Öz-Denetimli Öznitelikler Yardımıyla Videolarda Denetimsiz Nesne Keşfi

Nermin Samet

Agenda

- Self-supervised learning
    - SimCLR, DINO
- Self-supervised features for Unsupervised Object Discovery
    - TokenCut, LOST
- Self-supervised features for Unsupervised Object Discovery in Video

Self-Supervised Learning

- Learn representations from data without annotations
- Solve **pretext** tasks that obtain **good features** for downstream tasks
    - Labels for pretext are generated automatically
- Train with supervised learning objectives e.g. classification.

Pretext tasks for self-supervised learning

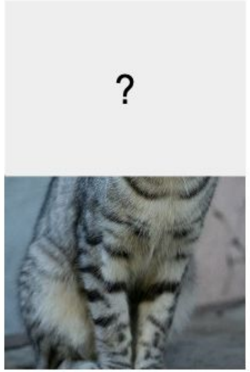
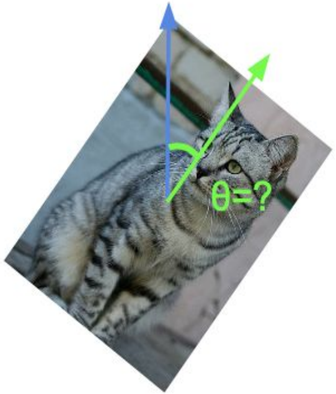


image completion



rotation prediction

“jigsaw puzzle”

colorization

Image Source

Problems with pretext tasks

Finding specific task oriented pretext is exhausting

Learned good features or representations may not generalize

Can we design a more general pretext task learning framework?

