UDC: 004.93 https://doi.org/10.15407/jai2024.04.133

O. Shabo[1], N. Shapoval[2]
[1,2]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
37 Beresteysky Avenue, Kyiv, 03056
[1]andriyshabo@gmail.com
[2]shovgun@gmail.com
[1]https://orcid.org/0009-0008-2661-4752
[2]https://orcid.org/0000-0002-8509-6886

SEMI-SUPERVISED LEARNING OF A VISION TRANSFORMER FOR THE TASK OF ROAD TRAFFIC SEGMENTATION IN AN UNSTRUCTURED ENVIRONMENT

Abstract. In the last few years, traditionally used for natural language processing tasks, recurrent neural networks have been replaced mainly by transformers. Thanks to the novel attention mechanism, they also sequentially receive text input but provide much better results than LSTM, GRU-based, or similar networks. Self-attention negates the problem of fading memory by allowing efficient evaluation of dependencies between distant tokens and provides a better means for parallelization for modern processing units like GPU. Until recently, the use of transformers for computer vision (CV) tasks was minimal. The biggest obstacles that hindered the progress in this field were immense computational complexity, the fact that the image is a grid, not a sequence-like text, and the lack of strong inductive bias, in other words, the ability to have a good grasp of local correlations, unlike their CNN counterparts. The latest slowed down the vision transformer (ViT) usage rate in semantic segmentation (SS) even more. However, it was recently shown that with sufficient data, Transformers could outperform CNN-based networks in image classification and, with the proper ViT structure, even in SS. A promising direction for providing a ViT with required training data is using semi-supervised learning (SSL), which allows for extracting helpful information from unlabeled data using only a small amount of labeled data. This approach is beneficial when solving the problem of SS since manually creating masks for images is very time-consuming. This paper proposes the robust semi-supervised ViT learning method using minimal labeled data. The combination of a strong augmentation pipeline and a dual teacher paradigm allows good performance for SS of road traffic in the unstructured environment without the need for extensive hyperparameter search.

Keywords: semi-supervised learning, vision transformer, semantic segmentation, unstructured environment.

Introduction

There are two main problems regarding ViT training and usage: large data requirements and significant computation expenses. The first problem was partly solved by self-supervised learning or supervised pretraining on large datasets like ImageNet and further finetuning the model on the necessary dataset for the required task [1]. However, it is still not straightforward how to use the unlabeled data from the domain of interest to boost the performance of the ViT further [2]. The second problem occurs mainly because computational complexity for self-attention increases quadratically with the number of tokens due to softmax [3].

Here, we show that it is possible to obtain diverse data using the pipeline of strong augmentations to obtain stable results for SS of the road traffic using ViT. We show that the proposed method is efficient in real-life unstructured road environments. The study is conducted using the EfficientViT [4] family transformer because, in actual conditions, there is often no access to powerful GPUs or similar computational units.

Statement of the problem

SS is a field of CV that has been extensively studied for almost two decades. The task is to classify each pixel without considering their belonging to a separate object. This problem is vital in environmental analysis and autonomous navigation [5]. It allows the model to perceive a 3D environment through a prism of the 2D images, as humans do. SSL compensates for the lack of labeled data, which is usually present in small quantities due to privacy, license, or ethical reasons. It uses unlabeled data, which can be obtained in abundance, to boost a model's performance further. Such a situation is more applicable to real-life case scenarios, particularly SS, since labeling here is rigorous and time-consuming [6].

One of the most popular SS datasets for urban traffic understanding is Cityscapes [7], which contains rich annotated data. However, this benchmark primarily consists of images with well-defined roads, pedestrian lanes, and ordered traffic. We evaluate the performance of the SSL method in our paper using the Indian Driving Dataset (IDD) [8] instead. The images in this dataset were collected from an unstructured road traffic environment in the Indian cities of Hyderabad and Bangalore and their suburbs. Unlike Cityscapes, there are mostly no well-defined car or pedestrian lanes; the traffic is denser and contains more diverse participants like motorcycles, auto rickshaws, or even cows. Using such data allows us to train and test a model in more real-life conditions, including riskier situations, making the model more robust to deviations from the standard environment.

