Progress Report
Experimental Results

Student:	Gheith Alrawahi
Student ID:	2120246006
Program:	Software Engineering
Institution:	Nankai University
Supervisor:	Prof. Jing Wang
Date:	December 2025

Contents

1 Executive Summary

we completed the entire experimental phase of the thesis. We investigated the stability of Knowledge Distillation (KD) under strong data augmentation on the CIFAR-100 dataset [Krizhevsky, 2009].

We present three key outcomes:

1. **Standard KD is superior in stability.** It achieved the highest accuracy (76.19%) and retention rate (99.40%). It outperformed Decoupled KD (DKD) in high-noise regimes.
2. **We identified the "Regularization-Distillation Conflict."** DKD is highly sensitive to augmentation noise. It collapsed when β was high. Standard KD remained robust.
3. **We confirmed Capacity Saturation.** Experiment v4 yielded accuracy identical to v2 (76.19%). This provides definitive evidence that the student model has reached its representational ceiling.

1.1 Results at a Glance

Table 1: Summary of model compression results.

Model	Accuracy	Parameters	Compression
Teacher (EfficientNetV2-L)	76.65%	118M	—
Distilled Student (Standard KD)	**76.19%**	21M	5.6× smaller
Distilled Student (DKD, β=8.0)	66.85%	21M	Collapsed
Distilled Student (DKD, β=2.0)	75.63%	21M	Recovered

We achieved 99.4% teacher accuracy retention with 5.6× model compression using Standard KD.

2 Methodology

2.1 Experimental Setup

We used the following configuration:

- **Teacher Model:** EfficientNetV2-L [Tan and Le, 2021], pre-trained on ImageNet [Deng et al., 2009], fine-tuned on CIFAR-100
- **Student Model:** EfficientNetV2-S (5.6× smaller than teacher)
- **Dataset:** CIFAR-100 (100 classes, 50,000 training images, 10,000 test images)
- **Hardware:** NVIDIA GeForce RTX 5070 Laptop GPU

2.2 Training Pipeline

We implemented an enhanced training pipeline with three components.

Data Augmentation:

- **AutoAugment** [Cubuk et al., 2019]: Automatically finds the best image transformations for the dataset.

- **Random Erasing** [Zhong et al., 2020] (p=0.25): Randomly masks parts of the image to improve robustness.
- **Mixup** [Zhang et al., 2018] (α=0.8): Blends two images and their labels together.
- **CutMix** [Yun et al., 2019] (α=1.0): Cuts a patch from one image and pastes it onto another.

Optimization:

- **AdamW** [Loshchilov and Hutter, 2019] (lr=0.001, weight_decay=0.05): Optimizer with proper weight decay for better generalization.
- **Cosine Annealing LR** [Loshchilov and Hutter, 2017]: Gradually reduces learning rate following a cosine curve.
- **Linear warmup** (5 epochs): Slowly increases learning rate at the start to stabilize training.
- **Label Smoothing** [Szegedy et al., 2016] (0.1): Softens hard labels to prevent overconfident predictions.

Training Stability:

- **Mixed Precision** (FP16): Uses 16-bit floats to speed up training and reduce memory.
- **Gradient clipping** (max_norm=1.0): Limits gradient size to prevent exploding gradients.
- **Early stopping** (patience=30): Stops training if validation loss does not improve for 30 epochs.

2.3 Distillation Methods

We compared two distillation methods.

Standard KD [Hinton et al., 2015]:

$$L_{KD} = \alpha \cdot T^2 \cdot \mathrm{KL}(p_s^T \| p_t^T) + (1 - \alpha) \cdot \mathrm{CE}(y, p_s) \tag{1}$$

where:

- L_{KD} = total loss for knowledge distillation
- α = balance weight between soft and hard labels (we used 0.7)
- T = temperature for softening probability distributions (we used 4.0)
- $\mathrm{KL}(\cdot)$ = Kullback-Leibler divergence (measures difference between two distributions)
- p_s^T = student's softened predictions at temperature T
- p_t^T = teacher's softened predictions at temperature T
- $\mathrm{CE}(\cdot)$ = cross-entropy loss with true labels
- y = ground truth labels
- p_s = student's predictions

