# Supplementary Materials

## A More Implementation Details

### A.1 Co-occurring Frequency among Categories

In this work, we propose SeCo to tackle the widespread co-occurrence issue in WSSS. Here, we calculate the co-occurring frequency among 20 categories in PASCAL VOC. As shown in Figure S1, each entry means the co-occurrence frequency between the query category(y-axis) and the key category(x-axis). The diagonal entries mean the single-category items. Almost all categories are obviously coupled with other categories. The {person} category has the highest co-occurring frequency with all categories. Other representative co-occurring pairs, such as {chair, sofa}, {bottle, dinning table} and {bus, car}, etc., also hold high co-occurring frequency. Therefore, apart from the incompleteness issue of CAM, the co-occurrence issue is another bottleneck of WSSS performance as well. SeCo acts in a 'separate and conquer' manner to decouple co-categories and gains the improvement.

### A.2 Training Configurations

For image decomposition, we subdivide integral images into patches with size $64 \times 64$. In practice, we automatically select 12 patches mainly around the foregrounds and group them into three types (i.e., background, category and uncertain) according to the generating threshold of category tags $\varphi$. The $\varphi$ is set as $0.2$ in our experiments. The capacity of the semantic patch reservoir is $4608$. The rectification threshold value $\sigma$ in the proposed tag rectification strategy is set as $0.1$. Encoders in the dual-teacher single-student framework all adopt ViT-B [3] as the backbone and are initialized with pre-trained weights on ImageNet [8]. Our decoder adopts a simple segmentation head with four $3 \times 3$ convolutional layers.

The temperature factors for the contrastive loss $L_{LiG}$ and $L_{LiL}$, i.e., $\tau_g$ and $\tau_l$, are set as $0.1$ and $0.08$, respectively. The loss weight factors $(\alpha, \beta, \gamma)$ are set as $(0.5, 0.5, 0.12)$. All the hyper-parameters above follow grid searching strategy. The learning rate in our experiments is set as $1e-6$. Following the training strategy in [9, 10], we use AdamW optimizer to train SeCo with a polynomial scheduler. The total training iteration for experiments on PASCAL VOC and MS COCO is set as 20k and 80k, respectively. For robustness, we update the category knowledge after $600$ and $6000$ iterations on PASCAL VOC and MS COCO, respectively. All experiments are conducted on RTX 3090 GPU.

### A.3 Settings for the Auxiliary Pseudo Mask

As we mentioned, we incorporate an auxiliary classification head to generate a category tag $t_i$ from the auxiliary pseudo mask and assign it to each local patch $x_i$. It guides the class-specific contrast to decouple co-contexts. The reason for such operation is that features from the last blocks of ViT intend to be over-smoothed [4, 11] and the auxiliary pseudo masks directly from the last block may be unreliable. In order to generate CAMs with diverse semantics at the beginning of our training, we generate the auxiliary pseudo masks from the intermediate block based on the observation that features from the intermediate blocks preserve more semantic diversity. We carry out detailed experiments to choose the best intermediate features, as reported in Table S1 (j). It turns out features from 10-th blocks of ViT-B (12 Transformer blocks in total) are competent to avoid the potential downside of ViT and generate more accurate category tags.

The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024 (CVPR 2024).

# B More Experiment Results

## B.1 Sensitivity of Hyper-parameter

**Patch Representation.** Here, the sensitivity analysis of four hyper-parameters for patch representation (i.e., the crop size of the local patches, the number of the local patches, the capacity of the reservoir, and the momentum to update the local teacher) are performed on VOC val setas reported in Table S1 (a-d). The patch size and the patch number are two key parameters as they directly affect the spatial separation of co-contexts in the image decomposition. SeCo keeps satisfying performance when the parameters change. Moreover, SeCo remains consistent with variations in reservoir size and momentum, which shows the robustness of our method. In our method, the default values of patch numbers, patch size, reservoir capacity and momentum are $12, 64 \times 64, 4608, 0.999$, respectively.

