

Supplementary Materials: Cross-View Consistency Regularisation for Knowledge Distillation

Anonymous Authors

This material provides supplementary content for our main paper. The pseudo-code of our method is provided in Section 1; the list of strong augmentation operations used is described in Section 2; further explanations on the feature-space consistency regularisation experiments presented in the main text are provided in Section 3; more ablation experiments and analyses are presented in Sections 4 and 5, respectively.

1 PSEUDO-CODE

In Algorithm 1, we provide the pseudo-code for the proposed CRLD algorithm. SLS(\cdot) denotes the confidence-based soft label mining operation using threshold τ , which produces a binary mask \mathbf{M} indicating the selected instance-wise predictions. Other notations in Algorithm 1 follow those defined in the main text.

2 LIST OF STRONG VIEW TRANSFORMATION OPERATIONS

Table 3 lists the image transformation operations used for strong view augmentation in CRLD. All transformations except for Cutout [4] are part of the RandAugment strategy initially proposed in [2]. In our experiments, $n = 2$ operations are randomly sampled from all 14 RandAugment transformation strategies, followed by Cutout. The strength (i.e., the operation parameter) v is set independently for each sampled operation and stochastically using the following equation:

$$v = v_{\min} + (v_{\max} - v_{\min}) * p \quad (1)$$

where v_{\min} and v_{\max} are the the lower and upper bounds of the parameter range for corresponding operations in Table 3; $p \in [0, 1]$ is a random number for stochastic parameter adjustment. For Cutout, its parameter v_{co} is generated by:

$$v_{\text{co}} = 0.5 \times p_{\text{co}} \quad (2)$$

where $p_{\text{co}} \in [0, 1]$ is another random number such that $v_{\text{co}} \in [0, 0.5]$ always holds.

3 FEATURE-SPACE CONSISTENCY REGULARISATION

In this section, we provide further details regarding our experiments on feature-space consistency regularisation described in Section 4.5 of the main text. Notation-wise, we use pool-feat to denote the pooled feature map by average pooling, right before the Softmax layer; we use feats-i to denote the feature map produced by the i th feature blocks immediately after the activation layer. Please refer to the *mdistiller* codebase for details.

In terms of network design, feature-based knowledge distillation method FitNets [6] adopts a convolutional regressor layer to adapt the student feature to the teacher feature. For a fair performance comparison, we follow this practice by employing two such layers, one for student’s predictions of the weakly-augmented view and the other for the strongly-augmented view.

Algorithm 1: The CRLD algorithm

Input: A batch of training samples \mathbf{x} & their labels \mathbf{y} ; weak augmentation $T_w(\cdot)$ & strong augmentation $T_s(\cdot)$; teacher network \mathcal{F}^T with parameters θ^T & student network \mathcal{F}^S with parameters θ^S

while model \mathcal{F}^S not converged **do**

for $i=1$ to step **do**

$\mathbf{p}_w^T = \mathcal{F}^T(T_w(\mathbf{x}); \theta^T)$ $\mathbf{p}_s^T = \mathcal{F}^T(T_s(\mathbf{x}); \theta^T)$

$\mathbf{p}_w^S = \mathcal{F}^S(T_w(\mathbf{x}); \theta^S)$ $\mathbf{p}_s^S = \mathcal{F}^S(T_s(\mathbf{x}); \theta^S)$

$\mathbf{M}_w = \text{SLS}(\mathbf{p}_w^T, \tau_w)$ $\mathbf{M}_s = \text{SLS}(\mathbf{p}_s^T, \tau_s)$

$\mathcal{L}_{CE} = \text{CE}(\mathbf{p}_w^S, \mathbf{y})$

$\mathcal{L}_{KD}^{WV} = \text{KLD}(\mathbf{p}_w^S, \mathbf{p}_w^T) \mathbf{M}_w + \text{KLD}(\mathbf{p}_s^S, \mathbf{p}_s^T) \mathbf{M}_s$

$\mathcal{L}_{KD}^{CV} = \text{KLD}(\mathbf{p}_s^S, \mathbf{p}_w^T) \mathbf{M}_w + \text{KLD}(\mathbf{p}_w^S, \mathbf{p}_s^T) \mathbf{M}_s$

$\mathcal{L}_{KD} = \mathcal{L}_{KD}^{WV} + \mathcal{L}_{KD}^{CV}$

$\mathcal{L}_{Overall} = \mathcal{L}_{CE} + \lambda_{KD} \mathcal{L}_{KD}$

 Update θ^S acc. to $\mathcal{L}_{Overall}$

end

end

Output: Well-trained model \mathcal{F}^S with parameters θ^S

Table 1: Ablation experiments on student’s self-supervised regularisation using Tiny-ImageNet.

