BPKD: Boundary Privileged Knowledge Distillation For Semantic Segmentation

Liyang Liu [1] Zihan Wang [2 3] Minh Hieu Phan [1] Bowen Zhang [1] Jinchao Ge [1] Yifan Liu [1*]

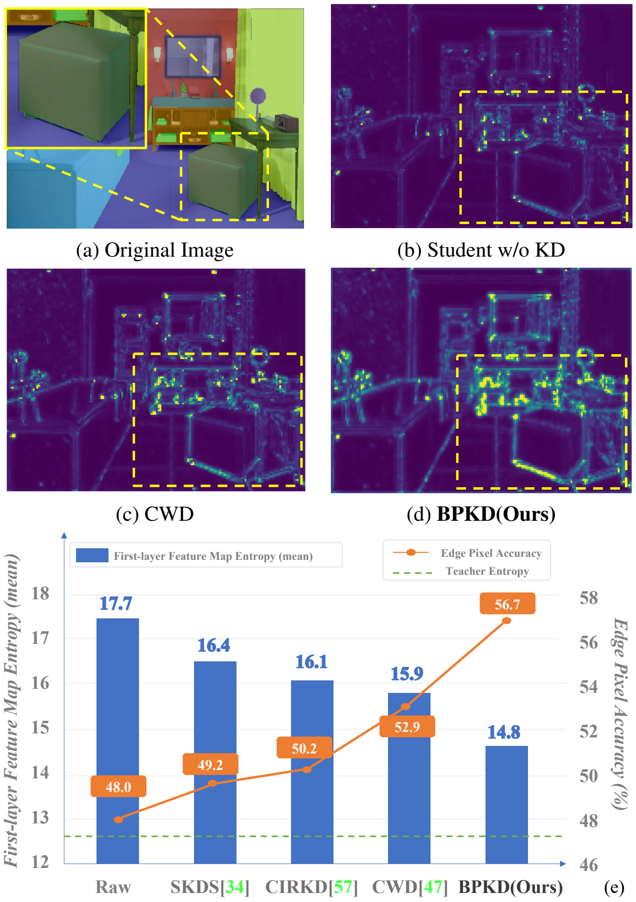
[1] The University of Adelaide, Australia [2] The University of Queensland, Australia
[3] CSIRO's Data61, Australia

Abstract

Current knowledge distillation approaches in semantic segmentation tend to adopt a holistic approach that treats all spatial locations equally. However, for dense prediction, students' predictions on edge regions are highly uncertain due to contextual information leakage, requiring higher spatial sensitivity knowledge than the body regions. To address this challenge, this paper proposes a novel approach called boundary-privileged knowledge distillation (BPKD). BPKD distills the knowledge of the teacher model's body and edges separately to the compact student model. Specifically, we employ two distinct loss functions: (i) edge loss, which aims to distinguish between ambiguous classes at the pixel level in edge regions; (ii) body loss, which utilizes shape constraints and selectively attends to the inner-semantic regions. Our experiments demonstrate that the proposed BPKD method provides extensive refinements and aggregation for edge and body regions. Additionally, the method achieves state-of-the-art distillation performance for semantic segmentation on three popular benchmark datasets, highlighting its effectiveness and generalization ability. BPKD shows consistent improvements across a diverse array of lightweight segmentation structures, including both CNNs and transformers, underscoring its architecture-agnostic adaptability. The code is available at `https://github.com/AkideLiu/BPKD`.

1. Introduction

Semantic segmentation is a complex computer vision task that involves assigning unique categories to each pixel of an input image. In recent years, deep learning models with large numbers of parameters have achieved remarkable performance in semantic segmentation [9, 10, 18, 38, 63, 64, 66]. However, such models are impractical for resource-constrained devices like mobile devices and robotics due to their high computational complexity [51, 55, 61]. To address


(a) Original Image

(b) Student w/o KD

(c) CWD

(d) **BPKD(Ours)**

Figure 1. Illustration of contextual information leakage. Above) The uncertainty maps are generated by computing the mean entropy from first-layer feature maps, employing different distillation strategies such as the raw student, CWD, and BPKD. Brighter colors indicate higher certainty. Below) Demonstrates the inverse correlation between mean entropy and pixel-level accuracy along edges. When learning edge pixels, the model aggregates contextual information from adjacent class categories. Previous whole-view distillation methods are affected by contextual information leakage, leading to high uncertainty in edge low-level features and low edge pixel accuracy.

[*] Corresponding author. Contact Email: akide.liu@adelaide.edu.au.

this issue, lightweight base models such as MobileNet [22], ShuffleNet [40], and EfficientNet [48] have been used for real-time semantic segmentation.

Designing compression and acceleration techniques for compact networks is challenging but crucial. Knowledge distillation approaches, such as those introduced in [19, 20, 28, 62], train a smaller student network to mimic the complex teacher network by minimizing the soft probabilities distance, typically measured by Kullback–Leibler (KL) divergence, between the student and teacher. In [31, 59, 65], authors have attempted to distill hidden knowledge by utilizing network and data relations, with a focus on classification tasks, achieving impressive results.

Pioneering knowledge distillation methods for semantic segmentation [34, 47, 53, 57] focus more on capturing the correlational information among pixels, channels, and images. Liu et al. [34] suggest that hidden knowledge in semantic segmentation is constructed by structured representation. Structured knowledge is more suitable for pair-wise similarity reduction and holistic distillation. IFVD [53] proposed to encode the knowledge according to the semantic masks. In CWD [47], authors refine distillation by emphasizing aligning the most salient region of each channel between the teacher and student.

In comparison to prior studies, which encompassed a variety of studies including [3, 26, 34, 47, 53, 57, 58], that predominantly concentrated on transferring knowledge representations across the entire image, the importance of distinct knowledge representations at different spatial locations has been neglected. When learning edge features, the model aggregates contextual information between adjacent class categories, leading to *contextual information leakage*. As shown in Fig. 1 b, c, d), current whole-view distillation methods exhibit high levels of uncertainty, as well as higher levels of entropy, at edge regions. Following prior works [1], we quantify the uncertainty by computing a mean entropy of first-layer features, capturing low-level textual representations. Fig. 1 e) shows that current methods suffer from higher uncertainties and lower accuracy at edge pixels, indicating the phenomenon of contextual information leakage. The low capacity of compact student networks further exacerbates this phenomenon, degrading segmentation details on the boundaries, especially for small object segmentation. However, delineating object's boundaries is mandatory for real-life applications such as localizing road boundaries for autonomous navigation [42] or segmenting tumors for treatment planning [33].

