

Perspective-Aware Teaching: Adapting Knowledge for Heterogeneous Distillation

Jhe-Hao Lin¹, Yi Yao¹ , Chan-Feng Hsu¹, Hong-Xia Xie² ,
Hong-Han Shuai¹ , Wen-Huang Cheng³ 

¹National Yang Ming Chiao Tung University ²Jilin University ³National Taiwan University

{jimmylin0979.11, leo81005.ee10, cfhsu311510211.ee11, hhshuai}@nycu.edu.tw

hongxiaxie@jlu.edu.cn, wenhuang@csie.ntu.edu.tw

Abstract

Knowledge distillation (KD) involves transferring knowledge from a pre-trained heavy teacher model to a lighter student model, thereby reducing the inference cost while maintaining comparable effectiveness. Prior KD techniques typically assume homogeneity between the teacher and student models. However, as technology advances, a wide variety of architectures have emerged, ranging from initial Convolutional Neural Networks (CNNs) to Vision Transformers (ViTs), and Multi-Level Perceptrons (MLPs). Consequently, developing a universal KD framework compatible with any architecture has become an important research topic. In this paper, we introduce a perspective-aware teaching (PAT) KD framework to enable feature distillation across diverse architectures. Our framework comprises two key components. First, we design prompt tuning blocks that incorporate student feedback, allowing teacher features to adapt to the student model’s learning process. Second, we propose region-aware attention to mitigate the view mismatch problem between heterogeneous architectures. By leveraging these two modules, effective distillation of intermediate features can be achieved across heterogeneous architectures. Extensive experiments on CIFAR, ImageNet, and COCO demonstrate the superiority of the proposed method. Our code is available at <https://github.com/jimmylin0979/PAT.git>.

1. Introduction

Knowledge distillation (KD) has received increasing attention from both academic and industrial researchers in recent years. The goal is to develop a lightweight and efficient student model by transferring knowledge from an already trained, larger teacher model. This knowledge may come in the form of predictions [17, 37], or intermediate features

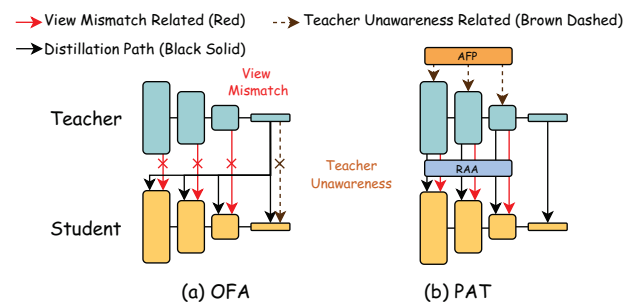


Figure 1. High-level comparison between former state-of-the-art OFA and our method PAT. Due to the view mismatch (red lines) and teacher unawareness (brown dashed lines) problems, OFA can only use final logits to supervise student intermediate features (black solid lines), thereby restricting its utility in downstream tasks. In contrast, PAT enables feature imitation by solving these two issues with RAA and AFP modules respectively.

[2, 3, 12], from the cumbersome pre-trained teacher model. Despite the significant progress in KD, there remains a gap in addressing cross-architecture distillation comprehensively. Existing KD approaches have predominantly focused on distillation between teacher and student models with homogeneous architectures, particularly feature-based approaches, leaving cross-architecture distillation relatively unexplored. For example, FitNet [30] achieves good results in CNN-to-CNN format but performs poorly when distilling in a heterogeneous CNN-to-ViT scenario, only getting 24.06% on CIFAR-100 with ConvNeXt-T teacher and Swin-P student model pair, which is lower than the basic logit-based method KD [10] by more than 50%.

Heterogeneous KD, however, offers numerous advantages. First, it helps overcome architecture-specific limitations, allowing the student model to mitigate weaknesses inherent to individual architectures and achieve a better performance than learning from a homogeneous teacher model. For instance, CNNs can assimilate a more comprehensive understanding from ViTs, conversely, ViTs can augment

their capacity for local feature extraction by leveraging insights from CNNs. Second, it enhances the robustness, as the student model learns from the varied error patterns and inductive biases of the teacher models, leading to better performance across different tasks and datasets. While a few studies have delved into the realm of heterogeneous architectures [20, 29, 38], it is noteworthy that these architectures often exhibit limitations in accommodating specific teacher-student pairings. As a result, a thorough restructuring of the link between the teacher and student models is essential whenever a new architecture is introduced into the system.

To overcome the existing limitations, OFA-KD [8] is the first to propose a generic framework for heterogeneous KD by projecting the intermediate student features into the logits space and subsequently aligning them with the final predictions of the teacher model. The projection into a latent space with reduced architecture-specific details allows the distillation process to concentrate on high-semantic information, thereby enhancing distillation performance across different architectures. However, this projection results in a loss of spatial information that is crucial for downstream tasks. Furthermore, distilling based on the final high-semantic output of the teacher model may not fully exploit the valuable intermediate features.

