Co-advise: Cross Inductive Bias Distillation

Sucheng Ren[1,5] Zhengqi Gao[2] Tianyu Hua[3,5] Zihui Xue[4] Yonglong Tian[2]
Shengfeng He[1*] Hang Zhao[3,5*]
[1]South China University of Technology [2]Massachusetts Institute of Technology
[3]Tsinghua University [4]The University of Texas at Austin [5]Shanghai Qi Zhi Institute

Abstract

The inductive bias of vision transformers is more relaxed that cannot work well with insufficient data. Knowledge distillation is thus introduced to assist the training of transformers. Unlike previous works, where merely heavy convolution-based teachers are provided, in this paper, we delve into the influence of models inductive biases in knowledge distillation (e.g., convolution and involution). Our key observation is that the teacher accuracy is not the dominant reason for the student accuracy, but the teacher inductive bias is more important. We demonstrate that lightweight teachers with different architectural inductive biases can be used to co-advise the student transformer with outstanding performances. The rationale behind is that models designed with different inductive biases tend to focus on diverse patterns, and teachers with different inductive biases attain various knowledge despite being trained on the same dataset. The diverse knowledge provides a more precise and comprehensive description of the data and compounds and boosts the performance of the student during distillation. Furthermore, we propose a token inductive bias alignment to align the inductive bias of the token with its target teacher model. With only lightweight teachers provided and using this cross inductive bias distillation method, our vision transformers (termed as CiT) outperform all previous vision transformers (ViT) of the same architecture on ImageNet. Moreover, our small size model CiT-SAK further achieves 82.7% Top-1 accuracy on ImageNet without modifying the attention module of the ViT. Code is available at https://github.com/OliverRensu/co-advise.

1. Introduction

Although convolutional neural network (CNN) has revolutionized the field of computer vision, it possesses certain limitations. Recent research interests have been intrigued in

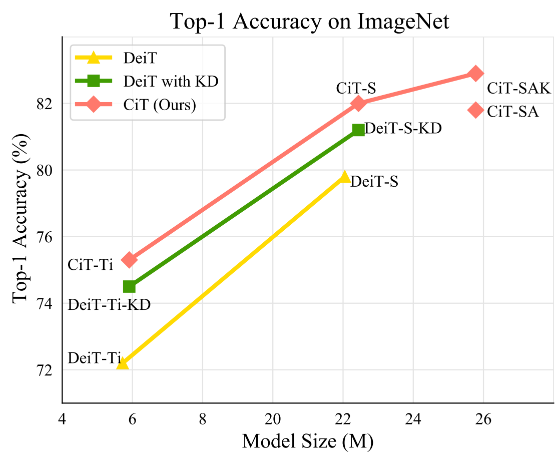


Figure 1. Comparison with DeiT. Here CiT-SA and CiT-SAK indicate models with token inductive bias alignment, without or with distillation. Our cross inductive bias distillation (CiT) outperforms DeiT where only lightweight teachers are provided. Combining with token inductive bias alignment, the performance of our method can be further improved.

replacing convolution layers with novel self-attention-based architectures. For instance, ViT [6] is a pure transformer without convolutional layers. Nevertheless, transformers have fewer inductive biases than CNNs (e.g., translation equivariance and locality) and thus suffer when the given amounts of training data are insufficient [6]. In this context, knowledge distillation technique [7, 16] is applied by DeiT [30] to assist the training of vision transformers. When the CNN teacher is powerful enough, transformers with such distillation [30] (*i.e.*, DeiT) can achieve competitive results as SOTA CNNs on ImageNet. However, DeiT has its own limitations: 1) The trained transformer is over-influenced by the inductive bias of the teacher CNN and mirrors its classification error; 2) DeiT requires the teacher CNN to be very large (e.g., RegNetY-16GF), which disturbingly brings about heavy computational overhead (e.g., training a RegNetY-16GF on ImageNet takes four times longer training time under the same training protocols than DeiT-S); 3) Class

*Corresponding authors: Shengfeng He (hesfe@scut.edu.cn), Hang Zhao (hangzhao@mail.tsinghua.edu.cn).

Figure 2. Class probabilities predicted by a CNN, an INN, a transformer without distillation, and a transformer distilling from both CNN and INN. CNN and INN come up with consistent (the first row) or complementary (the second and third rows) conclusions to correct transformer's prediction.

token and the Distillation token have different targets but share the same random initialization protocol.

In this paper, we argue that a heavy and highly-accurate teacher is not necessarily effective in teaching a "good" student transformer. Instead, the involved inductive bias plays a leading role. Our key observation is that models with different inductive biases tend to focus on diverse patterns despite that they are trained on the same dataset (see Figure 2). Namely, compared with naive teacher assembling, teachers of different inductive biases inherently make complementary assumptions of the data they see and focus on the data from various perspectives to attain diverse knowledge. They provide more precise, complementary and comprehensive descriptions of the data, which further compounds and boosts the performance of student during distillation. In contrast, teachers with similar inductive biases but different performance (e.g., ResNet-18 and ResNet-50) have little differences in data descriptions, and the student distilling from them have limited performance gain.