To tackle the issue of contextual information learning in existing methods, we propose a novel approach, termed Boundary Privileged Knowledge Distillation (BPKD). We divide the knowledge distillation process into two subsections: the edge distillation and the body distillation sections. Our proposed BPKD approach explicitly enhances the quality of edge regions and object boundaries by decoupling knowledge distillation and using teacher soft labels. The edge distillation loss involves spatial probability alignment and aggregation of contextual information to refine the boundaries. Furthermore, boundaries provide prior knowledge of the shape of an object's inner regions, and the body region can exploit this knowledge to eliminate high-uncertainty boundary samples and smooth the learning curves. Consequently, we observed that the object center received greater attention due to the implicit shape constraints, further improving segmentation in the body area.

Through empirical analysis, we have demonstrated that our proposed approach effectively guides the student network to learn from the teacher network's knowledge, resulting in improved segmentation performance. We evaluate our method over popular architectures on three segmentation benchmark datasets: Cityscapes [13], ADE20K [68], and Pascal Context [15]. Experimental results indicate that BPKD outperforms other state-of-the-art distillation approaches. Specifically, we reduce the disparity in performance between the student and teacher networks and exhibit competitive results in comparison to specialized real-time segmentation methods [16].

Our main contributions are summarized as follows:

- We show that current distillation methods suffer from contextual information leakage problems by analysing low-level feature uncertainty, leading to non-optimal segmentation performance at boundaries. To the best of our knowledge, this is the first paper identifying this critical problem within knowledge distillation literature for semantic segmentation.
- We propose a novel knowledge distillation method that separately focuses on distilling information related to the body and edge of objects. Our specialized edge loss function significantly enhances the quality of edge slices, while simultaneously imposing strong shape constraints on the body regions. This approach effectively minimizes the uncertainty prevents contextual information leakage in the distillation process and amplifies the focus on the inner region.
- Our method achieves state-of-the-art results on three popular benchmark datasets. We report an increase in the mean Intersection over Union (mIoU) by up to 4.02% when compared to the previous SOTA CWD. Additionally, we observe a remarkable enhancement in prediction quality in both edge and body regions, further demonstrating our effectiveness.

2. Related Work

Semantic Segmentation. Recent state-of-the-art approaches in semantic segmentation primarily leverage Fully Convolutional Networks (FCNs) [32, 39, 64]. Notable models like PSPNet [67] and DeepLab series [4–7] employ advanced techniques such as pyramid pooling modules (PPM) and atrous spatial pyramid pooling (ASPP) to capture multi-scale contexts. HRNet [51] further innovates with a parallel backbone for high-resolution feature maintenance. Despite their performance, these models are computationally intensive, limiting their applicability in real-time and edge-device scenarios. Consequently, lightweight models like ENet [43], SqueezeNet [25], and ESPNet [41] have gained traction. These models use strategies such as early down-sampling, filter factorization, and efficient spatial pyramids to reduce computational overhead. MobileNet variants [22,23,46] are also effective for efficient segmentation.

Edge Detection. Classical edge detection algorithms like Canny [2], Sobel [27], and Prewitt [44] have been retrofitted into modern deep learning architectures to achieve fine-grained segmentation. Deeply-supervised edge detection methods such as HED [56] and RCF [36] introduce multi-scale edge information directly into the segmentation pipelines. Similarly, models like CASENet [60] have advanced the state-of-the-art by fusing class-specific edges into segmentation algorithms, providing a dual benefit of detailed boundary representation and class differentiation. In parallel, techniques like edge-attention models [52] incorporate edge information by weighting features based on their boundary importance, enabling finer contour mapping in semantic segmentation.

Knowledge Distillation. Knowledge distillation (KD) aims to condense the learnings from one or more expansive teacher models into a streamlined student model [17, 20]. Predominantly employed in basic vision tasks, KD techniques can be taxonomized into response-based, feature-based, and relation-based paradigms. Response-based methods, chiefly initiated by Hinton *et al.* [20], minimize Kullback-Leibler divergence to convey implicit, high-value knowledge [19, 28, 62]. Feature-based approaches like FitNet [45] align internal feature activations between teacher and student, whereas relation-based methods [31, 59, 65] delve into inter-layer or inter-sample relationships. Nonetheless, traditional KD is largely skewed towards image classification, offering limited utility in pixel-level segmentation tasks.

Recent advancements have seen KD methods tailor-fit for semantic segmentation. Strategies such as structural knowledge distillation [34, 35] define segmentation as a structured prediction task, employing pair-wise similarities and holistic adversarial enhancements for knowledge transfer. Channel-wise distillation [47] concentrates on salient channel regions. Additional innovations include intra-class feature variation distillation [53], which amalgamates pixel-level and class-wise variation, and Cross-Image Relational distillation [57], which optimizes global semantic interconnections. Masked Generative Distillation [58] leverages teacher guidance for feature recovery. Recent work [3] suggests Pearson correlation as a viable KL divergence alternative. Empirical validation corroborates the efficacy of these specialized KD techniques in boosting semantic segmentation performance.

3. Methods

In this section, we first provide an overview of the workflow of the Boundary Privileged Knowledge Distillation (BPKD) framework (Section 3.1), followed by a detailed description of two key implementation aspects of our approach. Specifically, Section 3.2 outlines the edge knowledge distillation process, which involves pre-mask filtering and post-mask filtering. Section 3.3 introduces the distillation loss for body enhancements.

