Supplementary material
Self-Supervised Transformers for Unsupervised Object Discovery using Normalized Cut

Yangtao Wang[1], Xi Shen[2,3,*], Shell Xu Hu[4], Yuan Yuan[5], James L. Crowley[1], Dominique Vaufreydaz[1]
[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2] Tencent AI Lab [3] LIGM (UMR 8049) - Ecole des Ponts, UPE
[4] Samsung AI Center, Cambridge [5] MIT CSAIL

1. Analysis of backbones.

In Tab. 1, we provide an ablation study on different transformer backbones. The *"-S"* and *"-B"* are ViT small[2, 7] and ViT base[2, 7] architecture respectively. The *"-16"* and *"-8"* represents patch sizes 16 and 8 respectively. The *"MocoV3"* is another pre-trained self-supervised transformer model [3]. The τ value is set to 0.3 for MoCov3, while for Dino the best tau value is 0.2. We observe that although the result of MoCov3 is slightly worse than the results of TokenCut with Dino, MoCov3 still outperforms previous state-of-the-art, indicating that TokenCut can provide similar results when used with other self-supervised Transformer architectures. Besides, the results demonstrates that a patch size of 16 provides better results than a patch size of 8. Several insights can be found: i) TokenCut outperforms LOST for different backbones. ii) As LOST relies on a heuristic seeds expansion strategy, the performance varies significantly using different backbones. While our approach is more robust.

Table 1. **Analysis of different backbones.** We report CorLoc for unsupervised single object discovery on VOC07, VOC12, COCO20K.

Method	Backbone	VOC07	VOC12	COCO20K
LOST [12]	ViT-S/16 [2, 7]	61.9	64.0	50.7
TokenCut	MoCoV3-ViT-S/16 [3, 7]	66.2	66.9	54.5
TokenCut	ViT-S/16 [2, 7]	**68.8** (↑ **6.9**)	72.1 (↑ **8.1**)	58.8 (↑ **8.1**)
LOST [12]	ViT-S/8 [2, 7]	55.5	57.0	49.5
TokenCut	ViT-S/8 [2, 7]	67.3 (↑ **11.8**)	71.6 (↑ **14.6**)	**60.7** (↑ **11.2**)
LOST [12]	ViT-B/16 [2, 7]	60.1	63.3	50.0
TokenCut	ViT-B/16 [2, 7]	68.8 (↑ **8.7**)	**72.4** (↑ **9.1**)	59.0 (↑ **9.0**)

We provide another an ablation study on different backbones for weakly supervised object localization. Results are shown in Tab. 2. The "-S" and "-B" designate ViT small [2, 7] and ViT base [2, 7] architecture respectively. The "-16" and "-8" indicate patch sizes 16 and 8 respectively. For our approach, we report results with $\tau = 0.2$, which is the same on all the datasets. Note that LOST with ViT-S/8 achieves much worse results, because the seed expansion strategy in LOST relies on the top-100 patches which are with lowest degrees. When the total number of patches is large, the proposed seed expansion strategy is not able to cover entire objects. While our approach provides more robust performance on different datasets across different backbones.

[*] Corresponding Author

Table 2. **Analysis of backbones for weakly supervised object localization.** We report Top-1 Cls, GT Loc and Top-1 Loc on CUB [18] and Imagenet-1k [6] datasets.

Method	Backbone	τ	CUB [50], Acc. (%)			ImageNet-1K [11], Acc. (%)		
			Top-1 Cls	GT Loc	Top-1 Loc	Top-1 Cls	GT Loc	Top-1 Loc
LOST [12]	ViT-S/16 [2, 7]	-	**79.5**	89.7	71.3	**77.0**	60.0	49.0
TokenCut	ViT-S/16 [2, 7]	0.2	**79.5**	**91.8** (↑ **2.1**)	**72.9** (↑ **1.6**)	**77.0**	**65.4** (↑ **5.4**)	**53.4** (↑ **4.4**)
LOST [12]	ViT-S/8 [2, 7]	-	**82.3**	78.0	64.4	**79.4**	45.8	38.1
TokenCut	ViT-S/8 [2, 7]	0.2	**82.3**	**89.9** (↑ **11.9**)	**74.2** (↑ **9.8**)	**79.4**	**66.0** (↑ **20.2**)	**55.0** (↑ **16.9**)
LOST [12]	ViT-B/16 [2, 7]	-	**80.3**	**90.7**	**72.8**	**78.3**	58.6	48.3
TokenCut	ViT-B/16 [2, 7]	0.2	**80.3**	90.0 (↓ **0.7**)	72.5 (↓ **0.3**)	**78.3**	**63.2** (↑ **4.8**)	**52.3** (↑ **4.0**)

2. Analysis of bi-partition strategies.

In Tab. 3, 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 use python sklearn library for EM and K-means algorithm implementation. For EM algorithm, we set number of iteration to 300 and each component has its own general covariance matrix. The convergence threshold is set to 1e-3. For K-means algorithm, we use "k-means++" for initialization. The maximum number of iterations is set to 300. The convergence threshold is set to 1e-4. The result suggests that the simple mean value as the splitting point performs well for most cases. We have also tried to search for the splitting point based on the best Ncut(A,B) value. Due to the quadratic complexity, this approach requires substantially more computations. Thus, we finally obsolete it.

