

SUPPLEMENTARY: SINGLE TEACHER, MULTIPLE PERSPECTIVES: TEACHER KNOWLEDGE AUGMENTATION FOR ENHANCED KNOWL- EDGE DISTILLATION

Md Imtiaz Hossain, Sharmen Akhter, Choong Seon Hong^{*} & Eui-Nam Huh^{*}
Department of Computer Science & Engineering, Kyung Hee University, South Korea
{hossain.imtiaz, sharmen, cshong, johnhuh}@khu.ac.kr

CONTENTS

A	TeKAP using Different Methods of Noise Generation	3
B	Hyper-parameter Analysis	3
C	Effects of λ for different number of teachers	4
D	Effect of TeKAP on class imbalance datasets.	4
E	Effect of TeKAP with different variance σ.	5
F	Static vs Dynamic Noise	5
G	Comparison with teacher assistant based approach with narrow teacher assistant	5
H	Attention visualization and comparison	6
I	Theoretical Justification and insights into perturbation methods (Expanded)	7
I.1	Role of Perturbations in Enhancing Generalization	7
I.2	Theoretical Justification of Perturbations	7
I.3	Dynamic Perturbations: Mimicking an Ensemble	7
I.4	Analytical Comparison of Noise Types	7
I.5	Feature and Logit-Level Perturbations	8
J	Comparative Computational Complexity	8

^{*}Corresponding authors.

K How Inter-Class Diversity Works	8
L Datasets and Setups	8
M Experimental Details	9
M.1 Competing SOTA Methods	9
M.1.1 Noise Details	9
M.2 Hyperparameter: λ	9
M.3 Hyperparameter: α	9
M.4 Network Architectures	9
M.5 Implementation Details	10
M.5.1 For KD, CRD, TeKAP(F), TeKAP (L), TeKAP (F+L)	10
M.5.2 For DKD, MLKD, and CA-MKD	10
M.5.3 For TAKD	10
M.5.4 For DGKD	11

A TeKAP USING DIFFERENT METHODS OF NOISE GENERATION

Network	Noise	TeKAP (F+L)
resnet32x4-resnet8x4	Gaussian	75.98
	Uniform	75.71
WRN_40.2-WRN_40.1	Gaussian	74.41
	Uniform	74.26

Table 1: Effect of different noise augmentation techniques on TeKAP (F+L).

The results presented in Table 1 highlight the impact of different noise augmentation techniques, Gaussian and Uniform distribution, on the TeKAP (F+L) for two different network configurations: resnet32x4-resnet8x4 and WRN_40.2-WRN_40.1. Here F+L denotes the proposed approach TeKAP is trained with KD + CRD [Tian et al. \(2019\)](#) distillation approach. For both networks, the Gaussian noise augmentation technique outperforms the Uniform noise technique. Specifically, resnet32x4-resnet8x4 achieves a TeKAP (F+L) score of 75.98% with Gaussian noise, which is higher than its Uniform counterpart (75.71%). Similarly, for the WRN_40.2-WRN_40.1 network, Gaussian noise also leads to higher performance (74.41%) compared to Uniform noise (74.26%). These results suggest that Gaussian noise, which introduces a smoother form of perturbation to the data, may help the models generalize better by mimicking more natural variations in data, leading to slightly improved TeKAP scores. On the other hand, Uniform noise, which introduces more abrupt changes, appears to be less effective, resulting in marginally lower performance across both network configurations. This analysis indicates that the choice of noise technique can significantly influence the effectiveness of noise augmentation in training, with Gaussian noise proving to be more beneficial for model performance in the given scenarios.

B HYPER-PARAMETER ANALYSIS

Table 2 shows the effect of TeKAP(F+L) on CIFAR100 for different numbers of original and augmented teachers. First, we have trained resnet32x4 three times to achieve three pretrained teacher networks. While we have used TeKAP with one original teacher and three augmented teachers, we have achieved an accuracy of 75.98%. For two original pretrained teachers and three augmented teachers for every original teacher, i.e., six augmented teachers, the proposed TeKAP achieves a slightly better accuracy of 76.12%. The results for three original teachers and three augmented teachers per every original teacher, i.e., a total of nine augmented teachers and 76.19%. Hence the effect of TeKAP in ensemble learning is proved by these results. TeKAP successfully adds a positive impact on ensemble learning and increasing the number of teachers enhances diversity along with the original teacher which helps improve performance by the student.

