Knowledge Distillation with Ensemble Calibration

Ishan Mishra
Indian Institute of Technology
Jodhpur, India
mishra.10@iitj.ac.in

Riyanshu Jain
Indian Institute of Technology
Jodhpur, India
jain.59@iitj.ac.in

Dhruv Viradiya
Indian Institute of Technology
Jodhpur, India
viradiya.1@iitj.ac.in

Divyam Patel
Indian Institute of Technology
Jodhpur, India
patel.20@iitj.ac.in

Deepak Mishra
Indian Institute of Technology
Jodhpur, India
dmishra@iitj.ac.in

ABSTRACT

Knowledge Distillation is a transfer learning and compression technique that aims to transfer hidden knowledge from a teacher model to a student model. However, this transfer often leads to poor calibration in the student model. This can be problematic for high-risk applications that require well-calibrated models to capture prediction uncertainty. To address this issue, we propose a simple and novel technique that enhances the calibration of the student network by using an ensemble of well-calibrated teacher models. We train multiple teacher models using various data-augmentation techniques such as cutout, mixup, CutMix, and AugMix and use their ensemble for knowledge distillation. We evaluate our approach on different teacher-student combinations using CIFAR-10 and CIFAR-100 datasets. Our results demonstrate that our technique improves calibration metrics (such as expected calibration and overconfidence errors) while also increasing the accuracy of the student network.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Supervised learning**; *Regularization.*

KEYWORDS

Knowledge Distillation, Ensemble, Confidence Calibration, Augmentation

ACM Reference Format:
Ishan Mishra, Riyanshu Jain, Dhruv Viradiya, Divyam Patel, and Deepak Mishra. 2023. Knowledge Distillation with Ensemble Calibration. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP '23), December 15–17, 2023, Rupnagar, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3627631.3627647

1 INTRODUCTION

Machine learning algorithms, particularly large deep neural networks (DNNs), have witnessed a surge in their adoption across various real-life applications, driving substantial advancements in several fields. However, despite their impressive performance, DNNs face a significant challenge [22][1][8][18] when it comes to deploying them in safety-critical applications such as medical diagnosis, autonomous vehicles, and astronomy. The primary concern lies in ensuring the trustworthiness of DNN predictions, as inaccuracies or uncertainties in these high-risk applications can have severe consequences. DNN fails to capture the uncertainty associated with its predictions and often give overconfident predictions [7], thereby making them unreliable and limits their deployment in high-risk domains.

To address this challenge, it is essential to improve the calibration of DNNs [8] and reduce their tendency for overconfidence. Enhancing the calibration would enable DNN predictions to reflect the actual likelihood of correctness and accurately quantify the probability of misclassification. Confidence calibration ensures that the predicted confidence aligns with the accuracy of the model and is also important for model interpretability and explainability. Various advanced techniques like augmentations, and deep ensembles are proven to improve the calibration of the DNN while also improving its generalizability. However, applying these techniques incurs a computation overhead which can be a limiting factor for their practical use. This issue is easily addressed by Knowledge Distillation (KD) [24], which aims at distilling the dark knowledge of the large DNNs (teachers) [5] into a compact and shallow model (student). KD enables the training of smaller, more efficient neural networks without compromising much on accuracy, making it feasible to deploy deep learning models on mobile devices and in other resource-constrained environments. KD [24] has demonstrated advantages in a wide range of scenarios.

In this paper, we propose a simple yet effective approach to overcome the prevalent problem of poor calibration in DNN (student) via distillation. Our approach involves leveraging a framework that utilizes a set of well-calibrated teacher models, each trained using various data augmentation techniques, as the foundation to create an ensemble-based knowledge distillation model [25]. This novel approach enables the transfer of collective knowledge of the teacher models to a single student model, resulting in significantly enhanced performance compared to the standard vanilla knowledge distillation method. The intuition behind considering multiple well-calibrated teachers is that each teacher has been exposed to different data variations, enhancing the teachers' ability to generalize well to unseen examples. However, we cannot simply apply multiple augmentation techniques to improve the generalizability of the model. Fig 1 shows the effect of applying multiple augmentations over images. It results in loss of crucial features of

ICVGIP '23, December 15–17, 2023, Rupnagar, India

ACM ISBN 979-8-4007-1625-6/23/12.
https://doi.org/10.1145/3627631.3627647

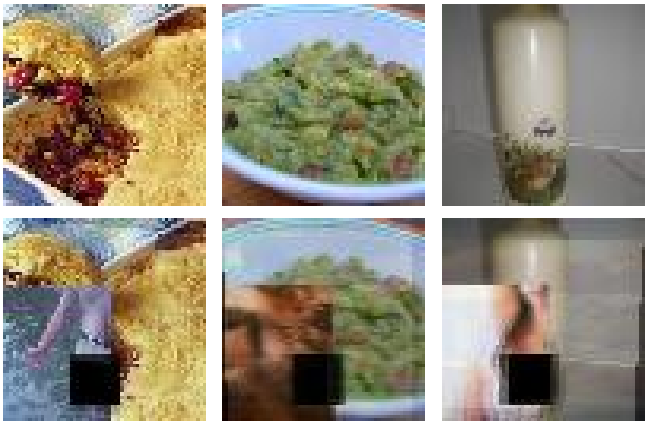


Figure 1: Progressive Augmentation Sequence Impact: In the top row, we present the unaltered original images, while the bottom row shows the outcomes after consecutive application of Mixup, CutMix, and Cutout techniques. Regrettably, the cumulative effect of these augmentations has resulted in a notable deterioration of crucial image features. Contrary to the intended regularization effect, the model's performance during training may adversely affect.

the image responsible for classification. Therefore, we utilize the generalizability of these augmentation techniques via distillation. By incorporating the collective dark knowledge of these well calibrated and generalized teachers, we ensure that the student model encompasses great generalizability and calibration.

