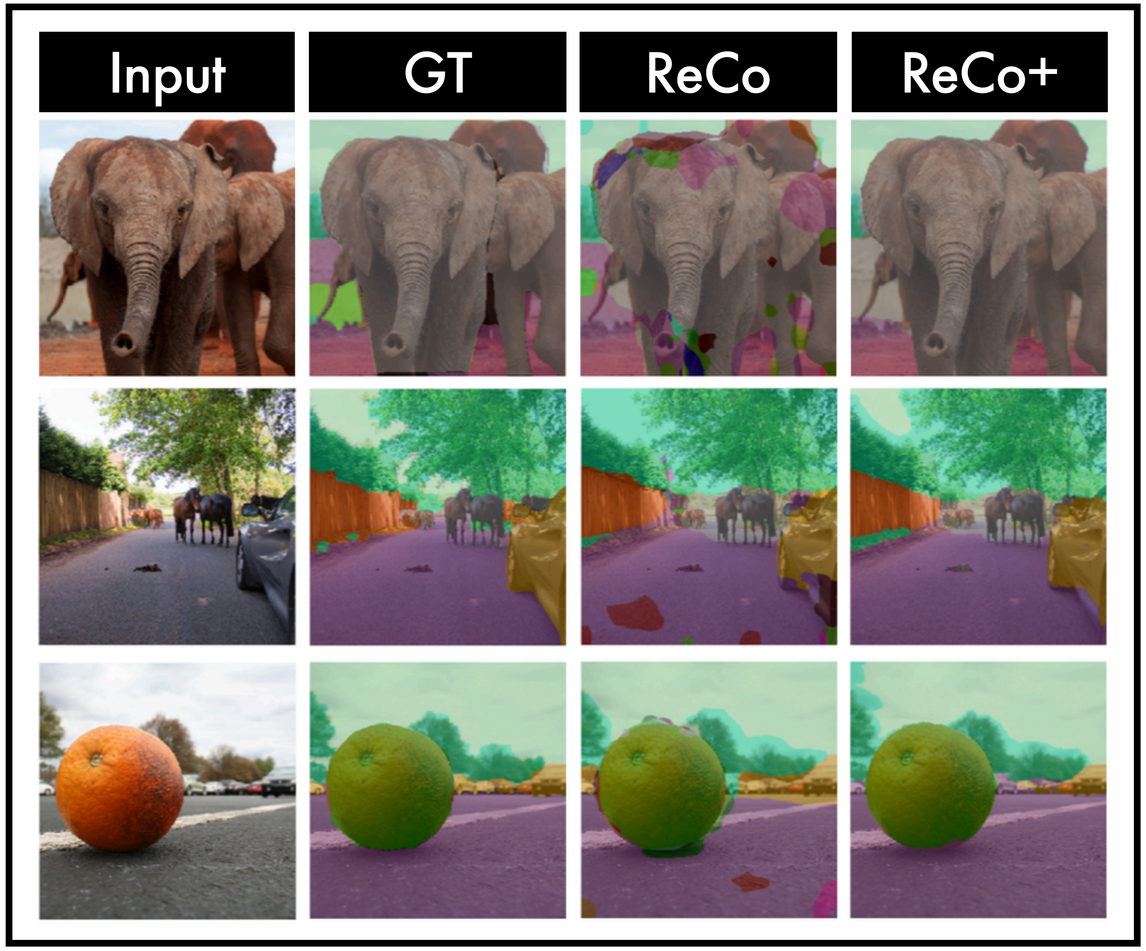
‡ vision-language pretraining

⋆ evaluate at original resolution

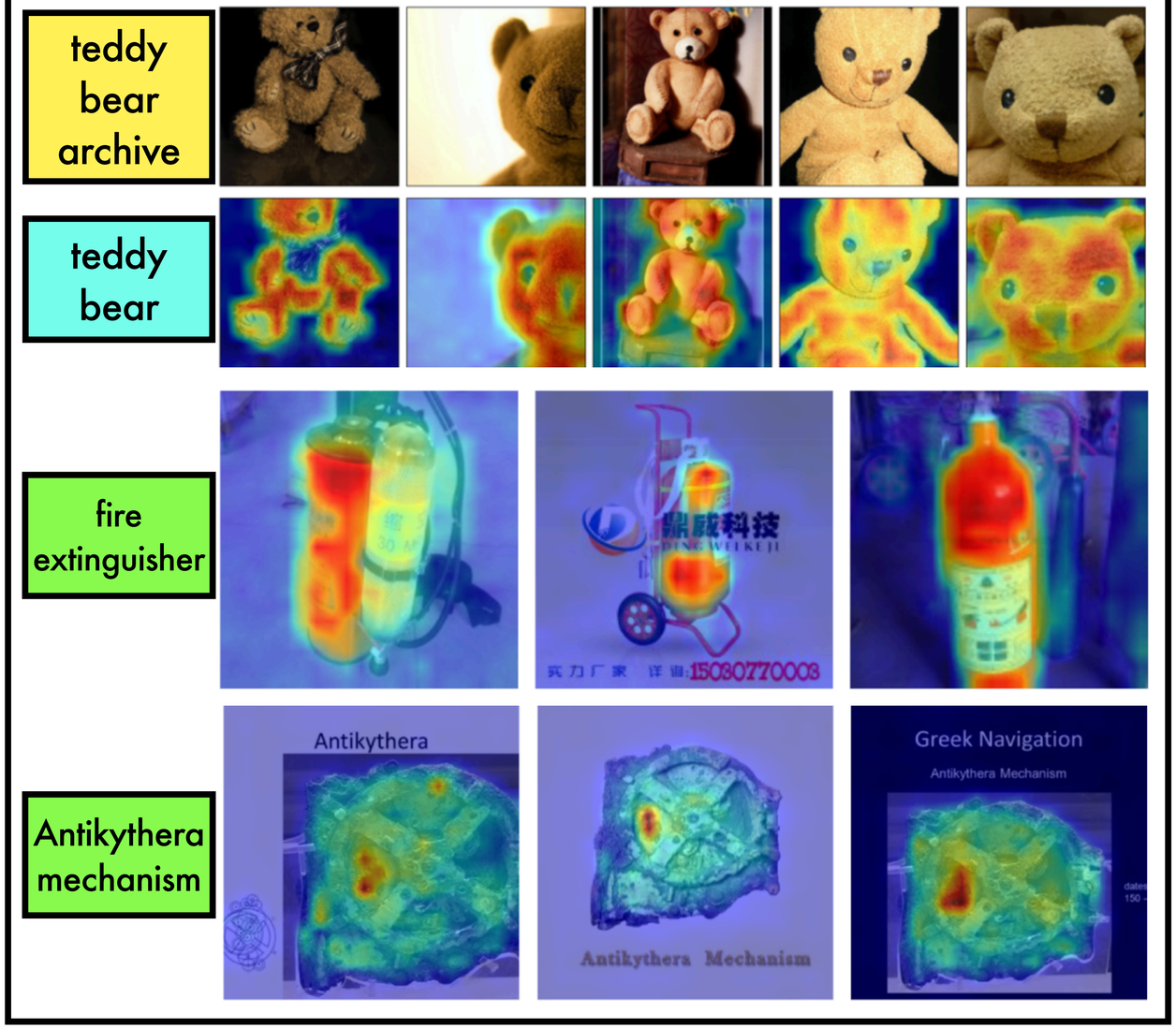
† Waymo Open training

References:
(DenseCLIP) C. Zhou et al., "Extract free dense labels from clip", arxiv (2021)
(ReCo) G. Shin et al., "ReCo: Retrieve and Co-segment for Zero-shot Transfer", NeurIPS (2022)
(IIC) X. Ji, "Invariant information clustering for unsupervised image classification and segmentation", CVPR (2019)
(MDC/PiCIE) J. Cho et al., "PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering", CVPR (2021)
(STEGO) M. Hamilton et al., "Unsupervised Semantic Segmentation by Distilling Feature Correspondences", ICLR (2022)
(D&S) A. Vobecky et al., "Drive&Segment: Unsupervised Semantic Segmentation of Urban Scenes via Cross-modal Distillation", arxiv (2022)
(Waymo Open) P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset", CVPR (2020)
(SegSort) J-J. Hwang et al., "Segsort: Segmentation by discriminative sorting of segments", ICCV (2019)
(HSG) T. Ke et al., "Unsupervised Hierarchical Semantic Segmentation with Multiview Cosegmentation and Clustering Transformers", CVPR (2022)

Qualitative Results



ReCo co-segmentations



References/image credits:
(ReCo) G. Shin et al., "ReCo: Retrieve and Co-segment for Zero-shot Transfer", NeurIPS (2022)

