Multi-Exit Semantic Segmentation Networks

Alexandros Kouris[1], Stylianos I. Venieris*[,1], Stefanos Laskaridis*[,1], Nicholas Lane[1,2]

[1]Samsung AI Center, Cambridge [2]University of Cambridge *Indicates equal contribution.
{a.kouris,s.venieris,stefanos.l,nic.lane}@samsung.com

Supplemental Material

1 Overview

In the supplementary material of our paper, we provide details about the experimental setup of our work, as well as additional qualitative and quantitative data further showcasing the quality of results from MESS networks and how different training/inference schemes perform in practice. Finally, we discuss limitations of the current methodology and potentially promising future work directions.

2 Experimental Configuration

2.1 Datasets

MS COCO: MS COCO [13] forms one of the largest datasets for dense scene understanding tasks. Thereby, it acts as common ground for pre-training semantic segmentation models across domains. Following common practice for semantic segmentation, we consider only the 20 semantic classes of PASCAL VOC [4] (plus a *background* class), and discard any training images that consist solely of background pixels. This results in 92.5k training and 5k validation images. We set crop size (b_R) to 520×520.

PASCAL VOC: PASCAL VOC [4] comprises the most broadly used benchmark for semantic segmentation. It includes 20 foreground object classes (plus a *background* class). The original dataset consists of 1464 training and 1449 validation images. Following common practise we adopt the augmented training set provided by [6], resulting in 10.5k training images. For PASCAL VOC, b_R is also set to 520×520.

2.2 Baselines

To compare our work's performance with the state-of-the-art, we evaluate against the following approaches:

Single-Exit Segmentation Backbones:

- **DRN**: Dilated Residual Networks [19] approach for re-using classification pre-trained CNNs as backbones for semantic segmentation, by avoiding loss of spatial information. We use an FCN head at the end.
- **DLBV3**: DeepLabV3 [19,2], one of the leading approaches in semantic segmentation, employing Atrous Spatial Pyramid Pooling (ASPP).
- **segMBNetV2**: The lightweight MobileNetV2 segmentor presented in [16] with an FCN head.

Mutli-Exit Segmentation SOTA:
– **LC**: The early-exit segmentation work Deep Layer Cascade [12]. **LC** proposes a pixel-wise adaptive propagation in early-exit segmentation networks, with confident pixel-level predictions exiting early.

NAS Segmentation SOTA:
– **AutoDLB**: The NAS-based segmentation approach Auto-DeepLab [14], employing a differential formulation for hierarchical NAS, leading to high search efficiency. We target the Auto-DeepLab-M variant.

Multi-Exit Network Training:
– **E2E**[1]: The conventional end-to-end training method for early-exit classification networks, introduced by MSDNet [8] and BranchyNet [17].
– **Frozen**[1]: The conventional frozen-backbone training method for early-exit classification networks, proposed by SDN [9] and HAPI [11].

Distillation-based Training:
– **KD**[1]: The originally proposed knowledge distillation technique of [7].
– **SelfDistill**[1]: The popular self-distillation approach for early-exit classification networks, utilised in [21,15,20].

2.3 Training Protocol

MESS instances are built on top of existing segmentation networks, spanning across the workload spectrum in the literature, *i.e.* from the computationally heavy [1] to the lightweight [16]. Through MESS, SOTA networks can be *further optimised for deployment efficiency*, demonstrating complementary performance gains by saving computation on easier samples. A key characteristic of the proposed two-stage MESS training scheme, is that it effectively preserves the accuracy of the final (baseline) exit, while boosting the attainable results to earlier segmentation exits. This is achieved by bringing together elements from both the end-to-end and frozen-backbone training approaches (Sec. 4.3). Thus, we consider the employed training scheme decoupled from the attainable comparative results, as long as both the baseline and the corresponding MESS instances share the same training procedure. As such, in order to preserve simplicity in this work, we use a straightforward training scheme, shared across all networks and datasets, and refrain from exotic data augmentation, bootstrapping and multi-stage pre-training schemes that can be found in accuracy-centric approaches.