3.1. BPKD Framework

Existing feature distillation techniques [47, 57] transfer the whole-view representations from the teacher while overlooking the effects of noisy edge features on the distillation process. In this framework, we carefully consider the sensitivity of edge representations and introduce the novel boundary-privileged knowledge distillation (BPKD) that transfers the knowledge in the body and the edge regions separately. Distilling edge regions individually enhances the quality of object boundaries explicitly. Furthermore, the edge distillation loss provides prior shape knowledge for the object's inner regions. For instance, given a vehicle's boundary constraint, the model can easily determine pixel categories for its inner region. The body distillation loss has two key benefits from the prior boundary knowledge: (1) reducing learning difficulty by mimicking the teacher's logit probability distribution since high-uncertainty boundaries are removed, and (2) leveraging higher attention on the object center through implicit shape constraints.

Our approach uses an edge detection technique to generate edge masks M_E for each class by processing the ground truth and the segmentation logit map. Let $\mathcal{Z} \in \mathbb{R}^{H \times W \times C}$ denote a network's logit map, where C corresponds to the number of channels and $H \times W$ represents the spatial resolution. The edge masks M_E are applied to separate the logit map $\mathcal{Z}$ into two components: the body component $\mathcal{Z}_B$ and the edge component $\mathcal{Z}_E$, which adhere to an additive rule, denoted by $\mathcal{Z} = \mathcal{Z}_B + \mathcal{Z}_E$. Our BPKD framework separately transfers the edge and body knowledge encoded in these two components to the student. As the edge slices have less amount of knowledge representations, we introduce a categorized awareness to balance the importance of different special perspectives. Together, these techniques play a crucial role in improving the overall performance of the model.

In this study, we propose a novel approach for decomposing the distillation loss into two distinct components, namely the body loss ℓ_B and the edge loss ℓ_E, as expressed by Equation 1. We include the body loss weight λ_b and

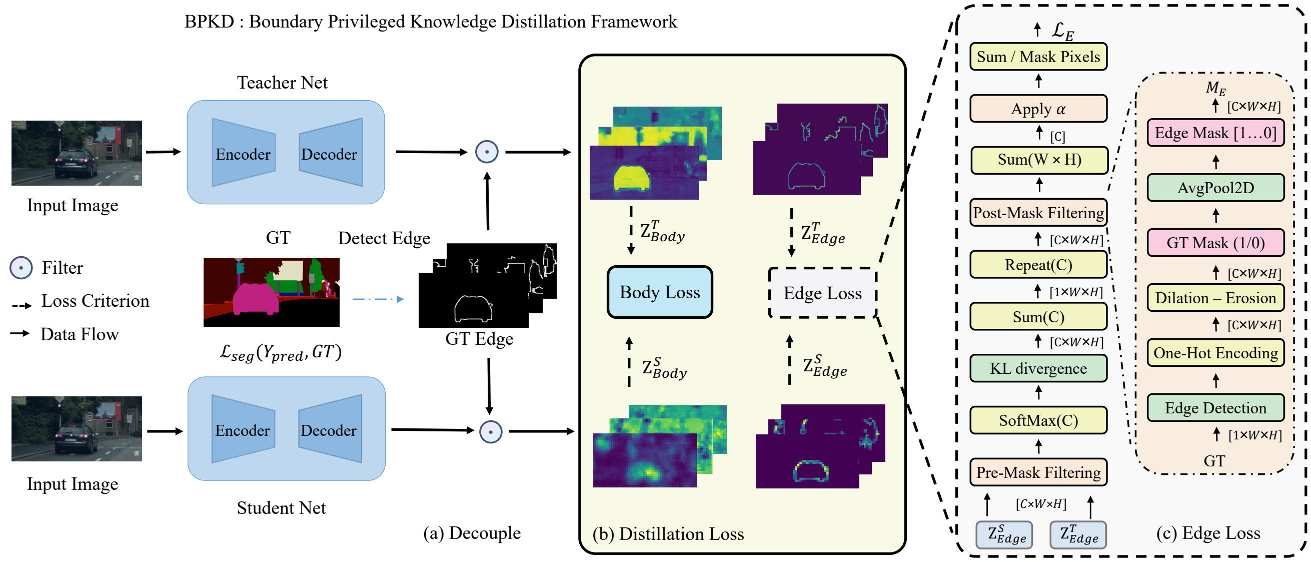


Figure 2. Illustration of our proposed **Boundary Privileged Knowledge Distillation** framework and architecture. (a) demonstrates the decoupling process that involves the edge detection on the ground truth to generate GT_{Edge} Mask, followed by applying the mask filter to obtain the Teacher and Student logits masks. This step ensures that the information from the boundary region is isolated and appropriately conveyed to the Student. (b) shows that distillation comprises two terms: body loss and edge loss. The body loss term captures the categorized similarities, whereas the edge loss term concentrates on the boundary regions' transfer. $\mathcal{Z}_E$ and $\mathcal{Z}_B$ is short terms for $\mathcal{Z}_{Edge}$ and $\mathcal{Z}_{Body}$. In $\mathcal{Z}^{S,T}$, S and T standard for student and teacher, respectively. (c) shows edge loss calculation is performed in two stages: pre-mask filtering and post-mask filtering. The pre-mask filtering step shapes the probability distribution to contain only edge information. Subsequently, the post-mask filtering step aggregates contextual information between adjacent categories to produce the final edge loss.

edge loss weight λ_e to control the contributions of each loss term. This decoupling strategy allows us to examine the sensitivity of edge learning in the knowledge distillation process, which has been overlooked by the literature. The loss objective is defined as:

$$\ell = \lambda_b \cdot \ell_B \left(\mathcal{Z}^S_B, \mathcal{Z}^T_B\right) + \lambda_e \cdot \ell_E \left(\mathcal{Z}^S_E, \mathcal{Z}^T_E\right). \tag{1}$$

3.2. Edge Knowledge Representation.

Our framework minimizes the discrepancy between the teacher's and the student's features on the edge areas. To achieve this, we extract an edge knowledge representation by using soft edge masks. This edge mask is applied on the logit map to produce a masked feature representation for teachers' and students' edge regions. The edge map M_E is created through two stages: Pre-Mask Filtering (PRM), which captures edge discrepancy for all classes, and Post-Mask Filtering (POM), which extracts edge discrepancy for each individual class. This approach allows a model to extract a more accurate and precise representation of the edge knowledge, leading to improved performance in details classification.

