FAKD: Feature Augmented Knowledge Distillation for Semantic Segmentation

Jianlong Yuan[1,2*†] Minh Hieu Phan[3*] Liyang Liu[3] Yifan Liu[3]
[1]Damo Academy, Alibaba Group [2]Hupan Lab [3]University of Adelaide

Abstract

In this work, we explore data augmentations for knowledge distillation on semantic segmentation. Due the capacity gap, small-sized student networks struggle to discover the discriminative feature space learned by a powerful teacher. Image-level augmentations allow the student to better imitate the teacher by providing extra outputs. However, existing distillation frameworks only augment a limited number of samples, which restricts the learning of a student. Inspired by the recent progress on semantic directions on feature space, this work proposes a feature-level augmented knowledge distillation (FAKD) which infinitely augments features along a semantic direction for optimal knowledge transfer. Furthermore, we introduce novel surrogate loss functions to distill the teacher's knowledge from an infinite number of samples. The surrogate loss is an upper bound of the expected distillation loss over infinite augmented samples. Extensive experiments on four semantic segmentation benchmarks demonstrate that the proposed method boosts the performance of current knowledge distillation methods without any significant overhead. The code will be released at FAKD.

1. Introduction

Semantic segmentation aims to assign a semantic label to every pixel in the image. The ability to extract fine-grained information makes segmentation vital for many real-world applications such as autonomous vehicles [1, 21], medical image diagnostics [15, 57], and aerial crop monitoring [35]. With the development of deep learning, semantic segmentation has made tremendous progress in recent years and achieved impressive results on large benchmark datasets [7, 9, 29, 72]. However, advanced segmentation models usually consume large memory footprints and have a low inference speed. This limits the potential for resource-constrained applications such as self-driving cars or assistive navigation robots.

To reduce computing costs, recent research applies knowledge distillation to transfer the knowledge of a cumbersome teacher network to a compact student [17, 24, 38, 51]. Conventional KD for image classification minimizes a Kullback–Leibler (KL) divergence between the output of the teacher and the student network for each example. However, mimicking every pixel's class distribution is inappropriate for semantic segmentation since the noise accumulated from the pixel-level activation may degenerate the performance. Structured knowledge distillation (SKD) [24,25] aggregates a sub-set of different spatial locations according pair-wise relations. IFVD [18] exploits the inter-class relations among different locations. CWD [38] proposes channel-wise knowledge distillation by minimizing the channel distribution, indicating the spatial location for each class. CIRKD [51] focuses on transferring structured pixel-to-pixel and pixel-to-region relations among the whole images. However, due to a low capacity, the student struggles to learn discriminative features captured by a powerful teacher. The lack of training data even exacerbates the performance gap between the student and the teacher. Fig. 1(b) illustrates this problem. A high-capacity teacher

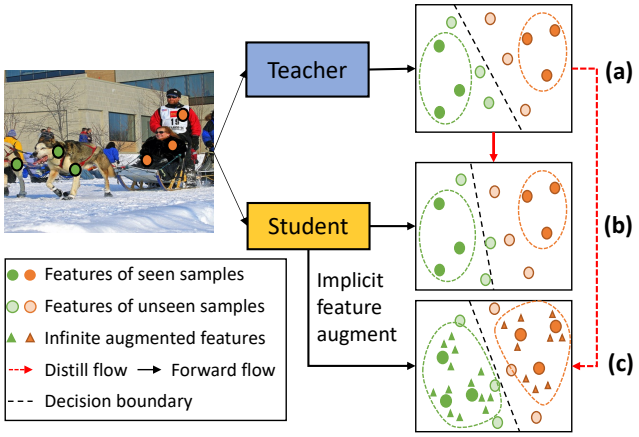


Figure 1. FAKD preserves teacher's boundaries by implicitly augmenting an infinite number of data. (a) The teacher's decision boundary generalizes unseen data. (b) Without sufficient examples, the student fails to preserve the teacher's decision boundary. (c) By implicitly augmenting an infinite number of features, the student can preserve the teacher's decision boundary and generalize on unseen samples.

*Equal contribution
†Corresponding Author

generalizes well on unseen data and discovers a decision boundary well-separating samples of two classes. With an insufficient number of training examples, the low-capacity student may not preserve this decision boundary, thus failing to generalize. Optimization based on a limited number of training examples makes the knowledge transfer more challenging.

To capture the data distribution on unseen data, augmentation is a prevalent strategy to improve the generalization of deep models. Concretely, conventional image-level data augmentation operators, including random crop, random scale, and flip [14], are widely applied in existing methods. However, the number of applicable augmentation operators for each image is limited. This may be insufficient for the student to mimic the diverse patterns of the teacher.

Recent studies [2, 22, 43, 46, 47] discover that deep features are usually linearized. Translating the feature in certain directions can produce features corresponding to the sample of the same class but with different semantics. Compared with image-level augmentation, feature-level augmentation can transform the semantics of the sample such as changing irrelevant appearance or varying the object textures. This work proposes to augment an infinite number of samples via feature augmentation to generate diverse and sufficient data for knowledge distillation. Specifically, we first capture the intra-class variation that contains the semantics of each class. Then we augment examples in feature space based on the feature variance of their classes.

To distill the teacher's knowledge via an infinite number of samples, we propose two novel loss objectives for the standard knowledge distillation [17] and for CWD [38]. Our proposed loss is an upper bound of the expected distillation objectives over an infinite number of examples. By minimizing the proposed loss functions, the student learns to imitate the teacher's outputs via infinite augmented samples, thus improving the student's generalization on unseen data.