To compare the influence of directly introducing inductive bias to the model and knowledge distillation, we propose a token alignment technique. Specifically, two tokens are used in DeiT, learning from a CNN teacher and golden labels, respectively. However, these two tokens share the same random initialization protocol, which we believe, actually limits the power of them to learn different targets. To make the representation power of tokens close to their corresponding teachers so that they could truly move towards their corresponding teachers, we propose token inductive bias alignment by further introducing inductive bias into tokens. In our experiments, we show that introducing the inductive bias to student model by our inductive bias alignment truly brings improvements on ImageNet. However, we also find that comparing with directly introducing the same inductive bias with the teacher model into the model by our inductive bias alignment, knowledge distillation helps the student to perform more similar to the teacher. Therefore, we find that although knowledge distillation cannot "transfer" inductive bias to the student, it helps the student to "inherit" more characteristics of the teacher.

Thanks to complementary inductive biases of convolution (spatial-agnostic and channel-specific) and involution (spatial-specific and channel-agnostic), our method only requires two super lightweight teachers (a CNN and an INN). In the distillation stage, the knowledge from teachers compensates each other and significantly prompts the accuracy of the student transformer. Our main observations of this paper are as follows:

- We observe that the intrinsic inductive bias of the teacher model matters much more than its accuracy.
- CNNs and INNs with different inductive biases are inclined to learn complementary patterns, while a vision transformer, a more general architecture with fewer inductive biases, can inherit knowledge from both.
- When several teachers with different inductive biases are provided, a student model with less inductive biases is more compatible to learn various knowledge.
- Compared with introducing the inductive bias into the transformer, knowledge distillation makes student transformer performs more similar to various inductive bias teachers.
- Our cross inductive bias vision transformers (CiT) outperform all previous vision transformers of the same architecture and only require super lightweight teachers with 20% and 50% parameters of the teacher in DeiT-Ti and DeiT-S, respectively.

2. Related Works

CNNs. Convolution operator was first proposed in [19] around thirty years ago. Its rejuvenation appears in the past decade, when deep CNNs (e.g., AlexNet [18], VGGNet [26], ResNet [11], EfficientNet [27]) led to an astonishing breakthrough in a great variety of tasks. The remarkable performance of CNNs origins from inherent characteristics (a.k.a. inductive biases) of the convolution operator such as translation equivariance [6] and spatial-agnostic [20]. On the other hand, its locality alternatively makes CNNs struggle to relate spatially-distant concepts, unless we deliberately increase the kernel size and/or model depth.

Transformers. Transformers, which first prevailed in natural language processing [32], has drawn attention in the computer vision community recently. The ViT proposed in [6] feeds 16×16 image patches into a standard transformer, achieving comparable results as SOTA CNNs on JFT-300M [6]. However, its superiority is at the expense of excruciatingly long training time and tremendous amount of labeled data. Most importantly, when insufficient amount of data are given, ViT only achieves modest improvement of accuracy. Furthermore, DETR and VT were proposed in [1] and [35], respectively. DETR [1] exploits bipartite matching loss and a transformer-based encoder-decoder structure in object detection task, while VT [35] represents images as semantic tokens and exploits transformers in image classification and semantic segmentation. Alternatively from theoretical perspective, it has been proven in [3] that the self-attention mechanism used in transformers is at least as expressive as a convolution layer.

INNs. Involution operator was proposed in [20, 33] lately. In a nutshell, convolution operator is spatial-agnostic and channel-specific, while an involution kernel is shared across channels and distinct in the spatial extent. In other words, involution attains precisely inverse inherent characteristics compared to convolution. As a result, it has the ability to relate long-range spatial relationship in an image. It is depicted in [20] that their involution-based RedNet consistently delivers enhanced performances compared with CNNs and transformers.

Knowledge Distillation. Knowledge distillation (KD) was first formulated in [16] as a strategy of model compression, in which a lightweight student is trained from a high-capacity teacher [31, 36]. Specifically, authors in [16] achieve this goal by minimizing the KL divergence of student's and teacher's probabilistic predictions. Afterwards, KD unfolds usefulness in various tasks such as privileged learning [21, 31], cross-modal learning [17, 36], adversarial learning [15, 24], contrastive learning [28], and incremental learning [23]. In relevance to our work, authors in [30] proposed to train transformers via a token-based KD strategy. By distilling from a large-scale and powerful CNN teacher, the resulting DeiT [30] can perform as well as CNNs on ImageNet, while the preceding ViT [6] cannot. Our method outperforms DeiT by distilling from two weak teachers with much fewer parameters, worse accuracy but different inductive bias.