During the edge detection process, we employ an adjustable Trimap algorithm [50] to extract edge representations denoted by Z_E from the ground truth (GT). Although the teacher's predictions could serve as an alternative source for generating an edge mask, they offer slightly reduced accuracy compared to using GT directly. To generate the binary edge mask GT_{edge}, we compute the difference between the dilation and erosion operations applied to the GT, formally expressed as $GT_{\text{edge}} = \text{dilation}(GT) - \text{erosion}(GT)$. The resultant binary mask $GT_{\text{edge}} \in \mathbb{R}^{C\times H\times W}$ undergoes average pooling to produce $M_E \in \mathbb{R}^{C\times H'\times W'}$, which shares the same shape as the logits prediction $\mathcal{Z}_{\text{pred}}$. The dimensions of M_E are governed by the output stride S of the segmentation network, specifically $W' = W/S$.

Pre Mask Filtering (PRM). To obtain the logits map, we apply M_E to both student and teacher logits: $\mathcal{Z}_E = \mathcal{Z}_{pred} \cdot M_E$. Specifically, we apply the edge mask for each channel C so that we can concentrate the logits for the overlap edge regions. An intuitive example is if there is a frame that displays a dog and a cat standing nearby, only the logits activation of the dog and the cat class will be considered and all other activation will be suppressed. Such an operation forces the student to focus more on the correlations between the adjacent ambiguous classes. A spatial-level KL divergence loss is applied to the filtered logits $\mathcal{Z}^S_E$ and $\mathcal{Z}^T_E$:

$$\varphi(\mathcal{Z}^T_E, \mathcal{Z}^S_E) = \sum_{c=1}^{C} \phi(\mathcal{Z}^T_{E,i}) \cdot \log \frac{\phi(\mathcal{Z}^T_{E,i})}{\phi(\mathcal{Z}^S_{E,i})}, \tag{2}$$

where ϕ is the softmax operation for each pixel. $\varphi(\mathcal{Z}^T_E, \mathcal{Z}^S_E)$ represents the edge-masked KL distances for all spatial locations.

Post Mask Filtering (POM). We further separate the edge loss for each class and perform normalization based on the edge area by Post Mask Filtering (POM). Let $\mathcal{Z}^T_{E,i,c}$ and $\mathcal{Z}^S_{E,i,c}$ denote the logits for the c-th class at pixel i in the teacher and student models, respectively. Let $M_{E,c}$ be the soft mask obtained by average-pooling the ground truth binary edge mask for the c-th class, and n_c denotes the number of non-zero pixels in $M_{E,c}$ for this class. Our POM term can be formulated as follows:

$$\ell_E = \sum_{c=1}^{C} \frac{\alpha_c}{n_c} \sum_{i=1}^{W\cdot H} \varphi(\mathcal{Z}^T_{E,i,c}, \mathcal{Z}^S_{E,i,c}) \cdot M_{E,c}, \tag{3}$$

By re-weighing the loss based on the edge area of each class, we prioritize the center of the edge, where the most important information is often located. This approach ensures that the student model focuses on learning the correct edge positions and shapes for each class.

The Soft Edge Masks M_E play a critical role in the Edge Loss, and our approach to generating them involves two specialized designs: 1) converting binary GT_E into a weighted discrete space, and 2) generating masks per channel instead of a unified mask. Directly applying binary masks may include unconfident bias, so we use average pooling to generate softer masks. We also carefully consider overlapping masks to minimize noise and uncertainty. Our mask design aims to exclusively include unconfident bias for minimizing knowledge distributions.

In summary, the proposed PRM and POM stages in the edge region refine the knowledge distillation process by identifying edge discrepancies for each class and applying a reweighting to the loss based on the edge area. This method guarantees that the student model learns the correct edge positions and shapes for each class, and provides the shape prior knowledge for body knowledge representation.

3.3. Body Knowledge Representation.

This section investigates the body knowledge distillation. Prior works consider whole-view distillation, which dilutes body knowledge with noisy representation on the edge. To overcome these challenges, we utilize the reversed edge binary mask to extract body masks. By removing the edge region, we exploit implicit shape constraints and reduce uncertainty, which allows the body loss to focus on assigning the large inner regions of objects to their corresponding categories. To achieve this, we propose a region alignment approach that synthesizes channel-level activations to obtain semantically rich sections. As we predefined that $\mathcal{Z} = \mathcal{Z}_B + \mathcal{Z}_E$. The body logits is obtained by $\mathcal{Z}_B = \mathcal{Z} \times (1 - M_E)$. As shown in the previous work [34, 47], a pixel-wise loss for the body region will bring in unexpected noise due to the hard constraints. Thus, we employ a loose constraint of the channel-wise distillation [47] for the body part. Body enhancement loss (BEL) is defined as:

$$\ell_B = \frac{\mathcal{T}^2}{C} \sum_{c=1}^{C} \sum_{i=1}^{W\cdot H} \phi(Z^T_{B,c,i}) \cdot \log\Big[\frac{\phi(Z^T_{B,c,i})}{\phi(Z^S_{B,c,i})}\Big], \tag{4}$$

4. Experiments

4.1. Experimental Setup

Dataset. We conduct the experiments on three benchmark datasets for semantic segmentation: Cityscapes [13], Pascal Context 2010 [15], and ADE20K [68].

ADE20K [68] contains 20k/2k/3k images for train/val/test with 150 semantic classes. It is constructed as the benchmark for scene parsing and instance segmentation.

Cityscapes [13] is an urban scene parsing dataset that contains 2975/500/1525 finely annotated images used for train/val/test. The performance is evaluated in 19 classes.

Pascal Context [15] provides dense annotations, which contain 4998/5105/9637 train/val/test images. We use 59 object categories for training and testing. Our results are reported on the validation set.

