Supplementary Material for: Unsupervised Domain Adaptation in Semantic Segmentation via Orthogonal and Clustered Embeddings

Marco Toldo, Umberto Michieli, Pietro Zanuttigh
Department of Information Engineering, University of Padova
{toldomarco,umberto.michieli,zanuttigh}@dei.unipd.it

1. Qualitative Results

In Figure 1 we show some additional qualitative results of the proposed model when adapting source knowledge either from the GTA5 or the SYNTHIA datasets to the Cityscapes one. In these figures we can appreciate a robust improvement with respect to the baselines: e.g., the person in row 1, the sidewalk in row 2, the traffic sign in row 3, the road in rows 4, 5, 7 and 8. Additionally, shapes and details are more refined and localized.

2. Ablation Studies

In this section we provide further analysis for each novel loss component of our complete framework.

2.1. Clustering Loss

The effect introduced by the clustering objective is investigated by means of the t-SNE tool [2]. The results are reported in Figure 2. In particular, we extract all feature vectors from a single target image and reduce their dimensionality from 2048 to 2, so that we able to visualize their disposition. Each single feature instance is then associated to a semantic class from the ground-truth segmentation map, with the categorization expressed by the standard color map we jointly report. To allow for the analysis of the aggregating effect of the clustering module, in Figure 2 we show the t-SNE plots in the *source only* scenario, when only $\mathcal{L}_{cl}$ is enabled and when all adaptation modules are turned on ($\mathcal{L}_{tot}$). The effect brought by the clustering constraint is twofold. From one hand, we can see how features of the same class are more tightly clustered when $\mathcal{L}_{cl}$ is enabled, effect which is even further amplified by the class-conditional structural regularization provided by the other components of our work (see $\mathcal{L}_{tot}$). For instance, this is particularly visible in *road* and *vegetation* in row 2. From the other hand, we can appreciate that features belonging to different classes are more easily spaced apart, as shown in *car* or *person* in row 1 or in *traffic light* in row 3.

2.2. Orthogonality Loss

We investigate the contribution of the orthogonality constraint $\mathcal{L}_{or}$ via a similarity score defined as an average class-wise cosine similarity measure. The cosine distance is first computed for every pair of feature vectors from a single target image. Then, the average values are taken over all features from the same class to get a score for each pair of semantic classes. The final values are computed by averaging over all images from the Cityscapes validation set.

In Figure 3 we analyze the orthogonalizing action in three different configurations and we can clearly see that $\mathcal{L}_{or}$ causes the similarity score to significantly increase on almost all the classes (higher similarity reflects into lower orthogonality). To show the effect of the orthogonality constraint alone, we compare the *source only* similarity score and the scenario with only $\mathcal{L}_{or}$ enabled (Figure 3a). To investigate its effect when all the loss components are enabled, we compared the similarity scores of the full approach with respect to *source only* (Figure 3b) and to the case where all components but the $\mathcal{L}_{or}$ are enabled (Figure 3c). The results are robust and coherent in showing an increased similarity score in the configurations where $\mathcal{L}_{or}$ is active.

The same considerations are visible in Figure 4 in which the matrices of class-wise similarity scores are reported for the three aforementioned scenarios. In particular, we can see how the diagonal is much brighter (i.e., high similarity for classes with themselves) when $\mathcal{L}_{or}$ is added, while off-diagonal entries are darker.

Our work was in part supported by the Italian Minister for Education (MIUR) under the "Departments of Excellence" initiative (Law 232/2016).