Contrastive Representation Learning

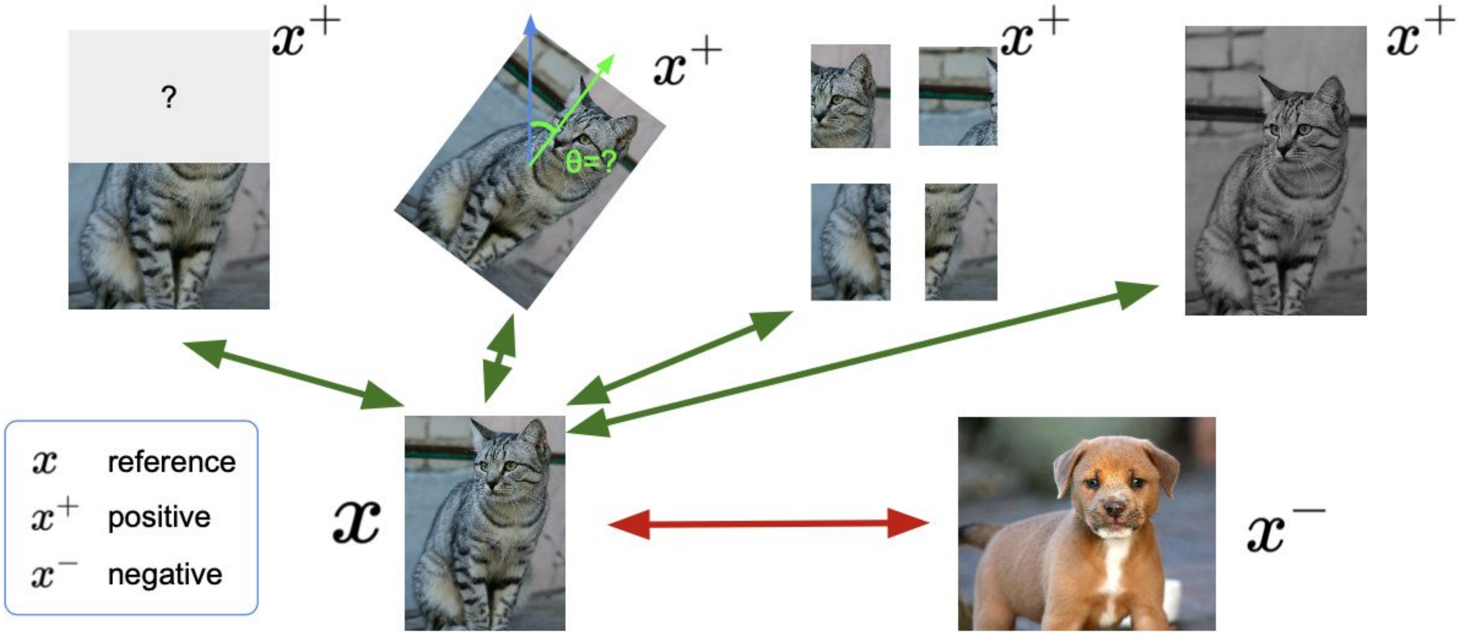


Image Source

SimCLR: A Simple Framework for Contrastive Learning of Visual Representations

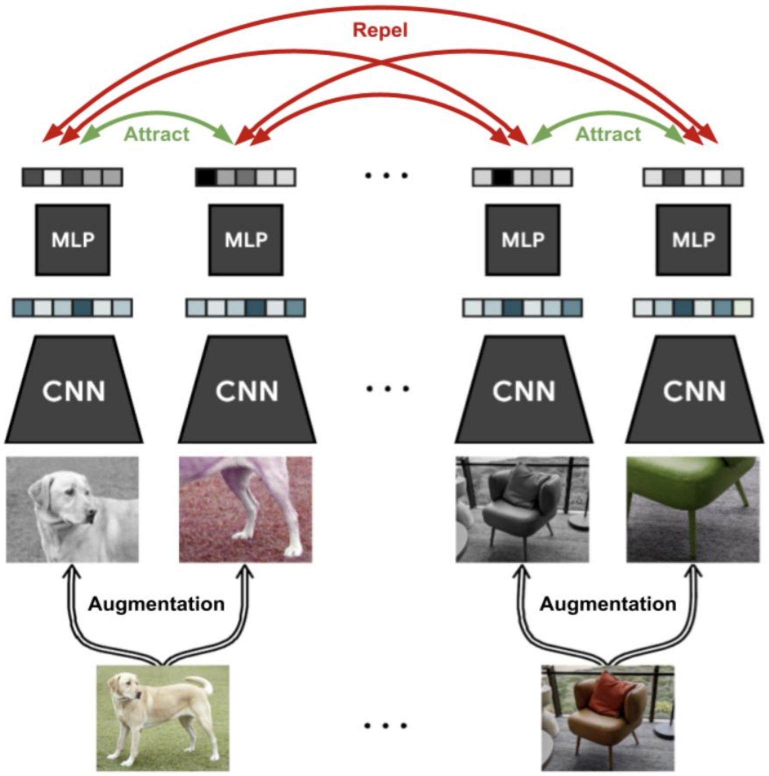


Image Source: https://simclr.github.io

Other Popular Contrastive Representation Learning

- MoCo: Momentum Contrast for Unsupervised Visual Representation Learning (CVPR’20)
- MoCo V2: Improved Baselines with Momentum Contrastive Learning (arXiv)
- CPC: Representation Learning with Contrastive Predictive Coding (arXiv)
