DISCRIMINATIVELY MATCHED PART TOKENS FOR POINTLY SUPERVISED INSTANCE SEGMENTATION –SUPPLEMENTARY MATERIAL–

Anonymous authors
Paper under double-blind review

1 ADDITIONAL ABLATION STUDIES

We further conduct ablation studies of DMPT on PASCAL VOC 2012, Fig. 1 and Table. 1.

Top-ranked tokens. In part token allocation and token-classifier matching procedures, we replace each part token ($i.e.$ the top-1 confident patch token) with Z top-ranked confident patch tokens. In Fig. 1(a), our method achieves the best performance 53.9% by selecting 30 top-ranked patch tokens ($Z = 30$). More patch tokens ($Z = 40/50$) would introduce background noise, leading to a 1.0∼1.5% performance drop.

Number of tokens for segmentation supervision. Instead of using the point within of top-1 confident patch token (matched part token), we utilize more top-ranked points for each object part to supervise instance segmenter. As shown in the Fig.1(b), selecting 3 top-ranked patch tokens (points) of each part as supervision achieves best performance 55.1% mAP_{50}. However, using more part points or part region does not bring significant performance gain but drop.

Model size of SAM. In Fig. 1(c), we use different model sizes for segment anything model when carrying out DMPT-SAM. With ViT-H as backbone, DMPT-SAM achieves 59.8% mAP_{50}, 1.4% higher than that with ViT-B (58.4%).

Binarization threshold. In Fig. 1(d), we conduct experiments on the threshold when binarizing attention map to obtain foreground patch tokens $M+$ (in Section 3.2 of main paper). The threshold value (0.3) reports the best performance. Higher value (0.5) causes serious part missing, and lower value (0.1) introduces background noise, both of which contribute to performance drop.

Clustering Method. The part token allocation in DMPT is related to mean-shift algorithm. Other clustering methods can also be used to perform this procedure. We conduct ablation studies to replace the clustering method in part token allocation with the K-Means algorithm and PST (Yang et al., 2022) to obtain the clusters and part tokens. In Fig. 1(e), it can be seen that mean-shift achieves almost comparable performance with PST and K-Means, indicating that the parameterized part-based modeling mechanism is robust to fine-grained semantics and background noise.

Efficiency. We report comprehensive index on computing overhead in Table. 1.

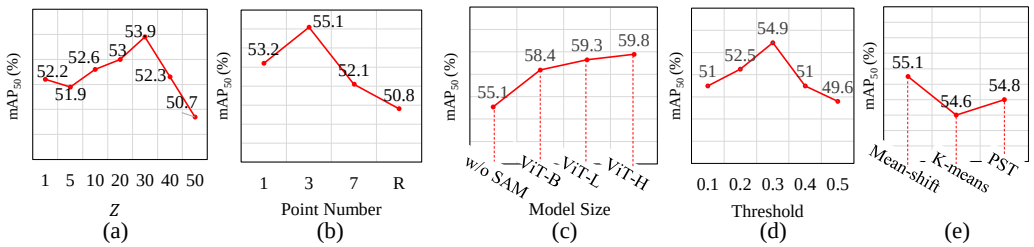

Figure 1: Ablation results. (a) Number of top-ranked tokens. (b) Number of matched tokens within each part for segmentation supervision ("R" denotes all tokens within the part cluster). (c) Model size of DMPT-SAM. (d) Experimental threshold for binarizing attention maps. (e) Clustering methods.

Method	GPU memory	Training Time	FPS	FLOPs	Backbone/Params.	mAP_{25}	mAP_{50}	mAP_{75}
BESTIE	12293M	34.2h	16.4	126.7G	HRNet-W48/63.6M	58.6	46.7	26.3
DMPT(ours)	15305M	9.7h	14.3	154.8G	ViT-S / 22.1M	69.8	55.5	27.8

Table 1: Comparison of GPU memory, computational complexity and training efficiency.

