OOD Generalization using Confidence Aware Multi-Teacher Knowledge Distillation

Abdullah Hashmat, Raafay Saeed Kazmi, Zammad Bin Ziyad Khan
Lahore University of Management Sciences
(25100148, 25100197, 25100124) @ lums.edu.pk
GitHub Repository Link

December 27, 2024

Abstract

Our project explores the challenge of OOD generalization in student models for two dominant cues -*texture and shape*- which often limit generalization in cue-conflict scenarios or under distributional shifts during image classification. We propose a multi-teacher single-student knowledge distillation framework where each teacher model is explicitly biased towards a specific visual cue. By combining the strengths of diverse architectures, such as Convolutional Neural Networks (CNNs) for texture, Vision Transformers (ViTs) for shape, and other models for additional biases, the student model learns to integrate and balance these biases. The framework incorporates weighted ensembling of multiple teacher models and intermediate level KD, enabling effective knowledge transfer. The goal is to train a compact and efficient student model that outperforms its individual teachers on cue-conflict datasets, demonstrating reduced bias. This work aims to provide new insights into reconciling architectural diversity and bias-aware training in Machine Learning.

1 Introduction

"The more you know, the more you know you don't know."

This quote from Aristotle perfectly encapsulates the essence of what we're tackling today. We strive to build complex models that learn from massive datasets, absorbing layers of knowledge. But as these models get smarter, they reveal gaps and unexpected behaviors when faced with unfamiliar data. This is where we realize that the more we train, the more we uncover how much there is left to learn, especially when tackling real-world challenges like out-of-distribution (OOD) generalization.

This can be accounted to the fact that these models often exhibit strong inductive biases towards visual cues such as color, texture, or shape. While these biases are integral for Machine Learning tasks and improve performance on certain datasets, they can limit generalization when data distributions shift or cues conflict. To address these challenges, we explore the potential of a **multi-teacher Knowledge Distillation** framework. Knowledge Distillaion (KD) has emerged as a powerful paradigm for transferring knowledge from one or more high-capacity teacher models to a compact student model, enabling the student to learn distilled insights from the teachers' expertise. Using multiple teacher models, each with distinct inductive biases, the student can synthesize and generalize

knowledge across diverse perspectives. This integration of diverse biases equips the student model to handle OOD scenarios more effectively, bridging the gap between theoretical advances in KD and practical demands for robustness in real-world applications.

Our project stands at the crossroads of two critical challenges in Machine Learning: OOD generalization and model compression. While the former focuses on enabling models to perform reliably under shifting data distributions, the latter seeks to distill the knowledge of large, complex models into smaller, more efficient ones without sacrificing performance. By combining these domains, our work explores how KD can serve as a bridge - integrating diverse insights from teacher models to train compact student models that are not only efficient but also resilient to distributional shifts.

2 Related Work

The approach of enabling the transfer of knowledge from a large, complex teacher model to a smaller, more efficient student model was initially formalized by Hinton [11], who proposed matching the softened logits of a teacher to the outputs of a student model, thus allowing the student to mimic the learned representations of the teacher. Multiple efforts have aimed to make student models more generalizable and better performing than their teacher counterparts. For example, self-distillation [6] allows a model to iteratively teach itself, enhancing generalization even without a larger teacher. Progressive distillation [18] refines knowledge representations during training, allowing students to surpass their teachers. Additionally, auxiliary supervision [19], such as data augmentation and feature alignment, further boost the performance and robustness of the student model.

An interesting area of emerging research is the use of **multiple teachers** to distill specialized knowledge to the student. This paper [17] demonstrated the potential of aggregating knowledge from multiple teachers through an ensemble of logits, thereby capturing diverse feature representations. Such multi-teacher approaches are particularly valuable when individual teachers encode distinct inductive biases. For example, studies [2] employing Convolutional Neural Networks (CNNs) as teachers emphasize spatially localized features, while transformers prioritize global contextual understanding. This diversity in learned features creates a compelling case for multi-teacher KD to build a more generalized and robust student model. Similarly, [14] applies multiteacher learning to multitask learning where each teacher corresponds to a particular task.

