Supplementary for A Closer Look at Self-training for Zero-Label Semantic Segmentation

Giuseppe Pastore[1], Fabio Cermelli[1,2], Yongqin Xian[3],
Massimiliano Mancini[4], Zeynep Akata[3,4,5], Barbara Caputo[1,2]
[1]Politecnico di Torino, [2]Italian Institute of Technology, [3]MPI for Informatics,
[4]University of Tübingen, [5]MPI for Intelligent Systems

1. Qualitative examples of pseudo-labeling.

In this section, we show additional qualitative results of pseudo-labeling. For each *original* image, *GT* is the actual ground truth y, while *labeled pixels* represents the annotation for seen classes y^s that the model sees before the pseudo-labeling.

PascalVOC12. Figure 1 and 2 and enrich the Figure 3 of the main paper with additional visual examples of pseudo-labels generated by our model for PascalVOC12 [3], both in the case where the background is ignored (Fig. 1) and where it is included (Fig. 2) during the training. In the first scenario, we can observe that while SPNet [4] correctly predicts the presence of the unseen classes *train*, *sofa* and *sheep*, a lot of pixels are assigned to the wrong class. As an example, *sheep* is wrongly predicted in the first row and *potted plant* is wrongly assigned to some pixels in the last one. STRICT instead reduces the noise by ignoring the most uncertain pixels (see the 5-th column) and expanding plus refining the correctly predicted regions with more iterations (*e.g.* last column).

We can see a similar behaviour in the second scenario. For instance, in the third row of Figure 2, SPNet wrongly assigned some of the *sofa* pixels to the *tv/monitor* class; STRICT reduces such noise and refines the area predicted as *sofa* even if still with some background pixels wrongly assigned to the *potted plant* class. The same effect is more visible in the second row, while in the first row is worth noticing how *potted plant* represents a challenging target also for STRICT, that is able to coarsely predict it only after some additional fine-tuning iterations (*e.g.* 9 in this case).

COCO-stuff. In Figure 3, we show qualitative results of the generated pseudo-labels on COCO-Stuff [2] dataset. As shown in Figure 3, while the pseudo-labeling with SPNet successfully identifies the classes present in the images, it does with a lot of noise. For instance, in the second row, some regions of the *playingfield* unseen class are assigned to the *road* and to the *grass* ones; in contrast, the consistency regularizer introduced by our STRICT approach reduces the quantity of misclassified pixels mostly by ignoring them. With more fine-tuning iterations, STRICT is also able to correctly classify these pixels.

2. More qualitative segmentation results.

In this Section, we extend the qualitative segmentation results reported in Figure 5 of the main paper.

PascalVOC12. Figure 4 shows the qualitative comparison of STRICT with ZS5 baseline on PascalVOC12 without the background as seen class. In these additional examples, we can observe again the superiority of our model over ZS5 in predicting both seen and unseen classes: therefore, ZS5 misclassifies the major part of *dog* and *chair* (seen classes) into the *sofa* (unseen class), while our model is able to correctly distinguish and predict them. This is also confirmed by the first row on the right, where ZS5Net misclassifies all the *table*'s pixels as a mix of *sofa* and *tv/monitor* unseen classes while our model correctly predicts them.

COCO-stuff. In Figure 5, we compare the qualitative results of SPNet (with self-training) with our STRICT on COCO-stuff. Both the methods correctly identify the categories shown in each image with few cases in which they predict absent classes (*e.g.* blue region in the last row or the green region on the first row on the left). Our model significantly limits the noisiness of the prediction: *e.g.* on the first image, the misclassified region produced by SPNet is completely removed by our model; the same happens for the last row on the left and for the third row on the right. However, COCO-stuff is particularly challenging, being a dense annotated dataset with a huge number of classes among objects and stuffs. Therefore, STRICT outperforms the other methods but, even if less than SPNet, still suffers from noisy predictions, especially on the identification of unseen stuff classes, as their extension includes various pixels with varying appearance (*e.g.* the second row on the left or the first row on the right).

STRICT strategy demonstrates to improve the performance on the GZLSS with astoninishing results; however, it

requires some form of regularization to address the dependency on the number of co-occurring pixels and a stronger embedding of the semantic knowledge for the dense annotated context with a large number of classes.

