Semantic Segmentation by Semi-Supervised Learning Using Time Series Constraint

Takahiro Mano[a], Sota Kato[b] and Kazuhiro Hotta[c]
Meijo University, 1-501 Shiogamaguchi, Tempaku-ku, Nagoya 468-8502, Japan

Keywords: Semantic Segmentation, Semi-Supervised Learning, Pseudo Label, Time Series Constraint.

Abstract: In this paper, we propose a method to improve the accuracy of semantic segmentation when the number of training data is limited. When time-series information such as video is available, it is expected that images that are close in time-series are similar to each other, and pseudo-labels can be easily assigned to those images with high accuracy. In other words, if the pseudo-labels are assigned to the images in the order of time-series, it is possible to efficiently collect pseudo-labels with high accuracy. As a result, the segmentation accuracy can be improved even when the number of training images is limited. In this paper, we evaluated our method on the CamVid dataset to confirm the effectiveness of the proposed method. We confirmed that the segmentation accuracy of the proposed method is much improved in comparison with the baseline without pseudo-labels.

1 INTRODUCTION

Semantic segmentation is the process of assigning a label to each pixel in an image. In general, image recognition using deep learning requires a large number of supervised images. Semantic segmentation requires pixel-level annotation in all images, and the cost of preparing a large number of supervised images is very high. For example, it is reported that it takes approximately 90 minutes to create one annotated image in the Cityscapes dataset (Cordts et al., 2016).

FCN (Long et al., 2015) which consists of all-layer convolution, SegNet (Badrinarayanan et al., 2017) and U-Net (Ronneberger et al., 2015) with encoder-decoder structures, and Deeplabv3+ (Chen et al., 2018) which is the extension of these methods have been proposed. But, they suffer from low accuracy when those methods are trained with only a small number of supervised images. Against this background, semi-supervised learning method using a small number of supervised and lots of unsupervised images has attracted attention. (Papandreou et al., 2015)

Recently, semi-supervised image segmentation (SSIS) methods (Chen et al., 2021; Ouali et al., 2020) has been proposed to train models with a limited number of labeled and unlabeled images. However, existing SSIS methods do not use a large number of unlabeled images and fail to take advantage of unlabeled images. The method (Chen et al., 2020) used additional teacher signals such as pseudo-labels for unlabeled images in order to improve the accuracy. However, these methods are not designed for video data and do not take into account the good property of time-series data, such that the images at time t and $t+1$ are similar to each other. We consider that time series constraints should be useful to improve the accuracy.

In this paper, we propose to a segmentation method from a small number of annotated images by using time series constraints effectively. When time-series information such as video is available, it is expected that images which are close in time-series are similar to each other, and pseudo-labels on those images are expected to be highly accurate. In other words, if pseudo-labels are assigned to the images in the order of time-series, it is possible to efficiently collect pseudo-labels with high accuracy. As a result, we can improve the segmentation accuracy from a small number of training images.

We conducted experiments on the Camvid dataset. We evaluated the accuracy when we train our method with approximately 1/91, 1/18, and 1/9 of all annotated images. Our method can use the images with pseudo-labels in the order of time-series for training. We confirmed the significant performance improvement by the proposed method. Specifically, the pro-

[a] https://orcid.org/0000-0003-2077-6079
[b] https://orcid.org/0000-0003-0392-6426
[c] https://orcid.org/0000-0002-5675-8713

Mano, T., Kato, S. and Hotta, K.
Semantic Segmentation by Semi-Supervised Learning Using Time Series Constraint.
DOI: 10.5220/0011721800003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 708-714
ISBN: 978-989-758-634-7; ISSN: 2184-4321

posed method improved the mIoU by 9.33%, 5.7%, and 3.86% respectively when we use approximately 1/91, 1/18, and 1/9 of all annotated images.

This paper is organized as follows. Section 2 presents related works. Section 3 explains the proposed method. Section 4 provides an experimental overview, and Section 5 shows the experimental results. Finally, Section 6 describes conclusions and future works.