Recent works include KD from specialized biased teachers to eliminate racial biases in the student model [4], by training teachers on a subset of biased data. However, the limitation lies in the homogeneous nature of teacher models which are used, hence model's inherent systematic biases are bound to be propagated to the student model as well. The lack of diversity among teacher models restricts the effectiveness of the distillation process in achieving a relatively unbiased and equitable student model. This paper [16] pre-trains multiple language models on a specialized NLP task and distills the information to a more generalized student model, however here as well the teacher and student model architectures are homogeneous and are limited to the transformer models only for NLP related tasks. [1] explores transferring specialized biases from teachers to students using various model architectures. However, it focuses on architectural inductive biases and does not address a multi-teacher setting. This limits its ability to explore how diverse teacher biases can be aggregated to create a more generalized student model.

Addressing these gaps, our work seeks to refine the multi-teacher KD paradigm by explicitly

leveraging the unique inductive biases of each teacher model while ensuring that these biases complement rather than conflict. By incorporating a dynamic ensemble weighting mechanism and distillation through logit matching [11] on a cue-conflict dataset [8], we aim to produce a student model that not only excels on IID data but also demonstrates better performance on OOD variations.

3 Methodology

Since, our work primarily focuses on addressing the challenge of OOD generalization during KD, we target two predominant cues in images: **texture and shape**. To tackle the issue of student model exhibiting strong biases of teacher models during KD, hindering its ability to generalize effectively in scenarios involving cue conflicts or distributional shifts, we employ **VGG16**, a CNN model with a known bias toward texture [10], and **ViT b16**, a vision transformer, which exhibits a bias toward shape [13], as teacher models.

In our framework, the student model learns to balance and integrate these distinct biases through **confidence-aware weighted logit matching**. The weighting is then dynamically determined based on the entropy levels of the teacher logits [12], enabling the student to adaptively prioritize information from the most confident teacher for each sample. This strategy ensures that the student model develops generalized representations that are less dependent on any single cue, resulting in improved OOD generalization.

3.1 Our Framework

In this section, we provide the motivation of learning teacher-dependent importance weight and an overview of the experimental framework.

3.1.1 Motivation and Overview

Imagine a classroom where one teacher is an expert in identifying textures, like the complex details of a fabric or the roughness of an object's surface. This teacher excels at recognizing local patterns and fine grained details but struggles to understand the overall structure or shape of objects. Meanwhile, another teacher specializes in recognizing shapes and is good at identifying the global outline of objects like the silhouette of a bird or the shape of a leaf. However, this teacher is less proficient at distinguishing fine textures and surface details between objects.

Now, consider training a student under the guidance of these two teachers. The goal is not to inherit the limitations of either teacher but instead to combine their strengths, so the student can be good at both texture recognition and shape identification. The student should perform well in scenarios where textures are ambiguous, shapes are distorted, or even when the data is OOD. On the whole, we aim to create a student who can efficiently handle both texture and shape based challenges, surpassing the capabilities of either teacher alone.

Our goal is to overcome these limitations by leveraging the complementary strengths of shape and texture biased models through a novel knowledge distillation framework. We use two teacher **models—ViT (shape)** and **CNN (texture)**—to transfer their biases to a student model via weighted, confidence-aware logit matching. Unlike traditional distillation, where the student inherits the biases of the teachers, our approach balances these biases, creating a more robust student capable of generalizing to **OOD data** and handling **cue-conflict** scenarios.

3.1.2 Confidence-Aware Weighting and Knowledge Distillation

Here we present our approach that combines entropy based confidence-aware weighting and KD to train an OOD generalizable student model.

Logit Matching Loss (KD Loss)

In Knowledge Distillation (KD), the goal is to transfer knowledge from a teacher model to a student model by aligning their logits (the raw outputs before applying the softmax). We modify this approach by using multiple teacher models and ensembled logits.

