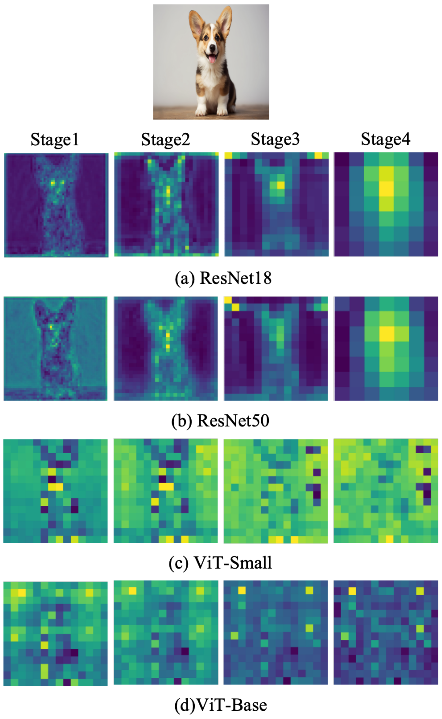
Fuse Before Transfer: Knowledge Fusion for Heterogeneous Distillation

Guopeng Li[1*] Qiang Wang[3] Ke Yan[3] Shouhong Ding[3] Yuan Gao[2†] Gui-Song Xia[2†]
[1]School of Computer Science, [2]School of Artificial Intelligence, Wuhan University
[3]Tencent YouTu Lab
{guopengli, guisong.xia}@whu.edu.cn, ethan.y.gao@gmail.com
{albertqwang, kerwinyan, ericshding}@tencent.com

Abstract

*Most knowledge distillation (KD) methods focus on teacher-student pairs with **similar architectures**, such as both being CNN models. The potential and flexibility of KD can be greatly improved by expanding it to Cross-Architecture KD (CAKD), where the knowledge of homogeneous and **heterogeneous** teachers can be distilled selectively. However, substantial feature gaps between heterogeneous models (*e.g., *ViT teacher *v.s.* CNN student) make CAKD extremely challenging, caused by the distinction of inherent inductive biases and module functions. To this end, we fuse heterogeneous knowledge before transferring it from teacher to student. This fusion combines the advantages of both cross-architecture inductive biases and module functions by merging different combinations of convolution, attention, and MLP modules derived directly from student and teacher module functions. Furthermore, heterogeneous features exhibit diverse spatial distributions, hindering the effectiveness of conventional pixel-wise MSE loss. Therefore, we replace it with a spatial-agnostic InfoNCE loss. Our method is evaluated across various homogeneous models and arbitrary heterogeneous combinations of CNNs, ViTs, and MLPs, yielding promising performance for distilled models with a maximum gain of 11.47% on CIFAR-100 and 3.67% on ImageNet-1K. Code is available at https://github.com/liguopeng0923/FBT.*



(a) ResNet18

(b) ResNet50

(c) ViT-Small

(d)ViT-Base

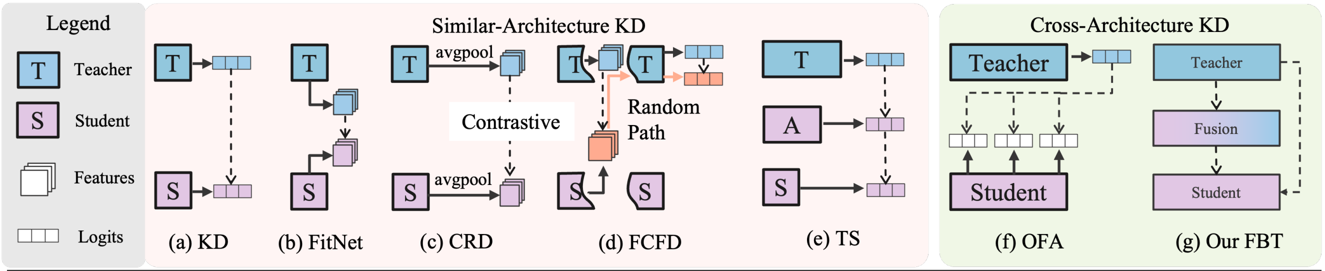
Figure 1. **Significant feature gaps among heterogeneous models.** For example, (a) CNN-based models [8] and (c) ViT-based models [6] have different features in different stages caused by different inductive biases and module functions.

1. Introduction

Knowledge Distillation (KD) [11, 29] has been demonstrated as a powerful method to transfer knowledge from a cumbersome teacher to a compact student. Compared to the model trained from scratch, the performance of distilled students usually improves significantly. Generally, knowledge transferred is derived from output logits (logits-based KD [32]) or intermediate features (feature-based KD [29]) of the teacher model. Therefore, it is intuitive to understand different teachers have different knowledge (logits or features) determined by their unique architectures [18].

Most existing KD methods focus on similar-architecture distillation [17, 29, 33] (called SAKD), *i.e.*, optional teachers are restricted to a limited scope with structures similar to the student model. This presents two principal limitations: **(1) Limited Potential:** Compared to the broader range of arbitrary teachers (including homogeneous and heterogeneous ones), the restricted scope of teachers in

[*] Work done during internship at Tencent YouTu Lab.
[†] indicates the corresponding authors.



	KD [11]	FitNet [29]	CRD [33]	FCFD [17]	TS [22]	OFA [7]	Ours
Knowledge to Distill	Logits	Feature	Feature	Feature	Logits	Logits	**Feature**
Generic	Yes	No	Yes	No	No	Yes	**Yes**
Scheme	T.-S.	T.-S.	T.-S.	T.-S.	T. → A. → S.	T.-S.	**T.-F.-S.**
Heterogeneous Inductive Bias Fusing	No	No	No	No	No	No	**Yes**
Heterogeneous Module Merging	No	No	No	Yes	No	No	**Yes**
Training Cost	**Very Low**	Middle	High	High	Middle	Middle	Low
Loss Function	$\mathcal{L}_{\text{KL}}$	$\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{InfoNCE}}$	$\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{KL}}$	$\mathcal{L}_{\text{OFA}}$	$\mathcal{L}_{\text{OFA}} + \mathcal{L}_{\text{InfoNCE}}$

Figure 2. **The taxonomy of our method.** Our methods are feature-based, generic, and three-level, which fuses heterogeneous inductive biases and module functions with an efficient fused model. Target-wise $\mathcal{L}_{\text{OFA}}$ and spatial-agnostic $\mathcal{L}_{\text{InfoNCE}}$ are more suitable for CAKD than $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{MSE}}$. To the best of our knowledge, our FBT is one of the pioneer works in feature-based generic distillation.

