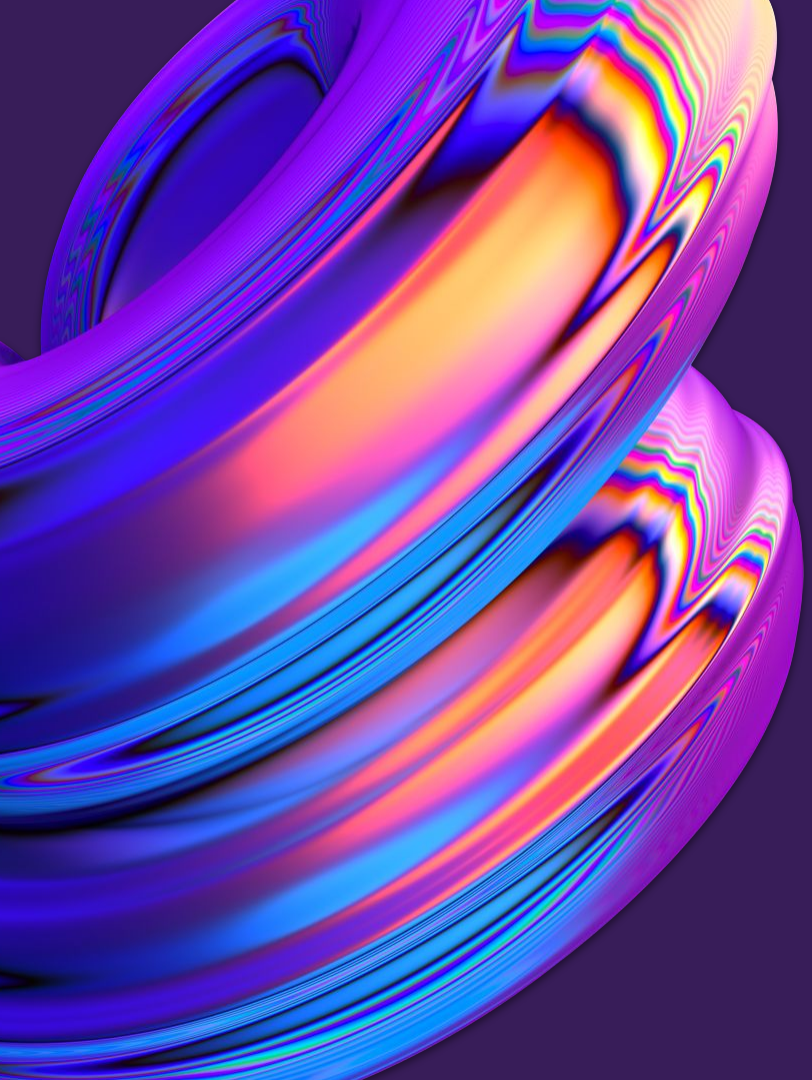


MERANTIX MOMENTUM

Recent Advances in Unsupervised Semantic Segmentation

Alexander Koenig
ML Researcher

August 24th 2023
Switzerland Innovation Park Basel

Semantic Segmentation

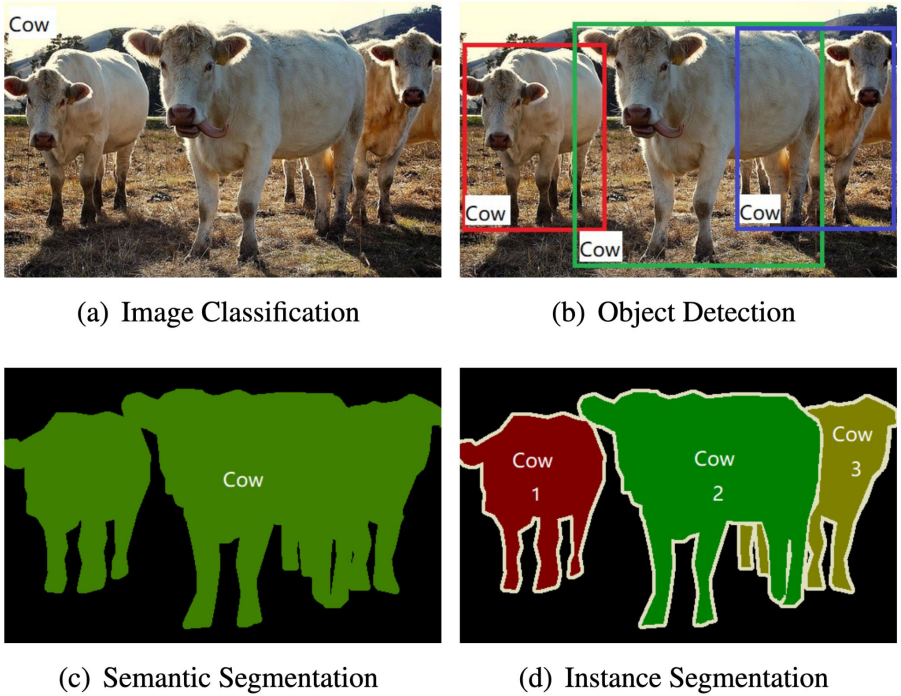


Figure: Taxonomy of computer vision tasks.

Wu et al. “Recent Advances in Deep Learning for Object Detection” Neurocomputing, 2020.

Examples of Semantic Segmentation

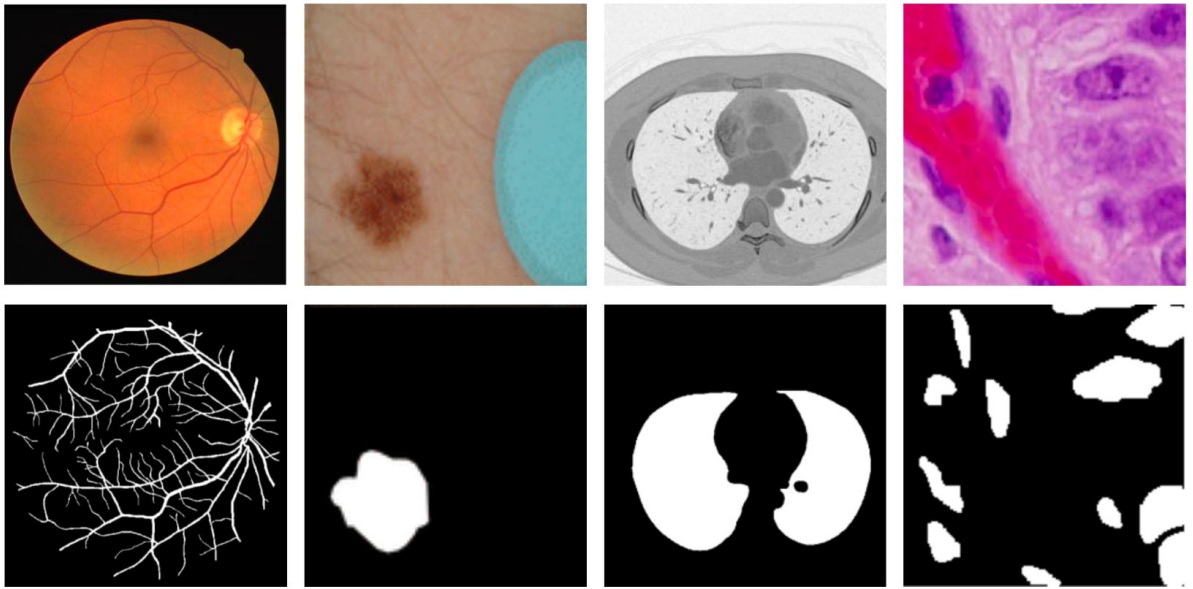


Figure: Application areas of segmentation in biomedicine.

Asadi-Aghbolaghi et al. “Multi-level Context Gating of Embedded Collective Knowledge for Medical Image Segmentation” arxiv, 2020.

Supervised vs Unsupervised Segmentation

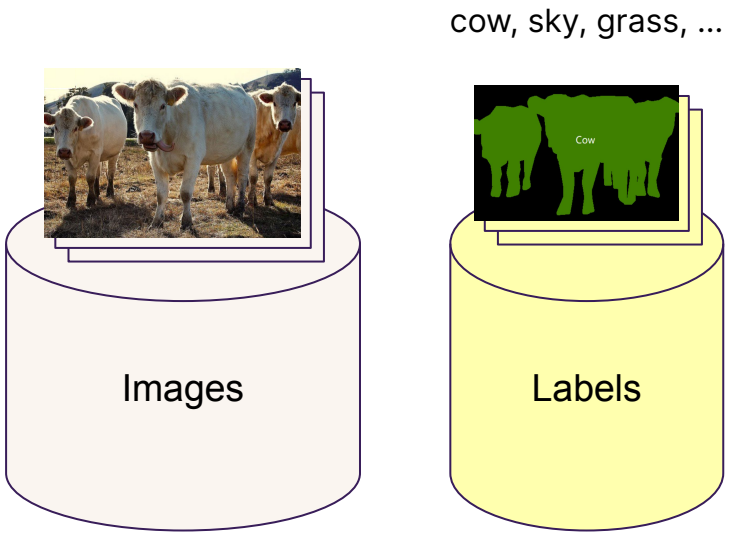


Figure 1: A *supervised dataset* relates each pixel to a fixed set of target classes.

Figure 2: A *unsupervised dataset* consists of just images without targets.

Wu et al. “Recent Advances in Deep Learning for Object Detection” Neurocomputing, 2020.

Unsupervised Semantic Segmentation

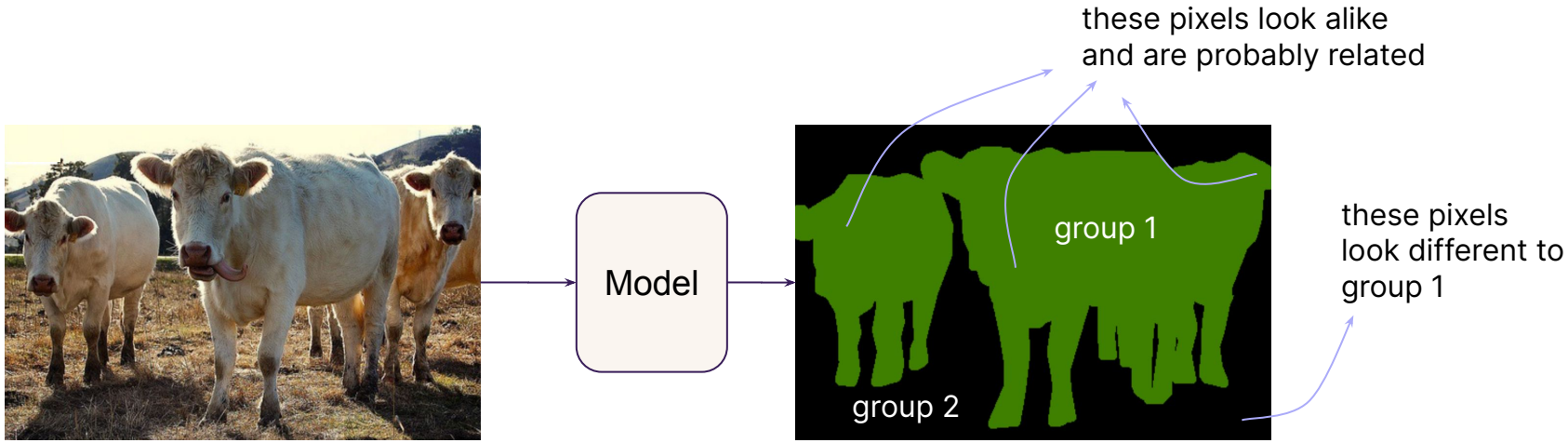


Figure: Perceptual grouping in unlabeled images.