The logit matching loss between the student model's logits $\mathbf{S}$ and the ensembled teacher logits $\mathbf{T}$ is computed using the Kullback-Leibler (KL) divergence [11] between the temperature-scaled softmax outputs:

$$L_{\text{distill}} = \text{KL}\left(\text{softmax}\left(\frac{\mathbf{S}}{T_s}\right), \text{softmax}\left(\frac{\mathbf{T}}{T_t}\right)\right) \times T_s^2$$

Where:

- $\mathbf{S}$ is the student model's logits.
- $\mathbf{T}$ is the ensembled logits from the teacher models.
- T_s is the temperature factor for the student model.
- T_t is the temperature factor for the teacher models.
- The KL divergence measures the difference between the two probability distributions.
- T_s^2 scaling factor is used to adjust for the softened logits.

This loss encourages the student model to match the softmax outputs of the ensembled teacher models, with temperature scaling making the knowledge transfer smoother.

Confidence-Aware Weighted Logit Ensemble

We use an ensemble of multiple teacher models to create the final teacher logits for distillation. Instead of simply averaging the logits, we weight each teacher's contribution based on their confidence level. Confidence is derived from the entropy of the teacher's output distribution.

1. Softmax: For each teacher model i, we compute the softmax probabilities $\mathbf{P}_i$ of the teacher's logits $\mathbf{T}_i$:

$$\mathbf{P}_i = \text{softmax}(\mathbf{T}_i)$$

2. Entropy: For each teacher's prediction $\mathbf{P}_i$, we compute the entropy $H(\mathbf{P}_i)$, which measures the uncertainty in the teacher's prediction, where ϵ is a small value to avoid division by zero or log of zero:

$$H(\mathbf{P}_i) = -\sum_j \mathbf{P}_i(j) \cdot \log(\mathbf{P}_i(j) + \epsilon)$$

3. Confidence Calculation: The confidence for each teacher is the inverse of their entropy. A teacher with lower entropy ie. more confident predictions will have a higher confidence score:

$$\text{Confidence}_i = \frac{1}{H(\mathbf{P}_i) + \epsilon}$$

4. Normalization of Confidence Scores: The confidence scores are normalized across all teacher models to ensure they sum up to 1:

$$\mathbf{w}_i = \frac{\text{Confidence}_i}{\sum_{i=1}^{n} \text{Confidence}_i}$$

5. Ensemble Logits Calculation: The final ensembled teacher logits $\mathbf{T}_{\text{ensemble}}$ are calculated as a weighted sum of the logits from all teacher models:

$$\mathbf{T}_{\text{ensemble}} = \sum_{i=1}^{n} \mathbf{w}_i \cdot \mathbf{T}_i$$

where $\mathbf{w}_i$ is the normalized weight based on each teacher's confidence, and $\mathbf{T}_i$ is the logits from teacher i.

This approach ensures that the ensembled logits give more weight to teachers that are more confident in their predictions. It contrasts with other methods where confident teachers might be given lower weight to reduce bias, but we instead emphasize the strong predictions from confident teachers in order to transfer their biases to the student.

Total Loss Calculation

The total loss function combines the Cross-Entropy (CE) loss and the logit matching distillation loss.

a) Cross-Entropy Loss: This loss measures how well the student model's predictions match the ground-truth labels. It is defined as:

$$L_{\text{CE}} = -\sum_{i} y_i \log(p_i)$$

Where:

- y_i is the ground-truth label, and
- p_i is the predicted probability for class i.

b) Logit Matching Loss: This is the distillation loss described earlier.

The **total loss** is a weighted sum of the Cross-Entropy loss and the distillation loss:

$$\boldsymbol{L}_{\text{total}} = (1 - \alpha) \cdot \boldsymbol{L}_{\text{CE}} + \alpha \cdot \boldsymbol{L}_{\text{distill}}$$

Where:

- α is a hyperparameter that controls the balance between the classification loss and the distillation loss. By adjusting α, we control how much emphasis is placed on learning from the teacher models versus the ground-truth labels.

This dynamic weighting allows the student model to learn from both the teacher models' knowledge and the ground-truth labels, ensuring that the student generalizes well to new data.