Implementation Details. Our implementation is based on the open-source toolbox MMSegmentation [11, 12] with PyTorch 1.11.0. We employ the standard data augmentation, including random flipping, cropping, and scaling in the range of [0.5, 2]. All experiments are optimized by SGD with a momentum of 0.9, and a batch size of 16. We use the crop of 512 × 512, 512 × 1024, and 480 × 480 for ADE20k, Cityscapes, and Pascal Context, correspondingly. We use an initial learning rate of 0.01 for ADE20K and Cityscapes. In addition, we use an initial learning rate of 0.004 for Pascal Context. The number of total training iterations is 80K. Following the previous methods [7, 67], we use the poly learning rate policy and report the single-scale testing result. We conduct all experiments on 4 NVIDIA A100 GPUs. All the distillation methods are trained with the same configurations.

Metrics. We set up a fair comparison by assigning identical parameters for each method with the same dataset. Mean Intersection-over-Union (mIoU), Trimap mIoU and pixel mean accuracy (mAcc) are employed as the main evaluation metrics. GFLOPs, FPS and No. Parameters are also reported for various student networks that we tested. All reported computational costs are measured using the fvcore. [1]

4.2. Compare with State-of-the-arts Methods

To ensure a fair comparison, we have re-implemented a number of previously proposed knowledge distillation methods, including those by [34, 47, 53, 57]. Subsequently, we benchmarked our BPKD method against various compact networks, such as PSPNet with ResNet18

[1] https://github.com/facebookresearch/fvcore

Table 1. Performance comparison of different distillation methods with state-of-the-art techniques. We test these methods on various segmentation networks for both student and teacher models, using datasets including Cityscapes [13], ADE20K [68], and Pascal Context [15]. The FLOPs and FPS are obtained on 512 × 512 resolutions. Our BPKD outperforms all previous methods in large margins across multiple datasets and network architectures. DLab refers to Deeplab architecture. HRV2P refers to HRNetV2p. MV2 refers to MobileNet v2.

Methods	FLOPs(G)	Param(M)	FPS(S)	ADE20K		Cityscapes		Pascal Context 59	
				80k 512*512		80k 1024*512		80k 480*480	
				mIoU(%)	mAcc(%)	mIoU(%)	mAcc(%)	mIoU(%)	mAcc(%)
T: PSPNet-R101 [67]	256.89	68.07	2.68	44.39	54.75	79.74	86.56	52.47	63.15
S:PSPnet-R18 [67]	54.53	12.82	15.71	33.30	42.58	74.23	81.45	43.79	54.46
SKDS [34]	54.53	12.82	15.71	34.49(▲1.19)	44.28	76.13(▲1.9)	82.58	45.08(▲1.29)	55.56
IFVD [21]	54.53	12.82	15.71	34.54(▲1.24)	44.26	75.35(▲1.12)	82.86	45.97(▲2.18)	56.6
CIRKD [57]	54.53	12.82	15.71	35.07(▲1.77)	45.38	76.03(▲1.80)	82.56	45.62(▲1.83)	56.15
CWD [47]	54.53	12.82	15.71	37.02(▲3.72)	46.33	76.26(▲2.03)	83.04	45.99(▲2.20)	55.56
BPKD(Ours)	54.53	12.82	15.71	**38.51(▲5.21)**	**47.70**	**77.57(▲3.34)**	**84.47**	**46.82(▲3.03)**	**56.29**
T:HRV2P-W48 [51]	95.64	65.95	6.42	42.02	53.52	80.65	87.39	51.12	61.39
S:HRV2P-W18S [51]	10.49	3.97	23.74	31.38	41.39	75.31	83.71	40.62	51.43
SKDS [34]	10.49	3.97	23.74	32.57(▲1.19)	43.22	77.27(▲1.96)	84.77	41.54(▲0.92)	52.18
IFVD [21]	10.49	3.97	23.74	32.66(▲1.28)	43.23	77.18(▲1.87)	84.74	41.55(▲0.93)	52.24
CIRKD [57]	10.49	3.97	23.74	33.06(▲1.68)	44.30	77.36(▲2.05)	84.97	42.02(▲1.40)	52.88
CWD [47]	10.49	3.97	23.74	34.00(▲2.62)	42.76	77.87(▲2.56)	84.98	42.89(▲2.27)	53.37
BPKD(Ours)	10.49	3.97	23.74	**35.31(▲3.93)**	**46.11**	**78.58(▲3.27)**	**85.78**	**43.96(▲3.34)**	**54.51**
T:DLabV3P-R101 [6]	255.67	62.68	2.60	45.47	56.41	80.98	88.70	53.20	64.04
S:DLabV3P-MV2 [46]	69.60	15.35	8.40	31.56	45.14	75.29	83.11	41.01	52.92
SKDS [34]	69.60	15.35	8.40	32.49(▲0.93)	46.47	76.05(▲0.76)	84.14	42.07(▲1.06)	55.06
IFVD [21]	69.60	15.35	8.40	32.11(▲0.55)	46.07	76.97(▲1.68)	84.85	41.73(▲0.72)	54.34
CIRKD [57]	69.60	15.35	8.40	32.24(▲0.68)	46.09	77.71(▲2.42)	85.33	42.25(▲1.24)	55.12
CWD [47]	69.60	15.35	8.40	35.12(▲3.56)	49.76	77.97(▲2.68)	86.68	43.74(▲2.73)	56.37
BPKD(Ours)	69.60	15.35	8.40	**35.49(▲3.93)**	**53.84**	**78.59(▲3.30)**	**86.45**	**46.23(▲5.22)**	**58.12**
T:ISANet-R101 [24]	228.21	56.80	2.35	43.80	54.39	80.61	88.29	53.41	64.04
S:ISANet-R18 [24]	54.33	12.46	17.34	31.15	41.21	73.62	80.36	44.05	54.67
SKDS [34]	54.33	12.46	17.34	32.16(▲1.01)	41.80	74.99(▲1.37)	82.61	45.69(▲1.64)	56.27
IFVD [21]	54.33	12.46	17.34	32.78(▲1.63)	42.61	75.35(▲1.73)	82.86	46.75(▲2.70)	56.4
CIRKD [57]	54.33	12.46	17.34	32.82(▲1.67)	42.71	75.41(▲1.79)	82.92	45.83(▲1.78)	56.11
CWD [47]	54.33	12.46	17.34	37.56(▲6.41)	45.79	75.43(▲1.81)	82.64	46.76(▲2.71)	56.48
BPKD(Ours)	54.33	12.46	17.34	**38.73(▲7.58)**	**47.92**	**75.72(▲2.10)**	**83.65**	**47.25(▲3.20)**	**56.81**