Goal: find perceptual groups that hold within and across images.

Wu et al. “Recent Advances in Deep Learning for Object Detection” Neurocomputing, 2020.

Properties of Self-Supervised Vision Transformers

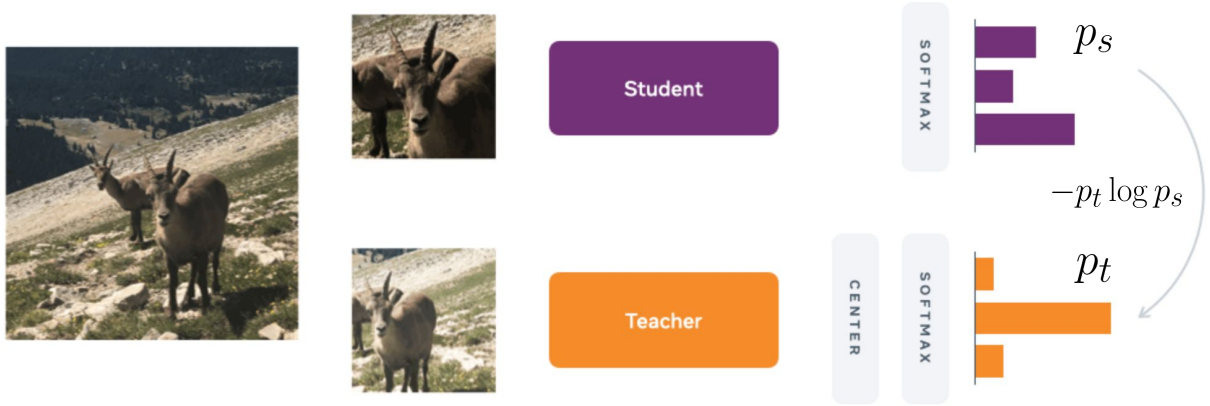


Figure 1: DINO visual pre-training strategy.

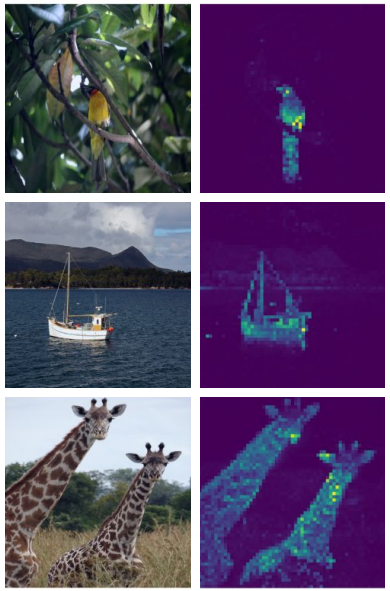
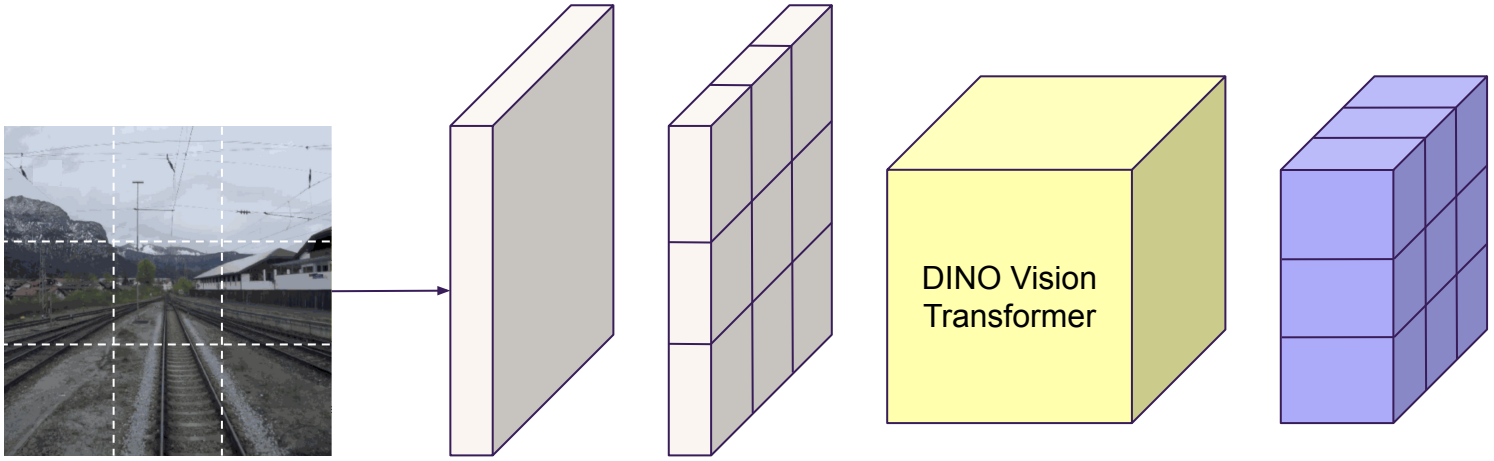


Figure 2: Last-layer attention of DINO pre-trained ViT.

Properties of Self-Supervised Vision Transformers



Key Insight: DINO pre-trained ViT produces dense visual features!

Properties of Self-Supervised Vision Transformers

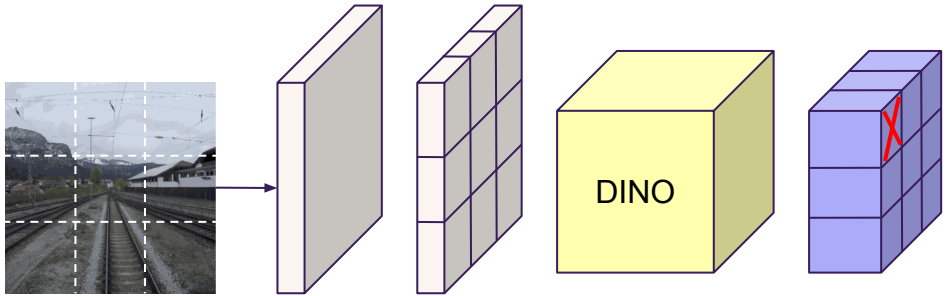


Image and
source token

Similarities
to source token

Properties of Self-Supervised Vision Transformers

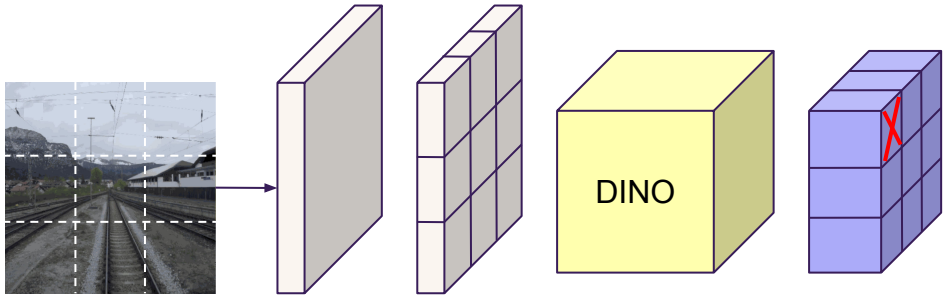
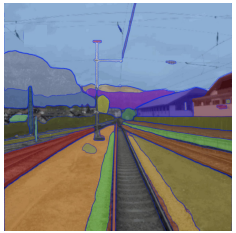