Optimization and Training

During training, the optimizer (AdamW) is used to minimize the total loss. The optimizer's update rule is:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \cdot \nabla_{\theta} L_{\text{total}}$$

Where:

- θ_{old} and θ_{new} are the parameters of the student model before and after the update,
- η is the learning rate, and
- $\nabla_\theta L_{\text{total}}$ is the gradient of the total loss with respect to the student model parameters.

The learning rate is adjusted during training using cosine annealing:

$$\eta_{\text{new}} = \eta_{\max} \cdot \frac{2}{1 + \cos\left(\frac{t}{T_{\max}}\pi\right)}$$

Where:

- $\eta_{\max}$ is the maximum learning rate, and
- $T_{\max}$ is the total number of epochs.

4 Experimental Setup

This section outlines the experimental setup used to evaluate our multi-teacher knowledge distillation framework.

4.1 Datasets

To comprehensively evaluate the performance of our multi-teacher knowledge distillation framework, we utilized the following datasets:

- **ImageNet-1k [5]:** We used ImageNet-1k for its diverse set of textures and shapes to establish a strong baseline for pretraining and fine-tuning. Additionally, pretrained models on ImageNet eminilated need to finetune models from scratch on this vast dataset.
- **Stylized ImageNet[9]:** This dataset was used to create **cue-conflict**, shape-texture, scenarios and evaluate model biases. Stylized ImageNet enabled us to evaluate the model's bias towards a certain cue.
- **Canny Edge ImageNet:** This variant of ImageNet consists of edge-detected images generated using the Canny edge detection algorithm . It was used to test the OOD generalization of the models by removing texture details and focusing on shape-based features.
- **Animals10 [3]:** This dataset was included due to its rich texture and varying shapes. It served as a test bed for evaluating the generalization capabilities of our student model on unseen data with high visual diversity.
- **Stylized Animals10:** This stylized version of Animals10 was used to evaluate model biases and OOD performance. It was used to evaluate biases of models and performanc in cue conflict scenarios.

4.2 Models

In our experiments, we initially explored a variety of architectures to assess their suitability as teacher and student models for our KD framework. The models evaluated included VGG16, VGG11, ResNet50, DenseNet121, Vision Transformer Base Patch (ViT-Base-Patch), and Vision Transformer Small Patch (ViT-Small-Patch). Based on these experiments, we selected the following models:

- **Teachers:**
  VGG16: A CNN model, pretrained on ImageNet-1K with a proven bias toward texture features [10], making it an ideal candidate for transferring texture-specific knowledge.
  ViT-Base-Patch: A transformer architecture pretrained on ImageNet-1k, known for its bias toward capturing global shape information [13].
- **Student:**
  ViT-Small-Patch: A compact Vision Transformer model with 14M parameters [15]. Through experimentation, we observed that this model neither strongly adopted a shape bias nor a texture bias, making it a suitable candidate for learning a balanced representation from the teachers.

The selection of these models was driven by their complementary inductive biases, which ensured that biases can be propogated from these models. By distilling knowledge from the texture-biased VGG16 and the shape-biased ViT-Base-Patch, we aimed to produce a student model that integrates both biases while remaining compact and efficient for OOD generalization.

4.3 Baselines

To evaluate the effectiveness of our proposed multi-teacher KD framework, we compare its performance against several baseline approaches:

- **Single Teacher KD [11]:** We perform KD using a single teacher model at a time. Specifically, the student model is trained using either the texture-biased VGG16 or the shape-biased Vision Transformer (ViT) as the teacher. This approach establishes a baseline for how well the student can generalize when learning from a single source of bias.
- **Multi-Teacher AVG-KD:** This baseline involves distilling knowledge from both teacher models using equal-weighted averaging of their logits. By combining the predictions of VGG16 and ViT without considering confidence levels or entropy, this approach provides insight into the limitations of naive ensembling for knowledge transfer.
- **Individual Teacher Model Performance:** The performance of the VGG16 and ViT models is evaluated independently on OOD datasets. This serves as a reference for comparing the student model's ability to generalize beyond the biases inherent to its teachers. These results also highlight the trade-offs between texture and shape biases in individual architectures.