Table 3. **Analysis of different bi-partition methods.** We report CorLoc for unsupervised single object discovery.

Bi-partition	VOC07	VOC12	COCO20K
Mean	**68.8**	**72.1**	58.8
EM	63.0	65.7	59.3
K-means	67.5	69.2	**61.6**

3. Datasets

We present in this section the details of the datasets used in our experiments:

- **VOC07 and VOC12** correspond to the training and validation set of PASCAL-VOC07 and PASCAL-VOC12. VOC07 and VOC12 contain 5 011 and 11 540 images respectively which belong to 20 categories. They are commonly evaluated for unsupervised object discovery [4, 15, 16, 17, 20].
- **COCO20K** consists of 19 817 randomly chosen images from the COCO2014 dataset [9]. It is used as a benchmark in [16] for a large scale evaluation.
- **CUB** consists of 200 bird species, including 6 033 and 5 755 images in training and test sets respectively, which is commonly used to evaluate weakly supervised object localization [1, 5, 13, 14, 23].
- **ImageNet** [6] is a widely used benchmark for image classification and object detection, which consists of 1 000 different categories. The number of images in training and validation sets are 1.3 million and 50,000 respectively. Each image contains a single object supposed to be detected. During the training, only class labels are available.
- **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 [21]. Following the previous works [10], we report the performance on the DUTS-test subset.
- **DUT-OMRON** [22] contains 5 168 images of high quality natural images for testing.

4. Visual results for unsupervised single object discovery on VOC07 and COCO12

We show visual results for unsupervised single object discovery on VOC07 [8] and COCO12 [9, 16], which are illustrated in Fig. 1 and Fig. 2 respectively.

For each dataset, we compare both attention maps and bounding box predictions among DINO [2], LOST [12] and TokenCut. The attention map for DINO is extracted from the CLS token attention map of the last layer of key features. The attention map for LOST is the inverse degree map used in LOST for detection. The TokenCut attention map is the second smallest eigenvector of Equation 2. These results show that TokenCut provides clearly better segmentation of the object.