2 RELATED WORKS

2.1 Semi-Supervised Semantic Segmentation

Semi-supervised learning is an intermediate method between supervised and unsupervised learning. Semi-supervised learning does not require annotation of all images, so we can reduce the annotation cost. Semi-supervised learning makes good use of unlabeled images to improve the accuracy. Specifically, it has been attracting attention because pseudo-labels can be assigned to unlabeled images.

Pseudo-labeling methods (Lee et al., 2013) created pseudo labels by assigning the most probable class to the pixel in an unlabeled image. In semi-supervised learning, pseudo-labeling methods (Chen et al., 2020; Zou et al., 2020) can be used to extend the training data using unlabeled images.

Semi-supervised semantic segmentation methods utilize Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) such as AdvSemiSeg (Hung et al., 2018) and S4GAN (Mittal et al., 2019). They use discriminators to distinguish the confidence maps from predictions for labeled and unlabeled data. Consistency-based methods include perturbations by CutMix (French et al., 2019) and ClassMix (Olsson et al., 2021) to the input image in order to enforce the consistency of predictions and intermediate features.

However, these methods have the problem that they are also not designed for video data and do not use the good property of time-series. Our proposed method uses time-series constraints that images at time t and $t+1$ is similar. By using this constraint effectively, we can predict more accurate pseudo-labels easily for unlabeled images in a video.

2.2 Contrastive Learning

Contrastive Learning (Hadsell et al., 2006) has successfully improved the accuracy for semi-supervised learning. Pixel-by-pixel learning was used in the method (Wang et al., 2021). The pixelwise contrastive loss is defined as

$$\mathcal{L}_i^{NCE} = \frac{1}{|\mathcal{P}_i|} \sum_{i^+ \in \mathcal{P}_i} -\log \frac{\exp(i \cdot i^+ / \tau)}{\exp(i \cdot i^+ / \tau) + \sum_{i^- \in \mathcal{N}_i} \exp(i \cdot i^- / \tau)} \quad (1)$$

where $\mathcal{P}_i$ and $\mathcal{N}_i$ denote pixel embedding collections of the positive samples and negative samples. i is a pixel, i^+ represents positive pixel and i^- represents negative pixel. Positive and negative pixels are learned contrastively not only in the same image but also in two images. The $\cdot$ means the dot product. We set the temperature τ to 0.1.

It is reported that contrastive learning improves the accuracy by learning rich semantic relationships between pixels across different images. In this paper, we also use contrastive learning between pixels to predict pseudo-labels with high accuracy because the model can learn semantic relationships between pixels across different images. In this study, after shuffling the annotated images and unlabeled images with pseudo-labels, pixel-wise contrast learning is performed between images in a mini-batch. In experiments, our method is trained by using both cross entropy loss and pixel-wise contrastive loss.

3 PROPOSED METHOD

In this section, We explain the details of the proposed method. section 3.1 describes how to create and learn pseudo-labels using time-series constraints in only one scene. Section 3.2 describes how to learn and create pseudo-labels in each scene. Section 3.3 explains the threshold when we predict pseudo-labels of unlabeled images in a video. Section 3.4 describes the loss function used in our method.

3.1 Semi-Supervised Learning with Time-Series Constraints

In the proposed method, the model is trained on some annotated images in the middle of a scene as shown in Figure 1. This is because we would like to use the time-series constraints effectively. In top row of Figure 1, five annotated images are used for training the model. The trained model is applied to two unlabeled images that are the closest in time-series to the annotated images because images that are chronologically adjacent to the annotated images are expected to be similar to the annotated images. Thus, we can assign highly accurate pseudo-labels to two images.