To effectively leverage the richly informative intermediate layer features and facilitate distillation across diverse architectures, we revisit the fundamental feature-based distillation pipeline to identify the potential issues. As illustrated in Fig. 1, apart from the different perspectives inherent in different architectural biases, which we refer to as the view mismatch problem, teacher unawareness is a notable concern, whereby the teacher model fails to adjust based on the student’s learning progress. Hence, a novel perspective-aware teaching (PAT) KD framework is proposed for distillation between heterogeneous architectures to tackle these issues. To address the view mismatch problem, a region-aware attention (RAA) module is proposed to let the student model learn how to reblend features from different patches and stages to achieve a similar perspective with the corresponding teacher features via the attention mechanism. Since the perspective is now similar, traditional stage-wise feature alignment across different architectures can be applied normally. For the teacher unawareness problem, an adaptive feedback prompt (AFP) module is introduced. The AFP module employs prompt tuning [24] to generate distillation-friendly teacher features with additional branches. Moreover, the difference of the feature map between the student and teacher model, referred to as feedback, is integrated as extra input to the prompt tuning blocks, enabling the teacher model to dynamically adapt its features in response to the student model’s learning trajectory.

Our contributions are summarized as follows:

- To address the view mismatch issue, we propose region-aware attention (RAA) to reblend student features to have a similar perspective with the teacher model.
- To address the teacher unawareness problem, we propose an adaptive feedback prompt (AFP) that enables the teacher model to dynamically adapt its features in response to the student model’s feedback.
- Extensive experiments manifest that PAT improves student model performance with gains of up to 16.94% on CIFAR-100 and 3.35% on ImageNet-1K and 3.50% performance gain on COCO.

2. Related Works

2.1. Homogeneous Knowledge Distillation

A majority of the KD methods [25, 37] predominantly concentrate on homogeneous architectures, particularly within feature-based approaches [30, 32, 36], as the intermediate features between such architectures exhibit greater similarity, thereby enhancing the efficiency of the distillation process. For instance, ReviewKD [3] introduces a review mechanism that leverages multi-level information from the teacher model to direct the single-level learning of the student model, achieving commendable performance in the distillation process between CNN-based models. The aforementioned methods exhibit efficacy in distillation processes involving homogeneous architectures. Nonetheless, they fail to consider heterogeneous architectures. As such, they may encounter challenges like view mismatch and teacher unawareness when attempting to transfer knowledge from a heterogeneous teacher model.

2.2. Generic Heterogeneous Knowledge Distillation

As more model architectures are developed, creating a KD framework that works with any architecture has become crucial. Different architectures have unique biases [27], which require careful design for KD and often limit it to specific teacher-student pairs [1, 20, 29, 39]. OFA-KD [8] solves this by converting student features into a logits space, allowing alignment across architectures by removing architecture-specific details, but it loses important spatial information in features. In this work, we aim to design a feature-based distillation framework that keeps rich spatial information, improving both classification performance and downstream tasks.

2.3. Adaptive Teacher in Knowledge Distillation

In typical KD approaches [3, 10, 17, 23], the teacher model is first trained in isolation and then used to distill the knowledge to the student model. This results in static knowledge transfer that ignores the student’s learning dynamics, often leading to suboptimal outcomes. This phenomenon has also been validated in [7], where they attempted to

improve student model learning by substituting the conventional teacher trained with cross-entropy loss with one trained using MSE loss. Although this approach yields better performance, training an entirely new teacher model often requires significantly higher computational resources. Another research direction draws inspiration from prompt tuning [13], exploring the possibility of leveraging task-specific prompts that can be adjusted to fine-tune the entire teacher model with minimal parameters [15, 38]. For instance, CAKD [38] integrates adaptable prompts into the teacher model, enabling it to manage distillation-specific knowledge while preserving its discriminative capability. In this paper, we propose using prompt tuning in a generic heterogeneous KD framework to create an adaptive teacher. Unlike previous methods that directly feed the teacher feature, we additionally incorporate student feedback to better align the adaptation with the student’s learning process, further improving the performance.

3. Methodology

In this section, we first review knowledge distillation (KD) and then introduce our PAT framework, which addresses existing issues such as view mismatch and teacher unawareness. We describe the adaptive feedback prompt (AFP) and region-aware attention (RAA) components, and summarize the objective function. An overview of the PAT framework is shown in Fig. 2.

3.1. Preliminaries

Knowledge Distillation (KD) [10] aims to effectively transfer knowledge from the pre-trained teacher model to a more compact student model. Among the various types of KD, logits and features are the most commonly used forms of knowledge to transfer.

Logits-based distillation trains the student model to mimic the predictive distribution of the teacher model. This process can be formulated as:

$$L_{LD} = L_{KL}(p^T, p^S), \quad (1)$$

where p^T, p^S represents the teacher and student model’s confidence across all classes, respectively, and L_{KL} denotes the KL-divergence function. For simplicity, we omit the temperature term here.