$\mathbf{p}_w^S - \mathbf{p}_w^T$	$\mathbf{p}_s^S - \mathbf{p}_s^T$	$\mathbf{p}_w^S - \mathbf{p}_s^T$	$\mathbf{p}_s^S - \mathbf{p}_w^T$	$\mathbf{p}_w^S - \mathbf{p}_s^S$	ResNet32×4
✓					60.32 / 82.81
✓				✓	54.87 / 79.50
✓	✓				60.69 / 83.11
✓	✓			✓	43.52 / 69.60
✓	✓	✓	✓		63.77 / 84.57
✓	✓	✓	✓	✓	63.25 / 83.87

Table 2: Ablation experiments on the strengths of view transformation on CIFAR-100 and Tiny-ImageNet.

Method	CIFAR-100	Tiny-ImageNet
w/o CVL	76.26	60.83 / 83.08
Weak-Weak	76.66	62.83 / 84.10
Strong-Strong	76.73	61.67 / 83.94
Strong-Weak	78.31	63.77 / 84.57

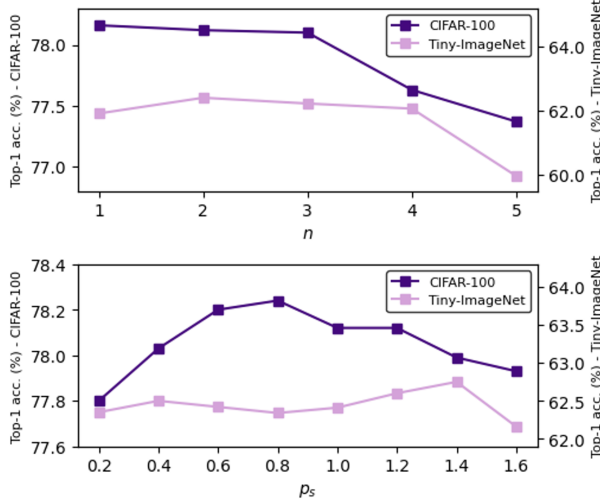
As for the loss function, following FitNets, we adopt the mean squared error (MSE) loss in place of the original Kullback-Leibler Divergence (KLD) loss for our consistency regularisation objectives.

4 ADDITIONAL ABLATION EXPERIMENTS

Effect of student self-supervision. We conduct further ablation experiments on the effect of student’s cross-view self-supervision on the Tiny-ImageNet dataset. As shown in Table 1, the inclusion of student’s self-supervision degrades the overall knowledge

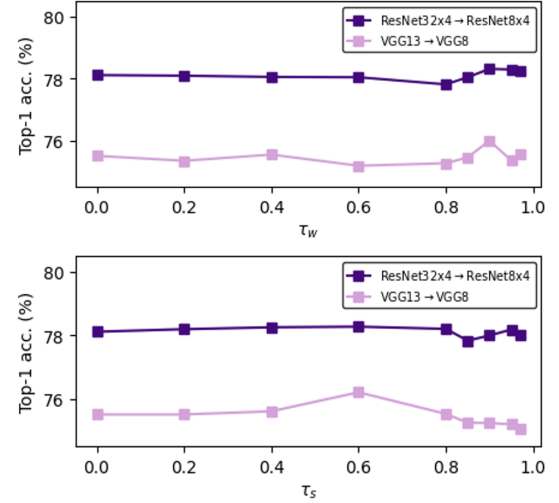
Table 3: List of transformation operations used for strong view transformation.

