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

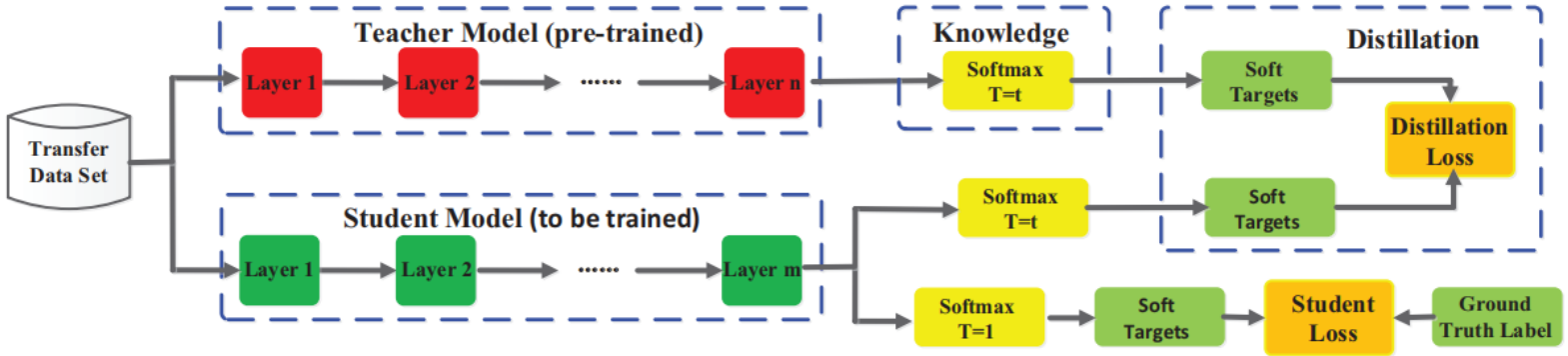

Fig. 4 The specific architecture of the benchmark knowledge distillation.

$$\mathcal{L}_{KD_i} = \mathcal{L}_{CE}(\boldsymbol{y}_i, \sigma(\boldsymbol{g}_i^s)) + T^2 \mathcal{L}_{KL}(\sigma(\boldsymbol{g}_i^t/T), \sigma(\boldsymbol{g}_i^s/T)),$$

$$\mathcal{L}_{FMD} = \sum_{(s_l,t_l)\in\mathcal{C}} Dist\left(Trans^t\left(F_{t_l}^t\right), Trans^s\left(F_{s_l}^s\right)\right),$$

$$\mathcal{L}_{total} = \sum_{i=1}^{b} \mathcal{L}_{KD_i} + \beta \mathcal{L}_{FMD},$$

Our contributions can be summarized as follows:

- We propose the knowledge distillation via a target-aware transformer, which enables the whole student to mimic each spatial component of the teacher respectively. In this way, we can increase the matching capability and subsequently improve the knowledge distillation performance.
- We propose the hierarchical distillation to transfer local features along with global dependency instead of the original feature maps. This allows us to apply the proposed method to applications, which are suffered from heavy computational burden because of the large size of feature maps.
- We achieve state-of-the-art performance compared against related alternatives on multiple computer vision tasks by applying our distillation framework.

