DeepGBASS: Deep Guided Boundary-Aware Semantic Segmentation

Qingfeng Liu, Hai Su, Mostafa El-Khamy, Kee-Bong Song
SOC Multimedia R&D, Samsung Semiconductor, Inc., USA
{qf.liu, hai.su, mostafa.e, keebong.s}@samsung.com

***Abstract*—Image semantic segmentation is ubiquitously used in scene understanding applications, such as AI Camera, which require high accuracy and efficiency. Deep learning has significantly advanced the state-of-the-art in semantic segmentation. However, many of recent semantic segmentation works only consider class accuracy and ignore the accuracies at the boundaries between semantic classes. To improve the semantic boundary accuracy, we propose low complexity Deep Guided Decoder (DGD) networks, trained with a novel Semantic Boundary-Aware Learning (SBAL) strategy. Our ablation studies on Cityscapes and the ADE20K-32 confirm the effectiveness of our approach with network of different complexities. We show that our DeepGBASS approach significantly improves the mIoU by up to 11% relative gain and the mean boundary F1-score (mBF) by up to 39.4% when training MobileNetEdgeTPU DeepLab on ADE20K-32 dataset.**

***Index Terms*—Semantic Segmentation, Deep Guided Decoder, Semantic Boundary-Aware Learning**

I. Introduction

Semantic segmentation is a fundamental task in many downstream computer vision tasks. Deep neural network based semantic segmentation is a dense prediction network that aims to classify each pixel in the input image into its corresponding predefined categories. For some tasks, e.g., content-aware Image Signal Processing (ISP) and autonomous driving, the accuracy in semantic boundary region is crucial. In particular for ISP chain in the mobile devices, many image enhancement operations rely on the boundary accuracy between different regions in the image.

One issue in many state-of-the-art segmentation networks is that most of them limits the networks capability to make accurate prediction around the semantic boundary regions due to the information loss of the low resolution feature map due to the downsampling inside the network. The boundary accuracy degradation will be more obvious with a low complexity network such as MobileNetEdgeTPU DeepLab model [7]. Another issue is that compared to the pixels near to the centers of objects, it is harder for the deep neural network to learn how to correctly predict the pixels near to the boundary regions. This is because: i) there are far fewer pixels in the boundary regions than the pixels located in the non-boundary regions therefore the training is dominated by the non-boundary pixels; ii) to correctly classify the boundary pixels requires both local and global information. However, it is challenging for the deep architecture to capture the local information. To counteract such problem, a common practice is to introduce a weight mask to weight the boundary pixels and non-boundary pixels differently so that the loss function can put more penalty on the boundary pixels during the training. However, such strategy can only achieve limited improvement because there is only uniform weight for the non-boundary regions and it does not penalize false boundaries in such regions.

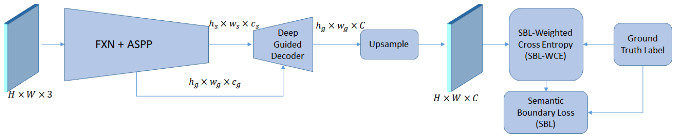

$H \times W \times 3$
FXN + ASPP
$h_s \times w_s \times c_s$
$h_g \times w_g \times c_g$
Deep Guided Decoder
$h_g \times w_g \times C$
Upsample
$H \times W \times C$
SBL-Weighted Cross Entropy (SBL-WCE)
Ground Truth Label
Semantic Boundary Loss (SBL)

Fig. 1: Overview of Deep Guided Decoder (DGD) and Semantic Boundary-Aware Learning (SBAL).

To address above issues, we propose our Deep Guided Boundary-Aware Semantic Segmentation (DeepBASS) model that consists of two novel modules: the Deep Guided Decoder (DGD) and the Semantic Boundary-Aware Learning (SBAL). First, the Deep Guided Decoder (DGD) refines the boundary accuracy of the low resolution semantic segmentation feature by using the low level feature from the Feature eXtraction Network (FXN) or backbone as the guidance to upsample the low resolution high level feature map to utilize the boundary structure in the low level feature. Second, the Semantic Boundary-Aware Learning (SBAL) trains the network with a combination of the channel-wise Semantic Boundary Loss (SBL) and SBL-Weighted Cross Entropy (SBL-WCE) loss. Both losses are calculated based on the boundary prediction errors. The SBAL only introduces extra computation during training and does not require extra computation during the inference.