Our approach comprehensively investigates the intricate interplay between data augmentation, knowledge distillation, and ensembling techniques, while considering crucial calibration metrics such as Expected Calibration Error (ECE) and Overconfidence Error (OE) [17] (Refer Appendix A for details). Our empirical findings substantiate the effectiveness of our proposed framework in significantly improving the calibration of student models. This breakthrough holds immense promise for the application of deep neural networks (DNNs) in safety-critical domains. Our research showcases the potential of harnessing an ensemble of well-calibrated teacher models in conjunction with knowledge distillation as a potent methodology for enhancing the calibration of student models. This pioneering work contributes to the advancement of calibration techniques and establishes a strong foundation for further exploration and adoption of DNNs in critical applications.

We investigate the hypothesis that using well-calibrated teacher models obtained through various data augmentation techniques results in improved calibration of student models through knowledge distillation. To evaluate this hypothesis, we use a framework that evaluates four data augmentation techniques, including Cutout [6], Mixup [22], CutMix [26], and AugMix [10], to create teacher networks on WideResNet and ResNet architectures. We seek to provide empirical evidence on the effectiveness of utilizing augmentation-trained teacher models for improved calibration in the context of knowledge distillation, thereby contributing to the understanding and advancement of calibration techniques in deep learning.

Our contributions are summarized as follows:

(1) We propose an ensemble-based approach to further distill the information obtained from well-calibrated teacher models into an efficient and well-calibrated student model.
(2) We demonstrate the efficiency of our proposed approach in enhancing the calibration of student models, as quantified by key metrics such as Expected Calibration Error (ECE), Overconfidence Error (OE), and Accuracy.
(3) We present an empirical study to demonstrate the effect of augmentation on single model, ensembles and distillation from a single calibrated model.
(4) We investigate the effects of weighted ensembling on the student models across various teacher-student configurations and datasets, including CIFAR-10 and CIFAR-100. Our study sheds light on the potential advantages of employing ensemble-based knowledge distillation with well-calibrated teachers to enhance the calibration of student models.

2 RELATED WORK

Data Augmentation Techniques: Deep neural networks are prone to overfitting. A common strategy to prevent overfitting is data augmentation, which seeks to actively add label-invariant modifications to training data. Data augmentation had greatly improved generalization performance. For image data, random left-right flipping and cropping are commonly used. Zhang et al. introduced a data augmentation approach named mixup [29] which linearly mixes two images with their labels combined using the same linear interpolation. In cutout [6], a selected number of randomly sized continuous sections are removed from the image to create a modified image for training. CutMix [26] is motivated by mixup and cutout, where the regions in an image are randomly cut and pasted among training images, and the ground truth labels are also mixed proportionally to the area of the regions. AugMix [10] tries to transform the input image and mix it with the original image thereby improving model robustness.

Knowledge Distillation: Knowledge distillation's primary objective is to use an effective pre-trained teacher model (or an ensemble of teacher models) [13] to direct the training of a student model. The fundamental concept of extracting knowledge from one model to another was first presented by Bucilu et al. in 2006 [2], later it was coined as knowledge distillation by Hinton et al.in 2015 [11]. [2] presented it as a fresh approach to model compression. It operates by progressively teaching a student network to replicate the actions of a larger network which is known as the teacher network.

Depending on the training process of the teacher and student, knowledge distillation is broadly classified into three distillation schemes. These schemes are: online [3][9][4], offline [11][12][28] and self-distillation [27][15][30]. Researchers have also proposed approaches to enhance the transfer of knowledge for e.g. using feature distance [19], similarity transfer [31], attention transfer [28], mutual information [23], etc. In our work, we are working with the offline distillation scheme.

Knowledge distillation with data augmentation: In the literature, various approaches have been proposed to demonstrate the effect of data augmentation on different distillation schemes

of knowledge distillation [27][20][14]. Wang et al. [24] proposed an approach to boost offline distillation by training the model on original as well as augmented inputs. In our approach, we are not using any augmentations during the distillation process. Moreover, we have analyzed the calibration effect of different augmentation techniques trained teacher models on knowledge distillation. Stanton et al. [21] have demonstrated the effectiveness of various data augmentation approaches and their effect on accuracy and calibration. However, they have not taken the effect of ensemble of teachers trained on different augmentation techniques on the student's accuracy and calibration. Instead, they have performed data augmentation while distilling the student network. Zhao et al. [31] proposed an approach that enhances the accuracy of the distilled student network by performing distillation only on the dataset generated by the augmentation techniques like Mixup and CutMix. The approach differs from Wang et al.[24] in the number of samples used during distillation. MixACM [16] improves the robustness of the distilled student network using a feature-based distillation loss. [5] have demonstrated the effect of knowledge distillation on teacher networks trained using various data augmentation schemes. They have concluded that distilling knowledge from these teachers may lead to student learning example-specific features. This leads to a loss in generalization and makes models more discriminative. Our work also validates this statement but to tackle such situation we proposed an approach to make the student learn from an ensemble of well-calibrated teachers. Our methodology is fundamentally built on top of the knowledge distillation and ensembling framework where we are targeting the better calibration in a student model by distilling the student model from ensemble of various calibrated teacher models obtained by multiple data augmentation techniques. We will describe our approach in detail in the upcoming section.

3 METHODOLOGY

Knowledge Distillation is a popular technique used to enhance the performance of shallow deep learning models by transferring the knowledge from a well-trained teacher network to a smaller student network. While using a well-calibrated teacher network trained using augmentation techniques such as Mixup, CutMix, etc., can improve the performance of the student network, it does not guarantee that the student network will also be well-calibrated [5].

One of the main challenges with combining advance augmentation techniques such as Mixup, Cutout, CutMix, and AugMix is that they use different loss functions and are applied on varying numbers of images (for example, CutMix needs atleast two images and AugMix needs only one). Moreover, applying all these augmentations sequentially on an image results in loss of distinctive features (see Figure 1), thereby affecting the performance of the teacher model trained using a combination of these augmentation techniques (Ref 4.1 for more details). To address this issue, we propose a novel approach that leverages the dark knowledge from teacher models trained using these advanced augmentation techniques to enhance the calibration of the student model without sacrificing accuracy. Our approach is simple yet effective and can be easily integrated into existing knowledge distillation frameworks. By using the dark knowledge from multiple teacher models, we effectively capture a diverse range of information about the input images. This enables the student model to learn from a wider variety of perspectives, which helps in improving its calibration. Moreover, our approach ensures that the student model retains the important and distinctive features of the input image, since the augmented input is not passed through the student model. Instead, the student model is regularized with the help of multiple teachers trained using different augmentation techniques.

