Supplementary Material for Domain Adaptive Knowledge Distillation for Unsupervised Semantic Segmentation

Divya Kothandaraman Athira Nambiar Anurag Mittal
ramandivya27@yahoo.in, {anambiar,amittal}@cse.iitm.ac.in
Indian Institute of Technology Madras, India

1. Implementation details

The segmentation network is optimized with Stochastic Gradient Descent (SGD) optimizer (with Nesterov acceleration), where the weight decay is 1e-4 and the momentum is 0.9. The initial learning rate is set at 2.5e-4 and is decreased with a polynomial decay of 0.9. To train the discriminator, we use the Adam optimizer with a learning rate of 10e-4. The polynomial decay is the same as that of the segmentation network. All our experiments are performed on a single NVIDIA GEForce 11 GB GPU, with a batch size of 2.

2. Quantitative results: Ablation studies

Table 1 shows class-wise performance for the ablation experiments conducted on the various distillation losses proposed. The ablation corresponds to the source distillation paradigm (paradigm a in the distillation paradigms figure), and the experiments have been conducted on the real-to-real case. To deduce the performance of the individual distillation loss functions, we conduct extensive ablation studies on the real-to-real adaptation case. We conduct these studies on distillation paradigm case (a). The rationale is that these results should scale for the other three cases as well.

2.0.1 Impact of various loss functions:

The impact of various loss functions has been discussed in the paper.

Class-wise performance: As with most segmentation models, we notice that our domain-adaptive distilled model performs particularly well on classes such as road, car, vegetation, sky, etc. which have a huge presence in the dataset. Rare classes such as trains have a high probability of being confused with bus; truck and bus can be confusing to differentiate - these can in fact be wrongly classified as cars; wall and fence can be ambiguous and so on. We also notice that detection of small objects like traffic signs and persons in some images gets missed out. This can be attributed to multiple reasons - size of the object, rare occurrence and nuanced boundaries. While our proposed model outperforms both the teacher and the student in most categories, the trends of these models across categories are very similar. Thus, we believe that these issues are innate to the baseline domain adaptation and segmentation models.

3. Qualitative Results

In this section, we present visual results for our proposed pipeline 'domain-adaptive distillation'. The evaluation is done on the target domain of the student network. The nomenclature is as follows:

- Image: Target domain input image on which evaluation is done
- GT: Corresponding ground truth
- Teacher: Teacher network output for the target domain image
- Student: Student network output for the target domain image
- Source distillation (a): Output of student network distilled as per distillation paradigm (a) (Source domain distillation), evaluated on the target domain image (All distillations are as per Fig. 2 in the paper)
- Target distillation (b): Output of student network distilled as per distillation paradigm (b) (Target domain distillation), evaluated on the target domain image
- Source + target distillation (c): Output of student network distilled as per distillation paradigm (c) (Source + target domain distillation), evaluated on the target domain image
- Target init distillation (d): Output of student network distilled as per distillation paradigm (d) (Target domain distillation, initialised with case (c)), evaluated on the target domain image

Table 1: Ablation studies: Impact of various distillation losses for source distillation on the real-to-real case.