We conducted extensive experiments on two popular datasets, i.e., Cityscapes [4] and a reduced ADE20K-32 [12]. In addition to mean Intersection over Union (mIoU), mean Boundary F1 score (mBF) [5] is used to measure the semantic boundary accuracy. For the two datasets, our approach achieves significant improvement in both mIoU and mBF with networks of different complexities.

The novelties of this paper are summarized as follows.

- Novel Deep Guided Decoder that refines and upsamples the intermediate high level semantic feature using the low level feature from the backbone as the guidance. In contrast, the prior art uses the high resolution input image as the guidance to refine the final output prediction map, which causes higher computational cost.
- A Semantic Boundary-Aware Learning strategy that combines SBL and SBL-WCE. Our approach can further enhance the networks capability to recognize the detailed image structures, thus achieves better accuracy in the semantic boundary regions.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

II. Related Works

In [9], Howard et al. introduced a hardware aware neural architecture search (NAS) framework to obtain the MobileNet-EdgeTPU, which achieved the state-of-the-art accuracy under the same multiplication-add count (MAC) on image classification tasks. Based on the MobileNetEdgeTPU, a light weight semantic segmentation model MNEdge DeepLab, is formed by integrating the decoder framework from the DeepLab network [2]. Liu et al. [11] proposed the attention based GSANet to improve the lightweight models based on the MobileNetEdgeTPU backbone network.

Recently, Wu et al. [14] proposed a guided filter module, which is reformulated as a fully differentiable block that could be integrated with the Convolutional Neural Networks (CNNs) and jointly optimized through end-to-end training. Our approach differs from this work. The original guided filter performs like a post processing module which will take the high resolution input image as the guidance and the low resolution network output as the feature to be refined. This will end up with high computational cost when the input image is very high resolution which is very common in nowadays mobile devices. In comparison, our proposed DGD is integrating the guided filter into the decoder of the network, where all the operations are applied with the downsampled low resolution network input. As a result, the proposed DGD will have much less computational cost and meanwhile, it will also take the benefit of the boundary information since the low level encoder feature is capable of encoding the structured information of the input image.

A few approaches have been proposed to improve the boundary accuracy of semantic segmentation, including i) to augment the semantic segmentation network with a boundary detection branch [1]; ii) multi-scale feature aggregation; and iii) multi-task training [15]. Another line of research aims to enhance the feature representation via the boundary information modulated feature propagation [3], [6]. The work closely related to our work is [13]. In [13], the semantic network is augmented with a semantic boundary detection branch. A dual task loss with L1 loss and a cross-entropy loss is proposed to further regularize the network training. Our SBAL is different from [13]. First, we propose to use the boundary prediction error to weight the cross entropy loss. Secondly, we remove the boundary detection branch and loss for the boundary detection branch to accelerate the inference and training.

III. Deep Guided Boundary-Aware Segmentation Networks

A. Deep Guided Decoder

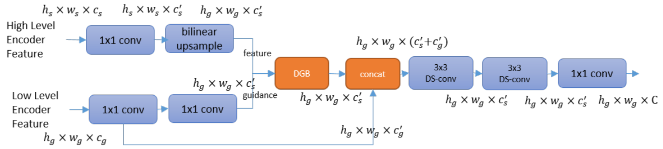

$h_s \times w_s \times c_s$
$h_s \times w_s \times c_s'$
$h_g \times w_g \times c_s'$
High Level Encoder Feature
1x1 conv
bilinear upsample
feature
$h_g \times w_g \times (c_s'+c_g')$
DGB
concat
3x3 DS-conv
3x3 DS-conv
1x1 conv
$h_g \times w_g \times c_s'$
guidance
Low Level Encoder Feature
1x1 conv
1x1 conv
$h_g \times w_g \times c_s'$
$h_g \times w_g \times c_s'$
$h_g \times w_g \times c_s'$
$h_g \times w_g \times C$
$h_g \times w_g \times c_g$
$h_g \times w_g \times c_g'$

Fig. 2: The architecture of the Deep Guided Decoder (DGD).