**Temperature Factor.** Temperature factor directly affects the sharpness of the learned contrastive representation. In Table S1 (e), we report the impact of global temperature factor $\tau_g$ in $L_{LiG}$ on the semantic performance of PASCAL VOC val set. In Table S1 (f), we report the impact of local temperature factor $\tau_l$ in $L_{LiL}$ on the semantic performance. It shows that SeCo remains consistent with variations of the temperature factors. In our experiments, $\tau_g = 0.1$, $\tau_l = 0.08$ can achieve the most satisfying performance.

**Loss Weights.** Table S1 (g) reports the impact of loss weights in $L_{SeCo}$ on the semantic prediction performance. $\alpha$ and $\beta$ are the weights for $L_{LiG}$ and $L_{LiL}$, respectively. It can be observed that $\alpha = 0.5$ and $\beta = 0.5$ achieves the best prediction performance on VOC val set, and SeCo can also yield satisfying results with other values.

**Tag Generating Threshold.** In our experiments, tag generating threshold $\varphi$ is leveraged to determine the type of each category tag according to the corresponding pseudo mask patch. Table S1 (h) shows that SeCo produces the best results when $\varphi$ is set as $0.2$.

**Tag Rectification Threshold.** We design a tag rectification threshold $\sigma$ to determine if the raw category tag is correct. Once the ratio of positive votes (i.e., the tag is not correct) to negative votes exceeds a threshold, we consider the tag noisy. As reported in Table S1 (i), SeCo produces the best results when $\sigma = 10$.

**Index of Intermediate Blocks.** We explore the impact of different intermediate blocks to generate auxiliary pseudo masks and report the influence on the performance of decoupling co-occurrence. As shown in Table S1 (j), the features from intermediate blocks help SeCo generate more diverse CAMs and allocate reliable category tags. The performance heavily drops when directly adopting the pseudo mask from the final block. It is observed that SeCo achieves the best performance with $\lambda = 10$.

## B.2 Convergence Speed

In Figure S3, we visualize the convergence of three methods (i.e., SeCo and the other two single-staged methods [9, 10]) in terms of mIoU as the number of iterations increases. It can be seen that our SeCo has a faster convergent speed and achieves more favorable performance.

## B.3 Quantitative Category-wise Performance

**Confusion Ratio Comparisons.** In Table S2, we specifically report the confusion ratio for each category of SeCo and other single-staged competitors [9, 10]. It demonstrates that our method consistently holds lower confusion ratio on 14 categories. With the same backbone as ToCo [10], our method shows significantly lower confusion ratio at those objects with high co-occurring frequency, such as {boat $(-79\%)$},{train $(-21\%)$},{aeroplane $(-12\%)$}, etc., which shows the superiority of SeCo suppressing false activation.

**Comparisons with Existing Methods Tackling Co-occurrence.** As reported in Table S3, we evaluate per-category performance with our method SeCo and other impressive methods tackling co-occurrence with external supervision, such as CDA [12], EPS [7], W-OoD [6]. For those most representative objects with high co-occurring frequency as shown in Figure S1, such as {chair}, {dining table}, {person}, {bottle}, SeCo consistently outperforms other competitors without any additional data or elaborate designs, which demonstrates the efficiency of SeCo tackling co-occurrence issue. On the other hand, SeCo is trained in a single-staged paradigm. SeCo not only has more efficient training process, but also remains superior over other single-staged and multi-staged methods on 11 categories.

### B.4 More Qualitative Results

More qualitative semantic segmentation results of co-occurrence cases from PASCAL VOC and MS COCO are presented in Figure S3 and Figure S4, respectively. SeCo holds superior performance on both datasets compared to the recent competitors [9, 10]. Both results show that SeCo accurately localise the co-occurrence objects by suppressing the false positives from backgrounds and foregrounds.