References

[1] Maxime Bucher, Tuan-Hung Vu, Mathieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *NeurIPS*, 2019. 4

[2] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. *CoRR*, abs/1612.03716, 2016. 1, 4, 5

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 1, 3, 4

[4] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata. Semantic projection network for zero- and few-label semantic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8248–8257, 2019. 1, 5

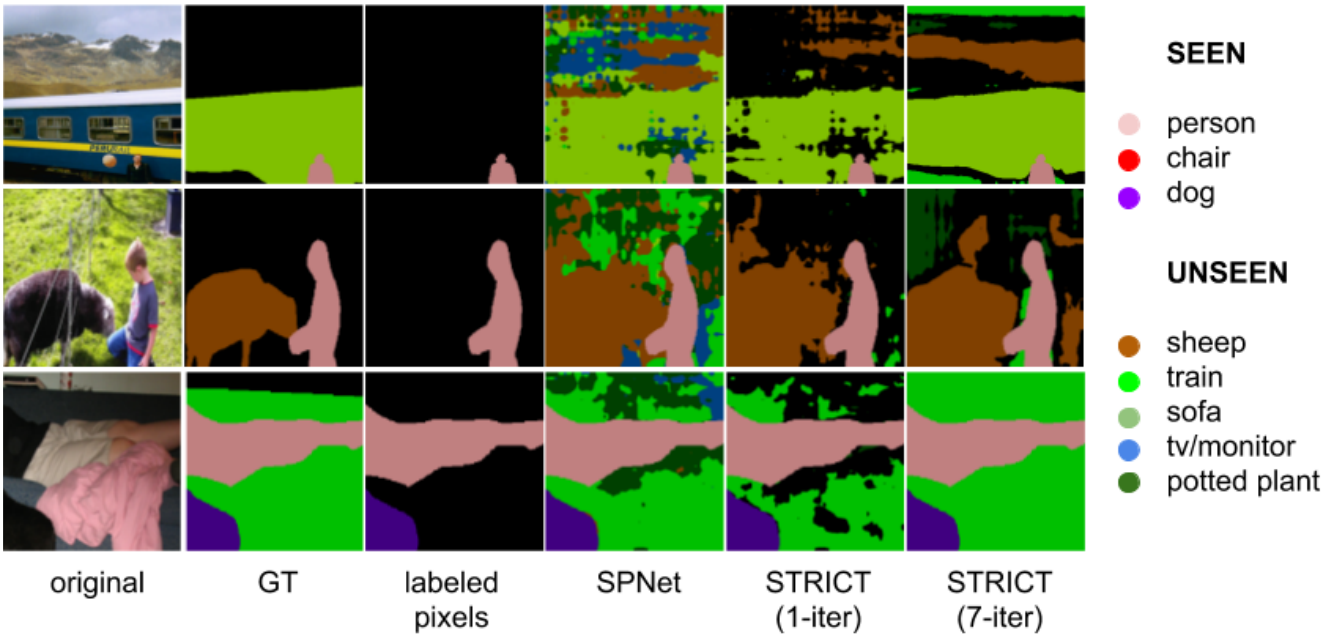

Figure 1: Supplementary qualitative pseudo-labeling results of STRICT on PascalVOC12 [3] without background as seen class. Train GT refers to labels for the unseen classes.