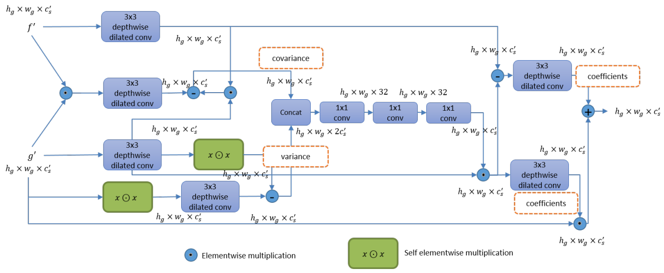

$h_g \times w_g \times c_s'$
f'
3x3 depthwise dilated conv
$h_g \times w_g \times c_s'$
covariance
3x3 depthwise dilated conv
$h_g \times w_g \times c_s'$
$h_g \times w_g \times c_s'$
Concat
1x1 conv
1x1 conv
1x1 conv
$h_g \times w_g \times 32$
$h_g \times w_g \times 32$
$h_g \times w_g \times 2c_s'$
$h_g \times w_g \times c_s'$
3x3 depthwise dilated conv
coefficients
$h_g \times w_g \times c_s'$
g'
$h_g \times w_g \times c_s'$
3x3 depthwise dilated conv
$x \odot x$
variance
$x \odot x$
3x3 depthwise dilated conv
3x3 depthwise dilated conv
coefficients
Elementwise multiplication
Self elementwise multiplication

Fig. 3: The implementation details of the Deep Guided Block (DGB).

We consider deep encoder-decoder architectures for image semantic segmentation networks. Inspired by [8], where the guided filter is used to model the output feature as a linear combination of guidance feature to preserve the edge information of the guidance feature. We propose the Deep Guided Decoder (DGD) to play the role of recovering the low-resolution features output from the encoder to full-resolution features, same as the input image. The DGD takes two inputs: the semantic feature as the low resolution high-level encoder-output feature $f \in R^{h_s \times w_s \times c_s}$, and the guidance feature as the high-resolution low-level encoder feature $g \in R^{h_g \times w_g \times c_g}$ from the FXN. To improve the boundary accuracy, the DGD uses the low-level encoder feature, which preserves the detailed edge information of the input, as a guidance to refine the high-level encoder feature in a locally linear fashion.

Figure 2 illustrates the detailed steps of the proposed DGD. The high-level encoder feature $f \in R^{h_s \times w_s \times c_s}$ which encodes the semantic information, will be upsampled to the same resolution as the low-level encoder feature $g \in R^{h_g \times w_g \times c_g}$ after applying 1×1 convolutions for dimension reduction. Denote the upsampled feature as $f^{'} \in R^{h_g \times w_g \times c_s^{'}}$ and the guidance as $g^{'} \in R^{h_g \times w_g \times c_s^{'}}$. Then, the Deep Guided (DG) block will be applied for the input $f^{'}$ and $g^{'}$, to obtain the refined feature $r \in R^{h_g \times w_g \times c_s^{'}}$. The detailed architecture of Deep Guided Block (DGB) is illustrated in Fig. 3. Similar to [14], the DGB first calculates the variance of the guidance feature $g^{'}$, and the covariance between the semantic feature $f^{'}$ and the guidance feature $g^{'}$. Then, the variance and covariance will be used to obtain the coefficients of the locally linear model. The coefficients are then applied to the guidance to output the refined semantic feature in a linear fashion. 2 3×3 depthwise separable convolutions are applied, followed by 1x1 convolution to obtain the low resolution logits.

Our proposed DGD only use the low level encoder feature as the guidance, which saves the computational cost since $H > h_g > h_s$ and $W > w_g > w_s$. In contrast, the prior art of applying guided filtering take the input image which has high resolution $H \times W$ as the guidance and use it to refine the output of other methods.

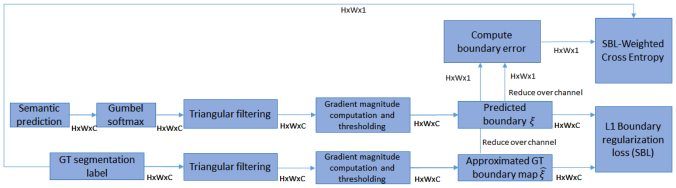


Fig. 4: The detailed steps in Semantic Boundary-Aware Learning.

B. Training Deep Guided Decoder Networks with SBAL

The DGD provides an efficient solution to improve semantic boundary accuracy by the utilization of the low-level encoder features with detailed low-level edge information. However, due to the fact that most prior arts of training semantic segmentation network only use a pixel-level class-prediction loss that only focus on the average class accuracy, it is not enough to achieve our goal of accurately predicting the boundaries between semantic classes without degrading the semantic class accuracy. Besides, the DGD might also introduce some unnecessary edge structures from the low level guidance feature.

