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 pipeline 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. Specifically, 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 × 256 and then cropped into 224 × 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 cropped as 512 × 512. When imposing our proposed CSD on MCTFormer and IRN, we set $\lambda_1 = 0.01$ and $\lambda_2 = 0.1$ 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.

© 2023. The copyright of this document resides with its authors. It may be distributed unchanged freely in print or electronic forms.
3 Algorithm Pipeline

Before the optimization pipeline of CSD, we first extract the original foreground mask $M_{fg}$ and background mask $M_{bg}$ 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 classes 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 $M_{fg}$ and corresponding background mask $M_{bg}$.

```plaintext
1: while not done do
2:   $\mathcal{X}^i_m, \mathcal{L}^i \leftarrow$ Load one multi-label sample;
3:   $\mathcal{X}^j_s, \mathcal{L}^j \leftarrow$ Resample one single-label image according co-occurrence;
4:   $\mathcal{X}_{fg}, \mathcal{X}_{bg} \leftarrow$ Crop foreground and background.
5:   $\mathcal{X}_{fg}^d \leftarrow$ Paste $\mathcal{X}_{fg}$ into $\mathcal{X}^i_m$;
6:   $\mathcal{X}_{bg}^d \leftarrow$ Paste $\mathcal{X}_{bg}$ into $\mathcal{X}^i_m$;
7:   $\mathcal{M}, \mathcal{M}_{fg}, \mathcal{M}_{bg} \leftarrow$ forward $\mathcal{X}_s, \mathcal{X}_{fg}, \mathcal{X}_{bg}$;
8:   $\mathcal{L}_{fg}, \mathcal{L}_{bg} \leftarrow$ KL($\mathcal{M}, \mathcal{M}_{fg}$), KL($\mathcal{M}, \mathcal{M}_{bg}$);
9:   $\mathcal{L}_{cls} \leftarrow$ CE($\mathcal{M}_m, \mathcal{L}^i$);
10:  Train Network $\leftarrow \mathcal{L}_{cls} + \mathcal{L}_{fg} + \mathcal{L}_{bg}$;
11:  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 [8] transfers the foreground prior from a simple single-label dataset to another complex multi-label dataset by adversarial learning [9]. 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


