Balancing Logit Variation for Long-tailed Semantic Segmentation

Yuchao Wang[1†] Jingjing Fei[2] Haochen Wang[3†] Wei Li[2]
Tianpeng Bao[2] Liwei Wu[2] Rui Zhao[1,2‡] Yujun Shen[4]

[1]Shanghai Jiao Tong University [2]SenseTime Research
[3]Institute of Automation, Chinese Academy of Sciences [4]CUHK

ycw991216@163.com shenyujun0302@gmail.com wanghaochen2022@ia.ac.cn
{feijingjing1, liwei1, baotianpeng, wuliwei, zhaorui}@sensetime.com

Abstract

Semantic segmentation usually suffers from a long-tail data distribution. Due to the imbalanced number of samples across categories, the features of those tail classes may get squeezed into a narrow area in the feature space. Towards a balanced feature distribution, we introduce category-wise variation into the network predictions in the training phase such that an instance is no longer projected to a feature point, but a small region instead. Such a perturbation is highly dependent on the category scale, which appears as assigning smaller variation to head classes and larger variation to tail classes. In this way, we manage to close the gap between the feature areas of different categories, resulting in a more balanced representation. It is noteworthy that the introduced variation is discarded at the inference stage to facilitate a confident prediction. Although with an embarrassingly simple implementation, our method manifests itself in strong generalizability to various datasets and task settings. Extensive experiments suggest that our plug-in design lends itself well to a range of state-of-the-art approaches and boosts the performance on top of them.[1]

1. Introduction

The success of deep models in semantic segmentation [12, 58, 71, 109] benefits from large-scale datasets. However, popular datasets for segmentation, such as PASCAL VOC [24] and Cityscapes [18], usually follow a long-tail distribution, where some categories may have far fewer samples than others. Considering the particularity of this task, which targets assigning labels to pixels instead of images, it is quite difficult to balance the distribution from the aspect of data collection. Taking the scenario of autonomous driving as an example, a bike is typically tied to a smaller image region (*i.e.*, fewer pixels) than a car, and trains appear more rarely than pedestrians in a city. Therefore, learning a decent model from long tail data distributions becomes critical.

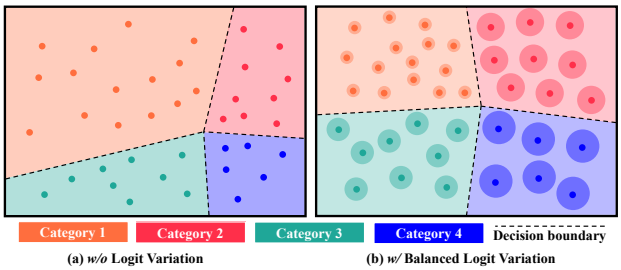


Figure 1. **Illustration of logit variation** from the feature space, where each point corresponds to an instance and different colors stand for different categories. (a) Without logit variation, the features of tail classes (*e.g.*, the **blue** one) may get squeezed into a narrow area. (b) After introducing logit variation, which is controlled by the category scale (*i.e.*, number of training samples belonging to a particular category), we expand each feature point to a feature region with random perturbation, resulting in a more category-balanced feature distribution.

A common practice to address such a challenge is to make better use of the limited samples from tail classes. For this purpose, previous attempts either balance the sample quantity (*e.g.*, oversample the tail classes when organizing the training batch) [8, 28, 36, 46, 75], or balance the per-sample importance (*e.g.*, assign the training penalties regarding tail classes with higher loss weights) [19,51,67,69]. Given existing advanced techniques, however, performance degradation can still be observed in those tail categories.

This works provides a new perspective on improving long-tailed semantic segmentation. Recall that, in modern pipelines based on neural networks [12,14,58,92,108,110], instances are projected to representative features before

[1]Code: https://github.com/grantword8/BLV.
[†]This work was done during the internship at SenseTime Research.
[‡]Rui Zhao is also with Qing Yuan Research Institute, Shanghai Jiao Tong University.

categorized to a certain class. We argue that the features of tail classes may get squeezed into a narrow area in the feature space, as the blue region shown in Fig. 1a, because of miserly samples. To balance the feature distribution, we propose a simple yet effective approach via introducing **balancing logit variation (BLV)** into the network predictions. Concretely, we perturb each predicted logit with a randomly sampled noise during training. That way, each instance can be seen as projected to a feature region, as shown in Fig. 1b, whose radius is dependent on the noise variance. We then propose to balance the variation by applying smaller variance to head classes and larger variance to tail classes so as to close the feature area gap between different categories. This newly introduced variation can be viewed as a special augmentation and discarded in the inference phase to ensure a reliable prediction.

We evaluate our approach on three different settings of semantic segmentation, including fully supervised [21, 90, 101, 109], semi-supervised [13, 87, 111, 115], and unsupervised domain adaptation [23, 52, 107, 113], where we improve the baselines consistently. We further show that our method works well with various state-of-the-art frameworks [2, 41, 42, 87, 92, 108] and boosts their performance, demonstrating its strong generalizability.

2. Related Work

Semantic segmentation. Network architecture for semantic segmentation has evolved for years, from CNNs [12, 58, 109] to Transformers [14, 21, 92, 108, 110]. Another line of research works focuses on enhancing the extracted representations like integrating attention mechanisms [27, 45, 50, 112] or context representations [55, 86, 100–102, 105] into segmentation models. BLV is complementary to these various frameworks and improves several state-of-the-art methods consistently.

Semi-supervised semantic segmentation. To alleviate the heavy need for large-scale annotated data, semi-supervised semantic segmentation has become a research hotspot. There are two typical frameworks for this task: consistency regularization [10, 26, 31] and self-training [3, 65, 77, 94]. Consistency regularization applies various perturbations [26] on training data and forces consistent predictions between the perturbed and the unperturbed input [31]. Self-training [43, 65, 87, 93, 97, 103, 116] uses the predictions from the pre-trained model as the "ground-truth" of the unlabeled data and then trains a semantic segmentation model in a fully-supervised manner. These two frameworks have no specialized operations for long-tail data. To this end, we provide a concise and generic approach that can be integrated into any framework.

Unsupervised domain adaptive semantic segmentation. UDA semantic segmentation aims at learning segmentation model that transfer knowledge from labeled source domain to unlabeled target domain. Early methods for UDA segmentation focus on enabling the model to extract to domain-invariant features. They align the cross-domain feature distribution at image level [29, 39, 74], feature level [7, 9, 53, 83] and output level [62, 83, 85] via image style transfer [32, 39, 47, 54, 98], image feature domain discriminator [30, 64, 66, 84, 88] or well-designed metrics [33, 49, 59]. Follow-up study [11, 107] suggests that the self-training-based pipeline leads to more consistent improvement. Recently, DAFormer [41] and HRDA [42] provide a self-training-based Transformer architecture together with many efficient training strategies, which can achieve consistent improvement over other competitors. BLV can be simply integrated into existing pipelines, and consistently improve their performance.