As shown in the bottom part of Figure 1, train-

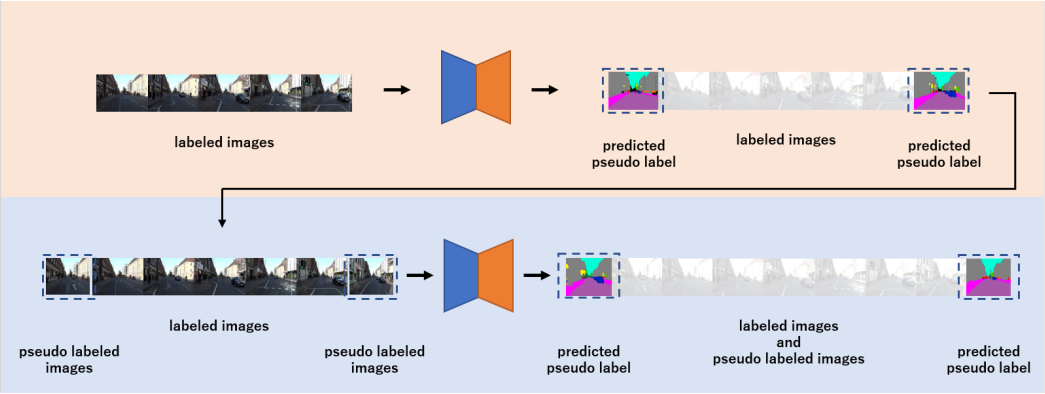

Figure 1: Creating pseudo labels using time-series constraints.

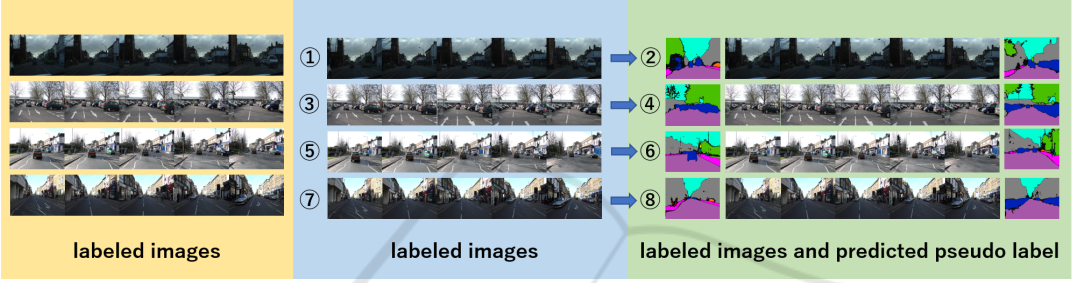

Figure 2: Scene-wise learning and prediction of pseudo labels.

ing is performed again by using the original annotated images and two pseudo-labeled images. The model is also applied to two unlabeled images that are the chronologically closest among the images that have not yet been assigned pseudo-labels. Then we assign pseudo-labels to two images. By repeating this process in the order of time-series, the accuracy of the proposed method is gradually improved.

3.2 Scene-Wise Learning and Pseudo Label Prediction

In general, in-vehicle video dataset such as CamVid dataset consists of some scenes. If we train a model using annotated images and pseudo-labeled images in all scenes simultaneously, we may not use time-series constraint effectively because the model is not specialized to one scene. Therefore, in this paper, we propose to learn and predict pseudo-labels for each scene as shown in Figure 2.

At first, the proposed method is trained using some annotated images in all scenes. In the Figure 2, there are four scenes. ① Next, we select one scene and the model is fine-tuning to the scene. ② Then, we predict the pseudo-labels of two images that are chronologically adjacent to the annotated images in the selected scene. ③ We pick up the different one scene from the first scene, and fine-tuning is performed. ④ We also perform the pseudo-label assignment. By fine-tuning the model for each scene, we can use time-series constraint effectively and the model can predict highly accurate pseudo-labels.

This process is continued for all scenes in the numbered order shown in Figure 2, and while increasing the number of pseudo-labeled images. Finally, the model is trained on the annotated images and all pseudo-labeled images obtained at the upper process.

3.3 Threshold

As described in previous section, pseudo-labels are assigned to unlabeled images using the time-series constraint. But, of course, all predictions are not correct. Thus, a threshold is needed to determine whether we should assign pseudo-label to each pixel based on the confidence through softmax function.

If the threshold is high, only the pixels that the model classifies with high confidence are assigned to pseudo-labels. However, if the threshold is too high, only a smaller number of pixels in an image can be pseudo-labeled.

On the other hand, if the threshold is low, we must assign pseudo-labels to the pixels with low confidence. But we can increase the number of pixels for training. Therefore, the setting of threshold is important for both accurate pseudo-labels and the num-

ber of pixels for training. In this paper, we tried two threshold values; a fixed value and an Average Predicted Value(APV) for each class.

