International Journal of Computer Vision (2023) 131:701–716
https://doi.org/10.1007/s11263-022-01720-7

SegMix: Co-occurrence Driven Mixup for Semantic Segmentation and Adversarial Robustness

Md Amirul Islam[1] · **Matthew Kowal**[2] · **Konstantinos G. Derpanis**[3] · **Neil D. B. Bruce**[4]

Received: 20 August 2021 / Accepted: 9 November 2022 / Published online: 12 December 2022

Abstract

In this paper, we present a strategy for training convolutional neural networks to effectively resolve interference arising from competing hypotheses relating to inter-categorical information throughout the network. In this work, this is accomplished for the task of dense image labelling by blending images based on (i) categorical clustering or (ii) the co-occurrence likelihood of categories. We then train a *source separation* network which simultaneously segments and separates the blended images. Subsequent feature denoising to suppress noisy activations reveals additional desirable properties and high degrees of successful predictions. Through this process, we reveal a general mechanism, distinct from any prior methods, for boosting the performance of the base segmentation and salient object detection network while simultaneously increasing robustness to adversarial attacks.

Keywords Categorical Mixup · Source Separation · Semantic Segmentation · Adversarial Robustness

1 Introduction

The advent of Deep Neural Networks (DNNs) has seen overwhelming improvement in dense image labeling tasks (Long et al., 2015; Noh et al., 2015; Badrinarayanan et al., 2017; Ghiasi & Fowlkes, 2016; Zhao et al., 2017; Islam et al., 2017; Chen et al., 2018; Islam et al., 2018; He et al., 2017; Li et al., 2016), however, for some common benchmarks (Everingham et al., 2015) the rate of improvement has slowed down. While one might assume that barriers to further improvement require changes at the architectural level, it has also been borne out that pre-training across a variety of datasets Russakovsky et al. (2015); Lin et al. (2014) can improve performance that exceeds improvements seen from changing the model architecture. However, there are challenging scenarios for which DNNs have difficulty regardless of pre-training or architectural changes, such as highly occluded scenes, or objects appearing out of their normal context (Singh et al., 2020). It is not clear though, for dense image labeling tasks, how to resolve these specific scenarios for more robust prediction quality on a per-pixel level.

In particular, one might expect that failures in correctly predicting labels for an image are more likely to be seen for challenging cases once a critical performance threshold has been reached.

A question that naturally follows from this line of reasoning is: How can the number of locally challenging cases be increased, or the problem made more difficult in general? In this paper, we address this problem using a principled approach to improve performance and that also implies a more general form of robustness.

In our work, the means of solving this problem takes a direct form, which involves training networks on specially designed training data of mixed images (see Fig. 1 (top and middle)) to simultaneously address problems of dense image labeling (Long et al., 2015; Chen et al., 2015; Noh

Communicated by Oisin Mac Aodha.

✉ Md Amirul Islam
amirul@cs.ryerson.ca

Matthew Kowal
m2kowal@eecs.york.ca

Konstantinos G. Derpanis
kosta@eecs.york.ca

Neil D. B. Bruce
brucen@uoguelph.ca

[1] Ryerson University, Vector Institute for AI, Toronto, Canada

[2] York University, Vector Institute for AI, Toronto, Canada

[3] York University, Samsung AI Research Center Toronto, Vector Institute for AI, Toronto, Canada

[4] University of Guelph, Vector Institute for AI, Guelph, Canada

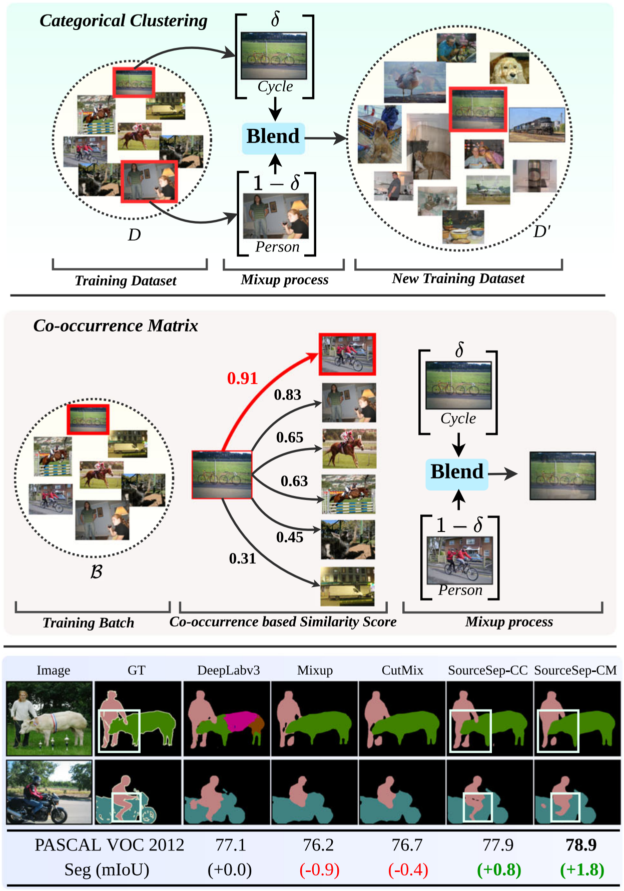

Image	GT	DeepLabv3	Mixup	CutMix	SourceSep-CC	SourceSep-CM
PASCAL VOC 2012 Seg (mIoU)		77.1 (+0.0)	76.2 (-0.9)	76.7 (-0.4)	77.9 (+0.8)	**78.9** (+1.8)

Fig. 1 Top: Overview of our category-specific image blending to create a new source dataset (D'). A segmentation network is trained with D' to simultaneously separate and segment both source and target images. Middle: Overview of our co-occurrence based image blending to create a mixed sample in the training batch, while training to simultaneously separate and segment both source and target images. Bottom: Results of our approach (SourceSep-CC and SourceSep-CM), Mixup (Zhang et al., 2018), and CutMix (Yun et al., 2019) on the PASCAL VOC 2012 (Everingham et al., 2015) segmentation task. Note that our methods substantially improve the overall performance

et al., 2015), and blind source separation (Georgiev et al., 2005; Huang et al., 2015). Humans show a surprising level of capability in interpreting a superposition (e.g., average) of two images, both interpreting the contents of each scene and determining the membership of local patterns within a given scene. The underlying premise of this work involves producing networks capable of simultaneously performing dense image labeling for pairs of images while also separating labels according to the source images. If one selects pairs on the basis of a weighted average, this allows treatment of the corresponding dense image labeling problem in the absence of source separation by extension. This process supports several objectives: (i) it substantially increases the number of occurrences that are locally ambiguous that need to be resolved to produce a correct categorical assignment, (ii) it forces broader spatial context to be considered in making categorical assignments, and (iii) it stands to create more powerful networks for standard dense labeling tasks and dealing with adversarial perturbations. The end goal of our procedure is to improve overall performance as well as increase the prediction quality on complex images (see Fig. 1 (bottom)), heavily occluded scenes, and also invoke robustness to challenging adversarial inputs.

The contribution of this paper extends from the approach presented in our prior work (Islam et al., 2020) which introduced a categorical clustering based mixup strategy to generate a new training dataset. In addition, we also proposed a source separation network (Islam et al., 2020) to simultaneously perform dense image labeling for pairs of images while also separating labels according to the source image classes. We extend our prior work in the following respects:

- We introduce a new and efficient *co-occurrence matrix* based *mixup* strategy which exploits co-occurrence likelihood of semantic categories from the dataset in the mixup process. This technique trains the network to separate semantic objects in commonly occurring complex scenes with high degrees of occlusion.
- We show, through extensive quantitative and qualitative experiments, that our newly introduced mixup technique outperforms our previous categorical clustering based technique (Islam et al., 2020) and recent mixing methods (Zhang et al., 2018; Yun et al., 2019) on the PASCAL VOC 2012 (Everingham et al., 2015) and MS-COCO (Lin et al., 2014) datasets, while simultaneously being less computationally expensive and maintaining robustness to adversarial attacks.
- We evaluate our newly introduced technique for an additional task, salient object detection, which shows improvements over the baselines.
- We provide an in-depth analysis and ablations of the introduced co-occurrence based mixup technique to show its influence in improving performance and robustness.

The paper is structured as follows: we discuss related work in Sect. 2. In Sect. 3, we first introduce two different image mixup techniques for the task of semantic segmentation followed by the source separation network for simultaneous dense labeling of two disparate images. Subsequently, we discuss the training procedure, and present the experimental results in Sect. 4. Finally, we provide extensive ablation studies in Sect. 5.

2 Related Work

Semantic Segmentation. Existing Convolutional Neural Network (CNN) based works (Long et al., 2015; Chen et al., 2015; Noh et al., 2015; Badrinarayanan et al., 2017; Ghiasi & Fowlkes, 2016; Chen et al., 2017; Takikawa et al., 2019) have shown widespread success on dense image prediction tasks (e.g., semantic segmentation). The feature representations produced in the top layers of shallower and deeper CNNs (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; Szegedy et al., 2015; He et al., 2016) carry a strong semantic representation, perhaps at the expense of retaining spatial details required for dense prediction due to the poor spatial resolution among deeper layers.