Long-tail learning. Since the long-tail phenomenon is common [96] in deep learning, the performance of the model tends to be dominated by the head category, while the learning of the tail category is severely underdeveloped. One intuitive solution to alleviate unbalanced data distribution is data processing, which typically consists of three ways: over-sampling [34, 35, 68, 89], under-sampling [6, 35, 76, 80] and data augmentation [15, 16, 56, 104]. Various methods have been proposed to alleviate the long-tail phenomenon in semantic segmentation, which can be mainly divided into three settings: fully supervised [5, 82], semi-supervised [25, 38, 44], and UDA [52, 73, 95, 114]. It is noteworthy that existing methods are usually limited to a specific setting and lack generalizability.

Noise-based augmentation. To improve model robustness and avoid over-fitting, augmenting data with noise [4, 20, 40] at image level or feature level is widely applied to model training. Techniques [60, 117] like Dropout [78], color jittering [1], gaussian noise, are the most common methods and proved to be simple yet efficient, but they might also introducing task-agnostic bias [99]. Another line of research aims to optimize the noise added in extracted features to "fool" the model [30, 48]. The optimized noise is defined as adversarial examples, which are commonly recognized as the hard sample for the model. Methods like *M2m* [46] and *AdvProp* [91] utilize adversarial examples to augment the training data and significantly improve model robustness. Prior arts focus on improving the robustness yet ignoring the prevalence of long-tail data, whereas our BLV can alleviate the feature squeeze caused by long-tail data effectively.

3. Method

In this section, we first formulate our problem mathematically and elaborate our approach detailly in Sec. 3.1. Then we specify how BLV can be used in three tasks where the settings are not exactly the same, *i.e.*, fully-supervised, semi-supervised, domain adaptive settings, in

Sec. 3.2, Sec. 3.3, Sec. 3.4, respectively.

3.1. Elaboration of BLV

Long-tailed label distribution is detrimental to the training of deep learning models. As Fig. 1a illustrated the total numbers of instances from tail categories are extremely much fewer when compared to head categories. As a result, they are squeezed into a very small area in the feature space, which means *the decision boundaries of these tailed categories can be severely biased*. Thus, at the inference stage, many similar data outside the distribution of the training tail category instances will be misclassified due to this squeeze. Next, we will elaborate on our approach.

Given a long-tailed training dataset with N labeled images of C categories: $D = \{(x^i_{image}, y^i_{image})\}^N_{i=1}$, where $y^i_{image} \in \{0, 1, ..., C-1\}$, our goal is to train a semantic segmentation model f_{model} with more balanced representations. To achieve this, we need to take a more fine-grained perspective. For segmentation tasks, the corresponding task-related instances are pixels, instead of images. Thus we can view the task as a multi-label classification task at the pixel level.

During the training stage, assuming there is an input data batch X_{batch} with a shape of $\{B, 3, H, W\}$ and its corresponding label Y_{batch}, where B is the batchsize and H, W denotes the size of the images, we can input it into the model f to get an output vector $\tilde{Z}_{batch}$. The shape of $\tilde{Z}_{batch}$ will be $\{B, C, H, W\}$ (we assume that H, W remain the same here for simplicity because $\tilde{Z}_{batch}$ can be upsampled to this size), where C is the number of categories.

From the view of instances (*i.e.*, pixels in segmentation task), we can reshape the output $\tilde{Z}_{batch}$ from $\{B, C, H, W\}$ into $\{B \times H \times W, C\}$. So for this batch, we have $B \times H \times W$ pixels and corresponding C-dimensional prediction for each of them. Taking pixel i as an example, its output $\tilde{Z}^i_{batch} = \left[z^i_0, z^i_1, \cdots, z^i_{C-1}\right]$. In order to calculate the cross entropy loss during training, we need to convert it into probabilities by the `softmax` formula Eq. (1).

$$p^i_k = \frac{e^{z^i_k}}{\sum_{j=0}^{C-1} e^{z^i_j}}, \tag{1}$$

where p^i_k denotes the probability of pixel i to be of category k and C is the number of categories. After obtaining the probabilities, common practices are to use them to calculate the Cross-Entropy Loss in Eq. (2).

$$L_{CE}(\tilde{Z}^i_{batch}) = -\sum_{k=0}^{C-1} y^i_k \log p^i_k, \tag{2}$$

where y^i_k is k-th term of the one-hot encoded ground truth $\left[y^i_0, y^i_1, \cdots, y^i_{C-1}\right]$.

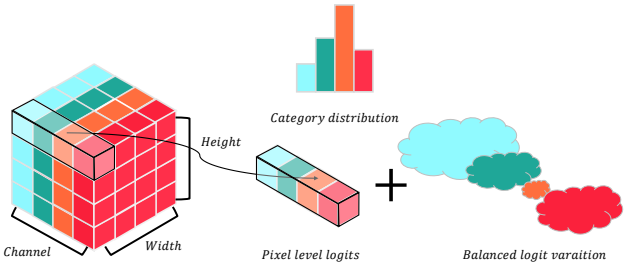


Figure 2. Diagram of the introduction of balanced logit variation, where we perturb the per-pixel logit with a category-specific noise. The noise variance is in inverse proportion to the category scale.

Every z^i_k is defined as the **logit** for the instance (*i.e.*, pixel) i. The step-by-step derivation from Eq. (1) to Eq. (2) depicts a direct relationship between the L_{CE} optimization and logit term z. Logit term z is critical to the long-tail problem. Because the dimensionality of logit z is consistent with the total number of categories and directly affects the computation of the loss, making it the most intuitive way to affect the size of the categorical area feature space. Then the crux lies in how to use logit to alleviate the long-tail problem. One intuitive way is simply rescale the logit according the category frequency [63]. However, semantic segmentation is an extremely instances-intensive task, so simply rescaling the logit fixedly according to the category frequency leads to overfitting problems.