SAKD may fail to include the optimal knowledge. For instance, as OFA [7] demonstrated, distilling knowledge from a heterogeneous ViT-Base to ResNet50 yields better performance compared to using a ResNet152 as the homogeneous teacher. **(2) Limited Flexibility:** The emergence of new models [21, 34] or the scarcity of homogeneous teachers in domain-specific tasks [14, 30] poses significant challenges in obtaining suitable homogeneous teachers, thereby impeding the applicability of SAKD. Thus, this paper tends to expand KD to cross-architecture KD (CAKD), broadening the scope of optional teachers and thus improving the potential and flexibility of KD [1, 11, 18].

In CAKD, the main challenge is that heterogeneous teachers and students have significant representation gaps, detailed in CKA analyses of OFA [7] and feature maps in Fig. 1. These gaps stem from inherent differences in inductive biases [27] and module functions [17]. **(1) Inductive biases:** As demonstrated in [23, 27], convolutional neural network-based models (CNNs) [8, 31] exhibit locality and translation-equivariance, while multi-head-self-attention-based models (MSAs) [6] and multilayer-perception-based models (MLPs) [34] depend on patchify and long-distance dependency. Consequently, CNN-generated features are located in local objects in Fig. 1 (a,b), but most MSA/MLP models generate global features in Fig. 1 (c,d). **(2) Module functions:** Varied module functions generate different features at different stages. For instance, features of shallow and deep layers in ViT have higher similarity than hierarchical CNN [23, 27] in Fig. 1(a,c).

To alleviate feature gaps in CAKD, as shown in Fig. 2 (g), we **F**use heterogeneous modules **B**efore **T**ransferring (FBT) by merging sequentially CNN/MSA/MLP modules derived from teachers and students. The archiercture of fused model is adaptive according to distilled model pairs.

Our FBT is well-motivated by the following popular beliefs: **(1) How do we fuse heterogeneous inductive biases?** As demonstrated in [15, 16, 23], CNNs and MSAs/MLPs are complementary. A fused model that uses CNNs in the early stages and MSAs/MLPs in the later stages can benefit from both local and global inductive biases. Compared to existing KD like Fig. 2 (a-f), our fused model in Fig. 2 (g) merges CNN and MSA/MLP modules, thereby reducing distillation gaps attributed to inductive biases. **(2) How do we fuse heterogeneous module functions?** As demonstrated in [3, 17], the disparity between heterogeneous features is also from different module functions, *i.e.*, how the models will read, decode, and process the inputs. Therefore, a fused model comprising student and teacher modules not only optimizes the functional similarity [17] between heterogeneous T.-S. pairs[1], but also introduces minimal additional learnable parameters. **(3) How do we align heterogeneous features spatially?** Widely used MSE loss aligns the pixel-wise features, which is inadequate for spatially diverse heterogeneous features. For example, (a) and (c) in Fig. 1 show distinct spatial distributions at four stages, which are hard to align pixel by pixel. To address this, we apply average pooling to smooth the spatial of features and utilize a spatial-agnostic loss [9] to align heterogeneous features.

In view of the above analysis, the taxonomy of our FBT is illustrated in Fig. 2. Our FBT falls under the category of *feature-based methods* for *generic distillation* with an *adaptive fusion* scheme. In our experiments, the proposed FBT greatly enhances the performance of student models in both CAKD and SAKD, achieving a maximum gain of 11.47% on the CIFAR100 and 3.67% on the ImageNet-1K.

[1] This paper shorten the teacher, fusion, and student by T., F., and S.

2. Releated Work

2.1. Taxonomy of our methods

As shown in Fig. 2, the majority of existing KD methodologies concentrate on homogeneous distillation by using a single projector (*e.g.*, single linear layer) to align the output logits [11, 12, 32], intermediate features [3, 17, 29], feature embeddings [33], and module functions [17] of T.-S. pairs, thanks to the highly-similar features between homogeneous T.-S. pairs. However, they fall short in addressing the complexities of heterogeneous distillation, where the distinct features between heterogeneous T.-S. pairs pose significant challenges. Although OFA [7] and [19] achieve consistent improvements for arbitrary T.-S. pairs, it does so at the expense of sacrificing feature information.

Additionally, several other works are pertinent to our method: (1) We note that some methods attempt to distill the knowledge between CNNs and MSAs [39], but they are tailored to specific T.-S. pairs rendering them impractical for our arbitrary T.-S. CAKD. (2) Certain methods apply progressive distillation to transfer the knowledge via a middle model [1, 18, 22], but they are progressive training strategies that are not training algorithms designed for transferring knowledge between heterogeneous T.-S. pairs. (3) While some logits-based methods can be easily applied to CAKD [7, 11, 32], they are suboptimal because they overlook the significance of feature-based knowledge [29]. (4) Some works [19, 37] attempt to align the heterogeneous features, but they also ignore the basic heterogeneous gaps in different inductive bias and module functions.

In this paper, our method dives into the nature of heterogeneous feature gaps (*i.e.*, caused by inductive bias and module functions) and introduces a simple fusion strategy to facilitate smoother feature transfer between heterogeneous T.-S. pairs. We hope this adaptive knowledge fusion strategy motivates more work in heterogeneous fusion.

