Revisiting Cross-Architecture Distillation: Adaptive Dual-Teacher Transfer for Lightweight Video Models

Ying Peng[1,2], Hongsen Ye[1,2], Changxin Huang[3], Xiping Hu[2], Jian Chen[1,*], Runhao Zeng[2,*]

[1]South China University of Technology,
[2]Artificial Intelligence Research Institute, Shenzhen MSU-BIT University,
[3]Shenzhen University
{py.yingpeng, runhaozeng.cs}@gmail.com, huxp@bit.edu.cn, ellachen@scut.edu.cn

Abstract

Vision Transformers (ViTs) have achieved strong performance in video action recognition, but their high computational cost limits their practicality. Lightweight CNNs are more efficient but suffer from accuracy gaps. Cross-Architecture Knowledge Distillation (CAKD) addresses this by transferring knowledge from ViTs to CNNs, yet existing methods often struggle with architectural mismatch and overlook the value of stronger homogeneous CNN teachers. To tackle these challenges, we propose a Dual-Teacher Knowledge Distillation framework that leverages both a heterogeneous ViT teacher and a homogeneous CNN teacher to collaboratively guide a lightweight CNN student. We introduce two key components: (1) Discrepancy-Aware Teacher Weighting, which dynamically fuses the predictions from ViT and CNN teachers by assigning adaptive weights based on teacher confidence and prediction discrepancy with the student, enabling more informative and effective supervision; and (2) a Structure Discrepancy-Aware Distillation strategy, where the student learns the residual features between ViT and CNN teachers via a lightweight auxiliary branch, focusing on transferable architectural differences without mimicking all of ViT's high-dimensional patterns. Extensive experiments on benchmarks including HMDB51, EPIC-KITCHENS-100, and Kinetics-400, demonstrate that our method consistently outperforms state-of-the-art distillation approaches, achieving notable performance improvements with a maximum accuracy gain of 5.95% on HMDB51.

1 Introduction

Video action recognition has received increasing attention in recent years due to its wide applications in intelligent surveillance, human–computer interaction, and autonomous driving. Vision Transformers (ViTs) (Arnab et al. 2021; Liu et al. 2022b; Tong et al. 2022), with their powerful global modeling capability, have achieved remarkable performance in this domain and have become one of the mainstream high-accuracy model families. However, the large parameter size and high computational cost of ViTs hinder their deployment on resource-constrained edge devices. In contrast, lightweight convolutional neural networks (CNNs) (Feichtenhofer 2020; Kondratyuk et al. 2021), are more suitable for practical deployment due to their efficient inference and hardware-friendliness, but they suffer from notable performance gaps compared to ViTs. Bridging this gap while maintaining model efficiency remains a core challenge.

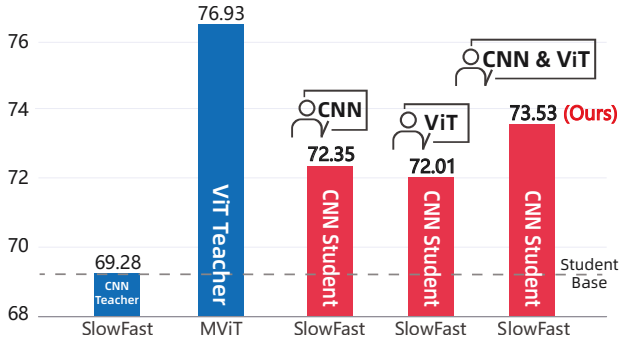

Figure 1: Motivation for our dual-teacher framework. A preliminary study shows that distillation from a weaker but structurally aligned CNN teacher can sometimes lead to better student performance than from a stronger, heterogeneous ViT teacher. This finding motivates the integration of both teacher types and highlights the critical role of architectural compatibility in cross-architecture knowledge transfer.

To address this issue, Cross-Architecture Knowledge Distillation (CAKD) has emerged as a promising solution, which aims to transfer knowledge from a powerful ViT teacher to a lightweight CNN student. However, due to the fundamental differences between ViTs and CNNs in terms of inductive biases, receptive fields, and architectural designs, their intermediate representations are difficult to align semantically and spatially. To overcome this challenge, existing CAKD methods often introduce feature projection modules to map heterogeneous features into a shared latent space. For example, CAKD (Liu et al. 2022a) incorporates complex projection modules involving Partially Cross Attention and Group-wise Linear Transformation for feature alignment, while OFA-KD (Hao et al. 2023) introduces auxiliary exits in the student model to project intermediate features into the logits space, thus bypassing structural mismatches. Although these approaches alleviate feature-level misalignment to some extent, they heavily rely on intricate alignment modules, suffer from limited training stability

*Corresponding Author

and scalability, and more importantly, overlook a valuable source of knowledge—homogeneous CNN teachers, which are structurally similar to the student.