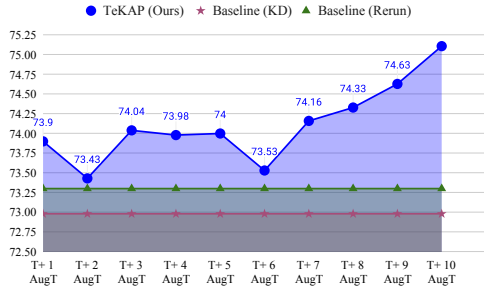


Figure 1: Effects of the number of augmented teachers on TeKAP on KD

Teacher	Accuracy
1 Original + 3 AugT	75.98
2 Original + 3 AugT	76.12
3 Original + 3 AugT	76.19

Table 2: The effects of multiple original teachers. We deploy three augmented teachers to every original teacher. resnet32x4 and resnet8x4 are considered teacher-student. Here, "Original Teacher" represents the teacher who is trained with 240 epochs. Three different original teachers use three different initializations (i.e., random seeds). AugT denotes the augmented teacher achieved by distorting the original teacher logits with random noise.

Fig. 1 shows the effect of our proposed method TeKAP(L) for different numbers of augmented teachers for a single teacher network on KD.

C EFFECTS OF λ FOR DIFFERENT NUMBER OF TEACHERS

The results in Table 3 demonstrate the effect of varying the noise weight (λ) and the number of augmented teachers (AugT) on the performance of the student model. For AugT=5, the accuracy consistently improves as λ increases, starting from 74.26% at $\lambda = 0.2$ and reaching 75.12% at $\lambda = 0.8$. This trend indicates that higher noise weights contribute positively to the student’s generalization by introducing greater diversity. Similarly, for AugT=10, the performance improves from 74.29% at $\lambda = 0.2$ to 74.98% at $\lambda = 0.8$, but the gains are less pronounced compared to AugT=5, suggesting a saturation effect with a larger number of augmented teachers. During this experiment, we set the value for σ and α to 1 and 0.1, respectively.

Comparing AugT=5 and AugT=10 across λ values, having more augmented teachers generally results in better performance, particularly at mid-range λ values such as $\lambda = 0.6$, where AugT=10 achieves an accuracy of 74.85%, slightly higher than 74.63% for AugT=5. However, at $\lambda = 0.8$, AugT=5 performs slightly better than AugT=10, achieving the highest accuracy (75.12%), likely due to the balance between noise diversity and stability. These results suggest that for smaller numbers of augmented teachers, higher λ values are beneficial, whereas for larger numbers of augmented teachers, moderate λ values strike the optimal balance between diversity and effective knowledge transfer, preventing over-saturation.

Number of AugT	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.8$
AugT=5	74.26	74.46	74.63	75.12
AugT=10	74.29	74.73	74.85	74.98

Table 3: Effect of the different values of λ (the weight to the noise terms). AugT denotes augmented teacher.

Methods	resnet32x4-resnet8x4	WRN_40_2-WRN_16_2	VGG13-VGG8
Baseline (KD)	41.71	52.08	47.52
+ TeKAP (Ours)	46.42	52.72	51.25

Table 4: Effect of TeKAP on class imbalance dataset. KD [Hinton \(2015\)](#) is used as the baseline distillation approach. We have used the class distribution of the CIFAR100 dataset that is described in Table 8.

D EFFECT OF TEKAP ON CLASS IMBALANCE DATASETS.

The results presented in Table 4 highlight the effectiveness of TeKAP(L) in addressing class imbalance in knowledge distillation tasks. TeKAP improves the performance of all three teacher-student pairs (resnet32x4-resnet8x4, WRN_40_2-WRN_16_2, and VGG13-VGG8) compared to the baseline knowledge distillation (KD) approach. Specifically, TeKAP boosts accuracy by 4.71% for resnet32x4-resnet8x4, 0.64% for WRN_40_2-WRN_16_2, and 3.73% for VGG13-VGG8. These results indicate that TeKAP is particularly effective in enhancing performance for models with lower baseline accuracy, though it also provides improvements for models with higher baseline accuracy. This suggests that TeKAP can effectively improve the performance under the class imbalance scenario, leading to enhanced generalization in knowledge distillation tasks.