Hyperpameters: All MESS and baseline models are optimised using SGD, starting from ImageNet [3] pre-trained backbones. The initial learning rate is set to lr_0=0.02 and *poly* lr-schedule ($lr_0 \cdot (1 - \frac{iter}{total\ iter})^{pow}$) [18] with pow=0.9 is employed. Training runs over 60k iterations in all datasets. Momentum is set to 0.9 and weight decay to 10^{-4}. We re-scale all images to a dataset-dependent base resolution b_R. During training we conduct the following *data augmentation* techniques: random re-scaling by 0.5× to 2.0×, random cropping (size: 0.9×) and random horizontal flipping ($p = 0.5$). For Knowledge Distillation, we experimentally set α to 0.5. For the overprovisioned network, we experimentally found N=6 to provide a good balance between search space granularity and size, for the examined backbones.

[1] Adapted for segmentation by incorporating dense predictions.

2.4 Inference Process

The main optimisation objective of this work is *deployment efficiency*. This renders impractical many popular inference strategies that are broadly utilised in the literature when optimising solely for accuracy, as they incur prohibitive workload overheads. As such, in this work, we refrain from the use of ensembles, multi-grid and multi-scale inference, image flipping, etc. Instead, in the context of this work, both MESS and baseline networks employ a straightforward single-pass inference across all inputs.

3 Comparison with Uniform Exit Architectures

Compared to a direct adoption of classification-based approaches that employ a uniform exit architecture across the depth of the backbone [8,17,9,11], MESS networks provide a significantly improved performance-accuracy trade-off that pushes the limits of efficient execution for semantic segmentation, by providing a highly customisable architectural configuration space for early exits, searched though our framework.

To demonstrate this, Fig. 7 illustrates the mean and per-label accuracy of multi-exit network instances, customised in view of requirements ranging between $1\times$ and $4.5\times$ lower latency compared to the original backbone. On the left side we depict baseline networks using a *uniform exit architecture* of FCN heads across all candidate exit points. In contrast, on the right side we examine MESS networks incorporating *tailored early-exit architectures* from the proposed search space. The results indicate significant accuracy (mIoU) gains by exploiting the proposed head, ranging up to **19.3 pp** (10.9 pp on avg.), across the examined latency budgets. This demonstrates that it is essential to re-design the segmentation heads for multi-exiting scenarios. Repeating the same experiment, but optimising for latency under an accuracy constraint, MESS reduces FLOPs by up to **3**$\times$ (2.4$\times$ on avg.), across varying accuracy targets.

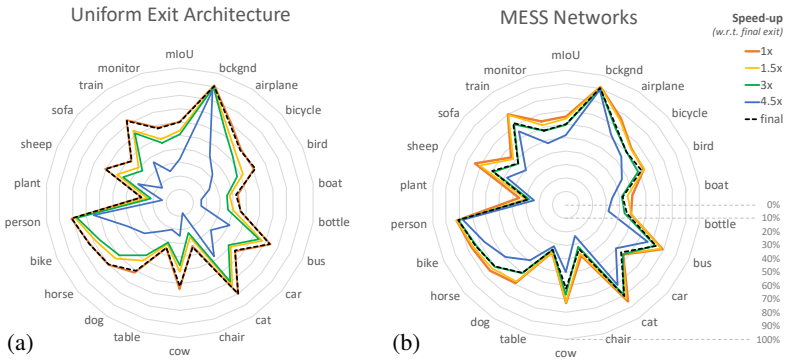

Fig. 7: Per-label IoU on MS COCO Validation Set of (a) uniform architecture early-exit networks and (b) MESS networks configured for different latency goals (expressed as speed-ups with respect to the final exit), for ResNet50. Each of the concentric spider graphs defines a distribution of exit rates over different heads, based on the confidence threshold determined by our search. The graphs have value both when studied in isolation, so as to monitor the behaviour of early exits across labels, and in comparison, to understand what our principled design approach offers.

