ClassMix: Segmentation-Based Data Augmentation for Semi-Supervised Learning

Viktor Olsson[1,2]*, Wilhelm Tranheden[1,2]*, Juliano Pinto[1], Lennart Svensson[1]
[1]Chalmers University of Technology, Gothenburg, Sweden
[2]Volvo Cars, Gothenburg, Sweden
{viktor.olsson.3,wilhelm.tranheden}@volvocars.com, {juliano,lennart.svensson}@chalmers.se

Abstract

The state of the art in semantic segmentation is steadily increasing in performance, resulting in more precise and reliable segmentations in many different applications. However, progress is limited by the cost of generating labels for training, which sometimes requires hours of manual labor for a single image. Because of this, semi-supervised methods have been applied to this task, with varying degrees of success. A key challenge is that common augmentations used in semi-supervised classification are less effective for semantic segmentation. We propose a novel data augmentation mechanism called ClassMix, which generates augmentations by mixing unlabelled samples, by leveraging on the network's predictions for respecting object boundaries. We evaluate this augmentation technique on two common semi-supervised semantic segmentation benchmarks, showing that it attains state-of-the-art results. Lastly, we also provide extensive ablation studies comparing different design decisions and training regimes.

1. Introduction

Semantic segmentation is the task of assigning a semantic label to each pixel of an image. This is an essential part of many applications such as autonomous driving, medical imaging and scene understanding. Significant progress has been made in the area based on fully convolutional network architectures [27, 3, 39]. When training deep learning models for semantic segmentation, a common bottleneck is the availability of ground-truth labels. In contrast, unlabelled data is usually abundant, and effectively leveraging it has the potential to increase performance with low cost.

Semi-supervised learning based on consistency regularization has recently seen remarkable progress for image classification [37, 33], utilizing strong data augmentations to enforce consistent predictions on unlabelled images. Augmentation techniques commonly used in classification have however proved ineffective for semi-supervised semantic segmentation [13, 29]. Recent works have addressed this issue by either applying perturbations on an encoded state of the network instead of the input [29], or by using the augmentation technique CutMix [38] to enforce consistent predictions over mixed samples [13, 21].

We propose a segmentation-based data augmentation strategy, ClassMix, and describe how it can be used for semi-supervised semantic segmentation. The augmentation strategy cuts half of the predicted classes from one image and pastes them onto another image, forming a new sample that better respects semantic boundaries while remaining accessible without ground-truth annotations. This is achieved by exploiting the fact that the network learns to predict a pixel-level semantic map of the original images. The predictions on mixed images are subsequently trained to be consistent with predictions made on the images before mixing. Following a recent trend in state-of-the-art consistency regularization for classification we also integrate entropy minimization [37, 33, 2, 1], encouraging the network to generate predictions with low entropy on unlabelled data. We use pseudo-labelling [23] to accomplish this, and provide further motivations for combining it with ClassMix. Our proposed method is evaluated on established benchmarks for semi-supervised semantic segmentation, and an ablation study is included, analyzing the individual impact of different design and experimental choices.

Our main contributions can be summarised as: (1) We introduce an augmentation strategy which is novel for semantic segmentation, which we call ClassMix. (2) We incorporate ClassMix in a unified framework that makes use of consistency regularization and pseudo-labelling for semantic segmentation. (3) We demonstrate the effectiveness of our method by achieving state-of-the-art results in semi-supervised learning for the Cityscapes dataset [6], as well as competitive results for the Pascal VOC dataset [10]. Code is available at `https://github.com/WilhelmT/ClassMix`.

*Equal contribution.

2. Related Work

For semantic segmentation, semi-supervised learning has been explored with techniques based on adversarial learning [19, 28, 31], consistency regularization [13, 21, 25, 30], and pseudo-labelling [12, 5]. Our proposed method primarily incorporates ideas from the latter two approaches, which are expanded upon in subsequent sections.

2.1. Consistency Regularization

The core idea in consistency regularization is that predictions for unlabelled data should be invariant to perturbations. A popular technique for classification is augmentation anchoring [37, 33, 1], where predictions performed on strongly augmented samples are enforced to follow predictions on weakly augmented versions of the same images. Our method utilizes augmentation anchoring in that consistency is enforced from unperturbed images to mixed images. Mixing images will create occlusions and classes in difficult contexts, hence being a strong augmentation.