- CLIP: Contrastive Language–Image Pre-training (ICML’21)

DINO: Self-**di**stillation with **no** labels

Emerging Properties in Self-Supervised Vision Transformers (ICCV'21)

- Utilize Vision Transformers (ViT)
- No contrastive learning i.e. no negative samples
- Two type of data crop augmentation:
  - Global crops: >50% of the image
  - Local crops: <50% of the image

DINO: Self-**di**stillation with **no** labels

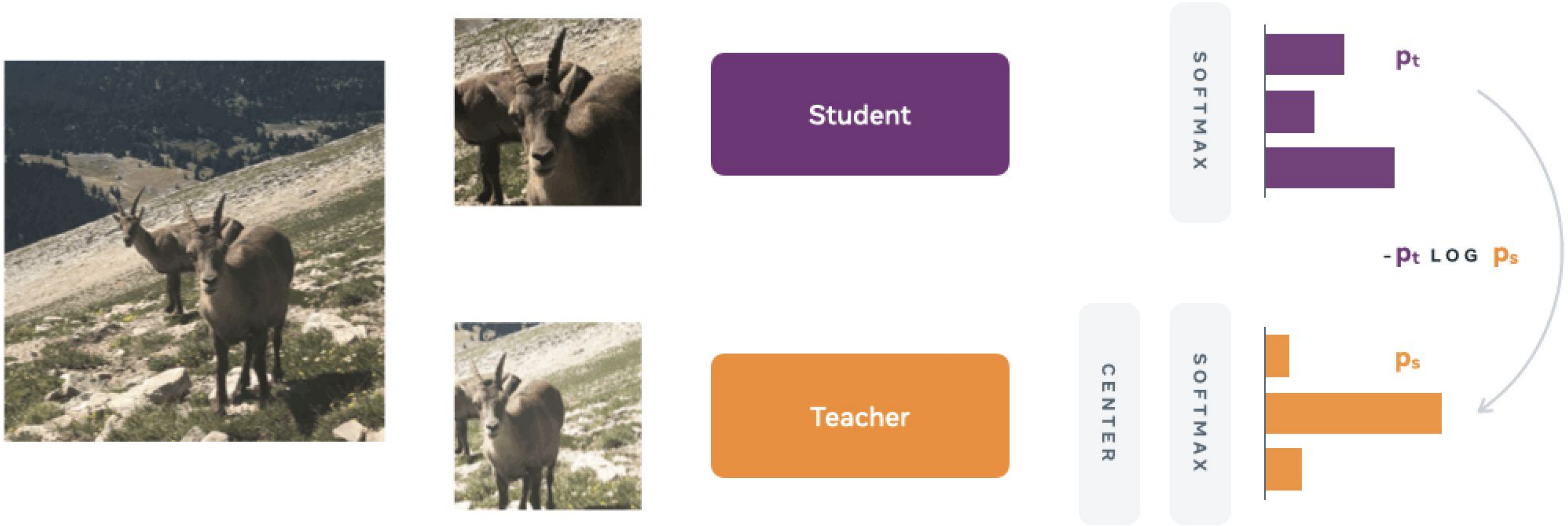
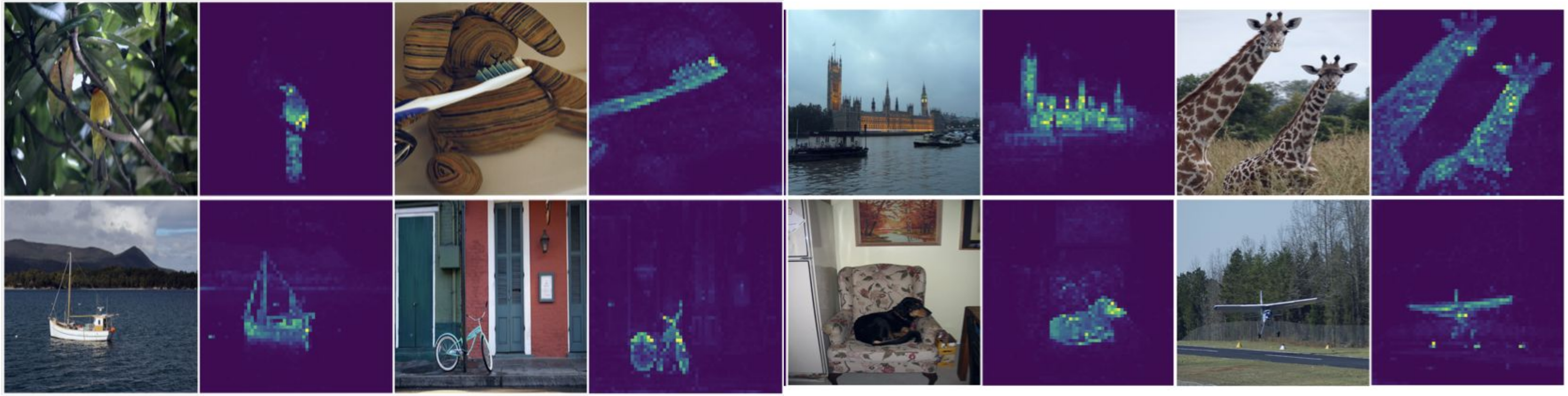