4.4 Evaluation Metrics

We use the following evaluation metrics to assess the performance of the models and measure the biases on two cues (Shape and Texture):

- **Accuracy:** The accuracy is measured by comparing the model's predictions against the true labels.
- **Cue Accuracy[7]:** This is defined as the ratio of predictions that match either the shape or texture label (i.e., correctly classified instances), which is computed as:

$$\text{Cue Accuracy}=\text{Shape Accuracy}+\text{Texture Accuracy}$$

- **Shape Bias:** The shape bias is calculated as the ratio of shape accuracy to cue accuracy:

$$\text{Shape Bias} = \frac{\text{Shape Accuracy}}{\text{Cue Accuracy}}$$

- **Texture Bias:** The texture bias is simply the complement of the shape bias:

$$\text{Texture Bias} = 1 - \text{Shape Bias}$$

4.5 Training Procedure and Knowledge Distillation

In our experiments, we began by fine-tuning both the teacher and student models on the respective datasets, following a standard training procedure. We used **10 epochs**, a learning rate of $\mathbf{1 \times 10^{-4}}$, and **batch size** of **32**. For the Animals10 dataset, both the teacher and student models were trained from scratch, while for ImageNet-1K, we used pretrained models and fine-tuned the classifiers on a subset of 16 superclasses from ImageNet, which were used in the cue-conflict dataset.

Initially, we measured the accuracies and biases of the models. After training, we observed that VGG16 exhibited a strong texture bias. To induce more shape bias in the Vision Transformer (ViT), we employed curriculum learning as follows:

- **Stage 1:** The model was trained on shape-biased data only.
- **Stage 2:** We mixed shape-biased data with the original data, ensuring that both models developed a shape bias.

For the student model, we used ViT-Small, which was initially untrained. The student was fine-tuned for **2 epochs** on the original data. For the remaining **8 epochs**, we applied our Confidence-Aware Knowledge Distillation (KD) framework to the student model.

Once training was completed, we measured the biases on the cue-conflict dataset to evaluate the shape and texture biases, as well as the accuracy on the original dataset to assess the impact of OOD generalization.

The experiments were conducted using various setups:

- Single Teacher to Student (using VGG16, ViT as the teacher),
- Multiple Teachers with Equal Weights to the Student (using VGG16 and ViT),
- Base Teachers
- Our Method (Confidence aware multi-teacher KD)

We compared the biases and accuracy results across these different setups to understand how each method influences the model's ability to generalize, especially in cue-conflict scenarios. Biases were recording across all these setups and accuracies we evaluated on the original, stylized and canny edge variants of **ImageNet1-k** and **Animals10**

.

5 Results and Findings

Firstly we conducted a series of baseline experiments to investigate the inductive biases of CNN and Vision Transformer (ViT) models. The accuracies for the standard, stylized, and Canny edge datasets were calculated, along with the biases (shape,texture) of the fine-tuned models using the CueConflict datasets respectively. Our results exhibited consistent trends across both datasets we tested it on ie: **ImageNet-1k** and **Animals-10**.

Models	**ImageNet-1K Accuracy**			**Animals-10 Accuracy**		
	Standard	Stylized	Canny Edge	Standard	Stylized	Canny Edge
VGG-16	**96.16%**	**34.62%**	18.33%	**73.22%**	**27.09%**	12.14%
VGG-11	93.42%	31.52%	12.15%	71.41%	20.19%	11.98%
ViT-Base Patch	86.44%	26.02%	**55.83%**	67.09%	12.07%	**35.67%**
ViT-Small Patch	82.14%	22.60%	47.29%	65.24%	9.89%	30.12%
ResNet50	86.54%	32.11%	14.03%	70.01%	23.44%	10.05%
DenseNet121	80.31%	28.19%	17.22%	69.16%	19.89%	10.63%

Table 1: Base performance comparison of different models on ImageNet-1K and Animals-10 datasets.

As seen in Table 1, the larger CNN model, VGG16, outperformed all other models on the standard dataset, achieving an accuracy of **96.16%** on ImageNet and **73.22%** on Animals-10. This performance can be attributed to the texture-rich nature of both datasets. On the stylized dataset, VGG16 again showed the highest accuracy as expected, achieving **34.62%** on ImageNet and **27.09%** on Animals-10, due to its ability to capture local features effectively.