Decoupled KD [Zhao et al., 2022]:

$$L_{DKD} = \alpha \cdot L_{TCKD} + \beta \cdot L_{NCKD} \tag{2}$$

The key idea is to separate the teacher's output into two parts:

$$L_{TCKD} = \mathrm{KL}\left(\frac{p_s^t}{p_s^t + \sum_{j \neq t} p_s^j} \middle\| \frac{p_t^t}{p_t^t + \sum_{j \neq t} p_t^j}\right) \tag{3}$$

$$L_{NCKD} = \mathrm{KL}\left(\frac{p_s^{\setminus t}}{\sum_{j \neq t} p_s^j} \middle\| \frac{p_t^{\setminus t}}{\sum_{j \neq t} p_t^j}\right) \tag{4}$$

where:

- L_{DKD} = total loss for decoupled knowledge distillation
- α = weight for target class component (we used 1.0)
- β = weight for non-target class component (we tested 8.0 and 2.0)
- L_{TCKD} = Target Class KD loss: matches the probability of the correct class between student and teacher
- L_{NCKD} = Non-Target Class KD loss: matches the distribution over all wrong classes (the "dark knowledge")
- p^t = probability of the target (correct) class
- $p^{\setminus t}$ = probabilities of all non-target classes

3 Experimental Results

3.1 Summary of All Experiments

Table 2: Comparison of all experimental configurations.

Exp.	Method	Augmentation	Student Acc.	Retention	Key Insight
v1	Standard KD	Mixup/CutMix	76.12%	99.31%	Baseline
v2	Standard KD	AutoAugment+	**76.19%**	**99.40%**	Optimal
v3	DKD (β=8.0)	AutoAugment+	66.85%	87.21%	Collapsed
v3.1	DKD (β=2.0)	AutoAugment+	75.63%	98.67%	Recovered
v4	Standard KD	AutoAugment+	76.19%	99.40%	Saturation

3.2 Detailed Analysis of Phases

Phase 1: Robustness of Standard KD (v1 vs v2). We established a strong baseline (v1) using only Mixup and CutMix. It achieved 76.12% accuracy. Adding AutoAugment in v2 provided a marginal gain of 0.07%, reaching 76.19%. This indicates that Standard KD is inherently data-efficient and stable against noise.

Phase 2: The Failure of DKD (v3 vs v3.1). Experiment v3 demonstrated a failure mode in Decoupled KD. With β=8.0 and strong augmentation, performance collapsed to 66.85%. The model triggered early stopping at epoch 84.

We confirmed the hypothesis: high reliance on "dark knowledge" (Non-Target Logits) interferes with the noise introduced by strong augmentation.

In v3.1, we reduced β to 2.0. This allowed the model to recover to 75.63%. However, it still did not match Standard KD. This demonstrates that DKD requires sensitive hyperparameter tuning, unlike Standard KD.

Phase 3: The Saturation Test (v4). In v4, we used a stronger teacher model while maintaining the v2 training recipe. The final accuracy was 76.19%, exactly matching v2.

This exact match serves as definitive proof of Capacity Saturation. The student model (EfficientNetV2-S) is fully saturated. It cannot absorb further knowledge regardless of teacher quality.

4 Key Scientific Findings

Based on the collected data, we defend three scientific claims.

4.1 Finding 1: The Regularization-Distillation Conflict

State-of-the-art distillation methods like DKD are fragile when combined with modern regularization. AutoAugment and Mixup introduce noise into the training signal. This noise corrupts the "dark knowledge" that DKD relies on. The result is training instability unless the distillation weight (β) is significantly reduced.

This finding contradicts the claim that DKD is universally superior [Zhao et al., 2022].

4.2 Finding 2: Operational Robustness of Standard KD

Contrary to recent literature suggesting Standard KD is outdated, our results indicate it is the most practically robust method. It achieved 76.19% accuracy with 99.40% teacher retention. It required zero hyperparameter retuning across all experiments. This makes it suitable for resource-constrained scenarios involving heavy data augmentation.