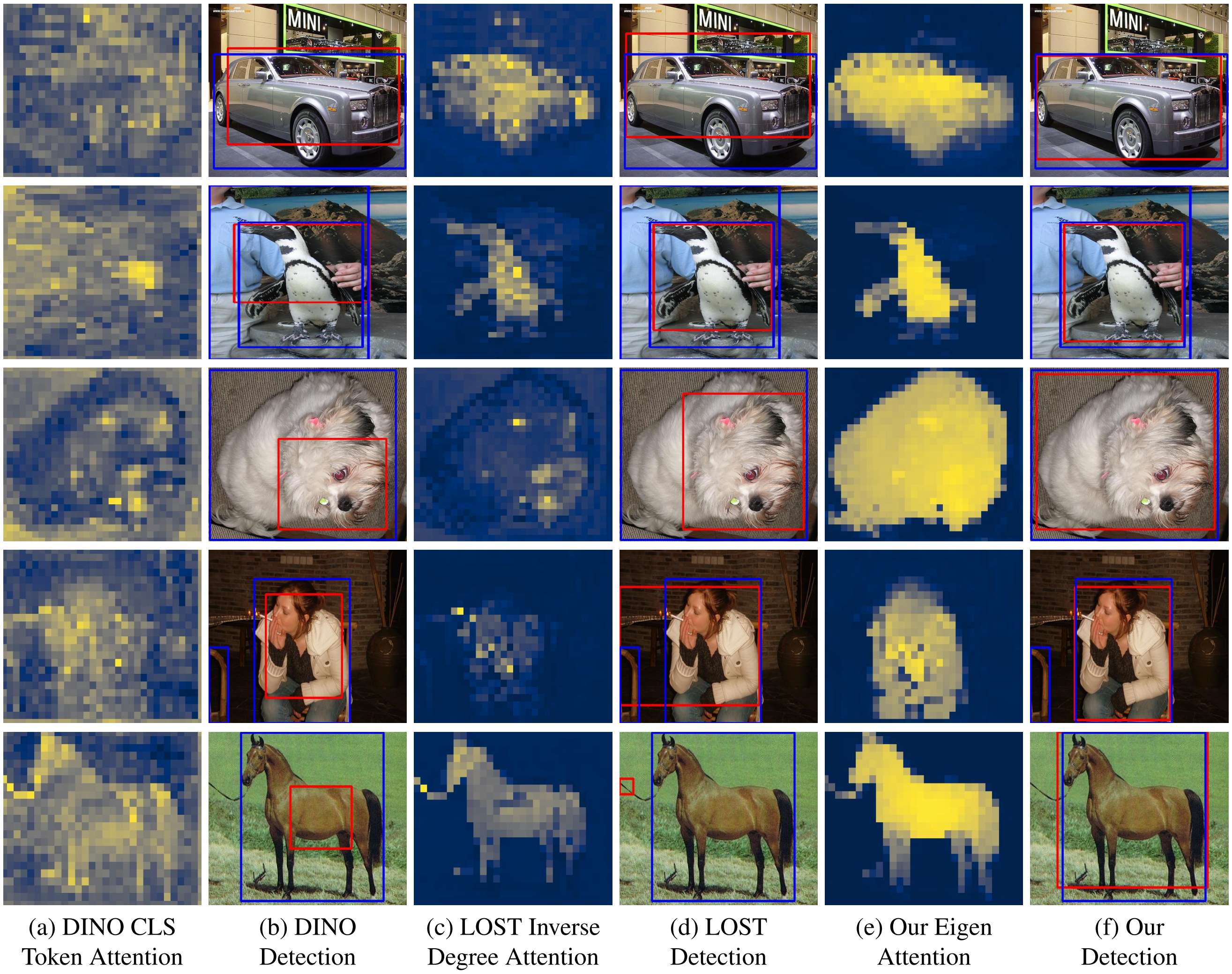
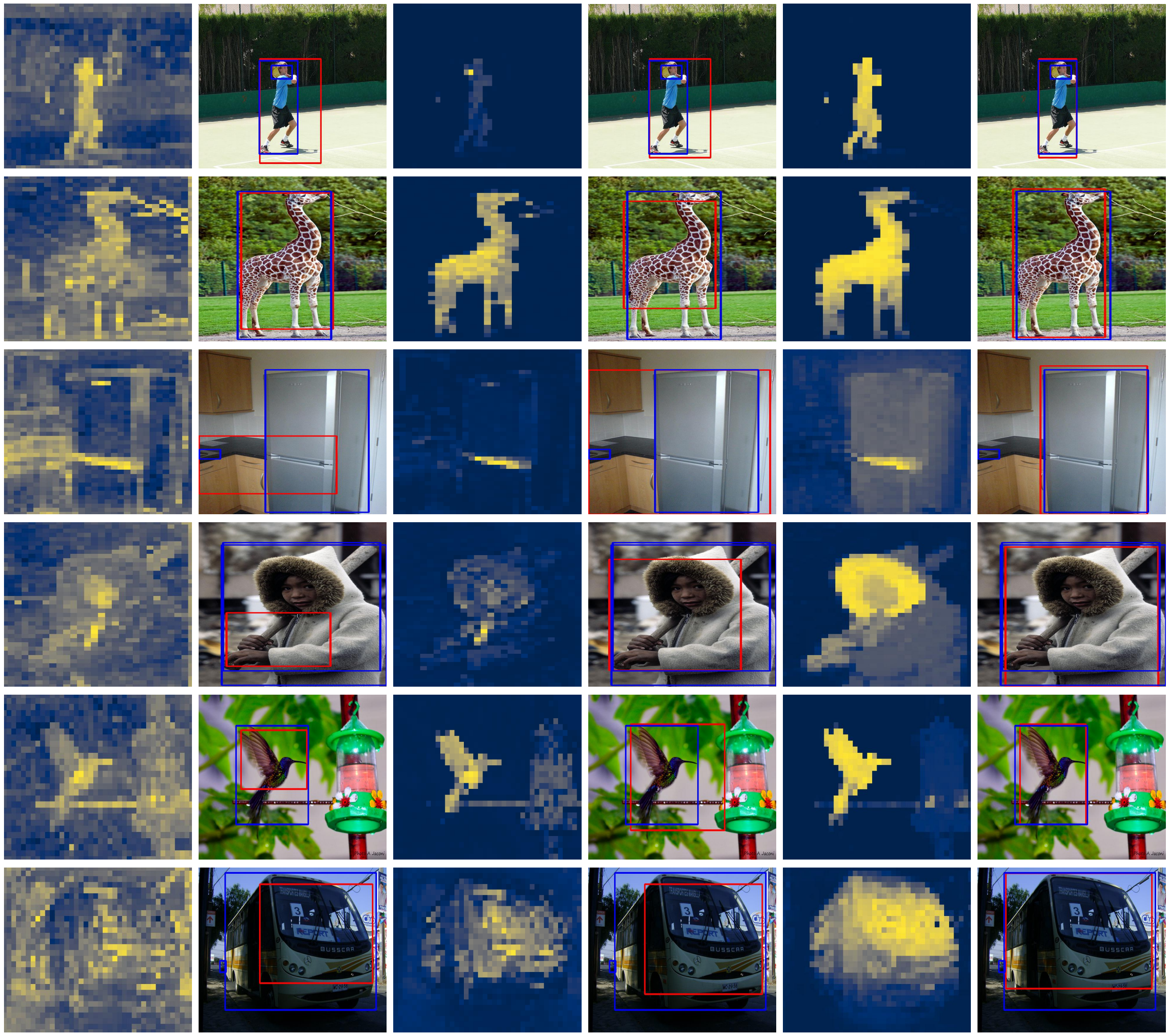


Figure 1. **Visual results of unsupervised single object discovery on VOC07 [8]** In (a), we show the attention of the CLS token in DINO [2] which is used for detection (b). LOST [12] is mainly relied on the map of inverse degrees (c) to perform detection (d). For our approach, we illustrate the eigenvector in (e) and our detection in (f). Blue and Red bounding boxes indicate the ground-truth and the predicted bounding boxes respectively.