E EFFECT OF TeKAP WITH DIFFERENT VARIANCE σ .

Variance	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Accuracy	74.89	74.79	74.35

Table 5: Effect of TeKAP with different variance σ . KD Hinton (2015) is used as the baseline distillation approach. We have used mean zero in all the cases.

These results suggest that, within the range of variances tested, increasing the noise variance does not significantly degrade performance. The accuracy only decreases marginally as the variance increases from 0.5 to 1.5, which indicates the robustness of TeKAP to noise. This behavior suggests that TeKAP can maintain competitive performance even with varying levels of noise in the teacher models, highlighting its resilience to noise during distillation. The consistent results across different variances also support the idea that TeKAP is stable and less sensitive to slight perturbations in the teacher’s logits. This stability is critical for practical applications where noise may be present in the data or models.

Table 5 summarizes the impact of different variances (σ) on the performance of TeKAP, using the CIFAR-100 dataset. The baseline distillation approach, knowledge distillation (KD), is used for comparison. As shown in the results, the accuracy of the model remains relatively stable across varying values of σ .

Specifically, when $\sigma = 0.5$, the model achieves an accuracy of 74.89%, slightly higher than the accuracy at $\sigma = 1$ (74.79%) and $\sigma = 1.5$ (74.35%).

F STATIC VS DYNAMIC NOISE

Methods	resnet32x4-resnet8x4	WRN_40_2-WRN_16_2	VGG13-VGG8
Baseline (KD)	73.33	74.92	72.98
+ TeKAP (Static-L)	73.74	74.66	73.29
+ TeKAP (Ours: Dynamic-L)	74.79	75.21	74.00

Table 6: Evaluations on the comparative effects between static and dynamic noise. KD Hinton (2015) is used as the baseline distillation approach. The experiment is done with three augmented teachers. We use $\sigma = 1$, $\lambda = 0.8$, and three augmented teachers. Gaussian noise is used to generate the noise.

and 72.98% accuracy, respectively. Incorporating static noise (TeKAP Static-L) shows minor improvements for resnet32x4-resnet8x4 and VGG13-VGG8, achieving 73.74% and 73.29%, but performs slightly worse (74.66%) for WRN_40_2-WRN_16_2, indicating its limited adaptability. Conversely, our proposed dynamic noise strategy (TeKAP Dynamic-L) consistently outperforms both static noise and baseline KD, achieving significant gains with accuracies of 74.79%, 75.21%, and 74.00%, respectively. This superiority stems from dynamic noise’s adaptability enabling robust generalization. These findings underscore the robustness and efficacy of dynamic noise in enhancing knowledge transfer during distillation, providing a compelling case for its application in improving student network performance across diverse architectures.

The results in Table 7 demonstrate the effectiveness of dynamic noise over static noise on the baseline Knowledge Distillation (KD) approach for three teacher-student pairs, resnet32x4-resnet8x4, WRN_40_2-WRN_16_2, and VGG13-VGG8 on CIFAR100 dataset. Baseline KD provides solid performance, achieving 73.33%, 74.92%,

G COMPARISON WITH TEACHER ASSISTANT BASED APPROACH WITH NARROW TEACHER ASSISTANT

The results in Table 7 compare our proposed **TeKAP** with the traditional teacher assistant-based knowledge distillation method, TAKD Mirzadeh et al. (2020). For these experiments, we used WRN_40_2 as the

teacher and WRN_16.2 and WRN_40.1 as the student networks. TAKD incorporates narrow teacher assistants (WRN_22.1, WRN_22.2, WRN_16.1, and WRN_16.2) to mediate knowledge transfer, while TeKAP directly distills knowledge from the teacher to the student without using any intermediate assistants. The performance of TeKAP consistently surpasses that of TAKD across all configurations.

For example, when WRN_40.1 is used as the student, TeKAP achieves a consistent accuracy of **73.80%**, compared to TAKD’s highest accuracy of **73.26%**. Similarly, with WRN_16.2 as the student and WRN_22.2 as the teacher assistant, TAKD achieves 75.02%, while TeKAP slightly improves it to **75.21%**. These results demonstrate the superior capability of TeKAP in directly transferring knowledge, avoiding the bottlenecks introduced by intermediate teacher assistants.