Our contributions can be summarized as follows

- We propose a novel feature augmented knowledge distillation (FAKD) method for semantic segmentation. The student mimics the teacher model with infinite augmented samples. To the best of our knowledge, this is the first work that explores feature-level augmentation for knowledge distillation in a semantic segmentation task.
- We derive the upper bound of the expected KL divergence loss for two types of distillation loss, i.e., standard distillation and channel-wise distillation loss. By minimizing the proposed loss objectives, the student can learn a diverse set of teacher patterns directly from infinite samples.
- Experimental results show that our FAKD improves the performance of various student network structures on four benchmark datasets: ADE20K, Pascal Context, Cityscapes, and Pascal VOC.

2. Related Work

Semantic segmentation can be regarded as a pixel-wise classification problem. The fully convolutional network (FCN) [27] is a pioneering work in the field of semantic segmentation, which can adapt to any scale input. To improve the performance of the segmentation network, many researchers improves FCN in different ways. In [5, 6, 33, 52, 56, 58, 67], the receptive field is enlarged to capture more details. In [12, 23, 53, 59, 60, 65, 74], the contextual information is combined to improve the semantic understanding of the model. In [3, 4, 8, 40, 55, 61, 69], the boundary information is considered to boost the segmentation accuracy further. In [10, 20, 45, 68, 71], some new attention modules are designed to enrich the representations of the feature map. Moreover, recent works [48, 50, 70] introduce transformer blocks into the Semantic Segmentation task, which boosts the performance with a large margin. Despite the state-of-the-art performance, segmentation models consume large memory footprint and have low inference speed. The high computational cost limits the application of the segmentation models on resource-limited mobile devices.

Efficient segmentation networks attract attention due to the need for real-time inference. Most of the works designs lightweight networks with low-cost operations. ENet [31] is a efficient segmentation network with early downsampling and lightweight decoder. ESPNet [28] introduces fast spatial pyramid of dilated convolution. ICNet [66] proposes a cascade structure to use low-resolution and high-resolution features. BiSeNet [54] combines a spatial path and a context path to process features efficiently.

Knowledge distillation has been extensively studied in recent years. The core idea of knowledge distillation is to transfer meaningful knowledge from a cumbersome teacher into a smaller and faster student. Most knowledge distillation methods for image classification networks can be categorized into three types: probability-based, feature-based and relation-based methods. Probability-based methods [17, 73] distill the output logits of the teacher network as soft labels to the student. Feature-based knowledge distillation methods [16, 36, 63] focuses on the feature maps. Finally, relation-based KD [30, 32, 34] aligns correlations among multiple instances between the student and teacher networks.

Several methods apply knowledge distillation for semantic segmentation. The pioneering work [24, 25] proposes Structured Knowledge Distillation (SKDS), which enables the output of the student model to transfer the structure knowledge among pixels from the teacher model to the stu-

dent model through pairwise relationships and adversarial training. Intra-class feature variations [49] learns the feature similarities between pixels of the same class to capture the structural knowledge for distillation. Channel-wise distillation proposed in [38] guides the student to mimic the teacher's channel-wise distribution, which indicates the salient image regions for each class. Cross-image relational KD [51] transfers structured pixel-to-pixel and pixel-to-region relations among the whole images. These methods develops complicated objective functions to align useful knowledge between the teacher and the student network. However, when training data is limited, aligning knowledge becomes difficult. Instead of developing sophisticated loss functions, our approach focuses on enriching the data source for effective knowledge distillation via a novel feature augmentation technique.

Data augmentations such as flipping, cropping and rotating are widely adopted techniques are widely applied to encourage the geometric invariance of deep networks [14, 39]. Recent techniques develop strong data augmentations that mix information in different images and appropriately adjust the ground truths. A classic example is Mixup [64] takes data from two images and blends them together in a way that combines each element of the images using a convex combination, as opposed to taking a portion from one image and inserting it into the other. CutMix [62] creates a new training sample by randomly combining two cropped training samples, rather than just blocking out a part of the image. Traditional methods such as color jittering or adding noise used in [39] could generate infinite samples in an *image* space. However, those methods are unaware of semantic information, which could alter the class identity [47]. In contrast, our method proposes to infinitely augment semantically meaningful samples in a *feature* space, preserving the class information.

3. Methods

This section revisits two distillation loss objectives which are the standard pixel-wise distillation (PD) proposed by Hinton et al. [17] and the channel-wise distillation (CWD) [38]. We then describe the proposed feature augmented knowledge distillation (FAKD). We derive the upper bound of the two distillation losses when distilling the knowledge via an infinite number of augmented features.

3.1. Revisit Knowledge Distillation

Hinton et al. [17] originally proposed knowledge distillation for the classification task. They minimize the KL divergence of the predictions between the teacher and the student model. For semantic segmentation, conventional methods [18, 24, 49] perform pixel-wise distillation, which minimizes the divergence of the prediction on every pixel between the student and the teacher. Later on, Liu et al. [24] proposed a channel-wise distillation (CWD), which aligns the output distribution for each channel rather than the pixel.