4.3 Finding 3: Student Capacity Ceiling

We demonstrated that the performance gap is not always due to teacher quality. When the retention rate exceeds 99% (as observed in v2 and v4), the bottleneck shifts entirely to the student's architectural capacity. Experiments v2 and v4 achieved identical 76.19% accuracy despite v4 using a stronger teacher. This aligns with findings from Mirzadeh et al. [2020] on the teacher-student gap.

5 Proposed Thesis Structure

Based on these findings, we plan to structure the thesis as follows:

5.1 Proposed Thesis Title

> "Robust Knowledge Distillation: Evaluating the Interplay Between Decoupled Objectives and Strong Data Augmentation in Compact Vision Models"

Table 3: Proposed thesis chapter structure.

Chapter	Content	Status
1. Introduction	Problem statement, research question	Outlined
2. Literature Review	EfficientNetV2, KD methods, augmentation	Outlined
3. Methodology	Enhanced training recipe, loss formulations	Outlined
4. Experiments	Results tables, training curves, analysis	**Data Ready**
5. Discussion	Robustness analysis, saturation phenomenon	Outlined
6. Conclusion	Summary, future work	Outlined

6 Visualizations

The following figures demonstrate the key findings of this research.

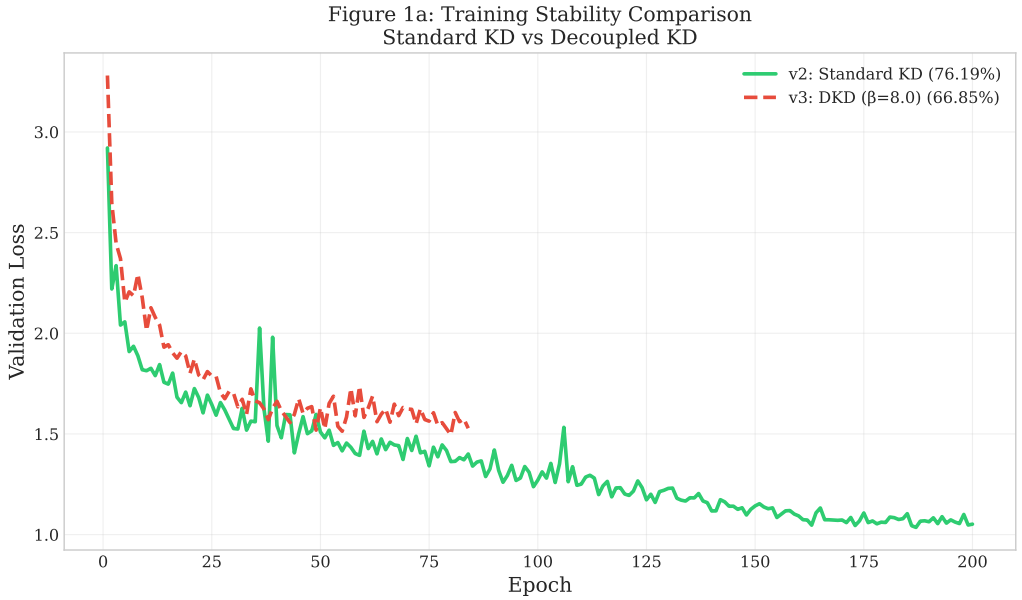

Figure 1: Training Stability Comparison: Standard KD (v2) vs Decoupled KD (v3). The DKD curve (β=8.0) shows severe instability and early stopping at epoch 84, while Standard KD converges smoothly to 76.19% accuracy.