In the upcoming subsection, we describe in detail the methodology adopted in this paper, ranging from the calibration and ensemble distillation techniques considered for the experiments, to the generalization measures we used to evaluate the student and teacher networks.

3.1 Calibration of the Teachers

Here, we describe the data augmentation approach we adopted to calibrate the teacher models, where each of these augmentation techniques and the corresponding loss used are discussed in detail.

Mixup: Mixup generates the augmented samples and their corresponding labels using the equation 1. It trains the model in VRM (Vicinal Risk Minimization) scenario. Let f_1 be the teacher model for mixup. In our training process, we have not used y_{mu}, instead, we have used a weighted cross entropy loss over y_1 and y_2 using the equation 2.

$$\begin{aligned} x_{mu} &= \lambda * x_1 + (1-\lambda) * x_2 \\ y_{mu} &= \lambda * y_1 + (1-\lambda) * y_2 \end{aligned} \tag{1}$$

$$\mathcal{L}_{mu} = \lambda * \mathcal{L}_{CE}(f_1(x_{mu}), y_1) + (1-\lambda) * \mathcal{L}_{CE}(f_1(x_{mu}), y_2) \tag{2}$$

where x_1 and x_2 are two randomly sampled input points, y_1 and y_2 are their associated one-hot encoded labels, f_1 is the mixup model, λ (in [0, 1]) is drawn from a $\beta(\alpha, \alpha)$ distribution and α is a hyperparameter.

Cutout: Cutout is a data augmentation that is inspired from the idea of dropout. It randomly masks patches from an image using the equation 3.

$$x_{co} = \mathrm{M} \odot x + (1 - \mathrm{M}) \odot Z \tag{3}$$

where x is the original image from the dataset, x_{co} is the cutout augmented image, Z is the zero matrix (black pixels) having same size as of input image x, M is the binary mask that denotes which pixels are to be replaced with black pixels. During the training, we used the simple cross-entropy loss using the equation 4.

$$\mathcal{L}_{co} = -f_2(x_{co}) \log\left(f_2(x_{co})\right) \tag{4}$$

where f_2 is the cutout model and x is the input image.

CutMix: CutMix is an augmentation technique inspired from Mixup and cutout that replaces random patches from an image. These replaced pixels are filled using the pixels of some other image in the dataset. This increases the number of informative pixels in the image, thereby making the model more robust and accurate. Mathematically, CutMix is described as follows:

$$\begin{aligned} x_{cm} &= \mathrm{M} \odot x_1 + (1 - \mathrm{M}) \odot x_2 \\ y_{cm} &= \lambda * y_1 + (1-\lambda) * y_2 \end{aligned} \tag{5}$$

where x_1 and x_2 are two samples from the dataset, y_1 and y_2 are the corresponding labels, **M** is the binary mask that indicates the cutout and the fill-in regions from the two randomly drawn images and $\lambda \in [0, 1]$, is drawn from a β distribution. The coordinates of

bounding boxes are $\mathbf{B} = (r_x, r_y, r_w, r_h)$ which indicates the cutout and fill-in regions in case of the images. The bounding box sampling is represented by:

$$\begin{aligned} r_x &\sim U(0, W), \quad r_w = W\sqrt{1-\lambda} \\ r_y &\sim U(0, H), \quad r_h = H\sqrt{1-\lambda} \end{aligned} \tag{6}$$

Here we have used the mixup loss as shown in equation 2.

AugMix: AugMix is a data processing technique that mixes randomly generated augmentations and improves model robustness. AugMix performs data mixing using the input image itself. It transforms (translate, shear, rotate and etc) the input image and mixes it with the original image. AugMix prevents the degradation of images while maintaining diversity as a result of mixing the results of augmentation techniques in a convex combination. It is distinguished at a high level by the combination of consistency loss and effortless augmentation operations. Here we have used the Jensen Shannon consistency loss along with the standard loss as described in [10], refer equation 7.

$$\mathcal{L} = \mathcal{L}_{CE} + \kappa * \mathcal{L}_{JS} \tag{7}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, $\mathcal{L}_{JS}$ is the Jensen Shannon loss (8) and κ is weighting hyper-parameter.

$$\mathcal{L}_{JS}(p_a; p_b; p_c) = \frac{1}{3}(\mathrm{KL}[p_a||M] + \mathrm{KL}[p_b||M] + \mathrm{KL}[p_c||M]) \tag{8}$$

where p_a, p_b, p_c are the probability distributions predicted by the model for original image x and two augmented images x_b and x_c and M is the mean of p_a, p_b and p_c.

3.2 Ensembling the teachers

We present an approach that deals with an ensemble knowledge distillation framework that improves classification performance and model generalization of small and compact networks by distilling knowledge from multiple teacher networks into a compact student network using an ensemble architecture. Each teacher is trained using a different augmentation technique and uses its own pre-defined (augmentation based) loss during the training. Our approach is shown in Figure 2.

Let (x,y) be the sample of the original dataset D, f_s be the student model and f_1, f_2, f_3, f_4 be the teacher models trained using augmentation techniques mixup, cutout, CutMix and AugMix respectively. We form an ensemble of teacher models. Each teacher model is trained with a different augmentation technique, enhancing their ability to capture diverse aspects of the data. The objective function of our approach minimizes the loss between the distribution of temperature scaled class probabilities of multiple teachers (z_1, z_2, z_3, z_4 where $z_i = \mathbb{P}(y|f_i, x)$) trained on different augmentation techniques and the student ($z_s = \mathbb{P}(y|f_s, x)$). This divergence loss helps the students to gain dark knowledge from multiple teachers. Mathematically,

$$\mathcal{L}_{div} = \sum_{i=1}^{4} \mathrm{KL}[z_s||z_i] \tag{9}$$

where $\mathcal{L}_{div}$ is the divergence loss. We also use the task-specific loss in addition to this divergence loss (equation 9). In classification tasks, the commonly used loss is cross-entropy loss defined as follows:

$$\mathcal{L}_{CE}(f_s, x, y) = -\sum ylog(f_s(x)) \tag{10}$$

The overall loss is calculated as:

$$\mathcal{L} = (1-\alpha) * \mathcal{L}_{CE} + \alpha * \mathcal{L}_{div} \tag{11}$$

Our approach is summarised in Algorithm 1.