Table 2. Performance comparison of transformers-based architecture vs. different distillation strategies. Standard mIoU and Trimap mIoU of Swin Transformers [37] and DeiT [49] with ViT Adapter (DeiT-Ada) [8] on ADE20K with UPerNet [54] decoder for 80K iterations. Distillation forward speed (DFS.), training time (TT.) and GPU memory footprint (GMem.) light our method have negligible computational cost. (DFS.) and (TT.) estimated on DeiT-Adapter with batch size = 16 with 4 GPUs and (GMem.) standard for per sample video memory allocation.

	DFS.(S)↑	TT.(H)↓	GMem.(G)↓	Swin↑	Trimap ↑	DeiT-Ada.↑	Trimap ↑
T:Base	9.52	11.26	8.32	50.13	40.10	48.80	39.72
S:Tiny	12.80	8.44	3.87	43.57	32.78	41.10	32.15
SKDS	8.72	11.36	4.45	43.58	33.04	41.90	32.25
IFVD	6.06	16.45	8.97	43.75	32.90	41.16	32.11
CIRKD	7.70	16.35	10.70	43.32	32.68	41.64	32.23
CWD	8.76	11.15	4.45	44.99	33.73	44.25	33.49
BPKD	7.84	13.49	5.49	**46.13**	**38.11**	**45.25**	**37.05**

backbone [67], HRNet-W18 [51], Deeplab-V3+ with MobileNetV2 backbone [46], ISANet with ResNet18 [24], Swin Transformers [37] with UPerNet [54] and DeiT [49]-Adapter [8] with UPerNet [54].

Performance. Table 1 reports our method's performance on the ADE20K validation set, where the proposed BPKD achieves state-of-the-art (SOTA) performance across multiple student networks. The distillation process enhances the mean Intersection over Union (mIoU) for student networks by up to 24.33%. Notably, BPKD consistently outperforms the current SOTA, CWD, by 3.87% across all evaluated network architectures. Additional results on the Pascal Context validation set indicate an average performance increase of 2% over SOTA methods. Further experiments presented in Table 2 reveal the method's efficacy on popular Transformer architectures like Swin and DeiT, where it outperforms CWD by up to 2.53%. These findings underscore the architecture-agnostic nature of our approach. In terms of computational efficiency, BPKD achieves a distillation forward speed up to 29.3%, training duration reduction of 21.2% and 21.9%, and a memory consumption (GMem.) reduction of 94.8% and 63.4%, compared to previous methods such as CIRKD and IFVD. Additionally, we register a 13% improvement in Trimap metrics over CWD on the Swin architecture, emphasizing the method's cost-

<table>
<tr><th colspan="7">Method</th></tr>
<tr><td colspan="3">Teacher: PSP-ResNet101
Student: PSP-ResNet18</td><td colspan="4">79.74%
Standard: 68.99%   Trimap: 55.34%</td></tr>
<tr><td colspan="3">Channel Wise Distillation [47]
Pixel Wise Distillation [34]</td><td colspan="4">Standard: 74.29%   Trimap: 57.34%
Standard: 69.33%   Trimap: 53.82%</td></tr>
<tr><td></td><td></td><td><b>Body(C)</b></td><td>Body(P)</td><td><b>Edge(P)</b></td><td>Edge(C)</td><td><b>Ours</b></td></tr>
<tr><td rowspan="2">mIoU(%)</td><td>Standard</td><td>74.17 (▲5.18)</td><td>72.70 (▲3.71)</td><td>71.63 (▲2.64)</td><td>66.83 (▼2.16)</td><td><b>75.94 (▲6.95)</b></td></tr>
<tr><td>Trimap</td><td>56.20 (▲0.86)</td><td>54.12 (▼1.22)</td><td>61.37 (▲6.03)</td><td>51.76 (▼3.58)</td><td><b>62.91 (▲7.57)</b></td></tr>
</table>

Table 3. The effectiveness of the decoupling whole-view knowledge representation. The results show knowledge representation for different spatial locations should be considered, separately. C and P denote channel-wise and pixel-wise knowledge distillation, respectively.

effectiveness with SOTA performance.

4.3. Ablation Study

In this section, we comprehensively evaluate our BPKD under different settings. All ablation experiments are conducted on the Cityscapes dataset with T: PSPNet-R101 and S: PSPNet-R18. To decrease computational costs, we adopt a streamlined training configuration, including crop size reduced to 512 × 512, and training schedule to 40k iterations. More experiment results are shown in the supplementary.

Effectiveness of Decoupled Knowledge. To verify the effectiveness of the proposed knowledge distillation approach, we evaluate the segmentation performance in the edge region in Table 3. We evaluated the Trimap mIoU metric [7] when using channel-wise and pixel-wise normalization for network distillation. Channel-wise normalization led to a decline in both standard mIoU by 2.16% and Trimap mIoU by 3.58%, indicating its sensitivity in edge regions. Conversely, pixel-wise distillation enhanced Trimap mIoU by 6.03%, benefiting from our specialized Edge loss design that refines boundary quality [29, 30]. The Body loss function shows an increase in standard mIoU by 5.18% when uses channel-wise and 3.71% when spatial-wise, emphasizing its effectiveness for body regions. However, it had a limited impact on Trimap's performance. In summary, our optimal approach, BPKD, achieved the best mIoU scores of $75.94(+6.95)\%$ and $62.91(+7.59)\%$ on standard and Trimap evaluations, respectively. This highlights BPKD's capability to refine semantic boundaries and body regions through pixel-level alignment and context aggregation.