(a) DINO CLS Token Attention

(b) DINO Detection

(c) LOST Inverse Degree Attention

(d) LOST Detection

(e) Our Eigen Attention

(f) Our Detection

Figure 2. **Visual results of unsupervised single object discovery on COCO20K [9, 16].** In (a), we show the attention of the CLS token in DINO [2] used for detection (b). LOST [12] mainly relies on the map of inverse degrees (c) to perform detection (d). For TokenCut, we illustrate the eigenvector in (e) and the detection in (f). Blue and Red bounding boxes indicate the ground-truth and the predicted bounding boxes respectively.

5. Visual results for weakly supervised object localizatio on CUB and Imagenet-1k

We present visual results for weakly supervised object localization on CUB [18] and Imagenet-1k [6] in Fig. 3 and Fig. 4 respectively.

For each dataset, we compare the attention map and bounding box prediction with LOST [12] and our approach. The eigenvector of TokenCut provides better segmentation on objects and leads to better detection results.

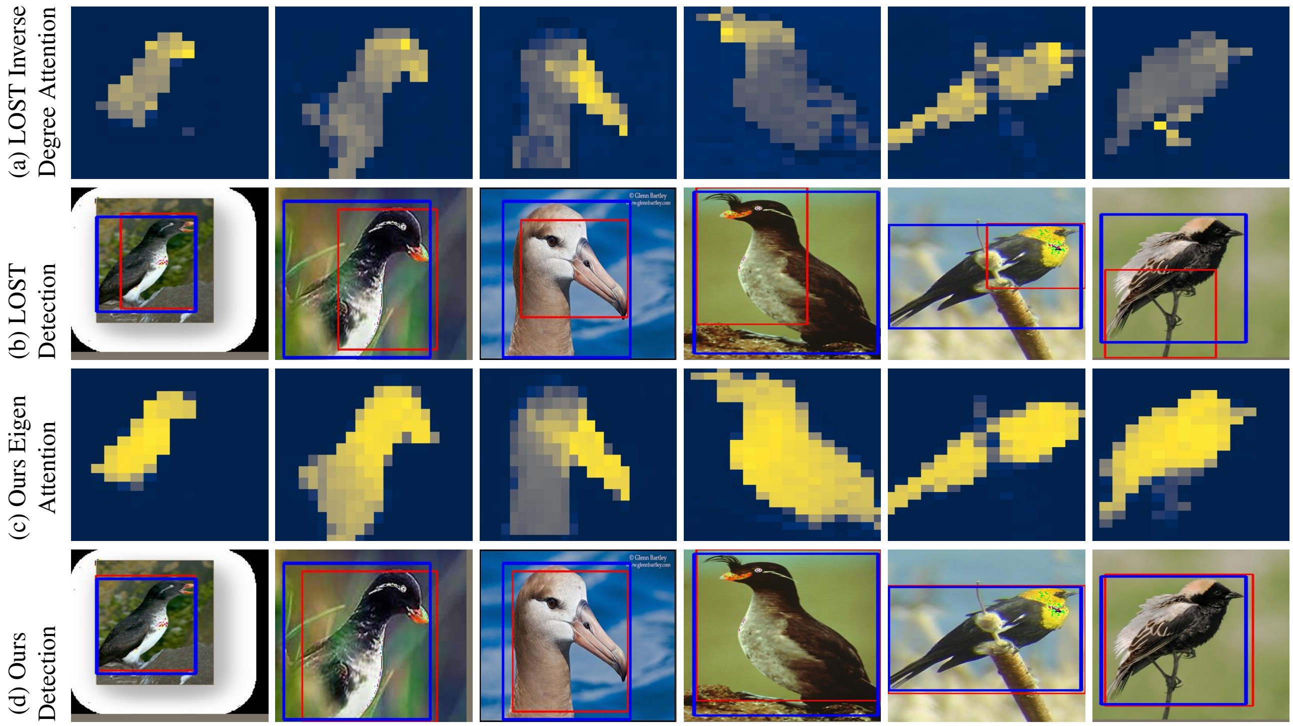


Figure 3. **Visual results for weakly supervised object localization on CUB [18].** In (a), we show the map of inverse degrees used to perform detection with LOST (b) [12]. For TokenCut, we illustrate the eigenvector in (c) used for detection in (d). Blue and Red bounding boxes indicate the ground-truth and the predicted bounding boxes respectively.