To further refine any possible noisy boundary predictions, we then propose to train the DGD with a novel Semantic Boundary Aware Learning that can explicitly model the semantic boundary information and remove those unnecessary noisy information introduced from the guidance feature. Specifically, we train the DGD using our SBAL as replacement to the pixel-level loss to learn the semantic boundary and the semantic class information concurrently. The SBAL consists of channel-wise Semantic Boundary Loss (SBL) and SBL-Weighted Cross Entropy (SBL-WCE) loss. Both are modulated by the prediction errors in the boundary regions. Compared to the conventional training strategy (standard Softmax Cross Entropy loss), our SBAL can further boost the networks capability to learn semantic boundary information, Figure 4.

The input to the SBL are the per class gradient magnitude map $\xi \in \mathbb{R}^{H\times W\times C}$ computed from the predicted semantic segmentation map, and the per class gradient magnitude map $\hat{\xi} \in \mathbb{R}^{H\times W\times C}$ computed from the ground truth semantic segmentation label. The loss is a L_1 loss computed between $\hat{\xi}$ and ξ. A threshold $T = 0.1$ is applied to ξ and $\hat{\xi}$ to obtain the spatial binary masks M_ξ and $M_{\hat{\xi}}$. The mask is used to mask the following loss via elementwise multiplication ($\odot$):

$$L_{SBL} = 0.5 * L_1 + 0.5 * L_2, \tag{1}$$

where $L_1 = \frac{\sum_c \sum_j \sum_i M_{\xi_{i,j,c}} |\xi_{i,j,c} - \hat{\xi}_{i,j,c}|}{\sum_c \sum_j \sum_i M_{\xi_{i,j,c}}}$, and $L_2 = \frac{\sum_c \sum_j \sum_i M_{\hat{\xi}_{i,j,c}} |\xi_{i,j,c} - \hat{\xi}_{i,j,c}|}{\sum_c \sum_j \sum_i M_{\hat{\xi}_{i,j,c}}}$ for $i \in \{0, 1, 2, \cdots, H-1\}$, $j \in \{0, 1, 2, \cdots, W-1\}$, and $c \in \{0, 1, 2, \cdots, C-1\}$.

The semantic logits are perturbed with a Gumbel Softmax, a differentiable surogate of Argmax operation [10] and filtered with triangular filter in both horizontal and vertical dimensions, and channel-wise using a kernel of bandwidth 9. The filtering works in a way similar to dilation operation and remains differentiable. The gradient magnitude computation layer uses the kernel $[-0.5, 0, 0.5]$. There are no learnable parameters in these three layers.

The inputs to the SBL-WCE include the predicted semantic boundary volume ξ and the semantic boundary volume $\hat{\xi}$ computed from the ground truth label. The discrepancy between the two boundary volumes is computed as follows for each pixel (i,j):

$$\begin{aligned} w_{i,j} = 0.5 * R_c(M_{\xi_{i,j,c}}) \| R_c(\xi_{i,j,c}) - R_c(\hat{\xi}_{i,j,c}) \| +, \\ 0.5 * R_c(M_{\hat{\xi}_{i,j,c}}) \| R_c(\xi_{i,j,c}) - R_c(\hat{\xi}_{i,j,c}) \|, \end{aligned} \tag{2}$$

where $R_c(\cdot)$ denotes logic OR operation across channel dimension. The obtained spatial weight is applied to the Softmax cross entropy with ground truth $t_{i,j,c}$ and network prediction $y_{i,j,c}$ for each pixel (i,j) at class c.

$$Loss_{SBL-WCE} = -\sum_{i=0}^{H-1} \sum_{j=0}^{W-1} w_{i,j} \sum_{c=0}^{C-1} t_{i,j,c} \log(y_{i,j,c}). \tag{3}$$

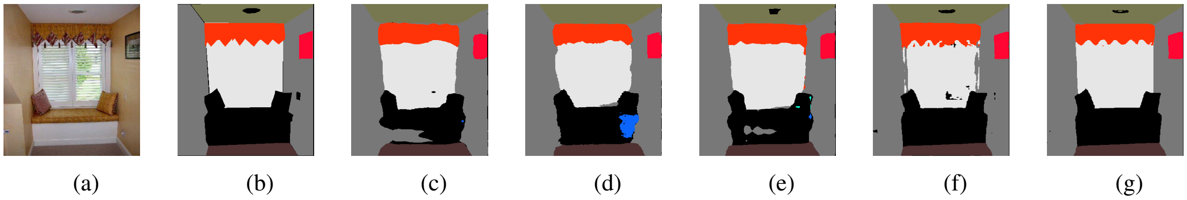
The total loss to train our DGD network is $L = \lambda_1 Loss_{SBL-WCE} + \lambda_2 L_{SBL}$.