Image Source: [DINO]

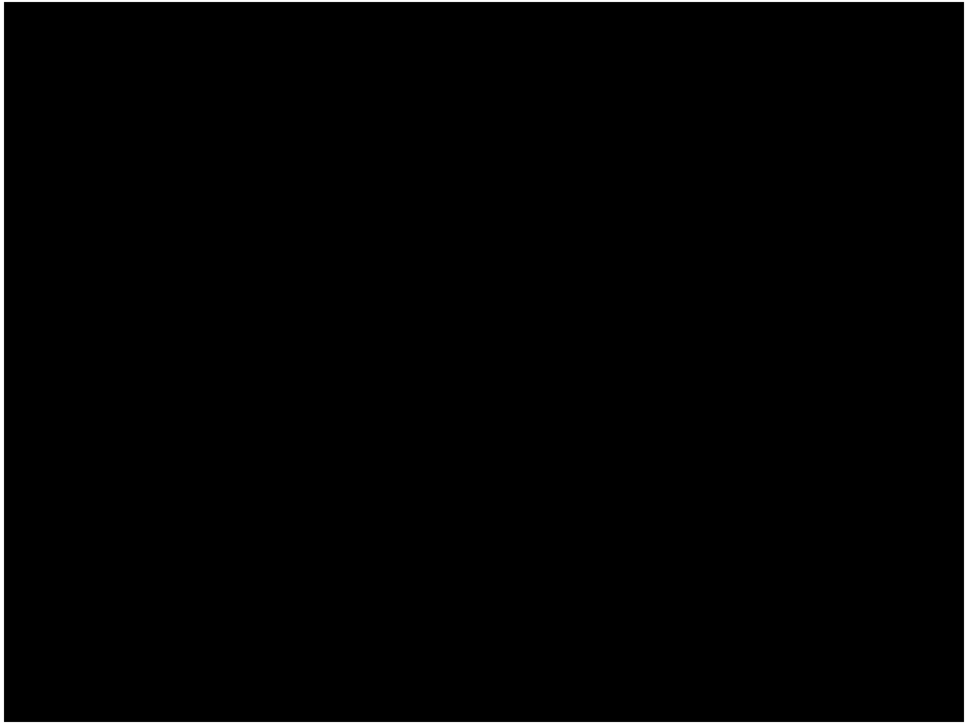
Self-attention visualizations



k-NN classification on ImageNet with a small ViT achieves 78.3% top-1 accuracy

Image Source: [DINO]

Self-attention visualizations in videos



Self-supervised features for Unsupervised Object Discovery

- **LOST:** Localizing Objects with Self-Supervised Transformers and no Labels, BMVC’21
- **TokenCuT**: Self-supervised Transformers for Unsupervised Object Discovery using Normalized Cut, CVPR’22

LOST: Localizing Objects with Self-Supervised Transformers and no Labels

Seed Expansion based on patch correlations within an image

- Divide the image into equal-sized patches
- Extract high-quality self-supervised DINO features for each patch
- Select a patch as a seed based on several criteria - a salient one
- Expand the seed adding other highly correlated patches