On the other hand, features-based distillation utilizes intermediate features from the teacher model to train the student model in a more fine-grained manner, which can be expressed as:

$$L_{FD} = \sum_{i=1}^4 D(F_i^T, M_i(F_i^S)), \quad (2)$$

where F_i^T, F_i^S denote the features of the teacher and student model in the i -th stage, respectively. All teacher and student models are divided into four stages. M_i represents

the transformation applied to the student feature at the i -th stage, and D denotes the distance function that measures the gap between the teacher and student features, which is typically the Mean Squared Error (MSE) function.

Overall, while logits-based distillation incurs minimal computational costs, its performance is generally inferior to that of the features-based approach. Despite this limitation, the architecture-agnostic nature of the predictive distribution in logits allows for straightforward application in heterogeneous KD scenarios. Conversely, features-based distillation requires careful designs to address potential bias misalignment but has the potential for superior performance, especially on downstream tasks. In this paper, we develop a generic feature-based distillation framework and mitigate the misalignments by introducing the region-aware attention (RAA) module and adaptive feedback prompt (AFP).

3.2. Region-Aware Attention (RAA) Module

A key issue in heterogeneous knowledge distillation is the difference in perspective between the teacher and student models, due to their distinct architectural receptive fields. For example, a ViT-based model begins with a broad view, whereas a CNN-based model builds up to a global understanding from local views, reaching full integration in the final layers. When trying to use the traditional stage-wise feature alignment method, this difference in perspective creates a significant challenge for the student model, hindering its ability to learn effectively from the teacher model.

To address this issue, we introduce a region-aware attention (RAA) module. This module leverages an attention mechanism to help the student model learn to integrate features from different regions across stages, thereby achieving a perspective comparable to the teacher model. Owing to the flexibility of the attention module, RAA can seamlessly integrate into any architecture without requiring a distinct design. The proposed RAA is shown in Fig. 2(a) and is composed of three steps. First, we patchify the student features from each stage i to $\frac{N_q}{4}$ patches to capture regional details within each stage, and then use convolution projectors C_i to map them to a predefined dimension $\mathbb{R}^{\frac{N_q}{4} \times d}$, where N_q is a hyperparameter that denotes the total number of queries across all four stages. These stage features are then concatenated as $F^S \in \mathbb{R}^{N_q \times d}$ as depicted in Eq. 3.

$$F^S = \text{Concat}(C_1(F_1^S), \dots, C_4(F_4^S)), \quad (3)$$

Second, We feed F^S into a self-attention module as expressed in Eq. 4, which is defined as:

$$F^{S'} = \text{Softmax}\left(\frac{(W_q F^S)(W_k F^S)^T}{\sqrt{d}}\right)W_v F^S, \quad (4)$$

where W_q, W_k, W_v are the learnable weights, and d is the channel dimension. F^S contains features from different