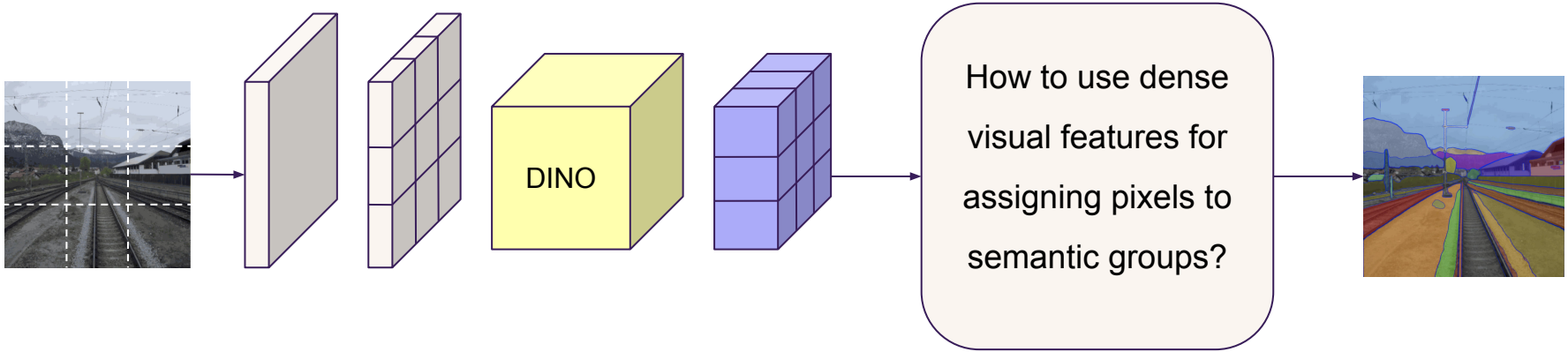
Image and
source token

Similarities
to source token

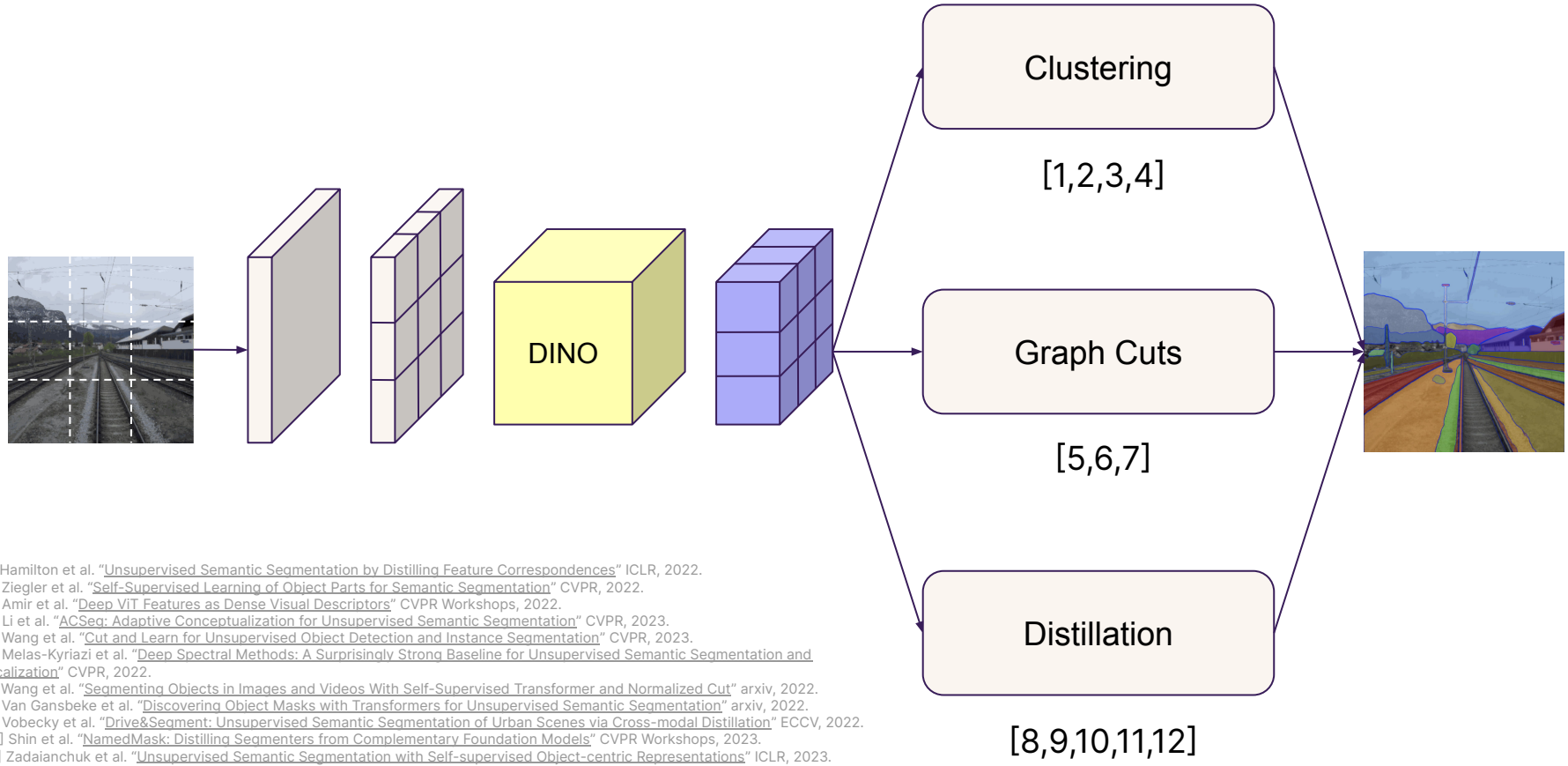


Desired semantic
segmentation

Unsupervised Semantic Segmentation

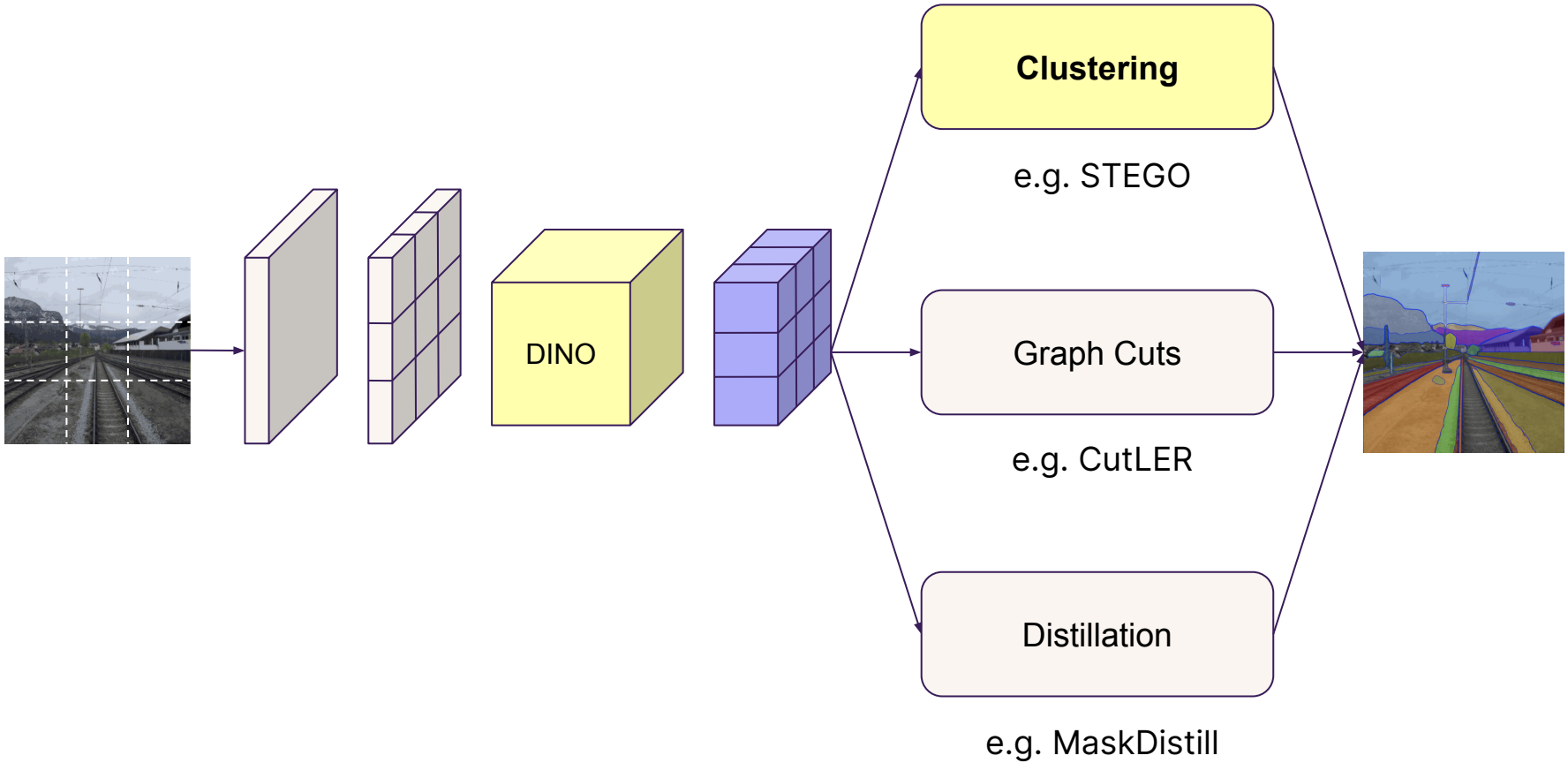


Unsupervised Semantic Segmentation



[1] Hamilton et al. "Unsupervised Semantic Segmentation by Distilling Feature Correspondences" ICLR, 2022.
[2] Ziegler et al. "Self-Supervised Learning of Object Parts for Semantic Segmentation" CVPR, 2022.
[3] Amir et al. "Deep ViT Features as Dense Visual Descriptors" CVPR Workshops, 2022.
[4] Li et al. "ACSeg: Adaptive Conceptualization for Unsupervised Semantic Segmentation" CVPR, 2023.
[5] Wang et al. "Cut and Learn for Unsupervised Object Detection and Instance Segmentation" CVPR, 2023.
[6] Melas-Kyriazi et al. "Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization" CVPR, 2022.
[7] Wang et al. "Segmenting Objects in Images and Videos With Self-Supervised Transformer and Normalized Cut" arxiv, 2022.
[8] Van Gansbeke et al. "Discovering Object Masks with Transformers for Unsupervised Semantic Segmentation" arxiv, 2022.
[9] Vobecky et al. "Drive&Segment: Unsupervised Semantic Segmentation of Urban Scenes via Cross-modal Distillation" ECCV, 2022.
[10] Shin et al. "NamedMask: Distilling Segmenters from Complementary Foundation Models" CVPR Workshops, 2023.
[11] Zadaianchuk et al. "Unsupervised Semantic Segmentation with Self-supervised Object-centric Representations" ICLR, 2023.
[12] Seitzer et al. "Bridging the Gap to Real-World Objectcentric Learning" ICLR, 2023.

Unsupervised Semantic Segmentation



Clustering – Intuition

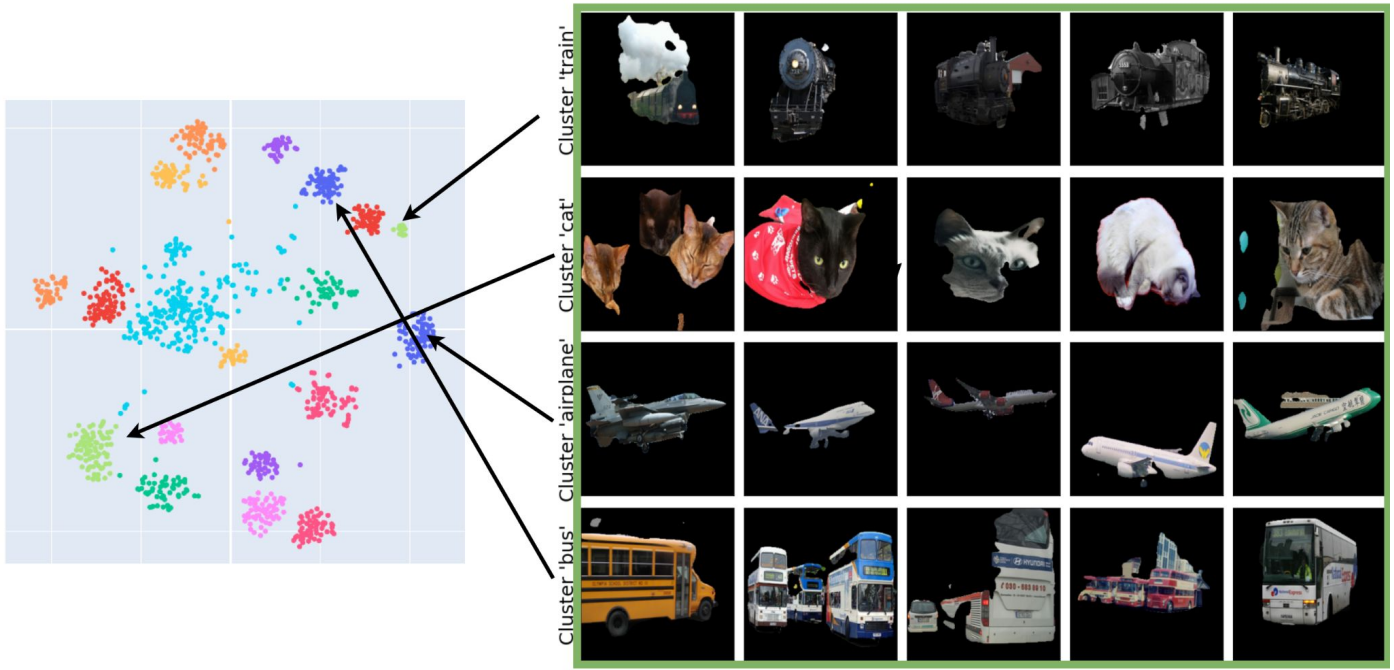


Figure: Pseudo-masks form clusters in DINO feature space.