LOST: Localizing Objects with Self-Supervised Transformers and no Labels

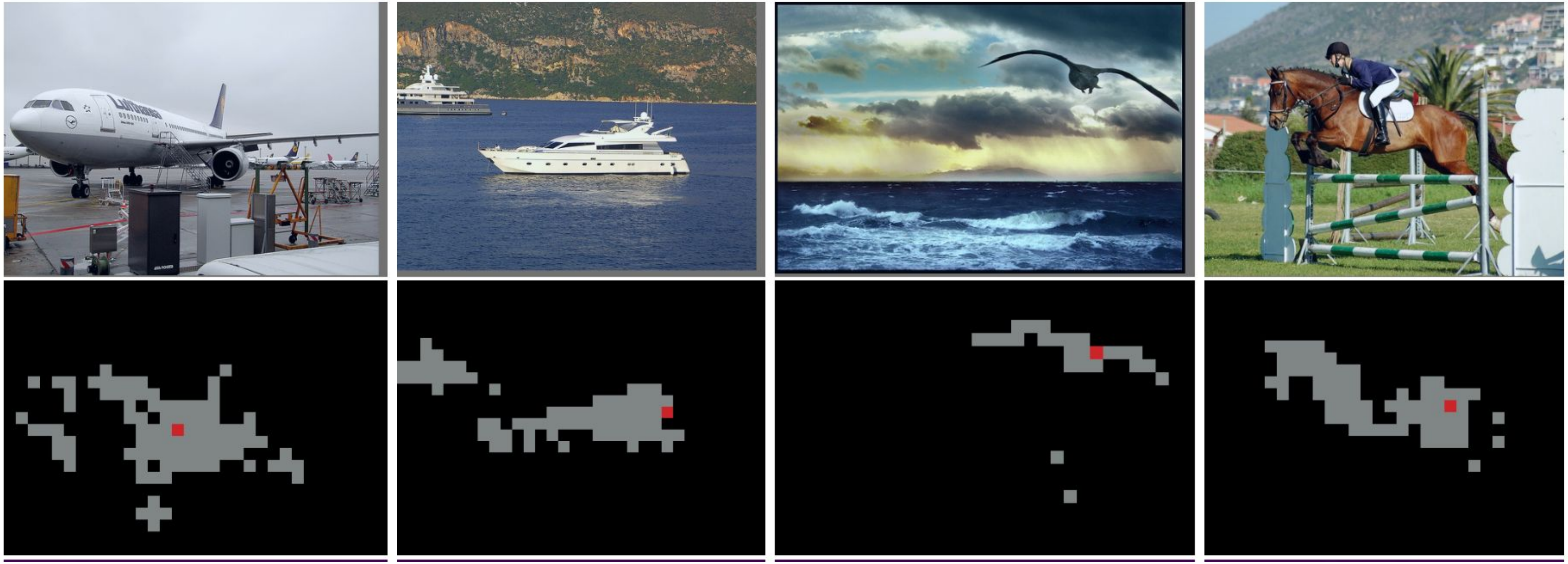


Image Source: [LOST]

TokenCuT: Self-supervised Transformers for Unsupervised Object Discovery using Normalized Cut

- Builds a weighted graph
  - visual tokens of ViT are nodes a
  - similarity scores between tokens are edges
- The segmentation problem is :
  - formulated as a normalized graph cut
  - solved using spectral clustering with eigen decomposition.

TokenCuT: Self-supervised Transformers for Unsupervised Object Discovery using Normalized Cut

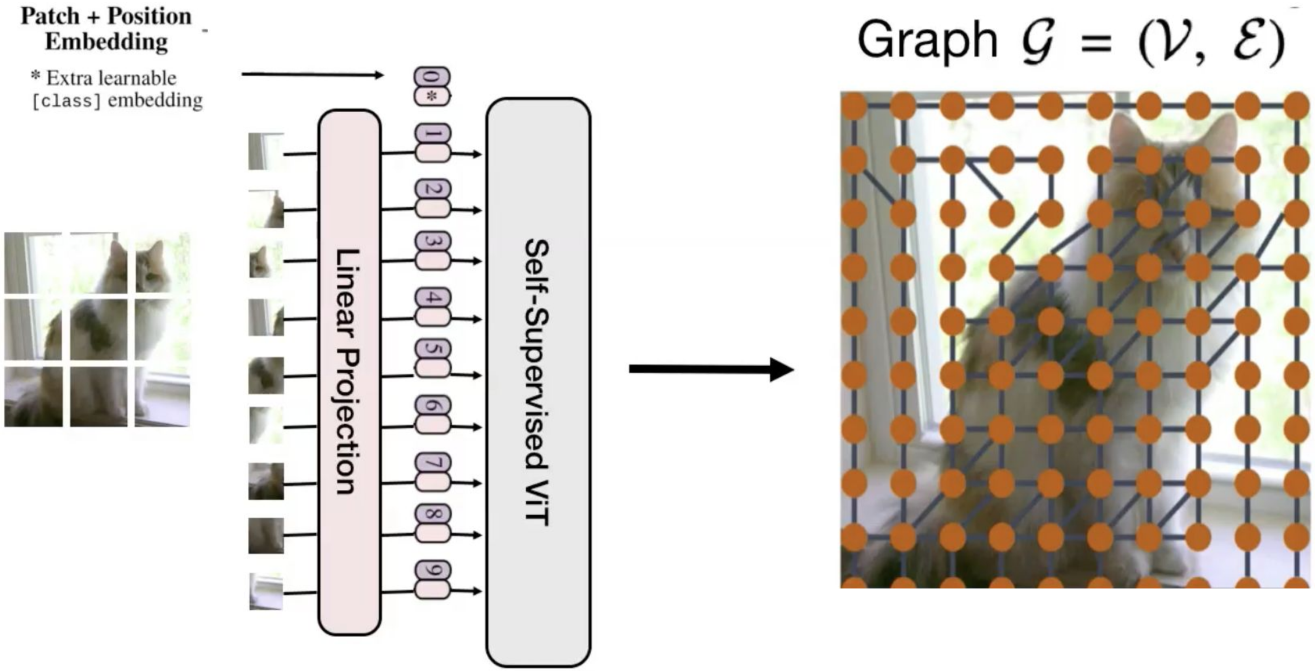
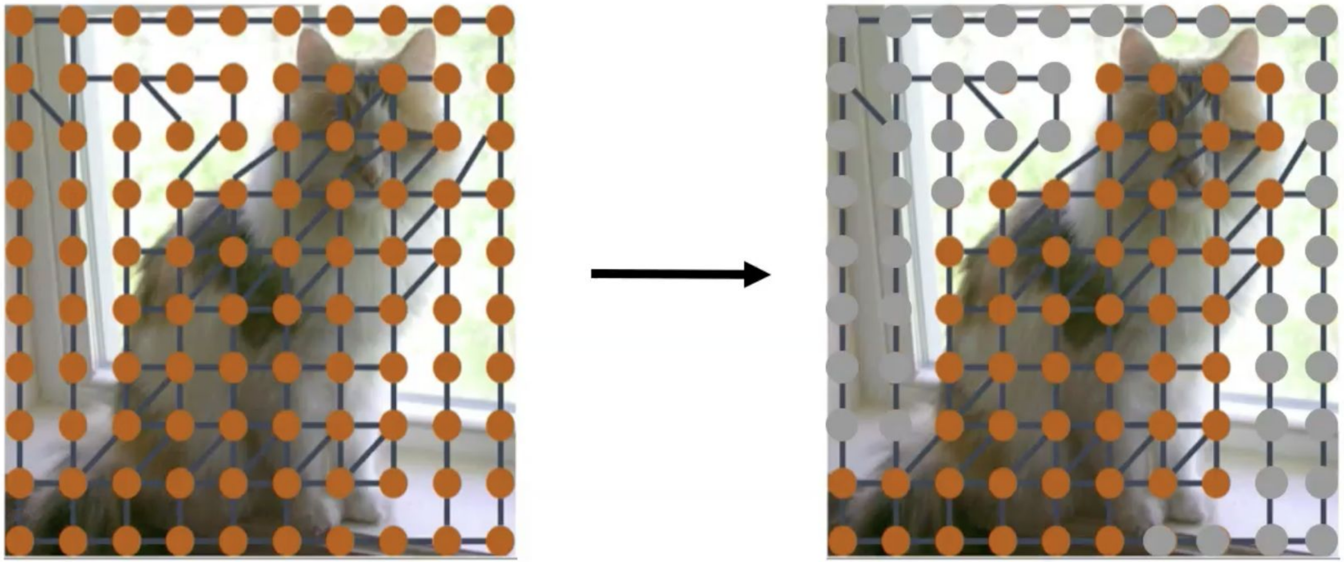