We also experimented with a similar kind of approach where weights are assigned to the divergence loss as shown in Equation 12.

$$\mathcal{L}_{div} = \sum_{i=1}^{4} w_i \mathrm{KL}[z_s||z_i] \tag{12}$$

where w_i is the ratio of ECE and OE of the corresponding teacher model.

Algorithm 1 Knowledge Distillation with Ensemble Calibration

Require: A pre-trained multiple teacher model f_1, f_2, f_3, f_4 trained using augmentation techniques mixup, cutout, CutMix and AugMix respectively
Require: The original training Dataset $(\mathbf{X}, y) \in D$, balancing factor α
Ensure: A compact student model S trained by all teacher models
Initialization: Student model S with parameters f_s
for `i = 1, ... , Max_epoch` **do**
  Sample a batch (x, y) from the training dataset D
  Generating student logits $z_s \leftarrow f_s(x)$
  for `j=1 to 4` **do**
    $z_j^f \leftarrow f_j(x)$
  end for
  Computing divergence loss $\mathcal{L}_{KL}$
  $\mathcal{L}_{KL} \leftarrow \sum_{j=1}^{4} KL(z_j^f||z_s)$
  Computing task-specific loss $\mathcal{L}_{CE}$
  $\mathcal{L}_{CE} \leftarrow -y \log(z_s)$
  Computing total loss:
  $\mathcal{L} \leftarrow \alpha * \mathcal{L}_{KL} + (1-\alpha) * \mathcal{L}_{CE}$
end for

4 EXPERIMENTS AND RESULTS

4.1 Effect of multiple augmentations

In this section, we conduct a comprehensive analysis of the impact of multiple augmentations on the performance of two deep learning models: ResNet-32x4 and WideResNet-40-2 (WRN40-2), trained on the CIFAR-10 dataset. The objective is to explore the effectiveness of applying multiple augmentations in a sequential manner. Additionally, we evaluated the generalizability power of these models by testing them on corrupted datasets. To achieve these objectives, we focus mainly on three augmentation techniques: cutout, CutMix, and mixup. It's noteworthy that these augmentations were applied subsequent to the application of standard transformations, including normalization using mean and standard deviation, random cropping of 32, random horizontal flipping, and random rotation of up to 15 degrees. The experiment settings are kept uniform throughout this experiment by keeping a batch size of 128 and using SGD optimizer with initial learning rate of 0.05, momentum of 0.9 and

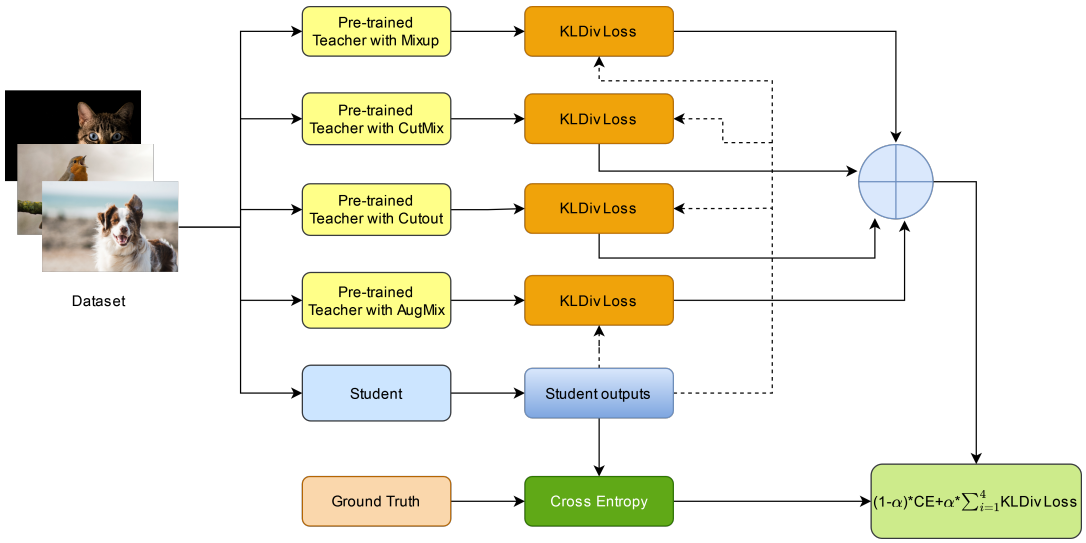


Figure 2: It illustrates our framework, which comprises one shallow student network and four pre-trained teacher networks of the same architecture. The teacher networks are trained using different augmentation techniques (Mixup, Cutout, CutMix, and AugMix) and four calibrated teacher models are produced. The KL-divergence loss is calculated between the logits of the teacher and the student. The Cross-entropy (CE) loss is calculated between student logits and ground truth labels. The final loss is the sum of all four KL-divergence losses and CE loss balanced by the factor α.

weight decay of 5×10^{-4}. The models are trained for a total of 120 epochs, the learning rate is multiplied by 0.1 at 50, 75, 100 epoch. In case of cutout, only 1 cut ($n_{holes} = 1$) is made of size 8x8, in case of mixup $\alpha = 0.3$ is used and for CutMix $\beta_{cm} = 1$ is used. If we increase the n_{holes} or hole size for cutout it will further degrade the performance as more loss of information will occur.