Let $\mathbf{S} \in \mathbb{R}^{M\times A}$ and $\mathbf{T} \in \mathbb{R}^{M\times A}$ be the A-dimension feature of the student and teacher, where M is the total number of pixels in the input image. The features are fed to a linear classifier to produce the logits $Z \in \mathbb{R}^{M\times C}$ representing the response for C different classes. Let $W = [w_1, ..., w_C] \in \mathbb{R}^{C\times A}$ and $B = [b_1, , .., b_C] \in \mathbb{R}^{C}$ denote the weights and bias of the student's classifier. The student's logit for class c at pixel i is defined as

$$Z^S_{i,c}(\tau) = (w_c^\top S_i + b_c^S). \tag{1}$$

Pixel-wise distillation computes the pixel-wise probability by applying Softmax across the channel dimension. It then minimizes the KL divergence between the teacher's and student's outputs:

$$L_{\texttt{PD}} = -\frac{\tau^2}{M}\sum_{i=1}^{M}\sum_{c=1}^{C}\frac{\exp Z^T_{i,c}}{\sum_{k=1}^{C}\exp Z^T_{i,k}/\tau}\log\frac{\exp Z^S_{i,c}/\tau}{\sum_{k=1}^{C}\exp Z^S_{i,k}/\tau}. \tag{2}$$

Channel-wise distillation [38] computes the channel-wise probability by applying Softmax across the spatial dimension. The KL divergence between the channel-wise probability is formulated as

$$L_{\texttt{CWD}} = -\frac{\tau^2}{M}\sum_{i=1}^{M}\sum_{c=1}^{C}\frac{\exp Z^T_{i,c}/\tau}{\sum_{j=1}^{N}\exp Z^T_{j,c}/\tau}\log\frac{\exp Z^S_{i,c}}{\sum_{j=1}^{N}\exp Z^S_{j,c}/\tau}. \tag{3}$$

where τ is the temperature factor. Note that the main difference between Eq. 3 and Eq. 2 lies in the normalization factor. PD normalizes the logits across the channel dimension, while CWD normalizes across the spatial dimension.

3.2. Feature Augmented Knowledge Distillation

Recent distillation methods develop sophisticated loss functions for effective knowledge transfer [18,24,25]. However, the student network still cannot accurately mimic the teacher's decision boundary due to the low capacity gap. To better learn the teacher's decision boundary, we propose to enrich training data by augmenting samples in the feature space. This section introduces the feature augmentation method and then derives two novel distillation loss objectives for training the student model via an infinite number of augmentations.

Data augmentation in the feature space. Recent methods [2, 22, 43, 46, 47] show that there exist semantic directions in the feature space, such that translating a sample in the feature space along that direction produces a sample of the same class but with different semantics. For example, by changing the latent code of a GAN-based model, we can modify the "style of the hair" for a person. The "style of the hair" is a meaningful direction for the 'person' class. Such semantic distributions can be implicitly represented by

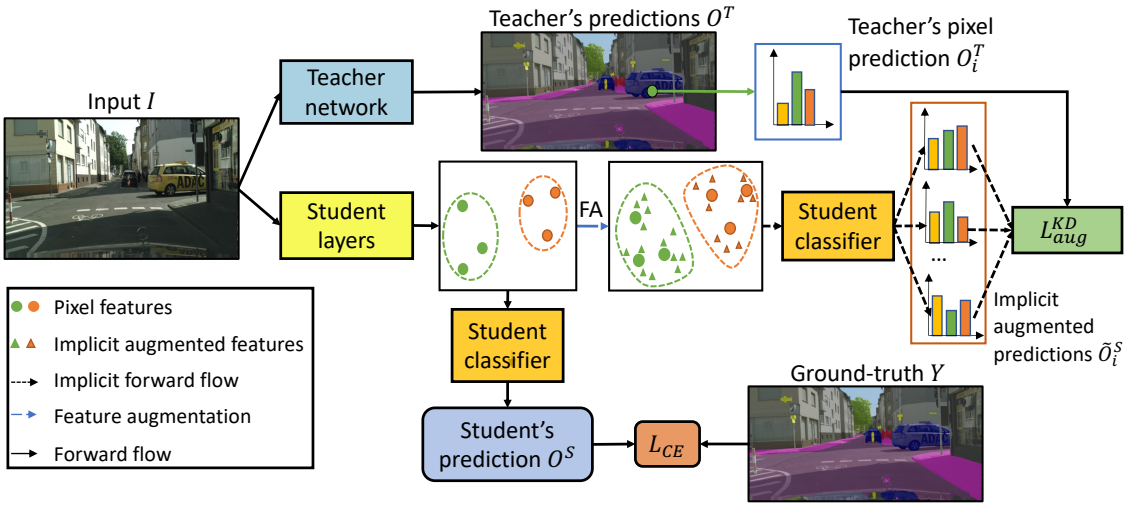


Figure 2. **Overview of the proposed feature augmented knowledge distillation (FAKD)**. The student is trained via the cross-entropy loss from the ground truth (orange box) and the proposed augmentation distillation from the teacher's outputs (green box). FAKD implicitly augments infinite samples in the feature space. To simulate the effects of infinite augmentation, our proposed loss function $L_{\text{aug}}^{\text{KD}}$ optimizes the upper bound of distillation loss between teacher's predictions O^T and student predictions on infinite augmented features $\tilde{O}^S$.

intra-class variations. This inspires us to augment samples by semantically transforming features along those meaningful directions.

Manually searching for useful semantic directions is expensive. To obtain meaningful directions, we sample a random vector from a normal distribution with a zero mean and a covariance proportional to the intra-class covariance. The intra-class covariance captures rich semantic information (e.g., 'hairstyle property') about a particular class (e.g., 'person' class). Thus, we can transform the student's features with the sampled vector to produce semantically meaningful training samples for effective KD of segmentation models.