Zadaianchuk et al. “Unsupervised Semantic Segmentation with Self-supervised Object-centric Representations” ICLR, 2023.

k-Means Clustering

Task: group data into clusters that have

- high intra-cluster similarity
- low inter-cluster similarity

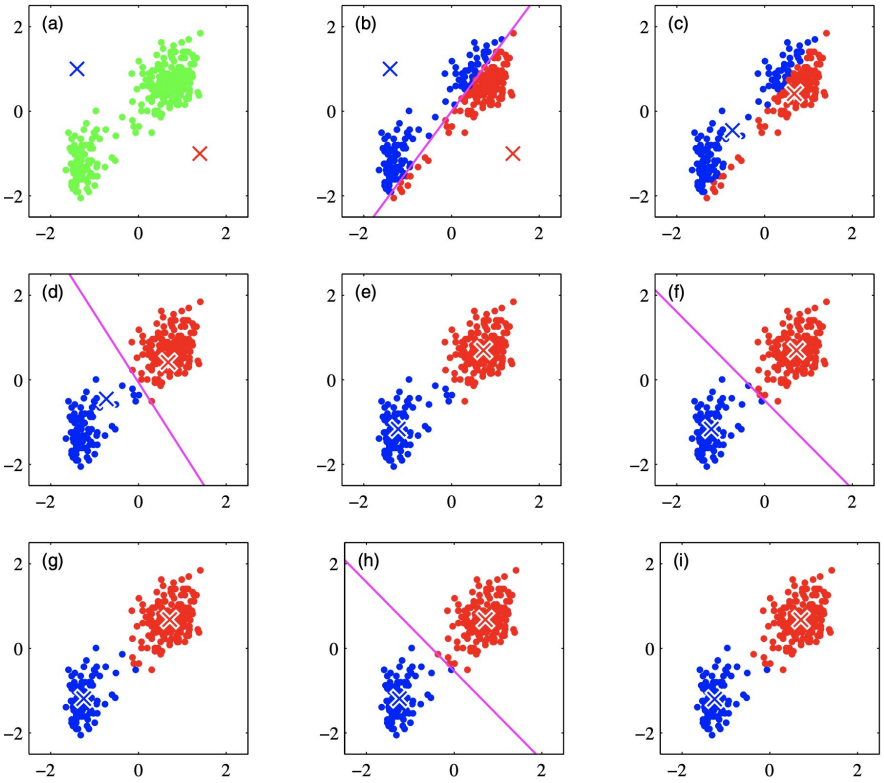


Figure: *k*-Means iterative assignment and updating of cluster centers.

Bishop “Pattern Recognition and Machine Learning” Springer, 2006.

Clustering – STEGO

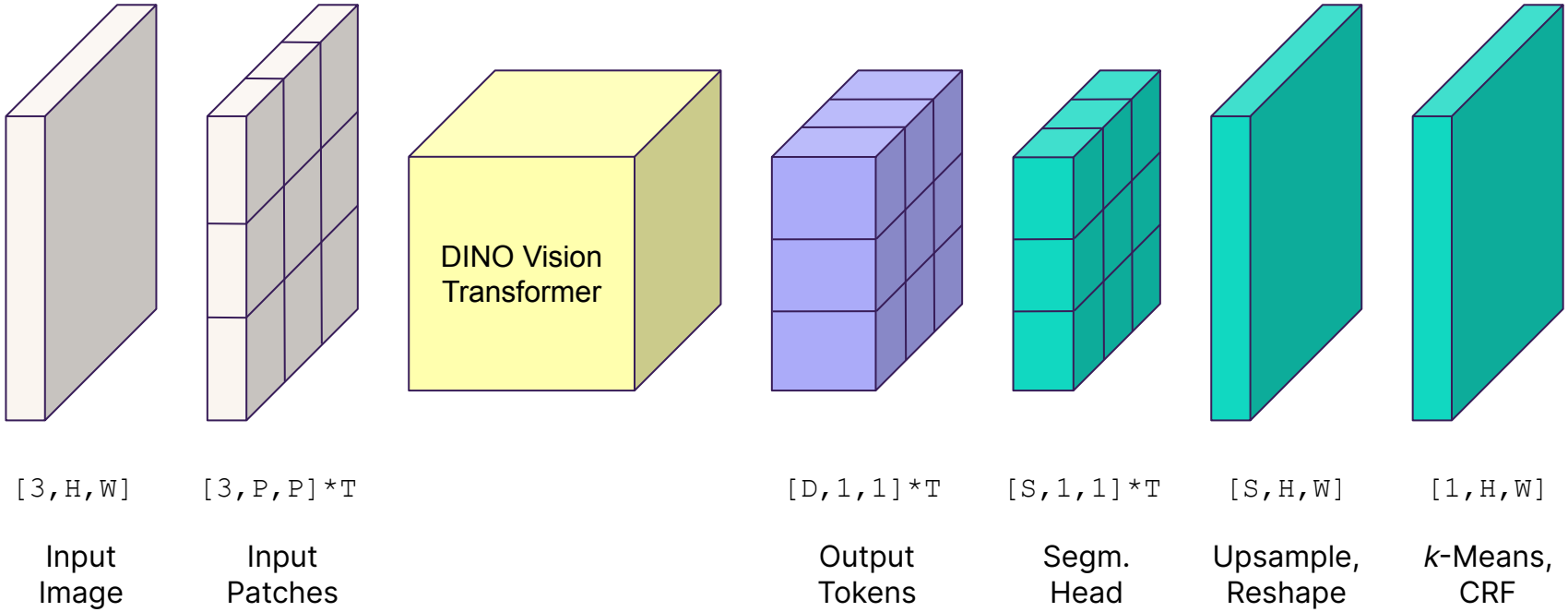


Figure: STEGO projects DINO features using a segmentation head and applies *k*-Means.

Hamilton et al. “Unsupervised Semantic Segmentation by Distilling Feature Correspondences” ICLR, 2022.
Koenig et al. “Uncovering the Inner Workings of STEGO for Safe Unsupervised Semantic Segmentation” CVPR Workshops, 2023.

Clustering – STEGO Results

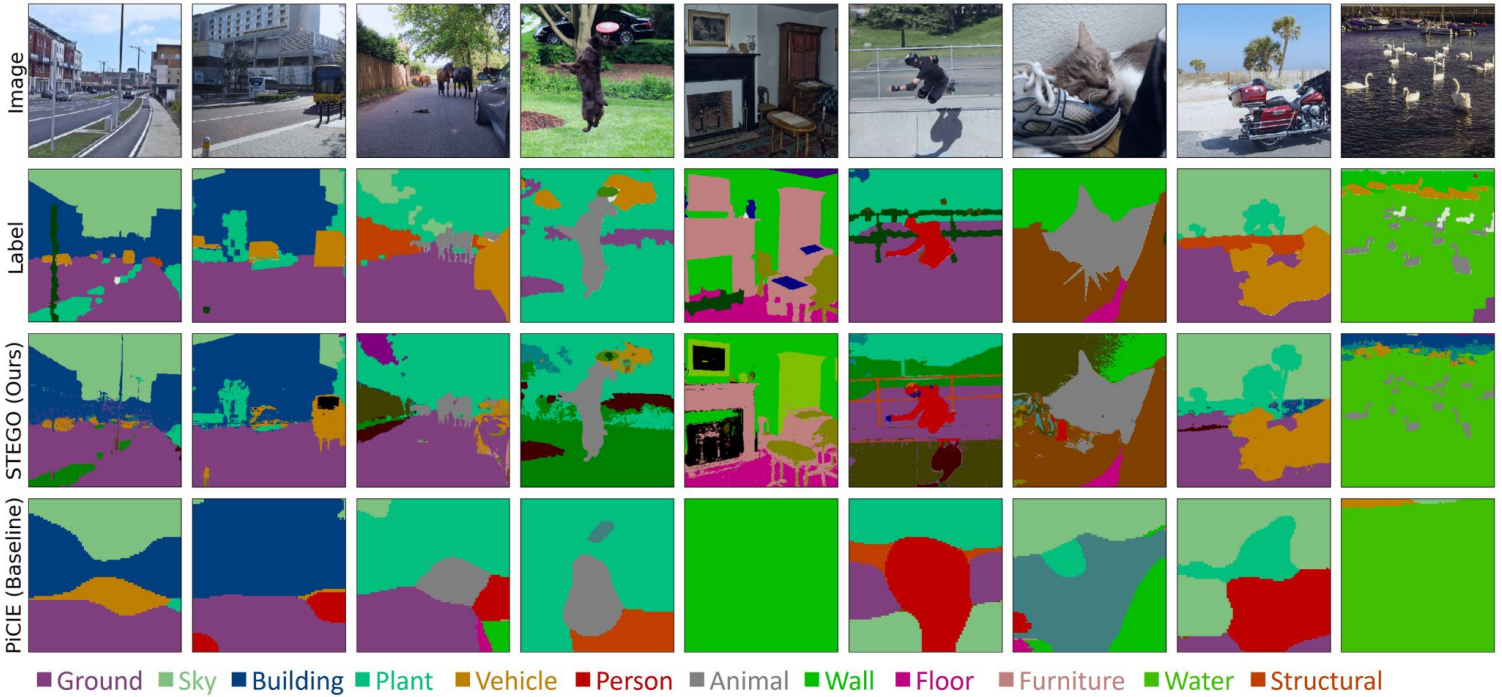


Figure: STEGO results on COCO-Stuff.

Hamilton et al. “Unsupervised Semantic Segmentation by Distilling Feature Correspondences” ICLR, 2022.

Clustering – Quantitative Results

Table 1: Comparison of unsupervised segmentation architectures on 27 class CocoStuff validation set. STEGO significantly outperforms prior art in both unsupervised clustering and linear-probe style metrics.