Until recently, consistency regularization had been successfully applied for semantic segmentation only in the context of medical imaging [25, 30]. Researchers pointed out the difficulties of performing consistency regularization for semantic segmentation, such as the violation of the cluster assumption, as described in [13, 29]. Ouali et al. [29] propose to apply perturbations to the encoder's output, where the cluster assumption is shown to hold. Other approaches [13, 21] instead use a data augmentation technique called CutMix [38], which composites new images by mixing two original images, resulting in images with some pixels from one image and some pixels from another image. Our proposed method, ClassMix, builds upon this line of research by using predictions of a segmentation network to construct the mixing. In this way, we can enforce consistency over highly varied mixed samples while a the same time better respecting the semantic boundaries of the original images.

2.2. Pseudo-labelling

Another technique used for semi-supervised learning is pseudo-labelling, training against targets based on the network class predictions, first introduced in [23]. Its primary motivation comes from entropy minimization, to encourage the network to perform confident predictions on unlabelled images. Such techniques have shown recent success in semi-supervised semantic segmentation [12, 5, 20]. Some methods of semi-supervised learning for classification [37, 33, 2, 1] integrate entropy minimization in the consistency regularization framework. This is achieved by having consistency targets either sharpened [37, 2, 1] or pseudo-labelled [33]. Our proposed method of consistency regularization also naturally incorporates pseudo-labelling, as it prevents predictions close to mixing borders being trained to unreasonable classes, which will be further explained in coming sections.

2.3. Related Augmentation Strategies

In the CutMix algorithm [38], randomized rectangular regions are cut out from one image and pasted onto another. This technique is based on mask-based mixing, where two images are mixed using a binary mask of the same size as the images. Our proposed technique, ClassMix, is based on a similar principle of combining images and makes use of predicted segmentations to generate the binary masks, instead of rectangles.

ClassMix also shares similarities with other segmentation-based augmentation strategies [9, 32, 8, 35, 11], where annotated single instances of objects are cut out of images, and pasted onto new background scenes. Our way of combining two images conditioned on the predicted semantic maps exploits the same idea of compositing images. However, in contrast to several existing techniques our proposed augmentation does not rely on access to ground-truth segmentation-masks, allowing us to learn from unlabelled images in the semi-supervised setting. Additionally, we perform semantic segmentation with multiple classes present in each image rather than single instances, allowing variety by randomizing which classes to transfer. As previously mentioned, ClassMix is formulated as a generalisation of CutMix, using a binary mask to mix two randomly sampled images. This means that we only distinguish between foreground and background when generating the binary mask and not when training on the mixed images. Segmenting both foreground and background images means that semantic objects recognized by the network do not only have to be invariant to their context, but invariant to a diverse set of occlusions as well.

3. Method

This section describes the proposed approach for semi-supervised semantic segmentation. It starts by explaining the data augmentation mechanism ClassMix, followed by a description of the loss function used, along with other details about the training procedure.

3.1. ClassMix: Main Idea

The proposed method performs semi-supervised semantic segmentation by using a novel data augmentation technique, ClassMix, which uses the unlabelled samples in the dataset to synthesize new images and corresponding artificial labels ("artificial label" in this context is used to refer to the target that is used to train on the augmented image in the augmentation anchoring setup). ClassMix uses two unlabelled images as input and outputs a new augmented image, together with the corresponding artificial label for it.

This augmented output is comprised of a mix of the inputs, where half of the semantic classes of one of the images are pasted on top of the other, resulting in an output which is novel and diverse, but still rather similar to the other images in the dataset.

Figure 1 illustrates the essence of how ClassMix works. Two unlabelled images, A and B, are sampled from the dataset. Both are fed through the segmentation network, f_θ, which outputs the predictions S_A and S_B. A binary mask M is generated by randomly selecting half of the classes present in the argmaxed prediction S_A and setting the pixels from those classes to have value 1 in M, whereas all others will have value 0. This mask is then used to mix images A and B into the augmented image X_A, which will contain pixels from A where the mask had 1's and pixels from B elsewhere. The same mixing is also done to the predictions S_A and S_B, resulting in the artificial label Y_A. While artifacts may appear because of the nature of the mixing strategy, as training progresses they become fewer and smaller. Additionally, consistency regulation tends to yield good performance also with imperfect labels and this is further confirmed by our strong results.

3.2. ClassMix: Details

Two other techniques were added on top of the version of ClassMix presented in the previous section for improving its performance. This subsection explains those changes and provides a detailed description of the final ClassMix algorithm in pseudocode, in Algorithm 1.

Algorithm 1 ClassMix algorithm

Require: Two unlabelled samples A and B, segmentation network $f_{\theta'}$.