Related work

SSL usage in ViTs is an underresearched problem, with relatively few papers published in the last few years. This is partly because ViTs are novel models that started to achieve better results than their CNN counterparts only recently and partly due to their weak inductive bias and high training data requirement [2]. In [9], instead of focusing on some SSL algorithm development, authors present novel automatic data selection that allows the formation of a small labeled set with samples diverse enough to represent the whole training dataset qualitatively. It also selects samples with a high degree of uncertainty to help the model learn harder off-road environment samples.

S4Former in [10] utilizes three novel modifications to introduce regularization into image, feature, and output aspects. For image regularization, PatchShuffle is used, an augmentation technique explicitly designed for the self-attention of ViT transformers. To perform feature regularization, a Patch-Adaptive Self-Attention, which dynamically increases value in an area with less confident predictions, is presented in the paper. Finally, a Negative Class Ranking Loss is suggested to preserve the distribution of negative classes in the model's output.

As has already been stated, the strengths of CNN and ViT are opposite and complementary: strong inductive bias, local awareness in CNN, and good global and context understanding in ViT. As a result, novel ViT architectures often incorporate CNN layers into their topology, frequently for high-dense prediction tasks thanks to their properties [1]. However, in [11], authors propose to decouple CNN and ViT architectures into two branches of Mean Teacher instead of training them simultaneously. Intra-model Local-Global Interaction is achieved thanks to the interaction of these two branches in the Fourier space. On the other hand, inter-model Class-wise Consistency is calculated as the difference between graphs of different branches based on patch and class center correlations.

The purpose of the study

Most of the SS experiments conducted nowadays are done under stable conditions. This study aims to show that a training pipeline based on the dual teacher paradigm can also effectively tackle the challenges of the unstructured road traffic environment in a frequent situation when only a small part of the data is labeled. To bring the study closer to realistic conditions, we utilize a resource-efficient transformer model with a few parameters that is more likely to be used in such applications. We study approaches to strong augmentation to achieve the best performance while dealing with training data scarcity.

Method

In the field of semi-supervised SS, the current prevalent approaches are student-teacher-based methods. To achieve stable results, we propose to apply the EMA-Teacher pipeline. However, as we found out, using this approach solely does not yield good results in the unstructured environment because of the widespread model collapse issue, when after some amount of training epochs, student and teacher models begin to make very similar predictions. To mitigate this issue, we incorporate an approach based on dual-teacher [12] instead. This technique utilizes two teachers that are switched each epoch, together with two strong augmentations – CutMix [13] and ClassMix [14]. When training models on datasets such as Cityscapes, it is not necessary to use strong augmentation since the general location of objects of a particular class is quite stable across images. The relatively small number of subclasses in each of the main classes and their nature in a structured environment allows efficient model training on the data in a reasonable amount. However, when working with a dataset in an unstructured environment, moreover, in the presence of a small amount of labeled data, it is essential to use strong augmentation so that the model can better distinguish different angles, positions,and variations of objects since the number of possible situations on the road and by the side in such a dataset increases significantly. To address this problem, we conduct experiments with different adaptations of CutMix and ClassMix to a semi-supervised SS. Performing augmentations when solving the SS task is complicated because, unlike image classification, where image augmentation does not affect the expected output of the model, affine and some other augmentations also impact the image's segmentation mask. To tackle the issues mentioned earlier, we propose a pipeline of the semi-supervised finetune depicted in Fig. 1. The CutMix augmentation is defined using formulae