Teacher	WRN_40.2					
Teacher Assistant	WRN_22.2	WRN_22.2	WRN_22.1	WRN_22.1	WRN_16.2	WRN_16.1
Student	WRN_16.2	WRN_40.1	WRN_16.2	WRN_40.1	WRN_40.1	WRN_40.1
TAKD	75.02	72.73	72.56	71.19	68.92	73.26
+ TeKAP (Ours)	75.21	73.80	75.21	73.80	73.80	73.80

Table 7: Comparison between TeKAP and the assistant teacher-based KD method TAKD. We have used KD loss for both methods (Hinton (2015)). We use $\sigma = 1$, $\lambda = 0.8$, and three augmented teachers. Gaussian noise is used to generate the noise for TeKAP. TeKAP outperforms TAKD without using any assistant teacher. We select WRN_40.2 as the teacher and WRN_16.2 and WRN_40.1 as the students. WRN_22.1, WRN_22.2, WRN_16.1, and WRN_16.2 are selected as the teacher assistant for TAKD. TeKAP does not use any assistant teachers. TeKAP transfers knowledge directly from the teacher to the student.

H ATTENTION VISUALIZATION AND COMPARISON

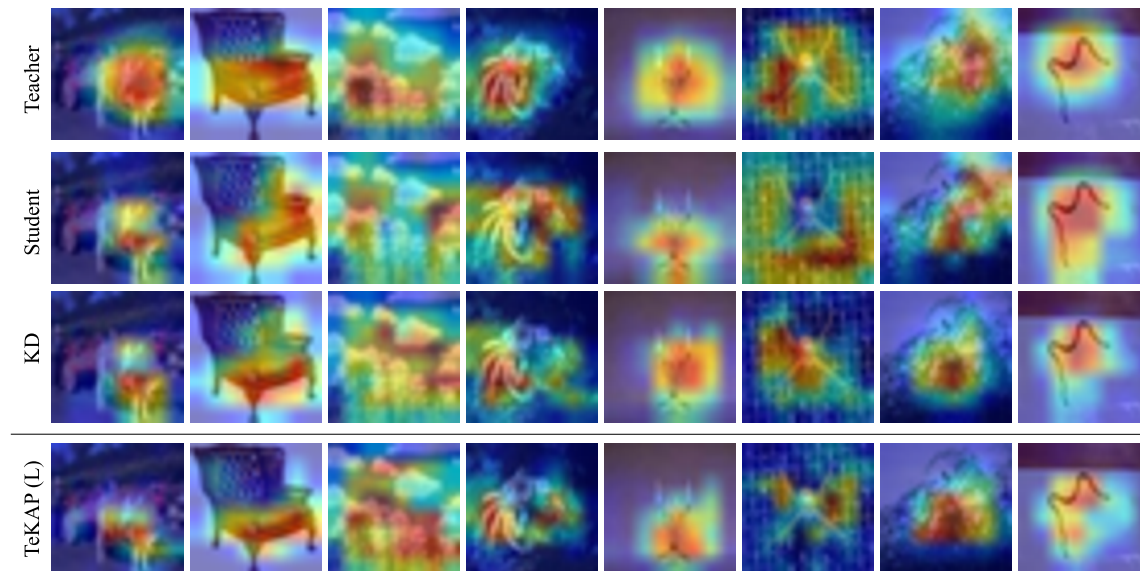


Figure 2: Visual comparison using gradient class activation maps (GradCAMs).

Fig. 2 shows the visual comparison among student networks regarding Gradient Class Activation Map, GradCAM Selvaraju et al. (2017). We select the images randomly from CIFAR100 datasets for this experiment. With the help of the augmented diversity, TeKAP-student achieves better attention to the targeted objects compared to the baseline and KD students. Both the teacher and student are trained on the CIFAR100 dataset. We consider resnet32x4 and resnet8x4 as the teacher and student, respectively.

