Supplementary Material: Knowledge Distillation via the Target-aware Transformer

Sihao Lin[1,3†], Hongwei Xie[2†], Bing Wang[2], Kaicheng Yu[2],
Xiaojun Chang[3§], Xiaodan Liang[4], Gang Wang[2]
[1]RMIT University [2]Alibaba Group [3]ReLER, AAII, UTS [4]Sun Yat-sen University
{linsihao6, hongwei.xie.90, Kaicheng.yu.yt, xdliang328}@gmail.com
{fengquan.wb, wg134231}@alibaba-inc.com, xiaojun.chang@uts.edu.au

1. Asset Usage

This work is built upon some public dataset and code assets. We appreciate their efforts. The benchmark dataset has been introduced in main paper. Here we list the URL, version, and license of the code assets that we used:

Table 1. Usage of Code assets.

Exp.	URL	Ver.	Licence
ImageNet	https://github.com/yoshitomo-matsubara/torchdistill	7b883ec	MIT
Cifar100	https://github.com/HobbitLong/RepDistiller	9b56e97	BSD 2-Clause
Pascal VOC	https://github.com/jfzhang95/pytorch-deeplab-xception	9135e10	MIT
	https://github.com/clovaai/overhaul-distillation	76344a8	MIT
COCOStuff10k	https://github.com/kazuto1011/deeplab-pytorch	4219467	MIT
	https://github.com/dvlab-research/ReviewKD	cede6ea	N/A

2. Additional Experiments

2.1. Comparison on COCOStuff10k

For the experiments of semantic segmentation, we have compared our method to a variety of stat-of-the-art methods in the Section 4 of the main paper. In terms of COCOStuff10k, since some methods do not support this dataset, we re-implement them and the result is presented on Table 2. We found that our method is competitive and it outperforms the comparison methods.

Table 2. Comparison (mIoU%) on COCOStuff10k.

	ICKD [2]	Overhaul [1]	Ours
ResNet18	27.22	27.86	28.75
MobileNetV2	26.64	26.96	28.05

§Corresponding Author.
†Equal contribution.

Table 3. Coefficients α and ϵ on different backbones on Cifar-100.

Teacher	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet110	ResNet32×4	VGG13
Student	WRN-16-2	WRN-40-1	ResNet20	ResNet20	ResNet32	ResNet8×4	VGG8
α	0.8	0.7	0.8	1	1	6	0.1
ϵ	4	3.6	0.4	0.75	1	39	8

Table 4. Adding $\mathcal{L}_{\mathrm{KL}}$ on Cifar100.

Teacher	WRN-40-2	ResNet110	ResNet32×4	VGG13
Student	WRN-16-2	ResNet20	ResNet8×4	VGG8
KD	74.92	70.67	73.33	72.98
FitNet+KD	75.12	70.67	74.66	73.22
AT+KD	75.32	70.97	74.53	73.48
SP+KD	74.98	71.02	74.02	73.49
CC+KD	75.09	70.88	74.21	73.04
RKD+KD	74.89	70.77	73.79	72.97
PKT+KD	75.33	70.72	74.23	73.25
NST+KD	74.67	71.01	74.28	73.33
CRD+KD	75.64	71.56	75.46	74.29
ICKD+KD	75.57	71.91	75.48	73.88
Ours+KD	**76.08**	**72.16**	**75.54**	**74.35**

2.2. Hyperparameters on Cifar-100

We used Bayesian optimization to obtain the weight factors α and ϵ in Eq. 9. Here we show the searching result on different backbones (See Table 3). We found that in most cases (4 out of 6), ϵ is greater than α, which indicates that our proposed objective is more important than the standard Cross-entropy during distillation. For instance, in the distillation VGG13→VGG8, ϵ is 8 and α is only 0.1. We also found that for the similar architectures, the searching result is similar, *e.g.*, when WRN-40-2 and ResNet110 are selected as teacher.

2.3. Adding KD loss on Cifar-100

We report the result of our method in Table 4 with $\mathcal{L}_{\mathrm{KL}}$ loss to compare with the baselines under the same settings. Our method with KD loss surpasses all the baselines again.

2.4. Feature Visualization

We further visualize the feature map and the associated TaT map to intuitively understand the functionality behind the proposed Target-aware Transformer. As exhibited in Figure 1, we visualize the feature maps of student before and after distillation, which are compared to the feature map of teacher. The teacher backbone is ResNet34 and student backbone is ResNet18. The input images are randomly selected from ImageNet validation set. While the 4-th block (*i.e.* distillation layer) of ResNet34 and ResNet18 has 512 channels, we visualize 64 channels for better visualization.

Obviously, the reconfigured student feature (3rd column) has a more similar pattern with teacher feature (4th column), which demonstrates that TaT can effectively adapt the student to mimic the teacher. In terms of the TaT map, which controls the intensity of semantic aggregation, it is close to the identity matrix. Recall that we apply the linear function $\phi(\cdot)$ on student feature f^s. And the TaT map will be further applied on $\phi(f^s)$ to reconfigure the student feature, which is lately asked to minimize the L_2 distance with teacher feature. When the TaT map is an identity matrix, it means that $\phi(f^s)$ can reconstruct the teacher feature on its own. However, since TaT map is not strictly the identity matrix, it indicates that each pixel of $\phi(f^s)$ still needs to *borrow* the semantic from other position (mostly neighborhood) to enhance itself. Indeed, by aggregating the semantic from neighbors, each pixel increases the receptive field and thus semantic capacity. This demonstrates the semantic mismatch between student and teacher due to the variation on network depth and width.

