



HAL
open science

TokenCut: Segmenting Objects in Images and Videos with Self-supervised Transformer and Normalized Cut

Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu,
James L Crowley, Dominique Vaufreydaz

► **To cite this version:**

Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, et al.. TokenCut: Segmenting Objects in Images and Videos with Self-supervised Transformer and Normalized Cut. 2022. hal-03765422v1

HAL Id: hal-03765422

<https://hal.science/hal-03765422v1>

Preprint submitted on 31 Aug 2022 (v1), last revised 30 Nov 2023 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TOKENCUT: SEGMENTING OBJECTS IN IMAGES AND VIDEOS WITH SELF-SUPERVISED TRANSFORMER AND NORMALIZED CUT

AUTHOR VERSION

Yangtao Wang^{1*}, Xi Shen^{2*}, Yuan Yuan³, Yuming Du⁴, Maomao Li², Shell Xu Hu⁵
James L. Crowley¹, Dominique Vaufreydaz¹

¹ Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

² Tencent AI Lab ³ MIT CSAIL ⁴ LIGM (UMR 8049) - Ecole des Ponts, UPE

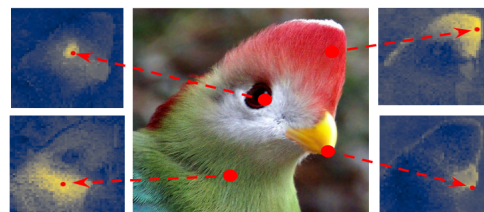
⁵ Samsung AI Center, Cambridge

Abstract

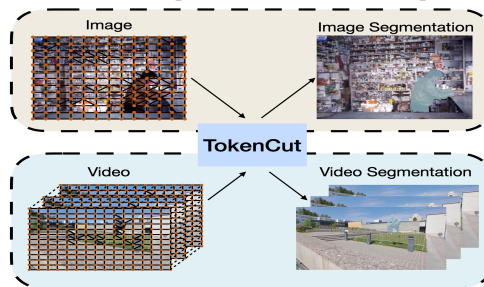
In this paper, we describe a graph-based algorithm that uses the features obtained by a self-supervised transformer to detect and segment salient objects in images and videos. With this approach, the image patches that compose an image or video are organised into a fully connected graph, where the edge between each pair of patches is labeled with a similarity score between patches using features learned by the transformer. Detection and segmentation of salient objects is then formulated as a graph-cut problem and solved using the classical Normalized Cut algorithm. Despite the simplicity of this approach, it achieves state-of-the-art results on several common image and video detection and segmentation tasks. For unsupervised object discovery, this approach outperforms the competing approaches by a margin of 6.1%, 5.7%, and 2.6%, respectively, when tested with the VOC07, VOC12, and COCO20K datasets. For the unsupervised saliency detection task in images, this method improves the score for Intersection over Union (IoU) by 4.4%, 5.6% and 5.2%. When tested with the ECSSD, DUTS, and DUT-OMRON datasets, respectively, compared to current state-of-the-art techniques. This method also achieves competitive results for unsupervised video object segmentation tasks with the DAVIS, SegTV2, and FBMS datasets. Our implementation is available at <https://www.m-psi.fr/Papers/TokenCut2022/>.

1 Introduction

Detecting and segmenting salient objects in an image or video are fundamental problems in computer vision with applications in real-world vision systems for robotics, autonomous driving, traffic monitoring, manufacturing, and embodied artificial intelligence [19, 67, 68]. How-



(a) Attention maps associated to different patches



(b) A unified method for image and video segmentation.

Figure 1: Attention maps associated with different patches highlight different regions of the object (Fig. 1a), which motivates us to build a unified graph-based solution for unsupervised image and video segmentation (Fig. 1b).

ever, current approaches rely on supervised learning requiring large data sets of high-quality, annotated training data [35]. The high cost of this approach becomes even more apparent when using transfer learning to adapt a pre-trained object detector to a new application domain. Researchers have attempted to overcome this barrier using active learning [1, 48], semi-supervised learning [8, 37], and weakly-supervised learning [27, 44, 76, 77] with limited results. In this paper, we report on results of an effort to use features provided by transformers trained with self-supervised learning, obviating the need for expensive annotated training data.

*Corresponding Author

Vision transformers trained with self-supervised learning [6, 15], such as DINO [6] and MAE [18] have been shown to outperform supervised training on downstream tasks. In particular, the attention maps associated with patches typically contain meaningful semantic information (Fig. 1a). For example, experiments with DINO [6] indicate that the attention maps of the class token highlight salient object regions. However, such attention maps are noisy and cannot be directly used to detect or segment objects.

The authors of LOST [49] have shown that the learned features from DINO can be used to build a graph and segment objects using the inverse degrees of nodes. Specifically, LOST employs a heuristic seed expansion strategy to accommodate noise and detect a single bounding box for a foreground object. We have investigated whether such learned features can be used with a graph-based approach to detect and segment salient objects in images and videos (Fig. 1b), formulating the segmentation problem using the classic normalised cut algorithm (Ncut) [46].

In this paper we describe TokenCut, a unified graph-based approach for image and video segmentation using features provided by self-supervised learning. The processing pipeline for this approach, illustrated in Fig. 2, is composed of three steps: 1) graph construction, 2) graph cut, 3) edge refinement. In the graph construction step, the algorithm uses image patches as nodes and uses features provided by self-supervised learning to describe the similarity between pairs of nodes. For images, edges are labeled with a score for the similarity of patches based on learned features for RGB appearance. For videos, edge labels combine similarities of learned features for RGB appearance and optical flow.

To cut the graph, we rely on the classic normalized cut (Ncut) algorithm to group self-similar regions and delimit the salient objects. We solve the graph-cut problem using spectral clustering with generalized eigen-decomposition. The second smallest eigenvector provides a cutting solution indicating the likelihood that a token belongs to a foreground object, which allows us to design a simple post-processing to obtain a foreground mask. We also show that standard algorithms for edge-aware refinement, such as Conditional Random Field [30] (CRF) and Bilateral Solver [5] (BS) can be used to refine the masks for detailed object boundary detection. This approach can be considered as a run-time adaptation method, because the model can be used to process an image or video without the need to retrain the model.

Despite its simplicity, TokenCut significantly improves unsupervised saliency detection in images. Specifically, it achieves 77.7%, 62.8%, 61.9% mIoU on the ECSSD [47], DUTS [64] and DUT-OMRON [72] respectively, and outperforms the previous state-of-the-art by a margin of 4.4%, 5.6% and 5.2%. For unsupervised video segmentation, TokenCut achieves competitive results on DAVIS [41], FBMS [40], SegTV2 [32]. Additionally, TokenCut also obtains important improvement on unsuper-

vised object discovery. For example, TokenCut outperforms DSS [38], which is a concurrent work, by a margin of 6.1%, 5.7%, and 2.6% respectively on the VOC07 [16], VOC12 [17], COCO20K [35].

In summary, the main contributions of this paper are as follows:

- We describe TokenCut, a simple and unified approach to segment objects in images and videos without using any human annotations.*
- We show that TokenCut significantly outperforms previous state-of-the-art methods unsupervised saliency detection and unsupervised object discovery on images. As a training-free method, TokenCut achieves competitive performance on unsupervised video segmentation compared to the state-of-the-art methods.
- We provide a detailed analysis on the TokenCut to validate the design of the proposed approach.