$S_A \leftarrow f_{\theta'}(A)$

$S_B \leftarrow f_{\theta'}(B)$

$\tilde{S}_A \leftarrow \arg\max_{c'} S_A(i,j,c')$ ▷ Take pixel-wise argmax over classes.

$C \leftarrow$ Set of the different classes present in $\tilde{S}_A$

$c \leftarrow$ Randomly selected subset of C such that $|c| = |C|/2$

For all i,j: $M(i,j) = \begin{cases} 1, \text{ if } \tilde{S}_A(i,j) \in c \\ 0, \text{ otherwise} \end{cases}$ ▷ Create binary mask.

$X_A \leftarrow M \odot A + (1-M) \odot B$ ▷ Mix images.

$Y_A \leftarrow M \odot S_A + (1-M) \odot S_B$ ▷ Mix predictions.

return X_A, Y_A

Mean-Teacher Framework. In order to improve stability in the predictions for ClassMix, we follow a trend in state-of-the-art semi-supervised learning [13, 2, 36] and use the Mean Teacher Framework, introduced in [34]. Instead of using f_θ to make predictions for the inputs images A and B in ClassMix, we use $f_{\theta'}$, where θ' is an exponential moving average of the previous values of θ throughout the optimization. This type of temporal ensembling is cheap and simple to introduce in ClassMix, and results in more stable predictions throughout the training, and consequently more stable artificial labels for the augmented images. The network f_θ is then used to make predictions on the mixed images X_A, and the parameters θ are subsequently updated using gradient descent.

Pseudo-labelled Output. Another important detail about ClassMix is that, when generating labels for the augmented image, the artificial label Y_A is "argmaxed". That is, the probability mass function over classes for each pixel is changed to a one-hot vector with a one in the class which was assigned the highest probability, zero elsewhere. This forms a pseudo-label to be used in training, and it is a commonly used technique in semi-supervised learning in order to encourage the network to perform confident predictions.

For ClassMix, pseudo-labelling serves an additional purpose, namely eliminating uncertainty along borders. Since the mask M is generated from the output prediction of A, the edges of the mask will be aligned with the decision boundaries of the semantic map. This comes with the issue that predictions are especially uncertain close to the mixing borders, as the segmentation task is hardest close to class boundaries [24]. This results in a problem we denote label contamination, illustrated in figure 2. When the classes chosen by M are pasted on top of image B, their adjacent context will often change, resulting in poor artificial labels. Pseudo-labelling effectively mitigates this issue, since the probability mass function for each pixel is changed to a one-hot vector for the most likely class, therefore "sharpening" the artificial labels, resulting in no contamination.

3.3. Loss and Training

For all the experiments in this paper, we train the parameters of the semantic segmentation network f_θ by minimizing the following loss:

$$L(\theta) = \mathbb{E}\left[\ell\Big(f_\theta(X_L), Y_L\Big) + \lambda\ell\Big(f_\theta(X_A), Y_A\Big)\right] . \quad (1)$$

In this expectation, X_L is an image sampled uniformly at random from the dataset of labelled images, and Y_L is its corresponding ground-truth semantic map. The random variables X_A and Y_A are respectively the augmented image and its artificial label, produced by the ClassMix augmentation method (as described in algorithm 1), where the input images A and B are sampled uniformly at random from the unlabelled dataset. (in practice the augmentations are computed by mixing all the images within a batch, for efficiency reasons; we refer interested readers to our code for further details). Lastly, λ is a hyper-parameter that controls the balance between the supervised and unsupervised terms, and ℓ