More visualization results of CAM on PASCAL VOC are shown in Figure S5. Our methods can precisely differentiate the co-occurring foregrounds and filter out the distracting backgrounds, which demonstrates the efficacy of SeCo tackling co-occurrence issue.

## References


[1] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4253–4262, 2020.

[2] Liyi Chen, Chenyang Lei, Ruihuang Li, Shuai Li, Zhaoxiang Zhang, and Lei Zhang. Fpr: False positive rectification for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1108–1118, 2023.

[3] 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. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Vision transformers with patch diversification. *arXiv preprint arXiv:2104.12753*, 2021.

[5] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021.

[6] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022.

[7] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5495–5505, 2021.

[8] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[9] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16855, 2022.

[10] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2303.01267*, 2023.

[11] Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen Lee, and James T Kwok. Revisiting over-smoothing in bert from the perspective of graph. *arXiv preprint arXiv:2202.08625*, 2022.

[12] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7004–7014, 2021.

Table S1: Impact of hyper-parameters on VOC val set. M: pseudo mask. Seg.: semantic prediction. Temp.: temperature. T.H.: threshold value.

(a) The patch size.

| Patch Size | M | Seg. |
|---|---|---|
| $32 \times 32$ | 72.8 | 72.4 |
| **64×64** | **75.2** | **74.0** |
| $96 \times 96$ | 73.8 | 72.6 |
| $112 \times 112$ | 72.6 | 71.9 |

(b) The patch number.

| Patch Number | M | Seg. |
|---|---|---|
| 16 | 73.4 | 72.8 |
| **12** | **75.2** | **74.0** |
| 10 | 73.6 | 72.9 |
| 8 | 72.1 | 71.3 |

(c) The reservoir capacity.

| Reservoir Size | M | Seg. |
|---|---|---|
| 2304 | 73.4 | 71.8 |
| **4608** | **75.2** | **74.0** |
| 6912 | 74.2 | 73.1 |
| 9216 | 74.5 | 73.3 |

(d) Momentum in EMA.

| Momentum | M | Seg. |
|---|---|---|
| 0.1 | 73.7 | 72.1 |
| **0.999** | **75.2** | **74.0** |
| 0.99 | 74.0 | 72.8 |
| 0.9 | 73.5 | 72.2 |

(e) Global Temperature Factor.

| Global Temp. $\tau_g$ | M | Seg. |
|---|---|---|
| **0.1** | **75.2** | **74.0** |
| 0.2 | 73.7 | 72.2 |
| 0.5 | 71.6 | 69.8 |
| 0.8 | 70.1 | 68.9 |

(f) Local Temperature Factor.

| Local Temp. $\tau_l$ | M | Seg. |
|---|---|---|
| **0.08** | **75.2** | **74.0** |
| 0.1 | 74.2 | 71.5 |
| 0.2 | 73.6 | 71.7 |
| 0.5 | 72.5 | 70.7 |

(g) Loss Weights.

| Loss Weights | $\alpha$ | $\beta$ |
|---|---|---|
| 0.1 | 71.2 | 73.4 |
| 0.3 | 70.7 | 72.7 |
| **0.5** | **74.0** | **74.0** |
| 0.8 | 72.7 | 71.9 |

(h) Generating Threshold of Tags.

| Generating T.H. $\varphi$ | M | Seg. |
|---|---|---|
| 0.01 | 75.0 | 73.4 |
| 0.1 | 74.8 | 73.5 |
| **0.2** | **75.2** | **74.0** |
| 0.5 | 72.9 | 70.2 |

(i) Rectifying Threshold of Tags.

| Rectifying T.H. $\sigma$ | M | Seg. |
|---|---|---|
| 1 | 71.7 | 70.1 |
| 5 | 73.1 | 72.2 |
| **10** | **75.2** | **74.0** |
| 20 | 74.6 | 72.6 |

(j) Block Index.