Although the overall accuracy of a homogeneous CNN teacher may be lower than that of a ViT, our preliminary experiments reveal that incorporating a CNN teacher can sometimes yield even greater performance gains. We attribute this to the structural consistency between the CNN teacher and student, which facilitates more effective feature transfer and guidance. This observation motivates us to rethink the ViT-dominant paradigm in CAKD and explore whether a more collaborative and compatible distillation framework can be designed to fully leverage both heterogeneous and homogeneous teachers.

To this end, we propose a novel Dual-Teacher Knowledge Distillation Framework, where a ViT teacher and a CNN teacher collaboratively guide a lightweight CNN student. However, introducing dual teachers raises two key challenges: **1) how can the student effectively select and balance the supervision from two structurally dissimilar teachers during training?** and **2) how can we extract complementary and architecture-relevant knowledge from both teachers, rather than forcing direct alignment with heterogeneous features?**

For the first challenge, we perform a statistical analysis[1] and observe that distillation gains are most significant when the teacher shows high confidence (i.e., low entropy) and there is a large discrepancy between teacher and student predictions (e.g., low cosine similarity). These conditions indicate that the teacher offers reliable yet novel knowledge for the student. Based on this, we propose a **Discrepancy-Aware Teacher Weighting (DATW)** mechanism, which computes the distillation target as a weighted combination of the two teachers' logits. The weights are dynamically assigned per sample, considering both teacher confidence and prediction discrepancy, allowing adaptive and informative supervision from the more helpful teacher.

To tackle the second challenge, we argue that directly aligning student features with those of a ViT teacher is suboptimal due to substantial architectural and representational differences. Instead, we ask: can the difference between heterogeneous teachers offer more targeted supervision? Since both the ViT and CNN teachers are more expressive than the student, the residual between their features naturally captures the student's structural deficiency—especially the global context strength of ViT. We propose a **Structure Discrepancy-Aware Distillation** mechanism that guides the student to predict this residual. A lightweight non-local module (Wang et al. 2018) takes the student's features as input and learns to approximate the difference between ViT and CNN teacher features. This encourages the student to internalize the architectural gap and extract the ViT's complementary knowledge. The auxiliary prediction branch is only used during training and is detached at inference.

We validate the effectiveness of our method on HMDB51, EPIC-KITCHENS-100, Kinetics-400 and Something-Something V2. Results demonstrate that the student models trained with our framework not only outperform existing methods, but even exceed the performance of both ViT and CNN teachers.

Our main contributions are summarized as follows:

- We propose a dual-teacher cross-architecture distillation framework for video action recognition, which is the first to jointly leverage both a heterogeneous ViT teacher and a homogeneous CNN teacher, breaking the ViT-only paradigm in existing CAKD methods.
- We introduce a discrepancy-aware teacher weighting mechanism that adaptively balances supervision from both teachers, enabling sample-specific guidance based on confidence and prediction disagreement.
- We design a structure discrepancy-aware distillation method that encourages the student to learn transferable differences between ViT and CNN features, improving global representation capability without incurring inference overhead. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art distillation techniques.

2 Related Work

2.1 Video Action Recognition

Video action recognition aims to classify human actions in videos. Early works like I3D (Carreira and Zisserman 2017), R(2+1)D (Tran et al. 2018), and SlowFast (Feichtenhofer et al. 2019a) advanced spatio-temporal modeling via 3D convolutions and dual-pathway designs. To further reduce overhead, methods such as TSM (Lin, Gan, and Han 2019), X3D (Feichtenhofer 2020), and MoViNet (Kondratyuk et al. 2021) propose lightweight temporal modules and compact architectures optimized for mobile devices.

In parallel, Vision Transformers (ViTs) have shown strong performance in video tasks by modeling long-range dependencies through spatio-temporal self-attention (Bertasius, Wang, and Torresani 2021; Liu et al. 2022c; Li et al. 2022b), often surpassing CNNs. Hybrid models like UniFormer (Li et al. 2022a) further integrate local convolution and global Transformer reasoning for enhanced performance.