3. Proposed Method

3.1. Cross Inductive Bias Teachers

DeiT [30], where the teacher model is a single convolution-based architecture, is limited by the knowledge of the teacher. A popular idea to go beyond the teacher performance is an ensembling of multiple teachers with different initializations [16]. However, those teachers with the same architecture have the same inductive biases, and consequently offer similar perspectives of data.

Table 1. Performance on ImageNet and Out-of-Distribution dataset of convolution and involution model "A", "R", "C" indicate ImageNet-A, R, C respectively. "mCE" indicate mean corruption error, and for convenience, we do not normalize them with AlexNet.

Model	ImageNet(%)	A (%) $\uparrow$	R(%) $\uparrow$	C(mCE) $\downarrow$
Convolution				
ResNet-18	68.74	2.60	31.90	65.58
ResNet-34	72.62	3.45	35.17	60.26
ResNet-50	75.57	2.60	35.61	59.15
ResNet-101	77.00	6.03	38.77	54.33
ResNet-152	77.96	7.73	40.72	53.18
Involution				
RedNet-26	75.19	5.49	33.33	61.09
RedNet-38	76.88	6.88	34.80	58.15
RedNet-50	77.72	7.64	35.72	56.03
RedNet-101	78.35	9.03	36.30	54.78
RedNet-152	78.54	9.24	36.84	53.58

When teachers have different inductive biases, the output distribution may vary distinctively as the different inductive biases inherently make the model biased towards different patterns. Such variation on output distribution may not be obvious if we use the top-1 accuracy to evaluate. For a better understanding, here we introduce out-of-distribution datasets [12–14] which are generated by applying different perturbations on ImageNet, *e.g.*natural adversarial examples (ImageNet-A), semantic shift (ImageNet-R), common image corruptions (ImageNet-C). As shown in Table 1, when the convolution model (ResNet) and involution model (RedNet) have similar accuracy on ImageNet like ResNet-50 and RedNet-26 or ResNet-101 and RedNet-38, but their performances vary on out-of-distribution dataset. This implies that if we take CNNs and INNs as teachers, CNN teachers will perform better on ImageNet-R/C but worse on ImagNet-A compared with INN teachers. This phenomenon also demonstrates that convolution and involution model may focus on different patterns and will drive different knowledge to the student model. In other words, the knowledge provided by cross inductive bias teachers can describe the data more precisely and comprehensively. In our later experiments, we show that our students will inherit the trend of teachers on out-of-distribution datasets: We match class toekn, Conv token and Inv token to golden labels, RegNet (CNN teacher), and RedNet (INN teacher), respectively. We observe that the Conv token and Inv token will perform similar to the CNN teacher and INN teacher respectively on out-of-distribution datasets.

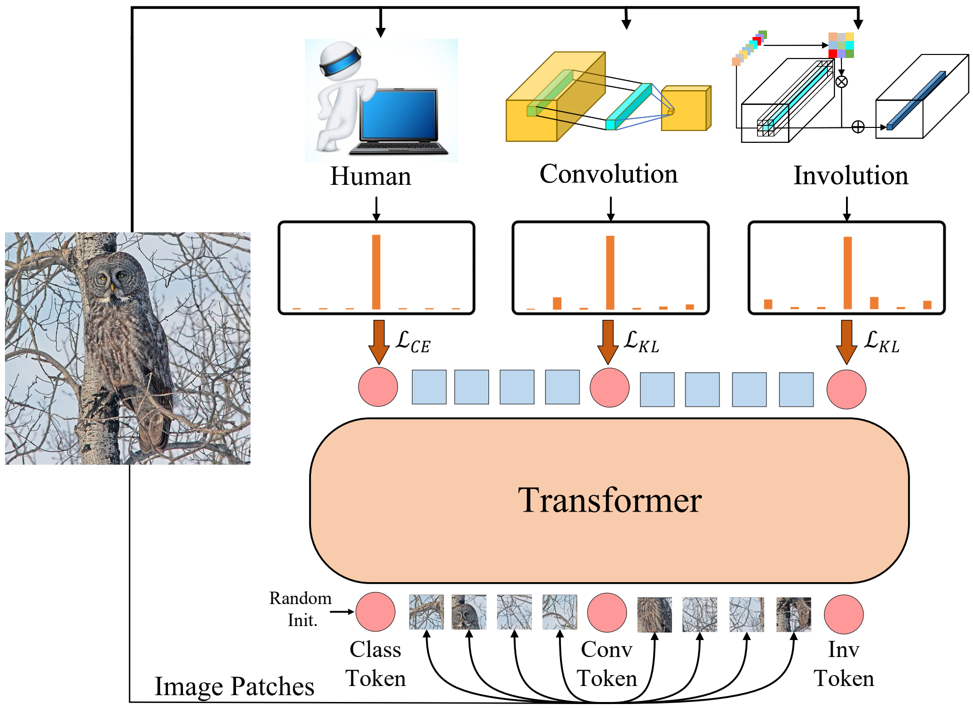


Figure 3. Schematic of our CiT. Given an image as input, human, convolution model and involution model will provide three similar (but slightly different) distributions to describe the image class. Our CiT model inherits the architecture of ViT but has two extra tokens (i.e., Conv token and Inv token) to learn from the convolution and involution teachers, respectively.