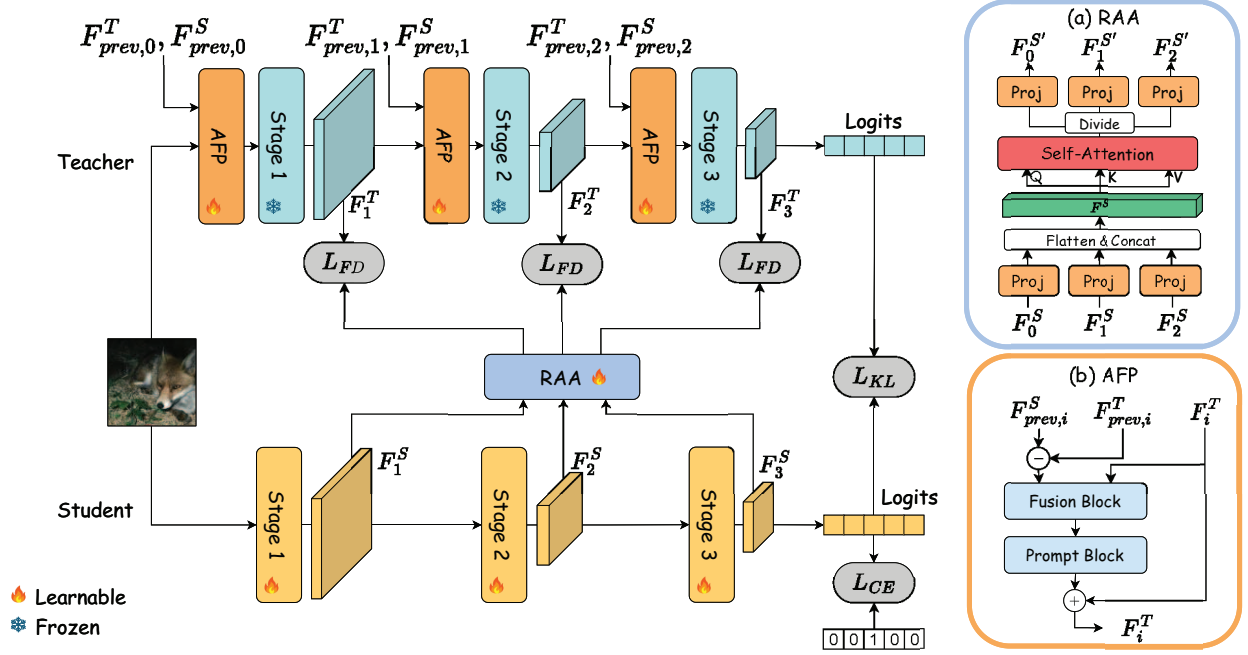


Figure 2. A comprehensive depiction of the general structure of our PAT framework and proposed modules. (a) RAA: student features across all stages are concatenated and fed into an attention module to learn how to integrate a new feature with a perspective similar to that of the teacher model. (b) AFP: a prompt-tuning method is introduced to modify the output stage features of the teacher model with respect to the student model’s learning process. *Note*: Only three stages are shown for convenience. In our experiment, all models are split into four stages.

patches and different stages. Through the attention mechanism, the student model learns how to focus on different spatial locations from different stages to blend a new feature with a perspective similar to that of the teacher model. Finally, we divide the blended student features $F^{S'}$ back into four stages after RAA and project each to the same size as the teacher. As the blended student features approximate the teacher’s perspective, standard feature matching can subsequently be applied as described in Eq. 2.

3.3. Adaptive Feedback Prompts (AFP) for Teacher

The existing KD paradigm typically involves training the teacher model in isolation. This independent training means that the teacher model does not account for the subsequent distillation process intended to train the student model. Consequently, the intermediate information generated by the teacher model may not be effective for distillation, leading to suboptimal performance. This problem is particularly pronounced in heterogeneous KD, as models with different architectures have varying inductive biases, making their features less compatible with the student model.

A naive solution is to optimize the teacher model during the distillation process simultaneously [6]. However, this method requires more computational cost and may lead to catastrophic forgetting [14] in the teacher model. To address

this problem, we propose an adaptive feedback prompts (AFP) module, which uses prompt tuning [13, 24] to fine-tune the teacher model’s knowledge with minimal parameters, aiming not only to eliminate information that is not conducive to distillation but also to make adjustments based on student feedback, while preserving the discriminative capacity of the teacher model.

In practice, the AFP modules are inserted before the start of each stage in the teacher model, and each consists of two blocks, prompt block PB , and fusion block FB , as shown in Fig. 2(b). To accommodate various architectures, we follow [24] to develop PB that integrates task-specific knowledge through an additional branch without the need to modify any architecture. Unlike the previous method, which directly feeds the teacher’s features into the PB [38], we introduce FB before the PB to incorporate the difference of the feature map between the student and teacher model, referred to as student feedback, into the teacher model’s features, aiming to make the output prompt able to reflect the student feedback. Specifically, let M_i^{AFP} denotes the transformation that transfers the student feature to the same dimension as the teacher’s at the i -th stage, we subtract the student features from the teacher features in each stage to create feedback as follows:

$$Feedback_i = M_i^{AFP}(F_{prev,i}^S) - F_{prev,i}^T. \quad (5)$$

This feedback identifies the locations and intensities where the teacher and student features are currently inconsistent.

During forward, the AFP block inserted before i -th stage takes the i -th stage teacher feature $F_{prev,i}^T$ and student feature $F_{prev,i}^S$ from the previous iteration, along with the current i -th stage teacher feature F_i^T as input, and outputs the final feedback enhanced teacher feature $F_i^{T'}$ with formula as follows:

$$F_i^{T'} = PB_i(FB_i(Concat(Feedback_i, F_i^T))). \quad (6)$$

We first concatenate the feedback with the teacher features and input them into a fusion block, resulting in a teacher feature that the feedback has enhanced. This feedback-enhanced feature is then sent to the prompt block to modify the teacher model’s output features. In this way, the teacher model can consider the student’s learning process via feedback, thus providing more friendly features to the student model during distillation. Note that the feedback is derived by subtracting the teacher and student features from the previous iteration, as utilizing the current teacher and student features to generate feedback may lead to the model easily adapting the teacher features based on this spatial-aligned feedback. Employing features from previous instances enables the model to understand the error patterns, thereby facilitating a beneficial adjustment for the student model.

Empirically, we find that the teacher features will be altered too much by the AFP modules in some teacher-student model pairs, even making the altered teacher features identical to the student’s, rendering feature matching useless. To mitigate this issue, we introduce a regularization loss L_{Reg} to avoid the teacher’s features from converging to the student’s features, thus preventing the generation of uninformative identity features by the teacher model that offers no value to the student model. The loss is defined as follows:

$$L_{Reg} = L_{KL}(p^T, p^{T'}), \quad (7)$$

where $p^{T'}$ is the predictive distribution of the teacher model incorporating the AFP modules. By ensuring that the KL-divergence distance between the predictive distributions with and without AFP modules remains minimal, we maintain the discriminative capacity of the teacher model, thereby averting the emergence of non-informative features.