Fixed Value

In general, a fixed value is used as threshold. This means that only pixels whose confidence exceeds the fixed threshold are assigned pseudo-labels. In this paper, the threshold for each class was empirically set.

Average Predicted Value

We propose a threshold value called Average Predicted Value (APV). When fixed values are used, it is difficult to assign pseudo-labels to pixels with low confidence, such as the rare class. To compensate for this, we use the average confidence of outputs of each class in an image as the threshold value. This allows approximately half number of pixels classified in each class to be assigned pseudo-labels. Since the distribution of the confidence levels varies for each image, we can change the threshold automatically for each image.

3.4 Loss

In this study, the model is trained with the following losses by using both annotated images and images with pseudo-labels. The loss for the proposed method can be formulated as follows.

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_i^{NCE} \quad (2)$$

where $\mathcal{L}_{ce}$ is cross entropy loss. $\mathcal{L}_i^{NCE}$ is pixel wise contrastive loss showin in equation (1). λ is a hyperparameter and is empirically set to 0.1.

4 EXPERIMENTAL SETTINGS

In this section, we explain the experimental settings. The dataset and training method are described in Section 4.1. Experimental conditions are described in Section 4.2. We explain the baseline method in Section 4.3.

4.1 Data Sets and Training Methods

The CamVid dataset (Brostow et al., 2009) used in the following experiment includes the images of 360 x 480 pixels and 11 classes. CamVid dataset consists of 367 training images, 101 validation images, and 233 test images. The original 367 training images consisted of 5 scenes, but the fifth scene was not used because the number of images in the scene was too small to increase the number of pseudo-labeled images sufficiently. Therefore, experiments were conducted using the remaining four scenes. In this paper, we used 1, 5, and 10 images from each of the four scenes for training as annotated images. Remaining images were used as the unlabeled images for pseudo labels. The 1, 5, and 10 annotated images were the center images in the order of each time-series scene.

We evaluate whether the proposed method improves the test mIoU. The training images, including the pseudo-labeled images, were randomly flipped left and right as data augmentation.

4.2 Network and Parameter

We used Deeplabv3+ with Resnet101 pre-trained on Imagenet dataset. The batchsize was set to 4, and SGD (Momentum=0.9, $weight_d = 1 \times 10^{-4}$) was used as the optimization method.

At first, the model is trained on the annotated images in four scenes for 200 iterations. Next, the trained model is fine-tuned to the specific scene for 200 iterations. The fine-tuned model is applied to two unlabeled images in the scene that are chronologically adjacent to the annotated images, and the pseudo-labels are assigned to the unlabeled images by using threshold. After we assign pseudo-labels to two unlabeled images in all scenes, the model is trained again using the images in all scenes for 200 iterations. The images with pseudo-labels are used for training in the next epoch. The process up to this point is defined as 1 epoch. In the second epoch, the model is trained for 200 iterations using the annotated images and pseudo-labeled images. This is repeated until all unlabeled images in scenes are used.

The initial learning rate was set to 0.001 for the backbone network parameters and 0.01 for the classifier parameters. Then, the learning rate was changed using the following equation. The learning rate was attenuated from the initial value at each iteration using Equation (3).

$$lr = lr_{base} \times (1 - iteration/200)^{0.9} \quad (3)$$

We used both cross entropy loss and pixelwise contrastive loss as the loss function. Mean IoU (mIoU) is used as evaluation measure.

4.3 Baseline Method

The baseline method is to train the same model as the proposed method with the same annotated images obtained from four scenes. The best model with the highest mIoU on validation set in 200 iterations is selected as the baseline model. In this experiment, the

Table 1: Comparison of the proposed method and baseline on Camvid test dataset.

