Supplementary Material for "Deep Features based on Contrastive Fusion of Transformer and CNN for Semantic Segmentation"

1 Evaluation Metrics

The Pixelwise Accuracy (PA) (1), Dice Coefficient (DC) (2) and mean Intersection over Union (mIOU) score (4) are used to evaluate the predicted mask with the ground truth. Particularly in the dataset with class imbalance, an image often has a tiny fraction of pixels for a few classes and all of the remaining images are other majority of classes. Since the accuracy score considers actual negative results, it will always produce an erroneous high score. On the other hand, due to their tendency to penalize false positives—a common occurrence in this dataset with extreme class imbalance, DC and mIoU are used here for semantic segmentation. The mIoU penalizes under and over-segmentation more than DC, which is the difference between the two metrics. Therefore, the value of mIOU is lower than the DC.

The expression for PA is as follows,

$$\text{PA} = \frac{\sum_{i=1}^{N} \mathbf{1}(y_i = \hat{y}_i)}{N}. \tag{1}$$

The expression for DC given as follows,

$$\text{DC} = \frac{2\sum_{i=1}^{N}(y_i \cdot \hat{y}_i)}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i}. \tag{2}$$

The expression of mIoU is given below,

$$\text{IoU}_c = \frac{\sum_{i=1}^{N}(y_i = c) \cdot (\hat{y}_i = c)}{\sum_{i=1}^{N}((y_i = c) \cup (\hat{y}_i = c))}, \tag{3}$$

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \text{IoU}_c. \tag{4}$$

In these equations, y_i represents the ground truth label for pixel i, $\hat{y}_i$ represents the predicted label or score for pixel i, N is the total number of pixels, C is the total number of classes, $(y_i = c)$ and $(\hat{y}_i = c)$ are indicator functions that evaluate to 1 if y_i or $\hat{y}_i$ equals c, and 0 otherwise.

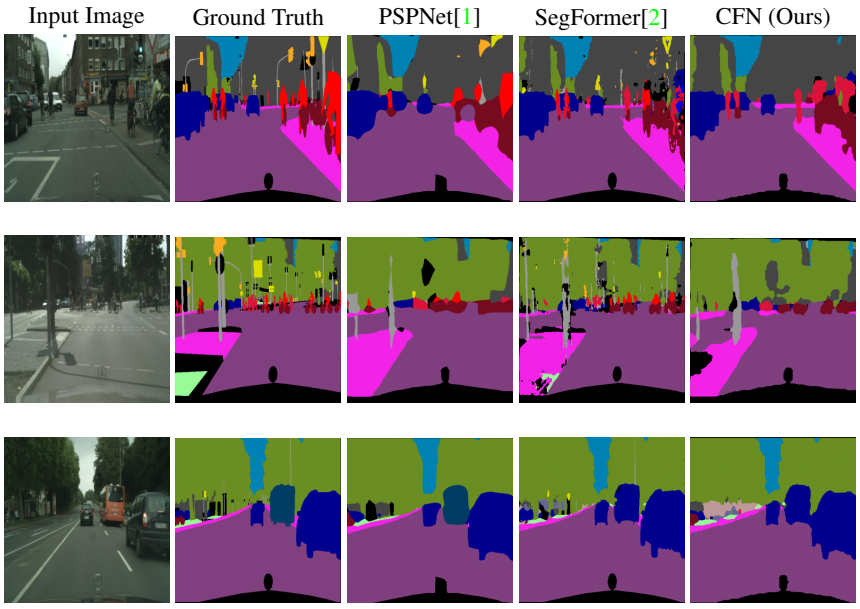

Figure 1: Qualitative comparison on validation images. PSPNet and SegFormer struggle with fine details (e.g., 'man' and 'bicycle' classes for Input Image 1). Our CFN outperforms them by benefiting from a mixture of losses, achieving superior performance. Zoom in for best view.

2 Experiments

2.1 Comparison with SOTA models

In contrast, transformer-based models, represented by SegFormer [2], exhibited more modest performance, with an mIOU of 28.9% and a DC of 50.2%. This suggests a potential need for additional training data or architectural refinements to enhance their effectiveness in semantic segmentation tasks. However, our CFN surpassed all these models. The student-teacher pair of CFN-HRNet demonstrates outstanding performance with an mIOU of 62.9% and a DC of 89.2%. This positions CFN as the leading choice among the evaluated models for semantic segmentation tasks.