In particular, atrous convolution (Chen et al., 2017; Yu & Koltun, 2016; Chen et al., 2018), encoder-decoder structures (Badrinarayanan et al., 2017; Noh et al., 2015) and pyramid pooling (Zhao et al., 2017; Chen et al., 2018) have been employed to decode low resolution feature maps, increase the contextual view, and capture context at different ranges of spatial precision, respectively.

Data Augmentation. Existing methods (Bishop, 1995; Krizhevsky et al., 2012; Hendrycks et al., 2019; Kim et al., 2020) introduced data augmentation based techniques to regularize the training of CNNs. These techniques regularize the models from over-fitting to the training distribution (e.g., categorical biases) and also improve the generalization ability by generating extra training samples given the original training set. Most commonly used data augmentation strategies are random cropping, horizontal flipping (Krizhevsky et al., 2012), and adding random noise (Bishop, 1995). Recently proposed data augmentation techniques, termed AugMix (Hendrycks et al., 2019) and PuzzleMix (Kim et al., 2020), were designed to improve the generalization performance and robustness against corruptions. However, these techniques are extensively evaluated for image classification and its unclear if these techniques will perform better for dense labeling tasks. In contrast, our proposed approach can be complementary to these techniques and could be applied in conjunction to further improve the dense labeling performance and robustness.

Mixup-based Augmentation. More closely related to our work, contributions (Yun et al., 2019; Zhang et al., 2018; Tokozume et al., 2018; Inoue, 2018; Cubuk et al., 2019; French et al., 2020; Harris et al., 2020; Chou et al., 2020) on data augmentation based techniques share a similar idea of mixing two randomly selected samples to create new training data for the image classification or localization task. Between-Class (BC) learning (Tokozume et al., 2018) showed that randomly mixing training samples can lead to better separation between categories based on the feature distribution. Mixup (Zhang et al., 2018) shares a similar idea of training a network by mixing the data that regularizes the network and increases the robustness against adversarial examples. Manifold Mixup (Verma et al., 2019) extends Mixup (Zhang et al., 2018) from input space to feature space and showed improvement on overall performance. Further, Guo et al. (2019) proposed an adaptive Mixup technique to prevent the generation of improper mixed data. CutMix (Yun et al., 2019) further proposed to overlay a cropped area of an input image to another. However, these methods randomly blend images and may generate non-optimal training samples according to object distributions, which might be problematic for more complex dense labeling tasks. Our proposed techniques aim to address this issue by utilizing category-level information in the mixup process. Our proposed framework differs from the above existing works in that: (i) the network performs simultaneous dense prediction and source separation to achieve superior dense labeling and adversarial robustness; whereas, other techniques are focused mainly on image classification or object localization while using a single output, (ii) previous methods either mix labels as the ground truth or use the label from only one sample, while we use both ground truth labels independently, and (iii) samples are chosen randomly for Mixup (Zhang et al., 2018) and CutMix (Yun et al., 2019) while we use two intuitive strategies (categorical clustering, Sect. 3.1.1 and Co-occurrence matrix, Sect. 3.1.2).

The groundwork for some of what is presented in this paper appeared previously (Islam et al., 2020), in which we introduced a categorical clustering based mixup strategy to generate a new training dataset followed by training a network for source separation with the ultimate goal of semantic segmentation. However, due to the size of the new training dataset, the computational load during training was significant. In this work, we address this training inefficiency issue by introducing an intuitive mixup technique which considers the co-occurrence likelihood of semantic categories before mixing two images. The main advantage of this new technique is the training sample generation is done online within a batch instead of creating a new large dataset offline. We provide an in-depth analysis of the newly introduced mixup technique to show its influence in improving performance and robustness.

3 Proposed Method

We propose a novel framework capable of solving the dense labeling problem. Our proposed framework consists of three key steps: (i) we first apply a blending technique (Sect. 3.1) on the training dataset either offline (Sect. 3.1.1) or online (Sect. 3.1.2), (ii) we train a CNN using the generated data that simultaneously produces dense predictions and source separations (Sect. 3.2), and (iii) we denoise the learned features from the source separation network by fine-tuning on standard data (Sect. 3.3).

3.1 Category-Dependent Image Blending

Recent works (Zhang et al., 2018; Yun et al., 2019; Tokozume et al., 2018; Inoue, 2018; Cubuk et al., 2019) simply mix two randomly selected samples to create new training data for image classification or object localization. Exploring a similar direction, we are interested in solving dense prediction tasks (e.g., semantic segmentation, salient object detection) in a way that provides separation based on mixed images. The traditional way (Zhang et al., 2018; Tokozume et al., 2018; Inoue, 2018) of combining two images is by a weighted average which implies that the contents of both scenes appear with varying contrast. Randomly combining two source images to achieve the desired objective is a more significant challenge than one might expect in the context of dense prediction. One challenge is the categorical bias of the dataset (e.g., mostly the *person* images will be combined with all other categories, since *person* is the most common category in PASCAL VOC 2012) across the newly generated training set. Previous methods (Zhang et al., 2018; Tokozume et al., 2018; Inoue, 2018), randomly select images to combine, results in a new data distribution with similar inherent biases as the original dataset.

Algorithm 1 Compute Co-occurrence Matrix

```
Input: Training Dataset $\mathcal{D}$, Category List $\mathcal{C}$, Total Category, $N$
Output: Co-occurrence Matrix, $\mathcal{C}_{ss}$
$\mathcal{C}_{ss} \leftarrow$ zeros($N$, $N$),
for $\mathcal{C}_i \in \mathcal{C}$ do
    for $\mathcal{I}_k \in \mathcal{D}$ do
        /* Compute the unique semantic categories */
        unique-cat ← unique(Ground-truth($\mathcal{I}_k$))
        if $\mathcal{C}_i \in$ unique-cat then
            Remove $\mathcal{C}_i$ from unique-cat
            for $m \in$ unique-cat do
                $\mathcal{C}_{ss}[\mathcal{C}_i, m] \leftarrow \mathcal{C}_{ss}[\mathcal{C}_i, m] + 1$
            end for
        end if
    end for
end for
```

To overcome these limitations, we introduce two different image blending techniques to create new training data, denoted as *categorical clustering* and *co-occurrence matrix*. In categorical clustering technique, we augment the training dataset to generate a new training set in a form that accounts for source separation and dense prediction. Categorical clustering combines images based on a uniform distribution across categories. For the co-occurrence matrix-based strategy, we consider the co-occurrence likelihood between semantic objects in the blending process to generate new training data. The main difference between these two blending techniques is the way that new training data is generated. The former one generates a new training dataset offline while the latter one blends images in the training batch. Thorough experimentation with our proposed mixing strategies show improvements in the network's ability to separate competing categorical features and can generalize these improvements to various challenging scenarios, such as segmenting out-of-context objects or highly occluded scenes. We believe that the issue of object-bias within a dataset will not be completely alleviated by our clustering based approach. However, our clustering based method combines objects which rarely co-exist together. This means that each combination of objects has a much larger number of examples in the training set than previously. So while there may be biases, the network will at least see numerous examples of every object-object combination.

3.1.1 Categorical Clustering

We first generate N (total number of categories in the dataset) different clusters of images, where each cluster contains images of a certain category from the training dataset. For each training sample in a cluster, we linearly combine it with a random sample from each of the $N - 1$ other clusters. For example, given a training sample $\mathcal{I}_a$ (dominant) from the *person* cluster, we randomly choose a sample $\mathcal{I}_b$ (phantom) from another categorical cluster and combine them to obtain a new mixed sample, $\mathcal{I}_{ss}$:

$$\mathcal{I}_{ss} = \delta * \mathcal{I}_a + (1 - \delta) * \mathcal{I}_b, \tag{1}$$

where δ denotes the randomly chosen weight that is applied to each image. We assign the weight such that the dominant image ($\mathcal{I}_a$) has more weight compared to the phantom one ($\mathcal{I}_b$). In our experiments, we sample δ uniformly from a range of $[0.7 - 1]$ for each image pair. Note that for one sample (e.g., the *person* cluster), we generate $N - 1$ new samples. We continue to generate new training samples for the other remaining images in the person cluster and perform the same operation for images in other clusters.

3.1.2 Co-occurrence Matrix

We propose an additional technique that blends images based on the co-occurrences among semantic categories (i.e., probability of appearing together in an image). Towards this goal, we first calculate the co-occurrence matrix, $\mathcal{C}_{ss} \in \mathbb{R}^{N \times N}$ (N= number of categories) using Algorithm 1 that contains the number of times two semantic categories co-occur within the training set.

Algorithm 2 describes the set of steps for generating new training samples in a batch based on the pre-computed co-occurrence matrix.