I THEORETICAL JUSTIFICATION AND INSIGHTS INTO PERTURBATION METHODS (EXPANDED)

I.1 ROLE OF PERTURBATIONS IN ENHANCING GENERALIZATION

Perturbation methods such as Gaussian and Uniform noise introduce diversity in the outputs of the teacher network, enabling the student to learn from multiple perspectives. This diversity acts as a regularization mechanism, reducing overfitting and encouraging the student network to generalize better. Gaussian noise, with its smooth and continuous distribution, promotes gradual variations, which align well with natural data patterns. On the other hand, Uniform noise introduces evenly distributed, abrupt changes, which are less effective in creating realistic variations but still useful for regularizing overconfident predictions.

I.2 THEORETICAL JUSTIFICATION OF PERTURBATIONS

Adding perturbations to teacher outputs expands the hypothesis space \mathcal{H} that the student can explore, increasing the Rademacher complexity $\hat{R}_n(\mathcal{H})$. The variance of the perturbed outputs, such as teacher logits $z_T^{(i)}(x) = (1 - \alpha)z_T(x) + \alpha\eta$, increases as:

$$\text{Var}[z_T^{(i)}(x)] = \text{Var}[z_T(x)] + \alpha^2\sigma^2, \quad (1)$$

where $\eta \sim \mathcal{N}(0, \sigma^2)$ for Gaussian noise. The additional variance encourages the student to generalize better by exposing it to a broader range of data representations. The generalization error (GE) bound is:

$$\text{GE} \leq \hat{L}_{\text{emp}}(S) + \sqrt{\frac{2\hat{R}_n(\mathcal{H})^2 \log(2/\delta)}{n}}, \quad (2)$$

where $\hat{L}_{\text{emp}}(S)$ is the empirical risk and δ is the confidence parameter. Higher diversity in teacher outputs, driven by noise, reduces empirical risk and improves generalization.

I.3 DYNAMIC PERTURBATIONS: MIMICKING AN ENSEMBLE

TeKAP refreshes noise dynamically during training, simulating an ensemble-like behavior. This process exposes the student to a continuously varying set of teacher outputs, preventing overfitting to spurious patterns and encouraging the discovery of robust decision boundaries. The dynamic noise ensures that the student benefits from a wide range of teacher perspectives throughout the training process.

I.4 ANALYTICAL COMPARISON OF NOISE TYPES

- Gaussian Noise ($\eta \sim \mathcal{N}(0, \sigma^2)$):
 - Smoother and more natural variations.
 - Better alignment with real-world data distributions.
 - Superior generalization due to gradual decision boundary shifts.

- Uniform Noise ($\eta \sim U(a, b)$):
 - Abrupt changes evenly distributed within an interval.
 - Effective for regularization but less aligned with natural variations.

From our experiments, Gaussian noise consistently achieves better student performance than Uniform noise, supporting its theoretical advantages.

I.5 FEATURE AND LOGIT-LEVEL PERTURBATIONS

At the feature level, perturbations such as $f_T^{(i)}(x) = \alpha\eta + (1 - \alpha)f_T(x)$ provide diverse intermediate representations, acting as a form of regularization. At the logit level, perturbations modify inter-class relationships, creating alternative supervisory signals for the student:

$$z_T^{(i)}(x) = \alpha\eta + (1 - \alpha)z_T(x). \quad (3)$$

These variations force the student to explore a broader set of decision boundaries, as evidenced by improved inter-class correlation alignment.

J COMPARATIVE COMPUTATIONAL COMPLEXITY

The training of a teacher ResNet32x4 in CIFAR100 using KD takes approximately 16 seconds per epoch. As we run for 240 epochs, the total time taken is $240 \times 16 = 64$ minutes using two 3080 NVIDIA GeForce GPUs. For multi-teacher or ensemble learning, we need to train multiple teachers. Let’s assume two teacher assistants of equal size, which takes $64 \times 2 = 128$ minutes (approx) for DGKD [Son et al. \(2021\)](#). In our approach, TeKAP takes 18 seconds per epoch, which is 72 minutes in total. As TeKAP avoids training multiple teachers it shows a significant reduction in computational complexity.