2 Additional Visualization Analysis

We provide additional visualization results of Fig.3 and Fig.4 in the main document. The results are shown in Fig. 2 and Fig. 3, respectively. We also visualize the instance segmentation results of the proposed DMPT, as shown in Fig. 4.

3 Details of Attention Map Generation and Pseudo Bounding-box Generation.

We implement a simple-yet-efficient method to generate pseudo bounding-boxes, which are used to supervise our detection head to output final detection results.

Attention Map Generation. In Sec.3.1 of the main document, a self-attention map A for each assigned query token is produced by a selection procedure. Firstly, we generate a set of self-attention maps $\mathbf{A} = \{A^l, l = 1, 2, ..., L\}$, where A^l denotes the attention map produced by l-th block in vision transformer and $L = 12$ the number of all blocks in vision transformers(Dosovitskiy et al., 2021). Additionally, $A^l_{i,j}$ is the attention value between patch token $\phi_{i,j}$ and the assigned query token in the l-th block. Following (Abnar & Zuidema, 2020), we update $A^l \in \mathbb{R}^{W \times H}$ as

$$A^l = A^L \otimes A^{L-1} \otimes ... \otimes A^l, \tag{1}$$

where $\otimes$ is the cross product.

Pseudo bounding-box generation. For each updated attention map A^l, we binarize it to a mask using empirically thresholds and generate a tight bounding-box B^l to enclose the maximum connected area on the foreground region. The pseudo bounding-boxes $\mathbf{B} = \{B^l, l = 1, 2, ..., L\}$ for each assigned query token (supervision point) are then used to extract object features using a 7×7 RoI-Align module (He et al., 2017). With object features as input, the dual-flow network, proposed by WSDDN (Bilen & Vedaldi, 2016), outputs two sets of classification scores for each supervision point. For one set of the classification scores, we use the $\mathrm{softmax}$ function over the dimension of category and obtain the classification probability $\mathbf{S}_{cls} \in \mathbb{R}^{L \times C}$, where L is the number of instances (pseudo bounding-boxes) of a bag (Chen et al., 2022) and C is the number of categories. Likewise, we do the same operation on the other set of classification scores over instance dimension and get the instance probability $\mathbf{S}_{ins} \in \mathbb{R}^{L \times C}$. Finally, we compute Hadamard product of classification probability and instance probability to get the object probability $\mathbf{S} \in \mathbb{R}^{L \times C} = \mathbf{S}^{cls} \odot \mathbf{S}^{ins}$. We select the bounding-box B with the highest object probability by $B = \arg\max_{B \in \mathbf{B}} \mathbf{S}$ as the pseudo bounding-box for its corresponding supervision point. To optimize the dual-stream network, we use binary cross-entropy loss (Bilen & Vedaldi, 2016) defined on the bag score, which is computed by summarizing the object probability over the instance dimension, as

$$\mathcal{L}_{BB} = \mathrm{BCE}(\sum_{l=1}^{L} \mathbf{S}, Y), \tag{2}$$

where Y is the binary label of the supervision point.