Algorithm 2 Co-occurrence based Image Blending

```
Input: Training Batch $\mathcal{B} = \{I, G\}$; Co-occurrence matrix $\mathcal{C}_{ss}$, $\alpha$, max unique Category threshold, $\gamma$
Output: New training batch $\mathcal{B}'$

if $\alpha > 0$ then
    $\delta \leftarrow$ random.beta($\alpha, \alpha$)                ▷ Generate $\delta$ from beta distribution if $\alpha > 0$
else
    $\delta \leftarrow 1$
end if
$\mathcal{B}' \leftarrow \{\}$,   $\mathcal{Y}_1' \leftarrow \{\}$,   $\mathcal{Y}_2' \leftarrow \{\}$
for $\mathcal{I}^k \in \mathcal{B}$ do
    similarity-score ← zeros(len($\mathcal{B}$))       ▷ Store similarity score with each sample in the batch other than $\mathcal{I}^k$
    mixed-category-list ← zeros(len($\mathcal{B}$))    ▷ Store total number of unique semantic categories
    for $\mathcal{I}^m \in \mathcal{B}$ do
        if $k \neq m$ then
            /* Compute the unique semantic categories */
            K-unique ← unique(Ground-truth($\mathcal{I}^k$))
            M-unique ← unique(Ground-truth($\mathcal{I}^m$))
            Remove the background class index and the ignore class index from both K-unique and M-unique
            if len(K-unique) $\geq 1$ & len(M-unique) $\geq 1$ then
                /* Initialize co-occurrence score to 0 */
                cooccurrence-score ← 0
                /* Compute total co-occurrence score */
                for i ← 1 to len(M-unique) do
                    for j ← 1 to len(K-unique) do
                        cooccurrence-score ← cooccurrence-score + $\mathcal{C}_{ss}$[M-unique[i]][K-unique[j]]
                    end for
                end for
                similarity-score [m] ← cooccurrence-score
                mixed-category-list [m] ← len(M-inique) + len(K-unique)
            end if
        end if
    end for
    top-sim-idx ← random(0, len($\mathcal{B}$))
    top-sim-idx ← argmax(similarity-score)             ▷ Choose the index with highest similarity score
    /* Restrict the mixing ratio if total number of unique semantic categories > $\gamma$ in the chosen pair */
    if mixed-category-list[top-sim-idx] $> \gamma$ then
        $\delta \leftarrow 0.9$
    end if
    $\mathcal{I}_{ss}^k \leftarrow \delta * \mathcal{I}^k + (1 - \delta) * \mathcal{I}^{top-sim-idx}$     ▷ Mix $\mathcal{I}^k$ with sample with highest similarity score
    $\mathcal{Y}_1' \leftarrow \mathcal{Y}[k]$,   $\mathcal{Y}_2' \leftarrow \mathcal{Y}[top - sim - idx]$     ▷ Choose the corresponding ground-truth segmentation map
    $\mathcal{B}' \leftarrow \{\mathcal{I}_{ss}^k, \mathcal{Y}_1', \mathcal{Y}_2'\}$     ▷ Mixed training sample in Batch, $\mathcal{B}'$
end for
```

In summary: for each training sample, I_a in a batch, B, of size n, we compute a scalar similarity score with the other $n - 1$ samples based on the pre-computed co-occurrence matrix. Note that we use similarity score to represent the similarity (i.e., how likely they will co-occur based on the data distribution) between two images based on the co-occurrence score. We pick the sample with highest similarity score, I_b, to be combined with the sample I_a. Finally, we apply Eq. 1 to generate a new blended training sample. Similar to the clustering based blending technique, we randomly choose δ and assign more weight on the dominant image, $\mathcal{I}_a$, compared to the phantom image, $\mathcal{I}_b$. The intuition of assigning higher weight on the dominant image is that the semantic segmentation task requires to learn the context of the semantic objects for accurate per-pixel labeling. However, blending two images with a large number of semantic categories with lower mixed ratio (i.e., assigning more weight on the *phantom* image) substantially increases the possibility of destroying the contextual information as well as introducing unlikely samples into the training set when considering the object distribution. For example, PASCAL VOC has very few images with 7+ objects in it, and therefore training on blended images with this many objects may add noise during training when more weight is assigned to the phantom image. Therefore, we choose a threshold for the maximum number of unique semantic categories, γ. If the total number of unique semantic categories in the chosen pair is greater than a certain threshold, we set the mixing ratio, δ to 0.9.

While there exist alternatives (Zhang et al., 2018; DeVries & Taylor, 2017; Yun et al., 2019) for combining pairs of images to generate a training set suitable for source separation training, our intuitive methods are simple to implement and achieve strong performance on a variety of metrics (see Sect. 4). Exploring further methods to combine and augment the training set is an interesting and nuanced problem to be studied further in the context of dense image labeling.

3.2 Source Separation Network

In this section, we present a fully convolutional source separation network in the context of dense prediction. Figure 2 illustrates the overall pipeline of our proposed method.

Overview and Notations. During training, our goal is to produce dense predictions of dominant, $\mathcal{I}_a$, and phantom, $\mathcal{I}_b$, images, given a blended image, $\mathcal{I}_{ss}$. Note that each blended image, $\mathcal{I}_{ss}$, in the new set is a weighted combination of dominant and phantom images $(\mathcal{I}_a, \mathcal{I}_b)$. We denote the dominant predictor as $\mathcal{F}_t(.)$ and phantom predictor as $\mathcal{F}_p(.)$.

3.2.1 Network Architecture

Figure 2 (left) reveals two key components of the source separation network including a *fully convolutional network* encoder and *source separator module* (SSM). Given a mixed image, $\mathcal{I}_{ss} \in \mathbb{R}^{h \times w \times c}$, we adopt DeepLabv3 (Chen et al., 2017) (f_{enc}) to produce a sequence of bottom-up feature maps. The SSM consists of two separate branches: (i) *dominant*, $\mathcal{F}_t(.)$, and (ii) *phantom*, $\mathcal{F}_p(.)$. Each branch takes the spatial feature map, $\hat{f}_b^i$, produced at the last block, `res5c`, of f_{enc} as input and produces a dense prediction for the dominant, $\mathcal{S}_t$, and the phantom, $\mathcal{S}_p$, image. Next, we append a *source separation head* (SSH) to generate a final dense prediction of categories for the dominant image. The SSH, $\mathcal{F}_{ss}$, simply concatenates the outputs of dominant and phantom branches followed by two 1 × 1 convolution layers with nonlinearities (ReLU) to obtain the final dense prediction map, $\mathcal{S}_{ss}$. The intuition behind the SSH is that the phantom branch may produce activations that are correlated with the dominant image, and thus the SSH allows the network to further correct any incorrectly separated features with an additional signal to learn from. Given a mixed image, $\mathcal{I}_{ss}$, the operations can be expressed as:

$$\hat{f}_b^i = f_{enc}(\mathcal{I}_{ss}), \quad \underbrace{\mathcal{S}_t = \mathcal{F}_t(\hat{f}_b^i)}_{\texttt{dominant}}, \quad \underbrace{\mathcal{S}_p = \mathcal{F}_p(\hat{f}_b^i)}_{\texttt{phantom}}, \tag{2}$$

$$\underbrace{\mathcal{S}_{ss} = \mathcal{F}_{ss}(\mathcal{S}_t, \mathcal{S}_p)}_{\texttt{source separation}} . \tag{3}$$

3.2.2 Training the Source Separation Network

The source separation network produces two dominant predictions, $\mathcal{S}_{ss}$ and $\mathcal{S}_t$, including a phantom prediction, $\mathcal{S}_p$; however, we are principally interested in the final dominant prediction, $\mathcal{S}_{ss}$. More formally, let $\mathcal{I}_{ss} \in \mathbb{R}^{h \times w \times 3}$ be a training image associated with ground-truth maps $(\mathcal{G}_a, \mathcal{G}_b)$ in the source separation setting. To apply supervision on $\mathcal{S}_{ss}$, $\mathcal{S}_t$, and $\mathcal{S}_p$, we upsample them to the size of $\mathcal{G}_a$. Then we define three pixel-wise cross-entropy losses, ℓ_{ss}, ℓ_t, and ℓ_p, to measure the difference between $(\mathcal{S}_{ss}, \mathcal{G}_a)$, $(\mathcal{S}_t, \mathcal{G}_a)$, and $(\mathcal{S}_p, \mathcal{G}_b)$, respectively. The objective function can be formalized as:

$$L_{stage1} = \ell_{ss} + \delta * \ell_t + (1 - \delta) * \ell_p, \tag{4}$$

where δ is the weight used to linearly combine images to generate $\mathcal{I}_{ss}$. Note that the network is penalized the most on the final and initial dominant predictions, and places less emphasis on the phantom prediction.

3.3 Feature Denoising Stage

While feature binding and source separation are interesting, the ultimate goal is to see improvement and robustness for standard images. For this reason, we mainly care about improving the overall dense prediction. To accomplish this, we further fine-tune our trained source separation model on the standard training set which we call the feature denoising stage. In this stage, as we feed a standard image to the network, the phantom predictor branch, $\mathcal{F}_{ph}$, has no supervisory signal, instead it acts as a regularizer. We propose the following technique to penalize the phantom prediction.