	4/367		20/367		40/367	
class	baseline	ours	baseline	ours	baseline	ours
sky	85.73	89.75 (+4.02)	89.57	89.76 (+0.19)	90.18	90.62 (+0.44)
building	67.91	73.10 (+5.19)	72.42	73.84 (+1.42)	74.33	75.15 (+0.82)
pole	4.08	12.67 (+8.59)	13.37	14.45 (+1.08)	14.23	16.65 (+2.42)
road	71.91	78.88 (+6.97)	78.53	87.67 (+9.14)	84.27	91.49 (+7.22)
sidewalk	16.15	39.60 (+23.45)	35.52	61.77 (+26.25)	53.67	73.71 (+20.04)
tree	55.17	67.38 (+12.21)	63.44	65.71 (+2.27)	65.55	67.67 (+2.12)
signal	1.11	14.35 (+13.24)	13.57	13.00 (-0.57)	31.35	30.56 (-0.79)
fence	0.00	0.16 (+0.16)	1.69	5.59 (+3.9)	8.95	13.66 (+4.71)
car	48.30	60.22 (+11.92)	62.31	73.17 (+10.86)	70.58	75.90 (+5.32)
pedestrian	3.10	19.94 (+16.84)	24.15	32.78 (+8.63)	36.56	42.14 (+5.58)
bicyclist	0.00	0.00 (+0.00)	0.01	0.00 (-0.01)	7.74	2.29 (-5.45)
mIoU	32.13	41.46 (+9.33)	41.33	47.07 (+5.74)	48.86	52.71 (+3.85)

effectiveness of the proposed method is demonstrated by comparing the test mIoU of the baseline method with the test mIoU of the proposed method when the number of pseudo-labeled images is increased.

5 EXPERIMENTAL RESULTS

We conducted an experiment to evaluate semi-supervised semantic segmentation on the CamVid dataset. Section 5.1 shows the accuracy and qualitative comparison of the baseline and the proposed method. Section 5.2 shows the results of ablation studies.

5.1 Comparison of the Proposed Method with Baseline

Table 1 shows the IoU of each class and mIoU by baseline and the proposed method. The column of baseline shows the IoU of each class by baseline. The column of ours show the IoU and improvement compared with baseline. 4/367, 20/367, and 40/367 indicate the number of annotated images actually used for training. For example, 4/367 means that we use only one annotated image in one of four scenes. 20/367 means that we use 5 annotated images in each scene.

By using the proposed method, we were able to significantly improve the mIoU compared to the baseline. When we used only one annotated image in each scene, the mIoU was improved 9.33%. The mIoU of our method at this setting was better than the mIoU of baseline at the setting that 5 annotated images per scene are used for training. The IoUs of sidewalk, tree, signal, and car are particularly improved. The classification of Pedestrian and Bicyclist is difficult because both Pedestrian and Bicyclist are very small and similar. In experiments, many Bicyclist was misclassified as Pedestrian. Thus, Bicyclist class had low accuracy and Pedestrian class was much improved.

Although our method improved the accuracy in all settings, the accuracy was significantly improved with fewer annotated images. Since the goal of semi-supervised learning is to learn with as fewer supervision as possible, the effectiveness of our proposed method is demonstrated.

Figure 3 shows the qualitative segmentation results of the baseline and the proposed method when we used 40 annotated images. Our proposed method generates more accurate segmentation results than the baseline. We see that sidewalks are predicted more accurately using the proposed method than baselines. In addition, there are many parts of the baseline where the pedestrian is not predicted, but the proposed method is able to predict the pedestrian more accurately. These results demonstrated the effectiveness of the proposed method.

5.2 Ablation Studies

Table 2: Ablation study on the loss function. We show the improvement from baseline in test mIoU by the proposed method when we used 20 annotated images.

$\mathcal{L}_{ce}$	$\mathcal{L}_i^{NCE}$	mIoU
✓		+2.27
✓	✓	+5.74

Table 3: Comparison of threshold when we used 20 annotated images.

threshold	0.2	0.4	0.6	0.8	APV
mIoU	43.25	43.33	43.99	44.77	47.07

Table 2 shows ablation study about the loss functions. Table shows the improvement from baseline in test