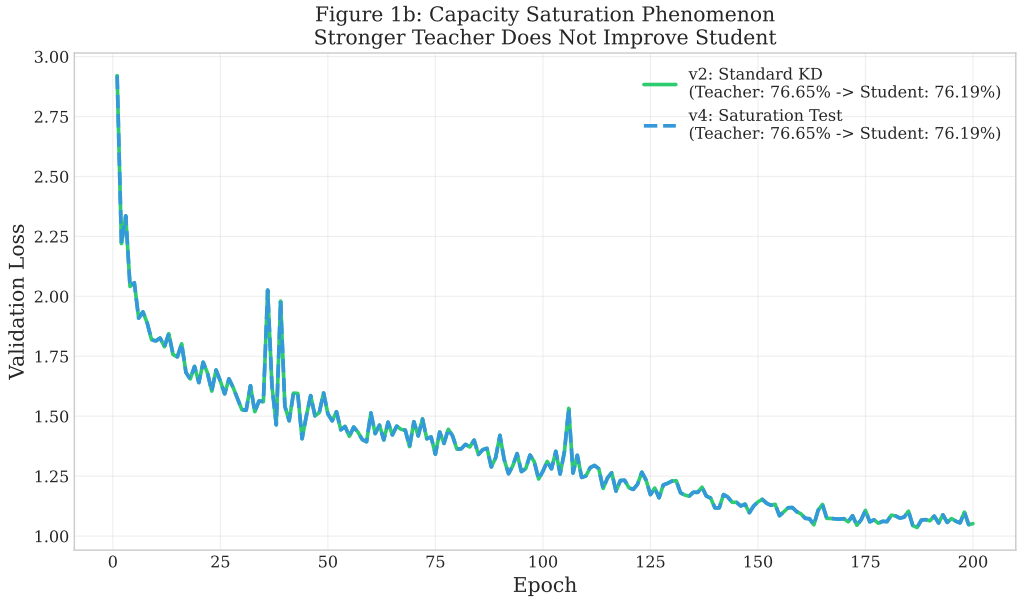

Figure 2: Capacity Saturation Phenomenon: Both v2 and v4 achieve identical 76.19% accuracy despite v4 using a stronger teacher. The overlapping curves confirm that EfficientNetV2-S has reached its representational ceiling.

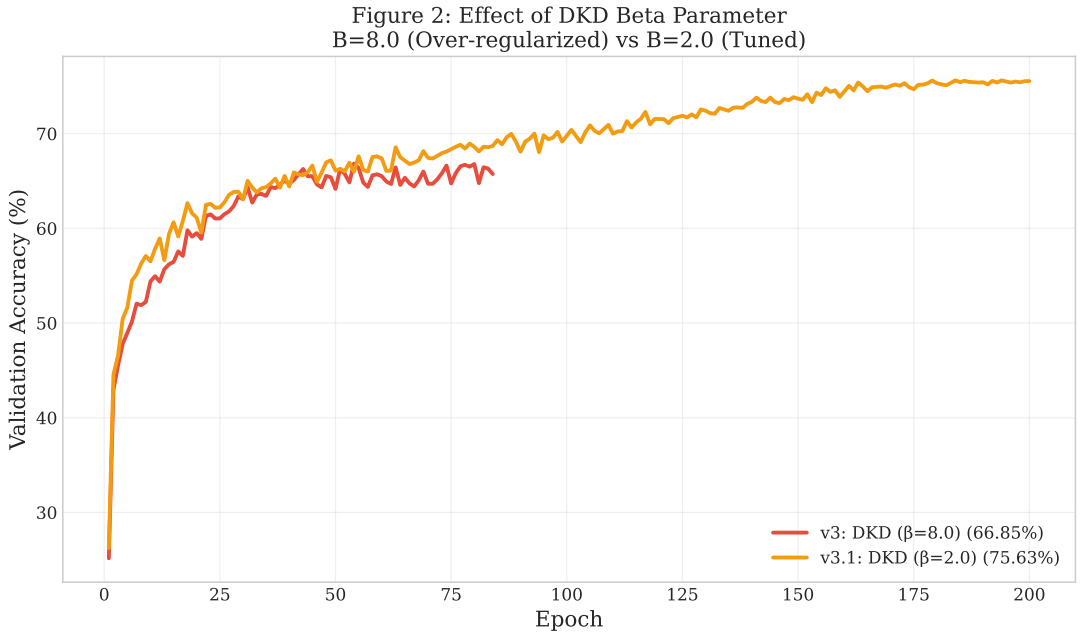

Figure 3: Effect of β on DKD Performance. With β=8.0, DKD collapses to 66.85% and triggers early stopping. Reducing β to 2.0 recovers performance to 75.63%, demonstrating the sensitivity of DKD to hyperparameter tuning.