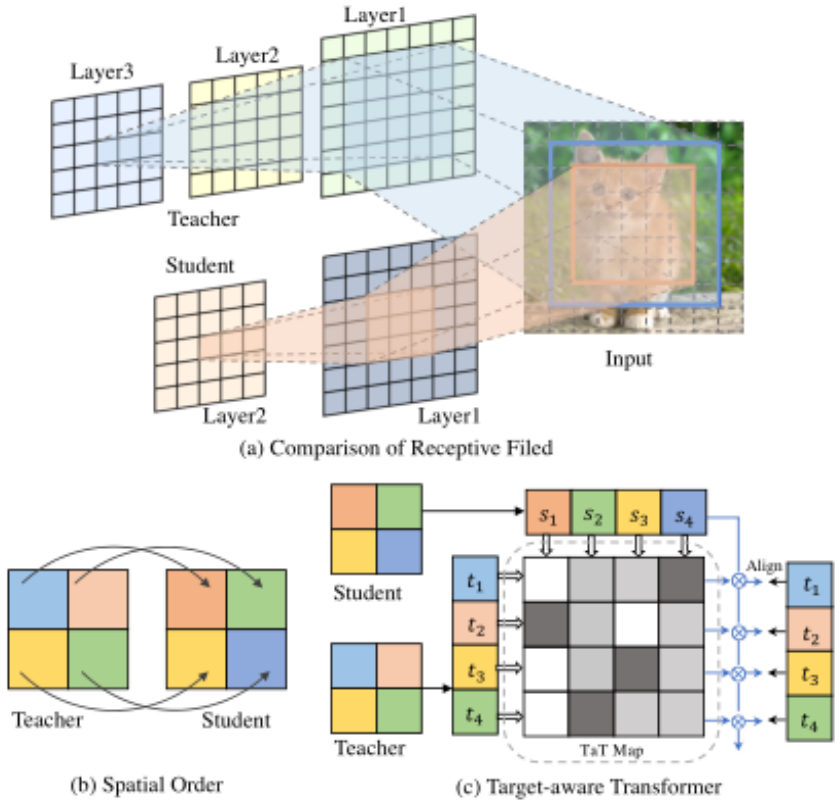

Figure 1. **Illustration of semantic mismatch.** Suppose that teacher and student are the 3-layers and 2-layers convnets with kernel size 3×3 and stride 1×1. (a) shows the receptive field of the middle pixel of the final feature map, where the blue box represents the teacher's receptive field and the orange box is that of the student's. Since teacher model has more convolutional operations, the resulting teacher feature map has a larger receptive field and thus contains richer semantic information. (b) Hence, directly regressing the student's and teacher's feature in a one-to-one spatial matching fashion may be suboptimal. (c) We proposed a one-to-all knowledge distillation via a target-aware transformer that can let the teacher's spatial components be distilled to the entire student feature maps.

将3维向量转为2维向量，N=H×W

$$f^s = \Gamma(F^S) \in \mathbb{R}^{N \times C}, \qquad f^{s\top} = [f_1^s, f_2^s, f_3^s, \ldots, f_N^s],$$
$$f^t = \Gamma(F^T) \in \mathbb{R}^{N \times C}. \qquad f^{t\top} = [f_1^t, f_2^t, f_3^t, \ldots, f_N^t].$$

使用TaT引导整个学生在其对应位置重建特征。在对齐目标的条件下，TaT应反映出与学生特征各组成部分的语义相似度。

$$\begin{aligned} W^i &= \sigma(\langle f_1^s, f_i^t \rangle, \langle f_2^s, f_i^t \rangle, \ldots, \langle f_N^s, f_i^t \rangle) \\ &= [w_1^i, w_2^i, \ldots, w_N^i], \end{aligned}$$

使用内积来测量语义距离，使用softmax函数进行归一化。

$$f_i^{s'} = w_1^i \times f_1^s + w_2^i \times f_2^s + \cdots + w_N^i \times f_N^s.$$

$$f_i^{s'} = \sigma(f^s \cdot f_i^t) \cdot f^s. \qquad f^{s'} = \sigma(\gamma(f^s) \cdot \theta(f^t)^\top) \cdot \phi(f^s),$$

Loss

$$\mathcal{L}_{\text{TaT}} = ||f^{s'} - f^t||_2. \qquad \mathcal{L} = \alpha \mathcal{L}_{\text{Task}} + \beta \mathcal{L}_{\text{KL}} + \epsilon \mathcal{L}_{\text{TaT}},$$