Sequential approach: In this approach, a sequence of augmentations is applied consecutively to a single image. The process begins with the application of mixup using a parameter value of $\alpha = 0.3$. Subsequently, cutout is employed on the resultant images, involving the creation of a single hole with dimensions 8x8. Finally, the images undergo CutMix using a parameter value of $\beta = 1$. This sequential combination will result in least loss of information. The composite loss term utilized in this method is the cumulative sum of all individual losses as defined in equation 13. Surprisingly, the outcomes indicate that rather than providing beneficial regularization effects, this approach has a detrimental impact on the model's performance (refer Table 1 and 2). The results shows that the sequential application of augmentations suffers from loss of information in images because of which the model is not able to generalise well and using more than one augmentation may not be as effective. The evaluation on the corrupted dataset (CIFAR10-C) shows that combining the augmentation techniques hurts the generalizability of the model.

$$\mathcal{L}_{total} = 1/3 \times [L_{CE}(f(x_{out}), y_A) + \mathcal{L}_m(x_{out}, y_A, y_B) + L_{cm}(x_{out}, y_A, y_C)] \quad (13)$$

where, x_{out} is the final image generated by combining all the augmentations, $f(x_{out})$ is the model's classification on the image x_{out}, y_A is the actual target, y_B is the label corresponding to the image used to create mixup augmentation and y_C is the label corresponding to the image used to create CutMix image.

Table 1: Results on the CIFAR-10 dataset. It is observed that Sequential approach degrades the performance of the model.

Model	RN32x4		WRN40-2	
Augmentation	Acc↑	ECE↓	Acc↑	ECE↓
none	94.26	0.0342	**95.18**	0.0306
mixup	95.23	0.0328	94.73	0.0429
cutout	95.54	**0.0223**	94.56	**0.0208**
CutMix	**95.82**	0.0248	94.8	0.0291
Sequential	92.5	0.3097	91.28	0.2687

Table 2: Results on the CIFAR10-C dataset. Avg.Acc and Avg. ECE are calculated over the 19 corruptions.

Model	RN32x4		WRN40-2	
Aug	Avg.Acc↑	Avg.ECE↓	Avg.Acc↑	Avg.ECE↓
none	78.35	0.1505	76.44	0.1559
mixup	**79.88**	**0.0832**	**78.02**	**0.0760**
cutout	76.34	0.1454	75.05	0.1341
CutMix	77.80	0.0887	75.15	0.0921
Sequential	72.98	0.1949	72.97	0.1676

4.2 Ensemble Distillation

To encapsulate the knowledge from various augmentation techniques, it's best to distill the model from an ensemble of teachers,

each trained on a different augmentation technique. We evaluate our ensemble approach on CIFAR-10 and CIFAR-100 image classification datasets. In our experiments, we have taken ResNet-32x4 and WideResNet-40-2 as the teachers and WideResNet-40-1, WideResNet-16-2, ResNet-8x4 as the students. All the teacher networks are trained on data augmented versions of the considered datasets with batch-size of 128. For all the experiments, SGD optimizer with weight decay = 5×10^{-4} and momentum = 0.9 is taken.

All the distillation (student) networks are trained with a initial learning rate of 0.05 (except for ShuffleNet where learning rate = 0.01), the batch size of 64, $\alpha = 0.8$ and temperature $\tau = 50$. For CIFAR-10, all the networks are trained for 240 epochs and learning rate is multiplied by 0.1 at 100, 150, 180, 210 epochs. For CIFAR-100, the number of epochs are 500 with initial learning rate 0.05, multiplied by 0.1 at 150, 180, 210 epoch. During distillation, standard augmentation is applied to the input images which includes random crop, horizontal flip, random rotation, and normalization. **It is crucial to emphasize that the experimental configuration utilized for the pre-training of the teacher and the subsequent distillation of the student differs from the experimental approach employed to analyze various augmentation techniques.**

We consider baselines as i.) student trained without distillation, ii.) student trained using vanilla knowledge distillation, iii.) student distilled using teachers that are pre-trained using mixup, CutMix, cutout and AugMix. For detailed training of the teachers, refer to Table 3 and 4. We have considered ECE (Expected Calibration Error), OE (Overconfidence Error) [17], and accuracy as metrics to evaluate the performance of the model (Refer Appendix A).

Table 3: CIFAR-10: Teacher Models.

Eval Metrics	No Aug	CutMix	Mixup	AugMix	Cutout
ResNet-32x4					
Accuracy	95.77	96.76	96.2	95.9	96.4
ECE	0.0265	0.0165	0.0393	0.02	0.0183
OE	0.0224	0.0058	0.0048	0.0156	0.0144
WideResNet-40-2					
Accuracy	95.22	95.99	95.33	95.3	95.59
ECE	0.0299	0.0288	0.0496	0.0185	0.0202
OE	0.026	0.0062	0.0033	0.0135	0.0164

Table 4: CIFAR-100: Teacher Models.

Eval Metrics	No Augm	CutMix	Mixup	AugMix	Cutout
ResNet-32x4					
Accuracy	77.14	80.49	79.71	77.82	79.13
ECE	0.0864	0.0277	0.0285	0.0648	0.0755
OE	0.0697	0.0150	0.0024	0.0481	0.0580
WideResNet-40-2					
Accuracy	75.15	78.14	76.48	75.98	76.92
ECE	0.1149	0.0331	0.0562	0.0588	0.0819
OE	0.0924	0.0179	0.0013	0.0410	0.0625

4.3 CIFAR-10

CIFAR-10 dataset consists of 60, 000 RGB images of size 32×32 in 10 different classes. We have reported the ECE and OE corresponding to the epoch having the best accuracy on the test dataset for all the configurations. Table 5 shows the result for CIFAR-10 dataset. We observe that our approach outperforms the baselines both in terms of ECE and accuracy for the ResNet-32x4/ResNet-8x4 combination. Also, the OE is reduced in our approach as compared to the student. However, the student distilled from a teacher trained on mixup augmented data has the best OE. This is due to the fact that mixup augmentation mainly focuses on minimizing the overconfidence of the model. For the WRN-40-2/WRN-16-2 and WRN-40-2/WRN-40-1, the accuracy of our approach is close to KD+cutout baseline with an improvement over the ECE and OE. Our weighted approach has a better ECE and OE as compared to our ensemble approach.