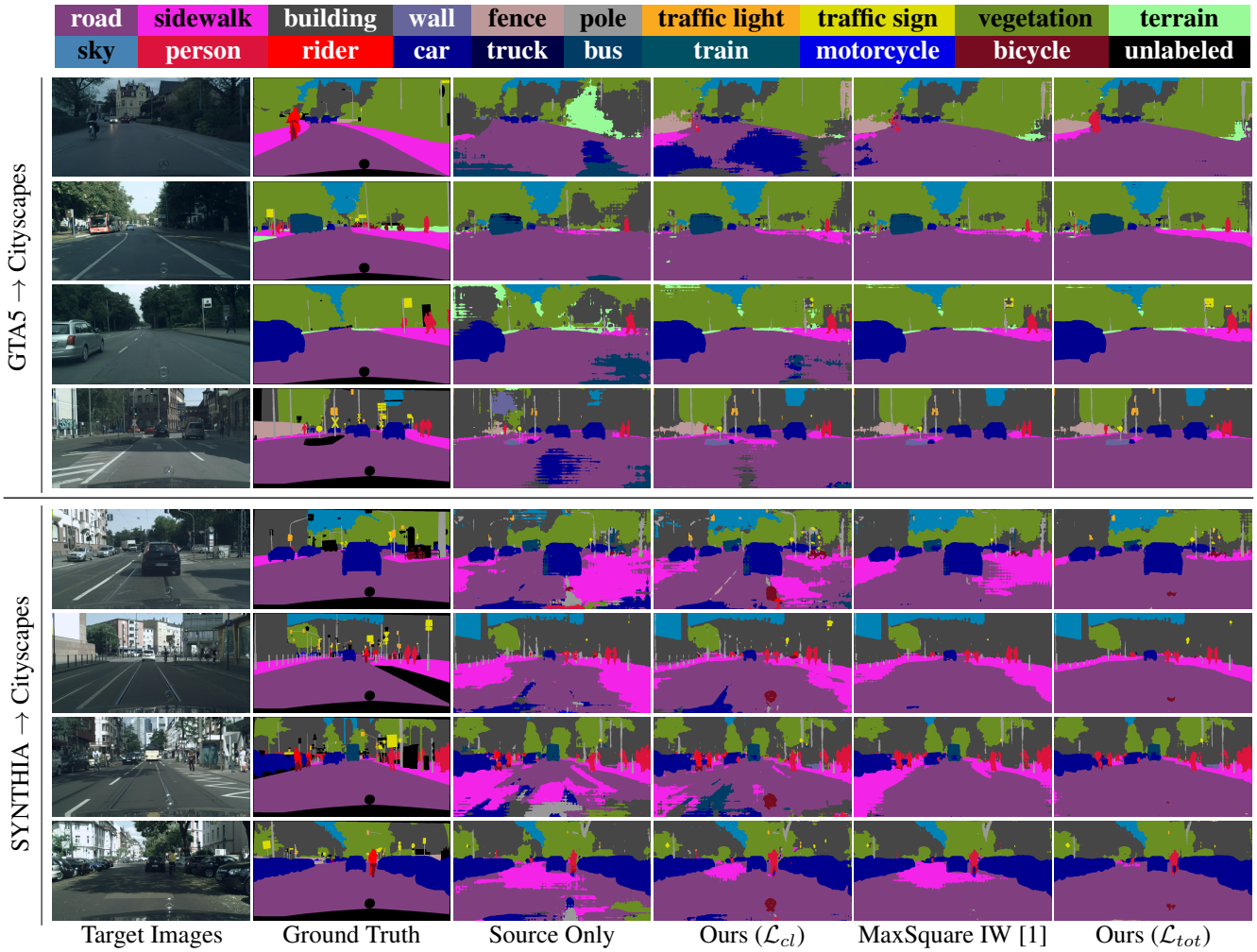

Figure 1: Semantic segmentation of some sample scenes extracted from the Cityscapes validation target dataset when adaptation is performed from the GTA5 (top) and SYNTHIA (bottom) source datasets and the DeepLab-V2 with ResNet-101 backbone is employed as segmentation network (*best viewed in colors*).

2.3. Sparsity Loss

To further investigate the contribution of the sparsity loss, $\mathcal{L}_{sp}$, we inspect how values of feature channels (i.e. single units in feature vectors) end up being distributed for different adaptation settings. In Figure 5 we plot the histogram distribution of all normalized feature activations with bin size set to 0.05 in linear scale and in log scale. Here, we can observe that adding $\mathcal{L}_{sp}$ leads to a greater number of occurrences of activations within 0 and 0.1 and within 0.95 and 1 than in the case without sparsity constraint. In the middle, instead, the opposite is true. We refer the reader to Eq. 7 of the main paper to certify that this is indeed what the sparsity loss was aiming to achieve. Namely, the sparsity constraint reduces class-wise the number of active feature channels pushing them either toward 0 (inactive features) or towards 1 (active features).