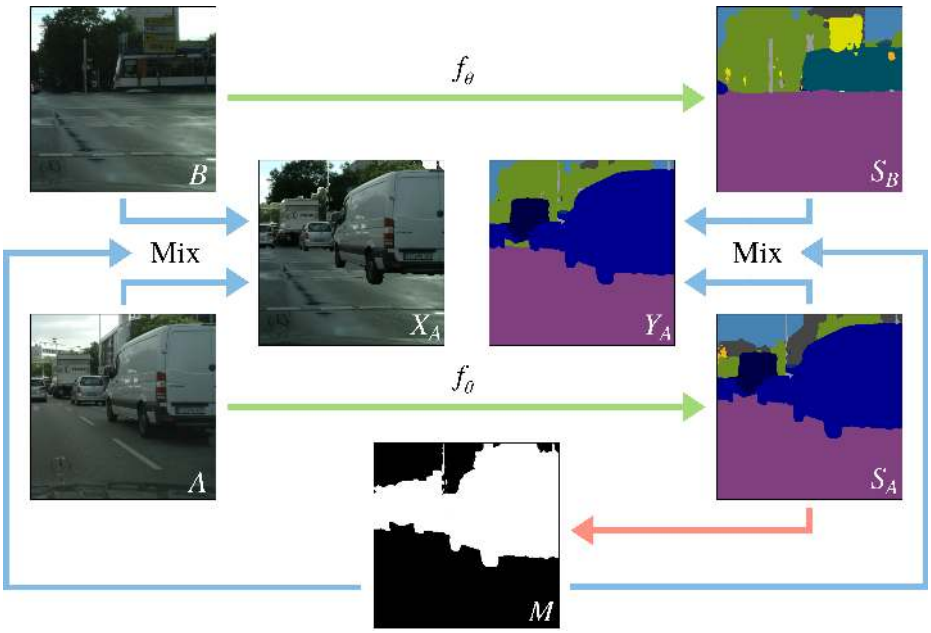

Figure 1: ClassMix augmentation technique. Two images A and B are sampled from the unlabelled dataset. Based on the prediction S_A of image A, a binary mask M is created. The mask is then used to mix the images A and B and their respective predictions into an augmented image X_A and the corresponding artificial label Y_A.

Figure 2: Toy example of label contamination with 3 different classes. Left: Ground-truth labels; red is class 1, blue is class 2. Middle: Prediction made by network; regions where the network is uncertain between classes 1 and 2 have a mix of red and blue colors. The decision boundary is marked with a white line. Right: The red class is pasted on top of a new image, which is comprised entirely of the third class. Note how the pasted class still brings some uncertainty of class 2 (blue) to the new image. This results in problematic artificial labels for training, since the context around the pasted object now changed to class 3.

is the cross-entropy loss, averaged over all pixel positions in the semantic maps, i.e.

$$\ell(S,Y) = -\frac{1}{W \cdot H} \sum_{i=1}^{W} \sum_{j=1}^{H} \left(\sum_{c=1}^{C} Y(i,j,c) \cdot \log S(i,j,c) \right), \tag{2}$$

where W and H are the width and height of the images, and $S(i,j,c)$ and $Y(i,j,c)$ are the probabilities that the pixel in coordinates i, j belongs to class c, according to the prediction S and target Y, respectively. We train θ by stochastic gradient descent on this loss, imposing batches with 50% labelled data and 50% augmented data.

It is beneficial to the training progress that the unsupervised weight λ starts close to zero, because initially the network predictions are of low quality and therefore the pseudo-labels generated will not be reasonable targets for training on the augmented images. As the predictions of the network improve, this weight can then be increased. This was accomplished by setting the value of λ for an augmented sample as the proportion of pixels in its artificial label where the probability of the most likely class is above a predetermined threshold τ. This results in a value between 0 and 1, which we empirically found to serve as an adequate proxy for the quality of the predictions, roughly in line with [13, 14].

4. Experiments

In order to evaluate the proposed method, we perform experiments on two common semi-supervised semantic segmentation datasets, and this section presents the results obtained. Additionally, an extensive ablation study for motivating our design decisions is also provided, where we further investigate the properties of ClassMix and its components.

4.1. Implementation Details and Datasets

Our method is implemented using the PyTorch framework and training was performed on two Tesla V100 GPUs. We adopt the DeepLab-v2 framework [4] with a ResNet101

Figure 3: Images and corresponding semantic maps from the Cityscapes dataset.

Figure 4: Images and corresponding semantic maps from the Pascal VOC 2012 dataset.

backbone [18] pretrained on ImageNet [7] and MSCOCO [26], identical to the ones used in [19, 12, 28]. As optimizer, we use Stochastic Gradient Descent with Nesterov acceleration, and a base learning rate of 2.5×10^{-4}, decreased with polynomial decay with power 0.9 as used in [4]. Momentum is set to 0.9 and weight decay to 5×10^{-4}.

We present results for two semantic segmentation datasets, Cityscapes [6] and Pascal VOC 2012 [10]. The Cityscapes urban scenery dataset contains 2,975 training images and 500 validations images. We resize images to 512×1024 with no random cropping, scaling or flipping, use batches with 2 labelled and 2 unlabelled samples and train for 40k iterations, all in line with [19]. For the Pascal VOC 2012 dataset we use the original images along with the extra annotated images from the Semantic Boundaries dataset [16], resulting in 10,582 training images and 1,449 validation images. Images are randomly scaled between 0.5 and 1.5 as well as randomly horizontally flipped and after that cropped to a size of 321×321 pixels, also in line with [19]. We train for 40k iterations using batches with 10 labelled and 10 unlabelled samples.