K HOW INTER-CLASS DIVERSITY WORKS

If two classes are strongly correlated in the teacher logits, random distortions will not eliminate this correlation but may perturb its exact magnitude or direction, leading to diverse interpretations of the relationship. Imagine teaching a concept by showing slightly varied examples, this helps learners generalize the concept rather than memorize specific instances. Similar to techniques like dropout (which can be considered implicitly network ensemble learning because every random dropping creates a different network structure), random feature distortion (considered as a diverse network as the outputs are slightly different so it is assumed they come from different networks) can force the model to adapt to a broader range of conditions. This diversity helps the student model avoid collapsing into a rigid interpretation of the teacher’s outputs.

L DATASETS AND SETUPS

We have evaluated our approach TeKAP on four standard benchmarks: (1) ImageNet-1K: (training images - 1.2 million; validation images - 50K; classes - 1K) [Deng et al. \(2009\)](#), (2) CIFAR100: (training images - 50K; validation images - 10K; classes - 100) [Krizhevsky et al. \(2009\)](#), (3) STL-10: (training images: 5K labelled from 10 classes and 100K unlabeled images; test set of 8K images) [Coates et al. \(2011\)](#), and (4) TinyImageNet: (training images - 500, testing images - 50; classes - 200) [Deng et al. \(2009\)](#). We adopt the hyperparameters and experimental setups from [Tian et al. \(2019\)](#). To generate random noise we have used `torch.rand()`, and `torch.randn()` functions of the Pytorch library.

M EXPERIMENTAL DETAILS

M.1 COMPETING SOTA METHODS

We have compared and employed our proposed approach TeKAP on the following SOTA methods:

- Knowledge Distillation (KD) (Hinton (2015))
- Contrastive Representation Distillation (Tian et al. (2019))
- Improved knowledge distillation via teacher assistant (TAKD) (Mirzadeh et al. (2020))
- Densely guided knowledge distillation using multiple teacher assistants (DGKD) Son et al. (2021)
- Decoupled Knowledge Distillation (DKD) (Zhao et al. (2022))
- Confidence-aware multi-teacher knowledge distillation (CA-MKD) (Zhang et al. (2022))
- Multi-level Logits Distillation (MLKD) (Jin et al. (2023))

M.1.1 NOISE DETAILS

We have evaluated two noise engines: 1) Gaussian Noise and 2) Uniform Noise.

- Gaussian Noise: For Gaussian noise, we have used zero mean and one standard deviation $\eta_i \sim \mathcal{N}(0, \sigma^2)$ to generate noise. We have used 0.1 and 0.9 weights for augmented and original teachers, respectively, to distort the teacher logits by the noise. For the ablation study, we also evaluate the significance of TeKAP for different values of variance $\sigma = [0.2, 0.4, 0.6, 0.8]$. We also examine the weights for original and augmented teachers for the ablation study.
- Uniform Noise: While generating random noise. The range to random noise is $[0,1)$. We have used 0.1 and 0.9 weights for augmented and original teachers, respectively, to distort the teacher logits by the noise.

M.2 HYPERPARAMETER: λ

The default value for λ is 0.8. However, we also show the ablation study for different values of λ , ($[0.2, 0.4, 0.6]$)

M.3 HYPERPARAMETER: α

We have used the value for $\alpha = 0.1$. However, we also examine the effect of TeKAP for different values of α ($[0.1, 0.3, 0.5, 0.7, 0.9]$) in the ablation study.

M.4 NETWORK ARCHITECTURES

We have followed identical network structures described in CRD Tian et al. (2019). The details can be followed as:

- **ShuffleNets**: Referring to Zhang et al. (2018) and Tan et al. (2019), ShuffleNetV1 and ShuffleNetV2 are lightweight architectures that are optimized for efficient training. In our work, these are adjusted to handle input dimensions of 32×32 .
- **MobileNetV2**: As presented in Sandler et al. (2018), we utilize MobileNetV2 with a width multiplier of 0.5 for our experiments.