2 Related Work

Self-supervised vision transformers. ViT [15] has shown that the transformer architecture [58] can be effective for computer vision tasks using supervised learning. Recently, many variants of ViT have been proposed to learn image encoders in a self-supervised manner. MoCo-V3 [7] demonstrates that using contrastive learning on ViT can achieve strong results. DINO [6] proposes to train transformers with self-distillation loss [21] and shows that the features learn through ViT contain explicit information useful for image semantic segmentation. Inspired by BERT, several approaches [4, 14, 20, 34] mask some tokens from the input and learn to recover the missing tokens in the output.

Unsupervised object discovery. Given a group of images, unsupervised object discovery seeks to discover and delimit similar objects that appear in multiple images. Early research [9, 22, 25, 26, 59] formulated the problem as a problem using an assumption on the frequency of object occurrences. Other researchers formulated object detection as an optimization problem over bounding box proposals [11, 53, 60, 61] or as a ranking problem [62]. Recently, LOST [49] significantly improved the state-of-the-art for unsupervised object discovery. LOST extracts features using a self-supervised transformer based on DINO [6] and designs a heuristic seed expansion strategy to obtain a single object region. As with LOST, TokenCut also uses features obtained with self-supervised learning. However, rather than relying on the attention map of some specific nodes, TokenCut forms a fully connected graph of image tokens, with edges labeled with a similarity score between tokens based on transformer features. The classical Ncut [46] algorithm is then used to detect and segment image objects.

*Our implementation is available at <https://www.mpsi.fr/Papers/TokenCut2022/>. An online demo is accessible at <https://huggingface.co/spaces/yangtaowang/TokenCut>.

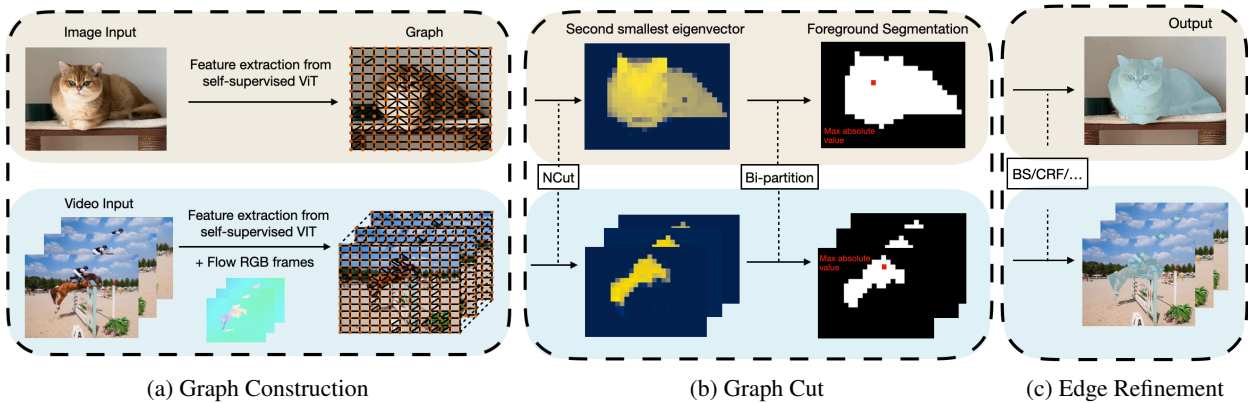


Figure 2: An overview of the TokenCut approach. The algorithm constructs a fully connected graph in which the nodes are image patches and the edges are similarities between the image patches using transformer features. Object segmentation is then solved using the Ncut algorithm [46]. Bi-partition of the graph using the second smallest eigenvector allows to detect foreground object. A Bilateral Solver [5] (BS) or Conditional Random Field [30] (CRF) can be used for edge refinement.

Unsupervised saliency detection. Unsupervised saliency detection seeks to segment a salient object within an image. Earlier works on this task [24, 33, 70, 80] use techniques such as color contrast [10], certain background priors [66], or super-pixels [33, 72]. More recently, unsupervised deep models [39, 78] propose to learn with noisy pseudo-labels generated from different handcrafted saliency methods. [63] shows that unsupervised GANs allow to differentiate between foreground and background pixels and generate high-quality saliency masks. In this work, we show that incorporating a simple post-processing step into our method for unsupervised object discovery can provide a strong baseline method for unsupervised saliency detection.

Unsupervised video segmentation. Given an unlabeled video, unsupervised video segmentation aims to generate pixel-level masks for the object of interest in the video. Prior works segment objects by selecting super-pixels [29], learning flattened 3D object representations [31], constructing an adversarial network to mask a region such that the model can predict the optical flow of the masked region [74], or reconstructing the optical flow in a self-supervised manner [71], etc. DyStaB [73] first partitions the motion field by minimizing the temporal consistent mutual information and then uses the segments to learn the object detector, in which the models are jointly trained with a bootstrapping strategy. The deformable sprites method (DeSprite) [75] trains a video auto-encoder to segment the object of interest by decomposing the video into layers of persistent motion groups. In contrast to these methods [71, 73, 74], our proposed method does not require prior training on videos. Compared with methods [29, 31] that do not train on videos, our method achieves superior performance.

3 Approach: TokenCut

In this section, we present TokenCut, a unified algorithm that can be used to segment the salient object in an image or moving objects in a video. Our approach, illustrated in Fig. 2, is based on a graph where the nodes are visual patches from either an image or a sequence of frames, and the edges are similarities between the features of the nodes extracted using self-supervised transformers.

This section is organised as follows: we first briefly review vision transformers and the Normalized Cut algorithm in Section 3.1.1 and Section 3.1.2. We then describe the TokenCut algorithm for object detection and segmentation in images and videos in Section 3.2.

3.1 Background

3.1.1 Vision Transformers

The Vision Transformer has been proposed in [15]. The key idea is to process an image with transformer [58] architectures using non-overlapping patches as tokens. For an image with size $H \times W$, a vision transformer takes non-overlapping $K \times K$ image patches as inputs, resulting in $N = HW/K^2$ patches. Each patch is used as a token, described by a vector of numerical features that provide an embedding. An extra learnable token, denoted as a class token CLS , is used to represent the aggregated information of the entire set of patches. A positional encoding is added to CLS token and the set of patch tokens, then they are fed to a standard transformer network with self-attention [58] and layer normalization [2].

The Vision Transformer is composed of several stacked layers of encoders, each with feed-forward networks and multiple attention heads for self-attention, paralleled with skip connections. For TokenCut algorithm, we use the Vision Transformer proposed in DINO [6], trained with