To this end, we propose to add variation into the network predictions (*i.e.* z here) in Eq. (3).

$$\hat{z}^i_k = z^i_k + \frac{c_k}{\max_{i=0}^{C-1} c_i}|\delta(\sigma)|, \quad c_k = \log \frac{\sum_{j=0}^{C-1} q_j}{q_k} \tag{3}$$

where q_k is the number of the instances with category k and δ is a gaussian distribution with a mean of 0 and standard deviation of σ. Eq. (3) is quite easy to understand, as it assigns smaller variation to the head categories and larger variation to the tail categories. By adding this variation, which is inversely proportional to the category scale, our method can be equivalent to *expanding the distribution of each instance over the feature space from a single point into a small region*. Therefore, when training under this setting, we can obtain a more category-balanced feature representation space. We give a more straightforward explanation of our approach in Fig. 2. Besides, the only hyper-parameter of Eq. (3) is the σ, making it easy to generalize to other tasks. The form of variation can actually be not limited to gaussian distribution, in the ablation experiment Sec. 4.5 we found that variation sampled from other distributions can also work. It is noteworthy that the introduced variation is discarded at the inference stage to facilitate a confident prediction. Next, we will elaborate on how to specifically apply Eq. (3) for different settings of long-tail semantic segmentation tasks.

3.2. BLV for Fully Supervised Segmentation

Since the labels of the training data are available in the fully supervised semantic segmentation task, the category-by-category distribution can be obtained easily when pre-processing the data. It should be noted that since the instances of the segmentation task are pixels, the number of pixels of each class needs to be counted before obtaining their distribution. We present all experimental results for this task in Sec. 4.1.

3.3. BLV for Semi-Supervised Segmentation

Semi-supervised semantic segmentation is a more challenging task, due to the fact that only a small portion of the training images are carefully labeled [106]. A simple approach is to equate the pixel-level category distribution of the labeled images with the category distribution of the whole training set (including both the labeled and the unlabeled images). This estimated distribution is quite inaccurate when the labeled/unlabeled division is very extremely unbalanced, for example, the $1/16(186)$ partition protocols in Sec. 4.2.

Therefore, we propose an epoch-based update strategy of the distribution to make it closer to the true distribution. Suppose after n epochs of training, we have a model f_n. For all the unlabeled training images set X^u_{image}, we infer the labels of all the images in X^u_{image} thus get its corresponding pseudo-label: $\hat{Y}^u_{image}$. Thus, we calculate the number of pixels in each category by the following formula Eq. (4).

$$q_k = \frac{\sum_{n=1}^{N}\sum_{m=1}^{H\times W}\mathbb{1}\left[\hat{y}^u_{nm}=k\right]}{\sum_{i=0}^{C-1}\sum_{n=1}^{N}\sum_{m=1}^{H\times W}\mathbb{1}\left[\hat{y}^u_{nm}=i\right]} \tag{4}$$

where q_k denotes the k-th category frequency, $\hat{y}^u_{nm}$ denotes the m-th element of the n-th pseudo-label and N is the number of the unlabeled images.

Eq. (4) will be used to calculate the updated category distribution $\{q_0, q_1, \cdots, q_{n-1}\}$ after every epoch. Then we can bring this estimated distribution into Eq. (3). With this design, our approach can efficiently and consistently improve the performance of semi-supervised semantic segmentation tasks.

3.4. BLV for UDA Segmentation

Unsupervised domain adaptive semantic segmentation attempts to train a model that works well on the target domain by using labeled source domain data and unlabeled target domain data. Since the goal is to improve the performance of the model on the target domain, yet the images on this domain are unlabeled, this poses a tricky problem.

A widely recognized perspective is to view UDA semantic segmentation as semi-supervised semantic segmentation [107]. Because the source domain data is naturally labeled, this perspective makes certain sense without considering the inter-domain gap. Therefore, we can estimate the distribution of the target domain data with Eq. (4).

However, for the UDA semantic segmentation, we propose a more concise way: viewing the category distribution of the source domain data as if it were the category distribution of the target domain. The data from the source domain are typical computer-rendered synthetic labeled images while the data from the target domain are generally a real-world collection of images. This difference makes the images of the two domains significantly different only in terms of style, and essentially identical in terms of contextual relationships and category distribution. In Sec. 4.3 we have experimentally demonstrated that this simple estimation method works.

4. Experiments

We present experimental results on three mainstream segmentation tasks: semantic segmentation, semi-supervised semantic segmentation, and domain adaptive semantic segmentation. Besides, we present comparisons with previous works towards class-imbalanced problems. The mean of Intersection over Union (mIoU) is adopted as the metric to evaluate all the results.

4.1. Towards Fully Supervised Setting

Datasets. We used the typical long-tailed dataset: Cityscapes [18]. Cityscapes is a driving dataset for semantic segmentation, which consists of 5000 high-resolution images for training and 500 images for validation. We first resize the training images into a resolution of 2048×1024, then crop them into 512×1024.

Implementation details. We used the mmsegmentation codebase [17] and trained all the models with 8 Tesla V100 GPUs. To validate the proposed method BLV, we apply our method BLV to various state-of-the-art semantic segmentation models including ResNet [37], Swin-Transformer [57], Mix-Transformer [92], ViT [21] based encoder with OCR-Head [101], K-NeT [108], PSPHead [109], Segformer-Head [92], UperHead [90] based decoders respectively. The batch size is set to 16 for all models. The training iterations are $160k$ for MiT-b0 + SegformerHead, $80k$ for Swin-T + K-Net and Vit-B16 + UperHead, $40k$ for all the other models. We use AdamW optimizer for three transformer-based models: Swin-T + K-Net, MiT-b0 + SegformerHead, and Vit-B16 + UperHead, with a learning rate of 6×10^{-5}, weight decay of 0.01, a linear learning rate warmup with $1.5k$ iterations and linear decay afterwards. For all of the other models, we use the same configuration: SGD optimizer with a learning rate of 0.01, a weight decay of 5×10^{-4}.

Results. Table Tab. 1 summarizes the detailed comparison results across different architectures. We observe that our

method boosts all of these baseline models consistently. Equipped with our method, these models gains $+0.72\%$, $+0.43\%$, $+0.55\%$, $+0.24\%$, $+0.47\%$, $+0.35\%$, $+1.20\%$ respectively without any additional model parameters. The performance gains on various models with different network structures, including CNN-based and Transformer based models, indicate that our method is universal and can be generalized to various segmentation models. To verify the effectiveness of BLV towards long-tail data, we compute mIoU on **9** tail categories: *Wall*, *T.light*, *Sign*, *Rider*, *Truck*, *Bus*, *Train*, *M.bike*, *Bike*. The "mIoU (tail)" column demonstrates that the BLV indeed boosts the performance of these tail categories by a large margin.