Let v denote the random vector sampled from a normal distribution $\sim \mathcal{N}(0, \Sigma_{y_i})$. Here, Σ_{y_i} is the class-conditional covariance matrix of all pixel-wise features having the same class y_i as pixel i. Following the method in [47], we compute the intra-class covariance Σ_{y_i} in an online fashion by aggregating pixel-wise feature statistics across all mini-batches. The augmented features $\tilde{S}_i$ is obtained by translating along the direction of v. Equivalently, the augmented features belong to a normal distribution with a mean equal to the original feature S_i:

$$\tilde{S}_i \sim \mathcal{N}(S_i, \lambda\Sigma_{y_i}), \tag{4}$$

where the coefficient λ controls the strength of data augmentation. When $\lambda = 0$, the normal distribution collapses into a Dirac delta distribution, i.e., no augmentation is performed. In the early epochs, the student does not fully capture the informative intra-class variations. Thus, we propose to incrementally increase the strength of data augmentation over time when the student learns more semantic features. Particularly, we apply a cosine schedule to slowly increase λ at the beginning and at the end for stable training:

$$\lambda = \lambda_{\text{end}} - \frac{1}{2}\lambda_{\text{end}}(1 + \cos(\pi t/T)), \tag{5}$$

where t and T are the current training iteration and the total number of iterations; λ_{end} sets the maximum value for λ at the last training iteration.

Knowledge distillation via infinite feature augmentations We perturb the pixel features S_i by N times to create N augmented features $\{\tilde{S}_i^1, ..., \tilde{S}_i^N\}$. The student classifier produces predictions $\tilde{O}_i^S$ from these augmented features. FAKD aligns the teacher's predictions O^T with the student's predictions on augmented features $\tilde{O}_i^S$. By distilling the knowledge via extra augmented data, the student can learn a diverse set of patterns from the teacher. Fig. 2 shows the overview of our proposed FAKD.

We reformulate the CWD loss in Eq. 3 to account for augmented examples. Let $O_{i,c}^T$ denote the channel-wise probability outputs of the teacher, $\tilde{Z}_{i,c}^{S,n}$ denote the logits for the n-th augmented features. The reformulated CWD loss in Eq. 3 is defined as

$$-\frac{\tau^2}{M}\sum_{i=1}^{M}\sum_{c=1}^{C}\frac{1}{N}\sum_{n=1}^{N}O_{i,c}^T \log\frac{\exp \tilde{Z}_{i,c}^{S,n}}{\sum_{j=1}^{M}\exp Z_{j,c}^{S,n}/\tau}. \tag{6}$$

With a large N, the information from the teacher can be leveraged more sufficiently. We consider an extreme case such that N grows to ∞. With infinite augmentations, the expectation over the set of all possible augmentations is optimized instead. Replacing $\tilde{Z}_{i,c}^{S,n} = w_c^\top \tilde{S}_i^n + b$ into Eq. 6 and taking expectation over ∞ augmented samples, the loss

function becomes

$$L_{\text{aug}}^{\text{CWD}} = -\frac{\tau^2}{M}\sum_{i=1}^{M}\sum_{c=1}^{C}\mathbb{E}_{\tilde{S}_i}\left[O_{i,c}^{T}\log\sum_{k=1}^{M}\exp\frac{w_c^{\top}(\tilde{S}_i-\tilde{S}_k)}{\tau}\right]. \tag{7}$$

Computing the loss function as in Eq. 7 is difficult. Hence, we derive an upper bound that is more tractable, given by the following proposition.

Proposition 1. *Suppose that* $\tilde{S}_i \sim \mathcal{N}(s_i, \lambda_i\Sigma_i)$*, we have*

$$L_{\text{aug}}^{\text{CWD}} \leq \frac{\tau^2}{C}\sum_{i=1}^{M}\sum_{c=1}^{C}O_{i,c}^{T} \log\left\{\sum_{k=1}^{M}\exp\left[\frac{w_c^{\top}(S_i-S_k)}{\tau}+\frac{w_c^{\top}(\lambda_c\Sigma_i+\lambda_k\Sigma_k)w_c}{2\tau}\right]\right\}. \tag{8}$$

The supplementary material provides detailed proof for Proposition 1. By minimizing the tractable surrogate loss in Eq. 8, the student learns to mimic the teacher via an infinite number of implicit augmentations.

With the same process, we derive the upper bound for pixel-wise distillation as follows.

$$L_{\text{aug}}^{\text{PD}} \leq \frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{C}O_{i,c}^{T} \log\left[\sum_{k=1}^{C}\exp(\Delta w_k S_i + b_k - b_c + \frac{\lambda}{2}\Delta w_k\Sigma_c\Delta w_k)\right], \tag{9}$$

where $\Delta w_k = w_k^{\top} - w_c^{\top}$. The proposed FAKD method is summarized in the Algorithm 1. Given a training set D and the coefficient λ for controlling augmentation's strength, we initialize the covariance matrix Σ for all classes with zero. For each mini-batch, we feed-forward training samples to obtain features, logits, and outputs. We update the intra-class covariance matrix Σ by aggregating features S within each class. The coefficient λ incrementally updates based on the current training step. Finally, parameters of the student model will be optimized with the corresponding loss.

Algorithm 1 The FAKD framework

Teacher: Teacher network with extraction function f_t and classification function g_t;
Student: Student network with extraction function f_s and classification function g_s;
Input: Training image dataset D, coefficient λ and a covariance matrix Σ initialized with zero;
for $step = 1, ..., n_{steps}$ **do**
  X, Y=Sample(D),
  $O^T = g_t(f_t(X))$,
  $S = f_s(X)$,
  $O^S = g_s(S)$,
  update coefficient λ with current $step$,
  update Σ with S, Y and $step$,
  compute augmentation loss L_{aug} based on Eq. 8 or. 9,
  final loss $= L_{\text{CE}}(O^s, Y) + L_{\text{aug}}$
  minimize the loss and update student parameters θ_s;
end

4. Experimental Results

4.1. Experimental Setup

Dataset. We employ four popular semantic segmentation datasets to conduct our experiments. ADE20K [72] contains 20k/2k/3k images for train/val/test with 150 semantic classes. Cityscapes [7] is an urban scene parsing dataset that contains 2975/500/1525 finely annotated images used for train/val/test. And the performance is evaluated on 19 classes. Pascal Context [29] provides dense annotations which contains 4998/5105/9637 train/val/test images. We use 59 object categories for training and testing. Pascal VOC contains 21 classes including 20 object categories and one background class. Following the procedure of [6, 67] , we use augmented data with annotation of resulting 10582, 1449, and 1456 images for train/val. Our results are all reported on the validation set.