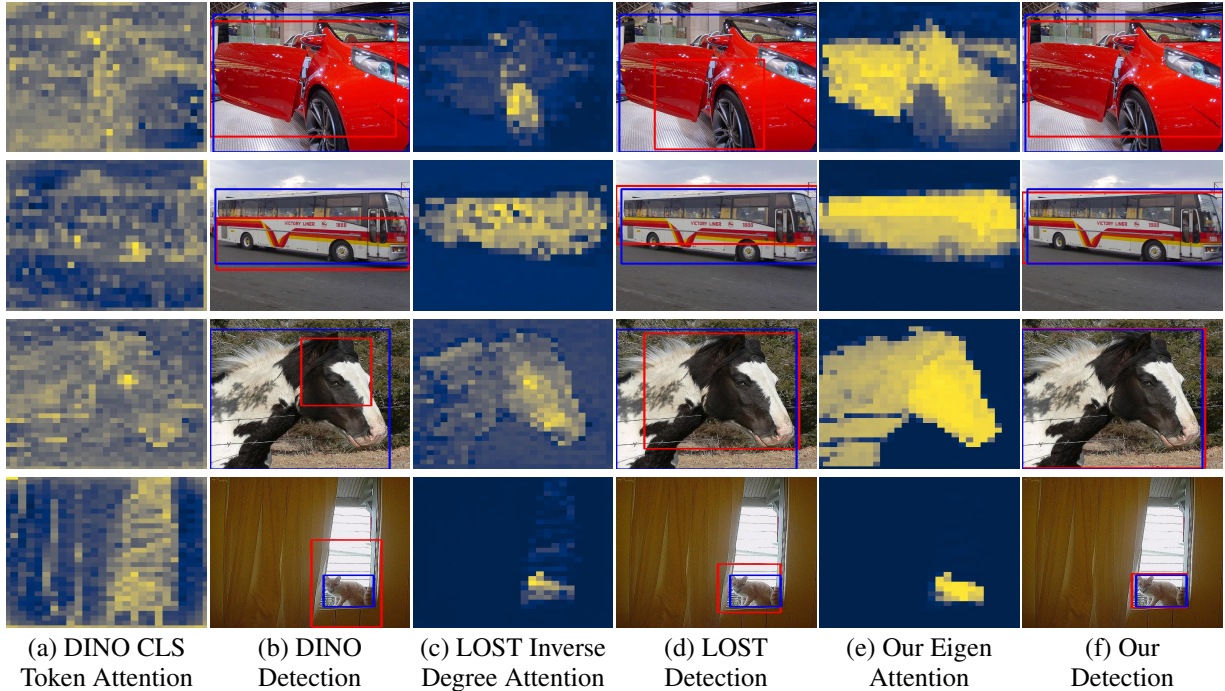


Figure 3: **Visual results of unsupervised single object discovery on VOC12.** In (a), we show the attention of the CLS token in DINO [6] which is used for detection (b). LOST [49] is mainly relied on the map of inverse degrees (c) to perform detection (d). For our approach, we illustrate the eigenvector in (e) and our detection in (f). **Blue** and **Red** bounding boxes indicate the ground-truth and the predicted bounding boxes respectively.

self-supervised learning. We extract latent features from the final layer as the input features for TokenCut.

3.1.2 Normalized Cut (Ncut)

Graph partitioning. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} and \mathcal{E} are sets of nodes and edges respectively. \mathbf{E} is the similarity matrix with $\mathbf{E}_{i,j}$ as the edge between the i -node v_i and the j -th node v_j . Ncut [46] is proposed to partition the graph into two disjoint sets \mathcal{A} and \mathcal{B} . Different to standard graph cut, Ncut criterion considers both the total dissimilarity between \mathcal{A} and \mathcal{B} as well as the total similarity within \mathcal{A} and \mathcal{B} . Precisely, we seek to minimize the Ncut energy [46]:

$$\frac{C(\mathcal{A}, \mathcal{B})}{C(\mathcal{A}, \mathcal{V})} + \frac{C(\mathcal{B}, \mathcal{B})}{C(\mathcal{B}, \mathcal{V})}, \quad (1)$$

where C measures the degree of similarity between two sets. , $C(\mathcal{A}, \mathcal{B}) = \sum_{v_i \in \mathcal{A}, v_j \in \mathcal{B}} \mathbf{E}_{i,j}$ and $C(\mathcal{A}, \mathcal{V})$ is the total connection from nodes in \mathcal{A} to all the nodes in the graph.

As shown by [46], the equivalent form of optimization problem in Eqn 1 can be expressed as:

$$\min_{\mathbf{x}} E(\mathbf{x}) = \min_{\mathbf{y}} \frac{\mathbf{y}^T (\mathbf{D} - \mathbf{E}) \mathbf{y}}{\mathbf{y}^T \mathbf{D} \mathbf{y}}, \quad (2)$$

with the condition of $\mathbf{y} \in \{1, -b\}^N$, where b satisfies $\mathbf{y}^T \mathbf{D} \mathbf{1} = 0$, where \mathbf{D} is a diagonal matrix with $\mathbf{d}_i = \sum_j \mathbf{E}_{i,j}$ on its diagonal.

Ncut solution with the relaxed constraint. Taking $\mathbf{z} = \mathbf{D}^{\frac{1}{2}} \mathbf{y}$, Eqn 2 can be rewritten as:

$$\min_{\mathbf{z}} \frac{\mathbf{z}^T \mathbf{D}^{-\frac{1}{2}} (\mathbf{D} - \mathbf{E}) \mathbf{D}^{-\frac{1}{2}} \mathbf{z}}{\mathbf{z}^T \mathbf{z}}. \quad (3)$$

Indicating in [46], the formulation in Eqn 3 is equivalent to the Rayleigh quotient [57], which is equivalent to solve $\mathbf{D}^{-\frac{1}{2}} (\mathbf{D} - \mathbf{E}) \mathbf{D}^{-\frac{1}{2}} \mathbf{z} = \lambda \mathbf{z}$, where $\mathbf{D} - \mathbf{E}$ is the Laplacian matrix and known to be positive semidefinite [42]. Therefore $\mathbf{z}_0 = \mathbf{D}^{\frac{1}{2}} \mathbf{1}$ is an eigenvector associated to the smallest eigenvalue $\lambda = 0$. According to Rayleigh quotient [57], the second smallest eigenvector \mathbf{z}_1 is perpendicular to the smallest one (\mathbf{z}_0) and can be used to minimize the energy in Eqn 3,

$$\mathbf{z}_1 = \operatorname{argmin}_{\mathbf{z}^T \mathbf{z}_0} \frac{\mathbf{z}^T \mathbf{D}^{-\frac{1}{2}} (\mathbf{D} - \mathbf{E}) \mathbf{D}^{-\frac{1}{2}} \mathbf{z}}{\mathbf{z}^T \mathbf{z}}.$$

Taking $\mathbf{z} = \mathbf{D}^{\frac{1}{2}} \mathbf{y}$,

$$\mathbf{y}_1 = \operatorname{argmin}_{\mathbf{y}^T \mathbf{D} \mathbf{1} = 0} \frac{\mathbf{y}^T (\mathbf{D} - \mathbf{E}) \mathbf{y}}{\mathbf{y}^T \mathbf{D} \mathbf{y}}.$$

Thus, the second smallest eigenvector of the generalized eigensystem $(\mathbf{D} - \mathbf{E})\mathbf{y} = \lambda\mathbf{D}\mathbf{y}$ is the real valued solution to the Ncut [46] problem.

3.2 TokenCut Algorithm

TokenCut algorithm consists of three steps: (a) Graph Construction, (b) Graph Cut, (c) Edge Refinement. An overview of the approach is shown in Fig. 2.

3.2.1 Graph construction

Image Graph. As described in Section 3.1.2, we consider a fully connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathbf{v}_i represents the feature vectors of the node v_i . Each patch is linked to other patches by labeled edges, \mathcal{E} . Edge labels represent a similarity score S .

$$\mathcal{E}_{i,j} = \begin{cases} 1, & \text{if } S(\mathbf{v}_i, \mathbf{v}_j) \geq \tau \\ \epsilon, & \text{else} \end{cases}, \quad (4)$$

where τ is a hyper-parameter and $S(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \mathbf{v}_j}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2}$ is the cosine similarity between features. ϵ is a small value $1e-5$ to assure a fully connected graph. Note that the spatial information has been implicitly included in the features, which is achieved by positional encoding in the transformer.