On the Canny edge dataset, ViT-Base achieved the highest accuracy, with **55.83%** on ImageNet and **35.67%** on Animals-10, due to its ability to capture global shape patterns. However, the performance of ViTs on stylized datasets was notably low, with ViT-Small achieving only **9.89%** on Animals-10, highlighting its limitations in handling stylized data.

Models	**ImageNet-1K Bias**		**Animals-10 Bias**	
	Shape Bias	Texture Bias	Shape Bias	Texture Bias
VGG-16	**0.14**	**0.86**	**0.20**	**0.80**
VGG-11	0.21	0.79	0.27	0.73
ViT-Base	**0.64**	**0.36**	**0.59**	**0.41**
ViT-Small	0.59	0.41	0.56	0.44
ResNet50	0.29	0.71	0.31	0.69
DenseNet	0.23	0.77	0.38	0.62

Table 2: Bias (Shape, Texture) comparison for ImageNet-1K and Animals-10 datasets.

In terms of model biases in Table 2, we observed consistent results across both datasets as well. The VGG16 model, pretrained on ImageNet, exhibited a strong texture bias of **86%**, with a minimal shape bias of **14%**. In contrast, the ViT-Base model demonstrated a much more balanced bias, with **64%** shape bias and **36%** texture bias. This trend was similarly reflected in the Animals-10 dataset, where CNNs exhibited strong texture bias, while ViTs were shape-biased.

Overall, our findings confirmed that CNN models, such as VGG16, are predominantly biased towards texture, whereas Vision Transformers like ViT-Base exhibit a greater reliance on shape features, providing a strong baseline for our distillation results. To our surprise, ViT-Small Patch exhibited a relatively balanced bias compared to all other models, with **59%** shape bias and **41%** texture bias on ImageNet, and **56%** shape bias and **44%** texture bias on Animals-10. However, the theme of being shape-biased is persistent here as well. These results set up a strong baseline for testing our proposed methodology.

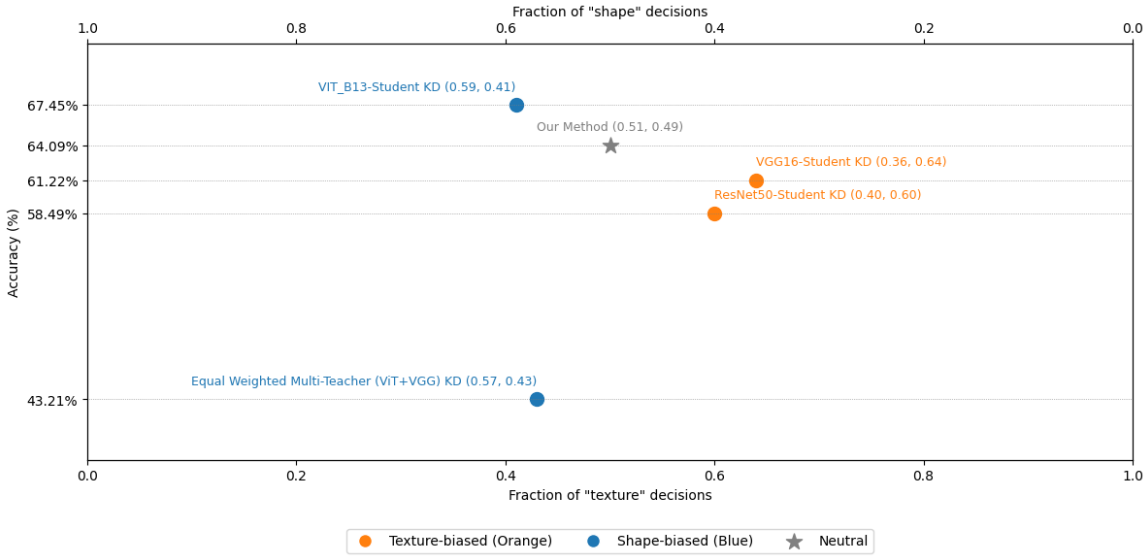

Figure 1: Texture vs. Shape Decision Fractions for various employed methods as baselines *(*Distillation was performed on limited epoch and smaller dataset due to compute and time constraints).*

In Figure 1, evaluation is conducted in terms of shape and texture biases alongside the accuracy of the model on the original dataset with a couple of baselines.