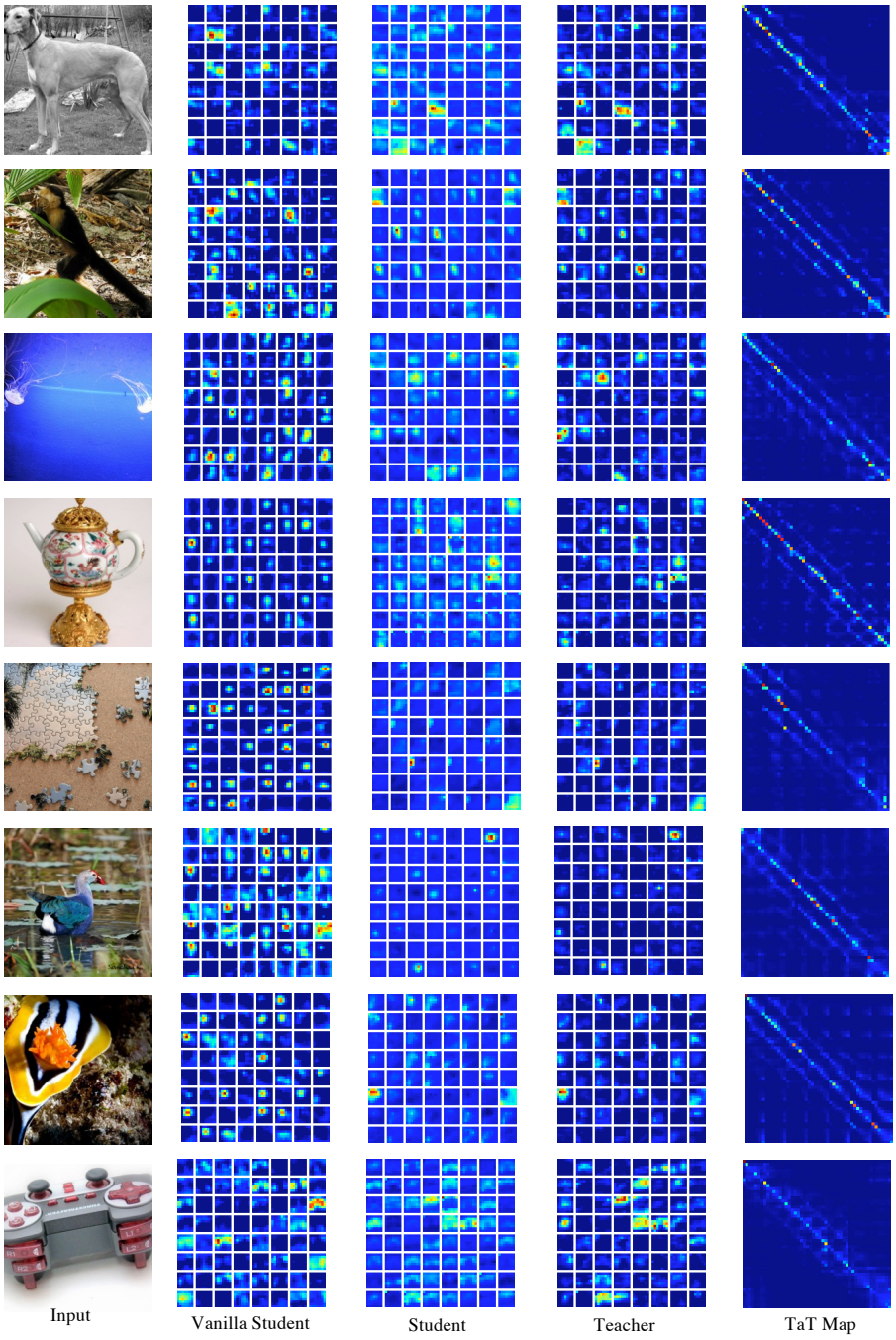

Figure 1. **Visualization of feature map and TaT map.** The input is selected from ImageNet validation set. The teacher backbone is ResNet34 and student backbone is ResNet18. The feature map of the distillation layer (4-th block) has been visualized. While there are 512 feature channels in total, we visualize 64 channels for better visualization. Through the Target-aware transformer, we found that the reconfigured student feature (3rd column) has a similar pattern with teacher feature (4th column). The associated TaT map has also been visualized, which indicates the student would aggregate the semantic mostly from neighbor to enhance its pixels.

References

[1] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, H. Park, N. Kwak, and J. Choi. A comprehensive overhaul of feature distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1921–1930, 2019.

[2] Li Liu, Qingle Huang, Sihao Lin, Hongwei Xie, Bing Wang, Xiaojun Chang, and Xiaodan Liang. Exploring inter-channel correlation for diversity-preserved knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8271–8280, October 2021.