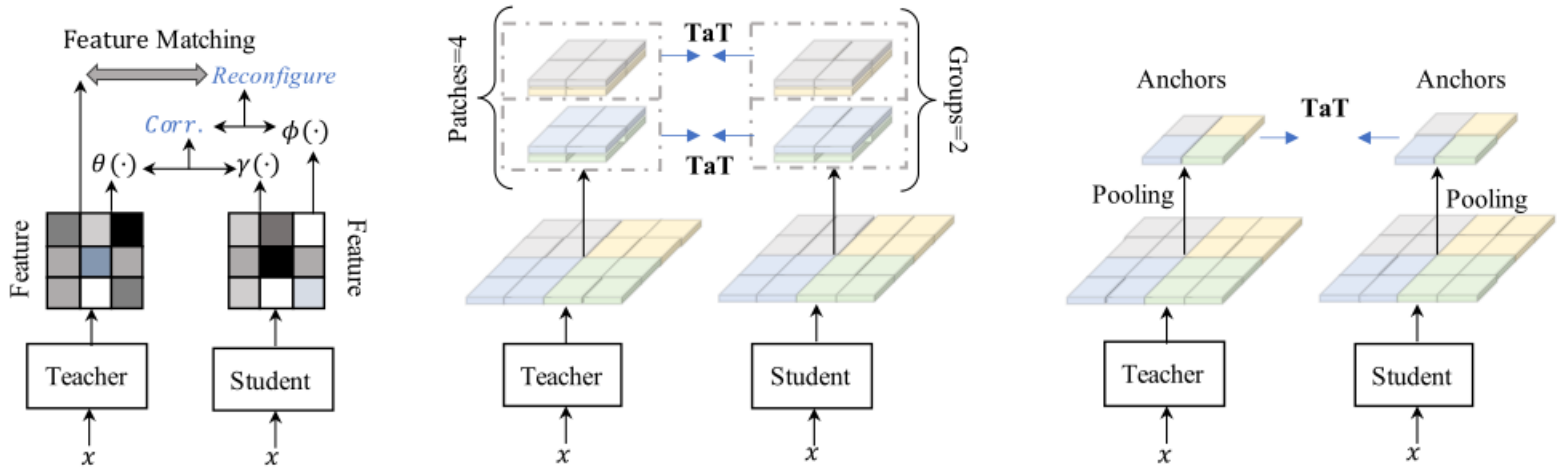

Figure 2. Illustration of our framework. (a) **Target-aware Transformer**. Conditioned on the teacher feature and the student feature, the transformation map *Corr.* is computed and then applied on the student feature to reconfigure itself, which is then asked to minimize the L_2 loss with the corresponding teacher feature. (b) **Patch-group Distillation**. Both teacher and student features are to be sliced and rearranged as groups for distillation. By concatenating the patches within a group, we explicitly introduce the spatial correlation among the patches beyond the patches themselves. (c) **Anchor-point Distillation**. Each color indicates a region. We use average pooling to extract the *anchor* within a local area of the given feature map, forming the new feature map of a smaller size. The generated anchor-point features will participate in the distillation.

ture F^S and teacher feature F^T, they are partitioned into $n \times m$ patches of size $h \times w$, where $h = H/n$, $w = W/m$. They are further arranged as g groups sequentially where each group contains $p = n \cdot m/g$ patches. Specifically, the patches in a group will be concatenated channel-wisely, forming a new tensor of size $h \times w \times c \cdot g$ that would be used for distillation lately. In this way, each pixel of the new tensor contains the features from p positions of the original feature, which explicitly includes the spatial pattern. There-

Patch-group Distillation 使学生能够模仿局部特征，而Anchor-point Distillation使学生能够学习粗糙特征的全局表示，它们是相互补充的。

$$\mathcal{L}_{\text{Seg}} = \alpha \mathcal{L}_{\text{CE}} + \delta \mathcal{L}_{\text{TaT}}^{\mathcal{P}} + \zeta \mathcal{L}_{\text{TaT}}^{\mathcal{A}}$$

Table 1. **Top-1 accuracy(%) on Cifar-100.** The loss term $\mathcal{L}_{\mathrm{KL}}$ in Eq. 9 is removed in this experiment.

Method	Network Architecture						
	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	ResNet56 ResNet20	ResNet110 ResNet20	ResNet110 ResNet32	ResNet32×4 ResNet8×4	VGG13 VGG8
Teacher	75.61	75.61	72.34	74.31	74.31	79.42	74.64
Vanilla	73.26	71.98	69.06	69.06	71.14	72.50	70.36
KD [19]	74.92	73.54	70.66	70.67	73.08	73.33	72.98
FitNet [36]	73.58	72.24	69.21	68.99	71.06	73.50	71.02
AT [50]	74.08	72.77	70.55	70.22	72.31	73.44	71.43
SP [43]	73.83	72.43	69.67	70.04	72.69	72.94	72.68
CC [35]	73.56	72.21	69.63	69.48	71.48	72.97	70.71
RKD [33]	73.35	72.22	69.61	69.25	71.82	71.90	71.48
PKT [34]	74.54	73.45	70.34	70.25	72.61	73.64	72.88
FSP [48]	72.91	NA	69.95	70.11	71.89	72.62	70.20
NST [20]	73.68	72.24	69.60	69.53	71.96	73.30	71.53
CRD [42]	75.48	74.14	71.16	71.46	73.48	75.51	73.94
ICKD [27]	75.64	74.33	**71.76**	71.68	73.89	75.25	73.42
Ours w/o $\mathcal{L}_{\mathrm{KL}}$	**76.06**	**74.97**	71.59	**71.70**	**74.05**	**75.89**	**74.39**

Table 2. Top-1 Accuracy(%) on ImageNet validation set. The ResNet34 is employed as the teacher backbone and the ResNet18 is selected as the student backbone. Our method can boost the performance of the tiny ResNet18 beyond 72% and outperforms other methods without $\mathcal{L}_{\mathrm{KL}}$.