4.2. Towards Semi-Supervised Setting

Datasets. Semi-Supervised semantic segmentation aims to learn a model with only a few labeled samples. Two typical benchmark datasets are usually used for validation: PASCAL VOC 2012 [24] and Cityscapes [18]. PASCAL VOC 2012 is a class-balanced and simpler dataset. Therefore, we mainly conduct experimental verification on Cityscapes. We follow the commonly used $1/16$, $1/8$, $1/4$ and $1/2$ partition protocols, that is, only the corresponding fraction number have labels, and the rest of the images are considered unlabeled. It is worth mentioning that our method adopts the generally used sliding window evaluation when evaluating.

Implementation details. The classical Self-Training [2] without any other tricks as the baseline due to its simplicity and to be consistent with our proposed method. The core of semi-supervised segmentation methods is the training strategy, not the network structure. So we use ResNet-101 [37] as the backbone and DeepLabv3+ [12] as the decoder. We use stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01, and weight decay as 0.0005. The momentum coefficient μ for Teacher model [81] updating is set to 0.999. The crop size is set as 769×769 and batchsize is set as 16.

Results. With our proposed BLV, Tab. 2 demonstrates that naive self-training framework achieves consistent performance gains over the naive self-training baseline by $+1.05\%$, $+1.26\%$, $+1.49\%$, $+0.99\%$ under $1/16$, $1/8$, $1/4$ and $1/2$ partition protocols. To verify the effectiveness of BLV towards long-tail data in semi-supervised segmentation task, we also list the "mIoU(tail)" column as the Sec. 4.1. This demonstrates that the BLV indeed improves the performance of the tail categories.

4.3. Towards UDA Setting

Datasets. Unsupervised Domain adaptive (UDA) semantic segmentation aims at transferring the knowledge from a source domain to a target domain. The source domain is a labeled dataset obtained from the synthetic images and the target domain is an unlabeled real image dataset. We use two synthetic datasets: GTA5 [70] and SYNTHIA [72] as source domains respectively and use real images from Cityscapes [18] as the target domain. In other words, We conduct experiments on two dataset settings: *GTA5* $\rightarrow$ *Cityscapes* and *SYNTHIA* $\rightarrow$ *Cityscapes*. It is worth mentioning that in *SYNTHIA* $\rightarrow$ *Cityscapes*, 16 and 13 of the 19 classes of Cityscapes are used to calculate mIoU, following the common practice [41]

Implementation details. We use the recent state-of-the-art frameworks DAFormer [41] and HRDA [42] as the baseline. In addition, since DAFormer [41] is a pure Transformer-based framework that integrates some effective training strategies, to verify our method, we also try to replace the model structure with ResNet101 [37] + DeepLabV2 [12]. This version of DAFormer based on the CNN structure also illustrates the generality of our method. In all UDA segmenetation experiments, AdamW [22] optimizer is utilized with a learning rate of 6×10^{-5} for the encoder and 6×10^{-4} for the decoder. This optimizer

Table 1. Experiments across architectures for fully semantic segmentation tasks on **Cityscapes** *validation* set. The green arrows indicate the relative improvement in performance.

Backbone	Decoder	mIoU	mIoU (tail)
HRNet-18 [79]	OCRHead [101]	79.22	63.51
	+ BLV	**79.94** ↑ **0.72**	**66.70** ↑ **3.19**
ResNet50 [12]	UperHead [90]	78.28	62.56
	+ BLV	**78.63** ↑ **0.35**	**64.57** ↑ **2.01**
ResNet50 [37]	PSPHead [109]	77.98	61.96
	+ BLV	**78.53** ↑ **0.55**	**63.34** ↑ **1.38**
ResNet101 [37]	UperHead [90]	79.41	64.68
	+ BLV	**79.88** ↑ **0.47**	**66.29** ↑ **1.61**
MiT-b0 [92]	SegformerHead [92]	76.85	67.58
	+ BLV	**77.09** ↑ **0.24**	**68.91** ↑ **1.33**
Swin-T [57]	K-NeT [108]	79.68	71.70
	+ BLV	**80.11** ↑ **0.43**	**72.94** ↑ **1.24**
Vit-B16 [21]	UperHead [90]	76.48	68.25
	+ BLV	**77.68** ↑ **1.20**	**70.63** ↑ **2.38**

Table 2. Experiments on semi-supervised semantic segmentation tasks on **Cityscapes** *validation* set. The green arrows indicate the relative improvement in performance.

Partition	Method	mIoU	mIoU (tail)
1/16 (186)	Self-Training	68.21	53.09
	+BLV	**69.26** ↑ **1.05**	**55.23** ↑ **2.14**
1/8 (372)	Self-Training	72.01	58.74
	+BLV	**73.27** ↑ **1.26**	**60.33** ↑ **1.59**
1/4 (744)	Self-Training	74.03	61.76
	+BLV	**75.52** ↑ **1.49**	**63.51** ↑ **1.75**
1/2 (1488)	Self-Training	77.99	65.96
	+BLV	**78.98** ↑ **0.99**	**67.24** ↑ **1.28**

Table 3. Comparison with state-of-the-art alternatives on *GTA5* → *Cityscapes* benchmark. The results are averaged over 3 random seeds. The top performance is highlighted in **bold** font. † indicates that the corresponding framework uses a CNN-based structure. ‡ indicates that the corresponding framework uses a Transformer-based structure.

Method	Road	S.walk	Build.	Wall	Fence	Pole	T.light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	mIoU
source only†	70.2	14.6	71.3	24.1	15.3	25.5	32.1	13.5	82.9	25.1	78.0	56.2	33.3	76.3	26.6	29.8	12.3	28.5	18.0	38.6
DAFormer†	94.6	66.5	87.9	39.5	33.7	38.5	49.6	60.0	88.0	**46.6**	88.3	**69.6**	44.4	**89.0**	46.8	**56.8**	0.0	17.8	44.3	55.9
DAFormer (*w/* BLV)	**94.9**	**68.2**	**88.8**	**40.9**	**37.1**	**42.6**	**52.1**	**62.1**	**88.3**	43.3	**89.3**	68.6	**44.5**	88.9	**56.0**	54.6	**3.8**	**38.6**	**58.3**	**59.0**
DAFormer‡	95.7	70.2	**89.4**	53.5	48.1	49.6	**55.8**	59.4	**89.9**	47.9	**92.5**	72.2	**44.7**	**92.3**	74.5	**78.2**	65.1	55.9	61.8	68.3
DAFormer (*w/* BLV)	**96.2**	**73.1**	89.3	**53.6**	**55.7**	**50.9**	55.7	**61.1**	89.7	**52.4**	92.3	**74.7**	43.5	91.6	**74.6**	77.4	**69.2**	**58.9**	**62.3**	**69.6**
HRDA‡	96.4	74.4	91.0	**61.6**	51.5	57.1	**63.9**	69.3	91.3	48.4	94.2	**79.0**	52.9	**93.9**	**84.1**	**85.7**	75.9	63.9	67.5	73.8
HRDA (*w/* BLV)	**96.7**	**76.6**	**91.5**	61.2	**56.9**	**59.4**	62.2	**72.8**	**91.5**	**51.2**	**94.3**	77.5	**54.7**	93.5	83.2	84.7	**79.7**	**68.1**	**67.6**	**74.9**

Table 4. Comparison with state-of-the-art alternatives on *SYNTHIA* → *Cityscapes* benchmark. The results are averaged over 3 random seeds. The mIoU and the mIoU* indicate we compute mean IoU over 16 and 13 categories, respectively. The top performance is highlighted in **bold** font. † indicates that the corresponding framework uses a CNN-based structure. ‡ indicates that the corresponding framework uses a Transformer-based structure.

