Supplementary material for: Cross-domain Semantic Decoupling for Weakly-Supervised Semantic Segmentation

Zaiquan Yang
zaiquanyangcat@gmail.com
Zhanghan Ke
kezhanghan@outlook.com
Gerhard Hancke
gp.hancke@cityu.edu.hk
Rynson Lau
rynson.lau@cityu.edu.hk

Department of Computer Science City
University of Hong Kong,
Hong Kong SAR, China

1 Overview

In Supplementary Material, we provide implementation details of the proposed CSD framework and the data loading and training procedure of the whole algorithm, which help future research and conform with reproducibility principles. We also give more discussion and comparison with other related work and demonstrate the superiority of the proposed method.

2 Implementation Details.

To validate the applicability of CSD, we deploy it on typical baseline methods (i.e., IRN [1] and MCTformer [8]). The general training pipline includes multi-label image classification, a pseudo-mask generation, and the final segmentation training three stages. We strictly follow the same settings (e.g., image augmentation) as reported in the official codes. Specially, for MCTFormer [8] baseline, Deit-S that pre-trained on ImageNet [3] is adopted as classification backbone with batch size as 64. Training images are resized to 256 $\times$ 256 and then cropped into 224 $\times$ 224. For IRN [1], ResNet50 [4] that pre-trained on ImageNet[3] is adopted as classification backbone with batch size as 16. Training images are croped as 512 $\times$ 512. When imposing our proposed CSD on MCTformer and IRN, we set λ_1 = 0.01 and λ_2 = 0.1 in order to keep balance with classification loss. As for the training epoch, learning rate, learning rate decay policy, weight decay rate, and optimizer, we follow the same setting as MCTformer and IRN. At test time of segmentation model, we used multi-scale testing and CRFs with the hyper-parameters suggested in [2] for post-processing.

3 Algorithm Pipeline

Before the optimization pipeline of CSD, we first extract the original foreground mask $\mathcal{M}_{fg}$ and background mask $\mathcal{M}_{bg}$ based on the baseline methods. Then during the pipeline of CSD, we augment the multi-label image by pasting the foreground mask and background mask of single-label image. By imposing the semantic activation consistency learning, we alleviate the coupling between multiple classses and obtain more precise semantic activation map as the pseudo labels of segmentation. We provide steps of the data loading and training in Algorithm 1.

Algorithm 1 Cross-domain Semantic Decoupling.

Input:

The training dataset images $\mathcal{X}$ and corresponding labels $\mathcal{L}$;

The foreground object mask $\mathcal{M}_{fg}$ and corresponding background mask $\mathcal{M}_{bg}$.

```
while not done do
    X_m^i, L^i ← Load one multi-label sample;
    X_s^j, L^j ← Resample one single-label image according co-occurrence;
    X_fg, X_bg ← Crop foreground and background.
    X_sm^fg ← Paste X_fg into X_m^i;
    X_sm^bg ← Paste X_bg into X_m^i;
    M_s, M_sm^fg, M_sm^bg ← forward X_s, X_sm^fg, X_sm^bg;
    L_bg, L_fg ← KL(M_s, M_sm^fg), KL(M_s, M_sm^bg) ;
    L_cls ← CE(M_m, L^i);
    Train Network ← L_cls + L_fg + L_bg ;
end while
```

while not done **do**
- $\mathcal{X}_m^i$, $\mathcal{L}^i \leftarrow$ Load one multi-label sample;
- $\mathcal{X}_s^j$, $\mathcal{L}^j \leftarrow$ Resample one single-label image according co-occurrence;
- $\mathcal{X}_{fg}$, $\mathcal{X}_{bg} \leftarrow$ Crop foreground and background.
- $\mathcal{X}_{sm}^{fg} \leftarrow$ Paste $\mathcal{X}_{fg}$ into $\mathcal{X}_m^i$;
- $\mathcal{X}_{sm}^{bg} \leftarrow$ Paste $\mathcal{X}_{bg}$ into $\mathcal{X}_m^i$;
- $\mathcal{M}_s$, $\mathcal{M}_{sm}^{fg}$, $\mathcal{M}_{sm}^{bg} \leftarrow$ forward $\mathcal{X}_s$, $\mathcal{X}_{sm}^{fg}$, $\mathcal{X}_{sm}^{bg}$;
- $\mathcal{L}_{bg}$, $\mathcal{L}_{fg} \leftarrow \mathrm{KL}(\mathcal{M}_s, \mathcal{M}_{sm}^{fg})$, $\mathrm{KL}(\mathcal{M}_s, \mathcal{M}_{sm}^{bg})$;
- $\mathcal{L}_{cls} \leftarrow \mathrm{CE}(\mathcal{M}_m, \mathcal{L}^i)$;
- Train Network $\leftarrow \mathcal{L}_{cls} + \mathcal{L}_{fg} + \mathcal{L}_{bg}$;

end while

4 Discussion and Comparison.

The proposed CSD framework propose a novel method designed for decoupling multiple target-objects from the cross-domain perspective. Our present dual background-foreground copy-and-paste scheme for balanced attention consistency. The benefits are twofold: the first is to avoid the over activation of foreground categories. The second is to promote the decoupling and differentiation between the background category and the other categories in the foreground. In prior work, CDA [7] also leverage the copy-and-paste for decoupling the high correlation between objects and their contextual background. AttBN [5] transfers the foreground prior from a simple single-label dataset to another complex multi-label dataset by adversarial learning [6]. However, they still cannot further narrow the gap between classification and segmentation tasks from the pixel level. In general, our CSD can effectively alleviate the pixel-wise coupling problem between all target-categories without introducing any extra data.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2209–2218, 2019.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Attention bridging network for knowledge transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5198–5207, 2019.

[6] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[7] 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.

[8] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022.