Ultimately, for more immediate visualization, we plot in the third image of Figure 5 the difference of the sparsity distributions with and without $\mathcal{L}_{sp}$. We can more easily verify that extremely low (in closest range to 0) and extremely high (in closest range to 1) bins have positive values while middle-range bins have negative values.

References

[1] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, pages 2090–2099, 2019.

[2] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008.

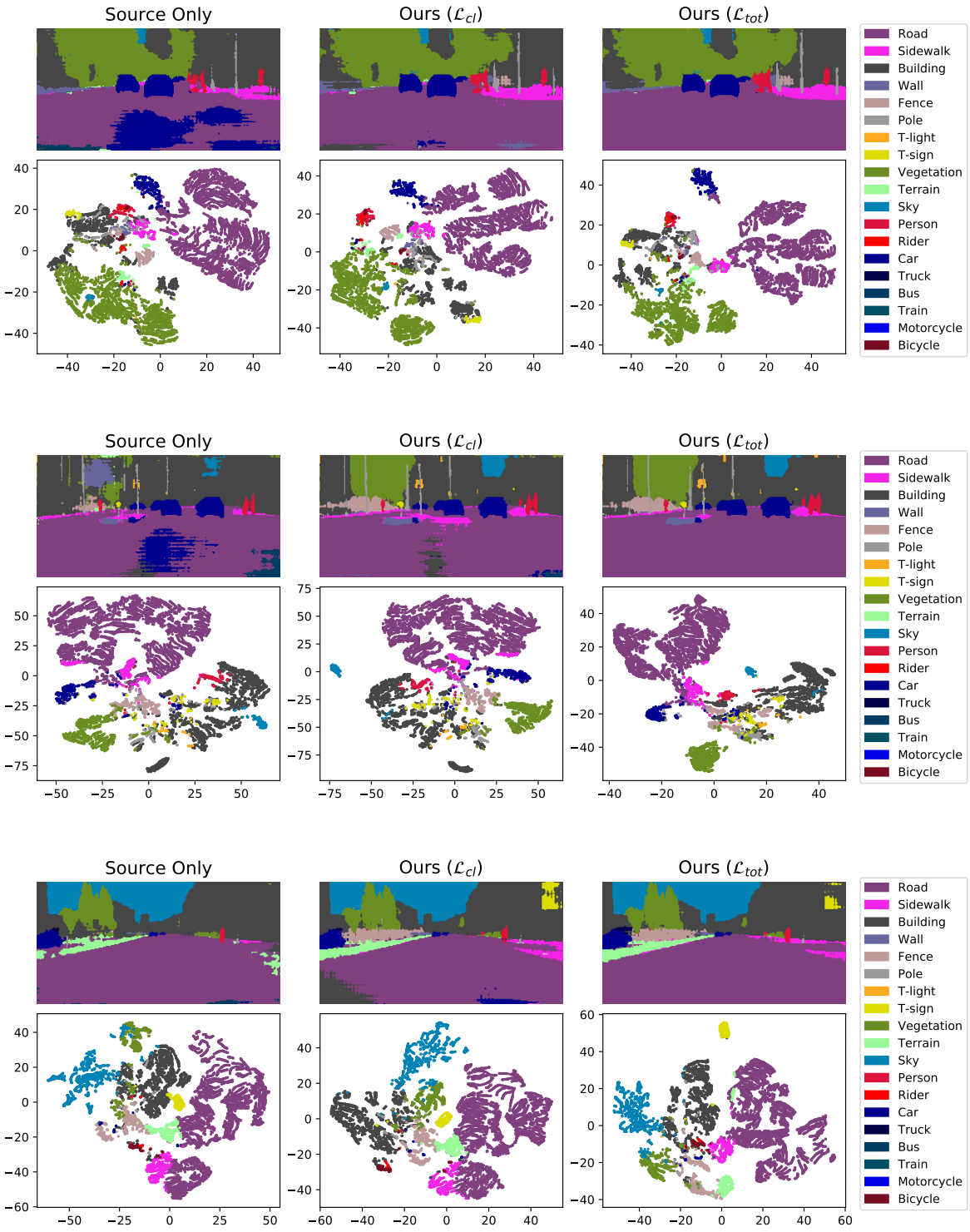

Figure 2: T-SNE computed over features of single images from the Cityscapes validation set when adapting from GTA5 (*best viewed in colors*).