Method	Vanilla	AT [50]	CRD [42]	SAD [22]	ICKD [27]	KR [7]	Ours	KD [19]	SCKD [3]	CC [35]	RKD [33]	Ours	Teacher
w/ $\mathcal{L}_{\mathrm{KL}}$								✓	✓	✓	✓	✓	-
Top-1	70.04	70.59	71.17	71.38	71.59	71.61	72.07	70.68	70.87	70.74	71.34	**72.41**	73.31

Table 3. **Comparing the semantic segmentation results (in mIoU%) of different methods on Pascal VOC.** We can observe that our method surpasses all previous baselines by a significant margin. Specifically, on the popular compact architecture MobilenetV2, our method improves the student by 5.39% comparing to the stand-alone training, and by 1.06% comparing to the state-of-the-art method ICKD. † indicates reproducing by training 100 epochs, using the official released code.

	ResNet18	MobilenetV2
Teacher	78.43	78.43
Student	72.07	68.46
KD [19]	73.74	71.73
AT [50]	73.01	71.39
FitNet [36]	73.31	69.23
Overhaul† [18]	73.98	72.30
ICKD [27]	75.01	72.79
Ours	**75.76**	**73.85**

Table 5. **Comparing the semantic segmentation results (in mIoU%) of different methods on COCOStuff10k.** As most baselines do not provide the code on the COCO dataset except KR, we only compare our method to KR in this case. We reproduce the baseline using the official code with the same training procedure. Our method surpasses the baseline by nearly 2%, and further demonstrates the effectiveness of our approach.

	Sutdent	KR [7]	Our	Teacher
ResNet18	26.33	26.73	**28.75**	33.10
MobilenetV2	26.29	26.63	**28.05**	33.10

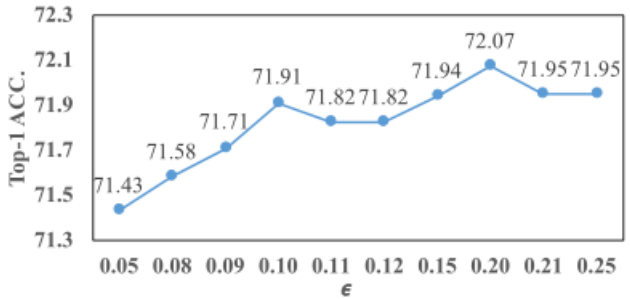

Figure 3. The performance of our model under different ϵ on ImageNet. Here the loss $\mathcal{L}_{\mathrm{KL}}$ is removed and α is set to 0.1.

Table 4. Non-parametric vs. parametric implementation of target-aware transformer on ImageNet, where check mark indicates applying linear function.

$\theta(\cdot)$	$\gamma(\cdot)$	Top-1 Acc.
		72.22
	✓	**72.41**
✓	✓	72.35

Table 7. Contribution of patch-group and anchor-point distillation. We observe that patch-group distillation presents more efficacy.

Anchor-point	Patch-group	mIoU
		72.07
✓		75.37
	✓	75.63
✓	✓	**75.76**

Table 8. Performance (%) and training time (minutes) of anchor-point distillation on Pascal VOC under different kernel sizes.

Pooling kernel	2×2	4×4	8×8	16×16
Training time	423	403	389	374
mIoU	**75.37**	75.27	74.79	74.56

Table 9. Performance (%) of patch-group distillation on Pascal VOC under different settings of patch size ($h \times w$). Groups is equal to patches $g = n \times m$.

Patch size	32×32	16×16	8×8	4×4
mIoU	75.33	75.45	**75.50**	75.47

Table 10. Performance (%) of patch-group distillation on Pascal VOC under different settings of groups where patch size is 8×8 and patch numbers is 256.

Groups	1	32	64	128	256
mIoU	75.26	75.57	**75.63**	75.62	75.50