3.4. Overall Objective Function

By integrating the proposed modules into the general knowledge distillation framework, the overall loss function of PAT is defined as:

$$L_{PAT} = L_{CE} + \alpha L_{KL} + \beta L_{FD} + \gamma L_{Reg}, \quad (8)$$

where L_{CE} is the standard cross-entropy loss, L_{KD} denotes the KL-divergence loss between p^T and p^S , and L_{FD}

denotes the feature matching loss between $F^{T'}$ and $F^{S'}$. The weights α , β , and γ balance the contribution of each term. For L_{FD} , we adopt the hierarchical context loss (HCL) from ReviewKD [3] as the distance function D , providing stronger hierarchical supervision.

4. Experiments

We conduct experiments on various tasks. First, we compare PAT with other KD methods on two image classification datasets. We then apply our method to object detection to show the pros of feature-based distillation. All results of PAT are reported by taking the average over 3 trials. Due to the space constraint, please refer to the supplementary for the implementation details.

Datasets Two image classification datasets and one detection dataset are adopted for our experiments. (1) CIFAR-100 [16] is a well-known dataset, containing 50K training and 10K test images of size 32×32 in 100 classes. (2) ImageNet [4] is a large-scale classification dataset, which provides 1.2M images for training and 50K images for validation over 1,000 classes. (3) COCO [18] is an 80-category object detection dataset, which contains 118K images for training and 5K images for validation.

Models We evaluate our PAT on various networks, including ResNet [9], MobileNetv2 [31], and ConvNeXt [22] as CNN-based models, ViT [5], DeiT [34] and Swin Transformer [21] as ViT-based models, and MLP-Mixer [33] and ResMLP [35] as MLP-based models.

Baselines We compare with various approaches that are feature-based, such as FitNet [30], CC [26], RKD [25], and CRD [32], as well as the baselines that are based on logits, which include KD [10], DKD [37], DIST [11] and the most recent state-of-the-art (SOTA) OFA-KD [8].

4.1. Results on Image Classification

On CIFAR-100, we compare our PAT method with eight other KD methods, encompassing a total of twelve heterogeneous teacher-student pairings. As shown in Table 1, PAT outperforms other methods on most teacher-student pairs, resulting in an average improvement of 8.17%. For instance, we achieve an accuracy of 83.50% on the ConvNeXt-T - ResMLP-S12 pair and 79.59% on the ConvNeXt-T - DeiT-T pair, surpassing not only the other feature-based methods but also obtaining an improvement of 2.28% and 3.83% compared to the prior SOTA logits-based method OFA. Specifically, PAT does not achieve as much improvement on CNN-based students as it does on ViT-based and MLP-based students. Unlike the other two types of students, CNN-based students need additional patch-embedding layers, we speculate that these additional parameters could result in prolonged training epochs, thus only having a small improvement in current settings.

Teacher	Student	From Scratch		Logits-based				Features-based				
		T.	S.	KD	DKD	DIST	OFA	FitNet	CC	RKD	CRD	PAT
<i>CNN-based students</i>												
Swin-T	ResNet18	89.26	74.01	78.74	80.26	77.75	<u>80.54</u>	78.87	74.19	74.11	77.63	81.22
ViT-S	ResNet18	92.04	74.01	77.26	78.10	76.49	80.15	77.71	74.26	73.72	76.60	80.11
Mixer-B/16	ResNet18	87.29	74.01	77.79	78.67	76.36	<u>79.39</u>	77.15	74.26	73.75	76.42	80.07
Swin-T	MobileNetV2	89.26	73.68	74.68	71.07	72.89	80.98	74.28	71.19	69.00	<u>79.80</u>	78.78
ViT-S	MobileNetV2	92.04	73.68	72.77	69.80	72.54	<u>78.45</u>	73.54	70.67	68.46	78.14	78.87
Mixer-B/16	MobileNetV2	87.29	73.68	73.33	70.20	73.26	78.78	73.78	70.73	68.95	78.15	<u>78.62</u>
<i>ViT-based students</i>												
ConvNeXt-T	DeiT-T	88.41	68.00	72.99	74.60	73.55	<u>75.76</u>	60.78	68.01	69.79	65.94	79.59
Mixer-B/16	DeiT-T	87.29	68.00	71.36	73.44	71.67	<u>73.90</u>	71.05	68.13	69.89	65.35	74.66
ConvNeXt-T	Swin-P	88.41	72.63	76.44	76.80	76.41	<u>78.32</u>	24.06	72.63	71.73	67.09	80.74
Mixer-B/16	Swin-P	87.29	72.63	75.93	76.39	75.85	78.93	75.20	73.32	70.82	67.03	<u>78.44</u>
<i>MLP-based students</i>												
ConvNeXt-T	ResMLP-S12	88.41	66.56	72.25	73.22	71.93	<u>81.22</u>	45.47	67.70	65.82	63.35	83.50
Swin-T	ResMLP-S12	89.26	66.56	71.89	72.82	11.05	<u>80.63</u>	63.12	68.37	64.66	61.72	80.94
<i>Average Improvement</i>				3.17	3.16	-2.31	<u>7.47</u>	-5.20	-0.33	-1.40	-0.02	8.17

Table 1. Result on CIFAR-100. Our results are the average over 3 trials. The highest results are indicated in bold, while the second-best results are underlined.