Video Graph. As with images, videos are presented as a fully connected graph where the nodes \mathcal{V} are visual patches and the edges \mathcal{E} are labeled with the similarity between patches. However, for videos, similarity includes a score based on both RGB appearance and a RGB representation of optical flow computed between consecutive frames [3]. The algorithm extracts a sequence of feature vectors using a vision transformer as described in Section 3.1.1. Let \mathbf{v}_i^I and \mathbf{v}_i^F denote the feature of i -th image patch and flow patch respectively. Edges is labeled with the average over the similarities between image feature and flow features, expressed as:

$$\mathcal{E}_{i,j} = \begin{cases} 1, & \text{if } \frac{S(\mathbf{v}_i^I, \mathbf{v}_j^I) + S(\mathbf{v}_i^F, \mathbf{v}_j^F)}{2} \geq \tau \\ \epsilon, & \text{else} \end{cases}. \quad (5)$$

Image feature provide segmentation semantically similar objects while flow features focus on moving objects. We provide a full analysis on the definition of edges in Section 4.4.

3.2.2 Graph Cut

On the constructed graph, we apply the Ncut algorithm and obtain the second smallest eigenvector of the generalized eigensystem, as described in Section 3.1.2. The second smallest eigenvector, which we refer to as the ‘‘eigen attention’’, highlights salient objects. We provide visualization of the attention map in Section 4. To segment objects, we propose to first bi-partition the graph, then determine which partition belongs to the foreground and finally select nodes that are from the same objects.

Bi-partition the graph. To partition the nodes into two disjoint sets, we simply leverage the average value of the second smallest eigenvector to cut the graph $\bar{\mathbf{y}}_1 = \frac{1}{N} \sum_i \mathbf{y}_1^i$. Formally, $\mathcal{A} = \{\mathbf{v}_i | \mathbf{y}_1^i \leq \bar{\mathbf{y}}_1\}$ and $\mathcal{B} = \{\mathbf{v}_i | \mathbf{y}_1^i > \bar{\mathbf{y}}_1\}$. Note that, we also tried classical clustering algorithms, such as K-means and EM, to cluster the second smallest eigenvector into 2 partitions. The comparison is available in Section 4.4, indicating that the mean generally provides better results.

Determine foreground. Given the two disjoint sets, we consider the partition with the maximum absolute value \mathbf{v}_{max} as the foreground. Intuitively, the foreground object should be salient thus less connected to the entire graph, *i.e.* $\mathbf{d}_i < \mathbf{d}_j$ if \mathbf{v}_i belongs to the foreground while \mathbf{v}_j is the background token. Therefore, the eigenvector of the foreground object should have a larger absolute value than the one of the background.

Select the object. In images, we are interested in segmenting a single object while the foreground might contain more than one connected regions. We finally select the connected component in the foreground containing the maximum absolute value \mathbf{v}_{max} as our final object. In videos, as we aims at segmenting all the moving objects, we simply take the entire foreground region as our final output.

3.2.3 Edge Refinement

The graph cut provides coarse masks of objects due to the large size of transformer patches, which can be easily refined using standard edge refinement approaches. In detail, we leverage the off-the-shelf edge-aware post-processing techniques such as Bilateral Solver [5] (BS), Conditional Random Field [30] (CRF) on top of the obtained coarse mask in order to generate a fine grained mask.

Implementation details. For our experiments, we use the ViT-S/16 model [15] trained with self-distillation loss (DINO) [6] to extract features of patches, if not explicitly mentioned. Following [49], we employ the key features of the last layer as the input features \mathbf{v} . Ablations on different features and ViT backbones are provided in Tab. 5. We set $\tau = 0.2$ for all image datasets and $\tau = 0.3$ for video datasets if not explicitly mentioned, the dependency on τ is provided in Section 4.4. In terms of running time, our implementation takes approximately 0.32 seconds to detect a bounding box of a single image with resolution 480×480 on a single GPU QUADRO RTX 8000.

To generate optical flow, we use two different approaches: RAFT [54] and ARFlow [36]. The first one is supervised and the second one is self-supervised. We extract the optical flow at the original resolution of the image pairs, with the frame gaps $n = 1$ for DAVIS [41] and SegTV2 [32] dataset and, for FBMS [40], we use $n = 3$ to compensate small motion. Due to the limit of computational resources, we construct the video graph with

Table 1: **Comparisons for unsupervised single object discovery.** We compare TokenCut to state-of-the-art object discovery methods on VOC07 [16], VOC12 [17] and COCO20K [35,61] datasets. Model performances are evaluated with CorLoc metric. “Inter-image Simi.” means the model leverages information from the entire dataset and explores inter-image similarities to localize objects.

Method	Inter-image Simi.	DINO [6] Feat.	VOC07 [16]	VOC12 [17]	COCO20K [35,61]
Selective Search [49,56]		-	18.8	20.9	16.0
EdgeBoxes [49,81]		-	31.1	31.6	28.8
Kim et al. [28,49]	✓	-	43.9	46.4	35.1
Zhang et al. [49,79]	✓	-	46.2	50.5	34.8
DDT+ [49,65]	✓	-	50.2	53.1	38.2
rOSD [49,61]	✓	-	54.5	55.3	48.5
LOD [49,62]	✓	-	53.6	55.1	48.5
DINO-seg [6,49]		ViT-S/16 [15]	45.8	46.2	42.1
LOST [49]		ViT-S/16 [15]	61.9	64.0	50.7
DSS [38]		ViT-S/16 [15]	62.7	66.4	56.2
TokenCut		ViT-S/16 [15]	68.8 (↑ 6.1)	72.1 (↑ 5.7)	58.8 (↑ 2.6)
LOD + CAD* [49]	✓	-	56.3	61.6	52.7
rOSD + CAD* [49]	✓	-	58.3	62.3	53.0
LOST + CAD* [49]		ViT-S/16 [15]	65.7	70.4	57.5
TokenCut + CAD* [49]		ViT-S/16 [15]	71.4 (↑ 5.7)	75.3 (↑ 4.9)	62.6 (↑ 5.1)

* +CAD indicates to train a second stage class-agnostic detector with “pseudo-boxes” labels.

maximum of 90 frames on the DAVIS dataset. For videos longer than 90 frames, we simply aggregate results using non-overlapping subgraphs with maximum video frames of 90.

4 Experiments

We evaluate our approach on three tasks: unsupervised single object discovery, unsupervised saliency detection and unsupervised video segmentation. We present results of unsupervised single object discovery in Section 4.1. The results of unsupervised saliency detection in Section 4.2. The results of unsupervised video segmentation in Section 4.3. We provide ablation studies in Section 4.4.

4.1 Unsupervised Single Object Discovery

Datasets. We evaluate our approach on three commonly used benchmarks for unsupervised single object discovery: VOC07 [16], VOC12 [17] and COCO20K [35,61]. VOC07 and VOC12 contain 5011 and 11540 images respectively which belong to 20 categories. COCO20K consists of 19817 randomly chosen images from the COCO2014 dataset [35]. VOC07 and VOC12 are commonly evaluated for unsupervised object discovery [11,60–62,65]. COCO20K is a popular benchmark for a large scale evaluation [61].

Evaluation metric. Following the previous researches [11,13,50,60–62,65], we report performance using the *CorLoc* metric for precise localization. We take one predicted bounding box in an image. For target image, *CorLoc* is one if the intersection over union (IoU) score between the predicted bounding box and one of the ground truth bounding boxes is superior to 0.5.

Quantitative Results. We evaluate the *CorLoc* scores in comparison with previous state-of-the-art single object discovery methods [28,49,56,61,62,65,79,81] on VOC07, VOC12, and COCO20K datasets. These methods can be roughly divided into two groups according to whether the model leverages information from the entire dataset and explores inter-image similarities. Because of quadratic complexity of region comparison among images, models with inter-image similarities are generally difficult to scale to larger datasets. The selective search [56], edge boxes [81], LOST [49] and TokenCut do not require inter-image similarities and are thus much more efficient. As shown in the Tab. 1, TokenCut consistently outperforms all the previous methods on all the datasets by a large margin. Particularly, TokenCut improves the concurrent work DSS [38] by 6.1%, 5.7% and 2.6% in VOC07, VOC12 and COCO20K respectively using the same ViT-S/16 features.