Model	Unsupervised Accuracy	Unsupervised mIoU	Linear Probe Accuracy	Linear Probe mIoU
ResNet50 (He et al., 2016)	24.6	8.9	41.3	10.2
MoCoV2 (Chen et al., 2020c)	25.2	10.4	44.4	13.2
DINO (Caron et al., 2021)	30.5	9.6	66.8	29.4
Deep Cluster (Caron et al., 2018)	19.9	-	-	-
SIFT (Lowe, 1999)	20.2	-	-	-
Doersch et al. (2015)	23.1	-	-	-
Isola et al. (2015)	24.3	-	-	-
AC (Ouali et al., 2020)	30.8	-	-	-
InMARS (Mirsadeghi et al., 2021)	31.0	-	-	-
IIC (Ji et al., 2019)	21.8	6.7	44.5	8.4
MDC (Cho et al., 2021)	32.2	9.8	48.6	13.3
PiCIE (Cho et al., 2021)	48.1	13.8	54.2	13.9
PiCIE + H (Cho et al., 2021)	50.0	14.4	54.8	14.8
STEGO (Ours)	**56.9**	**28.2**	**76.1**	**41.0**

Table 3: Results on the Cityscapes Dataset (27 Classes). STEGO improves significantly over all baselines in both accuracy and mIoU.

Model	Unsup. Acc.	Unsup. mIoU
IIC (Ji et al., 2019)	47.9	6.4
MDC (Cho et al., 2021)	40.7	7.1
PiCIE (Cho et al., 2021)	65.5	12.3
STEGO (Ours)	**73.2**	**21.0**

Hamilton et al. "Unsupervised Semantic Segmentation by Distilling Feature Correspondences" ICLR, 2022.

Pascal VOC12 Val

Method	mIoU
Sup. ResNet	18.5
Sup. ViT	21.1
DINO [6]	4.6
SwAV [25]	13.7
MoCo-v2 [25]	18.5
MaskContrast [49]	35.0†
Leopart (CBFE+CD)	**41.7**

Table 4. **Unsupervised semantic segmentation results.** We outperform other state-of-the-art methods by a large margin. † indicates result taken from [49].

Ziegler et al. "Self-Supervised Learning of Object Parts for Semantic Segmentation" CVPR, 2022.

Method	mIoU
IIC [20]	9.8
MaskContrast [43]	35.0
DSM† [32]	37.2 ± 3.8
Leopart [59]	41.7
TransFGU [53]	37.2
MaskDistill [44]	42.0
MaskDistill† [44]	45.8
ACSeg (*Ours*)	**47.1** ± 2.4

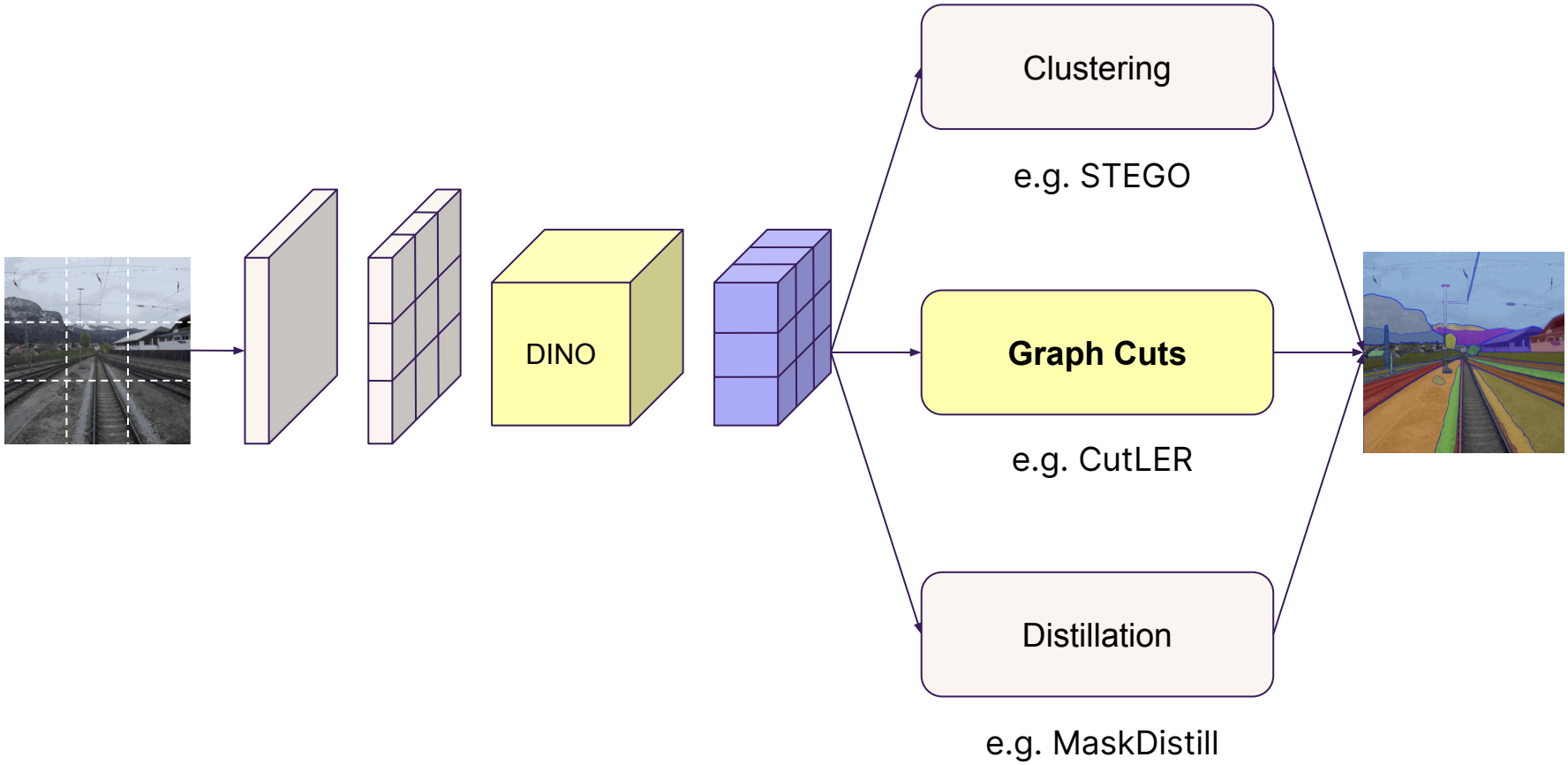
Table 1. **Unsupervised semantic segmentation results on PASCAL VOC.** † denotes results with re-training.

Method	mIoU
MoCo v2 [5]	4.4
IIC [20]	6.7
ImageNet [17]	8.9
DINO [4]	9.6
Modified DC [6]	9.8
PiCIE [6]	13.8
PiCIE+H [6]	14.4
ACSeg (*Ours*)	**16.4** ± 0.9

Table 2. **Unsupervised semantic segmentation results on COCO-Stuff-27 dataset.**

Li et al. "ACSeg: Adaptive Conceptualization for Unsupervised Semantic Segmentation" CVPR, 2023.

Unsupervised Semantic Segmentation

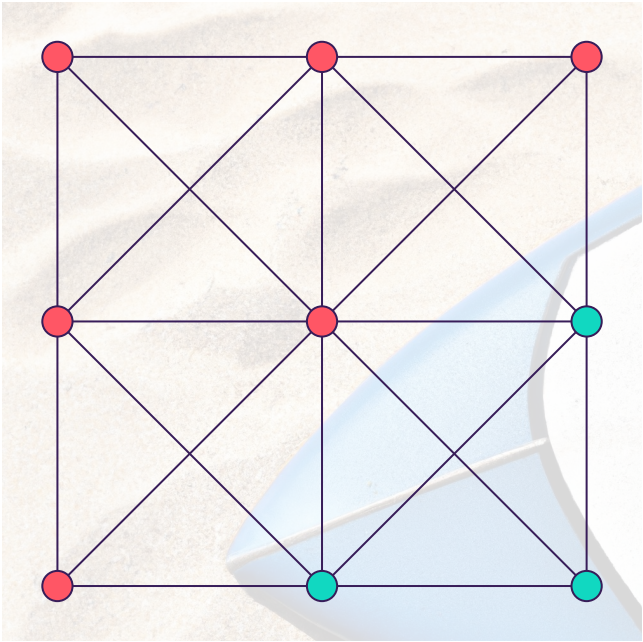


Minimum Graph Cuts

Figure: Example image.

Wu et al. “An Optimal Graph Theoretic Approach to Data Clustering” IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993.

Minimum Graph Cuts



- **Vertices** = Pixels
- **Edge weights** = similarity measure (e.g., function of pixel location and brightness)

Figure: Fully-connected, weighted graph.

Wu et al. “An Optimal Graph Theoretic Approach to Data Clustering” IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993.

Minimum Graph Cuts

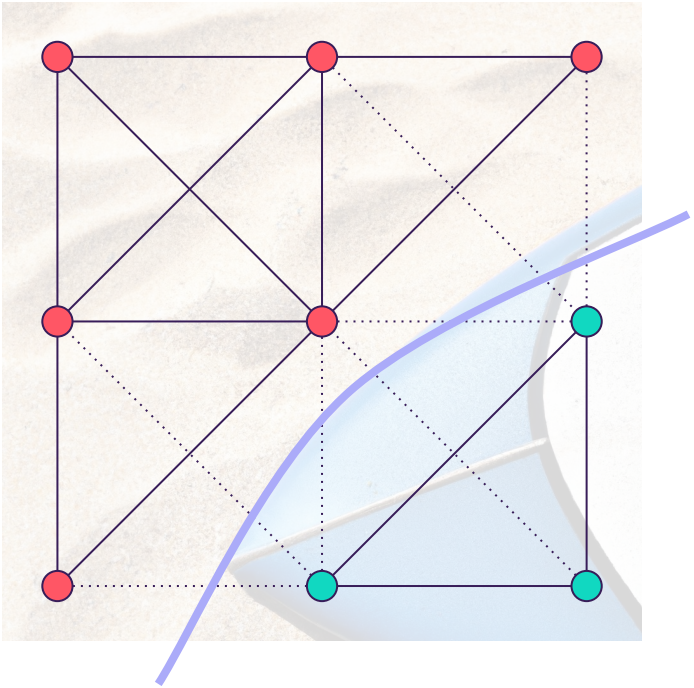


Figure: Minimum graph cut.

- **Vertices** = Pixels
- **Edge weights** = similarity measure (e.g. function of pixel location and brightness)
- **MinCut** [1]: bi-partition minimizing sum of cut edge weights
- **Problem**: MinCut favors small sets of isolated vertices
- **NormCut** [2]: normalizes by size of sub-graph
- Computed via spectral methods

[1] Wu et al. “An Optimal Graph Theoretic Approach to Data Clustering” IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993.
[2] Shi et al. “Normalized cuts and image segmentation” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000.