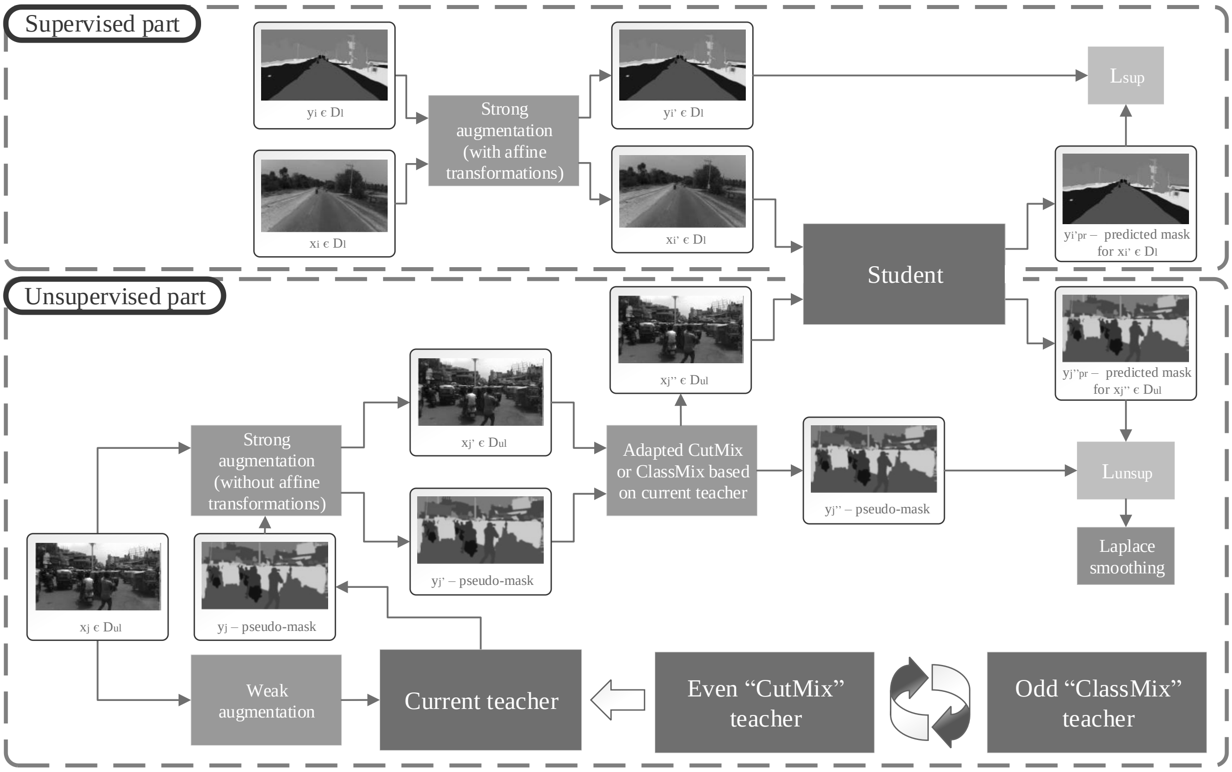

Fig. 1. General pipeline of the proposed SSL algorithm

(1)-(7) and consists of modifying each image from a batch by cropping a random rectangle from a respective image from the shuffled batch into the same position on the first image:

$$\lambda_i \sim \mathfrak{B}(4,4), \quad (1)$$

$$r_i^x \sim \mathcal{U}(0, W), \quad (2)$$

$$r_i^y \sim \mathcal{U}(0, H), \quad (3)$$

$$r_i^w = W\sqrt{1-\lambda_i}, \quad (4)$$

$$r_i^h = H\sqrt{1-\lambda_i}, \quad (5)$$

$$X_i = (1-\mathcal{B})\odot X_i + \mathcal{B}\odot X_{\sigma(I)_i}, \quad (6)$$

$$Y_i = (1-\mathcal{B})\odot Y_i + \mathcal{B}\odot Y_{\sigma(I)_i}, \quad (7)$$

where $\mathfrak{B}$ – is the beta distribution, H and W – are the height and width of an image, r_i^x and

r_i^y – are coordinates of the upper left cropped rectangle $\mathcal{B}$ corner, r_i^w and r_i^h – are the width and height of the $\mathcal{B}$, X_i and Y_i – are the ith image and its respective mask, '$\odot$' – is the pixel-wise multiplication, $\sigma(I)$ – is a random permutation of the index set $I = \{i \in N, 1 \leq i \leq \mathcal{N}\}$ and $\mathcal{N}$ is the batch size.