Image Source: [TokenCut]

TokenCut: Graph Bipartion with Normalized Cut*



*Shi, Jianbo, and Jitendra Malik. "Normalized cuts and image segmentation. TPAMI, 2000

Image Source: TokenCut

TokenCut: Graph Bipartion with Normalized Cut

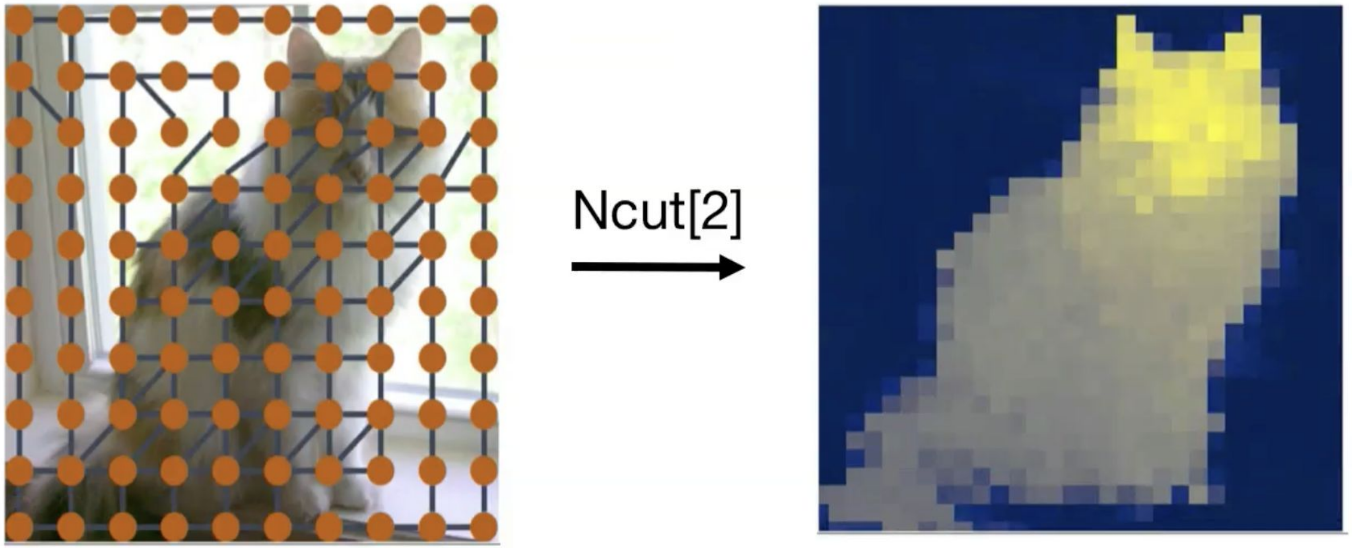


Image Source: TokenCut

TokenCut:Foreground/Background Segmentation

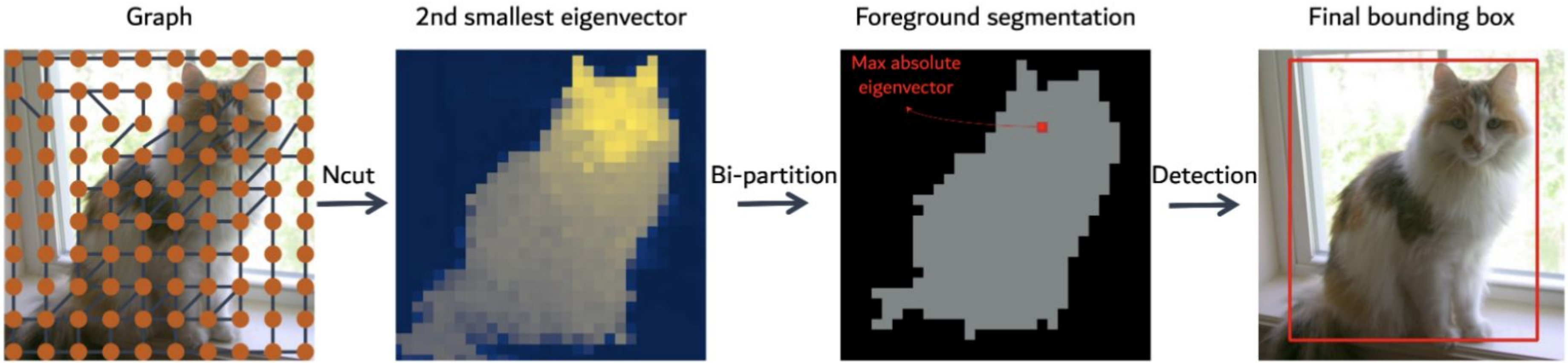


Image Source: [TokenCut]

Self-supervised features for Unsupervised Object Discovery in Videos

A Simple and Powerful Global Optimization for Unsupervised Video Object Segmentation

Georgy Ponimatkin[1] Nermin Samet[1] Yang Xiao[1] Yuming Du[1] Renaud Marlet[1,2] Vincent Lepetit[1]

[1]LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France [2]Valeo.ai, Paris, France

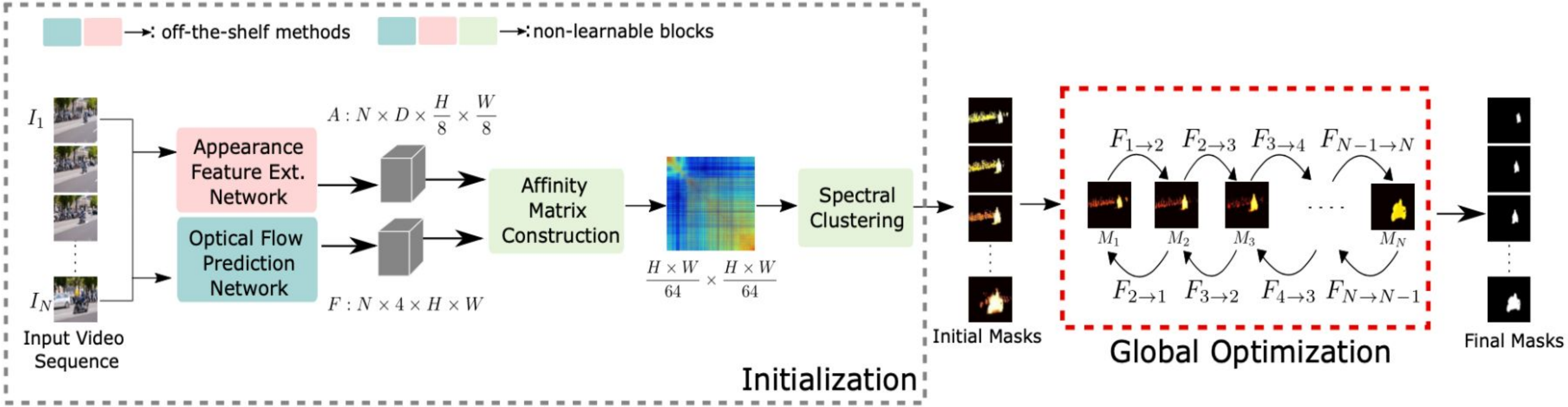
georgy.ponimatkin@enpc.fr

Code and supplementary material: https://ponimatkin.github.io/ssl-vos

A Simple and Powerful Global Optimization for Unsupervised Video Object Segmentation

- No training
- Fully unsupervised method
- No hyper-parameter tuning
- Generalizes well to new datasets without training
- SOTA results for unsupervised video object detection on two-benchmarks

Global Optimization for Unsupervised Video Object Segmentation



Objective Function

X_p optimized mask

$\tilde{X}_p$ initial mask

w warping function

CE cross-entropy

keep the masks close to their initialization

$$\mathcal{L}(\{\mathbf{x}_p\}_p) = \sum_p \lambda \, \mathrm{CE}(\widehat{\mathbf{x}}_p, \mathbf{x}_p) + \mathrm{CE}(\mathbf{x}_{p+1}, w_p^{p+1}(\mathbf{x}_p)) + \mathrm{CE}(\mathbf{x}_p, w_{p+1}^{p}(\mathbf{x}_{p+1})) \, ,$$

Forward flow consistency: warped the **mask p** by the optical flow from p to p + 1, and align it with the mask of frame p + 1

Backward flow consistency: warped the **mask p+1** by the optical flow from p+1 to p, and align it with the mask of frame p

A Simple and Powerful Global Optimization for Unsupervised Video Object Segmentation

Segmentation obtained by our spectral clustering formulation on the self-supervised image features from DINO in a **single image**

Segmentation obtained by our spectral clustering formulation using DINO and optical flow features from ARFlow from a **single image**.