3.2. Token Inductive Bias Alignment

Previous works [6, 30] use randomly initialized tokens to learn the label and distill from a CNN teacher. However, a randomly initialized token has limited power to learn a convolution teacher which has very specific inductive bias. To address this issue, we propose token inductive bias alignment, making tokens explicitly possessing different inductive biases so that they could move towards their corresponding teachers. Specifically, we have three kinds of teachers: human (i.e., golden labels), convolution teacher and involution teacher. Therefore, we have three tokens: Class token, Conv token, and Inv token. For the Class token, we simply apply truncated gaussian initialization [30] which makes this token have no inductive bias. To introduce corresponding inductive bias into the remaining two tokens, we combine token generation and patch embedding. Previous methods simply split image to non-overlap patches and use a linear projection to map these patches into tokens. We introduce convolution stem [9, 10] and involution stem to replace the linear projection. Then the Conv token and Inv token are the average pooled output of convolution and involution stem output respectively.

3.3. Cross Inductive Bias Distillation

The schematic of our CiT is demonstrated in Figure 3. Our learning objective is expressed as a weighted summation of two Kullback-Leibler divergence losses ($\mathcal{L}_{KL}$) and a cross-entropy loss ($\mathcal{L}_{CE}$):

$$\begin{aligned}\mathcal{L} =&\lambda_0 \mathcal{L}_{CE}(\sigma(\mathbf{z}_{s_{class}}), \mathbf{y}) \\ &+ \lambda_1\, \tau_1^2\, \mathcal{L}_{KL}[\sigma(\frac{\mathbf{z}_{s_{conv}}}{\tau_1}), \sigma(\frac{\mathbf{z}_{t_1}}{\tau_1})] \\ &+ \lambda_2\, \tau_2^2\, \mathcal{L}_{KL}[\sigma(\frac{\mathbf{z}_{s_{inv}}}{\tau_2}), \sigma(\frac{\mathbf{z}_{t_2}}{\tau_2})],\end{aligned} \tag{1}$$

where $0 < \tau_1, \tau_2 < \infty$ are hyper-parameters controlling the temperature of Softmax function σ [16]. $\mathbf{z}_{s_{class}}$, $\mathbf{z}_{s_{conv}}$, $\mathbf{z}_{s_{inv}}$ are the output of Class token, Conv token and Inv token. $\mathbf{z}_{t_1}$ and $\mathbf{z}_{t_2}$ denote logits of the CNN teacher and INN teacher, respectively. Here $0 \leq \lambda_0, \lambda_1, \lambda_2 \leq 1$ are weights balancing the importance of three loss terms.

4. Experimental Results

In Section 4.1, we describe our implementation details, and next compare our CiT with various transformers, convolution- and involution-based neural networks on ImageNet-1k [5] in Section 4.2. In the rest of this section, experiments are conducted on ImageNet-100 [34]. We analyze impacts of teacher performance and inductive biases to student performance in Section 4.3.1. Then we explain the advantage of choosing a transformer as student over CNNs and INNs in Section 4.3.1. To prove the efficiency of our co-advising strategy, we compare the prediction accuracy

Table 2. Comparison of teacher models used in DeiT [30] and CiT. DeiT uses a much larger and powerful convolution teacher, while CiT uses weak and small involution and convolution teachers.

Student	Teacher		
	Model	Param	Top-1 (%)
DeiT	RegNetY-16GF (Conv)	84M	82.9
CiT-Ti	RegNetY-600M (Conv)	6M	74.0
	RedNet-26 (Inv)	9M	76.0
CiT-S	RegNetY-4GF (Conv)	21M	79.9
	RedNet-101 (Inv)	26M	79.0

of models trained by our cross inductive bias distillation and naive multi-teacher distillation in Section 4.3.3. Finally, we study the influence of the inductive bias alignment on ImageNet and Out-of-Distribution datasets with or without distillation.

4.1. Implementation Details

For comparison purpose, following DeiT [30], we implement two variants of our model: (i) CiT-Ti has two hidden layers with dimensions of 192 and 12, respectively (each with three attention heads), and (ii) CiT-S has two hidden layers with dimensions of 384 and 12, respectively (each with six attention heads). (ii) CiT-SAK is the same as CiT-S except the token inductive bias Alignment. We use the same data augmentation and regularization methods described in DeiT [30] (e.g., Auto-Augment, Rand-Augment, mixup). The weights of our transformers are randomly initialized by sampling from a truncated normal distribution. We use AdamW [22] as optimizer with learning rate equal to 0.001 and weight decay equal to 0.05. For hyper-parameters in distillation, we set $\lambda_0 = \lambda_1 = \lambda_2 = 1$ and $\tau_1 = \tau_2 = 1$. During inference, we retrieve the value stored in the class token as the final output.