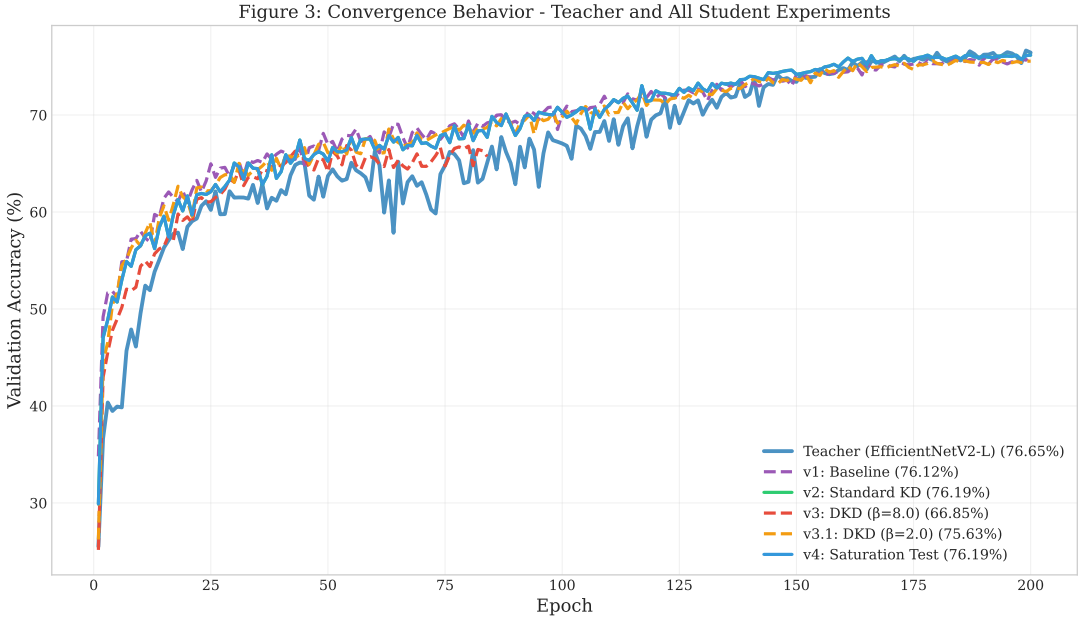

Figure 4: Convergence Behavior Across All Experiments. Standard KD (v2) achieves the highest accuracy (76.19%), while DKD with β=8.0 (v3) collapses to 66.85% due to over-regularization.