We also list a set of results that includes using a second stage unsupervised training strategy to boost the performance, which is referred to Class-Agnostic Detection (CAD) and proposed in LOST [49]. Precisely, we first compute K-means on all the boxes produced by the first stage single object discovery model to obtain pseudo labels of the bounding boxes. Then a classical Faster RCNN [43] is trained on the pseudo labels. As shown in Tab. 1, TokenCut with CAD outperforms the state-of-the-art by 5.7%, 4.9% and 5.1% on VOC07, VOC12 and COCO20k respectively, which demonstrates the efficiency of our method.

Qualitative Results. In Fig. 3, we provide visualization for DINO-seg [6], LOST [49] and TokenCut*. For each method, we visualize the heatmap that is used to per-

*More visual results can be found in the project webpage <https://www.m-psi.fr/Papers/TokenCut2022/>.

Table 2: **Comparisons for unsupervised saliency detection** We compare TokenCut to state-of-the-art unsupervised saliency detection methods on ECSSD [47], DUTS [64] and DUT-OMRON [72]. TokenCut achieves better results compared with other competitive approaches.

Method	ECSSD [47]			DUTS [64]			DUT-OMRON [72]		
	$maxF_{\beta}(\%)$	IoU(%)	Acc.(%)	$maxF_{\beta}(\%)$	IoU(%)	Acc.(%)	$maxF_{\beta}(\%)$	IoU(%)	Acc.(%)
HS [70]	67.3	50.8	84.7	50.4	36.9	82.6	56.1	43.3	84.3
wCtr [80]	68.4	51.7	86.2	52.2	39.2	83.5	54.1	41.6	83.8
WSC [33]	68.3	49.8	85.2	52.8	38.4	86.2	52.3	38.7	86.5
DeepUSPS [39]	58.4	44.0	79.5	42.5	30.5	77.3	41.4	30.5	77.9
BigBiGAN [63]	78.2	67.2	89.9	60.8	49.8	87.8	54.9	45.3	85.6
E-BigBiGAN [63]	79.7	68.4	90.6	62.4	51.1	88.2	56.3	46.4	86.0
LOST [45, 49]	75.8	65.4	89.5	61.1	51.8	87.1	47.3	41.0	79.7
LOST [45, 49]+BS [5]	83.7	72.3	91.6	69.7	57.2	88.7	57.8	48.9	81.8
DSS [38]	-	73.3	-	-	51.4	-	-	56.7	-
TokenCut	80.3	71.2	91.8	67.2	57.6	90.3	60.0	53.3	88.0
TokenCut + BS [5]	87.4 (\uparrow 3.7)	77.2	93.4	75.5	62.4	91.4	69.7 (\uparrow 11.9)	61.8	89.7
TokenCut + CRF [30]	87.4 (\uparrow 3.7)	77.7 (\uparrow 4.4)	93.6 (\uparrow 2.0)	75.7 (\uparrow 6.0)	62.8 (\uparrow 5.6)	91.5 (\uparrow 2.8)	69.2	61.9 (\uparrow 5.2)	89.8 (\uparrow 8.0)

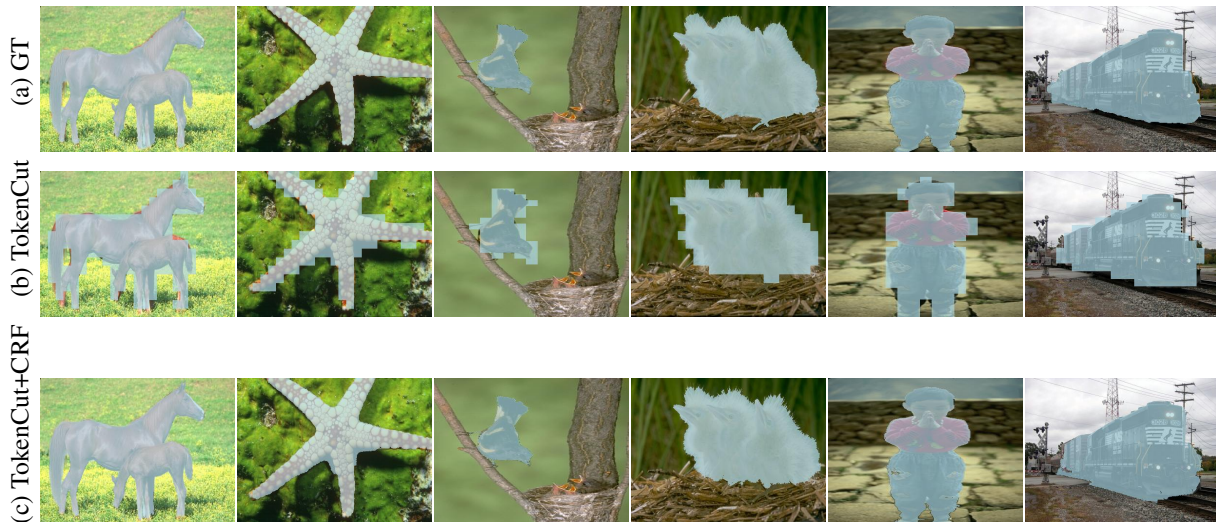


Figure 4: **Visual results of unsupervised segments on ECSSD [47]**. In (a), we show the ground truth. (b) is TokenCut coarse mask segmentation result. The performance of TokenCut + Bilateral Solver (BS) is presented in (c).

form object detection. For DINO-seg, the heatmap is the attention map associated to the CLS token. For LOST, the detection is mainly based on the map of inverse degree ($\frac{1}{d_i}$). For TokenCut, we display the second smallest eigenvector. The visual results demonstrate that our TokenCut can extract a high quality segmentation for the salient object. Compared with DINO-seg and LOST, TokenCut is able to extract a more complete segmentation as can be seen in the first and the third samples in Fig. 3. In some other cases, when all the methods have a high quality map, TokenCut has the strongest intensity on the object. This phenomenon can be viewed in the last sample in Fig. 3.

Internet Images. We further test TokenCut on Internet images*. The results are in Fig 5. It can be seen that even though the input images are with noisy backgrounds, our algorithm can still provide a precise attention map to

*We provide an online demo allowing to test Internet images: <https://huggingface.co/spaces/yangtaowang/TokenCut>.

cover the object and lead to accurate bounding box prediction, which demonstrates again the robustness of our approach again.

4.2 Unsupervised Saliency detection

Datasets. We validate the performance of the model on three datasets on unsupervised Saliency detection: Extended Complex Scene Saliency Dataset (ECSSD) [47], DUTS [64] and DUT-OMRON [72]. ECSSD contains 1 000 real-world images of complex scenes for testing. DUTS contains 10 553 train and 5 019 test images. The training set is collected from the ImageNet detection train/val set. The test set is collected from ImageNet test, and the SUN dataset [69]. Following the previous works [45], we report the performance on the DUTS-test subset. DUT-OMRON [72] contains 5 168 images of high quality natural images for testing.