However, the high computational and memory costs of ViTs limit their practical scalability. In contrast, CNNs offer low latency, efficient memory use, and strong hardware support (e.g., CUDA, TensorRT) but weaker global context modeling. This motivates using ViTs as teachers to strengthen lightweight CNNs via knowledge distillation. Yet cross-architecture distillation for video action recognition remains underexplored, especially when combining structurally aligned CNNs with architecturally distinct ViTs. Our work addresses this by leveraging their complementary strengths to guide efficient CNN students.

2.2 Knowledge Distillation (KD)

KD (Hinton, Vinyals, and Dean 2015; Wang et al. 2023; Yu et al. 2024; Li et al. 2025; Liu et al. 2025) has been extensively studied to transfer knowledge from large models to compact ones. Existing approaches span logits-based (Li et al. 2023; Mirzadeh et al. 2020; Zhang et al. 2018; Sun

[1]Please refer to the Appendix for more details.

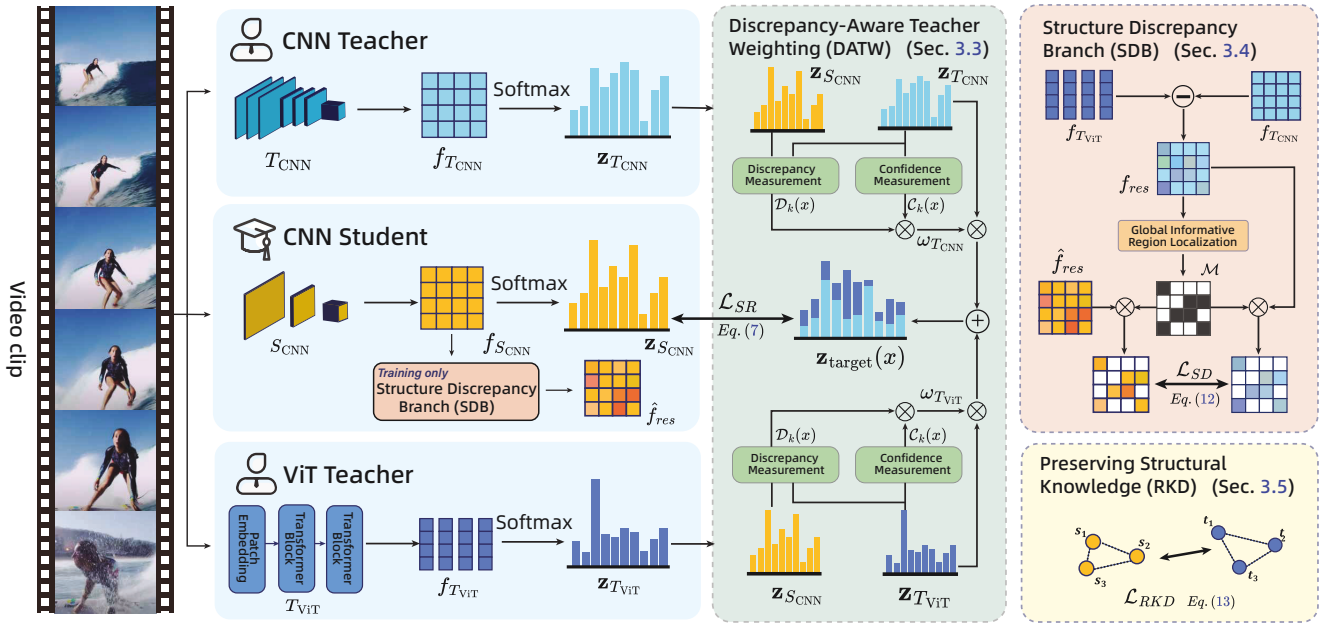

Figure 2: An overview of our dual-teacher distillation framework. Given a lightweight CNN student S_{CNN}, a ViT teacher T_{ViT}, and a CNN teacher T_{CNN}, we distill complementary knowledge via three components: 1) **Discrepancy-Aware Teacher Weighting** mechanism measures teacher confidence $\mathcal{C}_k(x)$ and prediction discrepancy $\mathcal{D}_k(x)$ for each sample, which are integrated to generate adaptive weights $\omega_k(x)$ for combining teacher logits. This enables the student to prioritize informative and reliable supervision on a per-sample basis. 2) **Structure Discrepancy Branch** predicts the feature residual $f_{T_{\text{ViT}}} - f_{T_{\text{CNN}}}$ using a Non-local module, enabling the student to capture ViT-specific global context cues. At inference, only S_{CNN} is used, introducing no extra computational overhead. 3) **Relational Knowledge Distillation** transfers architecture-agnostic structural knowledge.

et al. 2024), feature-based (Chen et al. 2021, 2022; Tian, Krishnan, and Isola 2019; Wang et al. 2024), and relation-based (Park et al. 2019; Yim et al. 2017; Peng et al. 2019) strategies. However, most are designed for homogeneous architectures (e.g., CNN-to-CNN) and assume strong structural alignment between teacher and student.