Penalize Phantom Activation. Along with ℓ_t, we propose a loss, ℓ_{PPA}, on the phantom prediction to penalize any activation (and suppress phantom signals and interference). The goal here is to push the output of the phantom branch to zero and suppress the phantom. The ℓ_{PPA} loss sums the absolute value of the confidence attached to categories and applies a log operation to balance the numeric scale with ℓ_t:

$$\ell_{\text{PPA}} = \log \sum_{\forall i \in h} \sum_{\forall j \in w} \sum_{\forall k \in c} \sigma(\mathcal{S}_p), \quad L_{stage2} = \ell_t + \ell_{\text{PPA}}, \tag{5}$$

where $\sigma(\cdot)$ is the ReLU function, which constrains the input to the log to be a positive value. In **Stage 1**, f_{enc}, $\mathcal{F}_t$, $\mathcal{F}_p$, and $\mathcal{F}_{ss}$ are trained in an end-to-end manner. Then, in **Stage 2**, f_{enc}, $\mathcal{F}_t$, and $\mathcal{F}_p$ are fine-tuned from the Stage 1 weights.

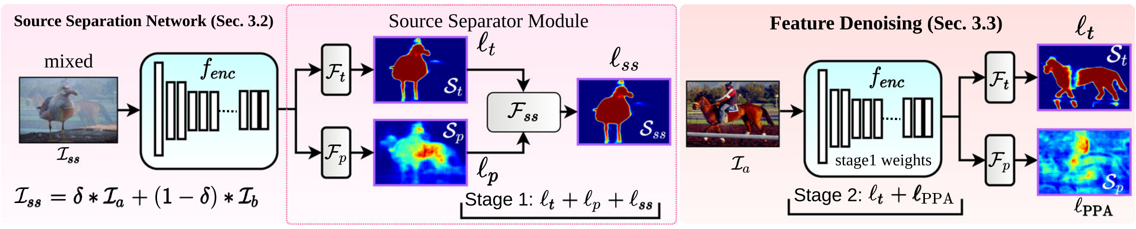

Fig. 2 An illustration of our proposed framework. At the data end, categorical collisions are created with a *dominant* ($\mathcal{I}_a$) and *phantom* ($\mathcal{I}_b$) image. **Stage 1:** The network is trained on mixed data ($\mathcal{I}_{ss}$) to perform simultaneous dense labeling and source separation. We use the labels of both *dominant* and *phantom* images as the targets for two separate output channels. **Stage 2:** Fine-tuning on standard data to further promote desirable properties along the two dimensions of base performance and robustness to perturbations. In this stage, the *phantom* activation of the second channel is suppressed. Confidence maps are plotted with the 'Jet' colormap, where red and blue indicates higher and lower confidence, respectively

4 Experiments

We first present results on the PASCAL VOC 2012 (Everingham et al., 2015) and MS-COCO (Lin et al., 2014) semantic segmentation datasets (Sect. 4.3). Unless otherwise stated, we use the DeepLabv3 (Chen et al., 2017) network without any bells and whistles (e.g., multi-scale processing, conditional random field) as our baseline model. We then show qualitative and quantitative evidence that our proposed mixing techniques improve the network's ability to segment highly occluded objects in complex scenes (Sect. 4.3.3), as well as objects found in out-of-context scenarios (Sect. 4.3.4). Throughout the experiments, we compare our methods to recent mixing strategies, CutMix (Yun et al., 2019) and Mixup (Zhang et al., 2018). Mixup and CutMix did not explicitly design their strategies for dense labeling; however, in CutMix, the authors used CutMix and MixUp for image localization and object detection tasks, so we view their strategies as a general data augmentation technique. Next, we evaluate the robustness of our methods to a variety of adversarial attacks (Sect. 4.4). We further apply our co-occurrence based image blending strategy for salient object detection task and compare the results with existing techniques (Sect. 4.5). Finally, we conduct an extensive ablation study (Sect. 5) to better tease out the underlying mechanisms giving performance boosts by evaluating the various image blending strategies and network architectures.

4.1 Implementation Details

We implement our proposed source separation networks using PyTorch (Paszke et al., 2017). We apply bilinear interpolation to upsample the predicted segmentation map before the losses are calculated. The *source separation* networks are trained using stochastic gradient descent for 50 epochs with momentum of 0.9, weight decay of 0.0005 and the "poly" learning rate policy (Chen et al., 2018) which starts at $2.5e^{-4}$. We use the same strategy during the feature denoising stage of training, but with an initial learning rate of $2.5e^{-5}$. During training, we apply random and center cropping to form 513×513 input images during training and inference, respectively. For a fair comparison, we implement and train Mixup (Zhang et al., 2018) and CutMix (Yun et al., 2019) using the same set of hyper-parameters. We report numbers for the following variants that are described in what follows: **DeepLabv3 + SourceSep-CC:** This network applies the categorical clustering based image blending with the DeepLabv3 based source separation network. **DeepLabv3 + SourceSep-CM:** This network uses the co-occurrence matrix based image blending with the DeepLabv3 based source separation network. **DeepLabv3 + Mixup:** This network uses the Mixup (Guo et al., 2019) technique with the DeepLabv3 network. **DeepLabv3 + CutMix:** This network applies the CutMix (Yun et al., 2019) technique with the DeepLabv3 network for the task of semantic segmentation.

4.2 Dataset and Evaluation Metrics

PASCAL VOC 2012: The PASCAL VOC 2012 dataset is considered the most popular semantic segmentation dataset, and includes 20 object categories and a background class. It consists of 1464 training images, 1449 validation images, and 1456 testing images. Following the current common practice (Chen et al., 2017; Long et al., 2015; Lin et al., 2017; Chen et al., 2018), we augment the training set using extra labeled PASCAL VOC images from Hariharan et al. (2011). We use the standard mean IoU metric to report semantic segmentation performance.

MS-COCO: MS-COCO 2014 (Lin et al., 2014) dataset is a large-scale challenging semantic segmentation dataset which includes 80 object categories and a background class. Following previous works (He et al., 2017; Ren et al., 2015), we train our model using the union of 80k train images and a 35k subset of val images (trainval35k), and report results on the remaining 5k validation set.

Table 1 Quantitative comparisons on PASCAL VOC 2012 val set

Backbone	Method	mIoU (%)
Res50	DeepLabv3-ResNet50 (Chen et al., 2017)	75.1
	DeepLabv3 + Mixup (Zhang et al., 2018)	73.6
	DeepLabv3 + CutMix (Yun et al., 2019)	75.1
	DeepLabv3 + SourceSep-CC	75.7
	DeepLabv3 + SourceSep-CM	**76.2**
Res101	DeepLabv3-ResNet101 (Chen et al., 2017)	77.1
	DeepLabv3 + Mixup (Zhang et al., 2018)	76.2
	DeepLabv3 + CutMix (Yun et al., 2019)	76.7
	DeepLabv3 + SourceSep-CC	77.9
	DeepLabv3 + SourceSep-CM	**78.9**

Our co-occurrence based source separation network outperforms the other mixing based techniques

4.3 Semantic Segmentation

4.3.1 Results on PASCAL VOC 2012

First, we show the improvements on segmentation accuracy by our methods on the PASCAL VOC 2012 validation dataset. We present a comparison of different baselines and our proposed approaches in Table 1.

As shown in Table 1, our image blending based source separation approaches improve the overall mIoU more than other approaches (Zhang et al., 2018; Yun et al., 2019). Additionally, the co-occurrence based blending technique (DeepLabv3+SourceSep-CM) marginally outperforms the categorical clustering based strategy (0.8% vs 1.8% improvement over the baseline DeepLabv3-ResNet101 method).

We further evaluate our approaches on the PASCAL VOC 2012 test set. Following prior works (Chen et al., 2018; Zhao et al., 2017; Noh et al., 2015), before evaluating our method on the test set, we first train on the augmented training set followed by fine-tuning on the original trainval set. As shown in Table 2, DeepLabv3 with categorical clustering based source separation network achieves 80.5% mIoU which outperforms the baseline. Additionally, co-occurrence based source separation network achieves 81.1% mIoU which marginally outperforms the baselines and the categorical clustering based source separation network.

Table 2 Quantitative comparisons of various mixing techniques on PASCAL VOC 2012 test set

Method	mIoU (%)
DeepLabv3-ResNet101 (Chen et al., 2017)	79.3
DeepLabv3 + Mixup (Zhang et al., 2018)	78.9
DeepLabv3 + CutMix (Yun et al., 2019)	80.2
DeepLabv3 + SourceSep-CC	80.5
DeepLabv3 + SourceSep-CM	**81.1**

Table 3 Quantitative comparisons of various mixing techniques on MS-COCO (Lin et al., 2014) dataset

Method	mIoU (%)
DeepLabv3-ResNet101 (Chen et al., 2017)	51.1
DeepLabv3 + Mixup (Zhang et al., 2018)	47.0
DeepLabv3 + CutMix (Yun et al., 2019)	49.5
DeepLabv3 + SourceSep-CM	**53.2**

Sample predictions of our methods and the baselines are shown in Fig. 3. As shown in Fig. 3, our proposed blending based source separation networks are very effective in capturing more distinct features for labeling occluded objects and plays a critical role in separating different semantic objects more accurately. Note the ability of our methods to segment scenes with a high degree of occlusion (see second last row in Fig. 3), thin overlapping regions (see top row), or complex interaction between object categories (see 6^{th} row). While other methods identify the dominant categories correctly, they often fail to relate the activations of smaller occluding features to the correct categorical assignments.