4.2. Comparison among Different Architectures

In this section, we compare accuracy of various convolution-, involution-, and transformer-based models on ImageNet-1k [5].

Teacher Model In Table 2, we compare teacher models used in DeiT [30] and our CiT. Different from DeiT, which uses a powerful convolution teacher RegNetY-16GF [25] with 84M parameters and top-1 accuracy of 82.9%, we choose a convolution teacher and an involution teacher who possess similar model sizes as the student transformer. We emphasize that the overall parameters of teacher models used in our CiT are still much fewer than those in DeiT, and that such small teachers significantly speed up the whole training process.

Results We report inference speed, top-1 accuracy of several models in Table 3. Compared with CNNs, when the model size is small (say around 6 million parameters), transformers do not reveal better performances. For instance, RegNet-600MF performs the best with top-1 accuracy equal to 76.0%, while DeiT-Ti, DeiT-Ti-KD, and our CiT-Ti achieve top-1 accuracy of 72.2% (−4.1%), 74.5% (−1.8%), and 75.3% (−1.0%), respectively. Namely, our CiT narrows the gap between the accuracy of CNNs and transformers in this context. When the model size grows, the accuracy of our CiT grows much faster than that of other models, and our CiT-S outperforms all other models at 20 million parameters. The performance of our CiT-S improves 2.6% over RegNet-4GF and 2.9% over RedNet-101.

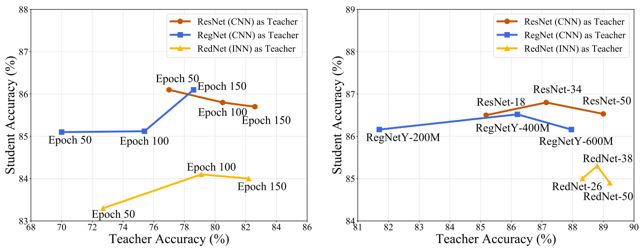
Table 3. Comparisons among different networks on ImageNet-1k [5]. Throughput is measured on a single RTX3090 with batch size of 64. CiT-SAK indicate the small size model with token alignment and knowledge distillation

	Model	Param (M)	Throughput (Images/s)	Top-1 (%)
CNN	ResNet-50 [11]	25.6	1349.4	76.2
	ResNet-101 [11]	44.5	799.4	77.4
	RegNetY-600MF [25]	6.1	1200.5	75.5
	RegNetY-4.0GF [25]	20.6	350.5	79.4
	RegNetY-8.0GF [25]	39.2	220.5	79.9
INN	RedNet-26 [20]	9.2	1820.9	73.6
	RedNet-50 [20]	15.5	1066.8	78.4
	RedNet-101 [20]	25.6	657.4	79.1
	RedNet-152 [20]	34.0	459.3	79.3
Trans.	ViT-B /16 [6]	86	166.88	77.9
	ViT-L /16 [6]	307	54.4	76.5
	DeiT-Ti [30]	5.0	3082.9	72.2
	Deit-S [30]	22	1562.0	79.8
	DeiT-Ti-KD [30]	6.0	3060.8	74.5
	DeiT-S-KD [30]	22	1546.1	81.2
	CiT-Ti (**Ours**)	6.0	3053.0	75.3
	CiT-S (**Ours**)	22	1564.1	82.0
	CiT-SAK (**Ours**)	26	1414.1	82.7

Compared with the recent transformer-based model ViT [6] (i.e., ViT-L /1 and ViT-B /16 in Table 3), our CiT-S requires about 4 times or 15 times fewer model parameters, while at the same time, achieves about 4.1% or 5.5% more accurate predictions. Furthermore, our CiT-S also outperforms the latest work DeiT-KD, even though DeiT-KD has a more potent teacher. Moreover, our CiT achieves similar inference speed as DeiT-KD or even slightly better: CiT-Ti and CiT-S improve 0.4% and 0.8% over the corresponding DeiT-KD of similar sizes. To sum up, the extra convolution and involution tokens boost the performance of student transformer almost without additional computation cost.

4.3. Ablation on Cross Inductive Bias Distillation

In this section, we keep the same Transformer as DeiT and perform all experiments on ImageNet-100.



(a) (b)

Figure 4. Prediction accuracy of Transformer-Ti distilled from different teachers on ImageNet-100. (a) We take ResNet-18, RegNetY-600M and RedNet-26 as teacher, the performances gap of teachers are different at different training epochs, but the students' performances almost keep unchanged from horizontal view. (b) Viewing horizontally reveals that the student's accuracy won't change much even though the teacher accuracy improves. Nevertheless, the vertical view demonstrates that teachers with same accuracy but belong to different kinds (e.g., CNNs or INNs) can yield students with different accuracy.