Proposed ProbCutMix modification uses formulae (8)-(10) instead of (1):

$$c_j^i = \max_{c=\overline{1,N_c}} p_c^{i,j}, \qquad (8)$$

$$\bar{c}_i = \frac{1}{HW} \sum_{j=1}^{HW} c_j^i, \qquad (9)$$

$$\lambda_i = \frac{\bar{c}_i}{\bar{c}_i + \bar{c}_{\sigma(I)_i}}, \qquad (10)$$

where $p_c^{i,j}$ – is the predicted probability of ith image jth pixel belonging to class c, N_c – is the total number of classes, c_j^i – is the model confidence in ith image jth pixel prediction.

As a result, the more confident the teacher is in occluding pseudo-mask, and the less confident the teacher is in occluded pseudo-mask – the bigger the cropped rectangle will be. This should result in a larger amount of high-confidence masks a teacher generates.

The ClassMix augmentation idea is to modify each image in a batch by pasting half of the classes chosen randomly from the respective image in the shuffled batch. In this study, we test the efficiency of the subsequent suggested modifications. In ObjectClassMix, we select half of the classes from the present in the image, excluding the 'road' and 'sky' classes. The essence of this approach is the desire to reduce the frequency of road and sky appearance, which takes up much space and blocks smaller and more challenging objects. MaxClassMix, where we select half of the classes with the highest class confidence defined by formulae (11)-(12) instead of doing it randomly, and ProbClassMix, where instead of equal probabilities of class selection, formulae (8), (11)-(13) define the probability, with both approaches generating augmented images that correspond to pseudo-masks with higher confidence:

$$\mathcal{C}_k^i = \left\{ j \,\middle|\, \underset{c=\overrightarrow{1,N_c}}{argmax}\, p_c^{i,j} = k \right\}, \qquad (11)$$

$$\bar{c}_k^i = \frac{1}{|\mathcal{C}_k^i|} \sum_{j \in \mathcal{C}_k^i} c_j^i, \qquad (12)$$

$$p_k^i = \frac{\bar{c}_k^i}{\sum_{k=1}^{N_c} \bar{c}_k^i}, \qquad (13)$$

where j – is a pixel, $\bar{c}_k^i$ – is the class k mean confidence in an ith image, p_k^i – is the probability of selecting class k in an ith image.

In this work, EfficientViT was chosen to assess the effectiveness of the SSL for road traffic SS in the unstructured environment for resource-efficient ViTs. EfficientViT family for high-resolution dense prediction significantly lowers computational complexity by introducing multi-scale linear attention instead of softmax and utilizing MBConv blocks [15] with small kernels. It also solves the problem of weak inductive bias in ViT models by using DWConv in its main EfficientViT Module with an attention mechanism. These factors allow the model to achieve SOTA performances on well-known datasets like Cityscapes and ADE20K while maintaining high throughput, thus making it a perfect candidate for autonomous driving solutions.

Another issue is that the teachers provide low-quality pseudo masks at the start of training, which hampers the training significantly due to the nature of ViT. This problem can be partially solved by gradually increasing the unsupervised loss coefficient starting from small values. However, we suggest performing supervised-only training first and then performing semi-supervised finetune instead, where the student and teachers use previously obtained weights. Such supervised pretraining is comparable to the HASSOD approach [16], where instead of using a supervised pre-trained teacher to obtain pseudo-masks, they are generated using feature similarity of patches in the backbone at first. Teacher model predictions replace initial masks in HASSOD only after a certain number of epochs.

Since ViT requires a large amount of diverse data and we have only a small amount of labeled data, we need to use a strong pool of augmentations, in particular, to achieve solid results from the supervised training part. We suggest using RandomResizeCrop with slight variations in scale and ratio, RandomHorizontalFlip, and RandAug [17] to boost the model generalization ability further. We noticed that the model quickly overfits when using a random ratio interval that differs quite a bit from the original. By sampling random augmentations from a predefined pool with random application magnitude, RandAug allows us to obtain significantly different results from the same input image each time. Since RandAug is mainly used only for classification tasks, to use it for segmentation, it should be adapted by applying the same augmentations to the mask as well as to the image, in case the transformation is affine. This set of augmentations refers to the "Strong augmentations" block in Fig. 1. As a "Weak augmentations" block, RandomHorizontalFlip and ColorJitter are used.