2.2. Hybrid Model

As illustrated in Fig. 1, different models exhibit different features caused by different inductive biases and module functions. [27] investigates the internal representation structures of ViT and CNN models, revealing significant differences between their heterogeneous features. [23] further provides some fundamental explanations for this phenomenon. Specifically, CNNs are data-agnostic and channel-specific high-pass filters, while MSAs are data-specific and channel-agnostic low-pass filters. Therefore, researchers [23] think CNNs and MSAs are complementary, which inspires them to design a hybrid model following the rules of "alternately replacing CNN blocks with MSA blocks from the end of a baseline CNN model". The hybrid model outperforms CNNs in both large and small data regimes [23]. Furthermore, the architecture of MLP models [34] is notably similar to ViTs not CNNs, so ConvMLP [15] also achieves advanced performance in basic visual tasks by the co-design of CNNs and MLPs. In a nutshell, hybrid CNN-MSA/MLP models improve performance and efficiency through the combination of different inductive biases and module functions.

Inspired by the design of the hybrid model, we mitigate the heterogeneous feature gaps by fusing heterogeneous knowledge between cross-architecture T.-S. pairs.

3. Method

3.1. Preliminaries

Existing KD methods perform well in homogeneous distillation, but they may fail in heterogeneous teachers and students. The primary reasons stem from fundamentally distinct feature and logit spaces, caused by different *inductive biases* and *module functions* of heterogeneous models.

Inductive Bias and module functions. Inductive bias refers to the set of assumptions that a model uses to make predictions on unseen data [28]. Module functions describe how a model reads, encodes, decodes, and processes the data [17]. Heterogeneous models exhibit different inductive biases and module functions. (1) CNN models [8] slide a set of learnable local kernels across the pixel-level image, focusing on local receptive fields. The weight-sharing kernels are applied across the entire image, providing the network with translation-equivariance to recognize an object regardless of location. (2) MSA models [6] split the input image into patches, and attention modules calculate the scores between the Query and Key to generate attention maps. This process, capturing long-distance dependency, allows the model to consider the global information from all patches. (3) MLP models [36] also begin by dividing the input image into patches. It then mixes global information along all patches' spatial and channel dimensions. In a nutshell, different inductive biases and module functions determine different distributions of generated features.

3.2. Adaptive Knowledge Fusion

As shown in Fig. 2, most methods usually apply a two-level paradigm in SAKD [11, 17, 29], *i.e.*, T.-S. scheme, to transfer directly the knowledge of teachers to students. Besides, some works apply progressive training strategy [18, 22] to transfer multi-teacher knowledge in SAKD. However, existing works [7, 11, 19, 37] fall into distillation with a common nature and ignore the unique requirements for specific model pairs, particularly for heterogeneous distillation with significant gaps. A natural question arises: *Can we design a common principle to satisfy varied requirements according to different model pairs?*

Motivated by module connections in FCFD [17], this paper introduces a fusion strategy (called fuse before transfer,

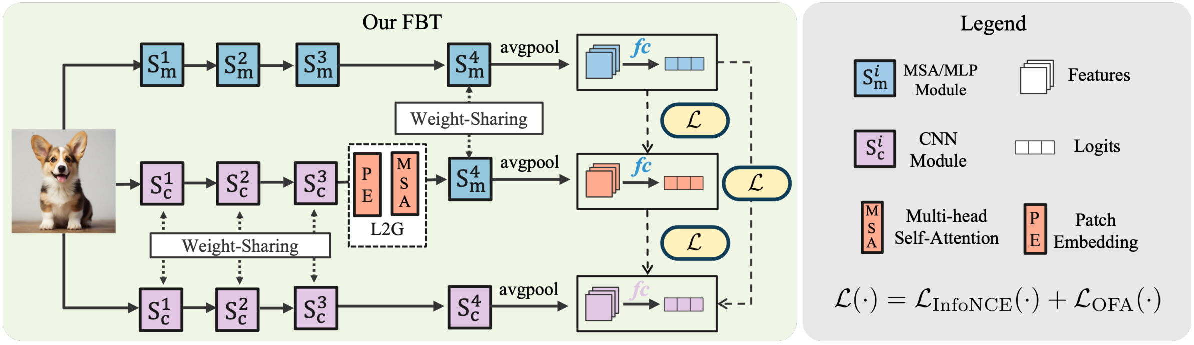


Figure 3. **Overall.** Firstly, our FBT fuses heterogeneous knowledge by merging the first three stages of CNNs, a projector L2G, and the last stage of MSAs/MLPs into the fused model. The fused model is adaptive and can be adjusted automatically according to different T.-S. pairs. Secondly, supervised by spatial-agnostic INFONCE loss [9] and target-wise OFA loss [7], the knowledge is transferred from the teacher to the fused model and student. All models are split into four stages following [7].

FBT), which obeys a common fusion principle but generates different fused models for different T.-S. pairs. Specifically, the knowledge is first fused by merging directly convolution and attention modules derived from both student and teacher module functions and then transferred by training a teacher-fusion-student scheme as follows:

$$\mathcal{L}_{\text{FBT}} = \mathcal{L}(\text{K}_\text{t}, \text{K}_\text{s}) + \mathcal{L}(\text{K}_\text{t}, \text{K}_\text{f}) + \mathcal{L}(\text{K}_\text{f}, \text{K}_\text{s}), \quad (1)$$

where $\mathcal{L}$ is our loss (details in Eq. (3)). K_t, K_f, and K_s denote the knowledge of the T., F., and S. model respectively.

Fusion Strategy. Our fusion connects the CNN modules and MSA/MLP modules derived from the students and teachers with a local-to-global (L2G) feature projector as shown in Fig. 3. Formulary, the logits output of our fusion can be described as follows:

$$p_\text{f}(x) = fc_\text{m} \circ \text{S}_\text{m}^4 \circ (\text{MSA} \circ \text{PE}) \circ \text{S}_\text{c}^3 \circ \text{S}_\text{c}^2 \circ \text{S}_\text{c}^1(x), \quad (2)$$

where x is the input image, S_c^i denotes CNN models, S_m^i denotes MSAs/MLPs, and fc_m denotes the fully-connected layers of MSAs/MLPs. To connect CNN and MSA/MLP modules, we propose an L2G module that includes a patch embedding [6] to convert the features into the required dimensions of subsequent MSA/MLP modules. Besides, to capture the long-distance dependency, our L2G also includes an MSA block to project the local features from CNN models to global receptive fields. For simplicity, the MSA module is a Swin block [20] in this paper. **Note that the fusion is also CNN-MSA/MLP when the teachers are CNNs and student-teacher when the model pairs are homogeneous** (details in our Appendix).