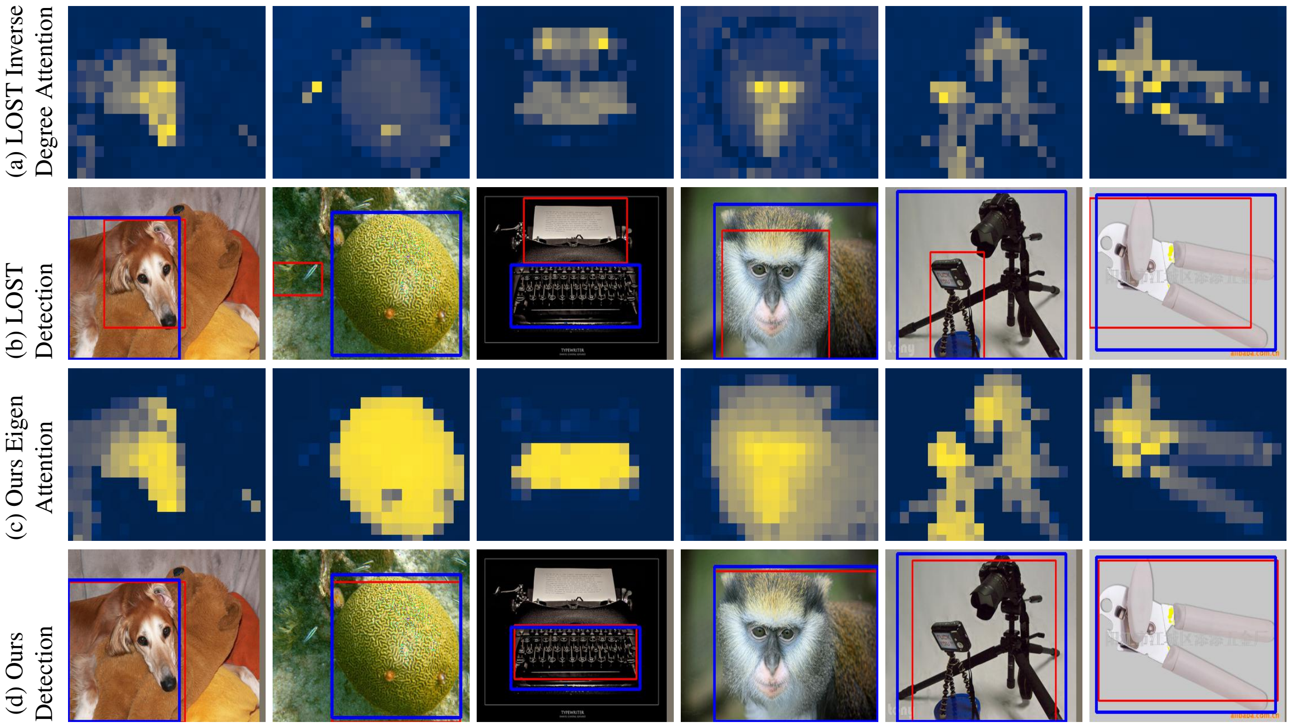


Figure 4. **Visual results of unsupervised single object discovery on Imagenet-1k [6].** In (a), we show LOST [12] the map of inverse degrees, which is used to perform detection (b). For TokenCut, we illustrate the eigenvector in (c) and the detection in (d). Blue and Red bounding boxes indicate the ground-truth and the predicted bounding boxes respectively.

6. Failure cases on CUB and Imagenet-1k

We illustrate additional failure cases in Fig. 5. Those failure cases can be organised into three categories: 1) Where TokenCut focus on the largest salient object, whereas the annotation is highlights a different object, shown in the first and the second column in Fig. 5. 2) Similar to LOST, Tokencut is not able to differentiate the connected objects, such as the third and the fourth column in Fig. 5. 3) In case of occlusion, neither LOST nor our approach can't detect the entire object, such as the last two columns in Fig. 5.