Figures 3 and 4 show example images of both datasets along with their corresponding ground truth semantic maps. It is clear that the images in Cityscapes contain a lot more classes in each image than the Pascal images do. At the same time the semantic maps are more consistent throughout the images in the Cityscapes dataset than between Pascal images, for example the road and sky classes are almost always present and in approximately the same place.

4.2. Results

Cityscapes. In Table 1 we present our results for the Cityscapes dataset, given as mean Intersection over Union (mIoU) scores. We have performed experiments for four proportions of labelled samples, which are given along with baselines that are trained in a supervised fashion on the corresponding data amounts. In the table we also provide results from four other papers, all using the same DeepLab-v2 framework. Hung et al. and Mittal et al. use an adversarial approach [19, 28], French et al. use consistency regularization [13] and Feng et al. use a self-training scheme [12]. We note that our results are higher for three out of four data amounts and that our improvement from the baseline result to the SSL result is higher for all amounts of training data.

The fact that we achieve, to the best of our knowledge, the best SSL-results on the Cityscapes dataset further supports that consistency regularization can be successfully applied to semi-supervised semantic segmentation. French et al. use a method similar to ours [13], where they enforce consistency with CutMix as their mixing algorithm, instead of our ClassMix. We believe that one reason for the higher performance of ClassMix is the diversity of the masks created. This diversity stems from the fact that each image includes many classes and that each class often contains several objects. Since there are many classes in each image, an image rarely has the exact same classes being selected for mask generation several times, meaning that the masks based on a given image will be varied over the course of training. Furthermore, since each class often contains several objects, the masks will naturally become very irregular, and hence very different between images; when using CutMix, the masks will not be nearly as varied.

We believe that another reason that ClassMix works well is that the masks are based on the semantics of the images, as discussed previously. This minimizes the occurrence of partial objects in the mixed image, which are difficult to predict and make learning unnecessarily hard. It also means that mixing borders will be close to being aligned with boundaries of objects. This creates mixed images that better respect the semantic boundaries of the original images. They are consequently more realistic looking than images created using, e.g., CutMix, and lie closer to the underlying data distribution.

A third reason for ClassMix performing well for Cityscapes may be that images are similar within the dataset. All images have the road class in the bottom and sky at the top, and whenever there are cars or people, for example, they are roughly in the same place. Another way to put this is that classes are not uniformly distributed across the image area, but instead clustered in smaller regions of the image, as is shown by the spatial distribution of classes in Figure 5. Because of this, objects that are pasted from one image to another are likely to end up in a reasonable

Table 1: Performance (mIoU) on Cityscapes validation set averaged over three runs. Results from four previous papers are provided for comparison, all using the same DeepLab-v2 network with ResNet-101 backbone.

Labelled samples	1/30	1/8	1/4	1/2	Full (2975)
Baseline	-	55.5%	59.9%	64.1%	66.4%
Adversarial [19]	-	58.8%	62.3%	65.7%	-
Improvement	-	3.3	2.4	1.6	-
Baseline	-	56.2%	60.2%	-	66.0%
s4GAN [28][1]	-	59.3%	61.9%	-	65.8%
Improvement	-	3.1	1.7	-	-0.2
Baseline	44.41%	55.25%	60.57%	-	67.53%
French et al. [13][1]	51.20%	60.34%	63.87%	-	-
Improvement	6.79	5.09	3.3	-	-
Baseline	45.5 %	56.7%	61.1%	-	66.9%
DST-CBC [12]	48.7 %	60.5%	**64.4%**	-	-
Improvement	3.2	3.8	3.3	-	-
Baseline	43.84%±0.71	54.84%±1.14	60.08%±0.62	63.02%±0.14	66.19%±0.11
Ours	**54.07%**±1.61	**61.35%**±0.62	63.63%±0.33	**66.29%**±0.47	-
Improvement	**10.23**	**6.51**	**3.72**	**3.27**	-

context, which may be important in the cases where objects are transferred without any surrounding context from the original image.

Pascal VOC 2012. In Table 2, we present the results from using our method on the Pascal VOC 2012 dataset. We compare our results to the same four papers as for Cityscapes. We note that our results are competitive, and that we have the strongest performance for two data amounts. There is, however, a significant difference in baselines between the different works, complicating the comparison of the results. In particular, French et al. have a significantly lower baseline, largely because their network is not pre-trained on MSCOCO, resulting in a bigger room for improvement, as shown in the table. We use a network pre-trained on MSCOCO because that is what most existing work is using. Our results can also be compared to those from Ouali et al., who obtain 69.4% mIoU on Pascal VOC 2012 for 1.5k samples using DeepLabv3 [29]. Our results for 1/8 (or 1323) labelled samples is higher, despite us using fewer labelled samples and a less sophisticated network architecture.