4.4 CIFAR-100

CIFAR-100 dataset is more challenging dataset as compared to CIFAR-10 dataset having 100 classes with 600 RGB images per class of dimension 32×32. The results are reported in Table 6. We observe that our approach surpasses all the baselines for ResNet-32x4/ResNet-8x4. The reliability plots in Figure 3 illustrate the calibration of the baselines along with our approach. Theoretically, a well-calibrated model has most of the density lying on the $y = x$ line, overconfidence is a situation where the density lies below the $y = x$ line, and underconfidence in the model is shown when density lies above the $y = x$ line. The reliability plots corresponding to our approach have most of the density along $y = x$ line that defines a well-calibrated model whereas all other techniques are either overconfident or underconfident. We see a major improvement in the accuracy along with better calibration for our ensemble approach. The OE of our weighted approach is close to the KD+mixup baseline with an improved ECE and accuracy. For WRN-40-2/WRN-16-2, we see a similar trend where the student model distilled using our ensemble approach outperforms the baselines in terms of accuracy and ECE. In case of WRN-40-2/WRN-40-1 configuration, our ensemble approach has the best accuracy among the baselines with a comparable ECE.

5 ABLATION STUDY

To delve deeper into the effect of ensemble on calibration, we ensembled calibrated models using various weight combinations in evaluation mode. The results are reported in Table 7.

The values 1,1,1,1,1 in the Table 7 shows the weightage given to the respective augmentation techniques while adding the softmax scores. Let p_1, p_2, p_3, p_4, p_5 be the softmax scores from the 5 different models, then final softmax score used during the evaluation is defined as :

$$\frac{w_1p_1 + w_2p_2 + w_3p_3 + w_4p_4 + w_5p_5}{w_1 + w_2 + w_3 + w_4 + w_5} \tag{14}$$

This resultant softmax value is used to calculate accuracy, ECE and OE.

We explored the accuracy-ECE correlation in combining pre-trained calibrated models, finding that while aggregation improves

Table 5: Experimental results for CIFAR-10 dataset.

Teacher	ResNet-32x4			WideResNet-40-2			WideResNet-40-2		
Student	ResNet-8x4			WideResNet-16-2			WideResNet-40-1		
Approach	Acc ↑	ECE ↓	OE ↓	Acc ↑	ECE ↓	OE ↓	Acc ↑	ECE ↓	OE ↓
Student	92.47	0.0323	0.0258	93.39	0.029	0.0239	93.66	0.0325	0.0268
Vanilla KD	93.34	0.0299	0.0244	94.25	0.0399	0.0345	94.2	0.045	0.0415
KD + Mixup	92.81	0.0377	**0.0002**	93.6	0.0303	**0.0018**	93.94	0.0309	**0.01**
KD + Cutout	93.08	0.0332	0.0266	**94.71**	0.03	0.0262	94.46	0.0383	0.0337
KD + CutMix	93.05	0.017	0.0044	94.16	0.0239	0.0193	93.7	**0.0278**	0.0226
KD + AugMix	93.35	0.023	0.0167	94.08	0.0343	0.0292	93.88	0.0349	0.0298
Ours (Ensemble)	93.65	0.0238	0.0164	94.61	0.0249	0.0198	**94.48**	0.0339	0.0294
Ours (Weighted)	**94.09**	**0.0174**	0.0093	94.52	**0.0219**	0.0138	94.09	0.0326	0.0266

Table 6: Experimental results for CIFAR-100 dataset.

Teacher	ResNet-32x4			WideResNet-40-2			WideResNet-40-2		
Student	ResNet-8x4			WideResNet-16-2			WideResNet-40-1		
Approach	Acc ↑	ECE ↓	OE ↓	Acc ↑	ECE ↓	OE ↓	Acc ↑	ECE ↓	OE ↓
Student	71.64	0.0772	0.056	72.48	0.087	0.0638	70.74	0.0834	0.0609
Vanilla KD	72.42	0.0659	0.0486	74.07	0.087	0.0672	73.24	0.1043	0.0826
KD + Mixup	71.91	0.084	**0.0055**	74.34	0.0554	0.0057	72.33	**0.0338**	0.0052
KD + Cutout	72.32	0.0665	0.0482	74.16	0.0638	0.047	73.53	0.0783	0.0573
KD + CutMix	71.51	0.068	0.0109	74.62	0.0541	0.014	71.12	0.0341	0.0049
KD + AugMix	72.97	0.0523	0.0346	74.28	0.0394	0.0246	73.11	0.0462	0.0301
Ours (Ensemble)	73.82	**0.0497**	0.0247	**75.24**	**0.0351**	0.0183	**74.34**	0.0355	0.0198
Ours (Weighted)	**74.21**	0.0648	0.0079	74.99	0.0623	**0.0007**	73.22	0.0637	**0.0004**

Table 7: Trade-off between ECE and Accuracy with different assigned weights for pre-trained models (WideResNet-40-2).

Assigned Weights					Acc	ECE	OE
No Aug	Mixup	CutMix	Cutout	AugMix			
1	1	1	1	1	81.75	0.0632	9.4×10^{-5}
3	1	1	3	1	80.71	0.0361	0.0015
2	1	1	6	1	79.8	0.0281	0.0048