4.3.2 Results on MS-COCO

Next, we show the improvements on segmentation accuracy on the MS-COCO dataset. We present a comparison of different baselines and our proposed approaches in Table 3.

As shown in Table 3, our co-occurrence based image blending approach (DeepLabv3+SourceSep-CM) outperforms the other mixup based approaches (Zhang et al., 2018; Yun et al., 2019) by a reasonable margin in terms of mIoU. Additionally, our co-occurrence based blending technique also outperforms the DeepLabv3-ResNet101 baseline method (2.1% improvement). The improvements on this large-scale challenging dataset further demonstrate the superiority of our approach.

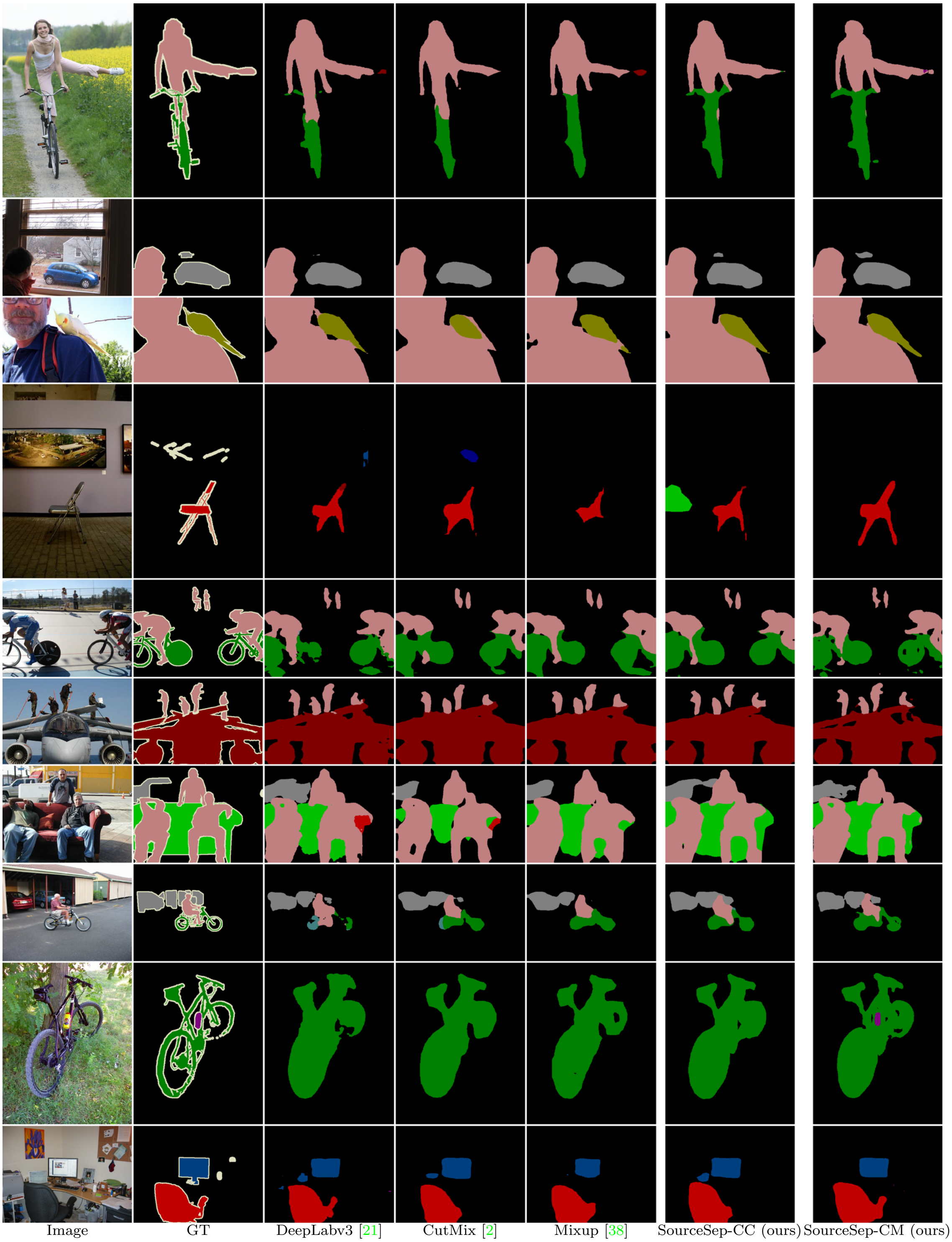

Fig. 3 Qualitative results on the PASCAL VOC 2012 val set

4.3.3 Segmenting Highly Occluded Objects in Complex Scenes

We argue that our mixing and source separation strategies are more powerful than existing mixing strategies in complex scenes with large amounts of occlusion. One reason for this is our categorical clustering based mixing strategy (Sect. 3.1) blends images based on categorical clusters with dynamic blending ratios. This means that the network will see more images with a wide array of categories blended together, as every category is guaranteed to be blended with every other category. Additionally, the co-occurrence based mixing strategy blends the images containing semantic objects which are likely to co-occur frequently (e.g., *person* and *motorcycle*). This strategy allows the network to learn stronger representations for objects in commonly occurring complex scenes. On the other hand, other strategies (Yun et al., 2019; Zhang et al., 2018) use two randomly selected images to blend. This means the statistics of the generated images will be largely driven by the statistics of the original dataset. Further, the *source separation module* (SSM) is specifically designed for separating features *before* the final layer of the network, allowing for finer details and semantics to be encoded into the target and *phantom* streams. For the other methods, they have a single prediction, which does not allow for these details to be separated early enough in the network to encode as much information as our method.

To substantiate this claim we evaluate each method under three specific data distributions that range in the amount of occlusion and complexity: (i) *Occlusion*: at least one object has occlusion with any other object (1-Occ) in an image and all objects have occlusion (All-Occ), (ii) *Number of Objects*: total number of object instances regardless of classes, and (iii) *Number of Unique Objects*: total number of unique semantic categories. The results are presented in Table 4. Our methods outperform the other mixing based methods in all cases. Interestingly, co-occurrence based blending techniques outperform the clustering based blending under most of the data settings. Note that the improvements on all occlusion and larger number of unique categories cases are particularly pronounced for our source separation models as the performance drop is substantially less than the other methods, when only considering images with many unique objects.

We next perform a cross-dataset experiment by taking our model trained on the PASCAL VOC 2012 training set and evaluate on the publicly available Out-of-Context (Choi et al.,

Table 4 Results on complex scenes in terms of mIoU, evaluated using various subsets from the PASCAL VOC 2012 val set

	Occlusion		Number of Objects				Number of Unique Objects		
	1-Occ	**All-Occ**	**1-Obj**	**2-Obj**	**3-Obj**	**4-Obj**	**2-Obj**	**3-Obj**	**4-Obj**
# of Images	1128	538	695	318	167	99	375	121	23
DeepLabv3	75.5	74.9	74.6	74.8	76.0	70.0	72.5	63.5	62.1
DeepLabv3 + Mixup	75.4	72.3	77.9	74.3	71.7	68.3	72.0	58.1	59.2
DeepLabv3 + CutMix	76.4	74.3	78.3	75.4	73.0	70.0	72.3	60.1	59.6
DeepLabv3 + SourceSep-CC	77.9	**76.1**	**80.7**	**77.2**	75.6	70.0	74.0	61.5	62.0
DeepLabv3 + SourceSep-CM	**78.7**	**76.1**	**80.7**	**77.2**	**77.9**	**72.1**	**75.4**	**65.6**	**63.6**

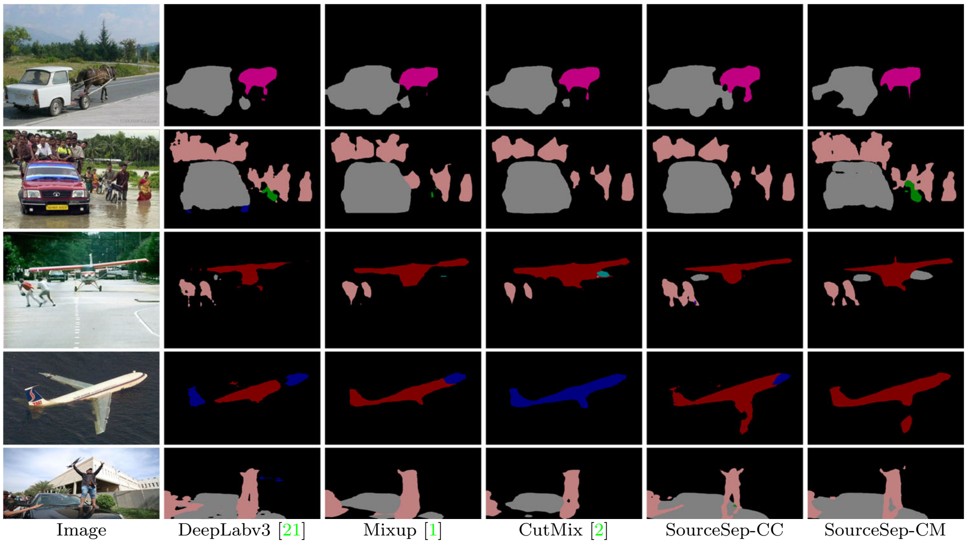

Fig. 4 Qualitative examples on the Out-of-Context (Choi et al., 2012) (top five rows) and UnRel (Peyre et al., 2017) (bottom three rows) datasets. Our proposed blending based source separation networks (SourceSep-CC and SourceSep-CM) generate higher quality segmentation maps compared to the baselines in the out-of-context scenarios