Evaluation metrics. For quantitative evaluation, we report the mean intersection over union (mIoU) and pixel accuracy (mAcc), which are standard metrics for segmentation evaluation. For qualitative evaluation, we visualize the segmentation results of all methods and present them in the supplementary materials.

Network architectures. On each dataset, the same series of teacher models and student models are used.

Implementation details. Following the standard data augmentation, we employ random flipping, cropping and scaling in the range of [0.5, 2]. All experiments are optimized by SGD with a momentum of 0.9, a batch size of 16, and 512×512 crop size. We use an initial learning rate of 0.01 for ADE20K, Cityscapes, and Pascal VOC. In addition, we use an initial learning rate of 0.004 for Pascal Context. The number of total training iterations is 40K. Following previous methods [6, 67], we use the poly learning rate policy where the current learning rate equals the base one multiplying $(1 - \frac{iter}{max_{iter}})^{0.9}$.

Baseline methods. On each dataset, we compare with state-of-the-art segmentation distillation methods including SKDS [24], IFVD [18] and CWD [38], CIRKD [51]. We re-run SKDS, IFVD, CWD and FAKD on 4 NVIDIA V100 GPUs. Due to the limitation of GPU memory, we re-run

Table 1. Performance comparison with state-of-the-art distillation methods over various student and teacher CNN based segmentation networks on ADE20K. All of the students' backbones except that with † are initialized with the pre-trained weights on ImageNet classification.

Teacher	PSPNet		PSPNet		HRNet		DeeplabV3Plus		ISANet	
Student	PSPNet-R18		PSPNet-18†		HRNet18s		Deeplab-MV2		ISANet-R18	
Methods	mIoU	mAcc(%)	mIoU	mAcc(%)	mIoU	mAcc(%)	mIoU	mAcc(%)	mIoU	mAcc(%)
Teacher	44.39	54.74	44.39	54.74	42.02	53.52	45.47	56.41	43.80	54.39
Student	29.42	38.48	17.11	22.99	28.69	37.86	22.38	31.71	27.68	36.92
SKDS	31.80	42.25	20.79	27.74	30.49	40.19	24.65	35.07	28.70	38.51
IFVD	32.15	42.53	20.75	27.60	30.57	40.42	24.53	35.13	29.66	39.80
CIRKD	32.25	43.02	22.90	30.68	31.34	41.45	25.21	36.17	29.79	40.48
CWD	33.82	42.41	25.14	34.13	31.36	39.68	26.89	35.79	34.69	43.05
Ours	**35.30**	**44.06**	**26.96**	**34.13**	**32.46**	**41.92**	**28.29**	**38.31**	**35.74**	**44.55**

Table 2. Performance comparison of different distillation methods over various Transformer architectures on ADE20K.

Teacher	Swin-Base		Segformer-B3		DeiT-Base	
Student	Swin-Tiny		Segformer-B0		DeiT-Tiny	
Method	mIoU(%)	mAcc(%)	mIoU(%)	mAcc(%)	mIoU(%)	mAcc(%)
Teacher	50.13	61.39	50.00	60.02	48.80	64.04
Student	43.57	51.70	38.00	48.76	41.10	49.10
SKDS	43.58	53.63	38.43	49.38	41.90	51.13
IFVD	43.75	54.03	38.72	49.65	41.16	50.79
CIRKD	43.32	54.10	38.88	49.70	41.64	52.66
CWD	44.99	56.01	38.91	49.22	44.25	53.24
Ours	**45.64**	**58.38**	**39.52**	**50.29**	**45.64**	**57.36**

CIRKD on 4 NVIDIA A100.

4.2. Comparisons with State-of-the-art Methods

ADE20K. For a fair comparison, we follow the experimental settings in [24, 38, 49, 51] and apply our FAKD to standard compact networks: PSPNet with ResNet18 backbone [67], HRNet-W18 [44], Deeplab-v3+ with MobileNetV2 backbone [37], ISANet with ResNet18 [19]. Table 1 presents the comparisons between the proposed FAKD with other state-of-the-art methods on various network structures. The proposed FAKD consistently boosts the performance of the student by 19.99%, 57.57%, 13.14%, 26.41%, and 29.12% on 5 different network structures. Compared to state-of-the-art (SOTA) distillation methods, our FAKD significantly outperforms the second-best method, i.e., channel-wise distillation (CWD) [38] by 4.38%, 7.24%, 13.14%, 26.41%, and 29.12% on all 5 network structures. Additionally, our method drastically improves the mIoU of PSPnet-R18† by 57.5% when the student model has not been pre-trained on the large-scale dataset, i.e., ImageNet.

We conduct experiments to benchmark distillation performance on three Transformer architectures: Swin Transformer [26], Segformer [50] and DeiT [42]. Table 2 shows that our FAKD consistently outperforms recent distillation methods on all Transformer models. Notably, our FAKD significantly increases mIoU of CWD and CIRKD by 3.14% and 9.61% when applied to DeiT, highlighting our method's effectiveness in state-of-the-arts Transformer models. Current SOTA methods introduces sophisticated distillation objectives; yet, low-capacity students struggle to mimic the classification ability of powerful teachers. To relieve the capacity gap problem, our method provides the student with a rich set of augmented data, enabling effective learning from a diverse set of teacher's feature patterns.