| Block Index $\lambda$ | M | Seg. |
|---|---|---|
| 12th | 36.4 | 35.9 |
| 11th | 66.5 | 65.1 |
| **10th** | **75.2** | **74.0** |
| 9th | 71.5 | 70.2 |

Table S2: Per-category confusion ratio comparison with recent methods [9, 10] on VOC val set.

| Methods | bkg | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | mCR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AFA [9]CVPR'22 | 0.05 | 0.12 | 2.16 | 0.14 | 0.42 | 0.20 | 0.10 | **0.06** | 0.20 | 1.09 | 0.07 | **0.16** | 0.12 | 0.13 | 0.27 | 0.34 | 0.39 | 0.09 | 0.57 | 0.63 | 0.49 | 0.36 |
| ToCo [10]CVPR'23 | 0.04 | 0.19 | **0.84** | 0.42 | 1.11 | **0.13** | 0.11 | 0.11 | 0.02 | 0.65 | 0.03 | 0.32 | **0.08** | 0.09 | 0.21 | 0.06 | 0.59 | 0.06 | 0.77 | 0.75 | **0.34** | 0.32 |
| **SeCo(Ours)** | **0.04** | **0.07** | 1.22 | **0.10** | **0.32** | 0.17 | **0.09** | 0.07 | **0.02** | **0.48** | **0.02** | 0.28 | 0.09 | **0.05** | **0.17** | **0.06** | **0.29** | **0.04** | **0.35** | **0.54** | 0.45 | **0.23** |

Table S3: Per-category performance comparison with recent methods that focus on tackling co-occurrence on VOC val set. IoU is the metric to validate the efficiency of tackling co-occurrence issue.

| Methods | bkg | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | mIou |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Multi-staged methods | | | | | | | | | | | | | | | | | | | | | | |
| CDA [12]CVPR'21 | 89.1 | 69.7 | 34.5 | 86.4 | 41.3 | 69.2 | 81.3 | 79.5 | 82.1 | 31.1 | 8.3 | 50.8 | 80.6 | 76.1 | 72.2 | 77.6 | 48.8 | 81.2 | 42.5 | 60.6 | 54.3 | 66.1 |
| AdvCAM [5]CVPR'21 | 89.5 | 76.9 | 33.5 | 80.3 | 63.7 | 68.6 | **89.7** | 77.9 | 87.6 | 31.6 | 77.2 | 36.2 | 82.6 | 78.7 | 73.5 | 69.8 | 51.9 | 81.9 | 43.8 | 70.9 | 52.6 | 67.5 |
| EPS [7]CVPR'21 | 91.7 | **89.4** | **40.6** | 84.7 | 67.0 | 71.6 | 87.8 | 82.7 | 87.4 | 33.6 | 81.9 | 37.3 | 82.5 | 82.9 | 76.6 | 82.8 | 54 | 79.7 | 39.1 | **85.4** | 51.7 | 71.0 |
| W-OoD [6]CVPR'22 | 91.0 | 80.1 | 34.1 | 88.1 | 64.8 | 68.3 | 87.4 | **84.4** | 89.8 | 30.1 | **87.8** | 34.7 | **87.5** | **85.9** | **79.8** | 75.0 | 56.4 | 84.5 | 47.8 | 80.4 | 46.4 | 70.7 |
| FPR [2]ICCV'23 | 91.4 | 81.8 | 35.1 | 82.4 | **68.7** | 73.7 | 88.8 | 80.5 | 85.9 | 33.3 | 82.4 | 45.3 | 82.5 | 81.6 | 72.9 | 78.5 | 50.7 | 82.6 | 46.5 | 83.1 | 49.1 | 70.3 |
| Single-staged methods | | | | | | | | | | | | | | | | | | | | | | |
| 1Stage [1]CVPR'20 | 88.7 | 70.4 | 35.1 | 75.7 | 51.9 | 65.8 | 71.9 | 64.2 | 81.1 | 30.8 | 73.3 | 28.1 | 81.6 | 69.1 | 62.6 | 74.8 | 48.6 | 71.0 | 40.1 | 68.5 | **64.3** | 62.7 |
| AFA [9]CVPR'22 | 89.7 | 79.3 | 30.3 | 79.8 | 64.6 | 62.0 | 82.3 | 66.5 | 80.5 | 29.6 | 83.9 | 45.0 | 80.2 | 76.0 | 70.1 | 76.1 | 51.8 | 84.8 | 44.6 | 59.6 | 52.8 | 66.0 |
| **SeCo(Ours)** | **92.5** | 86.3 | 39.8 | **88.8** | 68.4 | **78.5** | 88.1 | 80.1 | **90.4** | **38.3** | 84.5 | **52.4** | 86.9 | **85.9** | 73.5 | **84.4** | **62.4** | **89.6** | **57.4** | 62.2 | 62.6 | **74.0** |