- **VGG**: The VGG network used in our experiments, inspired by [Simonyan \(2014\)](#), is a modified version of the original model designed for ImageNet.
- **ResNet (ImageNet Style)**: Based on [He et al. \(2016\)](#), ResNet-d here refers to an ImageNet-style ResNet architecture employing Bottleneck blocks and additional channels.
- **ResNet (CIFAR Style)**: As per [He et al. \(2016\)](#), resnet-d is used to describe a CIFAR-style ResNet architecture consisting of three groups of basic blocks with 16, 32, and 64 channels, respectively. In this study, resnet8x4 and resnet32x4 refer to a version of this network that is four times wider, featuring 64, 128, and 256 channels in each block.
- **Wide Residual Network (WRN)**: Following [Zagoruyko \(2016\)](#), WRN-d-w denotes a wide residual network with a depth of d and a width factor of w .

M.5 IMPLEMENTATION DETAILS

M.5.1 FOR KD, CRD, TEKAP(F), TEKAP (L), TEKAP (F+L)

We have followed the identical implementation details similar to CRD [Tian et al. \(2019\)](#).

All the methods tested in our experiments utilize SGD for optimization.

- **For CIFAR-100**: The learning rate is initially set to 0.05 and reduced by a factor of 0.1 every 30 epochs after the first 150 epochs, continuing until the final epoch at 240. For MobileNetV2, ShuffleNetV1, and ShuffleNetV2, a learning rate of 0.01 is used, as grid search experiments identified this value as optimal for these architectures, whereas a learning rate of 0.05 proved better for the other models.
- **For ImageNet**: We adopt the standard PyTorch training protocol, extending the training period by 10 additional epochs. The batch size is set to 64 for CIFAR-100 and 256 for ImageNet.

The student model is trained using a combination of the cross-entropy loss and a knowledge distillation loss, expressed as:

$$L = (1 - \beta) \times L_{\text{cross-entropy}} + \beta \times L_{\text{distill}} \quad (4)$$

For the weight balancing factor β , we use the optimal values specified in the original papers where available. Otherwise, a grid search is conducted using WRN-40-2 as the teacher and WRN-16-2 as the student. The grid search results determine the β values employed for different objectives. For CRD $\beta = 0.8$, in general, $\beta \in [0.5, 1.5]$ works reasonably well. For KD $\beta = 0.9$ [Tian et al. \(2019\)](#).

M.5.2 FOR DKD, MLKD, AND CA-MKD

For DKD [Zhao et al. \(2022\)](#), MLKD [Jin et al. \(2023\)](#) and CA-MKD [Zhang et al. \(2022\)](#) we have followed the identical implementation details of the officially released code of the corresponding paper. We only distorted the teacher logits by noises and other setups are identical.

M.5.3 FOR TAKD

First, we train the vanilla teacher network. Second, we train the assistant teachers and then distil the assistant teacher knowledge to the student using KD. For both TeKAP and TAKD, we have used a similar KD approach (KD and CRD). We have evaluated this experiment in the code-based released by CRD [Tian et al. \(2019\)](#). The experimental setup here is identical. The teacher and teacher assistant size can be found in the corresponding results Tables.

M.5.4 FOR DGKD

We follow a similar phenomenon for both TAKD and DGKD. For DGKD we only impose the distillation approach inspired by the discussion in the paper. We have re-implemented the DGKD in the CRD [Tian et al. \(2019\)](#) code. First, we train the teacher assistants and run multiple teachers to transfer the densely connected knowledge to the student. The other experimental setup is identical to CRD [Tian et al. \(2019\)](#).

Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Frequency	500	50	50	50	500	500	500	50	50	50	50	500	50	50	500	500	50	500	500	500
Index	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
Frequency	500	500	500	50	500	500	50	50	500	500	500	500	500	500	50	50	50	500	500	500
Index	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
Frequency	50	50	50	50	50	50	50	50	500	50	50	50	500	500	500	50	500	500	500	500
Index	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
Frequency	500	500	500	500	500	500	50	50	50	50	50	500	500	500	50	50	50	500	500	500
Index	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99
Frequency	500	50	500	500	50	50	500	500	50	50	50	500	500	50	50	50	500	500	500	500

Table 8: Data Distribution of the CIFAR-100 dataset to evaluate the effect of TEeKAP on class imbalance dataset.

REFERENCES

- Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24276–24285, 2023.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.
- Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

- 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*, pp. 618–626, 2017.
- Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9395–9404, 2021.
- Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2820–2828, 2019.
- Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.
- Sergey Zagoruyko. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4498–4502. IEEE, 2022.
- Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.
- Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962, 2022.