4.3.1 Teacher Performance and Inductive Biases.

This section delves into the impacts of teacher's performance and inductive biases when distilling to a student transformer. For illustration purpose, we conduct an experiment on student's accuracy when it distills from different kinds of teachers. We take three kinds of teachers into consideration: convolution-based ResNet and RegNet, and involution-based RedNet. We choose CiT-Ti as student. During distillation, either a CNN teacher or an INN teacher (but not both) is provided, and thus one of the three tokens in CiT-Ti will be discarded in this experiment. From now on, this degenerated CiT-Ti will be referred to as Transformer-Ti. The results are reported in Figure 4.

As shown in Figure 4, if the teacher models share similar architecture (i.e., viewing horizontally in both (a) and (b)), the student model retains similar performance even though the teacher performances are boosted. For instance, in Figure 4(a), increasing training epochs leads to performance improvement of teacher models. Training extra 100 epochs helps the RegNet-200M teacher improve 9%, but the performance of the student transformer keeps hardly changed. Similar observation can be generalized to ResNet-18 and RedNet-26 teachers. In Figure 4(b), although the performances increase 6.5% from RegNet-200M to RegNet-600M, the performance of students remains still. This observation implies that the accuracy of the teacher model is not the most important factor determining the student's performance in this context. Namely, we are approaching saturation: when the accuracy of teacher model is sufficiently large, improving teacher accuracy won't result in the improvement of student model.

Table 4. Performances of different students distilled from involution and convolution teachers. When both involution- and convolution-based teachers are provided, Transformer-Ti becomes CiT-Ti.

Student	Teacher		Top-1
	ResNet-18	RegNet-26	(%)
ResNet-10			81.5
ResNet-10	✓		83.0
ResNet-10		✓	82.6
ResNet-10	✓	✓	83.4
Mixer-Ti			80.5
Mixer-Ti	✓		81.6
Mixer-Ti		✓	80.9
Mixer-Ti	✓	✓	82.3
Transformer-Ti			81.8
Transformer-Ti	✓		86.5
Transformer-Ti		✓	85.0
Transformer-Ti (Ours)	✓	✓	88.0

Alternatively, the vertical view of Figure 4 implies that we could resort to a teacher of a different type. For instance, when a teacher has similar performance but belongs to different kinds (*e.g.*, ResNet-18 and RedNet-26 with training 150 epochs in Figure 4(a), ResNet-50 and RedNet-50 in Figure 4(b)), the distilled student could possess relatively different performances. Our hypothesis is that different kinds of teachers have different inductive biases. Even trained on the same dataset, they tend to harvest different knowledge. During distillation, some knowledge might be easier to be understood and inherited by the student model, while others do not. Furthermore, in terms of the student performance, the inherent knowledge of the teacher model seems to weigh more than its accuracy.

4.3.2 Student Performance and Inductive Biases.

When distilling cross inductive knowledge to a student, the student needs to have few inductive biases to avoid overly inclining to a certain teacher. Moreover, the student model needs to have enough capability and model capacity to learn from its teachers. Based on these two considerations, we choose ResNet-10, Transformer-Ti, and Mixer-Ti [29] as students for testing purpose, and ResNet-18, RedNet-26 as teachers. ResNet-10 has stronger inductive biases than Transformer-Ti, and such inductive biases are similar to those of ResNet-18 and conflicts with those of RedNet-26. The results are reported in Table 4.

Our experiment results demonstrate that ResNet-10 distilling from two teachers attains a similar performance to that distilling from a single convolution-based ResNet-18. In contrast, Transformer-Ti can learn from both teachers and achieve higher performance (88%) than distilling from a single teacher. We believe the intrinsic reason is that a

Table 5. The output KL divergence. A smaller value indicates a larger similarity.

Student	ResNet-18	RedNet-26	Top-1 (%)
ResNet-10	0.261	0.274	83.4
Mixer-Ti	0.358	0.313	82.3
CiT-Ti conv token	0.255	0.290	87.1
CiT-Ti inv token	0.254	0.154	87.7

Table 6. Performances of various models on ImageNet-100. A check mark ✓ represents a teacher of the specified type is presented. ✓✓ indicates two architectural identical teachers with different initialization.

Student	Teacher			Top-1
	ResNet-18	ResNet-50	RedNet-26	(%)
ResNet-18				85.1
ResNet-50				89.0
RedNet-26				89.2
Transformer-Ti				81.8
Transformer-Ti	✓			86.5
Transformer-Ti		✓		86.6
Transformer-Ti			✓	85.0
Transformer-Ti	✓✓			87.2
Transformer-Ti	✓	✓		87.0
Transformer-Ti (Ours)	✓		✓	88.0

transformer possesses few inductive biases and the attention layer could not only perform convolution [4], but also has close relationship to involution [20].