Table 5 mIoU results on the PASCAL VOC 2012 val set, for the co-occurrence of the most salient person category with five other categories and the results when these five categories appear alone

	Co-occur with person					*Exclusive*				
	Horse	**Mbike**	**Bicycle**	**Bottle**	**Car**	**Horse**	**Mbike**	**Bicycle**	**Bottle**	**Car**
# of Images	32	34	30	20	45	44	23	29	35	45
DeepLabv3-ResNet101 (Chen et al., 2017)	87.9	81.6	77.7	**89.7**	**89.7**	90.9	91.5	60.4	85.4	96.0
DeepLabv3 + Mixup (Zhang et al., 2018)	86.9	82.8	76.5	87.6	86.2	92.5	93.0	60.0	80.6	95.5
DeepLabv3 + CutMix (Yun et al., 2019)	86.2	83.6	76.0	87.4	87.9	**94.1**	93.8	61.3	82.6	96.2
DeepLabv3 + SourceSep-CC	89.1	**87.2**	78.5	86.9	89.0	94.0	93.8	**61.5**	87.9	96.4
DeepLabv3 + SourceSep-CM	**89.9**	86.7	**79.3**	88.5	89.2	93.6	**95.1**	60.2	**88.2**	**96.6**

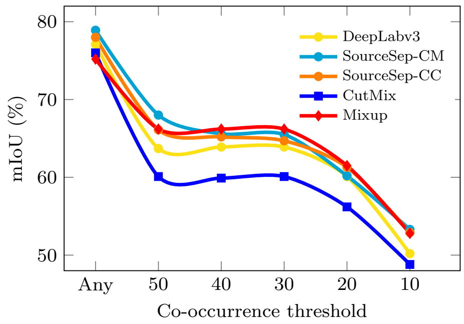

Fig. 5 Performance on images with various levels of object co-occurrence. *SourceSep-CC*, *SourceSep-CM*, and Mixup (Zhang et al., 2018) perform better on subsets of images with unlikely co-occurrences

2012) and UnRel (Peyre et al., 2017) datasets. Figure 4 visualizes how the segmentation models trained with only VOC 2012 co-occurring objects performs when objects appear without the context seen in training. Even with such challenging images with out of context objects (person *on top* of car (see second row)), our methods produce robust segmentation masks while the other mixing based methods fail to segment the objects with detail. The co-occurrence based source separation network also produces reasonable segmentation maps despite the nature of training where we blend images with semantic objects which are likely to co-occur. As the results shown in Fig. 4 are randomly sampled, there include examples of SourceSep-CM failing. Note that SourceSep-CC is trained on blended images with *unlikely* categorical combinations but SourceSep-CM is trained on images with *likely* categorical combinations. We believe this explains why SourceSep-CC may outperform SourceSep-CM on out-of-context data. We agree that DeepLabv3 performs surprisingly well on the visualized out-of-context data but since no segmentation labels exist for this dataset we are unable to back up our hypothesis quantitatively.

4.3.4 Segmenting Out-of-Context Objects

A model that heavily relies on context would not be able to correctly segment compared to a model that truly understands what the object is irrespective of its context. We argue that our clustering based mixing strategy performs better in out-of-context scenarios, as category-based mixing reduces bias in the dataset's co-occurrence matrix. In contrast, the co-occurrence based blending technique allows the source separation network to separate semantic objects which are likely to co-occur. We conduct two experiments to quantitatively evaluate each method's ability to segment out-of-context objects.

For the first experiment, we identify the top five categories that frequently co-occur with *person* in the training set, since person has the most occurrences with all other categories based on the co-occurrence matrix. We report performance in Table 5 on two different subsets of data: (i) *Co-occur with Person*: images with both the person and object in it, and (ii) *Exclusive*: images with only the single object of interest. As can be seen from Table 5, when bottle co-occurs with person all the methods are capable of segmenting bottle and person precisely, whereas the IoU for bottle is substantially reduced when bottle occurs alone. However, our proposed methods (especially the SourceSep-CC) successfully maintain performance on the exclusive case.

For the second out-of-context experiment, we first create different subsets of images from the VOC 2012 val set based on the training set's co-occurrence matrix. We select thresholds {50, 40, 30, 20, 10}, and only keep images which have objects that occur less than the chosen threshold. For instance, the threshold value 50 includes all the images where the co-occurrence value of object pairs is less than 50 (e.g., *cat* and *bottle* occur 18 times together, therefore images containing both will be in all subsets except the threshold of 10). Figure 5 illustrates the result of different baselines and our methods with respect to co-occurrence threshold. Our methods outperform the baseline DeepLabv3-ResNet101 for all the threshold values. Surprisingly, Mixup (Zhang et al., 2018) achieves very competitive performance under

Table 6 Adversarial segmentation robustness performance (mIoU) against the UAP (Moosavi-Dezfooli et al., 2017) and GD-UAP (Mopuri et al., 2018) attacks

Networks	Clean	**Adversarial Images**				
		UAP (Moosavi-Dezfooli et al., 2017)		GD-UAP (Mopuri et al., 2018)		
		ResNet	GNet	R-No	R-All	R-Part
DeepLabv3	75.9	59.1	63.6	66.7	63.9	64.0
+ Mixup	75.2	62.9	63.2	65.3	63.2	63.6
+ CutMix	76.2	60.9	64.3	64.2	62.5	62.2
+ SourceSep-CC	77.9	**69.1**	**70.2**	**68.2**	**67.0**	**67.2**
+ SourceSep-CM	**78.9**	63.2	67.1	67.2	65.0	64.9

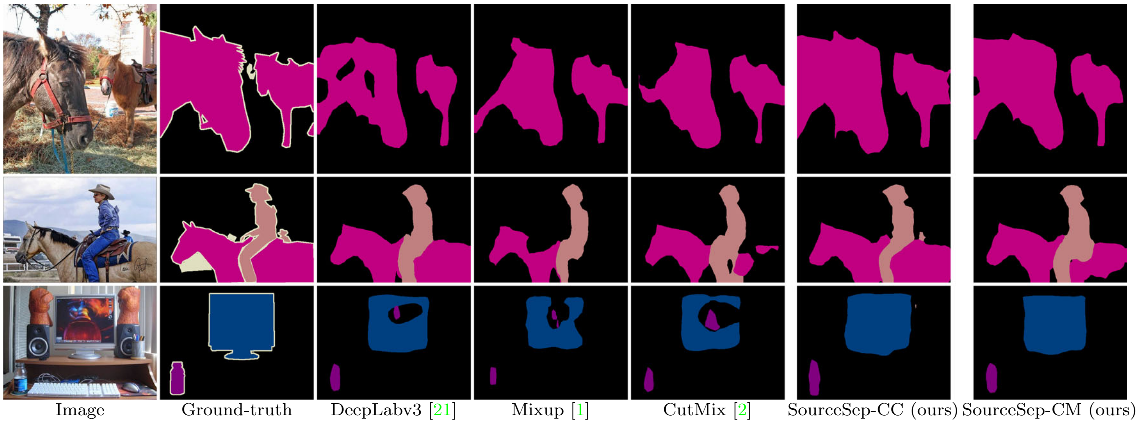

Fig. 6 Comparison of baselines and our methods when attacked by GD-UAP (Mopuri et al., 2018) under no data settings. Interestingly, the attack is more effective on the baselines compared to the network trained with our methods (SourceSep-CC and SourceSep-CM)

few co-occurrence thresholds. In addition, the co-occurrence based blending network marginally outperforms the clustering based technique which further strengthens the claim that our co-occurrence based blending technique allows the network to better separate semantic objects which are more likely to co-occur.

4.4 Adversarial Robustness

Existing defence mechanisms (Arnab et al., 2018; Xie et al., 2017; Guo et al., 2018; Madry et al., 2018) against adversarial attacks (Goodfellow et al., 2014; Kurakin et al., 2016; Moosavi-Dezfooli et al., 2016, 2017) attempt to reduce the impact of adversarial examples. Typically, adversarial defence mechanisms follow two main directions: (i) simply modifying the classifier to make it more robust, or (ii) transforming the adversarial examples in inference time. Even though our pipeline does not fall under either of these categories,we further claim our technique works as an implicit defense mechanism against adversarial images similar to Yun et al. (2019); Zhang et al. (2018); Inoue (2018); Cubuk et al. (2019); Harris et al. (2020); Chou et al. (2020). This is because the network optimization, in the form of source separation, enhances the capability of interacting with noisy features while imposing a high degree of resilience to interference from the superimposed image.