Transformation	Description	Param. Range
Autocontrast	Automatically adjusts image contrast by setting the darkest pixel to black and lightest to white	-
Brightness	Adjusts image brightness	[0.05, 0.95]
Color	Adjusts image colour balance	[0.05, 0.95]
Contrast	Adjusts image contrast	[0.05, 0.095]
Equalize	Equalises image histogram	[0, 1]
Identity	Keeps image unchanged	[0, 1]
Posterize	Reduces number of bits for each image channel	[4, 8]
Rotate	Rotates image	[-30, 30]
Sharpness	Adjusts image sharpness	[0.05, 0.95]
Shear_x	Shears image along horizontal axis	[-0.3, 0.3]
Shear_y	Shears image along vertical axis	[-0.3, 0.3]
Solarize	Inverts all image pixels above a given threshold	[0, 256]
Translate_x	Translates image horizontally	[-0.3, 0.3]
Translate_y	Translates image vertically	[-0.3, 0.3]
Cutout	Sets pixels side a random square path within image to gray	[0, 0.5]

**Figure 1: Sensitivity of CRLD against varying strengths of strong view transformation on CIFAR-100 and Tiny-ImageNet.**

distillation performance. More severe performance degradation is observed when less teacher supervision is imposed on the student. In our experiments, we also observe that the self-supervision loss oscillates dramatically in the initial stage of training, before quickly dropping to extremely small values. Our observations and conclusions on Tiny-ImageNet align with those made on CIFAR-100 [5] described in the main text.

Effect of view transformation pairs with different strengths. The success of CRLD hinges on a pair of strongly and weakly transformed images to establish cross-view consistency regularisation objectives. It is of interest to investigate to what extent the absolute and relative strengths in a pair of view transformations impact the subsequent cross-view learning. To this end, in Table 2 (ResNet32×4

**Figure 2: Sensitivity of CRLD against varying τ_w and τ_s values on CIFAR-100.**

as teacher and ResNet8×4 as student) we consider two additional cases: 1) using two independent weakly transformed views (denoted as “Weak-Weak”); 2) using two independent strongly transformed views (“Strong-Strong”), and compare them against the original strong-weak consistency regularisation design (“Strong-Weak”) and the baseline set-up without cross-view learning applied (“w/o CVL.”).

We easily draw the following conclusions: 1) Any form of cross-view learning, despite different view transformation strengths, leads to performance gains over the baseline, which again substantiates the effectiveness of our proposed cross-view consistency regularisation. 2) A pair of view transformations of identical strength results in degraded performance compared to the proposed strong-weak learning. We attribute this to additional dark information mined

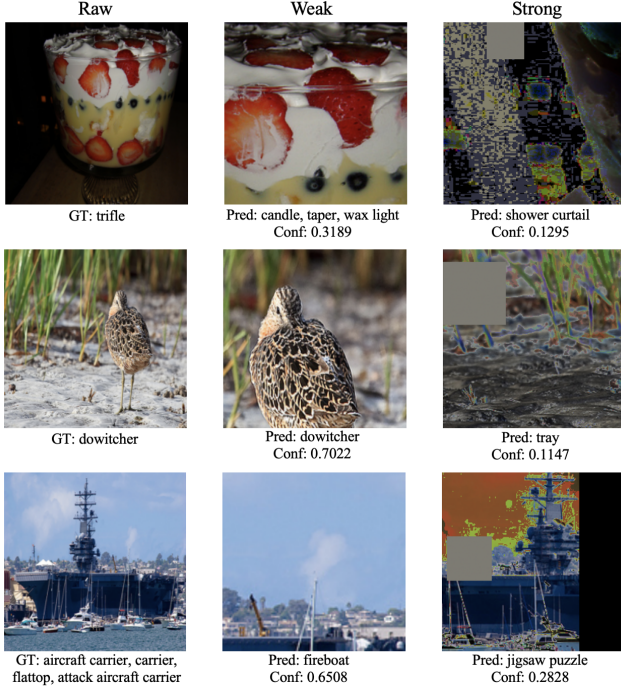


Figure 3: Examples of ImageNet [3] images transformed by the proposed weak and strong transformations and predictions made by a ResNet32×4 teacher.

and transferred across two different spaces of transformed images, compared to limited knowledge dug within a single space. 3) When using transformation pairs of the same strength, it is not decisive what strength level may be more beneficial – this may be dataset- and task-dependent.

Sensitivity to varying strengths of strong view transformation. Following previous investigations, we further carry out a set of experiments to probe into the impact of strong view transformation in different strengths on CRLD’s performance. First, we vary n , the number of view transformation operations randomly sampled and applied sequentially from all RandAugment operations in Table 3. As shown in the top figure in Figure 1, more strong view transformation operations degrade the performance of CRLD. This is expected since with an increasingly challenging strongly-augmented view, the teacher struggles to provide correct and beneficial soft predictions, and the student could be misled by a predominant amount of distracting and harmful signals from the teacher. Although the value of n can be tweaked for each dataset and even for each teacher-student configuration for further performance gains, we simply use $n = 2$ by default for simplicity.