Our results are not as strong for the Pascal dataset as they are for Cityscapes. We believe that this is largely because Pascal contains very few classes in each image, usually only a background class and one or two foreground classes. This means that the diversity of masks in ClassMix will be very small, with the same class or classes frequently being selected for mask generation for any given image. This is in contrast to Cityscapes as discussed above. The images in

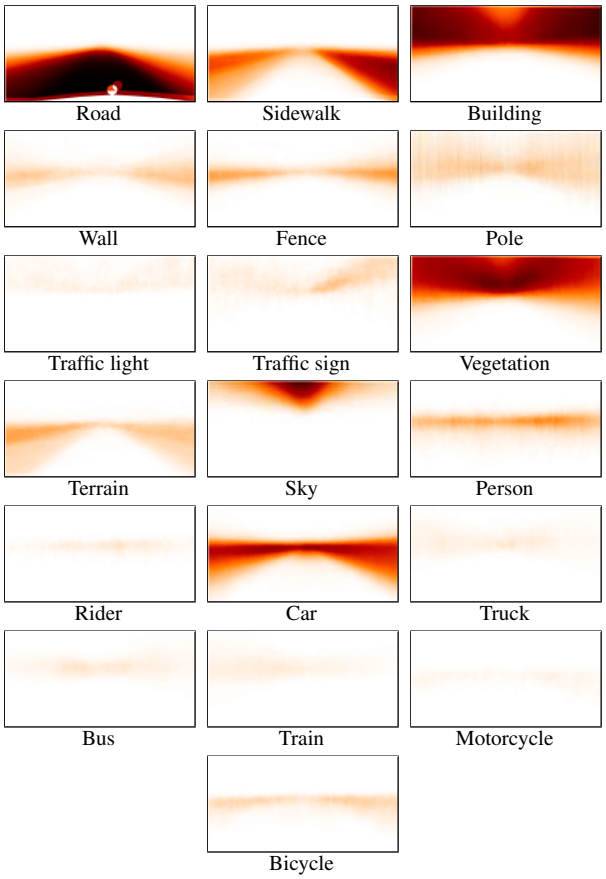

Figure 5: Spatial distribution of all classes in the Cityscapes training dataset. Dark pixels correspond to more frequent appearance.

[1] Same DeepLab-v2 network but with only ImageNet pre-training and not MSCOCO.

Table 2: Performance (mIoU) on the Pascal VOC 2012 validation set, results are given from a single run. Results from four previous papers are provided for comparison, all using the same DeepLab-v2 network with a ResNet-101 backbone.

Labelled samples	1/100	1/50	1/20	1/8	1/4	Full (10582)
Baseline	-	53.2%[2]	58.7%[2]	66.0%	68.3%	73.6%
Adversarial [19]	-	57.2%[2]	64.7%[2]	69.5%	72.1%	-
Improvement	-	4.0	6.0	3.5	**3.8**	-
Baseline	-	53.2%	58.7%	66.0%	-	73.6%
s4GAN [28]	-	63.3%	67.2%	**71.4%**	-	75.6%
Improvement	-	10.1	8.5	5.4	-	2.0
Baseline	33.09%	43.15%	52.05%	60.56%	-	72.59%
French et al. [13][3]	53.79%	64.81%	66.48%	67.60%	-	-
Improvement	**20.70**	**21.66**	**14.48**	**7.04**	-	-
Baseline	45.7%[4]	55.4%	62.2%	66.2%	68.7%	73.5%
DST-CBC [12]	**61.6%**[4]	65.5%	**69.3%**	70.7%	71.8%	-
Improvement	15.9	10.1	7.1	4.5	3.1	-
Baseline	42.47%	55.69%	61.36%	67.14%	70.20%	74.13%
Ours	54.18%	**66.15%**	67.77%	71.00%	**72.45%**	-
Improvement	11.71	10.46	6.41	3.86	2.25	-

the Pascal dataset are also not that similar to each other. There is no pattern to where in the images certain classes appear, or in what context, unlike Cityscapes. Therefore, pasted objects often end up in unreasonable contexts, which we believe is detrimental for performance. The patterns of where classes appear are made obvious by calculating the spatial distribution of classes, which is visualised for Pascal in Figure 6, and can be compared to Cityscapes in Figure 5. In these figures it is clear that the spatial distributions are much less uniform for Cityscapes than for Pascal. We note that in spite of these challenges for the Pascal VOC dataset, ClassMix still performs competitively with previous state of the art.

