Tilted Cross-Entropy (TCE): Promoting Fairness in Semantic Segmentation (Supplementary Material)

A. More Evaluation Results

Table 4: Performance comparison on ADE20k *validation* set, bottom 22 classes of MCCE.

Method	land	lake	shower	blanket	step	hill	bag	crt scrn.	tray	stage	hovel	dirt track
MCCE [29]	0.00	1.96	**2.62**	2.95	5.47	**5.52**	5.81	6.12	6.94	**7.66**	8.77	**10.37**
TCE $t=.1$	**3.58**	**34.94**	1.44	**7.13**	**7.54**	3.46	6.04	**8.50**	5.58	6.82	**9.14**	1.71
TCE $t=1$	1.08	14.64	2.30	4.60	3.19	4.76	**6.19**	0.86	**9.18**	7.37	8.49	2.04
continued	truck	river	bannister	canopy	glass	fountain	barrel	box	pole	tower	mIoU	mIoU(all)
MCCE [29]	10.70	**10.71**	11.55	**11.62**	12.63	12.67	17.68	**18.57**	19.37	19.48	9.51	**43.87**
TCE $t=.1$	12.85	2.88	12.27	9.37	12.83	**18.92**	f30.93	10.58	16.53	19.98	**11.56**	41.32
TCE $t=1$	**17.17**	10.49	**12.78**	10.51	**13.34**	15.87	**45.20**	17.36	**19.61**	18.74	11.17	41.77

Table 5: Performance comparison on ADE20k *validation* set, top 22 classes of MCCE.

Method	sky	p. table	bed	toilet	road	ceiling	car	building	tent	bus	floor	person
MCCE [29]	**93.92**	**91.99**	**86.39**	83.02	**82.70**	81.76	81.66	**81.41**	80.50	79.12	**78.91**	**78.42**
TCE $t=.1$	93.81	91.79	85.80	84.58	80.78	81.40	80.69	79.31	84.76	79.76	77.82	76.64
TCE $t=1$	**93.97**	90.27	85.33	**84.78**	80.97	**81.94**	**81.85**	80.18	**91.93**	**83.64**	77.47	77.27
continued	stove	wall	cradle	bathtub	tree	painting	dish w.	refrig.	runway	sink	mIoU	mIoU(all)
MCCE [29]	**76.14**	**74.68**	73.67	72.66	**71.14**	**70.96**	**70.70**	70.62	**70.48**	69.56	**78.20**	**43.87**
TCE $t=.1$	68.66	74.08	**80.17**	70.95	70.01	64.60	67.89	68.46	62.61	69.65	77.25	41.32
TCE $t=1$	70.84	73.92	77.89	**75.39**	70.59	66.24	64.31	**74.19**	66.98	**69.70**	**78.17**	41.77

Table 6: Fairness criteria for ADE20k [31].

	sorted 15%		(15^{th}perc., mIoU)		overall	
Method	bottom	top	bottom	top	worst	std.
MCCE [1]	9.51	78.20	(19.55, **9.51**)	(69.14, 78.20)	0.0	**21.95**
TCE $t=.1$	**11.56**	77.25	(15.31, 8.04)	(67.95, 77.62)	**1.44**	22.39
TCE $t=1$	11.17	78.17	(15.67, 7.99)	(67.26, 78.49)	0.86	22.79

We further assessed the impact of the proposed tilted cross-entropy (TCE) on yet another commonly adopted datasets for semantic segmentation; i.e., ADE20k [31]. ADE20k contains $20,100$ images for training and $2,000$ for validations from 150 object and stuff classes. As such, compared to Cityscapes [3], the typical scenes in ADE20k can be more complex in that they can potentially contain more target classes per image. For experiments on ADE20k, we have used the UPerNet [29] with ResNet-101 backbone as our reference implementation of multi-class cross-entropy (MCCE) and on top that we have implemented the proposed TCE. UPerNet is among the top performing model architectures for ADE20k. Following [29], we used minibatch SGD with learning rate $l_r = 0.01$ and momentum 0.9 for all models, and adjusted for a total minibatch size of 8. The reported results of [29] are based on our own trainings, for the sake of a fair comparison. Image crop size and other pre/post-processing parameters are set per default as suggested in [29]. Our evaluation strategy is exactly the same as explained for Cityspaces in Section 3 except that here we consider sorted 15% (bottom and top 22 classes) and bottom and top 15th percentile. This is because ADE20k contains much larger number of classes compared to Cityscapes (150 vs 19).

Table 4, compares the sorted mIoU breakdown of MCCE (UperNet [29]) for its bottom 15% (22) classes against the same model trained with TCE. Here, again we see improvement in the low-performing classes, which is also reflected in the mIoU of these 22 classes (for both $t = 0.1$ and 1). Conversely, the overall mIoU (denoted as mIoU(all)) has dropped for ADE20k irrespective of the choice of the tilting parameter t. To reiterate, TCE favors a less varied (and more fair) performance across classes, and not an improved overall mIoU. Table 5 provides the top part. As can be seen, here for most classes, MCCE outperforms TCE which confirms that the improvement in bottom (low-performing) classes is coming at the cost of performance degradation in top performing classes. The complete MIoU breakdown of ADE20k with 150 classes would not fit in two tables. Instead Table 6 summarizes the fairness measures for ADE20k. Here, the trend is less consistent compared to Cityscapes. TCE shows improvement in sorted percentage and overall fairness measures, but in percentile analysis this is not visible. This could be because typical ADE20k images contain more target classes. As such, in Algorithm 1, every time we