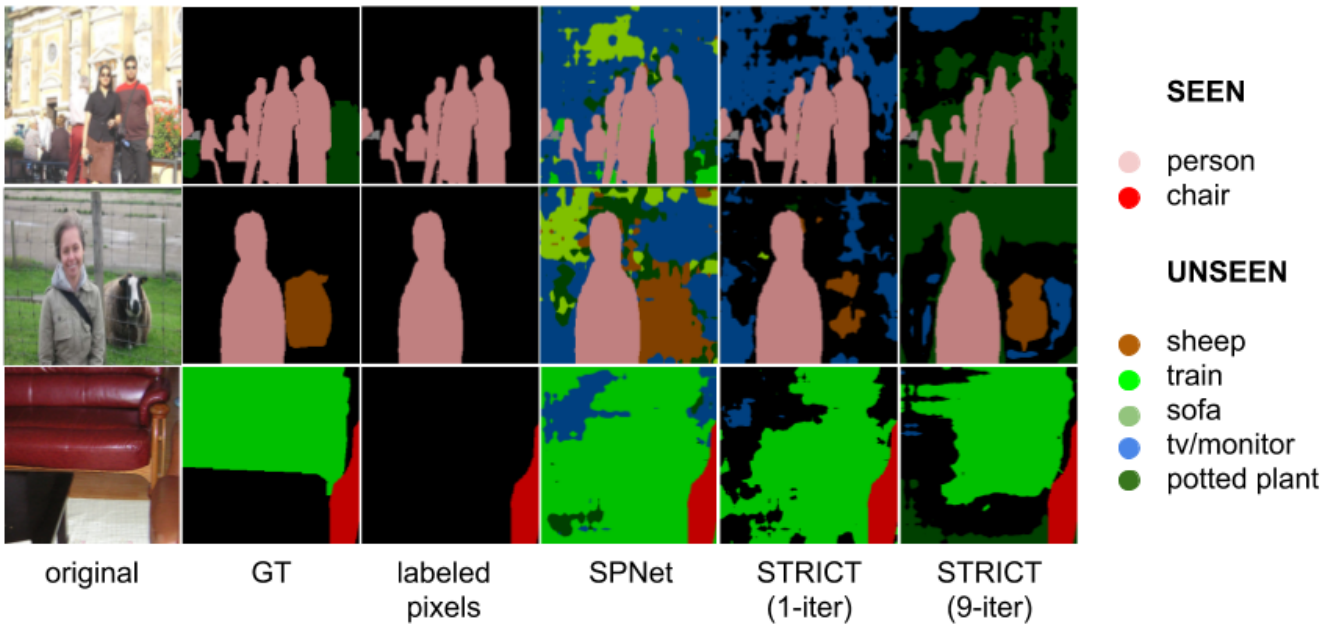

Figure 2: Supplementary qualitative pseudo-labeling results of STRICT on PascalVOC12 [3] with background as seen class. Train GT refers to labels for the unseen classes.

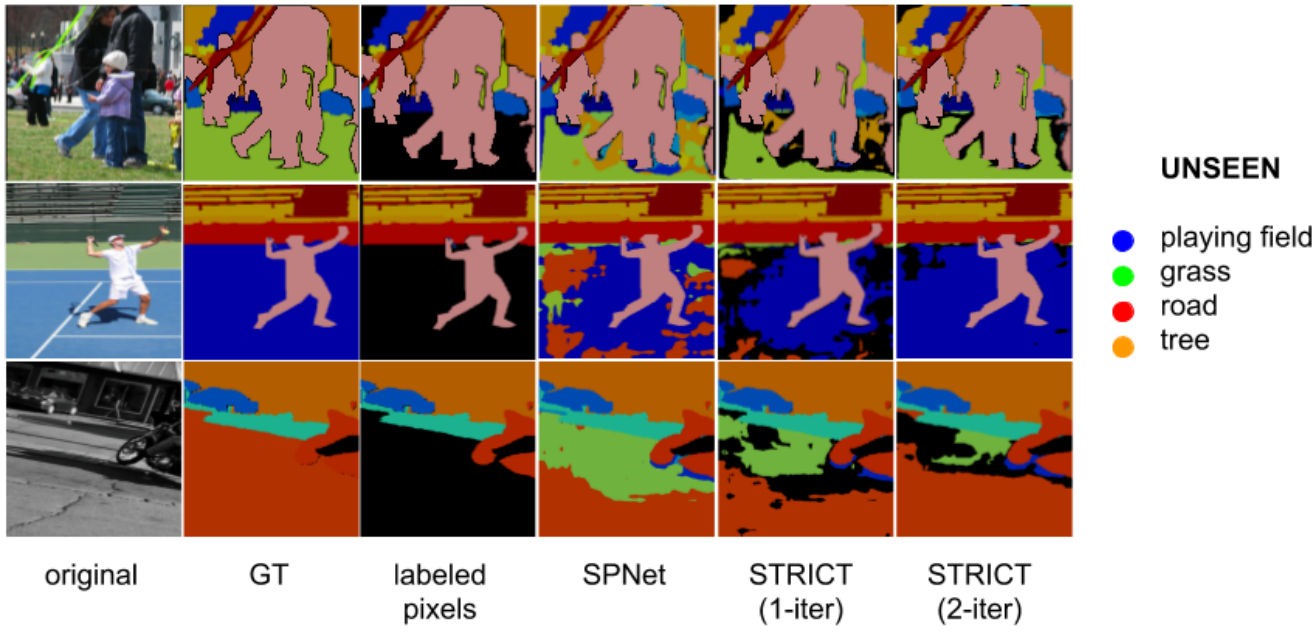

Figure 3: Supplementary qualitative pseudo-labeling results of STRICT on COCO-stuff [2]. Train GT refers to labels for the unseen classes.
Unseen classes: *playingfield*(blue), *road*, *suitcase*(light blue), *grass*(green), *tree*(ocra yellow),*wall-concrete*(light green).