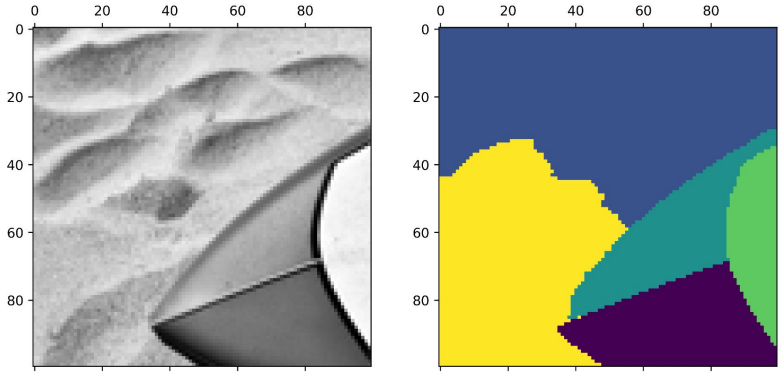
Spectral Clustering

Degree Matrix D (diagonal): sum of edge weights of node i to all other nodes.

$$\boldsymbol{d}(i) = \sum_j w(i,j)$$

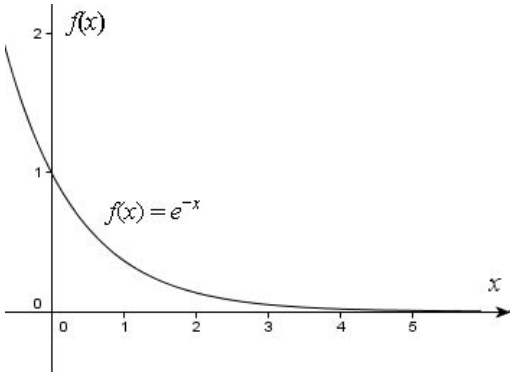
Laplacian matrix L: second smallest eigenvector (Fiedler vector) is used to bi-partition the graph.

L = D - W



Weight matrix W: measures similarity of all pixels.

$$w_{ij} = \underbrace{e^{\frac{-\|\boldsymbol{F}(i)-\boldsymbol{F}(j)\|_2^2}{\sigma_I}}}_{\text{feature similarity}} * \underbrace{\begin{cases} e^{\frac{-\|\boldsymbol{X}(i)-\boldsymbol{X}(j)\|_2^2}{\sigma_X}} & \text{if } \|\boldsymbol{X}(i)-\boldsymbol{X}(j)\|_2 < r \\ 0 & \text{otherwise,} \end{cases}}_{\text{spatial proximity}}$$



Normalized Cuts

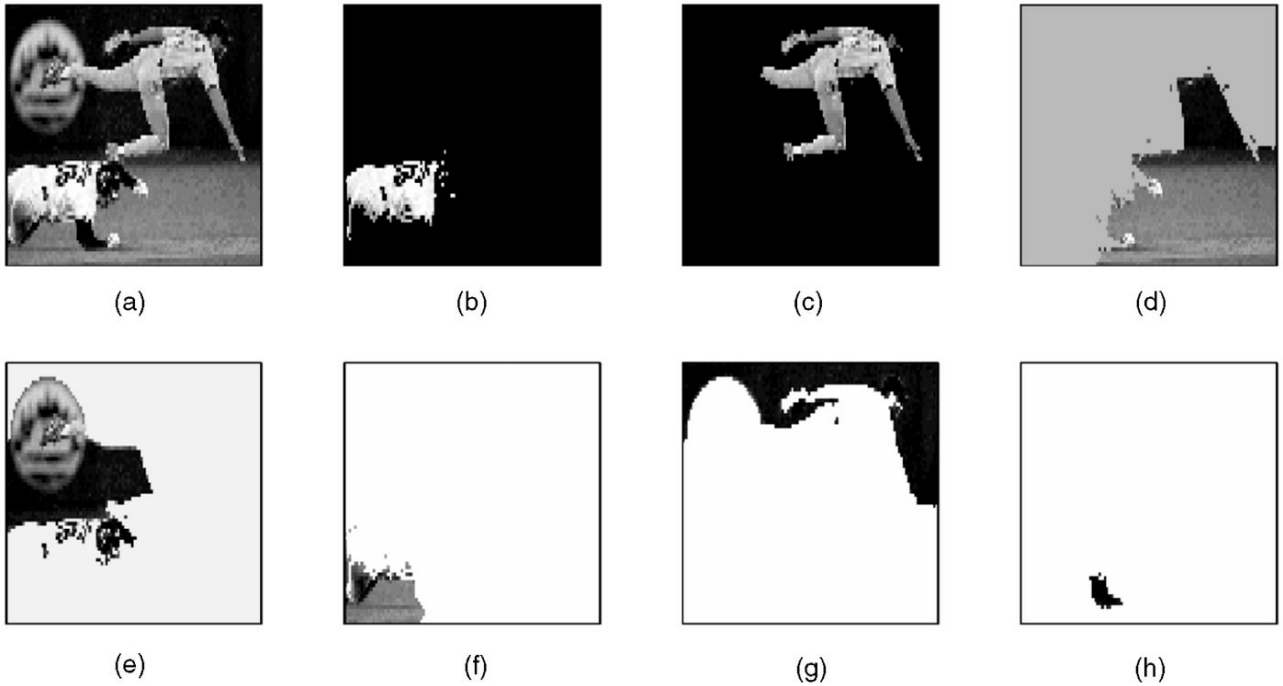


Figure: Normalized cuts detects 7 semantic groups.

Shi et al. “Normalized cuts and image segmentation”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000

Deep Spectral Methods

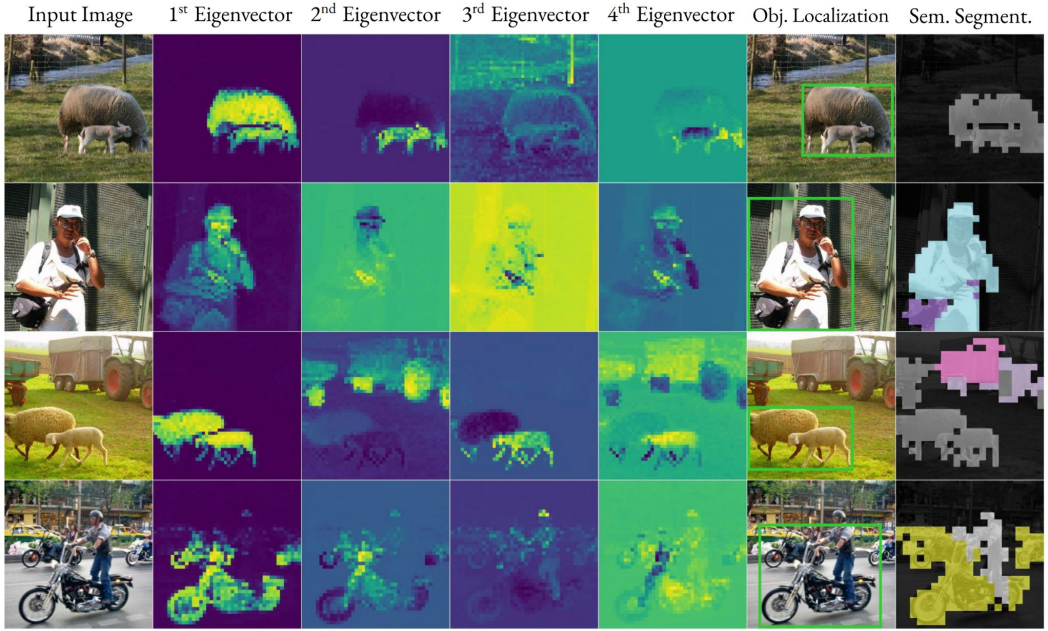


Figure: Spectral methods on color and deep features.

Melas-Kyriazi et al. “Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization” CVPR, 2022.

Cut-and-LEaRn (CutLER)

- **NormCut** produces bipartition of graph
- **Idea:** iteratively apply NormCut to masked similarity matrix to discover multiple objects / <u>object instances</u>

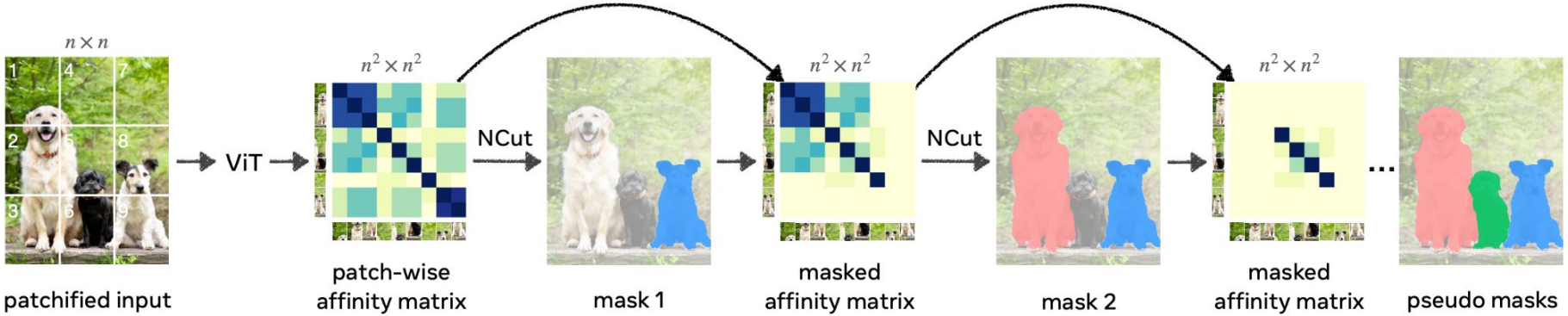


Figure: MaskCut of the CutLER algorithm uses NormCut on DINO features.

CutLER - Results

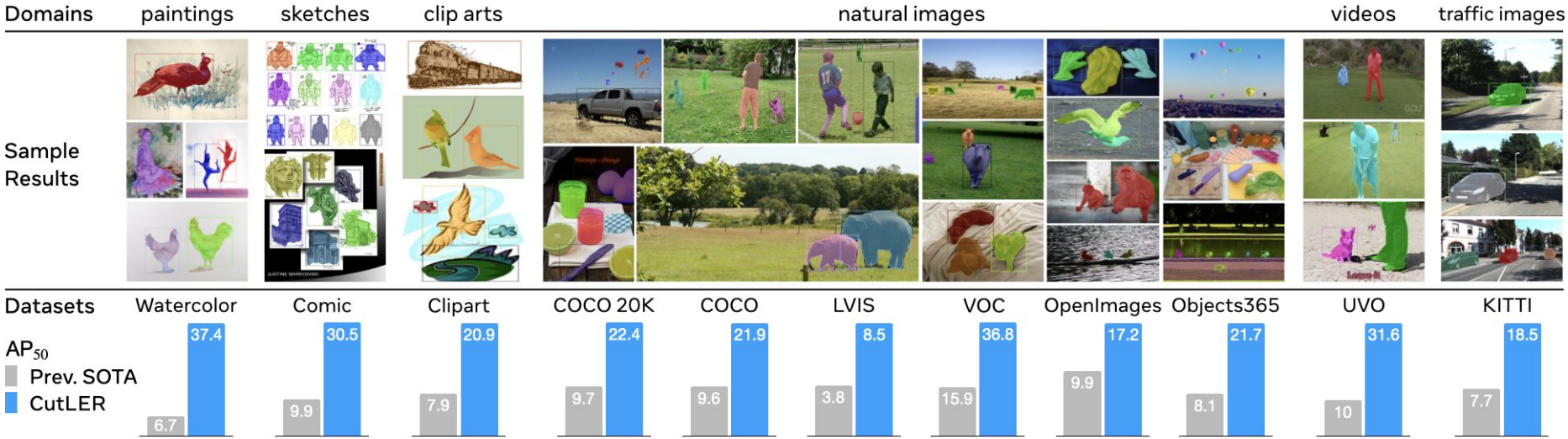


Figure: Zero-shot unsupervised object detection and instance segmentation. Comparison of CutLER [1] versus FreeSOLO [2].