Pascal Context. Table 3 summarizes our results on Pascal Context validation set. Our method achieves the best performance among all settings. It surpasses the competing method CWD [38] from 0.5% to 2% improvement on PSPnet-R18, HRNet18s, and Deeplab-MV2.

Table 3. Performance comparison with state-of-the-art distillation methods over various student and teacher segmentation networks on Pascal Context.

Teacher	PSPNet		HRNet		DeeplabV3Plus	
Student	PSPNet-R18		PSPNet18s		Deeplab-MV2	
Methods	mIoU	mAcc(%)	mIoU	mAcc(%)	mIoU	mAcc(%)
Teacher	52.47	63.15	51.12	61.39	53.20	64.04
Student	43.07	53.79	40.82	51.70	37.16	49.10
SKDS	43.93	54.01	42.91	53.63	39.18	51.13
IFVD	44.75	54.99	43.12	54.03	38.80	50.79
CIRKD	44.83	55.30	43.45	54.10	39.99	52.66
CWD	45.92	55.55	45.50	56.01	42.52	53.24
Ours	**46.37**	**56.39**	**45.61**	**56.13**	**43.44**	**54.29**

Cityscapes. Table 4 presents the comparative results on the Cityscapes dataset. Training via FAKD, the student's mIoU improves by 8.35%, 2.93%, 5.09%, 1.39%, respectively. The proposed FAKD achieves the best performance across various student networks. Particularly, our FAKD consistently outperforms CWD by up to 1.15% across all network structures. This shows that our method enhances the student's performance from natural scenes, i.e., ADE20k to road scenes, i.e., Cityscapes.

Pascal VOC. In Table 4, we show the segmentation performance of various distillation methods on Pascal VOC. The proposed FAKD consistently outperforms state-of-the-art KD methods.

Table 4. Performance comparison with state-of-the-art distillation methods over various student and teacher segmentation networks on Cityscapes dataset and Pascal VOC dataset.

Dataset	Teacher	PSPNet		HRNet		DeeplabV3Plus		ISNet	
	Student	PSPNet-R18		HRNet18s		Deeplab-MV2		ISANet-R18	
	Methods	mIoU	mAcc(%)	mIoU	mAcc(%)	mIoU	mAcc(%)	mIoU	mAcc(%)
Cityscapes	Teacher	79.74	86.56	80.65	87.39	80.98	88.70	80.61	88.29
	Student	68.99	75.19	73.77	82.89	70.49	80.11	71.45	78.65
	SKDS	69.33	75.37	74.75	83.23	70.81	79.31	70.65	77.53
	IFVD	71.08	77.46	75.33	83.83	71.82	80.88	70.30	77.79
	CIRKD	72.23	78.79	74.63	83.72	72.39	81.84	72.00	79.32
	CWD	74.29	80.95	75.54	84.08	73.35	82.41	71.61	80.02
	Ours	**74.75**	**82.00**	**75.93**	**84.75**	**74.08**	**83.83**	**72.44**	**81.04**
Pascal Context	Teacher	78.52	86.11	76.24	84.95	78.62	86.55	78.46	87.33
	Student	70.52	81.04	67.47	78.99	62.56	80.09	68.71	79.81
	SKDS	70.35	80.22	67.58	79.10	62.85	80.25	67.86	80.47
	IFVD	70.92	81.31	67.50	78.89	67.50	78.89	69.11	80.93
	CIRKD	70.13	80.24	67.36	79.22	63.57	79.63	69.0	80.83
	CWD	73.36	82.63	68.39	79.78	67.61	82.03	72.83	83.99
	Ours	**73.97**	**82.96**	**69.04**	**80.37**	**67.62**	**82.13**	**73.17**	**84.25**

4.3. Ablation Study

We conduct experiments to explore the effectiveness of our methods under different knowledge distillation settings. All ablation experiments are carried out on the ADE20k dataset, which is a standard benchmark for knowledge distillation. We choose PSPnet-R101 as the teacher and PSPnet-R18 as the student.

Effectiveness of augmentation loss. We analyze the effectiveness of the proposed augmentation loss objective by comparing the non-augmentation loss objectives, i.e., pixel distillation (PD) [17] and channel-wise distillation (CWD) [38] with their augmentation version proposed in ours, i.e., $L_{\texttt{PD}}^{\texttt{aug}}$, $L_{\texttt{CWD}}^{\texttt{aug}}$. Table 5 shows that our proposed $L_{\texttt{PD}}^{\texttt{aug}}$ and $L_{\texttt{CWD}}^{\texttt{aug}}$ improve upon the non-augmentation loss PD and CWD by 4.38% and 3.43%, respectively. We hypothesize that augmenting infinite samples as in our FAKD allows the student to preserve the decision boundary of the teacher model.

Table 5. Ablation study of different equations on ADE20K. Introducing infinite samples in ℓ_{PD} and ℓ_{CWD} both bring improvements.

Methods	mIoU	mAcc(%)
ℓ_{PD}	31.75	42.23
ℓ_{aug}^{PD}	32.84	42.33
ℓ_{CWD}	33.82	42.41
ℓ_{aug}^{CWD}	35.30	44.06

Analysis of augmentation strength coefficient λ. Our augmentation loss $L_{CWD}^{\texttt{aug}}$ has coefficient λ which controls the strength of augmentation. Here, λ starts with zero and incrementally increase to the maximum value λ_{end} via a cosine schedule. As shown in Table 6, we investigate the impact of the maximum λ in our FAKD. We find that $\lambda_{\text{end}} = 1.0$ is the best choice.

Table 6. Experiment for different λ_{end}, which is the maximum value for coefficient λ controlling the augmentation strength in our proposed augmentation loss.