The following considerations drive the design of our fusion: (1) Inductive biases of CNNs and MSAs/MLPs are complementary and hybrid CNN-MSA/MLP models demonstrate good performance [15, 23, 27]. Therefore, we replace the CNN blocks at the end of a baseline CNN model with MSA/MLP blocks which can benefit from both the local features in the early stages and the global information exchanges in the last stages. For example, the final feature appearances of the fusion are converted from the local CNN features to global receptive fields in Fig. 4. (2) Different models have different module functions, which can be aligned implictly by connecting different module functions in a single pipeline [17]. Therefore, we form our fusion by using CNN/MSA/MLP modules derived from students and teachers. As shown in Fig. 3, our fusion is mainly composed of weight-sharing modules $\text{S}_\text{c}^{1\to3}$ and S_m^4, which not only unifies different module functions but also introduces negligible additional learnable parameters. Besides, weight-sharing modules also ensure the bridge role of our fusion model in Tab. 6. (3) While alternately using CNN and MSA modules can improve the performance of the fusion [23], we only add one stage of MSA/MLP models following the first three stages of CNN models as illustrated in Eq. (2), which keeps both simplicity and adaption for different model pairs (details discussed in Sec. 4.4).

3.3. Spatial-Agnostic Knowledge Supervision.

As shown in Fig. 3, we only transfer the final features after average pooling and the logits for the following reasons. (1) Due to the weight-sharing between our fused model and T.-S., it combines actually different inductive biases only in the final features, not early and middle features. (2) As shown in Fig. 4 and Fig. 1, the final features of different models are also very different in spatial, so we smooth them by average pooling to mitigate the spatial gaps. The knowledge in Eq. (1) is formulated by $\text{K}_i = \{f_i, p_i\}, i = \text{t}, \text{f}, \text{s}$, where f_i and p_i denote the final features embeddings after average pooling and the output logits.

In this paper, we use spatial-agnostic InfoNCE loss $\mathcal{L}_{\text{InfoNCE}}$ [9, 33] and OFA loss $\mathcal{L}_{\text{OFA}}$ to supervise the transfer of features and logits respectively, motivated by the fol-

Teacher	Student	From Scratch		feature-based [24, 25, 29, 33]				logits-based [11, 12, 38]			CAKD	
		Teacher	Student	FitNet	CC	RKD	CRD	KD	DKD	DIST	OFA	**our FBT**
CNN-based students												
Swin-T	ResNet18	89.26	74.01	78.87	74.19	74.11	77.63	78.74	80.26	77.75	<u>80.54</u>	**81.61**
ViT-S	ResNet18	92.04	74.01	77.71	74.26	73.72	76.60	77.26	78.10	76.49	<u>80.15</u>	**81.93**
Mixer-B/16	ResNet18	87.29	74.01	77.15	74.26	73.75	76.42	77.79	78.67	76.36	<u>79.39</u>	**81.90**
Swin-T	MobileNetV2	89.26	73.68	74.28	71.19	69.00	79.80	74.68	71.07	72.89	<u>80.98</u>	**81.28**
ViT-S	MobileNetV2	92.04	73.68	73.54	70.67	68.46	78.14	72.77	69.80	72.54	<u>78.45</u>	**82.10**
Mixer-B/16	MobileNetV2	87.29	73.68	73.78	70.73	68.95	78.15	73.33	70.20	73.26	<u>78.78</u>	**80.83**
MSA-based students												
ConvNeXt-T	DeiT-T	88.41	68.00	60.78	68.01	69.79	65.94	72.99	74.60	73.55	<u>75.76</u>	**79.57**
Mixer-B/16	DeiT-T	87.29	68.00	71.05	68.13	69.89	65.35	71.36	73.44	71.67	<u>73.90</u>	**74.40**
ConvNeXt-T	Swin-P	88.41	72.63	24.06	72.63	71.73	67.09	76.44	76.8	76.41	<u>78.32</u>	**80.73**
Mixer-B/16	Swin-P	87.29	72.63	75.2	73.32	70.82	67.03	75.93	76.39	75.85	<u>76.65</u>	**78.44**
MLP-based students												
ConvNeXt-t	ResMLP-S12	88.41	66.56	45.47	67.70	65.82	63.35	72.25	73.22	71.93	<u>75.21</u>	**78.03**
Swin-T	ResMLP-S12	89.26	66.56	63.12	68.37	64.66	61.72	71.89	72.82	11.05	<u>73.58</u>	**77.20**
Average Improvements				−5.21	−0.33	−1.39	−0.02	+3.12	+3.16	−2.31	<u>+6.19</u>	**+8.38**

Table 1. **Top-1 accuracy (%) on CIFAR100.** All baseline results are from the paper or code of OFA [7]. Swin-P is a modified version of Swin-T[20] from OFA [7]. **Bold** denotes the best results and the second-best results are underlined.

lowing observations. (1) Wildly used MSE loss computes the pixel-wise metrics that are suitable for features having similar spatial information, but it will fail when the features are very spatially different (*e.g.*, FitNet with MSE loss gets only 24.06% top-1 accuracy when the teacher is ConvNeXt-T and the student is Swin-P in CIFAR100 Tab. 1). Consequently, we use a spatial-agnostic loss InfoNCE [9, 33] to transfer the structural information of feature embeddings [33], which captures complex interdependencies of features without spatial information. (2) As demonstrated in [7], the different inductive bias leads models to variant logit spaces. Therefore, $\mathcal{L}_{\mathrm{OFA}}$ enhances the information of the target class by adding a modulating parameter γ to the original KD loss, which prevents the student from learning incorrect information of the teacher. In a nutshell, our loss $\mathcal{L}$ is suitable for any representation distillation (*e.g.*, the superior consistently performance in Tab. 1, Tab. 2 and Tab. 3):