Parameter(λ)	mIoU	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MBike	Bike	mAcc
(i) KL divergence (L_{KL})																					
0.1	39.54	90.08	50.58	78.99	16.41	20.71	24.43	19.44	35.11	80.65	28.63	73.07	47	15.35	78.32	23.19	33.17	0.45	7.98	27.78	86.65
0.4	**40.27**	90.63	51.2	79.53	18.11	21.67	24.96	20.75	35.54	81.29	29.08	74.86	47.71	14.81	79.65	23.94	34.39	0.46	8.32	28.28	**87.15**
0.7	40.0	90.09	50.44	78.96	18.74	21.31	23.97	20.19	35.48	81.29	29.6	74.2	46.97	14.83	78.68	23.23	33.97	1.5	7.61	29.03	86.81
1.0	38.94	89.43	50.27	78.39	17.1	20.35	23.59	18.23	34.28	80.45	28.48	71.35	46.52	14.66	77.28	21.06	32.46	0.93	6.21	28.89	86.18
(ii) MSE loss (L_{MSE})																					
0.005	38.91	89.61	50.04	78.39	17.55	20.4	22.99	17.63	33.96	80.45	28.54	71.88	46.14	14.28	77.73	20.86	32.83	1.01	6.91	28.04	86.25
0.05	38.71	89.94	50.56	78.55	16.34	20.8	21.38	17.39	33.77	80.18	28.66	72.73	45.98	12.06	78.51	21.37	33.48	1.16	5.45	27.27	86.44
0.01	**39.86**	90.92	50.78	79.45	19.59	21.23	23.55	18.82	36.34	81.06	29.12	75.31	47.05	12.5	79.99	23.05	33.07	0.86	7.48	27.21	**87.17**
(iii) Cross entropy quasi teacher labels ($L_{CE-quasiT}$)																					
0.001	40.15	90.43	51.19	79.45	17.11	21.58	24.47	19.81	35.36	81.25	29.63	74.67	47.68	13.98	79.37	24.8	34.19	0.6	8.77	28.34	87.05
0.01	40.6	90.81	52.09	79.75	16.97	22.64	25.68	20.9	35.54	81.07	29.14	74.45	48.48	17.58	80.1	24.17	33.36	0.66	10.22	27.79	87.22
0.05	40.72	91.17	51.15	79.65	18.25	22.5	25.54	21.42	35.81	81.58	30.55	75.66	48.05	15.54	80.05	24.28	34.46	0.44	8.99	28.57	87.4
0.1	41.01	91.48	51.91	79.82	18.32	22.75	25.86	21.54	35.69	81.7	30.47	76.21	47.68	16.3	80.28	25.17	34.93	0.49	9.14	29.39	87.59
0.5	41.45	91.98	52.62	80	18.58	23.55	25.72	21.67	36.44	82.29	32.84	76.98	47.8	14.87	81.08	25.76	36.33	0.29	8.6	30.2	87.98
1.0	**42.18**	92.27	55.57	80.26	19.2	24.6	25.66	21.98	36.03	82.88	34.75	77.52	48.42	17.72	82.06	24.57	37.29	0.14	9.32	31.2	**88.28**
(iv) Combination of the loss terms KL, MSE (L_{KL}+L_{MSE})																					
0.4, 0.01	39.89	90.83	50.88	79.42	17.11	21.29	24.66	20.2	34.78	81.03	30.08	74.75	47.23	14.3	79.85	23.15	32.26	0.39	8.03	27.65	87.15
0.1, 0.01	**40.32**	91.04	51.11	79.82	17.35	21.55	25.57	20.58	34.46	81.26	29.92	75.67	47.59	15.76	80.55	24.04	33.79	0.3	8.63	27.1	**87.37**
0.7, 0.05	39.7	91.04	50.92	79.37	16.83	21.24	24.19	20.51	33.92	80.8	30.43	75.02	46.48	13.71	80.13	22.73	31.51	0.56	7.17	27.76	87.17
(v)Combination of the loss terms KL, CE-quasiT ($L_{KL} + L_{CE-quasiT}$)																					
0.1, 0.1	41.18	92.13	52.6	80.18	18.8	22.89	26.28	22.18	35.35	81.85	32.11	77.04	48	15.02	81.54	25.38	33.85	0.3	7.53	29.4	87.98
0.1, 1	**41.92**	92.4	53.8	80.44	18.62	24.52	25.61	22.17	36.95	82.63	33.45	78.14	48.18	14.92	81.72	26.63	35.9	0.09	8.95	31.43	**88.3**
(vi) Combination of the loss terms MSE, CE-quasiT ($L_{MSE} + L_{CE-quasiT}$)																					
0.01, 1	**42.2**	92.62	54.64	80.71	18.19	23.17	26.04	22.63	35.32	82.98	34.49	78.13	48.35	16.19	82.93	27.17	37.28	0.04	8.72	32.17	**88.54**
(vii)Combination ($L_{KL} + L_{MSE} + L_{CE-quasiT}$); λ_{KL} =0.1, λ_{MSE} = 0.01, $\lambda_{CE-quasiT}$ = 1.0																					
Case (a)	**42.33**	92.41	53.91	80.57	19.3	22.89	26.88	23.03	36.05	82.59	34.67	77.42	48.39	15.39	82.8	27.57	40.04	0.03	9.22	31.12	**88.42**