Fig. 1 illustrates sample semantic segmentation maps generated by our proposed approach and SOTA methods on the finely annotated validation dataset of the primary Cityscape dataset [3]. The results reveal that two models, PSPNet [1] and SegFormer [2], struggle to capture fine details for the 'man' and 'bicycle' class in the first sample image. Additionally, these two models fail to detect the 'unlabeled' class in the second sample image, whereas our model successfully identifies it. This success can be attributed to the mixture of losses utilized in our model. The contrastive loss captures common objects localized by the student-teacher network and updates the pa-

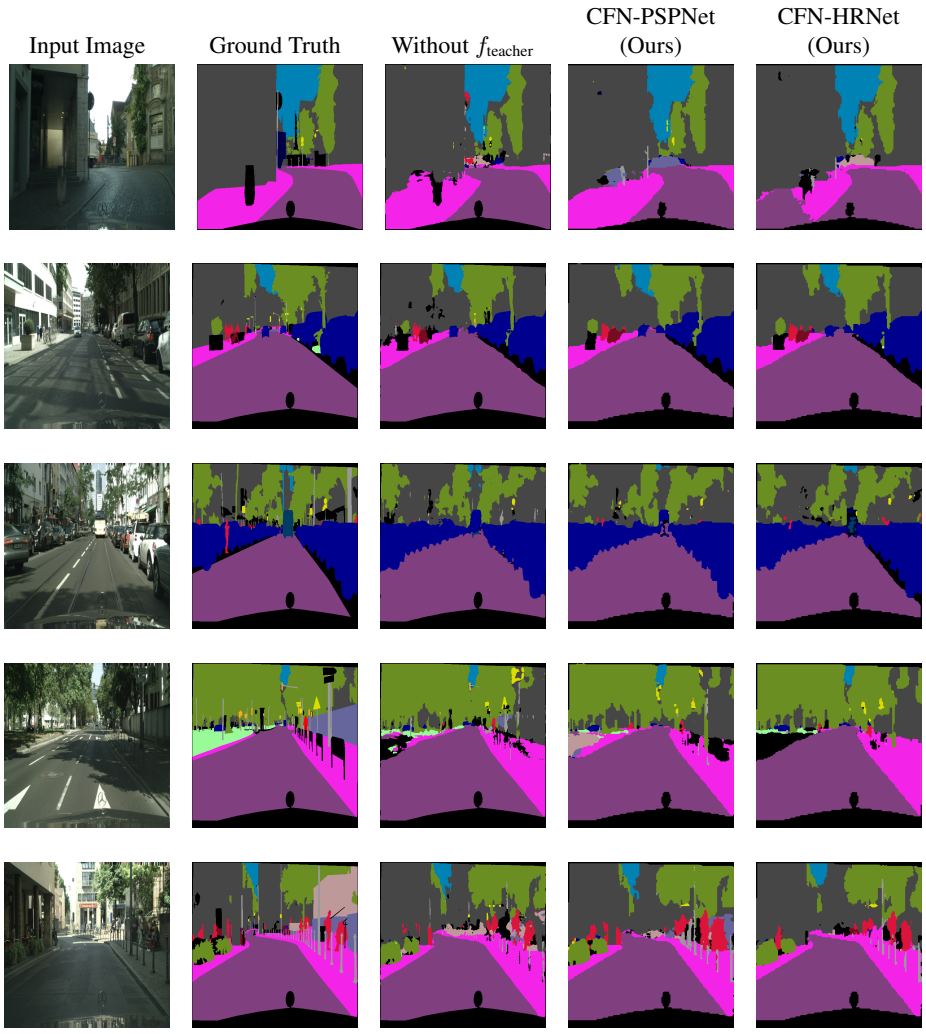

Figure 2: We conducted a qualitative comparison on validation images for two versions of the CFN model: CFN-PSPNet and CFN-HRNet. The predicted masks obtained from the model without f_{teacher} correctly segment larger classes such as roads, footpaths, and buildings. However, noticeable category noise is also present in the output. For a clearer view, please zoom in for the best visualization.

rameters of the student network. Furthermore, the dice and CCE losses in the overall loss equation compare the final segmentation maps with ground truths and refine the model.