In the first set of experiments, knowledge was distilled from a single teacher to a student model, we observe that teacher biases are consistently propagated to the student. In distillation from VGG-16 to a student model, the resulting student model exhibits an accuracy of **61.22%**, with a texture bias of **0.64** and a shape bias of **0.36**. This indicates that despite the student being a ViT (Vision Transformer) model, which typically emphasizes shape cues, the texture bias of the VGG-16 teacher is still inherited. A similar trend is observed for ResNet-50, another CNN-based teacher model.

On distilling from a transformer-based teacher (ViT-B16), the student model inherits a shape bias of **0.59**, while texture of only **0.41**. We achieved the highest accuracy at **67.45%** due to same architectures. These findings highlight the significant influence of teacher biases on student models during knowledge distillation.

The equal weighting method performed poorly, achieving an accuracy of only **43.21%**. This highlights the importance of employing dynamically weighted teacher contributions during distillation. Equal weighting fails to leverage the complementary strengths of multiple teachers, resulting in poor performance.

Our proposed dynamically weighted knowledge distillation method achieves a notable balance between accuracy and bias generalization as show by the **star** in Figure 1. Although there is a slight trade-off in accuracy which is **64.09%**. Our approach results in nearly equal biases, with a shape bias of **0.49** and a texture bias of **0.51**. This balance ensures that the distilled model is equally generalizable to both shape- and texture-based cues in out-of-distribution data.

Methods	ImageNet-1K	
	Stylized Accuracy	Canny Edge Accuracy
VGG-16 to Student KD	29.14%	17.34%
ViT B-16 to Student KD	21.89%	30.20%
Equal Weighted Multi-Teacher KD	11.43%	7.06%
Confidence Aware Multi-Teacher KD	**28.89%**	**31.12%**

Table 3: OOD performance of Student Model post Knowedgle Distillation.

As we can see in Table 3, in all other methods, the student performed better towards a certain cue only, either stylized or Canny edge, depending on the teacher. In our method, it performed well on both the stylized and Canny edge datasets. Specifically, it achieved an accuracy of **28.28%** on the stylized dataset and **31.12%** on the Canny edge dataset. Although there is a very minute drop in accuracy on the stylized dataset as compared to the VGG16-student model, where it was **29.14%**, we achieve a much more OOD generalizable student model, as it performs well on both datasets rather than just one of the cues.

6 Discussions

Our experiments reveal insights into the inductive biases of CNNs and Vision Transformers (ViTs) and the effects of knowledge distillation on student model. CNNs, like VGG-16, excel at handling texture-rich datasets, achieving the highest accuracy on standard datasets due to their ability to capture local texture features. However, their performance on Canny edge datasets is lower, highlighting a limitation in capturing global shape patterns.

ViTs, on the other hand, perform better on the Canny edge dataset, with ViT-Base achieving the highest accuracy of **55.83% on ImageNet**, showcasing their strength in shape recognition. However, they struggle with stylized datasets, with ViT-Small achieving just 9.89% on Animals-10, underlining their challenge in processing local texture cues.

Bias analysis confirmed that CNNs are predominantly texture-biased, while ViTs tend to be more shape-biased, for example ViT-Base, which showed a **64% shape bias**. Surprisingly, ViT-Small exhibited a relatively balanced bias, though shape features still dominated.

In terms of knowledge distillation, we found that teacher model biases are transferred to student models. Distilling from texture-biased models like VGG-16 resulted in students inheriting texture bias, while distilling from shape-biased models like ViT-Base led to shape bias in students. Our dynamically weighted distillation method performed best, achieving a balance between accuracy (64.09%) and bias generalization at **0.49 shape vs. 0.51 texture**, outperforming other methods like equal weighting, which performed poorly at **43.21%** accuracy.