3.1. BDD to cityscapes

This section has visual results for the real-to-real adaptation case: Berkeley Deep Drive to Cityscapes. (Fig. 1)

3.2. GTA5 to cityscapes

This section has visual results for the synthetic-to-real adaptation case: GTA5 to Cityscapes. (Fig. 2)

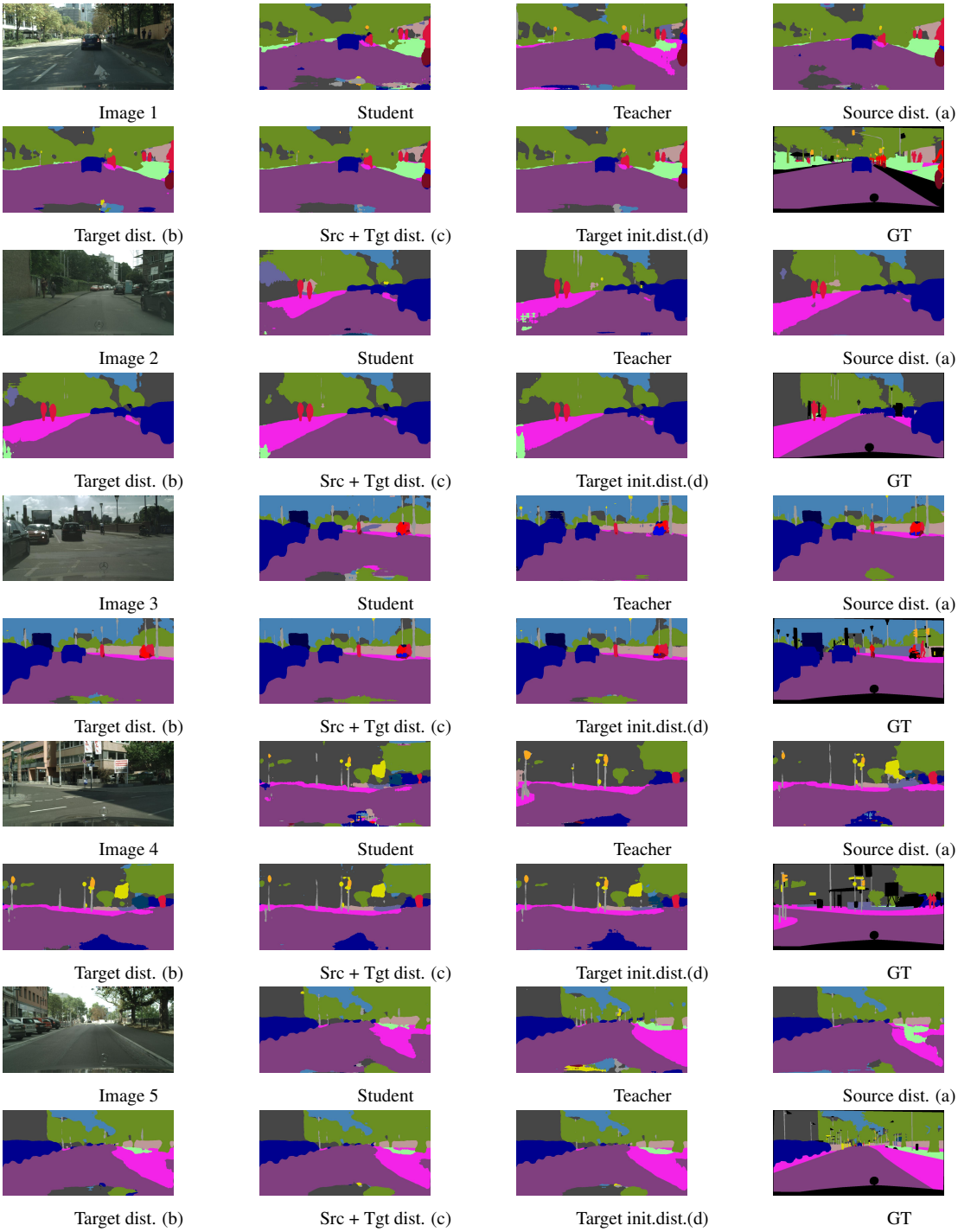

Figure 1: Visual results: BDD to Cityscapes

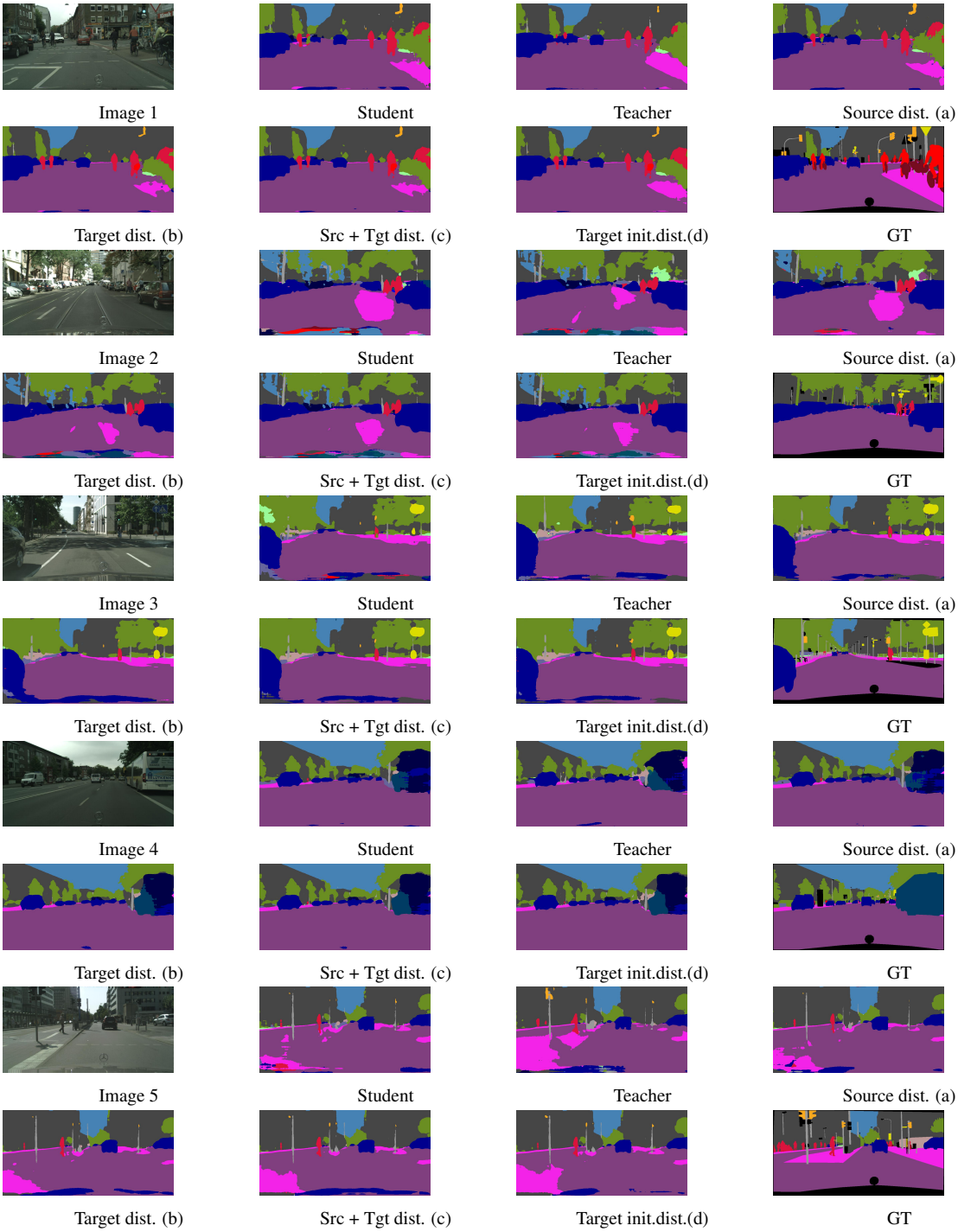

Figure 2: Visual results: GTA5 to Cityscapes