2.2 Visualization

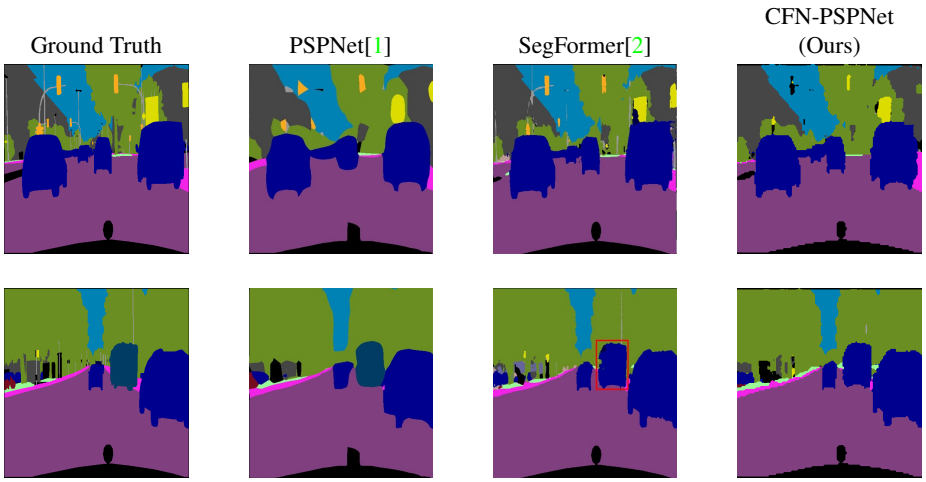

Figure 3: Comparison on validation images between PSPNet [1], SegFormer [2] and newly proposed CFN-PSPNet. Zoom in for best view.

We test our student model on the validation dataset of the main Cityscape dataset [3] after being trained on only the finely annotated training dataset. We gain the mIOU score of 62.20% and 62.86% for the teacher network of PSPNet and HRNet respectively. [1, 4] with its ability to capture fine-graded details through a pyramid and high-resolution representation learning, respectively, attains a comparable mIOU and DC. Among the three proposed mechanisms, CFN-PSPNet shows the best performance quantitatively. The qualitative comparisons can be done through figures as shown in Fig. 2.

Fig. 3 present an analysis of various models alongside the newly proposed CFN segmented results, highlighting their respective merits and shortcomings. CFN-HRNet and SegFormer demonstrate a propensity for generating segmentation maps devoid of pixel noise. However, the SegFormer being a transformer-based model, the mIOU is only 28.9%, highlighting the need for more data samples. The CNN-based PSPNet [1], trained on the finely annotated Cityscape dataset, exhibits a limitation in capturing fine details, as illustrated in Fig. 3. Despite being trained on a dataset with meticulous annotations, the model primarily captures coarse details and the general location of objects. Conversely, the performance of SegFormer [2] in this aspect is noteworthy. However, in Fig. 3, for the second image, both methods tend to misclassify the 'truck' as a 'car' class.

Taking into account the respective strengths of CNN-based architectures and transformer models, the CFN is developed. This approach reduces the need for additional datasets and enhances generalization capabilities. By employing contrastive loss, we effectively update the Transformer encoder network, improving its output. To facilitate reproducibility and further research, we have made our code publicly available. Researchers and practitioners can access our implementation to verify the results and explore the potential of CFN in various applications.

Transformer-based models, represented by SegFormer [2], exhibited more modest performance, with an mIOU of 28.9% and a DC of 50.2%. This suggests a potential need for additional training data or architectural refinements to enhance their effectiveness in semantic segmentation tasks. However, our CFN surpassed all these models. The student-teacher pair of CFN-HRNet demonstrates outstanding performance with an mIOU of 62.9% and a DC of 89.2%. This positions CFN as the leading choice among the evaluated models for semantic segmentation tasks.

Fig. 1 illustrates sample semantic segmentation maps generated by our proposed approach and SOTA methods on the finely annotated validation dataset of the primary Cityscape dataset [3]. The results reveal that two models, PSPNet [1] and SegFormer [2], struggle to capture fine details for the 'man' and 'bicycle' class in the first sample image. Additionally, these two models fail to detect the 'unlabeled' class in the second sample image, whereas our model successfully identifies it. This success can be attributed to the mixture of losses utilized in our model. The contrastive loss captures common objects localized by the student-teacher network and updates the parameters of the student network. Furthermore, the dice and CCE losses in the overall loss equation compare the final segmentation maps with ground truths and refine the model.

References

[1] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *arXiv*, 2016.

[2] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. Álvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *arXiv*, 2021.

[3] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CVPR*, 2016.

[4] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions," *arXiv*, 2019.