Teacher	Student	From Scratch		Logits-based				Features-based				
		T.	S.	KD	DKD	DIST	OFA	FitNet	CC	RKD	CRD	PAT
<i>CNN-based students</i>												
Swin-T	ResNet18	81.38	69.75	71.14	71.10	70.91	71.85	71.18	70.07	68.89	69.09	<u>71.54</u>
Mixer-B/16	MobileNetV2	76.62	68.87	71.92	70.93	71.74	<u>72.12</u>	71.59	70.79	69.86	68.89	72.22
<i>ViT-based students</i>												
ConvNeXt-T	DeiT-T	82.05	72.17	74.00	73.95	74.07	<u>74.41</u>	70.45	73.12	71.47	69.18	74.44
<i>MLP-based students</i>												
Swin-T	ResMLP-S12	81.38	76.65	76.67	76.99	77.25	<u>77.31</u>	76.48	76.15	75.10	73.40	77.59
<i>Average Improvement</i>				1.57	1.38	1.63	<u>2.06</u>	0.57	0.67	-0.53	-1.72	2.09

Table 2. Result on ImageNet. Our results are the average over 3 trials. The highest results are indicated in bold, while the second-best results are underlined.

Due to computational limitations, we select four teacher-student settings to experiment with on ImageNet. The results are presented in Table 2. With an increase in the volume of training data, the majority of methods exhibit a favorable average enhancement. Among them, PAT attains SOTA performance across most teacher-student pairs. For instance, we achieve 72.22% on MobileNetV2 and 77.59% on ResMLP-S12 when distilling from the teacher model Mixer-B/16 and Swin-T, which is 3.35% and 0.94% higher than the baselines, proving our generalizability on a larger dataset.

Previous feature-based methods achieve inferior perfor-

mance compared to logits-based methods, even getting a negative average improvement. Methods like CC, RKD, and CRD learn from the relationships between the data samples rather than directly imitating features, making them less vulnerable to view mismatch. However, they still underperform as they ignore the teacher unawareness problem. FitNet and our method both imitate the teacher model’s features, however, without considering the view mismatch problem, the imitation is less efficient and thus performs badly. Our performance suggests that by tackling the discrepancy in perspectives and lack of awareness in the instructor, a feature-imitated approach can achieve compara-

	Swin-T & ResNet18			Swin-T - MobileNetV2		
	mAP	AP50	AP75	mAP	AP50	AP75
Teacher	45.14	67.09	49.25	45.14	67.09	49.25
Student	33.26	53.61	35.26	29.47	48.87	30.90
KD	34.07	55.26	36.48	31.46	52.40	32.74
DKD	29.96	51.17	31.36	32.10	<u>53.82</u>	33.88
OFA	33.37	54.98	35.13	31.69	52.91	32.88
FitNet	<u>35.23</u>	<u>56.09</u>	<u>37.31</u>	<u>32.48</u>	<u>52.62</u>	<u>34.67</u>
PAT	35.62	56.67	38.04	32.97	54.18	35.08

Table 3. Results on COCO based on Faster RCNN [28] with FPN [19]. The highest results are indicated in bold, while the second-best results are underlined.

ble or even superior performance in heterogeneous KD scenarios.

4.2. Results on Object Detection

We extend our method to object detection and follow the setting of DKD [37] to perform distillation between the teacher’s and student’s backbone features. The results are presented in Table 3. Logits-based methods, as mentioned in DKD, can only achieve inferior performance as they disregard the intermediate features. For example, OFA can not reach good performance owing to the excessive focus on semantic details instead of spatial aspects, which is crucial for detection. Conversely, feature-based methods demonstrate superior performance, with our approach achieving SOTA results. We observe an enhancement of 2.36% for ResNet18 and 3.50% for MobileNetV2 when distilling from the Swin-T teacher model. This underscores that by mitigating view mismatch and teacher unawareness issues, the feature-mimicking technique can effectively leverage the abundant intermediate features for improved performance across classification and downstream tasks.

RAA	AFP w/o Feedback	AFP	Acc.
	Baseline (FitNet)		60.71
✓			70.12
✓	✓		79.13
✓		✓	79.59

Table 4. The effectiveness of our proposed RAA, AFP on CIFAR-100 with ConvNeXt-T - DeiT-T as the teacher-student pair.