[1] Wang et al. "Cut and Learn for Unsupervised Object Detection and Instance Segmentation" CVPR, 2023.
[2] Wang et al. "FreeSOLO: Learning To Segment Objects Without Annotations" CVPR, 2022.

Graph Cuts – Quantitative Results

Method	mIoU
Pretext task methods	
Co-Occurrence [40]	4.0
CMP [92]	4.3
Colorization [95]	4.9
Clustering/Contrastive methods	
IIC [41]	9.8
MaskContrast† [74]	35.0
Additional baselines	
Cluster-Patch	5.3
Cluster-Seg	12.1
Saliency-DINO-ViT-B†	30.1
MaskContrast-DINO-ViT-B†	31.2
Ours w/o self-training	30.8 ± 2.7
Ours	**37.2** ± 3.8

Table 4. **Semantic Segmentation on PASCAL VOC 2012.** We calculate the average and standard deviation of our method over four random seeds. † indicates methods using saliency networks that were initialized from supervised models.

Melas-Kyriazi et al. "Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization" CVPR, 2022.

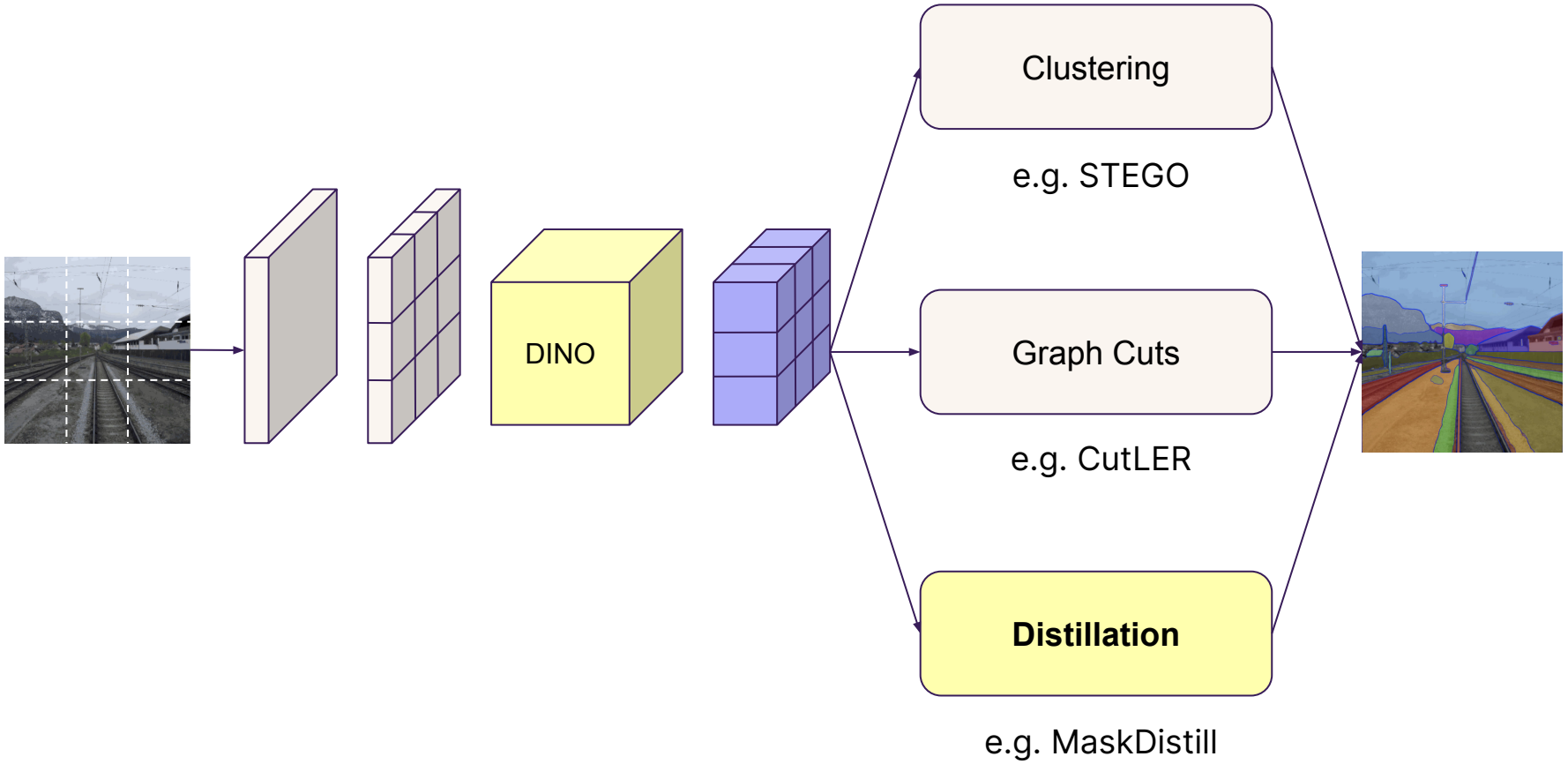
Instance Segmentation Results

Methods	Pretrain	Detector	Init.	COCO 20K						COCO val2017					
				AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}
non zero-shot methods															
LOST [38]	IN+COCO	FRCNN	DINO	-	-	-	2.4	1.0	1.1	-	-	-	-	-	-
MaskDistill [42]	IN+COCO	MRCNN	MoCo	-	-	-	6.8	2.1	2.9	-	-	-	-	-	-
FreeSOLO* [47]	IN+COCO	SOLOv2	DenseCL	9.7	3.2	4.1	9.7	3.4	4.3	9.6	3.1	4.2	9.4	3.3	4.3
zero-shot methods															
DETReg [3]	IN	DDETR	SwAV	-	-	-	-	-	-	3.1	0.6	1.0	8.8	1.9	3.3
DINO [7]	IN	-	DINO	1.7	0.1	0.3	-	-	-	-	-	-	-	-	-
TokenCut [50]	IN	-	DINO	-	-	-	-	-	-	5.8	2.8	3.0	4.8	1.9	2.4
CutLER (ours)	IN	MRCNN	DINO	21.8	11.1	10.1	18.6	9.0	8.0	21.3	11.1	10.2	18.0	8.9	7.9
CutLER (ours)	IN	Cascade	DINO	22.4	12.5	11.9	19.6	10.0	9.2	21.9	11.8	12.3	18.9	9.7	9.2
vs. prev. SOTA				**+12.7**	**+9.3**	**+7.8**	**+9.9**	**+6.6**	**+4.9**	**+12.3**	**+8.7**	**+8.1**	**+9.5**	**+6.4**	**+4.9**

Table 3. Unsupervised object detection and instance segmentation on COCO `20K` and COCO `val2017`. We report the detection and segmentation metrics and note the pretraining data (Pretrain), detectors, and backbone initialization (Init.). Methods in the top half of the table train on extra unlabeled images from the downstream datasets, while zero-shot methods in the bottom half only train on ImageNet. Despite using an older detector, CutLER outperforms all prior works on all evaluation metrics. *: results obtained with the official code and checkpoint. IN, Cascade, MRCNN, and FRCNN denote ImageNet, Cascade Mask R-CNN, Mask R-CNN, and Faster R-CNN, respectively.

Wang et al. "Cut and Learn for Unsupervised Object Detection and Instance Segmentation" CVPR, 2023.

Unsupervised Semantic Segmentation



MaskDistill

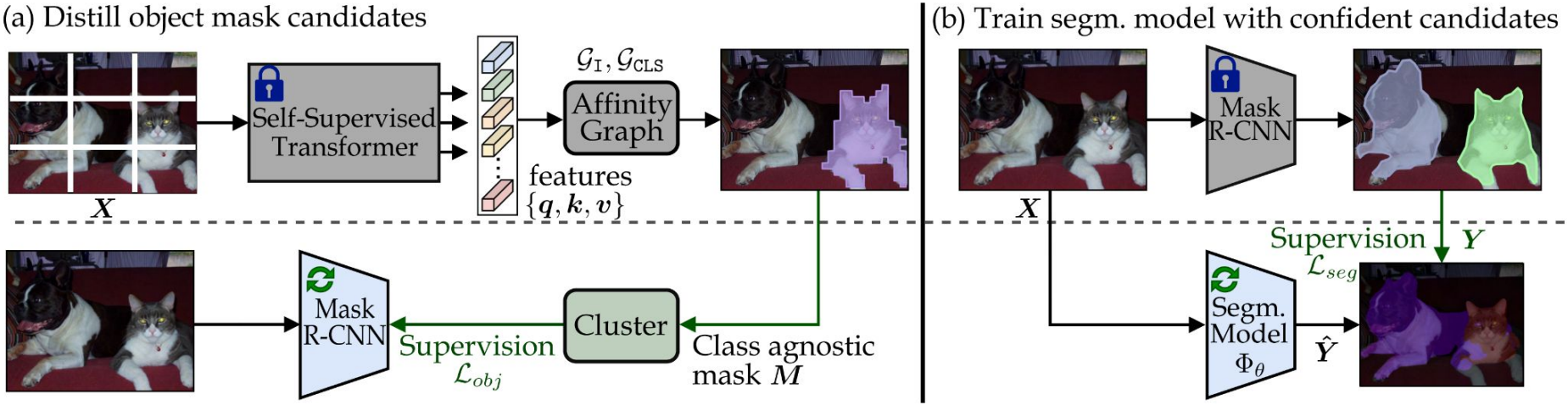


Figure: a) Extract object masks from DINO attention maps, cluster them, and use pseudo-ground-truth to train object mask predictor. b) Use only high-confidence predictions from fixed object mask predictor to generate targets for supervised segmentation model.