Limitations

Limitations of the ReCo framework and research methodology:

1. Some concepts are so rare that they do not appear in billion-image datasets — ReCo cannot segment these
2. Inference uses CLIP and visual encoder — computationally expensive — could address in future work
3. ImageNet is used for experiments — strong object-centric bias — may present optimistic assessment of ReCo generality
4. CLIP is expensive to retrain if new concept emerges — hard to update archives (e.g. for new products)
5. ReCo avoids pixel-level labels but design guided by ablations on labelled data — indirect supervision

References:
(ReCo) G. Shin et al., "ReCo: Retrieve and Co-segment for Zero-shot Transfer", NeurIPS (2022)

Related Work

Unsupervised semantic segmentation

- maximising mutual information — IIC (2019), AC (2020)
- metric learning across proposals — HG (2020), MaskContrast (2021)
- equivariance constraints — PiCIE (2021)
- distillation of feature correspondences — STEGO (2022)
- cross modal cues (vision and LiDAR) — D&S (2022)

ReCo: no Hungarian matching required to link predictions and labels

Weakly-supervised semantic segmentation

- pointing — Point-level (2016)
- sparse pixel labels — PixelPick (2021)
- scribbles — ScribbleSup (2016)
- extreme clicks — ExtremeClicking (2017)
- image-level labels — MIL (2015)

ReCo: can learn segmenter from any unlabelled collection of images

Zero-shot segmentation

- leverage word embeddings — OVSP (2017), ZS3Net (2019)
- leverage CLIP embeddings — DenseCLIP (2021), LSeg (2022)

ReCo focuses on zero-shot transfer

Directly comparable to "annotation-free" DenseCLIP

Co-segmentation

- classical approaches — Trust region graph cuts (2017)
- shared encoder networks — ABOCS (2018)
- iterative refinement — CycleSegNet (2021)
- weak supervision — Co-attention CNNs (2017)

ReCo can in principle use any co-segmentation approach

References:
(ReCo) G. Shin et al., "ReCo: Retrieve and Co-segment for Zero-shot Transfer", NeurIPS (2022)
(IIC) X. Ji, "Invariant information clustering for unsupervised image classification and segmentation", CVPR (2019)
(AC) Y. Ouali, "Autoregressive unsupervised image segmentation", ECCV (2020)
(HG) X. Zhang, "Self-supervised visual representation learning from hierarchical grouping", NeurIPS (2020)
(MaskContrast) W. Van Gansbeke et al., "Unsupervised semantic segmentation by contrasting object mask proposals", ICCV (2021)
(PiCIE) J. Cho et al., "PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering", CVPR (2021)
(STEGO) M. Hamilton et al., "Unsupervised Semantic Segmentation by Distilling Feature Correspondences", ICLR (2022)
(D&S) A. Vobecky et al., "Drive&Segment: Unsupervised Semantic Segmentation of Urban Scenes via Cross-modal Distillation", arxiv (2022)
(Point-level) A. Bearman et al., "What's the point: Semantic segmentation with point supervision", ECCV (2016)
(PixelPick) G. Shin et al., "All you need are a few pixels: semantic segmentation with PixelPick", ICCVW (2021)
(ScribbleSup) D. Lin et al., "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation", CVPR (2016)
(ExtremeClicking) D. Papadopoulos et al., "Extreme clicking for efficient object annotation", ICCV (2017)
(MIL) P. Pinheiro et al., "From image-level to pixel-level labeling with convolutional networks", CVPR (2015)
(OVSP) H. Zhao et al., "Open vocabulary scene parsing", ICCV (2017)
(ZS3Net) M. Bucher et al., "Zero-shot semantic segmentation", NeurIPS (2019)
(DenseCLIP) C. Zhou et al., "Extract free dense labels from clip", arxiv (2021)
(LSeg) B. Li et al., "Language-driven Semantic Segmentation", ICLR (2022)
(TRGC) C. Rother et al., "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs", CVPR (2006)
(ABOCS) H. Chen et al., "Semantic aware attention based deep object co-segmentation", ACCV (2018)
(CycleSegNet) C. Zhang et al., "Cyclesegnet: Object co-segmentation with cycle refinement and region correspondence", TIP (2021)
(Co-attention CNNs) K-J. Hsu et al., "Co-attention CNNs for unsupervised object co-segmentation", IJCAI (2018)