Final segmentation masks with our complete method:

After optimization **on the full frame sequence**, using DINO and ARFlow features.

Vanilla Spectral Clustering for Images

Affinity Matrix A:

- Each row and each column of A corresponds to an image location (HW x HW).
- A_{ij} shows how likely image location for row i and image location for column j belong to the same cluster.

Spectral clustering:

- A good segmentation mask can be obtained from the second largest eigenvector $\mathbf{X_2}$ of the normalized affinity matrix W: $\mathbf{W = D^{-1}A}$ where $\mathbf{D = diag(\{\sum_j A_{ij}\}_i)}$
- The coefficients of $\mathbf{X_2}$ correspond to image locations.
- By thresholding **X**, we get a binary mask segmented map

Extending Vanilla Spectral Clustering for Videos

Affinity Matrix A:

- Each row and each column of A corresponds to an image location in a video frame. NHW x NHW
- $\mathbf{A}_{\mathbf{(pi)(qj)}}$ **:** (pi) is a row corresponding to image location i in frame p, (qj) is a column corresponding to image location j in frame q
- $\mathbf{A}_{\mathbf{(pi)(qj)}}$ should express how likely image location i in frame p and image location j in frame q are to correspond to the same object.
- This is a very large A matrix and W matrix!
    - In the order of $10^9 \times 10^9$ for typical sequences in standard benchmarks.

Too large for storage and for eigenvector computation!

Solution

A new scalable spectral clustering formulation;

- We construct a tri-diagonal global affinity matrix
- We approximate eigenvectors by PIC* for each frame p independently.

Our spectral clustering: 0.1s/frame

Full eigenvector decomposition (from TokeCut): 17s/frame

*Frank Lin and William W. Cohen. Power Iteration Clus- tering. In *International Conference on Machine Learning (ICML)*, 2010

Tri-dioganal Global Affinity Matrix for Video Object Segmentation

$$\mathcal{A} = \begin{bmatrix} A_1 & \overline{F}_1 & 0 & \dots & 0 \\ F_1 & A_2 & \overline{F}_2 & \ddots & \vdots \\ 0 & F_2 & A_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \overline{F}_{N-1} \\ 0 & \dots & 0 & F_{N-1} & A_N \end{bmatrix}$$

Each element is a HW ×HW matrix.

Experiments

Datasets

- **DAVIS2016:** 50 sequences; 3455 frames in total
- **SegTrack-v2:** 14 sequences; 976 frames in total
- **FBMS59:** 59 sequences; 720 frames in total.

Metrics

- **Jaccard (J)**: the intersection-over- union between predicted masks and ground-truth masks
- **Contour accuracy (F)**
- **Computational cost:** On an Nvidia V100;
  - our initialization takes ~ 0.1 s/frame
  - optimization takes ~ 0.4 s/frame.

Comparison to the State of the Art

Method	Training on videos	Optical flow	Fully unsuperv.	DAVIS $\mathcal{J}\uparrow$	DAVIS $\mathcal{F}\uparrow$	STv2 $\mathcal{J}\uparrow$	FBMS59 $\mathcal{J}\uparrow$
(Mostly) Unsupervised methods:							
CUT [38]		LDOF [3]	✓	55.2	55.2	54.3	57.2
FTS [57]		LDOF [72]	✓	55.8	51.1	47.8	47.7
AMD [47]	✓		✓	57.8	-	57.0	47.5
MoSeg [87]	✓	RAFT [73]		68.3	61.1	58.6	53.1
CIS [90]	✓	PWCNet [71]		71.5	70.5	62.0	63.5
DS [92]	✓	RAFT [73]		79.1	-	72.1	71.8
DyStaB [89]	✓	PWCNet [71]		80.0	-	74.2	**73.2**
Ours		ARFlow [46]	✓	**80.2**	**77.5**	**74.9**	70.0

Ablation Experiments - I

(a) Different appearance features	($\mathcal{J}\uparrow$)	($\mathcal{F}\uparrow$)
(BT [RN-50] + ARFlow)	53.5	30.3
(BT [RN-50] + ARFlow) + Optimization (Eq. (1))	59.5 (**+6.0**)	48.2 (**+17.9**)
(SWAV [RN-50] + ARFlow)	48.0	27.2
(SWAV [RN-50] + ARFlow) + Optimization (Eq. (1))	53.6 (**+5.6**)	46.0 (**+18.8**)
(MoCo-v3 [ViT] + ARFlow)	58.0	35.3
(MoCo-v3 [ViT] + ARFlow) + Optimization (Eq. (1))	64.2 (**+6.2**)	61.1 (**+25.8**)
(DINO [ViT] + ARFlow)	72.1	72.5
(DINO [ViT] + ARFlow) + Optimization (Eq. (1))	76.8 (**+4.7**)	77.0 (**+4.5**)