Method	Road	S.walk	Build.	Wall*	Fence*	Pole*	T.light	Sign	Veget.	Sky	Person	Rider	Car	Bus	M.bike	Bike	mIoU	mIoU*
source only†	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6
DAFormer†	62.2	24.5	85.3	23.4	2.5	38.5	47.7	**51.1**	84.0	81.8	70.5	**41.3**	77.9	46.6	45.3	60.3	52.7	59.9
DAFormer (*w/* BLV)	**70.4**	**28.9**	**89.2**	**25.2**	**19.9**	**40.2**	**55.2**	50.3	**86.9**	**84.2**	**76.4**	40.5	**79.6**	**51.3**	**49.2**	**61.2**	**56.8**	**63.3**
DAFormer‡	84.5	40.7	88.4	41.5	**6.5**	50.0	55.0	54.6	86.0	89.8	73.2	**48.2**	87.2	53.2	53.9	61.7	60.9	67.4
DAFormer (*w/* BLV)	**86.7**	**44.9**	**89.0**	**43.2**	6.4	**52.1**	**60.0**	**54.9**	**88.2**	**91.3**	**74.9**	46.1	**88.6**	**55.6**	**55.0**	**62.3**	**62.5**	**69.0**
HRDA‡	85.2	47.7	88.8	49.5	4.8	**57.2**	65.7	60.9	85.3	92.9	**79.4**	52.8	89.0	**64.7**	63.9	64.9	65.8	72.7
HRDA (*w/* BLV)	**87.6**	**47.9**	**90.5**	**50.4**	**6.9**	57.1	64.3	**65.3**	**86.9**	**93.4**	78.9	**54.9**	**89.1**	62.9	**65.2**	**66.8**	**66.8**	**73.4**

is set to be with a weight decay of 0.01 along with a linear learning rate warmup with $1.5k$ iterations and linear decay afterward. During training, for the DAFormer-based methods, per batch input is set to be of two 512×512 random crops. For HRDA [42], whose main motivation considers the training image resolution, we adopt the settings consistent with the paper.

Results. Tab. 3 and Tab. 4 both suggest that our proposed BLV can consistently improve the performance of the UDA segmentation task. Our BLV advances the baseline frameworks DAFormer†, DAFormer‡, HRDA‡ with $+3.1\%$, $+1.3\%$, $+1.1\%$ respectively on *GTA5* → *Cityscapes* benchmark. BLV also advances DAFormer†, DAFormer‡, HRDA‡ with $+4.1\%$, $+1.6\%$, $+1.0\%$ on the mIoU evaluation of 16 categories and with $+3.4\%$, $+1.6\%$, $+0.7\%$ on the mIoU evaluation of 13 categories respectively on *SYNTHIA* → *Cityscapes* benchmark. From the per-category results in Tab. 4 and Tab. 3, we can observe that the improvement of our method for the overall mIoU comes from the improvement of IoU of the tail categories. We make this conclusion even more obvious by plotting the pixel-level category frequency versus performance improvement on tail categories in figure Fig. 3. Our BLV

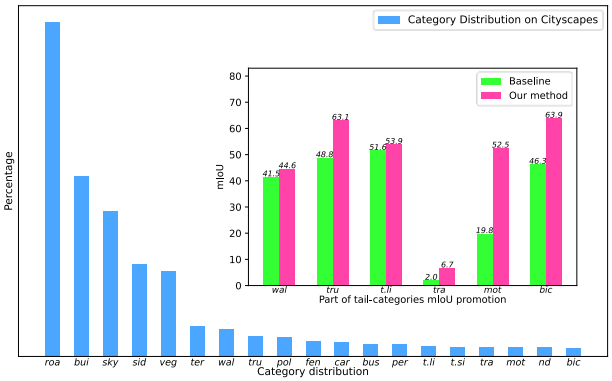


Figure 3. **Pixel-level category distribution versus performance improvement on multiple tail-categories.** This figure suggests that the performance improvement of our BLV comes mainly from the tail category.

achieves $+3.1\%$, $+14.3\%$, $+2.3\%$, $+4.7\%$, $+32.7\%$ and $+17.6\%$ boosts on "wall", "truck", "traffic light", "trailer", "motorcycle" and "bicycle", respectively, which happen to belong to the tail categories, indicating that our method optimizes the feature space of the tail category which leads to consistent performance improvements.

4.4. Comparisons with Long-tailed Methods

Implementation details. For the fully-supervised setting, we implement Lovász-Softmax [5] and Logit-Adjustment [63] loss on the ViT-B/16 + UperHead model. For the semi-supervised setting, we implement BLV and Logit-Adjustment [63] on the ResNet-50 + PSPNet model for a fair comparison with DARS [38]. For the UDA setting, we compare BLV with naive resampling the input instances, CBST [52], CLAN [61] and Logit-Adjustment [63] on *GTA5* $\rightarrow$ *Cityscapes*.

Results. We demonstrate more comparisons in Tab. 5. BLV achieves the mIoU of 77.7% on fully-supervised setting, 73.2% on semi-supervised setting and 59.0% on the domain adaptive setting. BLV also achieves the mIoU of 66.2% on fully-supervised setting, 59.3% on semi-supervised setting and 45.7% on the domain adaptive setting for the tail-categories. The overall better performance suggests BLV boosts the tail categories and outperforms other alternatives on different settings and benchmarks, which demonstrates its versatility and effectiveness.