Overall, OOD generalization was improved with our approach, as the student model performed well on both stylized and Canny edge datasets, unlike other methods that excelled on just one cue. Although there was a slight drop in accuracy on stylized data compared to VGG16-student (29.14%), our method achieved a more generalizable model across both datasets.

On the whole, CNNs and ViTs exhibit complementary biases, with CNNs favoring texture and ViTs favoring shape. By using dynamically weighted distillation, we effectively combine these strengths, resulting in a more robust and generalized model for diverse tasks.

7 Conclusion

In this study, we examined the inductive biases of CNNs and Vision Transformers and their impact on OOD generalization. Our results show that CNNs, like VGG-16 perform good in texture rich datasets, while ViTs perform better with shape based cues. We demonstrated that knowledge distillation effectively transfers these biases from teacher to student models, and our proposed dynamically weighted distillation method strikes a balance between accuracy and bias generalization. This approach outperformed traditional methods, leading to a student model that generalizes well across both stylized and Canny edge datasets, making it more apt for diverse tasks.

While our method achieves a balance between bias generalization and accuracy, there is a slight trade-off in the original dataset performance. This can be attributed to the nature of the heterogeneous setup and simple logit matching KD. We have observed how during distillation a teacher model can propogate its biases in student model even if architectures are different, hence, our method tackles the issue of mitigating these biases and created a well balanced student model.

Future work can focus on enhancing the dynamic weight allocation for teacher models and exploring hint-based distillation to improve accuracy and generalization. While hint-based distillation has previously been unsuccessful, if we succeed in aligning the representations of models like VGG and ViT-B16 into a common latent space, it could significantly improve accuracy and make our model more generalizable.

Limitations

While our method balances shape and texture biases, it does not outperform the Single-Teacher ViT in accuracy. This trade-off stems from challenges in heterogeneous architecture, where mismatched model capacities hinder effective learning. Discrepancies in the logits' scale and distribution due to different datasets can impair convergence. Additionally, the lack of feature-based distillation limits knowledge transfer, focusing only on final output probabilities, which may reduce accuracy on the original dataset, advocating for exploration in feature based and structural distillation in a hetrogenous environment.

References

[1] S Abnar, M Dehghani, and W Zuidema. Transferring inductive biases through knowledge distillation. arxiv 2020. *arXiv preprint arXiv:2006.00555*.

[2] Samira Abnar, Mostafa Dehghani, and Willem Zuidema. Transferring inductive biases through knowledge distillation. *arXiv preprint arXiv:2006.00555*, 2020.

[3] Tamara L Berg and David A Forsyth. Animals on the web. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1463–1470. IEEE, 2006.

[4] Eduarda Caldeira, Jaime S Cardoso, Ana F Sequeira, and Pedro C Neto. Mst-kd: Multiple specialized teachers knowledge distillation for fair face recognition. *arXiv preprint arXiv:2408.16563*, 2024.

[5] 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*, pages 248–255. Ieee, 2009.

[6] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pages 1607–1616. PMLR, 2018.

[7] Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Bianca Lamm, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Are vision language models texture or shape biased and can we steer them? *arXiv preprint arXiv:2403.09193*, 2024.

[8] R. Geirhos. Texture vs shape. `https://github.com/rgeirhos/texture-vs-shape`, 2020. Accessed: 2024-11-19.

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[10] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33: 19000–19015, 2020.

[11] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[12] Kisoo Kwon, Hwidong Na, Hoshik Lee, and Nam Soo Kim. Adaptive knowledge distillation based on entropy. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7409–7413. IEEE, 2020.

[13] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[14] Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*, 2019.

[15] Ross Wightman. timm: Pytorch image models. `https://huggingface.co/timm`, 2019.

[16] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. One teacher is enough? pre-trained language model distillation from multiple teachers. *arXiv preprint arXiv:2106.01023*, 2021.

[17] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1285–1294, 2017.

[18] Qiangqiang Yuan, Huanfeng Shen, Tongwen Li, Zhiwei Li, Shuwen Li, Yun Jiang, Hongzhang Xu, Weiwei Tan, Qianqian Yang, Jiwen Wang, et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote sensing of Environment*, 241:111716, 2020.

[19] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.