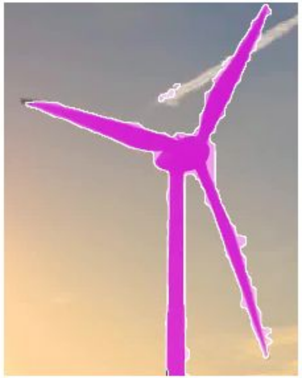
Ablation Experiments - II

(b) Different optical flows	($\mathcal{J}\uparrow$)	($\mathcal{F}\uparrow$)
DINO	61.2	65.8
TokenCut [82]	62.7	62.3
DINO + Optimization (Eq. (1))	66.7 (**+5.5**)	70.4 (**+4.6**)
(DINO + RAFT)	70.7	72.9
(DINO + RAFT) + Optimization (Eq. (1))	75.3 (**+4.6**)	76.2 (**+3.3**)
(DINO + ARFlow)	72.1	72.5
(DINO + ARFlow) + Optimization (Eq. (1))	76.8 (**+4.7**)	77.0 (**+4.5**)

Generalization to new datasets

MoSeg[*] performs poorly on a new sequence.

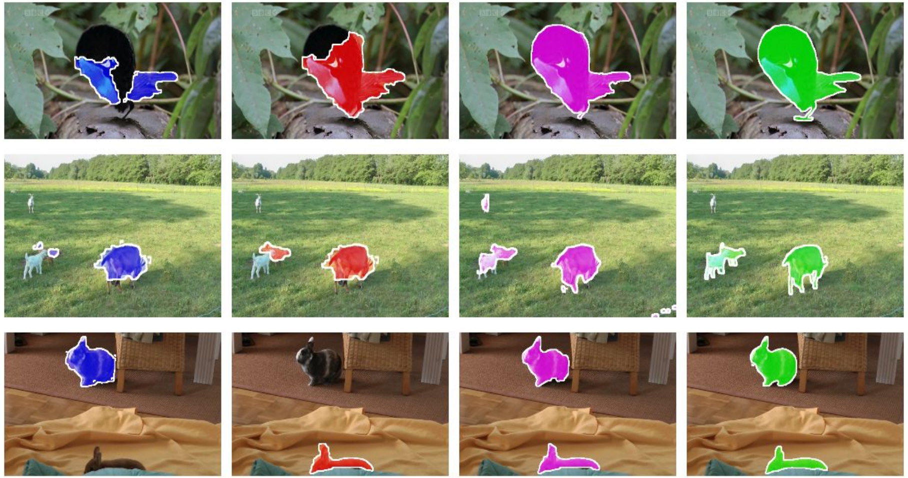
Spectral Clustering performs significantly better



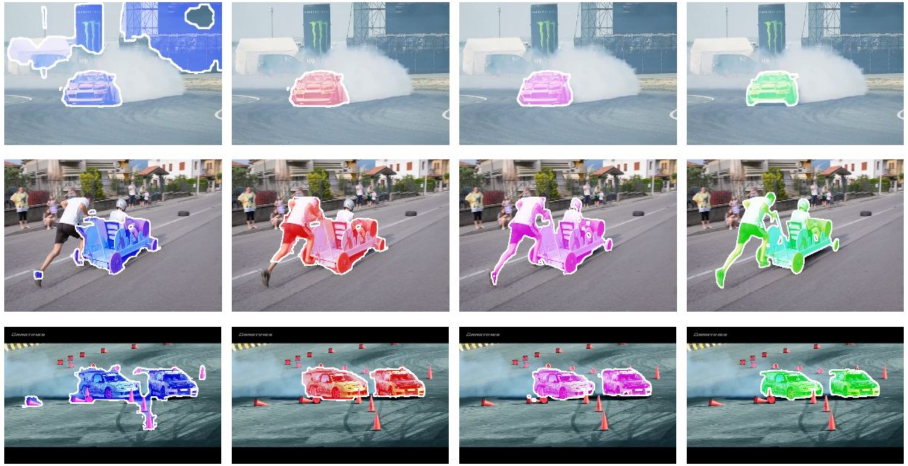
Result of our method.

* Charig Yang, Hala Lamdouar, Erika Lu,Andrew Zisserman, and Weidi Xie. Self-Supervised Video Object Segmentation by Motion Grouping. *ICCV*, 2021.

Qualitative Result

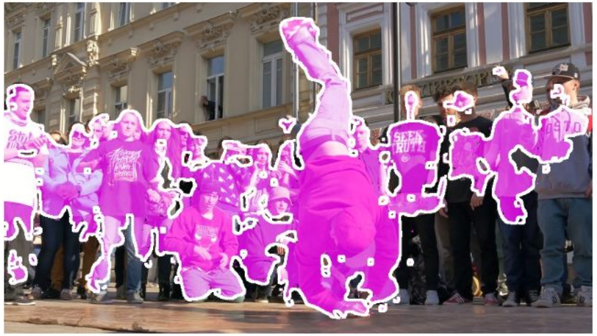


DINO | DINO + ARFlow | DINO + ARFlow + Opt | Ground Truth



DINO | DINO + ARFlow | DINO + ARFlow + Opt | Ground Truth

Failure Cases



over-segmentation

missing objects under occlusion

under-segmentation

G. Ponimatkin, N. Samet, Y. Xiao, Y. Du, R. Marlet and V. Lepetit "A Simple and Powerful Global Optimization for Unsupervised Video Object Segmentation", In IEEE Winter Conference on Applications of Computer Vision (WACV), 2023.

Github: https://github.com/ponimatkin/ssl-vos