4 Qualitative Evaluation

Fig. 8 depicts the qualitative difference of semantic map outputs with and without the proposed distillation mechanism incorporated during training the respective MESS models. The two samples of the figure show clearly the kind of per-pixel prediction errors that our PFD scheme tries to alleviate.

Fig. 9 demonstrates the quality-of-result for progressive segmentation outputs though a MESS network, for certain samples from MS COCO and PASCAL VOC. Table 8 shows the respective accuracy for the same set of images, along with the selected output, based on our exit policy.

Fig. 10 tells the story from a different standpoint, where we showcase how confident an exit of a MESS network is about its predictions. Concretely, we illustrate the (per-pixel) confidence heatmap for an early segmentation head and the final exit for certain samples of the same datasets, with and without Eq. (9). This demonstrates how our confidence-based mechanism works in the realm of semantic segmentation and the contribution our edge smoothing technique in the confidence of predictions along object edges. Similarly, Table 9 shows the single (per-image) confidence values for each prediction, obtained both through a baseline and the proposed method.

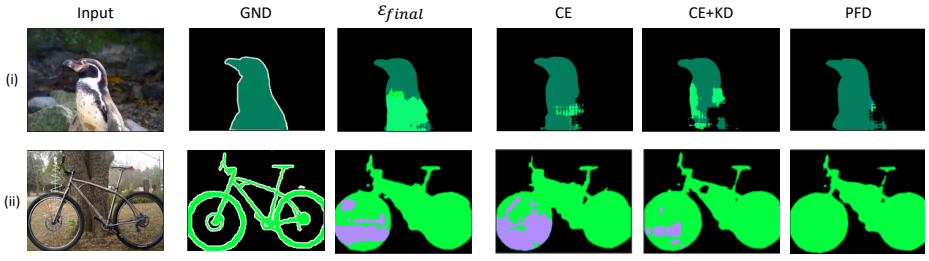

Fig. 8: Qualitative examples of semantic segmentation with different distillation schemes on ResNet50. From left to right we see the input image, the ground truth semantic map, the output of the final exit ($\mathcal{E}_{\text{final}}$) and the output of an early exit without distillation (CE), with baseline distillation (CE+KD) and with the proposed positive filtering distillation (PFD) approach. The proposed scheme aims to control the flow of information to early exits in knowledge distillation, "paying more attention" to pixels that are correctly predicted by the final exit during training, while avoiding to use contradicting CE and KD reference signals in the remainder of the image. The provided samples illustrate the kind of information that can be learned by means of the PFD scheme, even in the case of the original final exit being incorrect for certain pixels.

Fig. 9: Visualisation of per-exit semantic segmentation outputs of MESS ResNet50 for specific samples from MS COCO and PASCAL VOC. The first column represents the input image and last the ground truth labels. Intermediate columns represent the output per exit head.

Table 8: Per-exit semantic segmentation accuracy for samples of Fig. 9 from MS COCO and PASCAL VOC Validation Set. *mIoU* represents the normalised mean IoU per image and *pAcc* the pixel-accuracy (excl. True Positives on *background* class) for each exit head. We denote the selected exit for each sample, determined by the proposed MESS exit policy, with **bold** font. Different exits have different confidence thresholds, selected during search, so as to lead to ≤ 1 pp of accuracy degradation.