To enable fine-grained control over the strength of strong augmentations (*i.e.*, RandAugment operations and Cutout), we also introduce a probability multiplier p_s to tune the parameter value of each operation (listed in Table 3). p_s is introduced into Equations 3 and 4 as:

$$v = v_{\min} + (v_{\max} - v_{\min}) * p * p_s \quad (3)$$

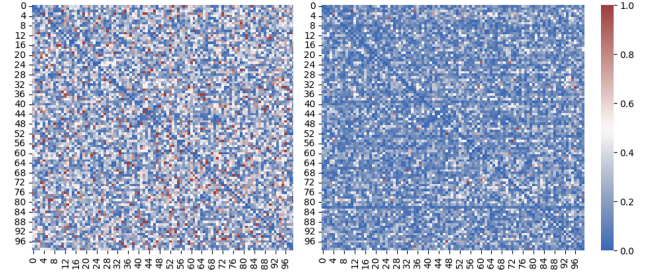


Figure 4: Class-wise similarity maps between teacher and student predictions by NormKD and CRLD on CIFAR-100.

and

$$v_{co} = 0.5 \times p_{co} * p_s \quad (4)$$

Note that a higher p_c value does not mean stronger transformation for all operations listed in Table 3. Nevertheless, larger p_c leads to more strongly transformed images on average, and we are interested in finding out how sensitive our method is to changes in the parameter values. From the bottom plot in Figure 1, we notice that the performance indeed varies with changing p_c . Overly large or small p_c tends to produce inferior performance, which echoes our findings in Table 2 and the above experiments on different n values. Besides, different datasets are observed to manifest different sensitivity patterns to p_c . More fine-grained control of the transformation parameters are left for future work.

Sensitivity to varying confidence thresholds. The confidence-based soft label mining mechanism essentially features a quantity-quality trade-off. With a higher threshold, we demand soft labels of higher quality but fewer of them are selected for knowledge transfer; with a lower threshold, we have richer knowledge in the form of teacher’s soft labels involved in the knowledge transfer, but their quality and reliability on average are lowered. Figure 2 visualises such a trade-off by plotting the performance of CRLD against different values of τ_w and τ_s .

We notice that the optimal trade-off point for τ_w is at a higher value. This is expected since the predicted confidence for the less challenging weakly-augmented view is much higher on average, which means a sufficient number of soft predictions of the teacher fall within the top-confidence interval. As such, setting a high τ_w ensures soft-labels are selected in high quality while also in ample quantity. By contrast, most teacher predictions for the strong view are less confident. A much smaller τ_s is required to ensure sufficient teaching signals. In practice, we set $\tau_w = 0.9$ and $\tau_s = 0.3$.

5 ADDITIONAL ANALYSES

Examples of challenging strongly augmented images. In Figure 3, we showcase some challenging examples produced by the proposed view transformation to support our motivation for confidence-based soft label selection. As can be seen, both weak and strong views can be misclassified by a well-trained ResNet32×4 teacher model. In particular, the strongly augmented view can be extremely challenging and sometimes almost completely indiscernible. Cross-view consistency imposed across these misleading predictions only

serves to harm the student’s learning, which is avoided by our proposed confidence-based soft label selection.

Teacher-student output correlations. To understand how well a trained student is able to mimic its teacher’s predictions, we compute and visualise the correlations between student’s and teacher’s predictions in the Euclidean space in Figure 4 (ResNet32×4 as teacher and ResNet8×4 as student). The left map corresponds to NormKD [1] and the right CRLD applied to NormKD. It is clear that with CRLD, the average distance between teacher and student predictions are significantly reduced for all categories on the test data — a compelling evidence of better distilled teacher knowledge and greater generalisation capabilities of the trained student.

REFERENCES

- [1] Zhihao Chi, Tu Zheng, Hengjia Li, Zheng Yang, Boxi Wu, Binbin Lin, and Deng Cai. 2023. NormKD: Normalized Logits for Knowledge Distillation. In *arXiv:2308.00520*.
- [2] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. 2020. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In *NIPS*.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.
- [4] Terrance DeVries and Graham W. Taylor. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. In *arXiv:1708.04552*.
- [5] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images.
- [6] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for Thin Deep Nets. In *ICLR*.