4.3. Ablation Study

We investigate our method further by training models with some components changed or removed, in order to see how much those specific components contribute to the performance of the overall algorithm. Additionally, we also experiment with additions to the algorithm, namely adding more augmentations and training for a longer time. Although such additions increase the final performance, they also make comparisons with other existing approaches unfair, which is why these results are not presented in subsection 4.2. The ablation results are presented in Table 3. All figures are from training a model with 1/8 labelled samples on the Cityscapes dataset, with the same settings as used for the main results except for the part being examined.

[2]As reported by [28].

[3]Same DeepLab-v2 network but with only ImageNet pre-training and not MSCOCO.

[4]Results for 100 (1/106) samples.

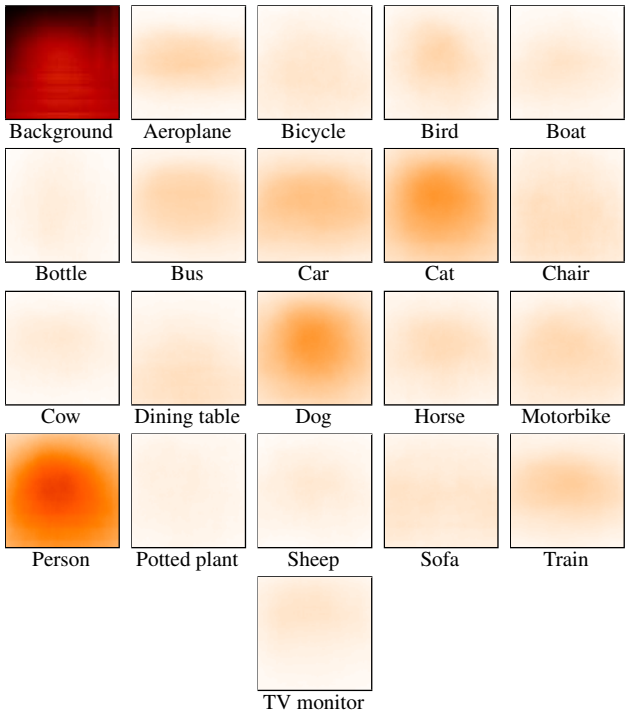

Figure 6: Spatial distribution of all classes in the Pascal training dataset. Dark pixels correspond to more frequent appearance.

First, we examine the effect of using different mixed sample data augmentations. Apart from ClassMix we try CutMix [38], as used for semi-supervised semantic segmentation in [13], and CowMix, introduced by French et al. [15]. We note that CowMix is very similar to the concur-

Table 3: Ablation study of the proposed method on the Cityscapes dataset. All results are mIoU scores averaged over three runs.

Settings	mIoU
Baseline	54.84%
Default SSL	61.35%
CowMix	60.37%
CutMix	59.12%
Pixel-wise threshold	58.61%
Sigmoid ramp up	60.58%
Constant unsupervised weight	60.58%
Squared error loss	58.74%
No Pseudo-label	60.15%
Random crop Baseline	56.42%
Random crop	62.16%
Extra augmentations	61.85%
80k iterations Baseline	55.05%
80k iterations	62.92%

rent FMix, introduced by Harris et al. [17]. As can be seen in Table 3, ClassMix performs significantly better than both other mixes. That CutMix performs the worst, we attribute to CutMix's masks being less varied than for the other methods. Both ClassMix and CowMix yield flexible masks and we speculate that ClassMix achieves higher results because the masks will follow semantic boundaries to a high degree, giving more natural borders between the two mixed images.

We try three different ways of weighting the unsupervised loss, additional to our default way of weighting it against the proportion of pixels that have a maximum predicted value above a threshold 0.968, as used in [13]. In contrast to this is pixel-wise threshold, where instead all pixels with predicted certainties below the threshold are masked and ignored in the loss. As can be seen in Table 3, using the pixel-wise threshold significantly lowers the results. We have found that this strategy masks almost all pixels close to class boundaries, as well as some small objects, such that no unsupervised training is ever performed on this kind of pixels, as also noted in [13]. Sigmoid ramp up increases the unsupervised weight λ from 0 to 1 over the course of the first 10k iterations, similarly to what was done in, e.g., [34, 22]. This yields results somewhat lower than in our default solution, and exactly the same results as when keeping the unsupervised weight at a constant 1.