Performance. In Table. 2 and 3, we conduct experiments on PASCAL VOC 2007, PASCAL VOC 2012, MS-COCO 2014 and MS-COCO 2017 to evaluate the quality of object localization of generated pseudo bounding-boxes. Whether train detectors in an end-to-end fashion or a re-training fashion, our approach outperforms all the state-of-the-art weakly/pointly object detection methods by a significant margin, which sets a novel baseline for pointly supervised object detection using vision transformer.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Box-supervised detectors							
RetinaNet (Lin et al., 2017)	ResNet-50	36.5	55.4	39.1	20.4	40.3	48.1
Faster R-CNN (Ren et al., 2015)	ResNet-50	37.4	58.1	40.4	21.2	41.0	48.1
Cascade R-CNN (Cai & Vasconcelos, 2018)	ResNet-50	40.3	58.6	44.0	22.5	43.8	52.9
RetinaNet (Lin et al., 2017)	ResNet-101	38.5	57.6	41.0	21.7	42.8	50.4
Faster R-CNN (Ren et al., 2015)	ResNet-101	39.4	60.1	43.1	22.4	43.7	51.1
Cascade R-CNN (Cai & Vasconcelos, 2018)	ResNet-101	42.0	60.4	45.7	23.4	45.8	55.7
Faster R-CNN (Ren et al., 2015)	ViT-Small	41.1	62.4	44.6	24.7	44.2	55.1
imTED (Zhang et al., 2022)	ViT-Small	43.1	64.1	46.9	25.0	46.5	58.6
Cascade R-CNN	ViT-Small	44.7	64.0	48.6	26.7	47.6	59.9
Cascade R-CNN (Cai & Vasconcelos, 2018)	ViT-Base	52.0	71.3	56.5	35.1	55.7	68.3
imTED (Zhang et al., 2022)	ViT-Base	52.2	72.8	57.1	36.0	55.0	67.5
Point-supervised detectors (retrained)							
UFO (Ren et al., 2020b)	VGG-16	13.5	27.9	-	-	-	-
UFO (Ren et al., 2020b)	ResNet-50	13.2	28.9	-	-	-	-
†P2BNet (Chen et al., 2022) + Faster R-CNN	ResNet-50	22.1	47.3	18.3	11.5	24.8	30.4
P2BNet (Chen et al., 2022) + Faster R-CNN	ResNet-50	24.0	49.9	20.3	11.5	26.4	34.1
P2BNet (Chen et al., 2022) + Faster R-CNN	ViT-Small	19.1	43.5	13.6	7.6	19.1	31.3
Ours + Faster R-CNN	ViT-Small	29.4	54.3	28.4	11.4	32.2	47.8
Ours + imTED (Zhang et al., 2022)	ViT-Small	29.6	54.4	28.5	11.3	32.3	48.5
Ours + imTED (Zhang et al., 2022)	ViT-Base	32.7	57.4	32.4	13.2	36.4	52.5
End-to-end point-supervised detectors							
‡P2BNet (Chen et al., 2022) + Faster R-CNN	ResNet-50	21.1	46.2	17.6	10.7	23.5	29.6
Ours + Faster R-CNN (Ren et al., 2015)	ViT-Small	24.6	47.0	23.2	8.0	27.5	38.9
Ours + imTED (Zhang et al., 2022)	ViT-Small	25.0	47.8	23.3	8.2	25.5	42.3
Ours + imTED (Zhang et al., 2022)	ViT-Base	28.9	54.9	27.5	9.4	34.1	47.1

Table 2: Performance of our pseudo bounding-box generation method evaluated on MS-COCO 2017 val set. † denotes using the pseudo center point (Chen et al., 2022) as supervision, and ‡ represents implementation with the official code.

REFERENCES

Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. In ACL, pp. 4190–4197, 2020.

Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In IEEE CVPR, pp. 2846–2854, 2016.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: delving into high quality object detection. In IEEE CVPR, pp. 6154–6162, 2018.

Pengfei Chen, Xuehui Yu, Xumeng Han, Najmul Hassan, Kai Wang, Jiachen Li, Jian Zhao, Humphrey Shi, Zhenjun Han, and Qixiang Ye. Point-to-box network for accurate object detection via single point supervision. arXiv preprint arXiv:2207.06827, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

Ross B. Girshick. Fast R-CNN. arXiv preprint arXiv:1504.08083, 2015.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In IEEE ICCV, 2017.

Zeyi Huang, Yang Zou, B. V. K. Vijaya Kumar, and Dong Huang. Comprehensive attention self-distillation for weakly-supervised object detection. In NeurIPS 2020, 2020.

Qifei Jia, Shikui Wei, Tao Ruan, Yufeng Zhao, and Yao Zhao. Gradingnet: Towards providing reliable supervisions for weakly supervised object detection by grading the box candidates. In AAAI, pp. 1682–1690, 2021.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In IEEE ICCV, pp. 2999–3007, 2017.

Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. Training object class detectors with click supervision. In IEEE CVPR, pp. 180–189, 2017.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In NIPS, pp. 91–99, 2015.

Method	Backbone	VOC2007 mAP_{50}	VOC2012 mAP_{50}	COCO2014 AP	COCO2014 AP_{50}
Box-supervised detectors					
Fast R-CNN (Girshick, 2015)	VGG-16	66.9	65.7	-	-
Faster R-CNN (Ren et al., 2015)	VGG-16	69.9	67.0	-	-
Faster R-CNN (Ren et al., 2015)	ResNet-50	80.4	75.3	35.5	56.7
Faster R-CNN (Ren et al., 2015)	ViT-Small	82.3	77.4	38.3	60.3
imTED (Zhang et al., 2022)	ViT-Small	82.5	78.0	40.6	61.7
imTED (Zhang et al., 2022)	ViT-Base	87.0	85.3	48.0	69.4
Image-supervised detectors					
OICR (Tang et al., 2017) + Fast R-CNN	VGG-16	47.0	42.5	7.7	17.4
PCL (Tang et al., 2020) + Fast R-CNN	VGG-16	-	-	9.2	19.6
C-MIL (Wan et al., 2019)	VGG-16	50.5	46.7	-	-
WSOD2 (Zeng et al., 2019)	VGG-16	56.0	52.7	10.8	22.7
UFO (Ren et al., 2020b)	VGG-16	-	-	10.8	23.1
GradingNet-C-MIL (Jia et al., 2021)	VGG-16	54.3	50.5	11.6	25.0
ICMWSD (Ren et al., 2020a)	VGG-16	54.9	52.1	11.4	24.3
ICMWSD (Ren et al., 2020a)	ResNet-50	-	-	12.6	26.1
CASD (Huang et al., 2020)	VGG-16	56.8	53.6	12.8	26.4
CASD (Huang et al., 2020)	ResNet-50	-	-	13.9	27.8
Point-supervised detectors (retrained)					
Click (Papadopoulos et al., 2017)	AlexNet	49.1	-	-	18.4
UFO (Ren et al., 2020b)	VGG-16	-	-	12.4	27.0
UFO (Ren et al., 2020b)	ResNet-50	-	-	12.6	27.6
P2BNet (Chen et al., 2022) + Faster R-CNN	ResNet-50	63.4‡	60.0‡	19.4	43.5
Ours + imTED (Zhang et al., 2022)	ViT-Small	77.0	72.5	26.2	49.3
Ours + imTED (Zhang et al., 2022)	ViT-Base	80.1	78.8	29.3	54.7
End-to-end point-supervised detectors					
Ours + imTED (Zhang et al., 2022)	ViT-Small	75.8	71.4	23.0	44.5
Ours + imTED (Zhang et al., 2022)	ViT-Base	79.2	77.1	27.1	51.2

Table 3: Performance of our pseudo bounding-box generation method evaluated on PASCAL VOC 2007 $test$ set, PASCAL VOC 2012 $test$ set and MS-COCO 2014 val set. ‡ denotes performance provided by authors.

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 IEEE CVPR, pp. 10595–10604, 2020a.

Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G. Schwing, and Jan Kautz. Ufo^2: A unified framework towards omni-supervised object detection. In ECCV, pp. 288–313, 2020b.

Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In IEEE CVPR, pp. 3059–3067, 2017.

Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan L. Yuille. PCL: proposal cluster learning for weakly supervised object detection. IEEE TPAMI, pp. 176–191, 2020.

Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-MIL: continuation multiple instance learning for weakly supervised object detection. In IEEE CVPR2019, pp. 2199–2208, 2019.

Boyu Yang, Fang Wan, Chang Liu, Bohao Li, Xiangyang Ji, and Qixiang Ye. Part-based semantic transform for few-shot semantic segmentation. IEEE TNNLS, 33(12):7141–7152, 2022.

Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. WSOD2: learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In IEEE ICCV, pp. 8291–8299, 2019.

Xiaosong Zhang, Feng Liu, Zhiliang Peng, Zonghao Guo, Fang Wan, Xiangyang Ji, and Qixiang Ye. Integral migrating pre-trained transformer encoder-decoders for visual object detection. arXiv:2205.09613, 2022.

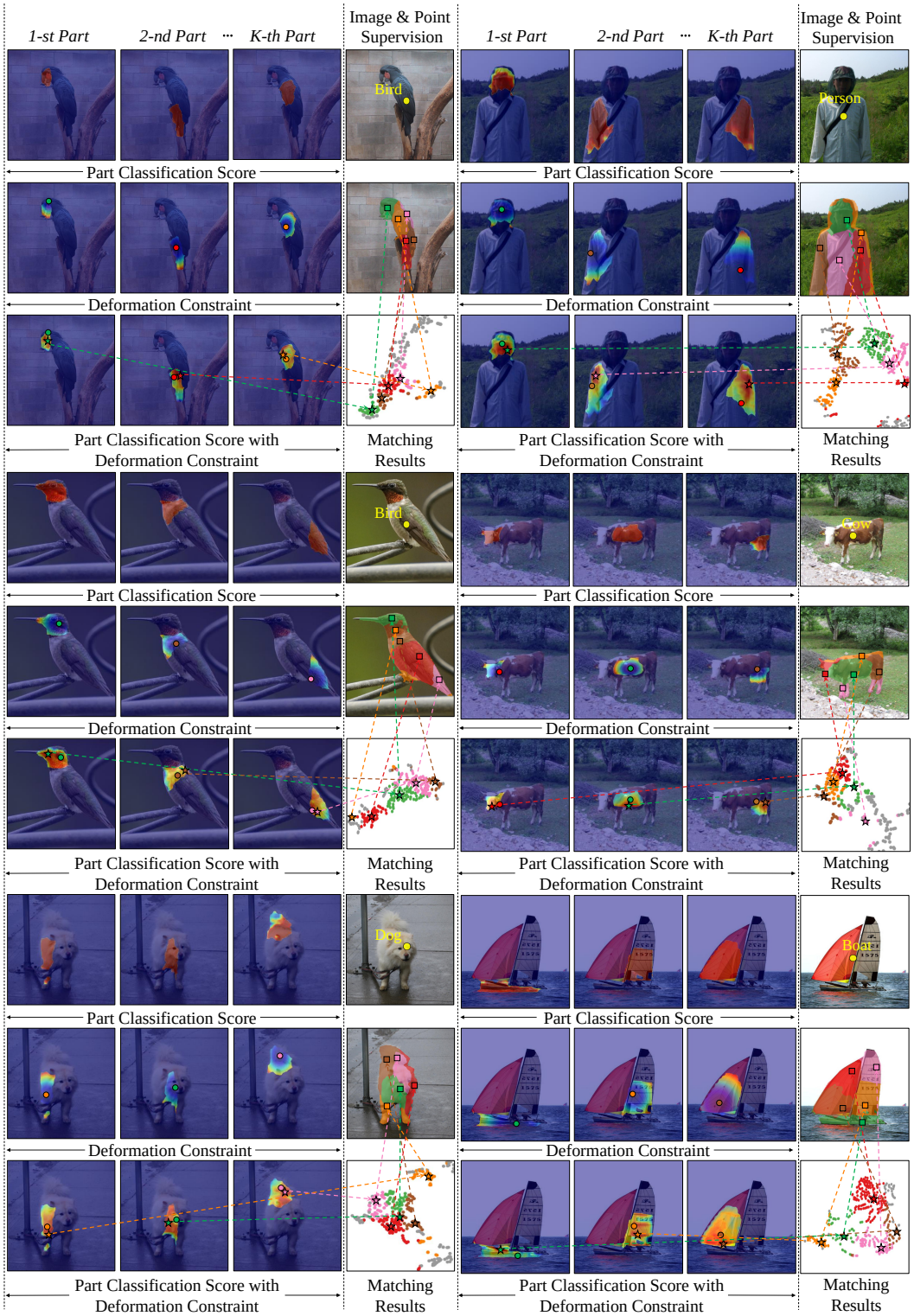