Different pixels should be handled in distinct ways during the computation of the unsupervised part of the loss. That is because pseudo masks are generated by the teacher model, which does not match the ground truth. As a result, the influence of pixels incorrectly classified by the teacher on the loss function prevents the model from pulling meaningful information from the correct pixels. It significantly distorts the distribution of objects in the image, which does not allow the model to improve performance when unlabeled data is added to the training. The popular classification task approach is filtering out pixels with a confidence lower than a certain threshold [18, 19]. However, this does not yield good results in SS because the threshold can vary significantly from batch to batch, and pixels that pass it can still contribute differently to loss. As a result, to weigh each pixel loss, we propose using Laplace smoothing [20]. This method consists of filtering out outliers of the q*th* quantile and multiplying the remaining ones by weights defined by formulae (8) and (14):

$$w_j^i = \frac{\left(c_j^i+\varepsilon\right)\mathbb{I}\left\{c_j^i>\tau\right\}}{\sum_{i=1}^{\mathcal{N}}\sum_{j=1}^{HW}\left(c_j^i+\varepsilon\right)\mathbb{I}\left\{c_j^i>\tau\right\}} \cdot \\ \cdot \sum_{i=1}^{\mathcal{N}}\sum_{j=1}^{HW}\mathbb{I}\left\{c_j^i>\tau\right\}, \qquad (14)$$

where i and j are image and pixel indices, respectively, $\mathbb{I}\{\cdot\}$ – is the boolean indicator, τ – is the q-th quantile of $\left\{c_j^i\right\}_{i=\overline{1,\mathcal{N}},\, j=\overline{1,HW}}$, ε – is the stability coefficient. We use $\varepsilon = 1$ and $q = 0.2$.

Table 1. Augmentation combinations results

	Supervised pretrain	Model 1	Model 2	Model 3	Model 4	Model 5
CutMix type	-	Ordinary	Prob	Prob	Prob	Prob
ClassMix type	-	Ordinary	Ordinary	Object	Prob	Max
mIoU	65.94	68.16	68.84	68.58	68.49	68.18
Dice score	77.05	78.8	79.39	79.22	79.14	78.88
Accuracy	76.71	77.9	78.85	78.57	78.46	78.85
Precision	77.47	79.98	80.2	80.12	80.07	79.08
mIoU gain	-	2.22	2.9	2.64	2.55	2.24

Table 2. Effectiveness of supervised pretrain

	Supervised pre-train	Semi-supervised only	Semi-supervised with supervised pre-train
mIoU	69.37	70.59	71.88
Dice	79.72	80.76	81.94
Accuracy	79.01	80.13	81.61
Precision	80.63	81.78	82.44
mIoU gain	-	1.22	2.51

Experiment

The training uses AdamW optimizer [21] and CosineDecay learning rate scheduler with a warmup of 10% steps using PyTorch. EfficientViT B1 weights from training on Cityscapes are utilized as a pre-trained backbone, and we initialize the head with random weights. We use the proposed IDD hierarchy with seven classes and 1/8 labeled images partition: 373 images as a labeled training set, 2603 for an unlabeled training set, 700 for validation, and 1048 for testing. The data is taken from IDD part 2, and we perform a center crop 960x1920 to maintain a 1:2 aspect ratio. We conduct a supervised pre-train using only the supervised part of the pipeline from Fig. 1 and supervised cross-entropy loss defined by the formula (15) and perform a semi-supervised finetune utilizing combined loss defined by formulae (15)-(17), and the whole pipeline from Fig. 1:

$$\mathcal{L}_{sup} = -\frac{1}{B_l}\sum_{i=1}^{B_l}\frac{1}{HW}\sum_{j=1}^{HW}\mathbb{1}\{y_j^i \neq$$

$$\neq ignore\_class\} log p_{y_j^i}^{i,j}, \qquad (15)$$

$$\mathcal{L}_{unsup} = -\frac{1}{B_{ul}}\sum_{i=1}^{B_{ul}}($$

$$\frac{1}{HW}\sum_{j=1}^{HW} log p_{y_j^i}^{i,j}), \qquad (16)$$

$$\mathcal{L}_{comb} = \mathcal{L}_{sup} + \mu\mathcal{L}_{unsup}, \qquad (17)$$

where B_l – is the labeled batch size, B_{ul} – is the unlabeled batch size, $p_{y_j^i}^{i,j}$ – is the predicted probability of the correct class y_j^i for a jth pixel in an ith sample in a batch taken from labeled and unlabeled batch for L_{sup} and L_{unsup}, respectively, ignore_class – is the class which marks pixels, that should not contribute to loss, for example after affine transformation, μ – is the unsupervised loss weight. We use $\mu = 5$.