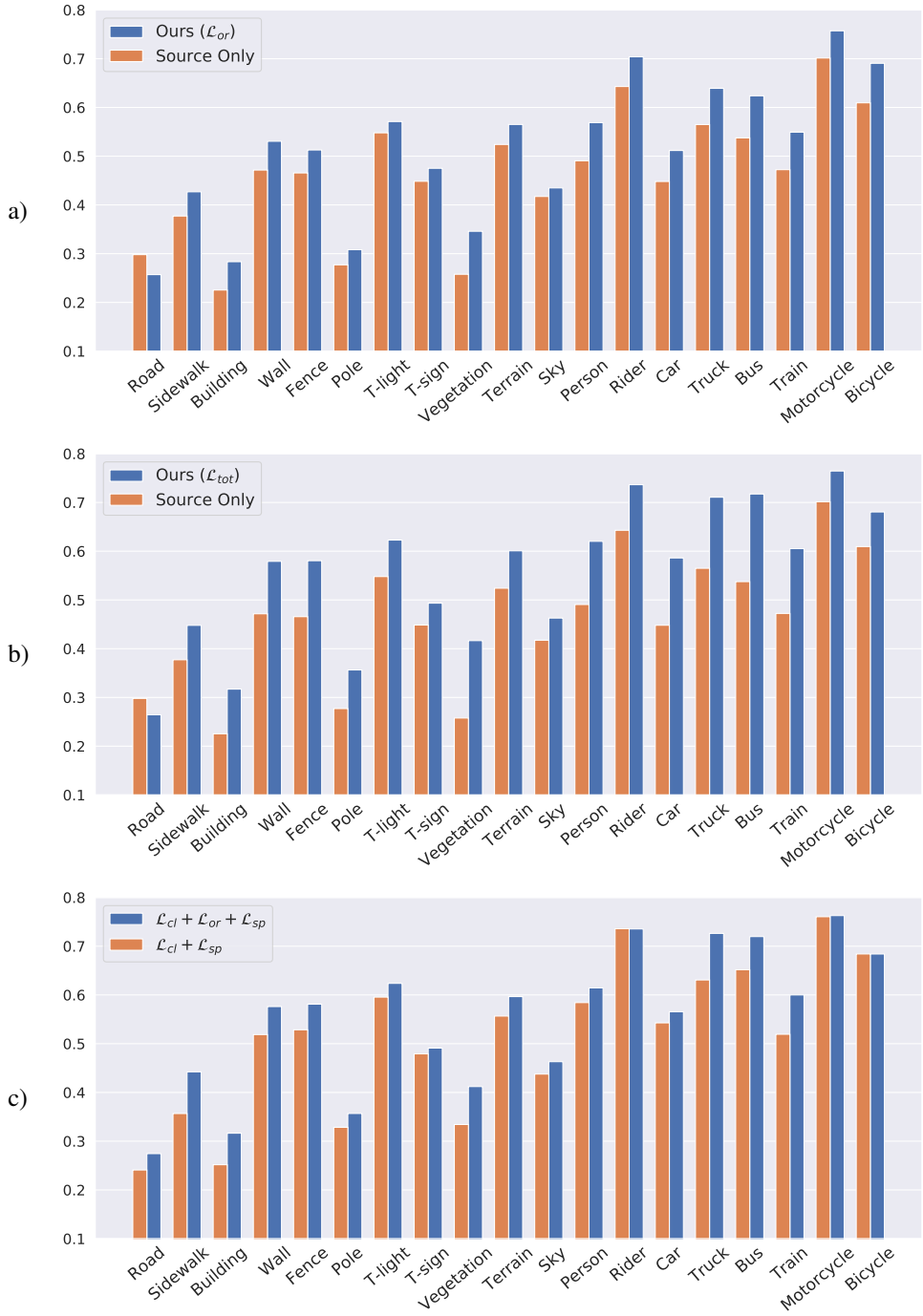

Figure 3: Similarity scores computed over all the images on the Cityscapes validation set when adapting from GTA5 to analyze the effect of the orthogonality constraint (*best viewed in colors*).