MaskDistill - Results

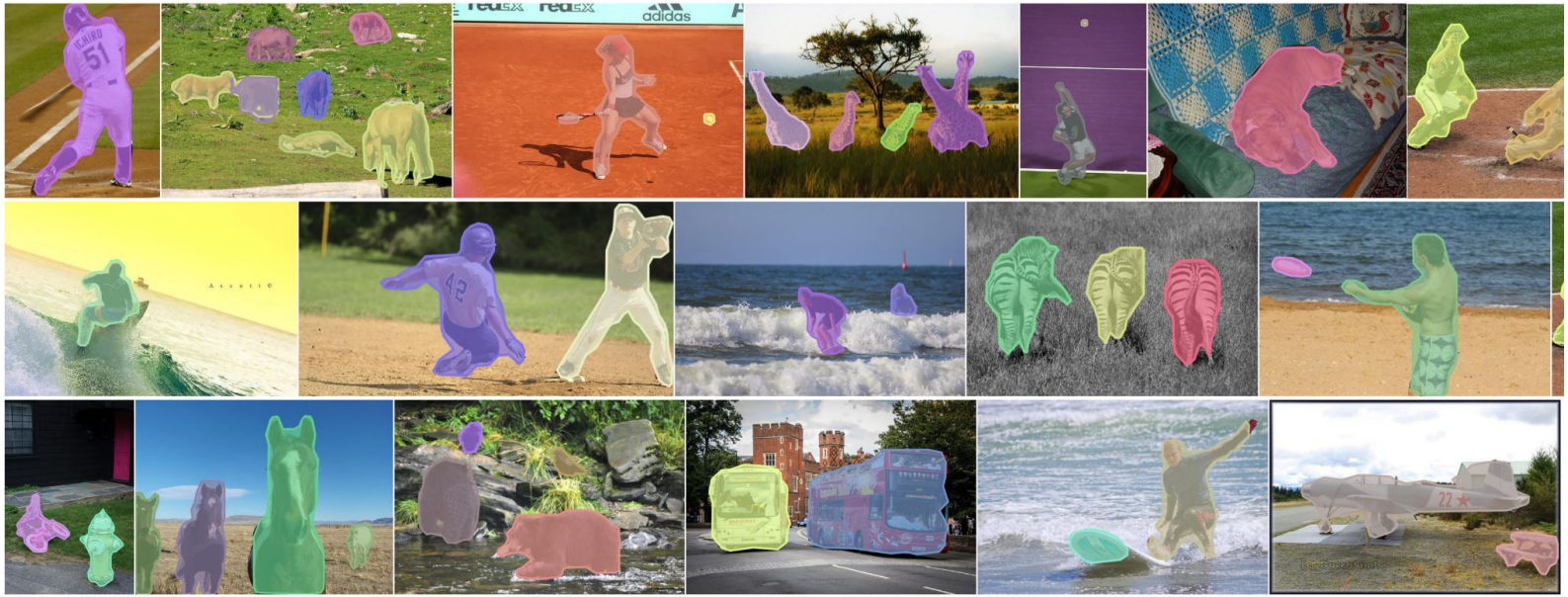


Figure: Instance segmentation on COCO with MaskDistill.

Van Gansbeke et al. “Discovering Object Masks with Transformers for Unsupervised Semantic Segmentation” arxiv, 2022.

Distillation – Quantitative Results

Table 2: SOTA comparison on PASCAL.

(a) Linear classifier. **(b)** Clustering.

Method	LC	Clustering
Proxy-tasks:		
Co-Occurence [42]	13.5	4.0
CMP [88]	16.5	4.3
Colorization [90]	25.5	4.9
Clustering:		
IIC [43]	28.0	9.8
Contrastive learning:		
Inst. Discr. [83]	26.8	4.3
MoCo [35]	45.0	3.7
InfoMin [72]	45.2	4.4
SwAV [11]	50.7	4.4
Handcrafted grouping priors:		
SegSort [40]†	36.2	-
Hierarch. Group. [91]†	48.8	-
MaskContrast [76]	58.4	35.0
MaskContrast [76]+CRF	59.5	-
MaskDistill	**58.7** (+0.3)	**45.8** (+10.8)
MaskDistill+CRF	**62.8** (+3.3)	**48.9** (+13.9)

[8] Van Gansbeke et al. "Discovering Object Masks with Transformers for Unsupervised Semantic Segmentation" arxiv, 2022.

(b) Unsup. Object Segmentation.

PASCAL VOC 2012 (mIoU)	
MaskContrast	35.0
COMUS	50.0
DINOSAUR	37.2 ±1.8

(c) Unsup. Scene Decomposition.

COCO-Stuff 27 (mIoU)	
SlotCon	18.3
STEGO	26.8
DINOSAUR	24.0 ±0.9

[12] Seitzer et al. "Bridging the Gap to Real-World Objectcentric Learning" ICLR, 2023.

Using COCO-81 here, most of others use COCO-Stuff-27

Table 3: Unsupervised semantic segmentation before and after self-learning evaluated by mIoU after Hungarian matching on the MS COCO *val* set. As discovered object category we count those categories with an IoU $> 20\%$ from all 81 categories. Also, we show IoU for categories that have corresponding cluster (i.e., with IoU larger than zero).

	all	discovered (with IoU≥ 20%)		have cluster (with IoU> 0%)	
	mIoU	number	mIoU	number	mIoU
Pseudo-masks	18.2	33	36.6	73	20.2
COMUS	19.6	34	40.7	60	26.5

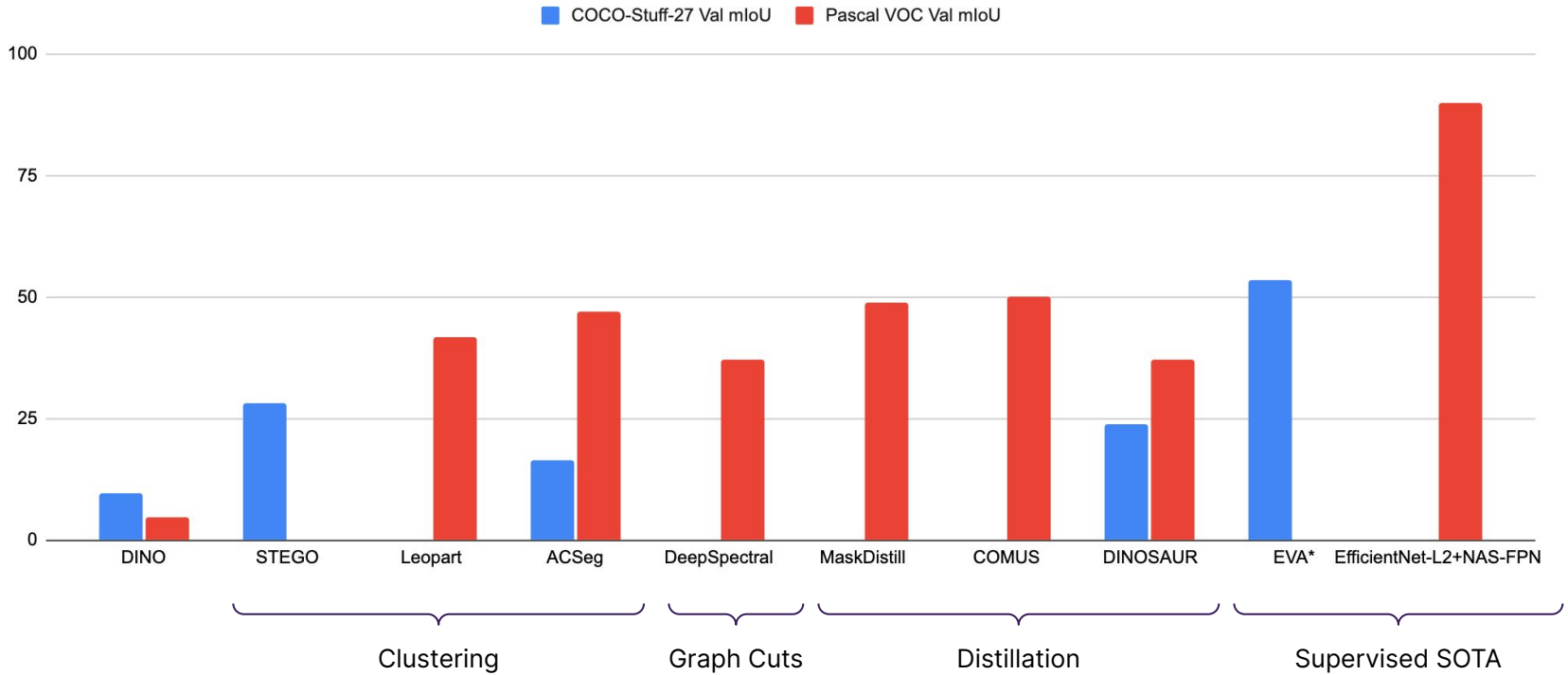
Table 1: Comparison to prior art and iterative improvement via self-training (evaluated by IoU after Hungarian matching) on the PASCAL 2012 *val* set. The results for SwAV and IIC methods are taken from MaskContrast paper. COMUS results are mean ± standard dev. over 5 runs.

Method	mIoU
Colorization (Zhang et al., 2016)	4.9
IIC (Ji et al., 2019)	9.8
SwAV (Caron et al., 2020)	4.4
MaskContrast (Van Gansbeke et al., 2021)	35.1
DeepSpectral (Melas-Kyriazi et al., 2022)	37.2 ± 3.8
DINOSAUR (Seitzer et al., 2023)	37.2 ± 1.8
Leopart (Ziegler & Asano, 2022)	41.7
Pseudo-masks (Iteration 0)	43.8 ± 0.1
COMUS (Iteration 1)	47.6 ± 0.4
COMUS (Iteration 2)	**50.0** ± **0.4**

[11] Zadaianchuk et al. "Unsupervised Semantic Segmentation with Self-supervised Object-centric Representations" ICLR, 2023.

SOTA Comparison

- Only methods included based on DINO
- Only models that evaluated mIoU scores (e.g. CutLER and MaskDistill do instance segmentation and evaluate AP)
- EVA * is evaluated on COCO-81 test (other algorithms usually evaluate with COCO-Stuff-27)



Future Directions

- Integrate advances in self-supervised learning (e.g., DINOv2 [1]) into existing methods
- Combat the curse of dimensionality: robustify data clustering in high dimensions and pre-process ViT features with dimensionality reduction (e.g., h-NNE [2], UMAP [3])
- Widen the image distribution: train a “biomedical” DINO ViT with various medical image types (X-Ray, Sonography, CT, …)
- Multimodal pre-training: leverage text supervision for “unsupervised” semantic segmentation (e.g., GroupViT [4], CLIPpy [5])
- Semi-supervised learning: efficient fine-tuning of pre-trained feature extractors with few labels
- Quantify zero-shot generalization abilities of supervised foundation models (e.g., Segment Anything [6]) to new domains such as biomedical imaging

[1] Oquab et al. “DINOv2: Learning Robust Visual Features without Supervision” arxiv, 2023.
[2] Sarfraz et al. “Hierarchical Nearest Neighbor Graph Embedding for Efficient Dimensionality Reduction” CVPR, 2022.
[3] McInnes et al. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction” arxiv, 2018.
[4] Ranasinghe et al. “Perceptual Grouping in Contrastive Vision-Language Models” ICCV, 2023.
[5] Xu et al. “GroupViT: Semantic Segmentation Emerges from Text Supervision” CVPR, 2022.
[6] Kirillov et al. “Segment Anything” arxiv, 2023.