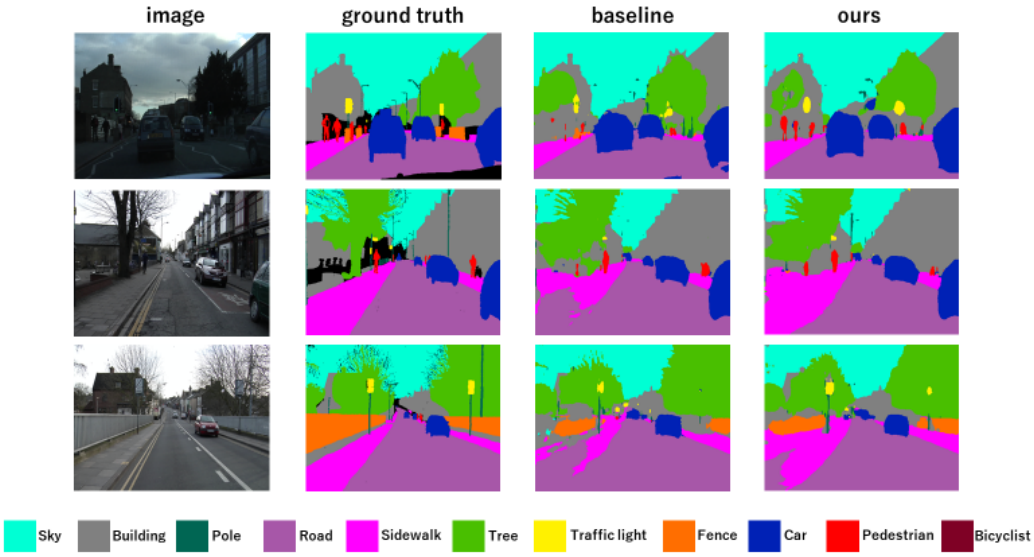

Figure 3: Comparison of qualitative results by the baseline and the proposed method when we used 40 labeled images for training.

mIoU by the proposed method when we used 20 annotated images in four scenes. We see that contrasitive loss improved the accuracy significantly.

Table 3 shows a comparison of the threshold when we use 20 annotated images. We see that the mIoU increases significantly when the Average Predicted Value(APV) of each class is used as the threshold value. If a fixed value is used for the threshold value, the accuracy is biased toward the classes with large training samples, because the classes with small training samples have a smaller probability than classes with large training samples. However, by using average predict value as the threshold value, we can use different threshold value for each class, and this improved the accuracy.

6 CONCLUSIONS

We proposed a semi-supervised semantic segmentation method that assigns pseudo-labels in chronological order and trains the model using those images step by step. We confirmed that our proposed method much improved test mIoU on the Camvid dataset in comparison with the baseline model.

Our proposed method used only the network's output to assign pseudo-labels. However, prior probability, which is calculated from past images in a time-series manner, may improve the accuracy of pseudo-labels. The usage of prior probability is one of the subject for future works.

ACKNOWLEDGEMENTS

This work is supported by JSPS KAKENHI Grant Number 21K11971.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495.

Brostow, G. J., Fauqueur, J., and Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97.

Chen, L.-C., Lopes, R. G., Cheng, B., Collins, M. D., Cubuk, E. D., Zoph, B., Adam, H., and Shlens, J. (2020). Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *European Conference on Computer Vision*, pages 695–714. Springer.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818.

Chen, X., Yuan, Y., Zeng, G., and Wang, J. (2021). Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223.

French, G., Laine, S., Aila, T., Mackiewicz, M., and Finlayson, G. (2019). Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Hung, W.-C., Tsai, Y.-H., Liou, Y.-T., Lin, Y.-Y., and Yang, M.-H. (2018). Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*.

Lee, D.-H. et al. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Mittal, S., Tatarchenko, M., and Brox, T. (2019). Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1369–1379.

Olsson, V., Tranheden, W., Pinto, J., and Svensson, L. (2021). Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378.

Ouali, Y., Hudelot, C., and Tami, M. (2020). Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684.

Papandreou, G., Chen, L.-C., Murphy, K. P., and Yuille, A. L. (2015). Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., and Van Gool, L. (2021). Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313.

Zou, Y., Zhang, Z., Zhang, H., Li, C.-L., Bian, X., Huang, J.-B., and Pfister, T. (2020). Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*.