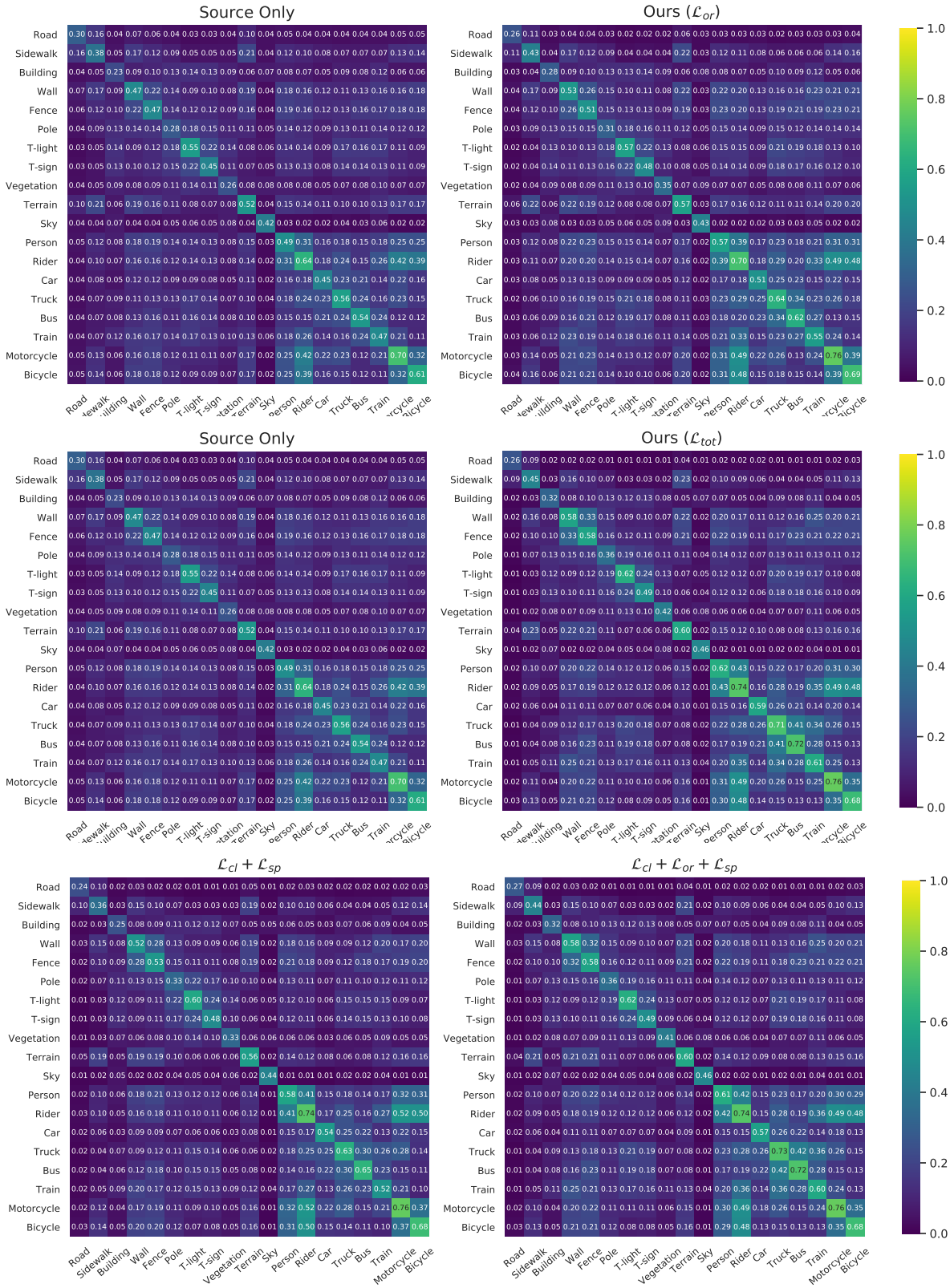

Figure 4: Class-wise similarity scores computed over images on the Cityscapes validation set when adapting from GTA5 (*best viewed in colors*).