λ_{end}	mIoU	mAcc(%)
0.5	35.06	43.91
1.0	35.30	44.06
1.5	34.63	42.97
2.5	31.95	39.04

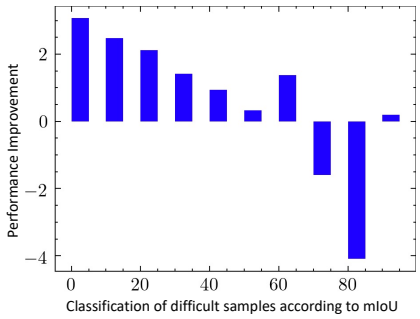


Figure 3. Performance improvement on different classes sorted by difficulty. The difficulty is ranked by the model performance, i.e., the lower the mIoU, the harder the class. The x-axis represents the difficulty. The y-axis represents the improvement of our FAKD over CWD.

Analysis of FAKD on hard examples. This section analyzes the efficiency of FAKD on hard examples. The sample's difficulty is determined by the model performance, i.e., the lower the mIoU, the harder the example. Fig. 3 analyzes the improvement of our FAKD upon CWD [38] on each class. Our method significantly improves the student's

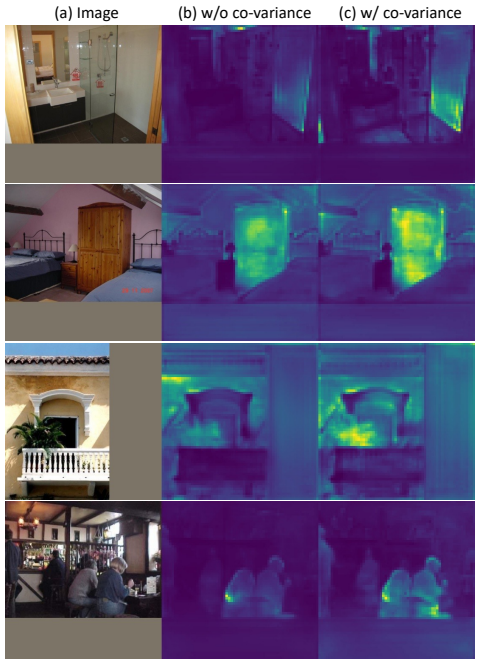
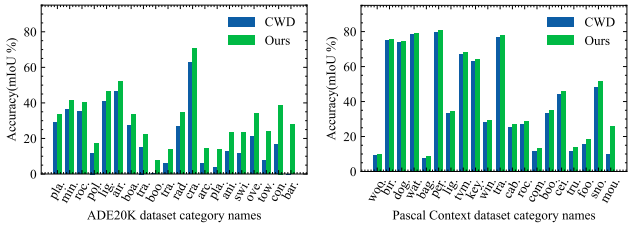


Figure 4. Visual samples of model's features with or without co-variance. By implicitly augmenting the feature with co-variance, the activation of the meaningful region is strengthened.

performance on difficult samples whose classification rates are less than 40% mIoU.



(a) ADE20k (b) Pascal Context

Figure 5. Per-class mIoU of CWD and our FAKD. Our FAKD improves more significantly on hard classes.

We also visualized the feature maps as shown in Fig. 4. Our proposed FAKD better highlights problematic regions. This show that augmenting data allows the student to focus more on learning hard pixels. Per-class IoU scores are shown in Figure 5.

Comparisons with image-level augmentations. We analyze the efficiency of our feature augmentation (FA) compared with image augmentation, as shown in Table 7. All methods apply standard weak data augmentations such as normalization, random flip, and padding. Image augmentation (IA) applies strong data augmentations (SDA) including color jittering, blurring, equalizing, and random grayscaling, which are widely used for learning discriminative representations [11, 13, 41]. Table 7 shows that our proposed FA is more efficient than IA. IA only has a limited number of operations. This leads to the sparse distribution of augmented samples, which causes difficulty for the student to learn. In contrast, our proposed FA produces an infinite number of samples. This creates the dense distribution of augmented data is dense, thus allowing the student to discover the teacher's diverse patterns.

Table 7. Performance comparison of different augmentation techniques for knowledge distillation. WDA: image-level weak data augmentation. SDA: image-level strong data augmentation. FA: the proposed feature augmentation.

Methods	WDA	SDA	FA	mIoU	mAcc(%)
CWD	✓	-	-	33.53	41.71
CWD	✓	✓	-	33.32	41.60
CWD (Ours)	✓	-	✓	**35.32**	**44.44**
CWD	✓	✓	✓	34.64	43.76

Effectiveness of FAKD under limited training data. We analyze the efficiency of the proposed data augmentation strategy for distillation under limited training data on ADE20K. Fig. 6 shows the correlation between student performance and the percentage of data to be used for training. The strong augmentation has incremental improvement when using less than 30% training data, but will have negative impacts when more training samples are avaiable. Our proposed FAKD is more effective than image-level augmentations under various limited data splits.

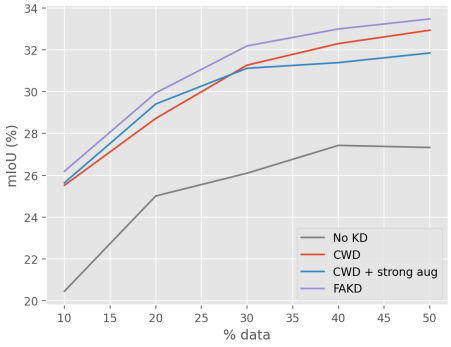


Figure 6. Performance of different methods under limited data.

5. Conclusion

This paper presents a novel feature augmented knowledge distillation (FAKD) method for semantic segmentation which increases the training samples for the student network in the feature space. Extensive experiments on four public segmentation datasets show the effectiveness of our FAKD for different network structures. Our method demonstrates superiority in long-tail problems. The performance for classes with fewer training samples has been improved by a large margin.