Sample	$\mathcal{E}_1$		$\mathcal{E}_2$		$\mathcal{E}_3$		$\mathcal{E}_4$		$\mathcal{E}_{final}$	
	mIoU	pAcc	mIoU	pAcc	mIoU	pAcc	mIoU	pAcc	mIoU	pAcc
(i)	94.79%	90.70%	**95.91%**	**92.69%**	95.90%	92.67%	95.83%	92.53%	95.82%	92.53%
(ii)	**92.89%**	**86.78%**	90.40%	82.12%	91.93%	84.98%	93.32%	87.58%	93.35%	87.62%
(iii)	84.61%	82.30%	88.68%	92.29%	**88.98%**	**92.77%**	89.16%	93.07%	88.19%	93.07%
(iv)	83.18%	68.73%	95.67%	91.82%	**98.51%**	**97.58%**	99.03%	98.53%	98.97%	98.45 %
(v)	73.06%	65.14%	81.11%	82.06%	96.68%	90.01%	**97.64%**	**98.31%**	90.78%	90.27%
(vi)	92.47%	92.87%	95.24%	95.34%	**95.46%**	**95.54%**	95.43%	95.53%	95.25%	95.34%
(vii)	93.07%	89.88%	**94.49%**	**92.72%**	94.59%	92.87%	94.56%	92.85%	94.54%	92.78%
(viii)	93.98%	90.75%	**95.58%**	**93.09%**	95.47%	93.12%	92.98%	93.89%	95.31%	92.10%

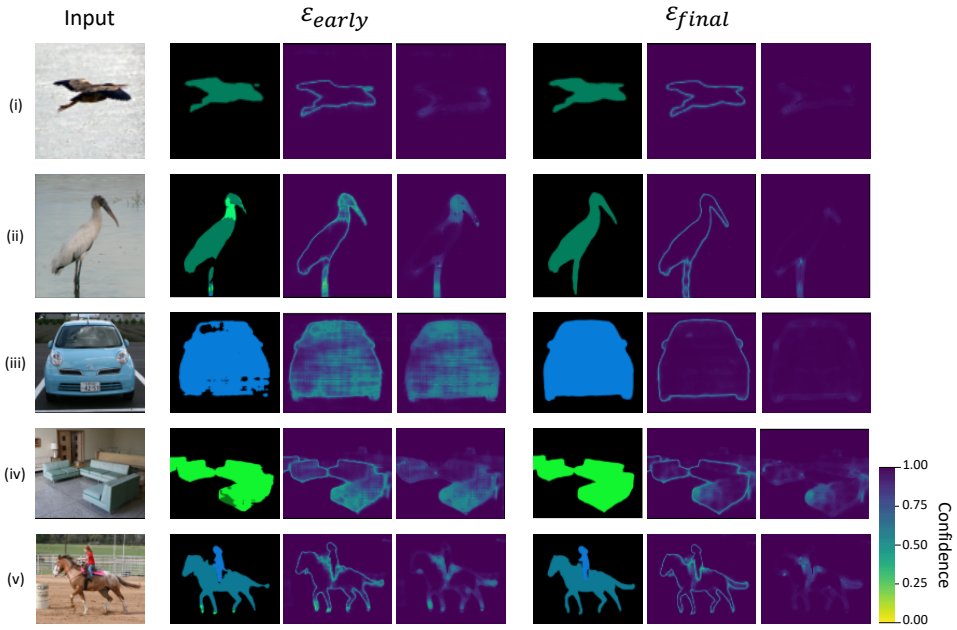

Fig. 10: Per-pixel confidence maps for an early and the final segmentation heads for ResNet50 with and without integrating Eq. (9). A solid purple colour is illustrating the best possible result, where the MESS network is most confident about the per pixel labels. With our edge smoothing technique (Section 5.2) we witness a more pragmatic confidence estimation for predictions along the edges of objects for both exits. This is especially important for early heads of MESS networks, as it helps distinguishing between "truly" under-confident and edge-rich predictions, leading to higher early-exit rates with respective latency gains.

Table 9: Per-image confidence values for samples illustrated in Fig. 10, reduced from the respective per-pixel confidence maps, through a baseline confidence-averaging approach and the proposed technique (considering the percentage of confident pixels at the output) with and without integrating the semantic edge confidence smoothing of Eq. (9). Using the proposed per-image confidence estimation methodology for dense predictions, a better separation is achieved between confident predictions (corresponding to higher quality-of-result outputs) and less confident predictions (prone to semantic errors).