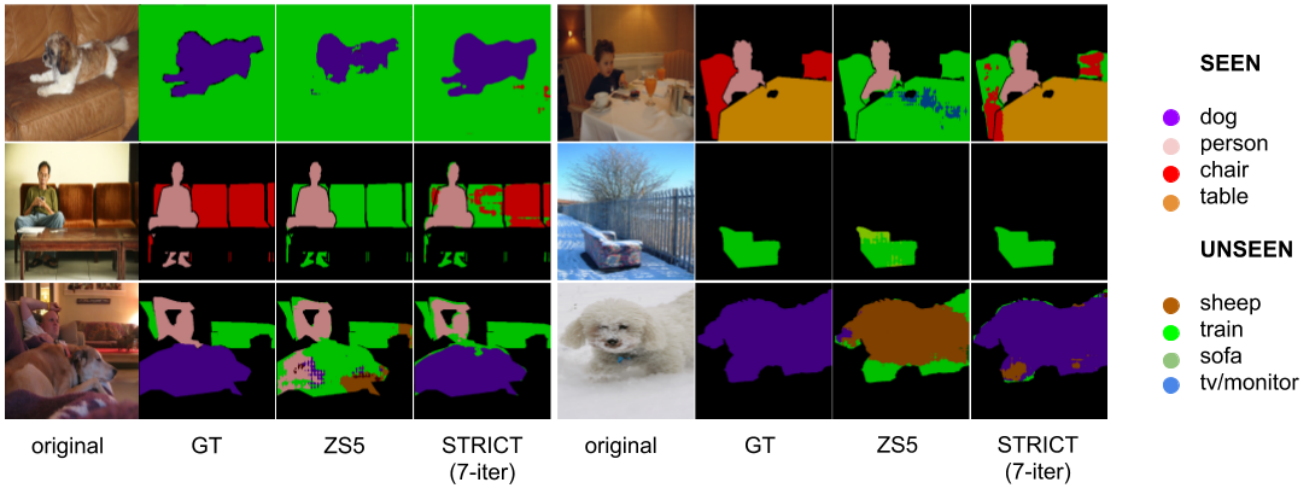

Figure 4: Qualitative comparison of STRICT results with ZS5 [1] on PascalVOC12 [3] without background as seen class.

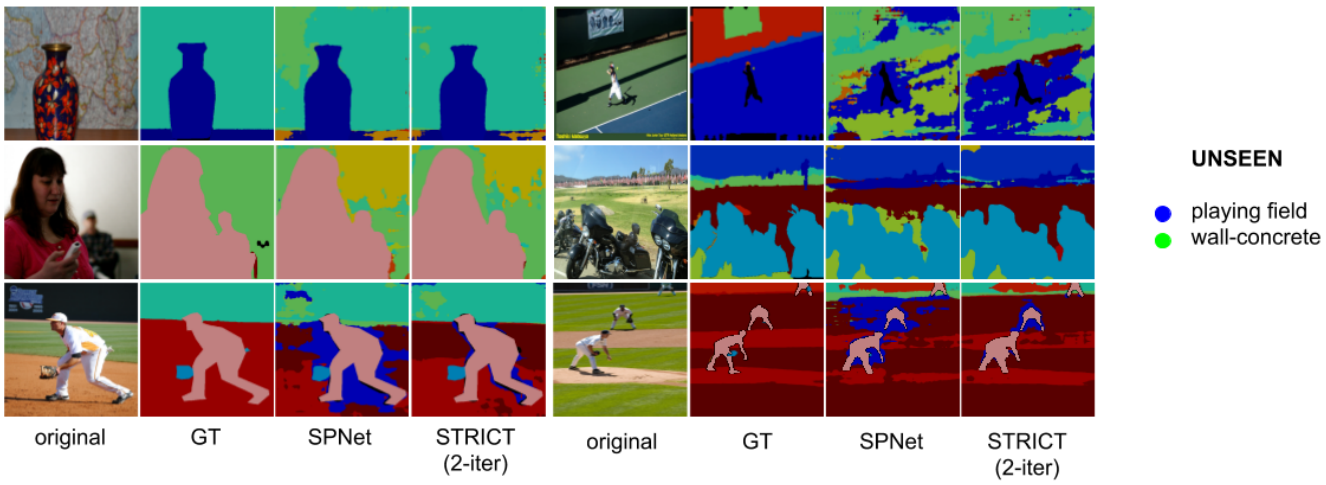

Figure 5: Qualitative comparison of STRICT results with calibrated SPNet [4] on COCO-stuff [2].