$$\left\{\begin{aligned} &\mathcal{L}_{\mathrm{OFA}}(p_{\mathrm{t}}, p_{\mathrm{s}}) = (1+p_{\mathrm{t}}^{\hat{c}})^{\gamma}\log(\frac{p_{\mathrm{t}}^{\hat{c}}}{p_{\mathrm{s}}^{\hat{c}}}) + \sum_{i=1, i\neq\hat{c}}^{C} p_{\mathrm{t}}^{c}\log(\frac{p_{\mathrm{t}}^{c}}{p_{\mathrm{s}}^{c}}), \\ &\mathcal{L}_{\mathrm{InfoNCE}}(f_{\mathrm{s}}, f_{\mathrm{t}}) = -\log\frac{\exp(f_{\mathrm{s}}\cdot f_{\mathrm{t}}^{+}/\tau_2)}{\sum_{i=0}^{F_{\mathrm{t}}}\exp(f_{\mathrm{s}}\cdot f_{\mathrm{t}}^{i}/\tau_2)}, \end{aligned}\right. \tag{3}$$

For each T.-S. pair, we transfer knowledge by $\mathcal{L}(\mathrm{K_t}, \mathrm{K_s}) = \mathcal{L}_{\mathrm{InfoNCE}}(f_{\mathrm{t}}, f_{\mathrm{s}}) + \mathcal{L}_{\mathrm{OFA}}(p_{\mathrm{t}}, p_{\mathrm{s}})$. Firstly, for $\mathcal{L}_{\mathrm{OFA}}$, the $\hat{c}$ and c denote the target class and predicted class of the input image. C is the all classes in the dataset. $\mathcal{L}_{\mathrm{OFA}}$ add a modulating parameter γ to enhance the target information when the teacher is not confident about the prediction. Secondly, for $\mathcal{L}_{\mathrm{InfoNCE}}$, f_{s} denotes an encoded student features by average pooling, and F_{t} is a set of encoded teacher features in a mini-batch. In F_{t}, only one positive sample f_{t}^{+} matches to f_{s}, *i.e.*, the student's and teacher's feature from the same image is a positive pair. The InfoNCE loss is low when the features of student f_{s} and teacher f_{t}^{+} are from the same image and high otherwise. This loss has been widely demonstrated for aligning different feature representations [9, 33].The temperature parameter τ_2 is learnable [26]. Lastly, the entire loss is $\mathcal{L}_{\mathrm{FBT}} = \mathcal{L}(\mathrm{K_t}, \mathrm{K_s}) + \mathcal{L}(\mathrm{K_t}, \mathrm{K_f}) + \mathcal{L}(\mathrm{K_f}, \mathrm{K_s})$, where the formulas of $\mathcal{L}(\mathrm{K_t}, \mathrm{K_f})$ and $\mathcal{L}(\mathrm{K_f}, \mathrm{K_s})$ is like $\mathcal{L}(\mathrm{K_t}, \mathrm{K_s})$.

4. Experiments

4.1. Implementary Details

Models. For a fair comparison, we evaluate our FBT using the same teacher-student pairs employed in OFA[7], including homogeneous distillation and heterogeneous combinations of CNNs, MSAs, and MLPs. Specifically, CNN models include ResNet [8], MobileNetv2 [31], and ConvNeXt [21]. MSA models cover ViT, DeiT [6, 35], and Swin [20], while MLP models consist of MLP-Mixer [34] and ResMLP [36].

Datasets. We use the CIFAR100 [13] and ImageNet-1K dataset [5] for evaluation. CIFAR100 consists of 50K training samples and 10K testing samples in a resolution of 32×32, while the ImageNet-1K dataset contains 1.2 million training samples and 50K validation samples with a resolution of 224×224. Since MSAs and MLPs accept image patches as input, we upsample the images in CIFAR100 to the resolution of 224×224 [7].

Baselines. In line with OFA [7], we choose several powerful KD methods as our baselines for comparison. Specifically, the feature-based methods include FitNet [29],

Teacher	Student	From Scratch		feature-based [24, 25, 29, 33]				logits-based [11, 12, 38]			CAKD	
		Teacher	Student	FitNet	CC	RKD	CRD	KD	DKD	DIST	OFA	**our FBT**
CNN-based models												
DeiT-T	ResNet18	72.17	69.75	70.44	69.77	69.47	69.25	70.22	69.39	70.64	71.01	**71.22**
Swin-T	ResNet18	81.38	69.75	71.18	70.07	68.89	69.09	71.14	71.10	70.91	71.76	**72.21**
Mixer-B/16	ResNet18	76.62	69.75	70.78	70.05	69.46	68.4	70.89	69.89	70.66	71.38	**71.44**
DeiT-T	MobileNetV2	72.17	68.87	70.95	70.69	69.72	69.6	70.87	70.14	71.08	71.39	**71.78**
Swin-T	MobileNetV2	81.38	68.87	71.75	70.69	67.52	69.58	72.05	71.71	71.76	72.32	**72.54**
Mixer-B/16	MobileNetV2	76.62	68.87	71.59	70.79	69.86	68.89	71.92	70.93	71.74	72.12	**72.31**
MSA-based Models												
ResNet50	DeiT-T	80.38	72.17	**75.84**	72.56	72.06	68.53	75.10	75.6	75.13	75.73	75.64
ConvNeXt-T	DeiT-T	82.05	72.17	70.45	73.12	71.47	69.18	74.00	73.95	74.07	74.41	**75.26**
Mixer-B/16	DeiT-T	76.62	72.17	74.38	72.82	72.24	68.23	74.16	72.82	74.22	74.46	**75.00**
ResNet50	Swin-N	80.38	75.53	76.83	76.05	75.90	73.90	77.58	76.24	77.29	77.76	**77.79**
ConvNeXt-T	Swin-N	82.05	75.53	74.81	75.79	75.48	74.15	77.15	77.00	77.25	77.5	**77.73**
Mixer-B/16	Swin-N	76.62	75.53	76.17	75.81	75.52	73.38	76.26	75.03	76.54	76.63	**76.87**
MLP-based models												
ConvNeXt-T	ResMLP-S12	82.05	76.65	74.69	75.79	75.28	73.57	76.87	77.23	77.24	77.26	**77.33**
Swin-T	ResMLP-S12	81.38	76.65	76.48	76.15	75.1	73.4	76.67	76.99	77.25	77.31	**77.42**
Average Improvements				+1.00	+0.56	−0.30	−1.65	+1.61	+1.05	+1.65	+2.05	+**2.31**

Table 2. **Top-1 accuracy (%) on ImageNet-1K.** All baseline results are from the paper or code of OFA [7]. Swin-N is a modified version of Swin-T[20] from OFA [7]. **Bold** denotes the best results, and the second-best results are underlined.