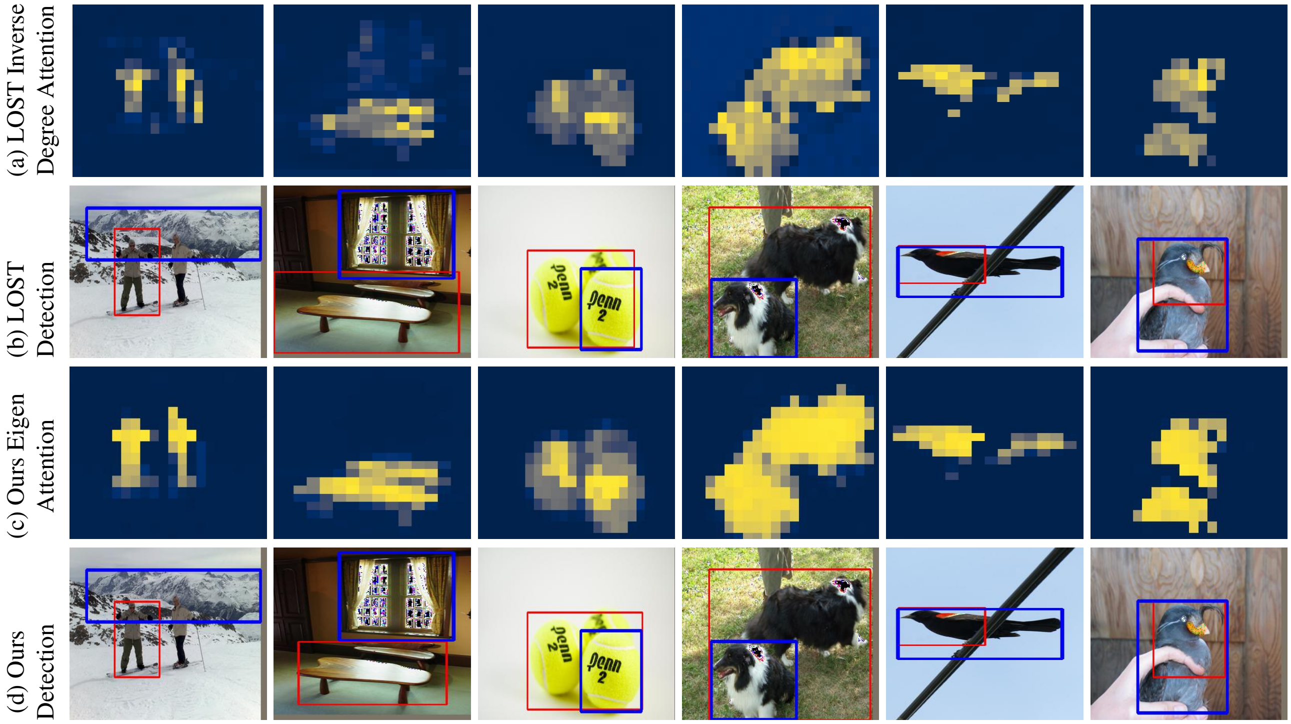


Figure 5. **Failure cases on Imagenet-1k [6] and CUB [18].** In (a), we show LOST [12] the map of inverse degrees, which is used to perform detection (b). For TokenCut, we illustrate the eigenvector in (c) and the detection in (d). Blue and Red bounding boxes indicate the ground-truth and the predicted bounding boxes respectively.

7. Analysis of Graph edge weight

In this section, we provide an ablation study on graph edge weight defininig on equation 4. We have tested to directly use the similarity score as edge weights (i.e., $\mathcal{E}_{ij} = S(x_i, x_j)$)). However, it is not possible because there may exist negative edge values, which violates the Normalized Cut algorithm assumption. Thus, we also tried thresholding the similarity score (i.e., $\mathcal{E}_{ij} = S(x_i, x_j)$ if $S(x_i, x_j) > \tau$, else ϵ). We obtain 68.9% on VOC07 dataset and 72% on VOC12 dataset, which is similar to our reported results.

8. Fine-tuning self-supervised transformers

For weakly supervised object localization, we use a pre-trained DINO model as our backbone and learn a linear classifier on the training set where we only have access to the class labels. We freeze the backbone weights and fine-tune a linear classifier. For CUB, We train with a SGD optimizer for 1000 epochs and set the batch size to 256 per GPU, distributed over 4 GPUs. The learning rate is linearly warmed during the first 50 epochs, then follows a cosine learning rate scheduler. We decay the learning rate from $\frac{\text{batch size}}{256} \times$5e-4 to 1e-6. The weight decay is set to 0.005. For ImageNet-1K, we use the models released by DINO. Other training setups and details can be found in the supplementary material.

9. Visual results for unsupervised saliency detecion on ECSSD, DUTS and DUT-OMRON

We present visual results for unsupervised saliency detecion on ECSSD [11], DUTS [19] and DUT-OMRON [22] in Fig. 6, 7 and 8 respectively.

For each dataset, we compare LOST segmentation, LOST + Bilateral Solver and our approch. The TokenCut provides better segmentation on objects. The performance is further improved with Bilateral Solver.