4.5. Ablation Studies

Exploration on the form of variations. In Tab. 6, we explore the influence of different perturbation forms on the experimental results. Results are obtained from the DAFormer† framework on *GTA5* $\rightarrow$ *Cityscapes setting*. "None-Variation" denotes the pure baseline. "Gaussian" variation parameters are illustrated in Sec. 3, For the "Uniform" term, We sample uniformly in the interval $[0, 1]$, which brings in the baseline with a boost of $+2.3\%$. For the "Beta" variation, we set the $\alpha = 0.5$ and $\beta = 0.5$, which advances the baseline by $+2.0\%$. For the "Exponential" variation, we set the $\lambda = 1$, which surpasses the baseline by $+2.6\%$ In order to ensure that the size of the perturbation is within a reasonable range, we clip all perturbations to the $[0, 1]$ interval. We empirically find that the "Gaussian" variation term outperforms all the other alternatives with a mIoU of 59.0%, while all the other variations advance the baseline. These suggest that various variations can improve the performance of the task, and the key to the improvement lies in the coefficients related to the category frequency rather than in the form of variation. This is more in line with our intuition. If the parameters of other variations are finetuned, better results may be obtained.

Exploration on the variance σ. As the only hyper-parameter needs to be carefully tuned, we ablate σ for potential generalized usage extended to other tasks. Tab. 7 gives experimental results on the influence of different σ with DAFormer † under two different adaptation settings: *GTA5* $\rightarrow$ *Cityscapes* and *SYNTHIA* $\rightarrow$ *Cityscapes*. BLV advances the baseline most by $+3.1\%$ when $\sigma = 6$ for *GTA5* $\rightarrow$ *Cityscapes* and by $+4.1\%$ when $\sigma = 4$ for *SYNTHIA* $\rightarrow$ *Cityscapes*. Although the different choices of σ affect the final performance, these gaps are quite trivial(the discrepancy between the maximum and minimum mIoU is within $+1\%$), which demonstrates that our BLV is robust to hyper-parameter choices to some extent and indicates its good scalability.

Exploration on components of BLV. Corresponding results are demonstrated in Tab. 8. "*w/o* variation" denotes our BLV without adding the variation in Sec. 3 rather a

Table 5. More comparisons with other long-tailed baselines on different semantic segmentation tasks. "RS" and "RW" denotes the naive resample and reweight trick respectively. * means the results come from our implementation. ° means the results come from the original papers.

Supervision	Fully			Semi		
Method	LA*	Lovász*	BLV	LA*	DARS°	BLV
mIoU	75.9	76.6	**77.7**	69.3	72.8	**73.2**
mIoU (tail)	62.4	63.9	**66.2**	55.7	58.4	**59.3**
Supervision	Domain Adaptive					
Method	RS	RW	CLAN°	CBST°	LA*	BLV
mIoU	56.2	56.4	43.2	45.9	56.5	**59.0**
mIoU (tail)	40.5	40.8	25.9	28.5	41.9	**45.7**

Table 6. **Ablation study on various types of variations.** "None-Variation" denotes the DAFormer† baseline. The green arrows indicate the relative improvement in performance.

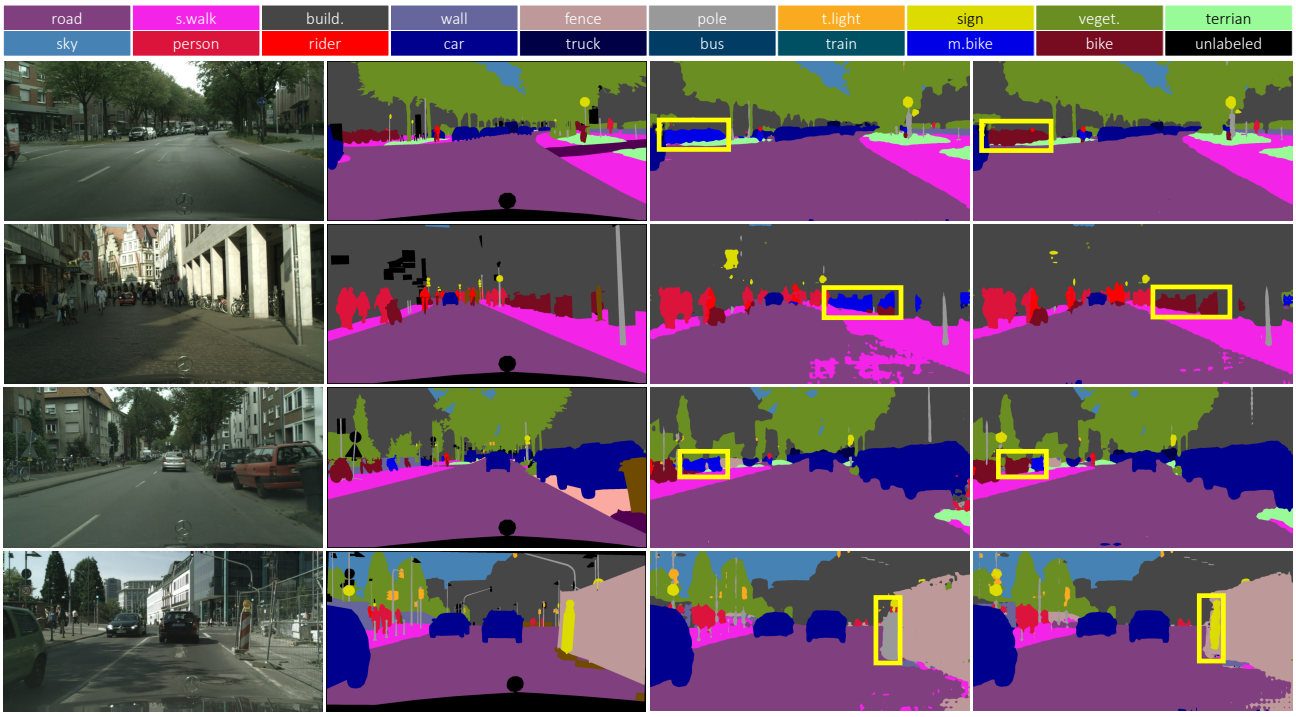
Variation	mIoU	
None-Variation	55.9	
Gaussian	**59.0**	↑ **3.1**
Uniform	58.2	↑ **2.3**
Beta	57.9	↑ **2.0**
Exponential	58.5	↑ **2.6**

Table 7. **Ablation study on the variance σ in Eq. (3),** which determines the overall magnitude of the variation.

Baseline	3	4	5	6	7
GTA5 $\rightarrow$ *Cityscapes*					
55.9	58.0	58.8	58.2	**59.0**	58.7
SYNTHIA $\rightarrow$ *Cityscapes*					
52.7	56.5	**56.8**	55.9	56.3	56.1

Table 8. **Ablation study on the components of BLV.** We ablate two components in Eq. (3). "*w/o* variation" indicates removing the $|\delta(\sigma)|$ term. "*w/o* variation" indicates removing the pixel-level category frequency term.