Sample	$\mathcal{E}_{early}$			$\mathcal{E}_{final}$		
	mean(c^{map})	Eq. (7)	Eq. (7) + (9)	mean(c^{map})	Eq. (7)	Eq. (7) + (9)
(i)	0.990	0.975	0.999	0.992	0.977	1.000
(ii)	0.968	0.920	0.946	0.983	0.951	0.988
(iii)	0.871	0.595	0.599	0.983	0.958	0.999
(iv)	0.905	0.725	0.734	0.958	0.874	0.918
(v)	0.967	0.913	0.939	0.975	0.924	0.966

5 Discussion and Future Work

MESS networks offer considerable computational gains by alleviating redundancies across the depth dimension of the backbone network, skipping unnecessary computation in a difficulty-aware manner. In contrast, other efficient model design methodologies, such as NAS [14], can attenuate redundancy in more dimensions (*e.g.* number of channels and spatial resolution) at the cost of prolonged training, search and inference times. With the two approaches having different benefits and capitalising on completely orthogonal directions to obtain efficiency gains (NAS focuses on uniformly eliminating redundancy *on the backbone*, whereas MESS *enhances a given backbone* with early exits to minimise computational redundancy in an adaptive-inference manner) future work could combine the two, by applying the MESS methodology on top of a NAS-crafted backbone, realising complementary performance gains. Moreover, softmax-based confidence can be an artificial proxy for measuring a network's uncertainty [11,10,5], applicable to classification. Thus, alternative, trainable exit policies and metrics can be a promising avenue of research.

References

1. Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
2. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.
3. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
4. 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 (IJCV)*, 88(2):303–338, 2010.
5. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning (ICML)*, 2017.
6. Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic Contours from Inverse Detectors. In *International Conference on Computer Vision (ICCV)*, pages 991–998, 2011.
7. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. In *NeurIPS 2014 Deep Learning Workshop*, 2014.
8. Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Weinberger. Multi-Scale Dense Networks for Resource Efficient Image Classification. In *International Conference on Learning Representations (ICLR)*, 2018.
9. Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. Shallow-Deep Networks: Understanding and Mitigating Network Overthinking. In *International Conference on Machine Learning (ICML)*, 2019.
10. Stefanos Laskaridis, Alexandros Kouris, and Nicholas D. Lane. Adaptive Inference through Early-Exit Networks: Design, Challenges and Directions. In *Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning (EMDL)*, page 1–6, 2021.

11. Stefanos Laskaridis, Stylianos I. Venieris, Hyeji Kim, and Nicholas D. Lane. HAPI: Hardware-Aware Progressive Inference. In *International Conference on Computer-Aided Design (ICCAD)*, 2020.
12. Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not All Pixels Are Equal: Difficulty-aware Semantic Segmentation via Deep Layer Cascade. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3202, 2017.
13. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.
14. Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 82–92, 2019.
15. Mary Phuong and Christoph H Lampert. Distillation-based Training for Multi-Exit Architectures. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1355–1364, 2019.
16. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.
17. Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2464–2469. IEEE, 2016.
18. Yanzhao Wu, Ling Liu, Juhyun Bae, Ka-Ho Chow, Arun Iyengar, Calton Pu, Wenqi Wei, Lei Yu, and Qi Zhang. Demystifying Learning Rate Policies for High Accuracy Training of Deep Neural Networks. In *IEEE International Conference on Big Data (Big Data)*, pages 1971–1980, 2019.
19. Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated Residual Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 472–480, 2017.
20. Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.
21. Linfeng Zhang, Zhanhong Tan, Jiebo Song, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. SCAN: A Scalable Neural Networks Framework Towards Compact and Efficient Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.