We investigate adjusting the unsupervised loss by changing our default cross-entropy loss with a squared error loss. The loss is summed over the class probability dimension and averaged over batch and spatial dimensions, in keeping with [13]. This yields results considerably lower than when using cross-entropy. We also try training with cross-entropy without using pseudo-labels, and instead merely softmax outputs as targets, which also lowers the results. This is likely because entropy minimization, here in the form of pseudo-labelling, helps the network generalize better, as seen in previous works [33, 23]. When not using pseudo-labels we also fail to avoid the problem of "label contamination", described in Figure 2, causing the network to be trained against unreasonable targets near object boundaries.

In our results for Cityscapes, the images are not cropped. Here, however, we investigate randomly cropping both labelled and unlabelled images to 512×512. This increases the performance of both baseline and SSL. The reason for this is likely that cropping adds a regularizing effect. The increase from cropping is larger for the baseline than for SSL, which is believed to be because the SSL solution already receives a regularizing effect from ClassMix, leaving less room for improvement.

Adding color jittering (adjusting brightness, contrast, hue and saturation) and Gaussian blurring also improves the results. These extra augmentations are applied as part of the strong augmentation scheme after ClassMix. This introduces more variation in the data, likely increasing the network's ability to generalize. It also makes the strong augmentation policy more difficult, in line with the use of augmentation anchoring.

Training for 80k iterations instead of 40k improves the results significantly, as can be seen in Table 3. It improves the results more for SSL than for the baseline, which is likely because there is more training data in SSL training, meaning that overfitting is a smaller issue.

5. Conclusion

In this paper we have proposed an algorithm for semi-supervised semantic segmentation that uses ClassMix, a novel data augmentation technique. ClassMix generates augmented images and artificial labels by mixing unlabelled samples together, leveraging on the network's semantic predictions in order to better respect object boundaries. We evaluated the performance of the algorithm on two commonly used datasets, and showed that it improves the state of the art. It is also worth noting that since many semantic segmentation algorithms rely heavily on data augmentations, the ClassMix augmentation strategy may become a useful component also in future methods. Finally, additional motivation for the design choices was presented through an extensive ablation study, where different configurations and training regimes were compared.

Acknowledgements. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *ArXiv*, abs/1911.09785, 2019.

[2] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5050–5060, 2019.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018.

[5] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Semi-supervised learning in video sequences for urban scene segmentation. *arXiv preprint arXiv:2005.10266*, 2020.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223. IEEE Computer Society, 2016.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[8] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 375–391, Cham, 2018. Springer International Publishing.

[9] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1310–1319. IEEE Computer Society, 2017.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[11] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[12] Zhengyang Feng, Qianyu Zhou, Guangliang Cheng, Xin Tan, Jianping Shi, and Lizhuang Ma. Semi-supervised semantic segmentation via dynamic self-training and class-balanced curriculum. *arXiv preprint arXiv:2004.08514*, 2020.

[13] Geoff French, Samuli Laine, Timo Aila, and Michal Mackiewicz. Semi-supervised semantic segmentation needs strong, varied perturbations. In *29th British Machine Vision Conference, BMVC 2020*, 2019.

[14] Geoffrey French, Michal Mackiewicz, and Mark H. Fisher. Self-ensembling for visual domain adaptation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[15] Geoff French, Avital Oliver, and Tim Salimans. Milking cowmask for semi-supervised image classification. *arXiv preprint arXiv:2003.12022*, 2020.

[16] Bharath Hariharan, Pablo Arbelaez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *International Conference on Computer Vision (ICCV)*, 2011.

[17] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, Adam Prügel-Bennett, and Jonathon Hare. Understanding and enhancing mixed sample data augmentation. *arXiv preprint arXiv:2002.12047*, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[19] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 65. BMVA Press, 2018.

[20] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5259–5270, 2019.

[21] Jongmok Kim, Jooyoung Jang, and Hyunwoo Park. Structured consistency loss for semi-supervised semantic segmentation. *ArXiv*, abs/2001.04647, 2020.

[22] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[23] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[24] Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6459–6468. IEEE Computer Society, 2017.

[25] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 63. BMVA Press, 2018.

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[28] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[29] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[30] Christian S Perone and Julien Cohen-Adad. Deep semi-supervised segmentation with weight-averaged consistency targets. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 12–19. Springer, 2018.

[31] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. Ke-gan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5237–5246, 2019.

[32] Tal Remez, Jonathan Huang, and Matthew Brown. Learning to segment via cut-and-paste. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 37–52, 2018.

[33] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[34] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 1195–1204, 2017.

[35] Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M. Rehg, and Visesh Chari. Learning to generate synthetic data via compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[36] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3635–3641. AAAI Press, 2019.

[37] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[38] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6022–6031. IEEE, 2019.

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.