5. Ablation Studies

5.1. Effectiveness of each proposed module

Experiments are conducted to evaluate the effectiveness, where the proposed modules are gradually integrated to measure their influence, and the results are summarized in Table 4. By leveraging the RAA module, the issue of view

Stage 1	Stage 2	Stage 3	Stage 4	Mem. (GB)	Acc.
Baseline (FitNet)				6.12	60.71
✓				10.76	75.43
	✓			11.55	77.34
		✓		10.30	76.26
			✓	9.01	69.65
✓	✓	✓		13.70	77.87
	✓	✓	✓	12.31	78.71
✓	✓	✓	✓	14.02	79.59

Table 5. Effect of number and position of AFP on CIFAR-100 with ConvNeXt-T - DeiT-T as the teacher-student pair.

N_q	Mem. (GB)	Acc.
Baseline (FitNet)	6.12	60.71
36	13.79	79.03
64	14.02	79.59
80	14.37	80.14
144	14.54	80.89

Table 6. Comparison of the N_q in RAA on CIFAR-100. We adopt ConvNeXt-T - DeiT-T as the teacher-student pair.

mismatch is resolved, leading to a significant improvement in performance compared to the baseline. When we further introduce the AFP module to adapt the teacher model, the student yields larger gains, and the performance becomes the best when student feedback is incorporated into the AFP.

5.2. Effectiveness of number and position of AFP

Since adopting AFP at every stage is beneficial yet introduces additional parameters, we conduct experiments to investigate the effectiveness of adopting AFP at each stage as in Table 5. Enhancements in performance are observed even with the adoption of AFP in just one of the stages, with optimal outcomes achieved when AFP is implemented across all stages but also with the highest memory usage. In our experiments, we adopt AFP at each stage of the teacher model by default to achieve the best performance.

5.3. Number of queries on RAA

In RAA, each stage feature of the student is projected into a predefined sequence of shape $\mathbb{R}^{\frac{N_q}{4} \times d}$ for attention computation. Increasing N_q allows finer subdivision of features, offering more detailed representations, but also leads to larger attention matrices and higher computational cost. We evaluate the impact of varying N_q from 36 to 144 on performance and memory usage. As shown in Table 6, performance improves steadily with larger N_q , accompanied by proportional memory growth. Notably, setting $N_q = 64$ achieves SOTA on CIFAR-100 with minimal memory usage. By default, we use $N_q = 64$ for CIFAR-100 and

Teacher Student	Swin-T ResNet18	ConvNeXt-T DeiT-T	ConvNeXt-T ResMLP-S12
KD	78.74	72.99	72.25
FitNet	78.87	60.78	45.47
PAT w/o KL Loss	79.14	75.98	81.79
PAT w/ KL Loss	81.22	79.10	83.50

Table 7. The effectiveness of loss L_{KL} on PAT on CIFAR-100.

Method	Extra Para. (M)	Epoch (s)	Acc.
KD	0	65.75	72.99
OFA	3.78	76.16	75.76
FitNet	1.11	90.45	60.78
PAT	14.48	208.57	79.59

Table 8. The complexity of KD methods with ConvNeXt-T - DeiT-T as the teacher-student pair on CIFAR-100.

$N_q = 144$ for larger datasets such as ImageNet and COCO.

5.4. The performance of PAT without KL loss

PAT is not purely feature-based, as it also incorporates a logits-based KL-divergence loss, as shown in Eq.8. This raises the question of whether performance gains stem primarily from the logits-matching loss rather than the feature-level alignment. To investigate, we train PAT without the KL loss. As shown in Table7, adding the KL loss improves performance. However, even without it, PAT still achieves performance close to SOTA. For best results, we include the KL loss by default to further enhance overall effectiveness.

5.5. Complexity Analysis

We also explore the complexity of our method. As shown in Table 8, it is found that logits-based methods, which rely solely on logits for distillation, require significantly fewer parameters compared to feature-based methods, leading to faster training times. In contrast, feature-based methods involve more parameters, which extends the training duration. Although PAT consumes the most parameters, it also achieves the highest performance. Notably, these additional parameters only appear during training and do not add any overhead to the student model during inference.

For the trade-off between the complexity and performance of PAT, as mentioned in section 5.2 and 5.3, we can achieve faster training by lowering the number of AFP modules, or by decreasing the queries number on RAA.

5.6. Visualization of attention map within RAA

Direct feature matching in heterogeneous KD is suboptimal due to differing inductive biases across architectures, leading to mismatched feature representations. To address this, PAT leverages attention to guide the student in blending features from different regions and stages, aligning its

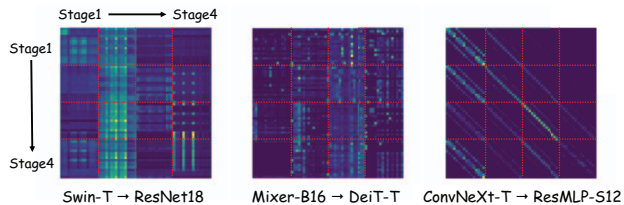


Figure 3. Visualizations of the attention map within RAA from different model pairs. Queries are sorted based on patch and stage position, as the black arrow shows.

perspective with the teacher. As shown in Fig. 3, attention patterns vary across teacher-student pairs. For example, the MLP-based student ResMLP-S12 learns from the CNN-based teacher ConvNeXt-T by attending to consistent spatial positions across stages, forming diagonal patterns that reflect local aggregation. In contrast, the CNN-based student ResNet18, learning from the ViT-based teacher Swin-T, captures features from neighboring patches, resulting in grid-like patterns indicative of global context aggregation. Similarly, the ViT-based student DeiT-T, when guided by the MLP-based teacher Mixer-B16, exhibits global attention with a strong focus on stage 3, suggesting that key information is concentrated there.