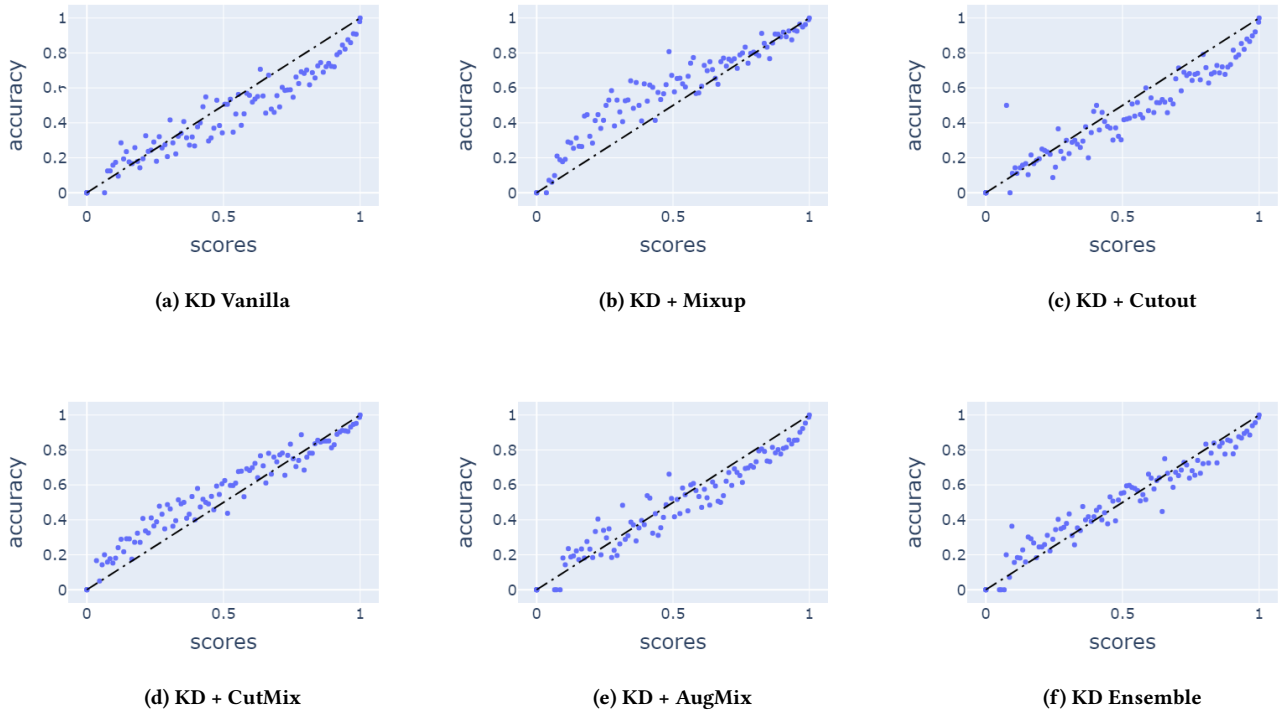
accuracy, it reduces ECE. To address this, optimal model combinations with specific weights are essential for ECE reduction (Table 7). Simple combination doesn't guarantee calibration; instead, weighted softmax averaging proves more effective. Notably, the weights lack a discernible pattern, precluding generalization.

We also perform experiments with different techniques to train the ensemble of all models with KD, which includes softmax averaging (taking average of softmax scores produced by the teachers and then evaluating loss with the student softmax scores), loss averaging (taking average of KL loss between student and different teachers), and taking weights (by default 1) while adding the loss. The results in Table 8 highlight that "Add loss" yields better accuracy and "Weighted add loss(ECE/OE)" yeilds better ECE and OE. The reasoning behind improvement of ECE and OE is as follows

Usually, DNNs are overconfident in nature and hence their ECE is also on a higher side. Now, applying Data Augmentation on such models controls their ECE and may decrease the OE also which means a model with better calibration but underconfident. If we distill a student from four models with controlled ECE, then ECE will have less effect on the ratio of ECE and OE but a model with a higher OE value will get less weightage and a model with low

Table 8: Experiments on different KD ensemble Techniques on CIFAR-100 with WideResNet-40-2 as teacher model, ShuffleNet V1 as student model(Refer Appendix B).

Technique	Best Acc	ECE	OE
Avg softmax	75.55	0.0707	0.0522
Avg loss	75.71	0.0752	0.0563
Add loss	**77.28**	0.061	0.046
Weighted Add loss (ECE/OE)	76.97	**0.0379**	**0.0224**
Vanilla KD	75.41	0.12	0.0992



(a) KD Vanilla **(b) KD + Mixup** **(c) KD + Cutout**

(d) KD + CutMix **(e) KD + AugMix** **(f) KD Ensemble**

Figure 3: This illustrates the reliability plots between accuracy and confidence of different augmentation techniques and KD ensemble using ResNet-32x4 as teacher model and ResNet-8x4 as student model on CIFAR-100 dataset.

OE value will get higher weightage because of the inverse relation. Therefore, the ratio of ECE and OE helps in distillation as the model learns more from a model with low OE instead of a model with high OE.

6 DISCUSSION

Our experimental findings shows the intricate interplay between augmentation techniques and the performance of deep learning models. Notably, advanced augmentation methods such as cutout, mixup, CutMix, and AugMix have consistently shown their capacity to yield well-calibrated and reliable models. However, our investigation into the fusion of these augmentation techniques has yielded unexpected outcomes. Surprisingly, the anticipated regularization benefits conferred by these augmentations appear to be outweighed by a counteracting factor: the loss of vital information (see Figure 1). This loss of information manifests as a degradation in model performance. Despite the individual strengths of these augmentation methods, their combined application seems to compromise the model's ability to retain crucial discriminative features, resulting in a noteworthy decline in performance. This intriguing finding underscores the delicate balance between regularization and feature preservation when utilizing multiple augmentation strategies simultaneously. Moreover, when evaluated on corrupted dataset, the results are not satisfactory. In light of these challenges, our study introduces the concept of knowledge distillation as a means to effectively harness the potential of augmentation techniques. Our experiments demonstrate that knowledge distillation, wherein a student model learns from an ensemble of well-calibrated teacher models, can be a powerful mechanism to combine the dark knowledge embedded in diverse augmentation strategies. By distilling this knowledge, the student model achieves enhanced generalizability as shown by the reliability plots (Fig 3).

7 CONCLUSION AND FUTURE WORK

In this paper, we introduce an approach for distilling the dark knowledge from an ensemble of teachers trained using various augmented data. Our experiments on CIFAR-10 and CIFAR-100 demonstrate that the student trained using our approach outperforms the student trained from scratch in terms of accuracy and calibration. We concluded that our approach outperforms the vanilla KD, and all other KD + individual data augmentation techniques in terms of ECE and Accuracy. It lowers down the ECE and increases the accuracy as a result we get a more accurate and calibrated student model. We also introduced a weighted ensemble version of our approach by taking the ratio of ECE and OE. We argue that this weighing scheme might be explored in the future by also taking the accuracy of various teachers into account.

ACKNOWLEDGMENTS

The authors would like to thank the Ministry of Electronics and Information Technology (MeitY), Government of India, for supporting the work.