	T.	S.	AT [2]	OFD [10]	CRD [33]	Review [3]	DKD[38]	DIST [12]	FCFD [17]	OFA [7]	Ours
(a)	73.31	70.66	70.69	70.81	71.17	71.61	71.70	72.07	72.24	72.10	**72.29**
(b)	76.61	68.58	69.56	71.25	71.37	72.56	72.05	73.24	73.37	73.28	**73.45**

Table 3. **Results in SAKD on ImageNet-1K.** The teacher and student are ResNet34 and ResNet18 in (a) and are ResNet50 and MobileNet in (b). As shown, our FBT method is still competitive in homogeneous distillation.

CC [25], RKD [24], and CRD [33], while the logits-based methods comprise KD [11], DKD [38], and DIST [12]. Originally, these methods were designed for SAKD, and thus OFA made some modifications to effectively apply them to CAKD scenarios. Although we are also relevant to [19, 37], they are not open-sourced and are hard to reproduce in the same experiments.

Training Protocols. Following the OFA [7], we utilize SGD optimizer for CNN-based students and AdamW optimizer for MSA- and MLP-based students. All models are trained for 300 epochs on the CIFAR100 dataset. As for the ImageNet-1K dataset, CNNs and MSA/MLP models are trained for 100 epochs and 300 epochs respectively. More details about training schedules are in Appendix A.

4.2. Main Results

Given extensive cross-architecture teacher-student model pairs, our FBT consistently achieves the best or most competitive performance on the CIFAR100 (+8.38% on average Top-1 accuracy) dataset and the ImageNet-1K dataset (+2.31% on average Top-1 accuracy).

Results on CIFAR100. To evaluate the performance in enough cross-architecture situations, as shown in Tab. 1, we conduct extensive experiments in 12 combinations of heterogeneous T.-S. models. We have the following important observations in this small-scale dataset.

Firstly, feature-based methods exhibit inferior performance on most occasions, *e.g.*, they have negative performance on average improvements, especially when facing the MSA/MLP student models. The reason is that, as discussed in Sec. 3.1, features of cross-architecture models are distinct because of different inductive bias and module functions, in which a naive feature projector struggles to address this dilemma in the small-scale datasets.

Secondly, FitNet [29] shows very poor performance when the teacher is ConvNeXt-T and the student is Swin-P, while the other feature-based methods obtain relatively normal performance. We believe that this limitation of FitNet stems from its use of the MSE loss to align features in a pixel-wise manner, while other methods solely transfer knowledge from the final feature embeddings or logits. In other words, applying pixel-wise MSE loss may not be suitable for spatially diverse feature maps of student and teacher models, as illustrated in Fig. 1 and Appendix E.

Lastly, OFA [7] yields significant and consistent improvements under all settings. However, these improvements come at the expense of structural feature information by projecting features to logits space. In contrast,

Methods	CIFAR100 [13]				ImageNet [5]			
	T.	S.	T.	S.	T.	S.	T.	S.
	Swin-T	ResNet18	ConvNeXt-T	Swin-P	Swin-T	ResNet18	ResNet50	DeiT-T
KD (Baseline)	78.74(-2.87)		76.44(-4.29)		71.14(-1.04)		75.10(-0.54)	
The architecture of fused model								
(A) w/o MSA and S_m^4 in Eq. (2)	75.95(-5.66)		77.65(-3.18)		70.86(-1.35)		74.67(-0.97)	
(B) w/o S_m^4 in Eq. (2)	77.21(-4.40)		77.84(-2.89)		71.78(-0.43)		75.14(-0.50)	
Loss functions								
(C) w/o $\mathcal{L}(K_t, K_f)$ in Eq. (1)	25.57(-56.04)		50.46(-30.27)		71.34(-0.87)		74.56(-1.08)	
(D) w/o $\mathcal{L}(K_f, K_s)$ in Eq. (1)	79.01(-2.60)		79.82(-0.91)		71.46(-0.75)		74.81(-0.83)	
(E) w/o $\mathcal{L}(K_t, K_s)$ in Eq. (1)	79.26(-2.35)		80.17(-0.56)		71.45(-0.76)		73.92(-1.72)	
(F) w/o $\mathcal{L}_{\text{InfoNCE}}$ in Eq. (3)	80.95(-0.66)		78.89(-1.84)		71.47(-0.74)		75.21(-0.43)	
(G) w/o $\mathcal{L}_{\text{OFA}}$ in Eq. (3)	77.91(-3.70)		80.32(-0.41)		70.37(-1.84)		72.13(-3.51)	
Ours	81.61		80.73		72.21		75.64	

Table 4. **Ablation study.** We evaluate the performance by removing some important components of our FBT and loss functions.

our framework bridges the cross-architecture representation gaps via a fused model and spatial-agnostic loss applied to spatial-smoothed features. Leveraging the two designs, our FBT achieves the best results in all T.-S. pairs in CAKD, obtaining an average gain of about 2.06% compared to the recent SOTA method OFA [7] on CIFAR100.

Results on ImageNet-1K. We also conduct extensive experiments on 14 combinations of cross-architecture T.-S. models on the large-scale ImageNet-1K dataset. Here, we observe that feature-based methods perform well when handling MSA/MLP students, for instance, distillation with FitNet [29] when the teacher model is ResNet50 and the student model is DeiT-T. This is opposite to our observations on CIFAR100. We argue that this discrepancy arises due to the data-hungry nature of MLP/MSA models and additional linear feature projectors [23], which are better suited for training on large-scale datasets. Even so, traditional feature-based methods still have negative impacts in some other situations, *e.g.*, FitNet yields 70.45% (-1.72%) when the teacher is ConvNeXt-T and the student is DeiT-T. In a nutshell, even in training with large-scale data, simple feature projectors are not sufficient to align features of cross-architecture T.-S. pairs.