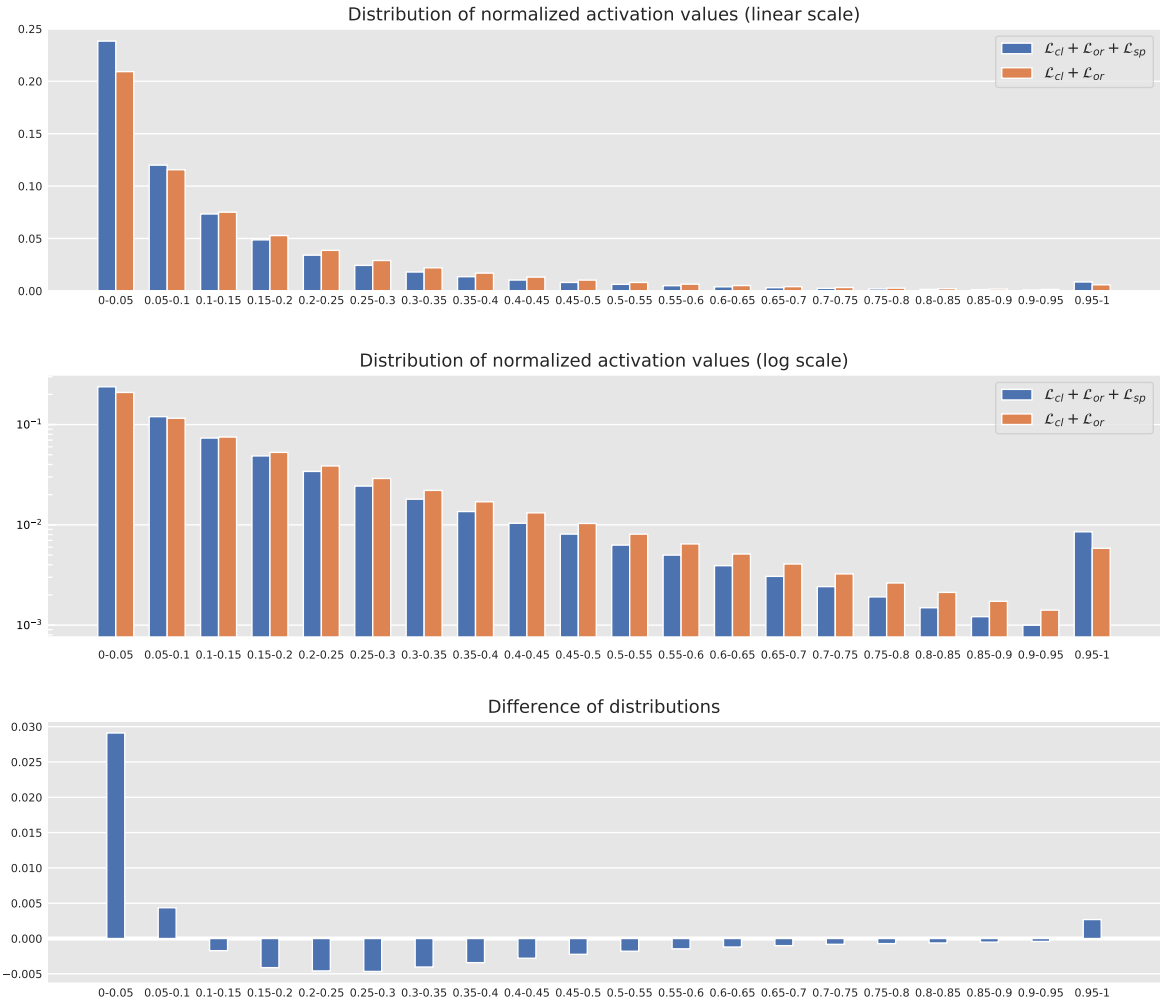

Figure 5: Analysis of the distribution of feature activations computed over all the images on the Cityscapes validation set when adapting from GTA5 (*best viewed in colors*).