This rises a natural question: An MLP possesses the fewest inductive biases, how about choosing it as student? To this end, we include the recent Mixer model [29], a pure multi-layer perceptron (MLP) structure, into comparison. For fairness of comparison, the Mixer-Ti used in our paper has 12 layers, and the hidden dimension is 192. As shown in Table 4, it indicates that without any distillation, Mixer-Ti and Transformer-Ti have similar performances. However, after distilling knowledge from teachers, Transformer-Ti gains more improvement than Mixer. This demonstrates the effectiveness of choosing transformer as a student.

The reason why Mixer-Ti doesn't gain as much as a Transformer through distillation will be clear if we compute the KL divergence between student's and teacher's outputs. As shown in Table 5, all values of KL divergence in Mixer-Ti are much larger than the others. It implies that Mixer-Ti doesn't have the ability to learn from the teacher when its model size is constrained to the same as its Transformer counterpart. On the contrary, compared with other students, CiT-Ti are more similar to teachers. Not surprisingly, the convolution token and involution token are more inclined to convolution and involution teacher, respectively, because our loss function in Eq (1) advocates them to mimic their corresponding teachers.

4.3.3 Naive Multi and Cross Inductive Bias Teachers.

In this section, we verify the effectiveness of our cross inductive bias distillation by comparing it with naive multi-teacher distillation. We implement three teachers: (i) ResNet-18, ResNet-50 are both convolution-based models. They have similar inductive biases, but different performances due to different model sizes. (ii) RedNet-26 is an involution-based model but with similar performance as ResNet-50. The results are illustrated in Table 6.

When Transformer-Ti distills from a single teacher, its performance gain is significant regardless the type of teacher. Specifically, after distilling from the convolution-based ResNet-18, Transformer-Ti can achieve about 86.5% top-1 accuracy on ImageNet-100, while after distilling from the involution-based RedNet-26, its performance gain is relatively moderate: achieving 85.0% top-1 accuracy.

When one more teacher is further allowed in distillation, interesting phenomenon occurs. If both teachers are convolution based (a.k.a. teacher ensembling [8]), the further performance improvement is limited (*e.g.* from 86.5% to 87.0% or 87.2%). In contrast, if we choose the additional teacher as the involution-based RedNet-26, the performance of Transformer-Ti rises to 88.0%. This justifies the effectiveness of providing two different types of teachers.

4.3.4 Effectiveness of Multiple Distillation Tokens.

In conventional knowledge distillation [16], one output token is used to fit the true label and teacher's logits simultaneously. However, such two objectives are sometimes in conflict [2]. As shown in Eq (1), we use different tokens to capture different knowledge provided by different teachers. Specifically, class , convolution and involution token learn from the true label, convolution teacher, and involution teacher, respectively. To evaluate the effectiveness of three tokens, we compare the accuracy of the learned Transformer with that trained via only one or two tokens. The results are reported in Table 7. When the number of tokens is one, distilling from two teachers with different inductive biases can bring considerable improvements, while only distilling from one teacher induces almost no positive result. With the same teachers, merely by increasing from one token to three, our method achieves a 4.5% accuracy improvement.

4.4. Ablation on Token Inductive Bias Alignments

In this section, we evaluate our token inductive bias alignments with or without knowledge distillation on ImageNet-1k and Out-of-Distribution datasets.

Inductive Bias Injection. We aim at align the inductive bias

Table 7. Performances of various models on ImageNet-100. A check mark ✓ represents a teacher of the specified type is presented.

Student		Teacher		Top-1 (%)
Model	Token	ResNet-18	RedNet-26	
Transformer-Ti	1			81.8
Transformer-Ti	1	✓		81.9
Transformer-Ti	1		✓	80.7
Transformer-Ti	1	✓	✓	83.5
Transformer-Ti (Ours)	3	✓	✓	88.0

Table 8. Performances of inductive bias injection on ImageNet-1k. A check mark ✓ represents a kind of inductive biases which are injecting into the transformer.

Model	Inductive Bias		Top-1 (%)
	Convolution	Involution	
Transformer-S			79.8
Transformer-S	✓		81.5
Transformer-S		✓	81.4
Transformer-S (Ours)	✓	✓	81.8

between the teachers and the corresponding tokens with token inductive bias alignments, but we find that simply inject the inductive bias will also brings significant improvements. As show in Table 8, if we inject involution or convolution, the performance will be improved 1.7 % and 1.6 % respectively. When we inject both two kinds of inductive bias, we are pleased to find they are compatible and complementary and can further improve the performance.