References

[1] Jose Manuel Alvarez, Theo Gevers, Yann LeCun, and Antonio M. Lopez. Road scene segmentation from a single image. In *European Conference on Computer Vision*, 2012. 1

[2] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *Proc. Int. Conf. Mach. Learn.*, pages 552–560. PMLR, 2013. 2, 3

[3] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Semantic segmentation with boundary neural fields. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3602–3610, 2016. 2

[4] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4545–4554, 2016. 2

[5] 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 TPAMI*, 40(4):834–848, 2017. 2

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. Eur. Conf. Comput. Vis.*, pages 801–818, 2018. 2, 5

[7] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3213–3223, 2016. 1, 5

[8] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *Proc. Int. Conf. Comput. Vis.*, pages 6819–6829, 2019. 2

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, June 2010. 1

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

[11] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Learning representations by predicting bags of visual words. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6928–6938, 2020. 8

[12] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7519–7528, 2019. 2

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9729–9738, 2020. 8

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 2, 3

[15] Yufan He, Dong Yang, Holger Roth, Can Zhao, and Daguang Xu. DiNTS: Differentiable neural network topology search for 3d medical image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5841–5850, 2021. 1

[16] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proc. Int. Conf. Comput. Vis.*, pages 1921–1930, 2019. 2

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3, 7

[18] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-region affinity distillation for road marking segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12486–12495, 2020. 1, 3, 5

[19] Lang Huang, Yuhui Yuan, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Interlaced sparse self-attention for semantic segmentation. *arXiv preprint arXiv:1907.12273*, 2019. 6

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

[21] Joel Janai, Fatma Guney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. *Foundations and Trends in Computer Graphics and Vision*, 12:1–308, 2020. 1

[22] Mu Li, Wangmeng Zuo, and David Zhang. Convolutional network for attribute-driven and identity-preserving human face generation. *arXiv preprint arXiv:1608.06434*, 2016. 2, 3

[23] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1925–1934, 2017. 2

[24] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2604–2613, 2019. 1, 2, 3, 5, 6

[25] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE TPAMI*, 2020. 1, 2, 3

[26] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10012–10022, 2021. 6

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3431–3440, 2015. 2

[28] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proc. Eur. Conf. Comput. Vis.*, pages 552–568, 2018. 2

[29] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 1, 5

[30] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3967–3976, 2019. 2

[31] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet a deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016. 2

[32] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proc. Int. Conf. Comput. Vis.*, pages 5007–5016, 2019. 2

[33] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4353–4361, 2017. 2

[34] Qi Qian, Hao Li, and Juhua Hu. Improved knowledge distillation via full kernel matrix transfer. In *Proc. SIAM Int. Conf. on Data Mining*, pages 612–620. SIAM, 2022. 2

[35] Ciaran Robb, Andy Hardy, John H Doonan, and Jason Brook. Semi-automated field plot segmentation from uas imagery for experimental agriculture. *Frontiers in Plant Science*, 11:591886, 2020. 1

[36] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4510–4520, 2018. 6

[38] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proc. Int. Conf. Comput. Vis.*, pages 5311–5320, 2021. 1, 2, 3, 5, 6, 7

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Represent.*, 2015. 3

[40] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proc. Int. Conf. Comput. Vis.*, pages 5229–5238, 2019. 2

[41] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020. 8

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 6

[43] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7064–7073, 2017. 2, 3

[44] 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. 6

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7794–7803, 2018. 2

[46] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing deep networks with semantic data augmentation. *IEEE TPAMI*, 2021. 2, 3

[47] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *Proc. Adv. Neural Inform. Process. Syst.*, 32:12635–12644, 2019. 2, 3, 4

[48] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8741–8750, 2021. 2

[49] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *Proc. Eur. Conf. Comput. Vis.*, pages 346–362. Springer, 2020. 3, 6

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Proc. Adv. Neural Inform. Process. Syst.*, 2021. 2, 6

[51] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12319–12328, 2022. 1, 3, 5, 6

[52] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3684–3692, 2018. 2

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

[54] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proc. Eur. Conf. Comput. Vis.*, pages 325–341, 2018. 2

[55] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1857–1866, 2018. 2

[56] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[57] Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L Yuille, and Daguang Xu. C2fnas: Coarse-to-

fine neural architecture search for 3d medical image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4126–4135, 2020. 1

[58] Jianlong Yuan, Zelu Deng, Shu Wang, and Zhenbo Luo. Multi receptive field network for semantic segmentation. In *Proc. Winter Conf. Appl. Comput. Vis.*, pages 1883–1892. IEEE, 2020. 2

[59] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. 2

[60] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 2

[61] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision*, pages 489–506. Springer, 2020. 2

[62] 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 *Proc. Int. Conf. Comput. Vis.*, pages 6023–6032, 2019. 3

[63] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 2

[64] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. Int. Conf. Learn. Represent.*, 2018. 3

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

[66] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proc. Eur. Conf. Comput. Vis.*, pages 405–420, 2018. 2

[67] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2881–2890, 2017. 2, 5, 6

[68] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proc. Eur. Conf. Comput. Vis.*, pages 267–283, 2018. 2

[69] Mingmin Zhen, Jinglu Wang, Lei Zhou, Shiwei Li, Tianwei Shen, Jiaxiang Shang, Tian Fang, and Long Quan. Joint semantic segmentation and boundary detection using iterative pyramid contexts. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13666–13675, 2020. 2

[70] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6881–6890, 2021. 2

[71] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13065–13074, 2020. 2

[72] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 633–641, 2017. 1, 5

[73] Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. *arXiv preprint arXiv:2102.00650*, 2021. 2

[74] Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Context-reinforced semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4046–4055, 2019. 2