We train models in all experiments for 80 epochs, with mIoU chosen as the primary metric for evaluation. Supervised pretraining was performed with batch size 16, backbone, and head learning rate 1e-4 and 1e-3, respectively. Semi-supervised finetune is performed with supervised batch size 2, unsupervised batch size 8, and backbone and head learning rate 5e-4 and 2e-3, respectively, with an EMA decay coefficient of 0.99. A lower learning rate is used for the backbone since it already has the initial ability to extract useful features. The results of different CutMix and ClassMix combinations are presented in Table 1, where we utilize 416x832 image size for training/inference. As we can see, the ProbCutMix and ClassMix combination have proven to provide the best results, so we use it to test the efficiency of proposed supervised pretraining usage before the semi-supervised finetune instead of using SSL solely. Since the teacher does not generate good-quality pseudo-masks at the start when not using supervised pre-train, we linearly increase the coefficient behind unsupervised loss from 0 to max throughout all training. We lower the batch size to 10 for supervised pretrain. Here, we used 608x1216 image size for training/inference, with results presented in Table 2. As we can see, our approach provides better results than using labeled data only, with supervised pre-train increasing mIoU value compared to using SSL pipeline only. Examples of segmentation are shown in the Fig. 2:

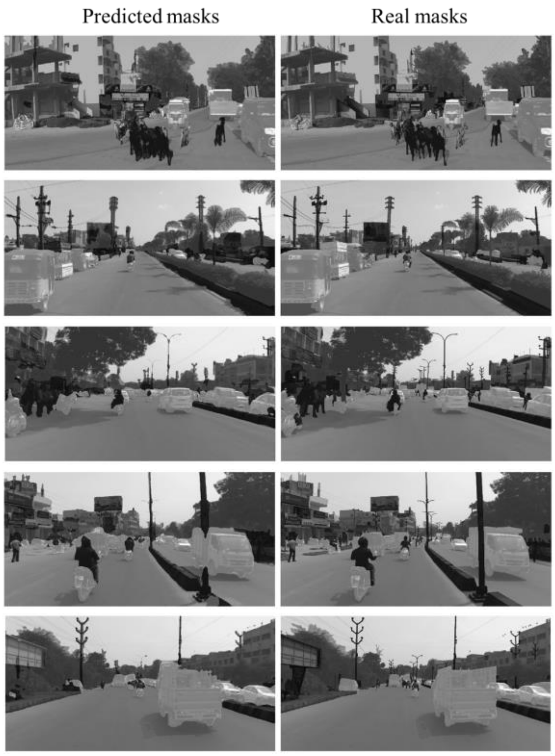

Fig. 2. Segmentation results

Conclusions

In this paper, we have studied the efficiency of the proposed framework for ViT semi-supervised road traffic SS. We have shown that utilizing the suggested ProbCutMix and standard ClassMix delivers the best results. This shows that it is much more essential for CutMix than ClassMix to consider the confidence of the teacher's prediction. Using a strong pipeline of augmentations, we can tackle the lack of training data and its high requirements for ViT. Although the augmentations we use for images deform it quite strongly, this approach was practical due to the unstructured nature of the environment itself. The proposed pipeline can achieve stable results without extensive hyperparameter search thanks to supervised pre-train and Laplace smoothing. As a further study, the suggested SSL method may be included in a resource-efficient autonomous driving system and adapted for mobile or edge devices.