IV. Experimental Results

A. Datasets, Metrics and Implementation Details

Cityscapes [4] dataset consists of 2,975, 500 and 1525 finely images for training, validation and test for 19 valid categories. The original ADE20K [16], [17] consists of 25,574 and 2000 labeled images for training and validation for 150 valid categories. In our experiment, we follow the setting in [12] designed for mobile platform scenarios. Specifically, only the most frequent 31 categories are considered as classes of interest (CoI). The remaining classes and the unlabeled class are combined into one class called Other class. All classes are included in training. We refer to this dataset as ADE20K-32. The mean Intersection over Union (mIoU) and the mean Boundary F1 score (mBF) [5] are used to measure the segmentation class accuracy and boundary accuracy, respectively.

All the trainings are initialized with the corresponding ImageNet pretrained checkpoints of the backbones. The augmentations including random cropping, random scaling in the range of [0.5, 2.0] and random horizontal flipping are applied to all the models. Polynomial learning rate decay with power 0.9, and weight decay 4×10^{-6} are used for all the models. All the models are trained for 35,000 warmup steps and 200,000 steps further. Crop size 512×1024 and 512×512 are used for Cityscapes dataset and ADE20K-32 dataset, respectively. In particular for ADE20K-32 dataset, we resize the longer dimension to 512 and padding the other dimension to 512 since the original images are of different sizes. As for Cityscapes dataset, MNEdge DeepLab based models used batch size 48, and Adam optimizer with base learning rate 0.002. The Xception65 DeepLabv3+ based models used batch size 24, and SGD optimizer with momentum 0.9 and base learning rate 0.007. As for ADE20K-32 dataset, MNEdge DeepLab



(a) (b) (c) (d) (e) (f) (g)

Fig. 5: Visualization of the results of two images from ADE20K-32. From (a) input image, (b) ground truth, (c) mnET, (d) +DGD, (e) +SBL, (f) +DGD+SBL, (g) +DGD+SBL+SBL-WCE

TABLE I: Ablation studies using MobilenetEdgeTPU (MNEdge) and Xception65 backbone on the Cityscapes dataset using single scale inference. $\Delta\%$ is the relative gain over the baseline

Model names	mIoU (%)	$\Delta\%$	mBF (%)	$\Delta\%$
MNEdge DeepLab	74.83	-	73.70	-
MNEdge DeepLab+DGD	76.28	1.9	75.54	2.5
MNEdge DeepLab+SBL	76.23	1.9	77.50	5.2
MNEdge DeepLab+DGD+SBL	76.96	2.8	79.11	7.3
MNEdge DeepLab +DGD+SBL+SBL-WCE (DeepGBASS)	**77.13**	3.1	**80.71**	9.5
Xception65 DeepLabv3+ [7]	78.79	-	77.86	-
Xception65 DeepLabv3+ +DGD+SBL+SBL-WCE (DeepGBASS)	**80.11**	1.7	**79.63**	2.3

based models used batch size 24, and the SGD optimizer with momentum of 0.9 and base learning rate 0.001. The Xception65 DeepLabv3+ based models used batch size 10 and the SGD with momentum 0.9 and base learning rate of 0.002.

B. Results on Cityscapes

Our ablation studies concluded from Table I that our approach can improve the boundary performance significantly for lightweight MNEdge backbone and it still provides improvement over the state-of-the-art DeepLab model using a heavier backbone Xception65.

In addition to show the effectiveness of our proposed DGD, we also investigated the possibility to reduce the complexity of DGD by reducing the channel size in the DGD. In Table II, we show that even our low complexity DGD-64 outperforms the DeepLabv3+ Decoder (DLv3D) especially in terms of both the boundary accuracy and complexity. In addition, we also observe that there is no performance drop when the DGD-256 is reduced to DGD-64, in comparison, there is some drop when DLv3D-256 is reduced to DLv3D-64. This is also a favorable property of our DGD and training with SBL. As a result, the reduced complexity enables more efficient inference on mobile devices.

