Supplementary for Perturbed and Strict Mean Teachers for Semi-supervised Semantic Segmentation

Yuyuan Liu[1] Yu Tian[1] Yuanhong Chen[1] Fengbei Liu[1]
Vasileios Belagiannis[2] Gustavo Carneiro[1]
[1] Australian Institute for Machine Learning, University of Adelaide
[2] Universität Ulm, Germany

1. Teacher-based Virtual Adversarial Training (T-VAT)

Virtual adversarial training (VAT) [1] aims to produce a gradient-oriented perturbation that maximises the divergence between the prediction distributions of the original and perturbed inputs. After maximising the divergence between the predictions of the original and its perturbed version, the classification boundary will move towards the classification boundary, producing a more challenging classification sample that can improve training generalization. During training, the teachers ensemble predictions yield more accurate performance than the student predictions, as shown in Fig. 3-(b). This demonstrates that the teachers ensemble results follow more closely the true classification boundary than the student predictions. Therefore, the adversarial samples provided by the teachers (i.e., T-VAT) can be more "accurate" to alter the student model's gradient than the adversarial samples provided by the student.

1.1. Implementation Details

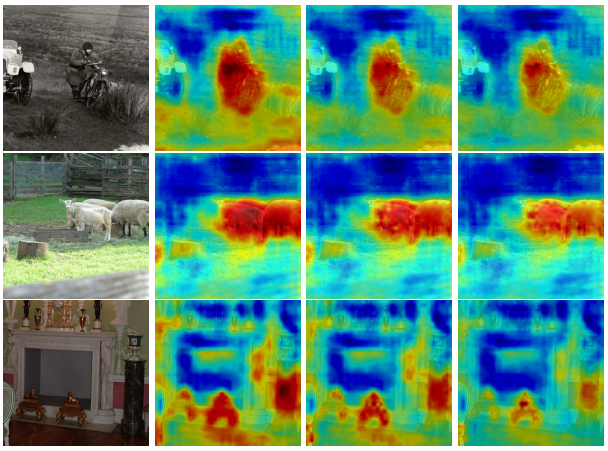
Our T-VAT results are produced based on the embedding produced by the student encoder, which is $\mathbf{z}^s = h_{\theta^s_h}(\mathbf{x})$. We first sample the noise $\mathbf{r}$ from a Gaussian distribution, which has the same dimensionality of $\mathbf{z}^s$. Then, we compute the gradient of the KL divergence, as in

$$\mathbf{g} = \nabla d\Big(0.5 \times g_{\theta^{t1}_g}(\mathbf{z}^s) + 0.5 \times g_{\theta^{t2}_g}(\mathbf{z}^s), \\ 0.5 \times g_{\theta^{t1}_g}(\mathbf{z}^s + \mathbf{r}) + 0.5 \times g_{\theta^{t2}_g}(\mathbf{z}^s + \mathbf{r})\Big). \quad (1)$$

Lastly, we obtain the teacher-gradient adversarial perturbation $\mathbf{r}_{adv}$, which is normalised as follows

$$\mathbf{r}_{adv} = \epsilon \frac{\mathbf{g}}{||\mathbf{g}||}, \quad (2)$$

where we use $\epsilon = 2$ for all experiments. We inject the perturbation $\mathbf{r}_{adv}$ into the original embedding with $\mathbf{z}^s = \mathbf{z}^s + \mathbf{r}_{adv}$.



(a) Input (b) No perturbation (c) VAT (d) T-VAT

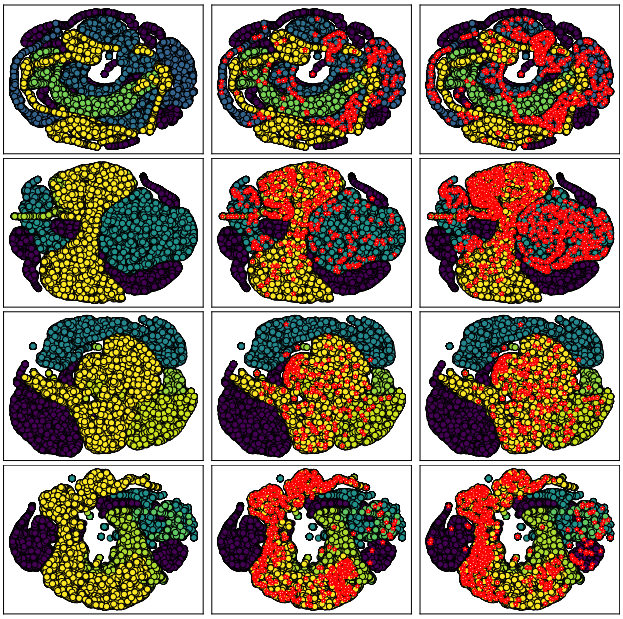
Figure 1. The regional contribution to the feature of interest that visualised by Grad-cam [3]. The red region corresponds to high contribution. **Less red regions indicate better impact for the perturbation.**

1.2. T-SNE Visualisation

In Fig. 2, we visualise the perturbation impact using T-SNE. We randomly sample $25,000$ correctly predicted pixels based on 5 different classes (each class represented by a different colour) in (a). Then we apply the VAT noise using the student model in (b) and our T-VAT noise in (c) to observe the perturbations of the original results (from correct to incorrect). T-VAT demonstrates more challenging predictions compared with the VAT (with more incorrect predictions), which shows better perturbation efficiency.

2. Ramp-up Function

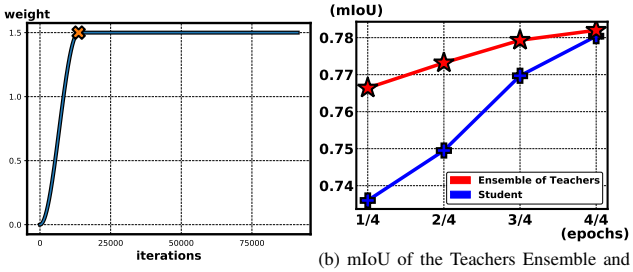
Due to the prediction instability of the unlabelled data in the early training stages, we introduce a ramp-up weight to be applied to the unsupervised loss [2]. In Fig. 3-(a), we illustrate our Gaussian ramp-up function applied for the weight β in Eq. (1) of the main paper. The final weight is



(a) No perturbation (b) VAT (c) T-VAT

Figure 2. **T-SNE visualization.** We visualise the incorrect predictions after the feature perturbation, shown with red stars. More incorrect predictions demonstrates more effective perturbations.

1.5 and the ramp-up length of 12 is used for all the experiments.



(a) Ramp-up for β in Eq. (1) (b) mIoU of the Teachers Ensemble and Student

Figure 3. (a) **Ramp-up function for β in Eq. (1).** This diagram shows the ramp-up curve that we use to weight the unsupervised loss as a function of the iterations. (b) **mIoU of the Teachers Ensemble and Student** on the validation set during the training process.

3. Single teacher Inference

After the training, two teachers will converge into same local minimal and the single teacher's performance will be similar with the ensemble result.

Table 1. DeeplabV3+ based ResNet50 mIoU results on Pascal VOC2012 with the ensemble model inference (1st row – paper results), and single teacher model inference (2nd row).

ratio	1/16	1/8	1/4	1/2
Ensemble (paper)	72.83	75.70	76.43	77.88
Single model	72.79	75.59	76.26	77.71

References

[1] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 1

[2] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 1

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1