Baseline	*w/o* variation	*w/o* balance	BLV
GTA5 $\rightarrow$ *Cityscapes*			
55.9	56.5	56.8	**59.0**
SYNTHIA $\rightarrow$ *Cityscapes*			
52.7	53.9	54.5	**56.8**



(a) Image (b) Ground Truth (c) Baseline (d) BLV

Figure 4. **Qualitative results on Cityscapes val set**. Baseline and BLV are trained on *GTA5*→ *Cityscapes* benchmark of unsupervised domain adaptive semantic segmentation task. (a) Input images. (b) Ground truth annotations for the corresponding images. (c) Result of the DAFormer baseline. (d) Result of our method (DAFormer + BLV). Yellow rectangles highlight the promotion of segmentation results by our method on tail categories.

constant value adjustment for each category. "*w/o* balance" denotes our BLV without adding categorical balance coefficient in Sec. 3 rather a fixed-scale adjustment for each category. Tab. 8 demonstrates that either the "*w/o* variation" or "*w/o* balance" can boost the baseline non-trivially on both the *GTA5* $\rightarrow$ *Cityscapes* by $+0.6\%$, $+0.9\%$ and the *SYNTHIA* $\rightarrow$ *Cityscapes* settings by $+1.2\%$, $+1.8\%$, although not as significant as our BLV. This suggests two conclusions: 1) Adding category-agnostic consistent variation to logits can indeed optimize the representation space to a certain extent, but it cannot completely solve the adverse effects of long-tailed data. 2) Adding static category-related adjustments can also alleviate this problem, but this cannot enrich training instances thus leading to potential overfitting problems while the variation term of BLV can be used to avoid this problem efficiently. This ablation experiment demonstrates the necessity of all components of our proposed BLV.

4.6. Visualization

Fig. 4 shows the result of our method on the Cityscapes val set. With this visualization, we prove that overlaying our method to the baseline is effective in alleviating category confusion, so our method achieves better performance. More details can be demonstrated by the yellow rectangle highlighting part in Fig. 4c and Fig. 4d. (*i.e.* the pixel misclassified in the baseline are corrected by balancing logit variation.)

5. Conclusion

In this paper, we propose the BLV, a simple yet effective plug-in design for various kinds of long-tail semantic segmentation tasks. We introduce category scale-related variation during the model training stage. This variation is inversely proportional to the frequency of occurrences of instances, which effectively closes the gap between the feature area of different categories. Extensive experiments on fully supervised, semi-supervised, and unsupervised domain adaptive semantic segmentation tasks suggest our method can boost performance. Compared with other methods towards alleviating the class-imbalance issues, our BLV is better and more concise and general. Furthermore, sufficient ablation experiments as well as intuitive visualization results prove the necessity of individual components and the effectiveness of our method.

Discussion. One necessary premise of BLV is that the category frequencies need to be known. It is unlikely to be satisfied in some tasks like unsupervised semantic segmentation and domain generalized semantic segmentation.

References

[1] Mahmoud Afifi and Michael S Brown. What else can fool deep learning? addressing color constancy errors on deep neural network performance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 243–252, 2019. 2

[2] Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Emilie Devijver, and Yury Maximov. Self-training: A survey. *arXiv preprint arXiv:2202.12040*, 2022. 2, 5

[3] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014. 2

[4] Yoshua Bengio, Frédéric Bastien, Arnaud Bergeron, Nicolas Boulanger-Lewandowski, Thomas Breuel, Youssouf Chherawala, Moustapha Cisse, Myriam Côté, Dumitru Erhan, Jeremy Eustache, et al. Deep learners benefit more from out-of-distribution examples. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 164–172. JMLR Workshop and Conference Proceedings, 2011. 2

[5] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4413–4421, 2018. 2, 7

[6] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018. 2

[7] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 1

[9] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[10] Huaian Chen, Yi Jin, Guoqiang Jin, Changan Zhu, and Enhong Chen. Semi-supervised semantic segmentation by improving prediction confidence. *IEEE Transactions on Neural Networks and Learning Systems*, 13(1), 2021. 2

[11] Lin Chen, Zhixiang Wei, Xin Jin, Huaian Chen, Miao Zheng, Kai Chen, and Yi Jin. Deliberated domain bridging for domain adaptive semantic segmentation. *arXiv preprint arXiv:2209.07695*, 2022. 2

[12] 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, 2017. 1, 2, 5

[13] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2613–2622, 2021. 2

[14] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 1, 2

[15] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *Eur. Conf. Comput. Vis.*, pages 95–110. Springer, 2020. 2

[16] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *Eur. Conf. Comput. Vis.*, pages 694–710. Springer, 2020. 2

[17] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 4

[18] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1, 4, 5

[19] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9268–9277, 2019. 1

[20] Jun Ding, Bo Chen, Hongwei Liu, and Mengyuan Huang. Convolutional neural network with data augmentation for sar target recognition. *IEEE Geoscience and remote sensing letters*, 13(3):364–368, 2016. 2

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 4, 5

[22] Timothy Dozat. Incorporating nesterov momentum into adam. In *Int. Conf. Learn. Represent. Worksh.*, 2016. 5

[23] Ye Du, Yujun Shen, Haochen Wang, Jingjing Fei, Wei Li, Liwei Wu, Rui Zhao, Zehua Fu, and Qingjie Liu. Learning from future: A novel self-training framework for semantic segmentation. *Adv. Neural Inform. Process. Syst.*, 2022. 2

[24] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 1, 5

[25] Jiashuo Fan, Bin Gao, Huan Jin, and Lihui Jiang. Ucc: Uncertainty guided cross-head co-training for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9947–9956, 2022. 2

[26] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *Brit. Mach. Vis. Conf.*, 2020. 2

[27] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3146–3154, 2019. 2

[28] Mikel Galar, Alberto Fernández, Edurne Barrenechea, and Francisco Herrera. Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recog.*, 46(12):3460–3471, 2013. 1

[29] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, 2014. 2

[31] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. *CAP*, 367:281–296, 2005. 2

[32] Shaohua Guo, Qianyu Zhou, Ye Zhou, Qiqi Gu, Junshu Tang, Zhengyang Feng, and Lizhuang Ma. Label-free regional consistency for image-to-image translation. In *Int. Conf. Multimedia and Expo*, 2021. 2

[33] Xiaoqing Guo, Chen Yang, Baopu Li, and Yixuan Yuan. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[34] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5356–5364, 2019. 2

[35] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239, 2017. 2

[36] Jingyu Hao, Chengjia Wang, Heye Zhang, and Guang Yang. Annealing genetic gan for minority oversampling. *arXiv preprint arXiv:2008.01967*, 2020. 1

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 4, 5

[38] Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *Int. Conf. Comput. Vis.*, pages 6930–6940, 2021. 2, 7