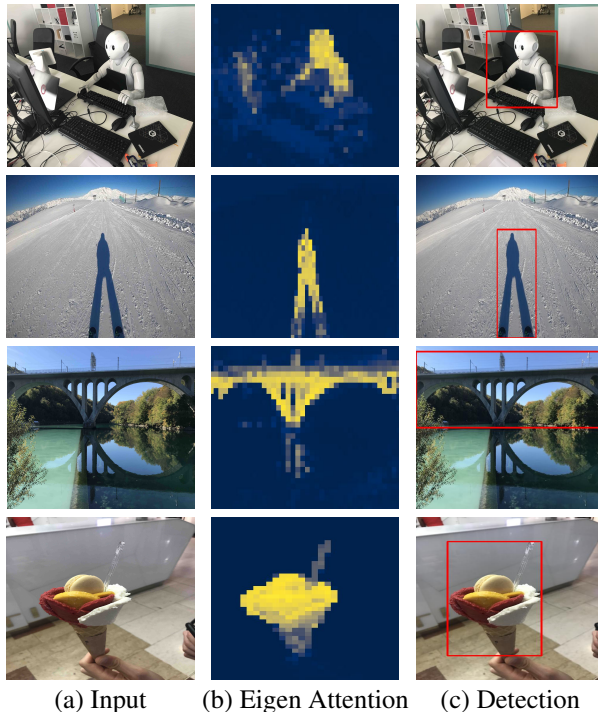


Figure 5: **Visualization of images from the Internet.** We show the input images, our eigen attention, and final detection in (a), (b), and (c) respectively.

Evaluation Metrics. We report three standard metrics: F-measure, IoU and Accuracy. *F-measure* is a standard measure in saliency detection. It is computed as $F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}$, where the Precision and Recall are defined based on the binarized predicted mask and ground truth mask. The $max F_\beta$ is the maximum value of 255 uniformly distributed binarization thresholds. Following previous works [45, 63], we set $\beta = 0.3$ for consistency. *IoU* (Intersection over Union) score is computed based on the binary predicted mask and the ground-truth, the threshold is set to 0.5. *Accuracy* measures the proportion of pixels that have been correctly assigned to the object/background. The binarization threshold is set to 0.5 for masks.

Results. The qualitative results are in Tab. 2. TokenCut significantly outperforms previous state-of-the-art. Adding BS [5] or CRF [30] refines the boundary of an object and further boosts the TokenCut performance, which can also be seen from the visual results presented in Fig. 4.

4.3 Unsupervised Video Segmentation

Datasets. We further evaluate TokenCut on three commonly used datasets for unsupervised video segmentation: DAVIS [41], FBMS [40] and SegTV2 [32]. DAVIS contains 50 high-resolution real-world videos, where 30 for training and 20 for validation. The pixel-wise anno-

Table 3: **Comparisons for unsupervised video segmentation.** We report *Jaccard index* and compare TokenCut to state-of-the-art unsupervised video segmentation methods on DAVIS [41], FBMS [40] and SegTV2 [32]. TokenCut achieves competitive results comparing with other competitive approaches.

Method	Flow	Training	DAVIS [41]	FBMS [40]	SegTV2 [32]
ARP [29]	CPM [23]		76.2	59.8	57.2
ELM [31]	Classic+NL [51]		61.8	61.6	-
MG [71]	RAFT [54]	✓	68.3	53.1	58.2
CIS [74]	PWCNet [52]	✓	71.5	63.5	62
DyStaB [73]*	PWCNet [52]	✓	80.0	73.2	74.2
DeSprite [75]‡	[54]	✓	79.1	71.8	72.1
TokenCut	RAFT [54]		64.3	60.2	59.6
TokenCut + BS [5]	RAFT [54]		75.1	61.2	56.4
TokenCut + CRF [30]	RAFT [54]		76.7	66.6	61.6
TokenCut	ARFlow [36]		62.0	61.0	58.9
TokenCut + BS [5]	ARFlow [36]		73.1	64.7	54.6
TokenCut + CRF [30]	ARFlow [36]		74.4	69.0	60.8

*: [73] is trained on DAVIS and evaluated on FBMS and SegTV2;

‡: [75] is optimized for each video separately.

tations is depicted the principle moving object within the scene for each frame. FBMS consists of 59 multiple moving object videos, providing 30 videos for testing with a total of 720 annotation frames. SegTV2 contains 14 full pixel-level annotated video for multiple objects segmentation. Following [71], we fuse the annotation of all moving objects into a single mask on FBMS and SegTV2 datasets for fair comparison.

Evaluation metrics. We report performance using *Jaccard index*. Jaccard index measures the intersection of union between an output segmentation M and the corresponding ground-truth mask G , which has been formulated as $\mathcal{J} = \frac{|M \cap G|}{|M \cup G|}$.

Results. We compare TokenCut to the state-of-the-art unsupervised video segmentation results in Tab. 3. Our approach achieves competitive performances for this task. Note that DyStaB [73] needs to be trained on the entire DAVIS training set and use the pretrained model for evaluation on FBMS and SegTV2 datasets. DeSprite [75] learns an auto-encoder model to optimize on each individual video. While our approach is training free and generalized well on all three datasets. The visual results are illustrated in Fig. 6, our approach can precisely segment moving objects even in some challenging occlusions, and adding CRF as a post-processing allows obtaining better segmentation on the boundary*.

4.4 Analysis and Discussion

Impact of τ . In Tab. 4, we provide an analysis on τ defined in Eqn 4. The results indicate that the effects of variations in τ value are not significant and that a suitable threshold is $\tau = 0.2$ for image input and $\tau = 0.3$ for video input.

*The segmentation results of entire videos can be found in our project webpage: <https://www.m-psi.fr/Papers/TokenCut2022/>.

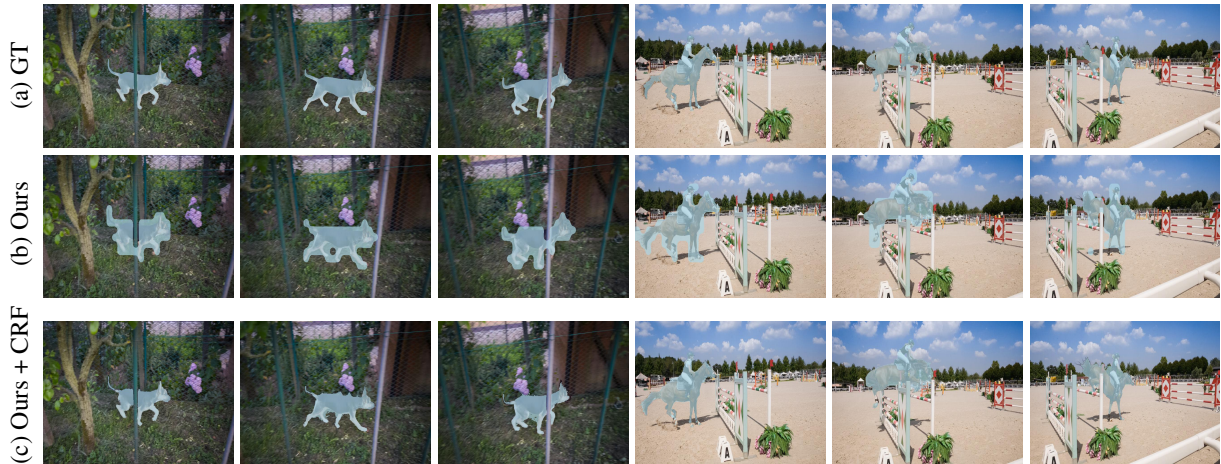


Figure 6: **Visual results of unsupervised video segmentation on DAVIS [12].** In (a), we show the ground truth segmentation. For TokenCut, we illustrate its coarse mask in (b) and refinement results with CRF in (c).

Table 4: **Analysis of τ .** We report CorLoc for unsupervised single object discovery on VOC07 [16], VOC12 [17] and COCO20K [35, 61] datasets, and Jaccard index on DAVIS [41].