In this paper, besides the feature projectors L2G, our fused model also includes modules derived from students and frozen teachers. In other words, our fused model achieves a more important task, *i.e.*, aligning the knowledge of student functions to match the frozen teacher functions. Therefore, our FBT leads to superior and stable performance on extensive combinations of cross-architecture models in the small-scale and large-scale dataset. Besides, compared to leveraging four intermediate features of SOTA [7], our FBT achieves more competitive performance by only leveraging the final features.

Results in SAKD. As shown in Tab. 3, we compare the distilled results of AT [2], OFD [10], CRD [33], Review [3], DKD [38], DIST [12], FCFD [17] and OFA [7] on ImageNet-1k dataset. Compared to the recent works in homogeneous distillation (FCFD [17]) and heterogeneous distillation (OFA [7]), our FBT has a competitive performance. This is because our fusion strategy is also beneficial for aligning different modules in homogeneous distillation.

4.3. Ablation Study

Knowledge Fusion. In Tab. 4 (A-B), we remove the module S_m^4 in (B) and the MSA module in (A), the performance of different teacher-student pairs drops significantly. This demonstrates the power of fusing the inductive biases by adding MSA modules following the CNN modules and fusing the module functions by adding S_m^4. Besides, different MSA blocks have different functions for different T.-S. pairs (details in Appendix F), so we use the Swin block as our MSA block in L2G for simplicity.

Knowledge Transfer. (1) In Tab. 4 (C), we remove the transfer from the teacher to fused model, *i.e.*, $\mathcal{L}(K_t, K_f)$, which makes the fused model learn no correct knowledge and then transfers the incorrect knowledge to the students. So the distilled students have poor performance. (2) We remove the transfer from the fused model to students in Tab. 4 (D), *i.e.*, $\mathcal{L}(K_f, K_s)$. Due to the gaps between heterogeneous students and teachers that are not mitigated without our fusion, the final performance of students is poor too. (3) We remove the transfer from the teacher model to students in Tab. 4 (E), *i.e.*, $\mathcal{L}(K_t, K_s)$. In this case, although the performance of distilled students is good in some situations, it is not optimal compared to our FBT. This is because our fusion inevitably damages some knowledge from the teachers, and some easy knowledge is more suitable to transfer directly by the T.-S. scheme without a middle bridge. Totally, the proposed fusion strategy and knowledge transfer path is powerful for heterogeneous distillation.

Knowledge Supervision. In Tab. 4, we remove the feature loss $\mathcal{L}_{\text{InfoNCE}}$ (F) and the logit loss $\mathcal{L}_{\text{OFA}}$ (G), and different T.-S. pairs have different performance drops on CIFAR100 and ImageNet-1K. For instance, $\mathcal{L}_{\text{OFA}}$ is more important when the teacher is Swin-T and the student is ResNet18

Teacher	fused models with the same length			The same CNN modules	
(A): ViT-S	$\mathrm{S}_{\mathrm{c}}^{1} \rightarrow \mathrm{S}_{\mathrm{m}}^{2\rightarrow fc}$	$\mathrm{S}_{\mathrm{c}}^{1\rightarrow 2} \rightarrow \mathrm{S}_{\mathrm{m}}^{3\rightarrow fc}$	$\mathrm{S}_{\mathrm{c}}^{1\rightarrow 4} \rightarrow \mathrm{S}_{\mathrm{m}}^{fc}$	$\mathrm{S}_{\mathrm{c}}^{1\rightarrow 3} \rightarrow \mathrm{S}_{\mathrm{m}}^{2\rightarrow fc}$	$\mathrm{S}_{\mathrm{c}}^{1\rightarrow 3} \rightarrow \mathrm{S}_{\mathrm{m}}^{3\rightarrow fc}$
	81.5 / 80.56	**82.3** / 79.28	81.15 / 79.27	81.07 / 80.18	82.14 / 78.56
	The same MSA modules			Ours	
(B): Swin-T	$\mathrm{S}_{\mathrm{c}}^{1} \rightarrow \mathrm{S}_{\mathrm{m}}^{4\rightarrow fc}$	$\mathrm{S}_{\mathrm{c}}^{1\rightarrow 2} \rightarrow \mathrm{S}_{\mathrm{m}}^{4\rightarrow fc}$	$\mathrm{S}_{\mathrm{c}}^{1\rightarrow 4} \rightarrow \mathrm{S}_{\mathrm{m}}^{4\rightarrow fc}$	$\mathrm{S}_{\mathrm{c}}^{1\rightarrow 3} \rightarrow \mathrm{S}_{\mathrm{m}}^{4\rightarrow fc}$	
	80.84 / 80.23	81.7 / 80.93	80.11 / 79.98	81.93 / **81.61**	

Table 5. **Different fusions.** The student is ResNet18 in CIFAR100. $\mathrm{S}_{\mathrm{c}}^{1\rightarrow 2} \rightarrow \mathrm{S}_{\mathrm{m}}^{3\rightarrow fc}$ denotes we fuse the first two stages of CNN models and the remain parts start from the third stage of MSA models. The others are similar to this definition.

Teacher	Student	From Scratch		FBT		
		T.	S.	T.	F.	S.
Swin-T	ResNet18	81.38	69.75	81.38	**76.91**	72.21
Mixer-B/16	MobileNetV2	76.62	68.87	76.62	**73.85**	72.31
Mixer-B/16	DeiT-T	76.62	72.17	76.62	**76.30**	75.00
ConvNeXt-T	ResMLP-S12	82.05	76.65	82.05	**81.20**	77.33

Table 6. The performance of our fusion model (F.) is between that of the teacher (T.) and student (S.).