Adversarial Attacks. We generate adversarial examples using various techniques, including the Universal Adversarial Perturbation (UAP) (Moosavi-Dezfooli et al., 2017) and Generalizable Data-free Universal Adversarial Perturbation (GD-UAP) (Mopuri et al., 2018) under different settings. We use publicly available computed universal perturbations from these methods to generate adversarial examples for the PASCAL VOC 2012 val set. For UAP, which is a black-box attack, we generate adversarial images with both ResNet152 and GoogleNet based universal perturbations. GD-UAP is a grey-box attack, as it generates a perturbation based on the source data (VOC 2012 train set) and the backbone network (ResNet101). For GD-UAP, we compare different levels of adversarial attack strength by generating the perturbation based on various amounts of source data information.

Robustness of Segmentation Networks. We evaluate the robustness of different methods to adversarial examples and show how source separation-driven training learns to substantially mitigate performance loss due to perturbation. Table 6 shows the robustness of different baselines and our approaches on the PASCAL VOC 2012 validation dataset. In general, DeepLab-based methods (Chen et al., 2018) achieve

Table 7 Quantitative comparison (in terms of max F_β and MAE) with recent methods. Down arrow means lower is better and up arrow means higher is better

Methods	ECSSD (Yan et al., 2013)	
	F_β ↑	MAE ↓
DeepLabv3-ResNet50 (Chen et al., 2017)	0.906	0.045
DeepLabv3 + Mixup (Zhang et al., 2018)	0.893	0.057
DeepLabv3 + CutMix (Yun et al., 2019)	0.903	0.050
DeepLabv3 + SourceSep-CM	**0.909**	**0.043**

Table 8 Significance of feature denoising stage (DN)

Methods	mIoU
DeepLabv3-ResNet50 (Chen et al., 2017)	75.9
DeepLabv3 + SourceSep-CC (w/o DN)	75.4
DeepLabv3 + SourceSep-CC (w/ DN)	**75.7**
DeepLabv3 + SourceSep-CM (w/o DN)	76.1
DeepLabv3 + SourceSep-CM (w/ DN)	**76.2**
DeepLabv3-ResNet101 (Chen et al., 2017)	77.1
DeepLabv3 + SourceSep-CC (w/o DN)	76.4
DeepLabv3 + SourceSep-CC (w/ DN)	**77.9**
DeepLabv3 + SourceSep-CM (w/o DN)	78.3
DeepLabv3 + SourceSep-CM (w/ DN)	**78.9**

It is clear that the feature denoising stage further improves the overall performance under both mixing techniques

Table 9 Performance comparison with and without the source separation head (SSH) in the source separator module

Methods	mIoU
DeepLabv3-ResNet101 (Chen et al., 2017)	77.1
DeepLabv3 + SourceSep-CC (w/o SSH)	76.1
DeepLabv3 + SourceSep-CC (w/ SSH)	**76.4**
DeepLabv3 + SourceSep-CM (w/o SSH)	77.9
DeepLabv3 + SourceSep-CM (w/ SSH)	**78.3**

Including the source separation head marginally improves the overall performance for both techniques. Note, we report the numbers without the denoising stage

higher mIoU for the segmentation task on clean examples and are also shown to be more robust to adversarial samples compared to the shallower networks (Arnab et al., 2018). In the case of black-box attacks, the adversarial examples originally generated by UAP on ResNet152, are less malignant when the clustering based blending method is applied in the source separation network, while being effective in substantially reducing the performance of other methods.

When we apply a gray-box attack under the setting (R-All), where VOC 2012 training data and the ResNet101 network are used to generate the perturbation, DeepLabv3 and Mixup show robustness against adversarial examples which is improved by applying our blending strategies. Surprisingly, the performance of CutMix is substantially reduced when tested against adversarial samples generated by GD-UAP. Similarly, we find that DeepLabv3, Mixup, and CutMix are also vulnerable to adversarial cases under the *R-No* and *R-Part* settings, where no data and partial data is used, respectively to generate the perturbations. Notably, DeepLabv3+SourceSep-CC and DeepLabv3+SourceSep-CM exhibit significant robustness to extreme cases which further reveals the importance of source separation training pipeline to successfully relate internal activations corresponding to common sources in the adversarial images. In general, the SourceSep-CC network shows more robustness than the SourceSep-CM network under various adversarial settings. The reason behind the greater robustness is that the clustering based technique allows the source separation network to be trained on a larger set of noisy mixed data, while the co-occurrence based method allows mixing only between images which have semantic objects that are likely to co-occur.

Figure 6 depicts the outputs of baselines and our approaches to the GD-UAP attack on the PASCAL VOC 2012 validation set. It is clear that our proposed approaches are more robust against the GD-UAP attack compared to the baseline methods. These observations and results on different attacks reveal that the relative ranking of adversarial robustness for the different networks is improved with the addition of our proposed blending based source separation training.

4.5 Results on Salient Object Detection

We further validate our proposed co-occurrence based mixing technique on the salient object detection (SOD) task and present a comparison with existing mixing methods (Yun et al., 2019; Zhang et al., 2018) in Table 7. Similar to the task of semantic segmentation, we train the DeepLabv3-ResNet50 (Chen et al., 2017) network with various mixing strategies on the DUT-S dataset (Wang et al., 2017) and evaluate on ECSSD dataset (Shi et al., 2016). Since DUT-S dataset does not provide any semantic segmentation ground-truth, we can not directly apply our co-occurrence based image blending technique during training. Towards this goal, we first generate pseudo semantic labels for DUT-S by simply passing the images to the DeepLabv3-ResNet50 network trained on PASCAL VOC 2012 dataset for semantic segmentation task.

Table 10 Influence of co-occurrence based image blending techniques and other components on improving overall performance. DN denotes feature denoising stage

Methods	mIoU
DeepLabv3-ResNet50 (Chen et al., 2017)	75.1
+ Mixup (Zhang et al., 2018)	73.6
+ Mixup (Zhang et al., 2018) + Co-occurrence	74.2
+ Mixup (Zhang et al., 2018) + Co-occurrence + SourceSep	76.1
+ Mixup (Zhang et al., 2018) + Co-occurrence + SourceSep + DN	**76.2**
DeepLabv3-ResNet101 (Chen et al., 2017)	77.1
+ Mixup (Zhang et al., 2018)	76.2
+ Mixup (Zhang et al., 2018) + Co-occurrence	77.2
+ Mixup (Zhang et al., 2018) + Co-occurrence + SourceSep	78.3
+ Mixup (Zhang et al., 2018) + Co-occurrence + SourceSep + DN	**78.9**

While the boundaries of the generated pseudo-labels are not perfect, the predicted class labels can still be used as image-level labels in the image blending process.

From Table 7, it can be seen that our DeepLabv3+ SourceSep-CM method outperforms or achieves competitive performance compared to the baseline methods.

5 Ablation Studies

In this section, we examine the variants of our proposed pipelines by considering three different settings: (i) effectiveness of the feature denoising stage and source separation head (ii) influence of the co-occurrence based mixing technique, and (iii) impact of choosing maximum semantic categories in the co-occurrence based blending.

5.1 Feature Denoising and Source Separation Head

We examine the effectiveness of the feature denoising (DN) stage and report results in Table 8. Interestingly, using the DeepLabv3-ResNet101 (Chen et al., 2017) network as the backbone, the clustering based mixing approach exhibits larger improvement with the addition of the denoising stage than the co-occurrence based technique (1.5% vs. 0.6% improvement). The reason behind the larger improvement is that the clustering based technique allows the source separation network to be trained on a larger set of noisy mixed data, while the co-occurrence based method allows mixing only between images which have semantic objects that are likely to co-occur. This is why, with a deeper backbone network (e.g., DeepLabv3-ResNet101), the source separation training with categorical clustering is more noisy which allows the denoising stage to improve the performance more substantially.

We also conduct experiments (see Table 9) varying the source separator module, including the source separation head (SSH). It is clear that the overall performance of DeepLabv3-ResNet101 based source separation networks can be marginally improved with the addition of a source separation head (0.3% and 0.4% improvement respectively). We believe the source separation head allows the network to make a more informed final prediction based on the source *and* the phantom activations, and therefore learns to identify harmful features at inference time, leading to a more accurate prediction.

5.2 Influence of Co-occurrence based Mixup

We further tease out the importance of our proposed co-occurrence based technique by simply applying it with an existing mixup technique. Table 10 presents quantitative results comparing different components. DeepLabv3-ResNet50 with Mixup (Zhang et al., 2018) achieves 73.6% mIoU. The performance is improved by 0.6% when we apply co-occurrence matrix based blending with Mixup. The source separation training pipeline further improves the overall performance by 1.9% which is further improved by 0.1% by applying the denoising stage. From the results, it is clear that co-occurrence based blending has an clear influence on improving the segmentation performance. As shown in Table 10, the results are consistent when we use DeepLabv3-ResNet101 as the backbone.