τ	CorLoc			Jaccard Index
	VOC07 [16]	VOC12 [17]	COCO20K [35, 61]	DAVIS [41]
0	67.4	71.3	56.1	70.7
0.1	68.6	72.1	58.2	74.6
0.2	68.8	72.1	58.8	75.8
0.3	67.7	72.1	58.2	76.7

Backbones. In Tab. 5, we provide an ablation study on different transformer backbones. The “-S” and “-B” are ViT small [6, 15] and ViT base [6, 15] architecture respectively. The “-16” and “-8” represents patch sizes 16 and 8 respectively. The “DeiT” is pre-trained supervised transformer model. The “MoCov3” [7] and “MAE” [20] are pre-trained self-supervised transformer model. We optimise τ for different backbones: τ is set to 0.3 for MoCov3 and MAE, while for DINO and Deit τ is set to 0.2. Several insights can be found: 1) TokenCut is not suitable for supervised transformer model, while self-supervised transformers provide more powerful features allowing completing the task with TokenCut. 2) As LOST [49] relies on a heuristic seeds expansion strategy, the performance varies significantly using different backbones. While our approach is more robust. Moreover, as no training is required for TokenCut, it might be a more straightforward evaluation for the self-supervised transformers.

Bi-partition strategies. In Tab. 6, we study different strategies to separate the nodes in our graph into two groups using the second smallest eigenvector. We consider three natural methods: mean value (Mean), Expectation-Maximisation (EM), K-means clustering (K-means). We have also tried to search for the splitting point based on the best Ncut energy (Eqn 1). Note this approach

Table 5: **Analysis of different backbones.** We report CorLoc for unsupervised single object discovery on VOC07 [16], VOC12 [17] and COCO20K [35, 61] datasets.

Method	Backbone	VOC07 [16]	VOC12 [17]	COCO20K [35, 61]
LOST [49]	DINO-S/16 [6, 15]	61.9	64.0	50.7
TokenCut	DeiT-S/16 [15, 55]	2.39	2.9	3.5
TokenCut	MoCoV3-S/16 [7, 15]	66.2	66.9	54.5
TokenCut	DINO-S/16 [6, 15]	68.8 (\uparrow 6.9)	72.1 (\uparrow 8.1)	58.8 (\uparrow 8.1)
LOST [49]	DINO-S/8 [6, 15]	55.5	57.0	49.5
TokenCut	DINO-S/8 [6, 15]	67.3 (\uparrow 11.8)	71.6 (\uparrow 14.6)	60.7 (\uparrow 11.2)
LOST [49]	DINO-B/16 [6, 15]	60.1	63.3	50.0
TokenCut	MAE-B/16 [15, 20]	61.5	67.4	47.7
TokenCut	DINO-B/16 [6, 15]	68.8 (\uparrow 8.7)	72.4 (\uparrow 9.1)	59.0 (\uparrow 9.0)

is computational expensive due to the quadratic complexity. The result suggests that the simple mean value as the splitting point performs well for most cases.

Video input. We also study the impact of using RGB or Flow for video segmentation. The quantitative results are Tab. 7. We can see constructing graph with RGB and Flow together can importantly improve the performances on DAVIS [41]. On FBMS [40] and SegTV2 [32], due to low quality of optical flow, the performances are not stable. Using both RGB appearance and flow lead to slightly improvement without the edge refinement, while a slightly worse results than using RGB appearance only after edge refinement. Some qualitative results are illustrated in Fig. 7. We can see how RGB frame and optical flow are complementary to each other: in the first row, the target moving person shares semantically similar features to other audiences and leveraging only RGB frames would produce a mask cover all the persons; in the second row, the flow also has non-negligible values on the surface of the river, thus using only flow leads to worse performances.

Video graph. In Tab. 8, we provide an analysis for different ways to construct. For edges, we also consider

Table 6: **Analysis of different bi-partition methods.** We report CorLoc for unsupervised single object discovery on VOC07 [16], VOC12 [17] and COCO20K [35, 61] datasets.

Bi-partition	VOC07	VOC12	COCO20K
Mean	68.8	72.1	58.8
Energy (Eqn 1)	67.3	69.7	-
EM	63.0	65.7	59.3
K-means	67.5	69.2	61.6

Table 7: **Analysis of video input.** We report Jaccard index for video segmentation on DAVIS [41], FBMS [40] and SegTV2 [32] with using input. “RGB + Flow” refers to using both video RGB frame and optical flow as input. “RGB” and “Flow” present using only either RGB frames or optical flow as input. “CRF” indicates whether edge refinement step using CRF [30] is executed.

Input	CRF	DAVIS [41]	FBMS [40]	SegTV2 [32]
RGB		51.8	58.4	59.3
Flow		49.9	48.3	46.7
RGB + Flow		64.3	60.2	59.6
RGB	✓	62.2	67.5	63.7
Flow	✓	63.1	50.2	50.2
RGB + Flow	✓	76.7	66.6	61.6

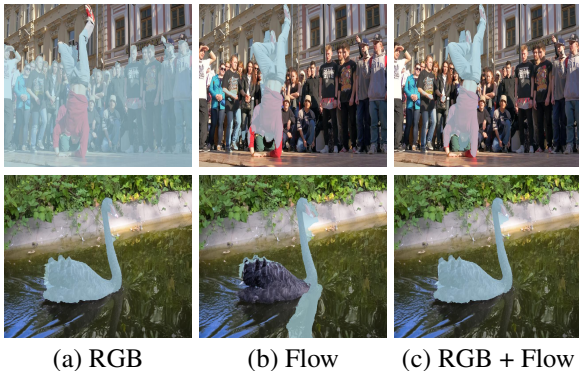


Figure 7: **Visualization on DAVIS [41] using different inputs.** We show segmentation results with RGB, Flow and RGB + Flow in (a), (b) and (c) respectively.

the minimum and maximum values between the flow and RGB similarities. For nodes, a natural baseline is to build a graph for each single frame. We can see that the optimal choice is to leverage the average value of the flow and RGB similarities (Eqn. 4) and build a graph for an entire video.

Limitations. Despite the good performance of the TokenCut proposal, it has several limitations, where we show several failure cases in Fig. 8: i) As seen in the 1st row, TokenCut focuses on the largest salient part in the image, which may not be the desired object. ii) Similar to LOST [49], TokenCut assumes that a single salient

Table 8: **Analysis of video graph.** We report Jaccard index (\mathcal{J}) for video segmentation on DAVIS [41] with different video graphs. “single frame” represent creating the graph for each frame separately.

Nodes	Edges	DAVIS (\mathcal{J})
Video	$\min(S(\mathbf{v}_i^I, \mathbf{v}_j^I), S(\mathbf{v}_i^F, \mathbf{v}_j^F))$	73.7
Video	$\max(S(\mathbf{v}_i^I, \mathbf{v}_j^I), S(\mathbf{v}_i^F, \mathbf{v}_j^F))$	71.1
Video	$\frac{S(\mathbf{v}_i^I, \mathbf{v}_j^I) + S(\mathbf{v}_i^F, \mathbf{v}_j^F)}{2}$	76.8
Single Frame	$\frac{S(\mathbf{v}_i^I, \mathbf{v}_j^I) + S(\mathbf{v}_i^F, \mathbf{v}_j^F)}{2}$	76.4

object occupies the foreground. If multiple overlapping objects are present in an image, both LOST and our approach would fail to detect one of the object, as displayed in the 2nd row. iii) For object detection, neither LOST nor our approach can handle occlusion properly, which is showed in 3rd row.