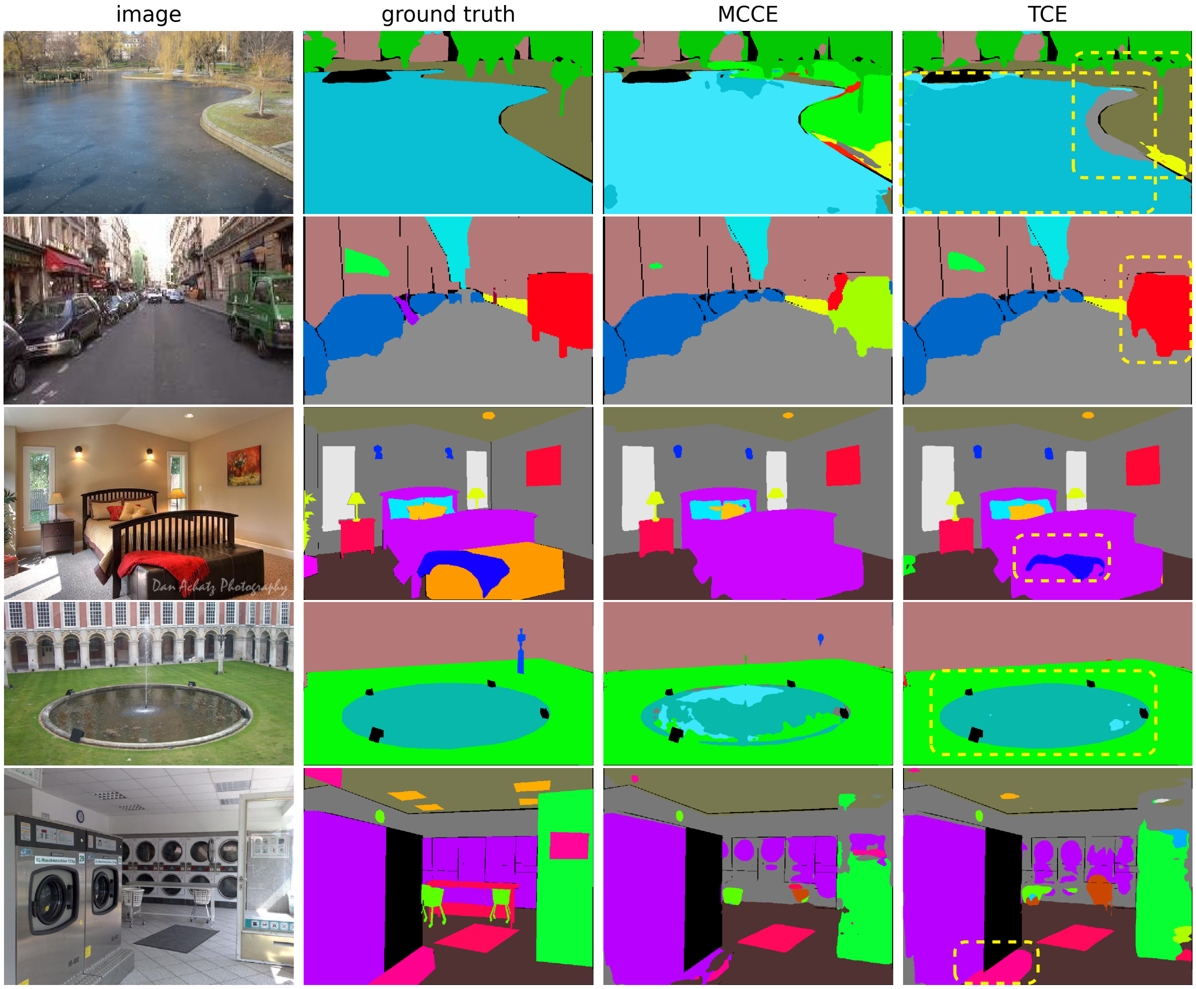

Figure 2: Impact of TCE on improving low-performing classes of MCCE. Best view in color with 300% zoom. From the top row to bottom, tilted cross-entropy (TCE) is providing a more consistent label map compared to standard multi-class cross-entropy (MCCE) for "lake", "truck", "blanket", "fountain" and "step" all which lie within the low-performing bottom 15% (22) classes of ADE20k.

samples from $\mathcal{D}_c^t$ to improve class c, we also involve several other classes. A possible remedy could be to tilt at class level per image. Finally, Figure 2 provides further qualitative examples illustrating the impact of TCE on ADE20k dataset. From the top row to bottom, TCE is providing a more consistent label map compared to MCCE for "lake", "truck", "blanket", "fountain" and "step" all which lie within the low-performing bottom 15% (22) classes of ADE20k.

B. Acknowledgment

The authors would like to thank Shell Global Solutions International B.V. and Delft University of Technology (TU Delft) for their support and for the permission to publish this work. The authors also extend their appreciation to Ahmad Beirami from Facebook AI for helpful discussions on TERM, and Jan van Gemert from CV Lab at TU Delft for constructive feedback.