References

1. Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. *ACM Computing Surveys*. https://doi.org/10.1145/3505244
2. Cai, Z., Ravichandran, A., Favaro, P., Wang, M., Modolo, D., Bhotika, R., Tu, Z., & Soatto, S. (2022). *Semi-supervised Vision Transformers at Scale*. arXiv. https://arxiv.org/pdf/2208.05688
3. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. https://doi.org/10.1109/iccv48922.2021.00986
4. Cai, H., Li, J., Hu, M., Gan, C., & Han, S. (2023). EfficientViT: Lightweight Multi-Scale Attention for High-Resolution Dense Prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. https://doi.org/10.1109/iccv51070.2023.01587
5. Csurka, G., Volpi, R., & Chidlovskii, B. (2022). Semantic Image Segmentation: Two Decades of Research. *Foundations and Trends® in Computer Graphics and Vision*, *14*(1-2), 1–162. https://doi.org/10.1561/0600000095
6. Chen, Y., Mancini, M., Zhu, X., & Akata, Z. (2022). Semi-Supervised and Unsupervised Deep Visual Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–23. https://doi.org/10.1109/tpami.2022.3201576
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2016.350
8. Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., & Jawahar, C. V. (2019). IDD: A Dataset for Exploring Problems of Autonomous Navigation in Unconstrained Environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. https://doi.org/10.1109/wacv.2019.00190
9. Singh, A., Singh, K., & Sujit, P. (2021). *OffRoadTranSeg: Semi-Supervised Segmentation using Transformers on OffRoad environments*. arXiv. https://arxiv.org/pdf/2106.13963
10. Hu, X., Jiang, L., & Schiele, B. (2024). Training Vision Transformers for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4007–4017). https://openaccess.thecvf.com/content/CVPR2024/papers/Hu_Training_Vision_Transformers_for_Semi-Supervised_Semantic_Segmentation_CVPR_2024_paper.pdf
11. Huang, H., Xie, S., Lin, L., Tong, R., Chen, Y.-W., Li, Y., Wang, H., Huang, Y., & Zheng, Y. (2023). SemiCVT: Semi-Supervised Convolutional Vision Transformer for Semantic Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr52729.2023.01091
12. Na, J., Ha, J.-W., & Chang, H. J. (2023). Switching Temporary Teachers for Semi-Supervised Semantic Segmentation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine (Ed.), *Advances in Neural Information Processing Systems* (D. Han & W. Hwang, Corresponding author; Vol. 36, pp. 40367–40380). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/7eeb42802d3750ca59e8a0523068e9e6-Paper-Conference.pdf
13. Yun, S., Han, D., Chun, S., Oh, S. J., Yoo, Y., & Choe, J. (2019). CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE. https://doi.org/10.1109/iccv.2019.00612
14. Olsson, V., Tranheden, W., Pinto, J., & Svensson, L. (2021). ClassMix: Segmentation-Based Data Augmentation for Semi-Supervised Learning. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. https://doi.org/10.1109/wacv48630.2021.00141
15. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2018.00474

16. Cao, S., Joshi, D., Gui, L., & Wang, Y.-X. (2023). HASSOD: Hierarchical Adaptive Self-Supervised Object Detection. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine (Ed.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 59337–59359). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/b9ecf4d84999a61783c360c3782e801e-Paper-Conference.pdf

17. Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE. https://doi.org/10.1109/cvprw50498.2020.00359

18. Jiang, T., Chen, L., Chen, W., Meng, W., & Qi, P. (2023). ReliaMatch: Semi-Supervised Classification with Reliable Match. *Applied Sciences*, *13*(15), 8856. https://doi.org/10.3390/app13158856

19. Zhu, L., Ke, Z., & Lau, R. (2023). *Towards Self-Adaptive Pseudo-Label Filtering for Semi-Supervised Learning*. arXiv. https://arxiv.org/pdf/2309.09774

20. Jin, Y., & Lin, D. (2022). Semi-Supervised Semantic Segmentation via Gentle Teaching Assistant. In *Advances in Neural Information Processing Systems* (J. Wang, Corresponding author; Vol. 35, pp. 2803–2816). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/12d286282e1be5431ea05262a21f415c-Paper-Conference.pdf

21. Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. У *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

The article has been sent to the editors 08.10.24.
After processing 16.10.24.
Submitted for printing 30.12.24.