5.3 Impact of Choosing Maximum Semantic Categories in Co-occurrence based Mixup

Existing mixing based techniques have been applied mostly on datasets where there exists one dominant semantic object (e.g., ImageNet (Russakovsky et al., 2015), CIFAR-10). However, semantic segmentation datasets naturally contain images with more than one category in complex scenes. Therefore, randomly combining two source images based on the co-occurrence likelihood during training to achieve the desired objective is still a more significant challenge than one might expect in the context of dense labeling. For instance,

Table 11 We examine the influence of changing the maximum number of unique semantic objects during the blending process

Methods	mIoU
DeepLabv3-ResNet50 (Chen et al., 2017)	75.1
DeepLabv3 + SourceSep-CM (max = 2)	75.9
DeepLabv3 + SourceSep-CM (max = 3)	**76.1**
DeepLabv3 + SourceSep-CM (max = 4)	75.6
DeepLabv3 + SourceSep-CM (max = ∞)	76.1
DeepLabv3 + SourceSep-CM (max = ∞) + γ-thres	**76.1**
DeepLabv3-ResNet101 (Chen et al., 2017)	77.1
DeepLabv3 + SourceSep-CM (max = 2)	77.2
DeepLabv3 + SourceSep-CM (max = 3)	77.6
DeepLabv3 + SourceSep-CM (max = 4)	77.7
DeepLabv3 + SourceSep-CM (max = ∞)	77.0
DeepLabv3 + SourceSep-CM (max = ∞) + γ-thres	**78.3**

Note, we report the number without any denoising stage. 'γ-thres' refers to the default model which allows the blended images over the maximum threshold but sets the mixing ratio, γ, to 0.9

if we blend two images containing three and four semantic objects with a lower mixing ratio (i.e., assign more weight to the *target* image), there is a high chance that the mixed image will lack context. Also, images with seven unique objects are extremely rare or non-existent in the datasets we explore, and therefore this type of image may be too different from the target distribution. To explore this issue, we first restrict the number of unique semantic categories to be blended (e.g., we do not blend images if the maximum threshold is exceeded). We additionally try a strategy where we mix the images, but set the mixing ratio to a constant value (0.9) if this threshold is surpassed. The intuition is that, for a pair of images where the total number of unique semantic categories is higher than the threshold, we want to reduce the amount of blending by assigning more weight to the dominant image.

Table 11 presents the results of choosing different thresholds in the co-occurrence based mixing process. It is clear that restricting the mixing ratio when the maximum number of unique objects in the blended image is greater than a certain threshold achieves higher mIoU compared to other alternatives.

6 Discussion and Conclusion

Training with the categorical clustering and a co-occurrence based source separation pipeline enables learning resilient features, separating sources of activation, and resolving ambiguity with richer contextual information. Although DeepLabv3 is a powerful segmentation network, there are cases (see Fig. 7) where background objects are correctly classified (car and plane) but other semantic categories are not separated correctly due to high degrees of occlusion (person on the stairs, see Fig. 7 right). In contrast, the source separation based learning approaches are highly capable of resolving such cases by learning to separate source objects and tying them to specific regions.

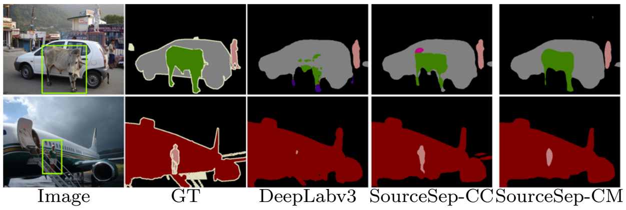

Fig. 7 Two challenging images where semantic objects are highly occluded. While pixels belonging to the dominant semantic categories are identified correctly, the prediction fails to relate activation tied to smaller occluding features to correct categorical assignments. This is resolved when trained using our proposed mixing based source separation networks

In summary, we have presented two approaches to train CNNs based on the notion of source separation. This process includes, as one major component, careful creation of categorical collisions in data during training. This results in improved segmentation performance, and also promotes significant robustness to adversarial perturbations. Denoising in the form of fine-tuning shows further improvement along both these dimensions.

References

Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. In *ICLR*.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., Zisserman, A. (2015). The Pascal visual object classes challenge: A retrospective. *IJCV*.

Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *CVPR*.

Noh, H., Hong, S., Han, B. (2015). Learning deconvolution network for semantic segmentation. In *ICCV*.

Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. In: *TPAMI*.

Ghiasi, G., & Fowlkes, C. C. (2016). Laplacian pyramid reconstruction and refinement for semantic segmentation. In *ECCV*.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Pyramid scene parsing network. In *CVPR*.

Islam, M. A., Rochan, M.M., Bruce, N. D. B., Wang, Y. (2017). Gated feedback refinement network for dense image labeling. In *CVPR*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *TPAMI*.

Islam, M. A., Kalash, M., Bruce, N. D. (2018). Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *CVPR*.

He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. In *ICCV*.

Li, K., Hariharan, B., Malik, J. (2016). Iterative instance segmentation. In *CVPR*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. (2015). ImageNet large scale visual recognition challenge. *IJCV*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *ECCV*.

Singh, K. K., Mahajan, D., Grauman, K., Lee, Y.J., Feiszli, M., Ghadiyaram, D. (2020). Don't judge an object by its context: Learning to overcome contextual bias. In *CVPR*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2015). Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*.

Georgiev, P., Theis, F., & Cichocki, A. (2005). Sparse component analysis and blind source separation of underdetermined mixtures. *TNN, 16*(4), 992–996.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P. (2015). Joint optimization of masks and deep recurrent neural networks for monaural source separation. *ASLP*.

Islam, M. A., Kowal, M., Derpanis, K. G., Bruce, K. G. (2020). Feature binding with category-dependant mixup for semantic segmentation and adversarial robustness. In *BMVC*.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587.

Takikawa, T., Acuna, D., Jampani, V., Fidler, S. (2019). Gated-SCNN: Gated shape CNNs for semantic segmentation. In *ICCV*.

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *NIPS*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. In *CVPR*.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *ICLR*.

Bishop, C. M. (1995). Training with noise is equivalent to tikhonov regularization. *Neural computation*.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., Lakshminarayanan, B. (2019). AUGMIX: A simple data processing method to improve robustness and uncertainty. In *ICLR*.

Kim, J.-H., Choo, W., Song, H. O. (2020). Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *ICML*.

Tokozume, Y., Ushiku, Y., Harada, T. (2018). Between-class learning for image classification. In *CVPR*.

Inoue, H. (2018). Data augmentation by pairing samples for images classification. arXiv:1801.02929.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. In *CVPR*.

French, G., Aila, T., Laine, S., Mackiewicz, M., Finlayson, G. (2020). Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. In *BMVC*.

Harris, E., Marcu, A., Painter, M., Niranjan, M., Hare, A. P.-B. J. (2020). FMix: Enhancing mixed sample data augmentation. arXiv preprint arXiv:2002.12047

Chou, H.-P., Chang, S.-C., Pan, J.-Y., Wei, W., Juan, D.-C. (2020). Remix: Rebalanced mixup. In *ECCVW*.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y. (2019). Manifold mixup: Better representations by interpolating hidden states. In *ICML*.

Guo, H., Mao, Y., Zhang, R. (2019). Mixup as locally linear out-of-manifold regularization. In *AAAI*.

DeVries,T., Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A. (2017). Automatic differentiation in pytorch.

Lin, G., Milan, A., Shen, C., Reid, I. (2017). RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*.

Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J. (2011). Semantic contours from inverse detectors. In *ICCV*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Choi, M. J., Torralba, A., & Willsky, A. S. (2012). Context models and out-of-context objects. *Pattern Recognition Letters*.

Peyre, J., Sivic, J., Laptev, I., Schmid, C. (2017). Weakly-supervised learning of visual relations. In *ICCV*.

Arnab, A., Miksik, O., Torr, P. H. (2018). On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*.

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A. (2017). Adversarial examples for semantic segmentation and object detection. In *ICCV*.

Guo, C., Rana, M., Cisse, M., van der Maaten, L. (2018). Countering adversarial images using input transformations. In *ICLR*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *ICLR*.

Goodfellow, I. J., Shlens, J., Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv:1412.6572.

Kurakin, A., Goodfellow, I., Bengio, S. (2016). Adversarial examples in the physical world. arXiv:1607.02533.

Moosavi-Dezfooli, S.-M., Fawzi, A., Frossard, P. (2016). DeepFool: a simple and accurate method to fool deep neural networks. In *CVPR*.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P. (2017). Universal adversarial perturbations. In *CVPR*.

Mopuri, K. R., Ganeshan, A., Babu, R. V. (2018). Generalizable data-free objective for crafting universal adversarial perturbations. *TPAMI*.

Yan, Q., Xu, L., Shi, J., Jia, J. (2013). Hierarchical saliency detection. In *CVPR*.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X. (2017). Learning to detect salient objects with image-level supervision. In *CVPR*.

Shi, J., Yan, Q., Xu, L., Jia, J. (2016). Hierarchical image saliency detection on extended CSSD. *TPAMI*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.