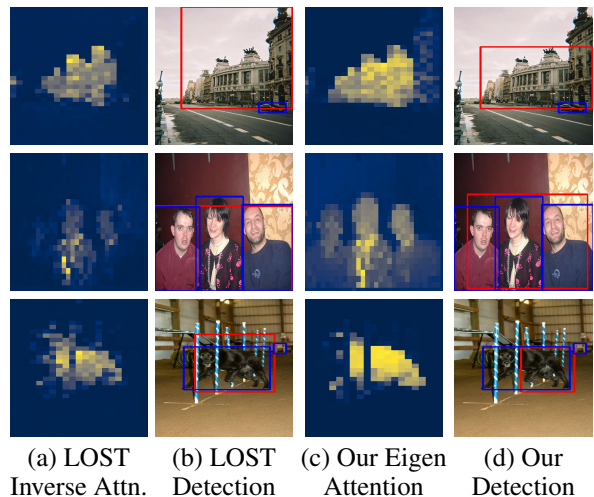


Figure 8: **Failure cases on VOC12 (1st and 2nd row) and COCO (3rd row).** LOST [49] mainly relies on the map of inverse degrees (a) to perform detection (b). For our approach, we illustrate the eigenvector in (c) and our detection in (d). Blue and Red bounding boxes indicate the ground-truth and the predicted bounding boxes respectively.

5 Conclusion

This paper describes TokenCut, an unified and effective approach for both image and video object segmentation without the need for supervised learning. TokenCut uses features from self-supervised transformers to constructs a graph where nodes are patches and edges represent similarities between patches. For videos, optical flow is incorporated to determine moving objects. We show that salient objects can be directly detected and delimited using the Normalized Cut algorithm. We evaluated this approach on unsupervised single object discovery, unsuper-

vised saliency detection, and unsupervised video object segmentation, demonstrating that TokenCut can provide a significant improvement over previous approaches. Our results demonstrate that self-supervised transformers can provide a rich and general set of features that may likely be used for a variety of computer vision problems.

Acknowledgements This work has been partially supported by the MIAI Multidisciplinary AI Institute at the Univ. Grenoble Alpes (MIAI@Grenoble Alpes - ANR-19-P3IA-0003), and by the EU H2020 ICT48 project Humane AI Net under contract EU #952026.

References

- [1] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López. Active learning for deep detection neural networks. In *ICCV*, 2019. 1
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv*, 2016. 3
- [3] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 2011. 5
- [4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv*, 2021. 2
- [5] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *ECCV*, 2016. 2, 3, 5, 7, 8
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 3, 4, 5, 6, 9
- [7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 2, 9
- [8] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021. 1
- [9] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. *PAMI*, 2020. 2
- [10] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 2014. 3
- [11] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *CVPR*, 2015. 2, 6
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 9
- [13] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010. 6
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2018. 2
- [15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2, 3, 5, 6, 9
- [16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>. 2, 6, 9, 10
- [17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>. 2, 6, 9, 10
- [18] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *CVPR*, 2022. 2
- [19] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 2013. 1
- [20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv*, 2015. 2, 9
- [21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv*, 2015. 2
- [22] Kuang-Jui Hsu, Yen-Yu Lin, Yung-Yu Chuang, et al. Co-attention cnns for unsupervised object co-segmentation. In *IJCAI*, 2018. 2
- [23] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *CVPR*, 2016. 8
- [24] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013. 3
- [25] Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010. 2
- [26] Armand Joulin, Francis Bach, and Jean Ponce. Multi-class cosegmentation. In *CVPR*, 2012. 2
- [27] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. *ICLR*, 2021. 1
- [28] Gunhee Kim and Antonio Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *NeurIPS*, 2009. 6
- [29] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *CVPR*, 2017. 3, 8
- [30] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *NIPS*, 2011. 2, 3, 5, 7, 8, 10
- [31] Dong Lao and Ganesh Sundaramoorthi. Extending layered models to 3d motion. In *ECCV*, 2018. 3, 8
- [32] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013. 2, 5, 8, 9, 10
- [33] Nianyi Li, Bilin Sun, and Jingyi Yu. A weighted sparse coding framework for saliency detection. In *CVPR*, 2015. 3, 7
- [34] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. In *NeurIPS*, 2021. 2

- [35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 6, 9, 10
- [36] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *CVPR*, 2020. 5, 8
- [37] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021. 1
- [38] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *CVPR*, 2022. 2, 6, 7
- [39] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mumtadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. In *NeurIPS*, 2019. 3, 7
- [40] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *TPAMI*, 2013. 2, 5, 8, 9, 10
- [41] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2, 5, 8, 9, 10
- [42] Alex Pothen, Horst D Simon, and Kang-Pu Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM journal on matrix analysis and applications*, 1990. 4
- [43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 2015. 6
- [44] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Yong Jae Lee, Alexander G Schwing, and Jan Kautz. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *CVPR*, 2020. 1
- [45] Xi Shen, Alexei A Efros, Armand Joulin, and Mathieu Aubry. Learning co-segmentation by segment swapping for retrieval and discovery. *arXiv preprint arXiv:2110.15904*, 2021. 7, 8
- [46] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *TPAMI*, 2000. 2, 3, 4, 5
- [47] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *TPAMI*, 2015. 2, 7
- [48] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewlet: Active learning with viewpoint entropy for semantic segmentation. In *CVPR*, 2020. 1
- [49] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021. 2, 4, 5, 6, 7, 9, 10
- [50] Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013. 6
- [51] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *CVPR*, 2010. 8
- [52] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 8
- [53] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *CVPR*, 2014. 2
- [54] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 5, 8
- [55] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 9
- [56] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *IJCV*, 2013. 6
- [57] Charles F Van Loan and G Golub. Matrix computations. *The Johns Hopkins University Press*, 1996. 4
- [58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3
- [59] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *CVPR*, 2011. 2
- [60] Huy V Vo, Francis Bach, Minsu Cho, Kai Han, Yann Lecun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *CVPR*, 2019. 2, 6
- [61] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *ECCV*, 2020. 2, 6, 9, 10
- [62] Huy V Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. *arXiv*, 2021. 2, 6
- [63] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. In *ICML*, 2021. 3, 7, 8
- [64] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 2, 7
- [65] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 2019. 6
- [66] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *ECCV*, 2012. 3
- [67] Bichen Wu, Forrest Iandola, Peter H Jin, and Kurt Keutzer. Squeezednet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *CVPRW*, 2017. 1
- [68] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018. 1
- [69] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 7
- [70] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, 2013. 3, 7
- [71] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *CVPR*, 2021. 3, 8

- [72] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 2, 3, 7
- [73] Yanchao Yang, Brian Lai, and Stefano Soatto. Dystab: Unsupervised object segmentation via dynamic-static bootstrapping. In *CVPR*, 2021. 3, 8
- [74] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3, 8
- [75] Vickie Ye, Zhengqi Li, Richard Tucker, Angjoo Kanazawa, and Noah Snavely. Deformable sprites for unsupervised video decomposition. *CVPR*, 2022. 3, 8
- [76] Yuan Yuan, Xiaodan Liang, Xiaolong Wang, Dit-Yan Yeung, and Abhinav Gupta. Temporal dynamic graph lstm for action-driven video object detection. In *ICCV*, 2017. 1
- [77] Yuan Yuan, Yueming Lyu, Xi Shen, Ivor Tsang, and Dit-Yan Yeung. Marginalized average attentional network for weakly-supervised learning. In *ICLR*, 2019. 1
- [78] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *CVPR*, 2018. 3
- [79] Runsheng Zhang, Yaping Huang, Mengyang Pu, Jian Zhang, Qingji Guan, Qi Zou, and Haibin Ling. Object discovery from a single unlabeled image by mining frequent itemsets with multi-scale features. *TIP*, 2020. 6
- [80] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014. 3, 7
- [81] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 6