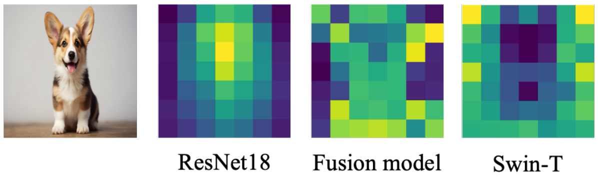


Figure 4. The final spatial distribution of the T., F., and S. is different. So we smooth them for alignments in $\mathcal{L}_{\text{InfoNCE}}$.

on CIFAR100, while $\mathcal{L}_{\text{InfoNCE}}$ is more important when the teacher is ConvNeXt-T and the student is Swin-P. In other words, our $\mathcal{L}_{\text{InfoNCE}}$ and $\mathcal{L}_{\text{OFA}}$ are complementary for different T.-S. pairs, so utilizing them jointly will make various distillations promising. Besides, we demonstrate that MSE loss is not suitable for some T.-S. pairs compared to our INFONCE loss [33] in Sec. 4.2 and Appendix E.

4.4. Discussion

Fuse with different modules. We compare the performance when we fuse different modules of students and teachers in Tab. 5. Specifically, we conduct nine different fusions between the student ResNet18 and the teacher (A) ViT-S / (B) Swin-T on CIFAR100. As shown in Tab. 5, the best result is 82.3% when the teacher is ViT-S and the fusion is $\mathrm{S}_{\mathrm{c}}^{1\rightarrow 2} \rightarrow \mathrm{S}_{\mathrm{m}}^{3\rightarrow fc}$ and is 81.61% when the teacher is Swin-T and the fusion is $\mathrm{S}_{\mathrm{c}}^{1\rightarrow 3} \rightarrow \mathrm{S}_{\mathrm{m}}^{4\rightarrow fc}$. For the hybrid fusion, CNN modules and MSA modules are complementary and both play important roles [4, 23]. In this paper, although different T.-S pairs have different optimal fusions, we add an MSA/MLP stage following three CNN stages for fusion in most situations, *i.e.*, $\mathrm{S}_{\mathrm{c}}^{1\rightarrow 3} \rightarrow \mathrm{S}_{\mathrm{m}}^{4\rightarrow fc}$, which ensures simplicity and is adaptive for different model pairs.

The performance and features of our fused model. Firstly, as shown in Tab. 6, our fused model delivers a middle performance between students and teachers, thereby demonstrating its role as a knowledge-fusion bridge. Secondly, as shown in Fig. 4, the final features of our fusion model are global, *i.e.*, our fused model combines the knowledge from different inductive biases and module functions in final feature spaces. Lastly, the features of the student, fusion, and teacher models are spatially different in Fig. 4, so it is reasonable to smooth them before transferring in Eq. (3). Our fusion model plays an important role for bridging the knowledge transfer between T.-S. pairs.

5. Conclusion

Limitation and Future Work. (1) It is noteworthy that for certain specific models, such as the extensively studied ResNet18, the performance resulting from distillation by a heterogeneous teacher is inferior to that achieved by a homogeneous teacher. Although our current focus is on the generalizability of various heterogeneous models and yields significant performance improvements in CAKD, a promising avenue for future research may involve additional prior when specific teacher-student pairs are predefined. (2) Our FBT may disrupt heterogeneous features' spatial alignments. This limitation could be mitigated by aligning extra spatial-level distributions (rather than pixel-level).

Conclusion. This paper introduces a novel knowledge **F**usion scheme **B**efore **T**ransferring (FBT), which is adaptive according to different model pairs and enhances the efficacy of heterogeneous distillation. Our FBT integrates diverse inductive biases and module functions by fusing CNN/MSA/MLP modules derived from students and teachers, thereby improving the feature transfer among heterogeneous models. Besides, we replace pixel-wise MSE loss with spatial-agnostic loss, which mitigates heterogeneous feature gaps in spatial. Extensive experiments demonstrate our FBT is more powerful than most homogeneous and heterogeneous methods on CIFAR100 and ImageNet-1K.

6. Acknowledgement

This work is supported by National Natural Science Foundation of China grants under contracts NO.62325111, NO.62306214, and NO.U22B2011.

References

[1] Shengcao Cao, Mengtian Li, James Hays, Deva Ramanan, Yu-Xiong Wang, and Liangyan Gui. Learning lightweight object detectors via multi-teacher progressive distillation. In *International Conference on Machine Learning*, pages 3577–3598. PMLR, 2023. 2, 3

[2] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 7

[3] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 6, 7

[4] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 8

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 5, 7

[6] 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 *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 3, 4, 5

[7] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3, 4, 5, 6, 7

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3, 5

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4, 5

[10] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6, 7

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3, 5, 6

[12] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 5, 6, 7

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 7

[14] Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16719–16729, 2024. 2

[15] Jiachen Li, Ali Hassani, Steven Walton, and Humphrey Shi. Convmlp: Hierarchical convolutional mlps for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4

[16] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 2

[17] Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. Function-consistent feature distillation. *International Conference on Learning Representations (ICLR)*, 2023. 1, 2, 3, 4, 6, 7

[18] Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *arXiv:2103.04062*, 2021. 1, 2, 3

[19] Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, Jingting Ding, and Liang Li. Cross-architecture knowledge distillation. In *Proceedings of the Asian conference on computer vision*, pages 3396–3411, 2022. 3, 6

[20] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 4, 5, 6

[21] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5

[22] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2020. 2, 3

[23] Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3, 4, 7, 8

[24] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 6

[25] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 5, 6

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 5

[27] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3, 4

[28] Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, and Hang Zhao. Co-advise: Cross inductive bias distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[29] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 2, 3, 5, 6, 7

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[31] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5

[32] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15731–15740, 2024. 1, 3

[33] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[34] Ilya O. 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3, 5

[35] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 2021. 5

[36] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 3, 5

[37] Hongjun Wu, Li Xiao, Xingkuo Zhang, and Yining Miao. Aligning in a compact space: Contrastive knowledge distillation between heterogeneous architectures. *arXiv preprint arXiv:2405.18524*, 2024. 3, 6

[38] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5, 6, 7

[39] Borui Zhao, Renjie Song, and Jiajun Liang. Cumulative spatial knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6146–6155, 2023. 3