Compare Edge Filter Locations. From Table 4, the numerical results demonstrated the effectiveness of our proposed method. We further analyzed the impact of applying the edge filter for different locations. Applying the Pre Mask filter, the performance slightly improved by 1.98% compared to the student without distillation. In contrast, Post Mask filtering improves the performance by 2.64%, because POM extracts edge discrepancy specifically for each type of class. The Body Enhancement Module takes care of non-edge information during our distillation setting, and the performance is raised by 5.18%. Afterwards, we explore the performance by applying PRM or POM to BEL. The combination of three terms archives best mIoU that 75.94% on the Cityscapes validation set.

Method	mIoU(%)	IMP.(%)	mAcc(%)
Teacher	79.74	-	86.56
ResNet18	68.99	-	75.19
+ PRM	70.37	▲1.98	76.95
+ POM	71.63	▲2.64	78.47
+ BEL	74.17	▲5.18	80.47
+ PRM + BEL	74.81	▲5.92	81.52
+ POM + BEL	74.62	▲5.63	80.98
+ PRM + POM + BEL	**75.94**	**▲6.95**	**82.62**

PRM: Pre-Mask Filtering   **BEL**: Body Loss   **POM**: Post-Mask Filtering

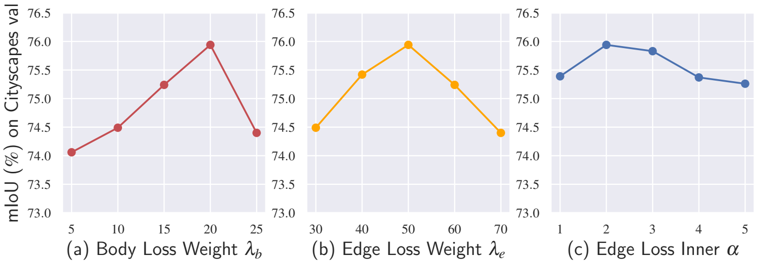
Table 4. The different locations apply the mask in the proposed method. (IMP.) refers to the improvement achieved by the student network.

Impact of Different Edge Widths. Edge width is pivotal in the Edge Detection Module, intricately influencing both the quality and the computational demands of the edge detection process. Specifically, a larger edge unit results in a wider edge, which incorporates more pixels into the edge loss. While this increase can potentially augment the robustness of edge detection, it also presents a trade-off: the wider the edge includes pixels from the body region, which could dilute the precision of edge-specific features. We systematically analyze the impact of varying edge widths on the distillation performance, as summarized in Table 5. Our analysis reveals that an edge width of 7 units is optimal, leading to a significant performance improvement of 6.95%.

Impact of Hyperparameters. In the optimization process, the parameters λ_b and λ_e serve as weighting factors for the body and edge loss functions, respectively. These hyperparameters are pivotal in balancing the focus between semantic boundary refinement and overall body region accuracy. A higher value of λ_e places more emphasis on edge quality, conversely, a larger λ_b value makes the model more attuned to the larger body regions and potentially boosts the standard mIoU. The edge loss incorporates an internal parameter α, serving as a class-wise balancing factor. This parameter is particularly instrumental in mitigating class imbalance

Edge Width	mIoU	IMP.	mAcc
3	73.89	▲4.90	81.24
5	74.26	▲5.27	82.00
7	**75.94**	**▲6.95**	**82.62**
10	74.70	▲5.71	82.13
15	74.36	▲5.37	82.02

Table 5. Performance comparison for various different edge width with PSPNet-R18 on the Cityscapes validation set, for a fair comparison, we rerun the same setting 3 times and measure mean mIoU for evaluation.



(a) Body Loss Weight λ_b (b) Edge Loss Weight λ_e (c) Edge Loss Inner α

Figure 3. Impact of the (a) Body Loss weight λ_b and (b) Edge Loss weight λ_e and (c) Edge Loss Inner weight α on Cityscape val. We found the optimal combination by board range study that $\lambda_b = 20, \lambda_e = 50, \alpha = 2$.

by intra-adjusting the contribution of each class to the overall loss. As shown in Fig 3, we investigate the impact of loss weights in our BPKD and find $\lambda_b = 20$, $\lambda_e = 50$ and $\alpha = 2$ is the best choice.

CAM Visualization Analysis. Figure 4 illustrates the explicit refinement of semantic boundaries on multiple objects. The shape constraints are evident, such as the strong attention given to the pillow outlines. Despite BPKD distillation, the student network cannot perfectly segment the horse in the second row, but it has highly attended to the bone silhouette. This shows that BPKD has tried its best to distill knowledge to the student network, but due to its limited size and capacity, the student can only learn the surface-level capacity of the teacher. The body sections have been smoothly affiliated to a single category, reducing high-uncertainty edge and incorporating shape prior knowledge from the edge loss pressure. More qualitative segmentation results in supplement visually demonstrate our BPKD's effectiveness for both tiny and large objects with explicit boundaries enhancement.

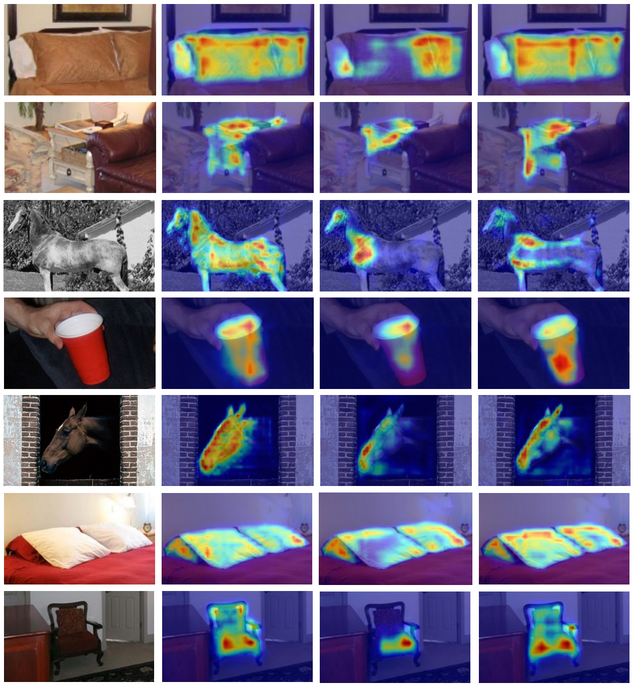

(a) Image (b) Teacher (c) W/o distillation (d) **BPKD(Ours)**

Figure 4. Comparison of CAM visualizations among (b) the teacher model, (c) the student model without distillation, and (d) the BPKD model. Activation maps were extracted from the last block of the corresponding ResNet backbones using HiResCAM [14]. The results indicate that BPKD shows superior refinement of boundaries and higher attention to semantic bodies. For better visualization, zoom-in is recommended.