[39] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Int. Conf. Mach. Learn.*, 2018. 2

[40] Lasse Holmstrom, Petri Koistinen, et al. Using additive noise in back-propagation training. *IEEE transactions on neural networks*, 3(1):24–38, 1992. 2

[41] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 5

[42] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. *Eur. Conf. Comput. Vis.*, 2022. 2, 5, 6

[43] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. In *Adv. Neural Inform. Process. Syst.*, 2021. 2

[44] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021. 2

[45] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 603–612, 2019. 2

[46] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13896–13905, 2020. 1, 2

[47] Myeongjin Kim and Hyeran Byun. Learning texture invariant representation for domain adaptation of semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[48] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2

[49] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[50] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018. 2

[51] Mingchen Li, Xuechen Zhang, Christos Thrampoulidis, Jiasi Chen, and Samet Oymak. Autobalance: Optimized loss functions for imbalanced data. *Adv. Neural Inform. Process. Syst.*, 34:3163–3177, 2021. 1

[52] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11593–11603, 2022. 2, 7

[53] Shuang Li, Binhui Xie, Bin Zang, Chi Harold Liu, Xinjing Cheng, Ruigang Yang, and Guoren Wang. Semantic distribution-aware contrastive adaptation for semantic segmentation. *arXiv preprint arXiv:2105.05013*, 2021. 2

[54] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[55] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 603–619, 2018. 2

[56] Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. Breadcrumbs: Adversarial class-balanced sampling for long-tailed recognition. In *Eur. Conf. Comput. Vis.*, pages 637–653. Springer, 2022. 2

[57] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, pages 10012–10022, 2021. 4, 5

[58] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 1, 2

[59] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Int. Conf. Mach. Learn.*, 2015. 2

[60] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019. 2

[61] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2507–2516, 2019. 7

[62] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[63] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *Int. Conf. Learn. Represent.*, 2021. 3, 7

[64] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Adv. Neural Inform. Process. Syst.*, 2016. 2

[65] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[66] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[67] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 735–744, 2021. 1

[68] Junran Peng, Xingyuan Bu, Ming Sun, Zhaoxiang Zhang, Tieniu Tan, and Junjie Yan. Large-scale object detection in the wild from imbalanced multi-labels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9709–9718, 2020. 2

[69] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Int. Conf. Mach. Learn.*, pages 4334–4343. PMLR, 2018. 1

[70] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Eur. Conf. Comput. Vis.*, 2016. 5

[71] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241. Springer, 2015. 1

[72] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 5

[73] Congcong Ruan, Wei Wang, Haifeng Hu, and Dihu Chen. Category-level adversaries for semantic domain adaptation. *IEEE Access*, 7:83198–83208, 2019. 2

[74] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[75] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. *Eur. Conf. Comput. Vis.*, 2016. 1

[76] Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. Class-wise difficulty-balanced loss for solving class-imbalance. In *Proceedings of the Asian conference on computer vision*, 2020. 2

[77] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[78] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 2

[79] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5693–5703, 2019. 5

[80] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11662–11671, 2020. 2

[81] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 5

[82] Junjiao Tian, Niluthpol Chowdhury Mithun, Zachary Seymour, Han-Pang Chiu, and Zsolt Kira. Striking the right balance: Recall loss for semantic segmentation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 5063–5069. IEEE, 2022. 2

[83] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[84] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Int. Conf. Comput. Vis.*, 2019. 2

[85] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[86] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui

Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3349–3364, 2020. 2

[87] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[88] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[89] Jialian Wu, Liangchen Song, Tiancai Wang, Qian Zhang, and Junsong Yuan. Forest r-cnn: Large-vocabulary long-tailed object detection and instance segmentation. In *ACM Int. Conf. Multimedia*, pages 1570–1578, 2020. 2

[90] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Eur. Conf. Comput. Vis.*, pages 418–434, 2018. 2, 4, 5

[91] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 819–828, 2020. 2

[92] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Adv. Neural Inform. Process. Syst.*, 2021. 1, 2, 4, 5

[93] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[94] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *Int. Conf. Mach. Learn.*, 2021. 2

[95] Jihan Yang, Ruijia Xu, Ruiyu Li, Xiaojuan Qi, Xiaoyong Shen, Guanbin Li, and Liang Lin. An adversarial perturbation oriented domain adaptation approach for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12613–12620, 2020. 2

[96] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *Int. J. Comput. Vis.*, pages 1–36, 2022. 2

[97] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[98] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[99] Shi Yin, Chao Liu, Zhiyong Zhang, Yiye Lin, Dong Wang, Javier Tejedor, Thomas Fang Zheng, and Yinguo Li. Noisy training for deep neural networks in speech recognition. *ICASSP*, 2015(1):1–14, 2015. 2

[100] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12416–12425, 2020. 2

[101] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Eur. Conf. Comput. Vis.*, pages 173–190. Springer, 2020. 2, 4, 5

[102] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context for semantic segmentation. *Int. J. Comput. Vis.*, 129(8):2375–2398, 2021. 2

[103] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Int. Conf. Comput. Vis.*, 2019. 2

[104] Yuhang Zang, Chen Huang, and Chen Change Loy. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *Int. Conf. Comput. Vis.*, pages 3457–3466, 2021. 2

[105] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7151–7160, 2018. 2

[106] Man Zhang, Yong Zhou, Jiaqi Zhao, Yiyun Man, Bing Liu, and Rui Yao. A survey of semi-and weakly supervised semantic segmentation of images. *Artificial Intelligence Review*, 53(6):4259–4288, 2020. 4

[107] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12414–12424, 2021. 2, 4

[108] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Adv. Neural Inform. Process. Syst.*, 34:10326–10338, 2021. 1, 2, 4, 5

[109] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1, 2, 4, 5

[110] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6881–6890, 2021. 1, 2

[111] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 7273–7282, 2021. 2

[112] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13065–13074, 2020. 2

[113] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma.

Uncertainty-aware consistency regularization for cross-domain semantic segmentation. *Computer Vision and Image Understanding*, page 103448, 2022. 2

[114] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2

[115] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *Int. Conf. Learn. Represent.*, 2020. 2

[116] Simiao Zuo, Yue Yu, Chen Liang, Haoming Jiang, Siawpeng Er, Chao Zhang, Tuo Zhao, and Hongyuan Zha. Self-training with differentiable teacher. *arXiv preprint arXiv:2109.07049*, 2021. 2

[117] Richard M Zur, Yulei Jiang, Lorenzo L Pesce, and Karen Drukker. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. *Medical physics*, 36(10):4810–4818, 2009. 2