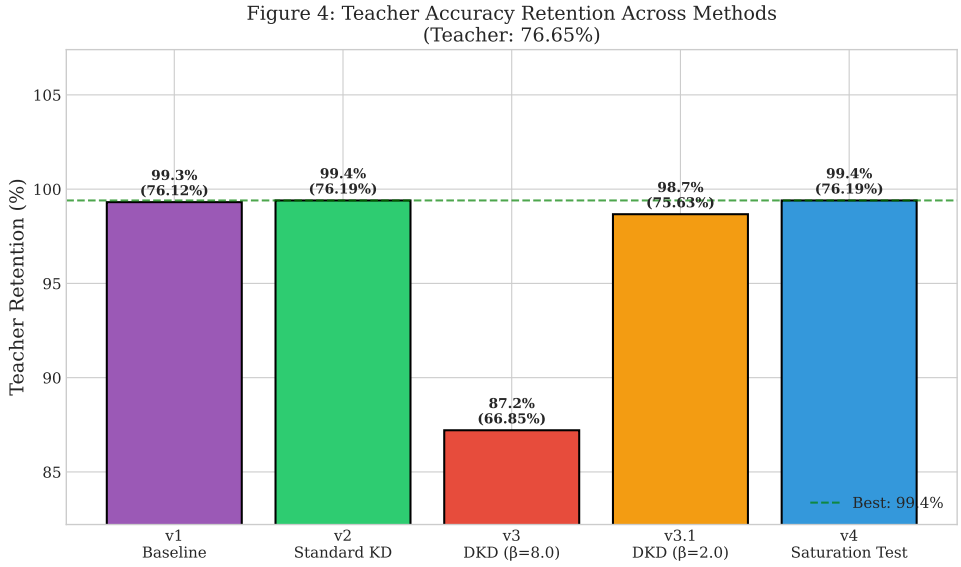

Figure 5: Teacher Accuracy Retention Across Methods. Standard KD (v2 and v4) achieves the highest retention rate (99.40%), while DKD with β=8.0 collapses to 87.21%.

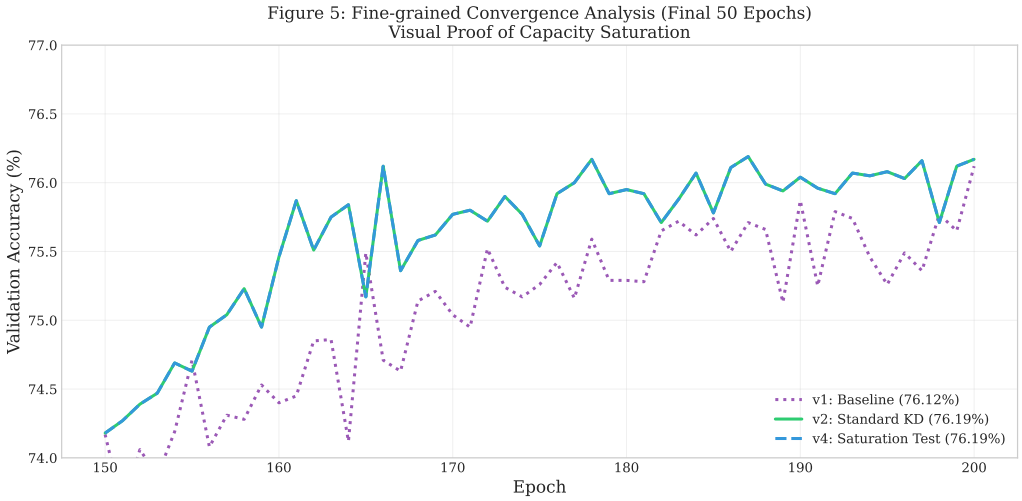

Figure 6: Zoomed Convergence Plot (Final 50 Epochs): v2 and v4 curves overlap perfectly, providing visual proof of capacity saturation. The student model cannot improve beyond 76.19% regardless of teacher quality.

7 Next Steps

With all experiments concluded, the focus now shifts to thesis writing.

1. **Write Chapter 4 (Results):** Generate high-resolution plots and formalize comparison tables.
2. **Write Chapter 5 (Discussion):** Contextualize findings within existing literature and discuss implications for Edge AI deployment.
3. **Finalize Chapter 3 (Methodology):** Complete mathematical formulations and implementation details.

8 Questions for Supervisor

1. Is the scope of comparing Standard KD vs DKD sufficient for the thesis?
2. Do you have suggestions for strengthening the saturation phenomenon argument?

References

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2019). AutoAugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123. https://arxiv.org/abs/1805.09501

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. https://ieeexplore.ieee.org/document/5206848

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint*, arXiv:1503.02531. https://arxiv.org/abs/1503.02531

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto. https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1608.03983

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1711.05101

Mirzadeh, S. I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., and Ghasemzadeh, H. (2020). Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198. https://arxiv.org/abs/1902.03393

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. https://arxiv.org/abs/1512.00567

Tan, M. and Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 10096–10106. https://arxiv.org/abs/2104.00298

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032. https://arxiv.org/abs/1905.04899

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1710.09412

Zhao, B., Cui, Q., Song, R., Qiu, Y., and Liang, J. (2022). Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11953–11962. https://arxiv.org/abs/2203.08679

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. (2020). Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008. https://arxiv.org/abs/1708.04896