Figure 2: Visualization of part token allocation (heat-maps in 1-3 and 5-7 columns) and token-classifier matching results (4 and 8 columns). (Best viewed in color)

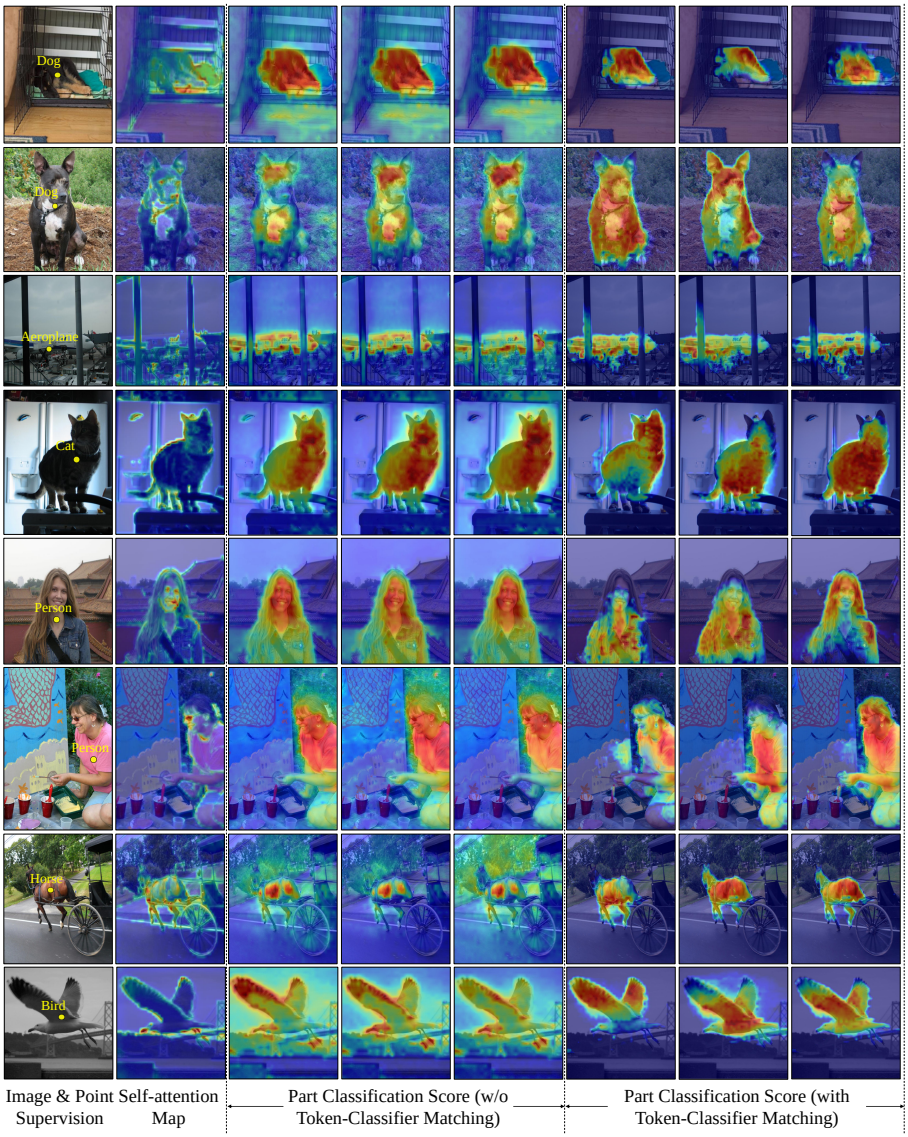

Figure 3: Visualization of the self-attention map (column 2), activation map of part classifier trained without token-classifier matching (columns 3-5), and activation map of part classifier trained with token-classifier matching (columns 6-8). (Best viewed in color)