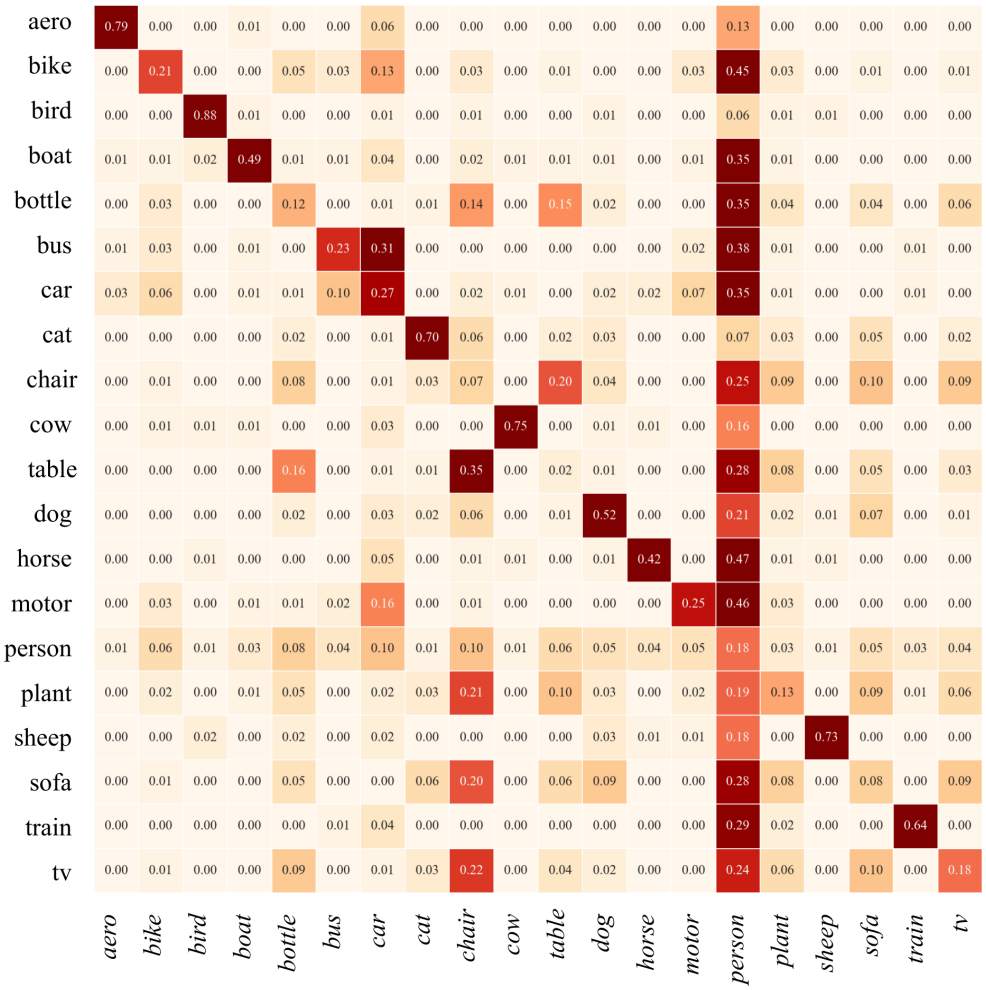


Figure S1: Co-occurrence frequency matrix among category labels of PASCAL VOC train set. Each entry means the co-occurrence frequency between the query category(y-axis) and the key category(x-axis). The diagonal entries mean the single-category items.