C. Results on ADE20K-32

The experimental results shown in Table III confirmed that our approach can achieve 11.1% relative improvement in mIoU and 39.4% improvement in mBF. The ablation studies from Table III also show the effectiveness of each module of our approach. The segmentation results with our different models on sample images from the validation set are shown in Figure 5. For the model using heavier backbone, Xception65 Deeplabv3+ + DGD+SBL+SBL-WCE obtains 4.0% relative improvement in mIoU and 12.9% in mBF as shown in Table III.

TABLE II: Investigation of DGD channel size with Mobilenet-EdgeTPU as backbone on the Cityscapes dataset using single scale inference with 1024 × 2048 input resolution. DGD-k means the channel size ($c_s^{'}$) used in DGD is k. DLv3D-k is defined similarly.

Model	mIoU (%)	mBF (%)	GFLOPs	Params(M)
DLv3D-256	75.82	74.48	142	2.95
DLv3D-64	75.50	74.31	104	2.72
DGD-256	76.28	**75.54**	156	3.01
DGD-64	**76.32**	75.37	108	2.73

TABLE III: Ablation results of DGD and SBAL on DeepLab using MNEdge and Xception65 FXNs on ADE20K-32 dataset. $\Delta\%$ is the relative gain over the baseline

Model names	mIoU (%)	$\Delta\%$	mBF (%)	$\Delta\%$
MNEdge DeepLab	53.64	-	31.52	-
MNEdge DeepLab+DGD	55.14	2.8	35.00	11.0
MNEdge DeepLab+SBL	57.04	6.3	37.62	19.4
MNEdge DeepLab+DGD+SBL	59.23	10.4	43.64	38.5
MNEdge DeepLab +DGD+SBL+SBL-WCE (DeepGBASS)	**59.6**	11.1	**43.94**	39.4
Xception65 DeepLabv3+	60.77	-	41.63	-
Xception65 DeepLabv3+ +DGD+SBL	62.98	3.6	46.78	12.4
Xception65 DeepLabv3+ +DGD+SBL+SBL-WCE (DeepGBASS)	**63.18**	4.0	**47.00**	12.9

V. Conclusion

We introduced DeepGBASS, a framework to improve the class and boundary accuracy of semantic segmentation. Our ablation studies showed that each of the two components of DeepGBASS can independently improve both class and boundary accuracy, and can be applied to different network architectures. DeepGBASS with MobileNetEdgeTPU and Xception feature extractors can improve the accuracy of the state-of-the-art Deeplab networks without notable additional complexity on ADE20K-32 and Cityscapes datasets.

References

[1] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4545–4554, 2016.

[2] 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.

[3] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. CSPN++: learning context and resource aware convolutional spatial propagation networks for depth completion. In *The Thirty-Fourth Conference on Artificial Intelligence, AAAI 2020*, pages 10615–10622. AAAI Press, 2020.

[4] 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.

[5] Gabriela Csurka, Diane Larlus, Florent Perronnin, and France Meylan. What is a good evaluation measure for semantic segmentation?. In *BMVC*, volume 27, pages 10–5244, 2013.

[6] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6819–6829, 2019.

[7] Google. Mobilenetedgetpu deeplab. https://github.com/tensorflow/models/blob/master/research/deeplab/g3doc/model_zoo.md, 2020.

[8] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(6):13971409, 2013.

[9] 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.

[10] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[11] Qingfeng Liu, Mostafa El-Khamy, Dongwoon Bai, and Jungwon Lee. Gsanet: Semantic segmentation with global and selective attention. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1471–1475. IEEE, 2020.

[12] V. Reddi, D. Kanter, P. Mattson, J. Duke, T. Nguyen, R. Chukka, K. Shiring, K. Tan, M. Charlebois, W. Chou, M. El-Khamy, et al. MLPerf mobile inference benchmark: Why mobile AI benchmarking is hard and what to do about it. *arXiv preprint arXiv:2012.02328*, 2020.

[13] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5229–5238, 2019.

[14] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1838–1847, 2018.

[15] Y. Zhao, J. Li, Y. Zhang, and Y. Tian. Multi-class part parsing with joint boundary-semantic awareness. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9176–9185, 2019.

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017.

[17] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.