Tokens on Out-of-Distribution Dataset. Inductive bias is the set of assumptions predefined in the model, it is hard to say 'transfer' or 'inherent' inductive bias by knowledge distillation without model modification. However, the tokens in our distilled student performs more similar to the corresponding teachers with different inductive bias comparing with simply inject some inductive bias into the model. According to the results on Table 1, convolution perform better on ImageNet-R and C but worse on ImageNet-A comparing with involution when models have similar performance on ImageNet. As shown in Table 9, when we simply inject the inductive bias to the tokens which inherent the inductive bias of teachers but different tokens share same learning targets (Random w/o KD and Align w/o KD), such modification truly brings some differences but is too limited. When the situation goes to knowledge distillations (Random w/o KD and Random w/ KD), there is no inductive bias injected into the student model, but thanks to the different knowledge, the student model perform much similar to the teachers than simply inject the inductive bias. Specifically, convolution teacher perform better than the involution teacher on ImageNet-R

Table 9. Performance of transformer w/ and w/o knowledge distillation and inductive bias alignments on Out-of-Distribution datasets.

Model		ImageNet ↑	A ↑	R↑	C ↓
Random w/o KD	Conv Token	79.80	18.36	42.35	41.36
	Inv Token	79.80	18.35	42.35	41.35
Random w/ KD	Conv Token	81.43	16.18	45.08	39.58
	Inv Token	81.89	18.80	44.43	40.95
Align w/o KD	Conv Token	81.72	24.89	41.88	38.54
	Inv Token	81.74	24.88	41.76	38.56
Align w/ KD	Conv Token	82.11	23.58	47.41	38.11
	Inv Token	82.51	25.15	46.81	38.04

and C but worse on ImageNet-A. The tokens in our student inherent the characteristics and the Conv token perform better than Inv token on ImageNet-R and C but worse on ImageNet-A. Finally, when the knowledge distill and token inductive bias alignments combine together (Random w/o KD and Align w/ KD), our student inherent the characteristics of the teacher best.

5. Conclusion

In this paper, we introduce a cross inductive bias transformer (CiT) by distilling from teacher networks with diverse inductive biases. Compared with distilling from convolution teacher, cross inductive bias teachers provide different perspectives of data and avoid that student is over biased toward single teacher. In our experiments, we find that the teacher inductive biases play a more critical role than the teacher performance in knowledge distillation. Furthermore, we delve into the student model's inductive biases, and the capability of imitating teachers and the transformer shows its superiority in these two aspects comparing with Mixer and ResNet. Finally, we evaluate the effectiveness of token alignment, and prove the distillation help student perform more similar to teachers, and the distillation help student perform best together with the token alignment.

Limitations. We need to independently train our two lightweight teachers, although the total training time is still much less than that of the heavy teacher in DeiT. In theory, our method is compatible with more cross inductive bias teachers. More suitable teachers other than CNNs and INNs will be explored in our future work.

Acknowledgement. This project is supported by the National Natural Science Foundation of China (No. 61972162); Guangdong International Science and Technology Cooperation Project (No. 2021A0505030009); Guangdong Natural Science Foundation (No. 2021A1515012625); Guangzhou Basic and Applied Research Project (No. 202102021074); and CCF-Tencent Open Research fund (RAGR20210114).

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 3

[2] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *ICCV*, pages 4794–4802, 2019. 7

[3] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*, 2019. 3

[4] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *ICLR*, 2020. 7

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 4, 5

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 3, 4, 5

[7] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pages 1607–1616. PMLR, 2018. 1

[8] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pages 1607–1616. PMLR, 2018. 7

[9] Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. *arXiv preprint arXiv:2104.01136*, 2021. 4

[10] Jianyuan Guo, Kai Han, Han Wu, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Cmt: Convolutional neural networks meet vision transformers. *arXiv preprint arXiv:2107.06263*, 2021. 4

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 5

[12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 3

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019. 3

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 3

[15] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge distillation with adversarial samples supporting decision boundary. In *AAAI*, volume 33, pages 3771–3778, 2019. 3

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 3, 4, 7

[17] Judy Hoffman, Saurabh Gupta, Jian Leong, Sergio Guadarrama, and Trevor Darrell. Cross-modal adaptation for rgb-d detection. In *ICRA*, pages 5032–5039, 2016. 3

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25:1097–1105, 2012. 2

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[20] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. *arXiv preprint arXiv:2103.06255*, 2021. 2, 3, 5, 7

[21] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015. 3

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. 5

[23] Umberto Michieli and Pietro Zanuttigh. Knowledge distillation for incremental learning in semantic segmentation. *CVIU*, 205:103167, 2021. 3

[24] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 582–597, 2016. 3

[25] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *CVPR*, 2020. 5

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[27] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019. 2

[28] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2020. 3

[29] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision, 2021. 6, 7

[30] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1, 3, 4, 5

[31] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, pages 1365–1374, 2019. 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3

[33] Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. Carafe: Content-aware reassembly of features. In *ICCV*, October 2019. 3

[34] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. 119:9929–9939, 13–18 Jul 2020. 4

[35] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 3

[36] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. *arXiv preprint arXiv:2103.14431*, 2021. 3