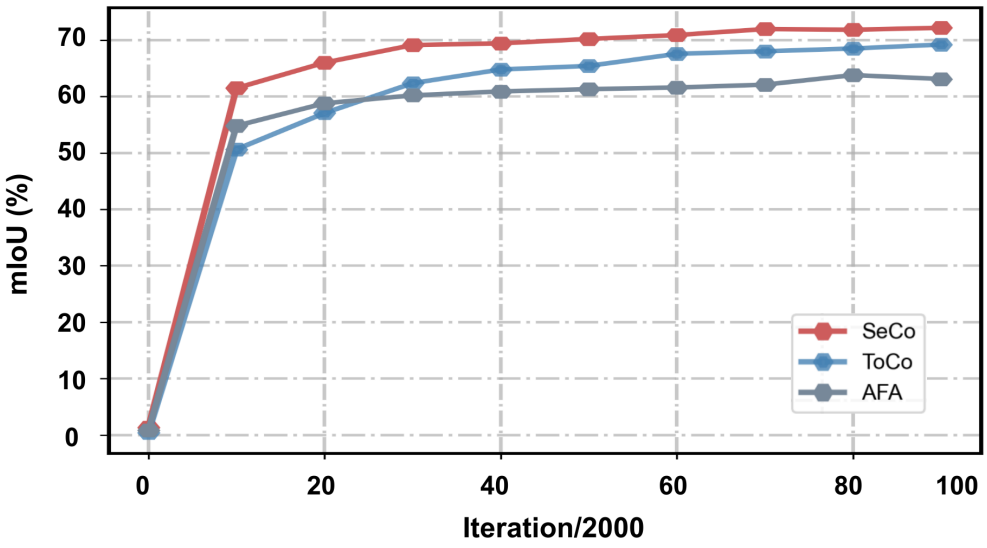


Figure S2: Performance of convergence speed of SeCo compared to AFA [9] and ToCo [10]. The experiment is implemented on VOC val set and the semantic segmentation result is evaluated with mIoU.

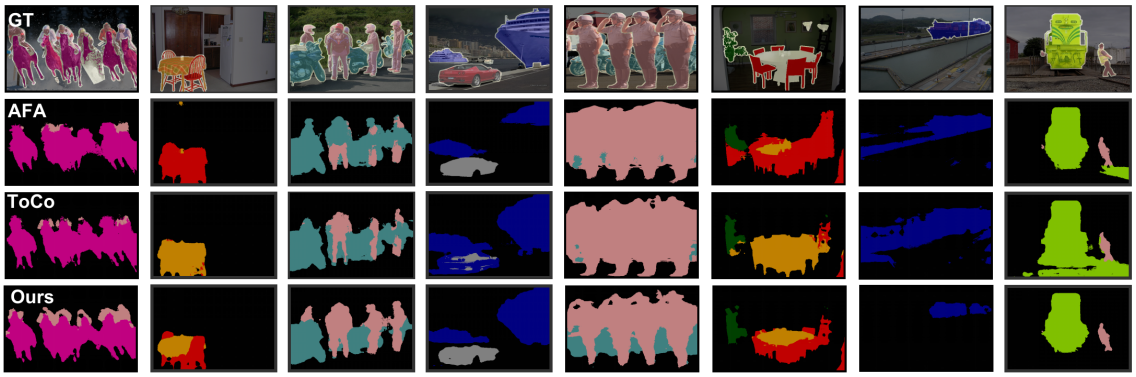


Figure S3: Qualitative segmentation performance on PASCAL VOC. The comparisons are conducted among AFA [9], ToCo [10] and ours. SeCo precisely differentiates the co-occurring foregrounds and filters out distracting backgrounds.