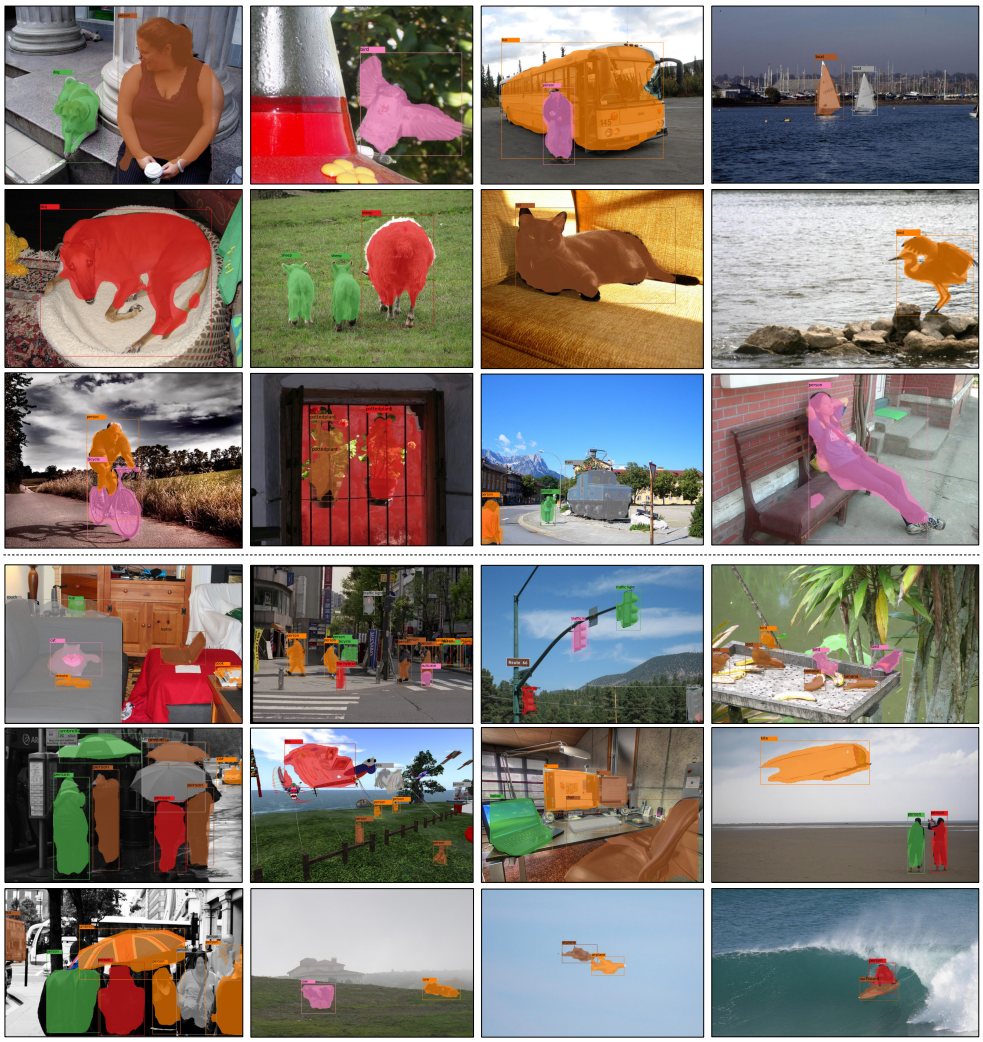

Figure 4: Visualization of instance segmentation results on PASCAL VOC 2012 val set (upper) and MS-COCO 2017 $tets\text{-}dev$ set (lower).