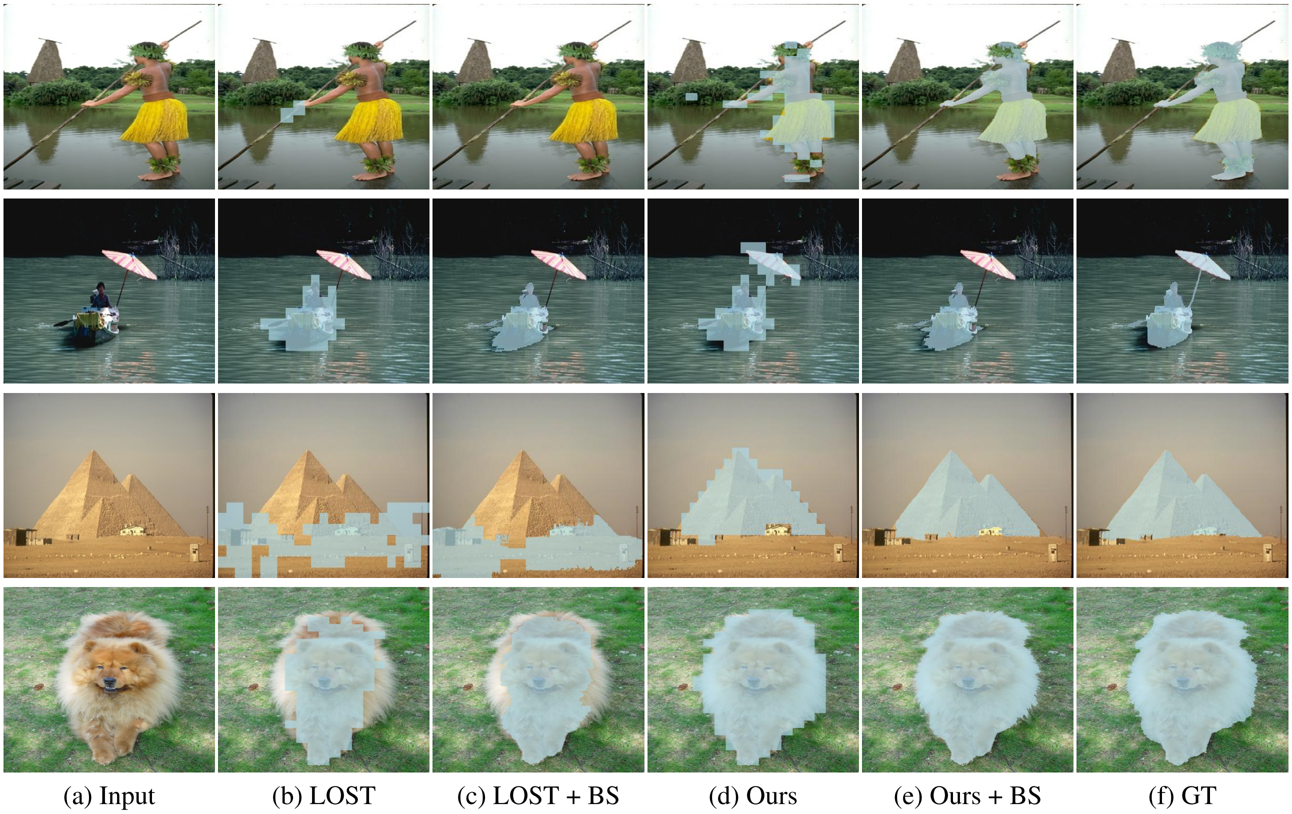
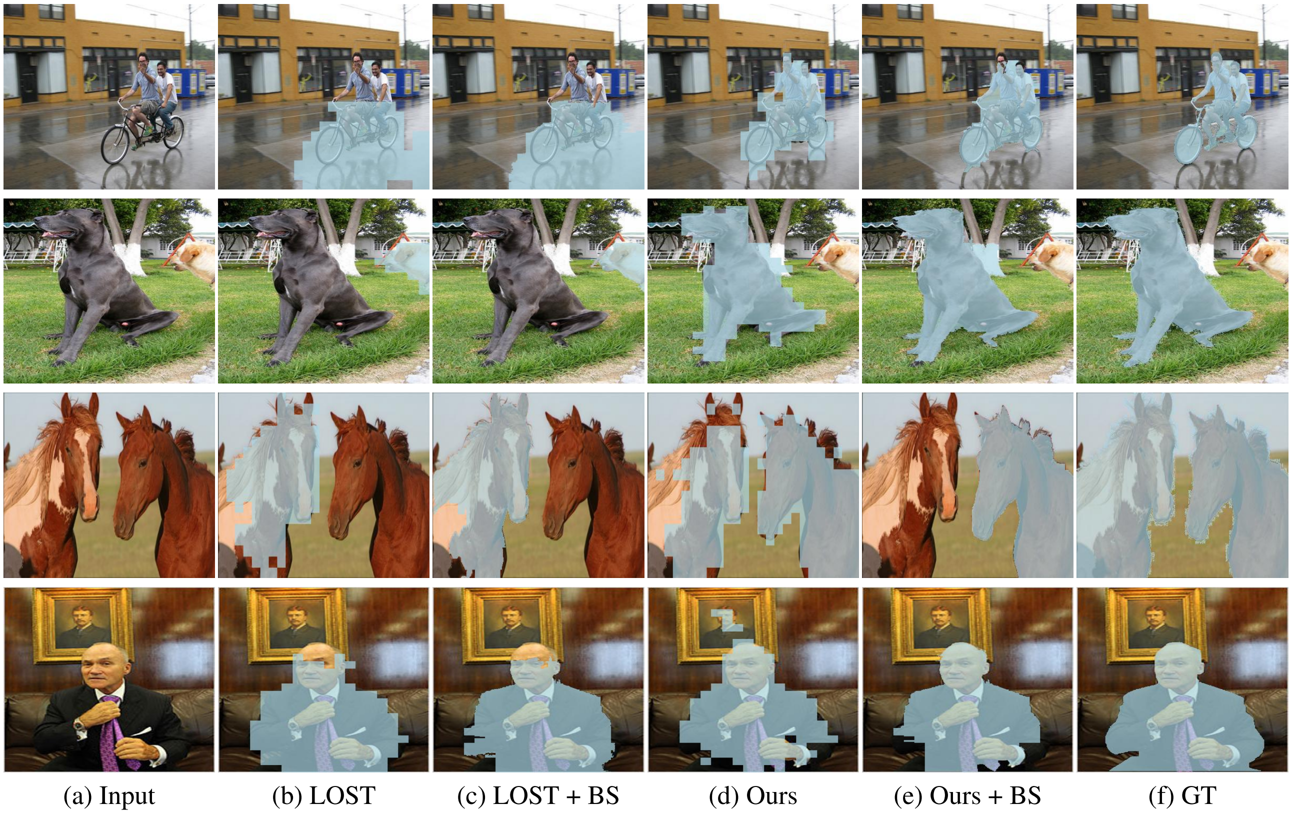
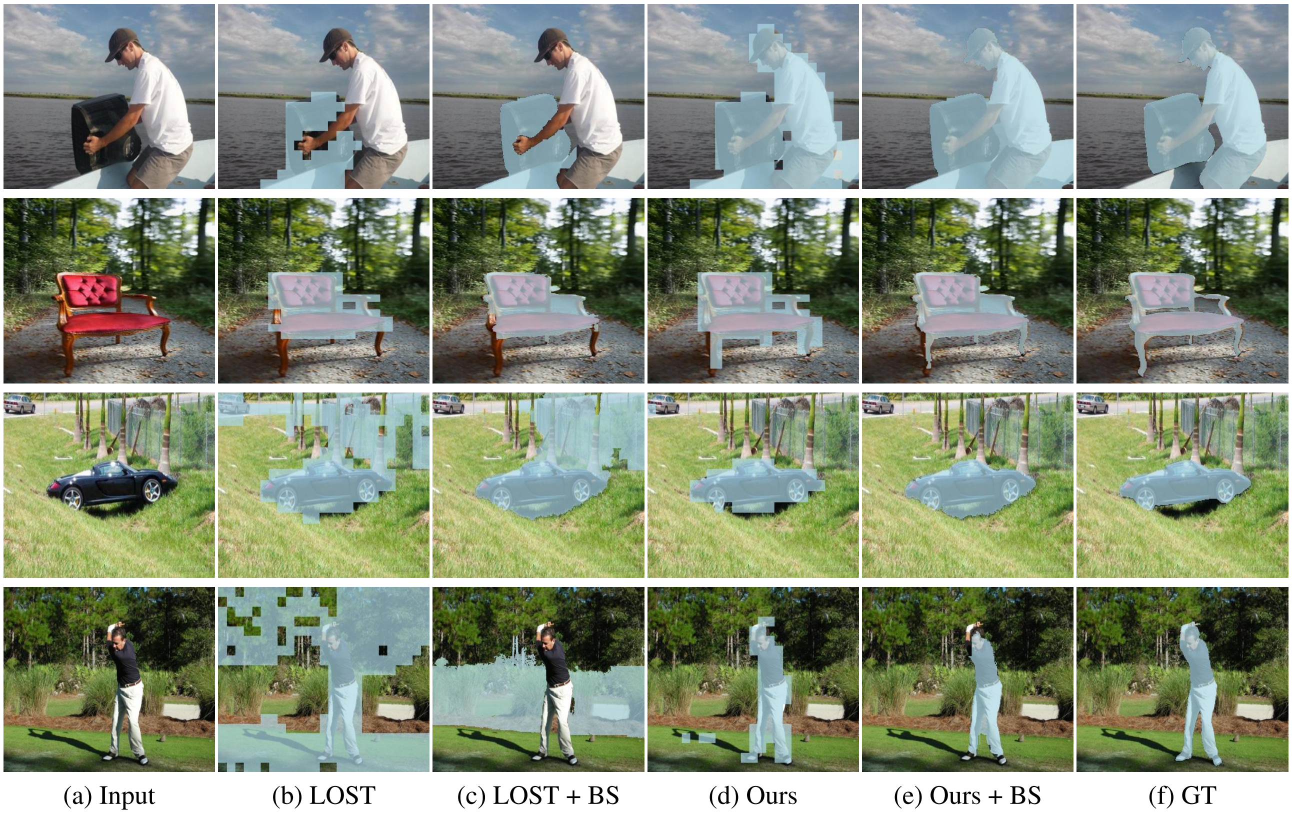


Figure 6. **Visual results of unsupervised segments on ECSSD [11]**



(a) Input (b) LOST (c) LOST + BS (d) Ours (e) Ours + BS (f) GT

Figure 7. **Visual results of unsupervised segments on DUTS [19]**



(a) Input (b) LOST (c) LOST + BS (d) Ours (e) Ours + BS (f) GT

Figure 8. **Visual results of unsupervised segments on DUT-OMRON [22]**

References

[1] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *ECCV*, 2020. 2

[2] 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. 1, 2, 3, 4

[3] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 1

[4] 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

[5] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, 2019. 2

[6] 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. 2, 4, 5, 6

[7] 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. 1, 2

[8] 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. 3

[9] 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. 2, 3, 4

[10] 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. 2

[11] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *TPAMI*, 2015. 7

[12] 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. 1, 2, 3, 4, 5, 6

[13] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 2

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[15] 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

[16] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *ECCV*, 2020. 2, 3, 4

[17] Huy V Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. *arXiv*, 2021. 2

[18] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. 2, 4, 5, 6

[19] 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. 7, 8

[20] 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. 2

[21] 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. 2

[22] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 2, 7, 8

[23] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2