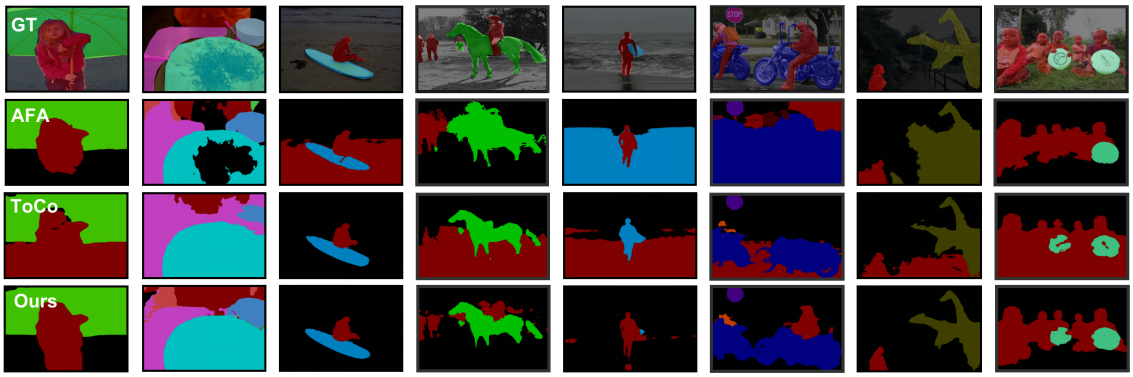


Figure S4: Qualitative segmentation performance on MS COCO. The comparisons are conducted among AFA [9], ToCo [10] and ours. SeCo accurately localises the co-occurring objects.

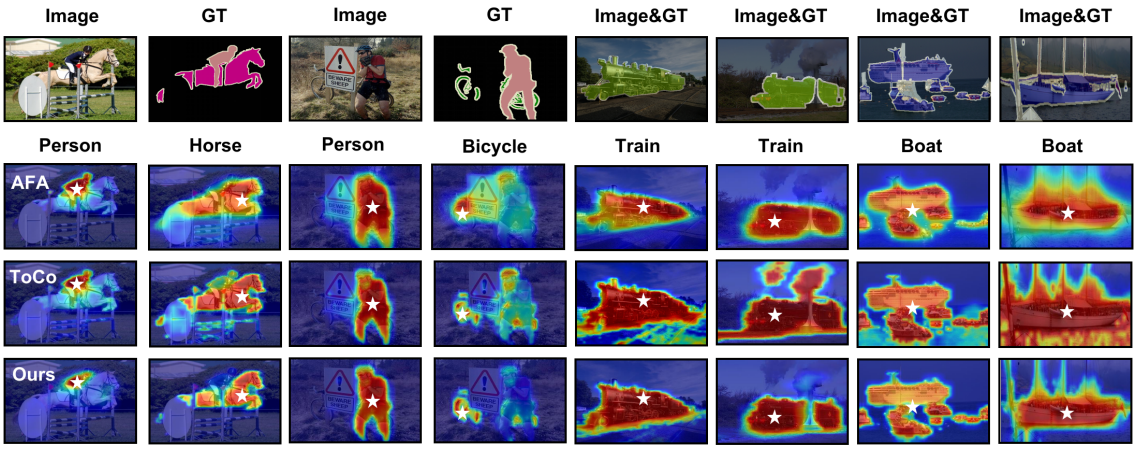


Figure S5: Qualitative CAM visualization of SeCo, AFA [9], and ToCo [10]. SeCo effectively suppresses the false positive pixels from backgrounds and foregrounds