REFERENCES

[1] Murat Seckin Ayhan and Philipp Berens. 2018. Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks. In *Medical Imaging with Deep Learning*. https://openreview.net/forum?id=rJZz-knjz

[2] Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model Compression. (2006), 535–541. https://doi.org/10.1145/1150402.1150464

[3] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. 2020. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3430–3437.

[4] Inseop Chung, Seonguk Park, Jangho Kim, and Nojun Kwak. 2020. Feature-map-level Online Adversarial Knowledge Distillation. 119 (13–18 Jul 2020), 2006–2015. https://proceedings.mlr.press/v119/chung20a.html

[5] Deepan Das, Haley Massa, Abhimanyu Kulkarni, and Theodoros Rekatsinas. 2020. An empirical analysis of the impact of data augmentation on knowledge distillation. *arXiv preprint arXiv:2006.03810* (2020).

[6] Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017).

[7] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2021. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342* (2021).

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[9] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. 2020. Online Knowledge Distillation via Collaborative Learning. (June 2020).

[10] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)* (2020).

[11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).

[12] Zehao Huang and Naiyan Wang. 2019. Like What You Like: Knowledge Distill via Neuron Selectivity Transfer. https://openreview.net/forum?id=rJf0BjAqYX

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

[14] Xingjian Li, Haoyi Xiong, Chengzhong Xu, and Dejing Dou. 2021. SMILE: Self-Distilled MIxup for Efficient Transfer LEarning. *arXiv preprint arXiv:2103.13941* (2021).

[15] Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. 2020. Self-Distillation Amplifies Regularization in Hilbert Space. 33 (2020), 3351–3361. https://proceedings.neurips.cc/paper/2020/file/2288f691b58edecadcc9a8691762b4fd-Paper.pdf

[16] Awais Muhammad, Fengwei Zhou, Chuanlong Xie, Jiawei Li, Sung-Ho Bae, and Zhenguo Li. 2021. MixACM: Mixup-Based Robustness Transfer via Distillation of Activated Channel Maps. *Advances in Neural Information Processing Systems* 34 (2021), 4555–4569.

[17] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring Calibration in Deep Learning.. In *CVPR Workshops*, Vol. 2.

[18] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf

[19] Adriana Romero, Samira Ebrahimi Kahou, Polytechnique Montréal, Y. Bengio, Université De Montréal, Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *in International Conference on Learning Representations (ICLR*.

[20] Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, and Yandong Guo. 2022. Online Distillation with Mixed Sample Augmentation. *arXiv preprint arXiv:2206.12370* (2022).

[21] Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. 2021. Does knowledge distillation really work? *Advances in Neural Information Processing Systems* 34 (2021), 6906–6919.

[22] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems* 32 (2019).

[23] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*.

[24] Huan Wang, Suhas Lohit, Michael Jones, and Yun Fu. 2020. Knowledge distillation thrives on data augmentation. *arXiv preprint arXiv:2012.02909* (2020).

[25] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1285–1294.

[26] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6023–6032.

[27] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. 2020. Regularizing Class-Wise Predictions via Self-Knowledge Distillation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Sergey Zagoruyko and Nikos Komodakis. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *ICLR*. https://arxiv.org/abs/1612.03928

[29] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1Ddp1-Rb

[30] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3713–3722.

[31] Haoran Zhao, Kun Gong, Xin Sun, Junyu Dong, and Hui Yu. 2021. Similarity Transfer for Knowledge Distillation. *arXiv preprint arXiv:2103.10047* (2021).

A EVALUATION METRICS

For the evaluation purpose, we consider ECE (Expected Calibration Error), OE (Overconfidence Error) [17], and Accuracy to evaluate the model's performance on the test dataset. Let X_n be the set of samples whose prediction scores (the winning softmax score) fall into bin n. The accuracy and confidence of X_n are calculated as: Our test sample is first divided into n number of bins. Then, based on each sample's probability, we place it into one of the bins. Finally, we calculate the accuracy of the bin as the number of correct samples contained in the bin, and the confidence of the bins displays the average probability of the samples present in the bin.

$$\text{acc}(X_n) = \frac{1}{|X_n|} \sum_{i \in X_n} 1(\hat{y}_i = y_i) \tag{15}$$

where $\hat{y}_i$ is the predicted label and y_i is the actual label

$$\text{conf}(X_n) = \frac{1}{|X_n|} \sum_{i \in X_n} \hat{p}_i \tag{16}$$

where $\hat{p}_i$ is the prediction score of sample i.

The **Expected Calibration Error** (ECE) is calculated as:

$$\text{ECE} = \sum_{i=1}^{n} \frac{|X_i|}{N} \big|\text{acc}(X_i) - \text{conf}(X_i)\big| \tag{17}$$

We also calculate an additional calibration metric – the **Overconfidence Error** (OE) – as follows

$$\text{OE} = \sum_{i=1}^{n} \frac{|X_i|}{N} \big|\text{conf}(X_i) \, \text{x} \max(\text{conf}(X_i) - \text{acc}(X_i), 0)\big| \tag{18}$$

where, n is the total number of bins and N is total length of the test dataset.

B STUDENT ANALYSIS

In this section, we observe the impact of various advance augmentation techniques like cutout, mixup, cutmix, augmix, if applied directly on the student model. The results also demonstrate that the regularization effect cause by these augmentation techniques is dependent on the model's architecture. The results are poor when compared with our knowledge distillation based approach.

Table 9: It shows the effect of the data augmentation techniques when applied directly on the student model. We have kept all the hyper-parameters same as KD while training the model. The result corresponds to ShuffleNet V1 model.

Technique	Accuracy	ECE	OE
No augmentation	71.45	0.0890	0.0646
Mixup	71.25	0.0454	0.0016
CutMix	72	0.0302	0.0092
Cutout	67.32	0.13	0.10
AugMix	68.47	0.07	0.047

We use batch size as 128 along with an initial learning rate of 0.01 to train the Shufflenet V1 student model and obtained the result shown in Table 9. The model was trained for 500 epochs and the learning rate is multiplied by 0.1 at 150, 180, 210 epochs. For all the experiments shown in Table 9, SGD optimizer with weight decay = 5×10^{-4} and momentum = 0.9 is taken.