5. Conclusion

This work presents a novel boundary-privileged knowledge distillation (BPKD) method for semantic segmentation, which transfers the cumbersome teacher model's body and edge knowledge to the compact student model, separately. Extensive experiments demonstrate that the knowledge representation in the body and the edge regions should be considered differently. Due to different intrinsic properties, the edge region needs to focus on distinguishing between uncertain classes for each pixel while the body region needs to focus more on localizing and connecting object structures. Experimental results show that the proposed distillation method consistently outperforms state-of-the-art methods on various public benchmark datasets. The overall mIoU and the performance in the edge region are both improved by a large margin.

6. Acknowledgments

Y. Liu acknowledges the support of start-up funding from The University of Adelaide for their participation in this work. We express our gratitude to The University of Adelaide High-Performance Computing Services for providing the GPU Compute Resources, and to Mr. Wang Hui and Dr. Fabien Voisin for their valuable technical support. We would like to thank Mr. Hanwen Wang for providing the visualization collection used in the supplementary.

This work stands as a testament to the dedication and expertise of Dr. Yifan Liu. We are deeply grateful to Dr. Liu for her unparalleled contribution and leadership throughout this project. Her insights and guidance have been invaluable. We wish to honour and remember Dr. Yifan Liu for her enduring impact and legacy in this field.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. 2

[2] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 3

[3] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *arXiv preprint arXiv:2207.02039*, 2022. 2, 3

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 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 transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 6

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2, 5, 7

[8] 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. 6

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1

[10] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1

[11] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 5

[12] MMRazor Contributors. Openmmlab model compression toolbox and benchmark. https://github.com/open-mmlab/mmrazor, 2021. 5

[13] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 5, 6

[14] Rachel Lea Draelos and Lawrence Carin. Hirescam: Faithful location representation in visual attention for explainable 3d medical image classification. *arXiv preprint arXiv:2011.08891*, 2020. 8

[15] 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. 2, 5, 6

[16] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9725, 2021. 2

[17] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[19] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787, 2019. 2, 3

[20] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. 2015. 2, 3

[21] Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, and Chen Change Loy. Inter-region affinity distillation for road marking segmentation. In *CVPR*, pages 12486–12495, 2020. 6

[22] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 2, 3

[23] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[24] 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

[25] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 3

[26] Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. Structural and statistical texture knowledge distillation for semantic segmentation. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16876–16885, 2022. 2

[27] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988. 3

[28] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 2, 3

[29] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009. 7

[30] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 7

[31] Seung Hyun Lee, Dae Ha Kim, and Byung Cheol Song. Self-supervised knowledge distillation using singular value decomposition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–350, 2018. 2, 3

[32] Akide Liu and Zihan Wang. Cv 3315 is all you need: Semantic segmentation competition. *arXiv preprint arXiv:2206.12571*, 2022. 2

[33] Xiaofeng Liu, Fangxu Xing, Georges El Fakhri, and Jonghye Woo. Self-semantic contour adaptation for cross modality brain tumor segmentation. In *IEEE International Symposium on Biomedical Imaging*, pages 1–5. IEEE, 2022. 2

[34] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 2, 3, 5, 6, 7

[35] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 3

[36] Yueming Liu, Xiaomei Yang, Zhihua Wang, Chen Lu, Zhi Li, and Fengshuo Yang. Aquaculture area extraction and vulnerability assessment in sanduao based on richer convolutional features network model. *Journal of Oceanology and Limnology*, 37(6):1941–1954, 2019. 3

[37] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[38] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1

[39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[40] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 2

[41] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018. 3

[42] Shivam K Panda, Yongkyu Lee, and M Khalid Jawed. Agronav: Autonomous navigation framework for agricultural robots and vehicles using semantic segmentation and semantic line detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6271–6280, 2023. 2

[43] 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. 3

[44] Judith MS Prewitt et al. Object enhancement and extraction. *Picture processing and Psychopictorics*, 10(1):15–19, 1970. 3

[45] 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. 3

[46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3, 6

[47] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 2, 3, 5, 6, 7

[48] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[49] 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

[50] Jue Wang, Michael F Cohen, et al. Image and video matting: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(2):97–175, 2008. 4

[51] 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 transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1, 3, 6

[52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[53] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *European Conference on Computer Vision*, pages 346–362. Springer, 2020. 2, 3, 5

[54] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 6

[55] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 1

[56] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 3

[57] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022. 2, 3, 5, 6

[58] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pages 53–69. Springer, 2022. 2, 3

[59] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4133–4141, 2017. 2, 3

[60] Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5964–5973, 2017. 3

[61] Yuhui Yuan, Xiaokang Chen, Xilin Chen, and Jingdong Wang. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019. 1

[62] 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, 3

[63] Bowen Zhang, Liyang Liu, Minh Hieu Phan, Zhi Tian, Chunhua Shen, and Yifan Liu. Segvitv2: Exploring efficient and continual semantic segmentation with plain vision transformers. *IJCV*, 2023. 1

[64] Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, and Yifan Liu. Segvit: Semantic segmentation with plain vision transformers. *arXiv preprint arXiv:2210.05844*, 2022. 1, 2

[65] Chenrui Zhang and Yuxin Peng. Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. *arXiv preprint arXiv:1804.10069*, 2018. 2, 3

[66] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2736–2746, 2022. 1

[67] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2, 5, 6

[68] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 5, 6