6. Conclusion

In this paper, we propose a perspective-aware teaching (PAT) framework to facilitate feature-level supervision in the context of heterogeneous KD. First, we introduce region-aware attention to blend the stage features of the student model to align with those of the teacher model, thereby alleviating the issue of view mismatch prevalent between the two models. Moreover, we incorporate the prompt tuning technique to address the teacher unawareness problem and enable to adapt the teacher model’s feature utilizing feedback from the student model with minimal parameters. Through the integration of these two components, we can realize feature-level distillation in heterogeneous KD scenarios and achieve promising performance in classification and downstream tasks.

7. Acknowledgments

This work is partially supported by the National Science and Technology Council, Taiwan, under Grant: NSTC-114-2640-E-A49-011, NSTC-114-2218-E-A49-017, and NSTC-112-2628-E-002-033-MY4, and was financially supported in part by the Center of Data Intelligence: Technologies, Applications, and Systems, National Taiwan University (Grants: 114L900901/114L900902/114L900903), from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education, Taiwan.

References

- [1] Samira Abnar, Mostafa Dehghani, and Willem Zuidema. Transferring inductive biases through knowledge distillation. *arXiv preprint arXiv:2006.00555*, 2020.
- [2] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *AAAI*, pages 7028–7036, 2021.
- [3] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, pages 5006–5015, 2021.
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- [6] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *CVPR*, pages 11017–11026, 2020.
- [7] Shayan Mohajer Hamidi, Xizhen Deng, Renhao Tan, Linfeng Ye, and Ahmed Hussein Salamah. How to train the teacher model for effective knowledge distillation. In *ECCV*, 2024.
- [8] Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. In *NeurIPS*, 2023.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [11] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *NeurIPS*, 2022.
- [12] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. *AAAI*, 35:7945–7952, 2021.
- [13] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022.
- [14] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L. Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *AAAI*. AAAI Press, 2018.
- [15] Gyeongman Kim, Doohyuk Jang, and Eunho Yang. Promptkd: Distilling student-friendly knowledge for generative language models via prompt tuning. *arXiv preprint arXiv:2402.12842*, 2024.
- [16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [17] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *AAAI*, pages 1504–1512, 2023.
- [18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.
- [19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017.
- [20] Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, Jingting Ding, and Liang Li. Cross-architecture knowledge distillation. In *ACCV*, pages 3396–3411, 2022.
- [21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 9992–10002, 2021.
- [22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11966–11976, 2022.
- [23] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, number 04, pages 5191–5198, 2020.
- [24] Xing Nie, Bolin Ni, Jianlong Chang, Gaofeng Meng, Chunlei Huo, Shiming Xiang, and Qi Tian. Pro-tuning: Unified prompt tuning for vision tasks. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023.
- [25] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pages 3962–3971, 2019.
- [26] B. Peng, X. Jin, D. Li, S. Zhou, Y. Wu, J. Liu, Z. Zhang, and Y. Liu. Correlation congruence for knowledge distillation. In *ICCV*, pages 5006–5015, Los Alamitos, CA, USA, 2019. IEEE Computer Society.
- [27] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *NeurIPS*, 2021.
- [28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*. Curran Associates, Inc., 2015.
- [29] Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, and Hang Zhao. Co-advise: Cross inductive bias distillation. In *CVPR*, pages 16752–16761, 2022.
- [30] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.
- [31] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.

- [32] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020.
- [33] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Peter Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-mixer: An all-MLP architecture for vision. In *NeurIPS*, 2021.
- [34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021.
- [35] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5314–5321, 2023.
- [36] Yuzhu Wang, Lechao Cheng, Manni Duan, Yongheng Wang, Zunlei Feng, and Shu Kong. Improving knowledge distillation via regularizing feature norm and direction. In *ECCV*, 2024.
- [37] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, pages 11943–11952, 2022.
- [38] Weisong Zhao, Xiangyu Zhu, Zhixiang He, Xiao-Yu Zhang, and Zhen Lei. Cross-architecture distillation for face recognition. In *Proceedings of the ACM International Conference on Multimedia (MM)*, page 8076–8085, New York, NY, USA, 2023. Association for Computing Machinery.
- [39] Xu Zheng, Yunhao Luo, Pengyuan Zhou, and Lin Wang. Distilling efficient vision transformers from cnns for semantic segmentation. *arXiv preprint arXiv:2310.07265*, 2023.