Cross-architecture distillation (CAKD), especially from ViTs to CNNs, is more challenging due to the architectural gap. Recent CAKD works (Liu et al. 2022a; Hao et al. 2023; Zhao et al. 2023) attempt to bridge this via projection modules or shared latent spaces. Yet, these designs often require complex feature alignment and ignore the representational strengths of structurally aligned CNN teachers.

Notably, in video action recognition, cross-architecture KD remains underexplored. Existing methods primarily distill from a single ViT teacher and often overlook the value of combining CNN teachers for structure-compatible guidance. We addresses these gaps by introducing a dual-teacher framework that integrates both ViT and CNN teachers, using structure-aware mechanisms to fuse global and local inductive biases and better support lightweight CNN students.

3 Proposed Method

3.1 Preliminaries and Motivation

Cross-Architecture Knowledge Distillation (CAKD) typically aims to transfer knowledge from a powerful ViT teacher (T_{ViT}) to a lightweight CNN student (S_{CNN}). Given an input video $V \in \mathbb{R}^{T\times H\times W\times C}$, where T, H, W, and C denote the number of frames, height, width, and channels of the video clip, respectively. The ViT teacher first embeds the input into a sequence of patch tokens and processes them through Transformer blocks, yielding features $f_{T_{\text{ViT}}} \in \mathbb{R}^{N\times D}$, where N is the number of tokens and D is the embedding dimension. In contrast, the CNN student extracts spatial-temporal features $f_{S_{\text{CNN}}} \in \mathbb{R}^{T'\times H'\times W'\times D'}$ via 3D convolutions. Both features are then forwarded to task heads to produce final logit outputs $\mathbf{z}_{T_{\text{ViT}}}$ and $\mathbf{z}_{S_{\text{CNN}}}$.

CAKD typically operates on two levels: **logits-based** distillation and **feature-based** distillation. The former encourages the student to mimic the softened output distribution of the teacher to learn inter-class relationships:

$$\mathcal{L}_{\text{KD}}^{\text{logits}}(\mathbf{z}_{T_{\text{ViT}}}, \mathbf{z}_{S_{\text{CNN}}}) = \tau^2 \cdot \text{KL}(\sigma(\mathbf{z}_{T_{\text{ViT}}}/\tau) \parallel \sigma(\mathbf{z}_{S_{\text{CNN}}}/\tau)), \tag{1}$$

where $\sigma(\cdot)$ is the softmax function and τ is the temperature parameter. The latter aims to encourage the student to mimic the teacher's intermediate representations. The loss is typically defined as:

$$\mathcal{L}_{\text{KD}}^{\text{feat}}(f_{T_{\text{ViT}}}, f_{S_{\text{CNN}}}) = \|\phi(f_{S_{\text{CNN}}}) - f_{T_{\text{ViT}}}\|^2 \tag{2}$$

where $\phi(\cdot)$ is an optional projection module used to align feature dimensions when necessary.

Motivation Existing CAKD methods face two key limitations. **First**, overlooking structurally homogeneous CNN

teachers, which, despite lower accuracy than ViTs, can more effectively guide CNN students due to shared architectural priors (see Figure 1). **Second**, enforcing full imitation of ViT features, risking architectural noise and suboptimal learning. These insights motivate us to integrate both teacher types, enabling the student to benefit from the structural alignment of CNNs and the representational strength of ViTs through complementary, targeted supervision.

3.2 General Scheme

To address the aforementioned limitations, we propose a **Dual-Teacher Knowledge Distillation (DT-KD)** framework that jointly leverages a heterogeneous ViT teacher (T_{ViT}) and a homogeneous CNN teacher (T_{CNN}) to guide a lightweight CNN student (S_{CNN}). This framework addresses two key challenges: **C1**–how to adaptively balance supervision from structurally dissimilar teachers, and **C2**–how to extract complementary, architecture-relevant knowledge without enforcing direct feature alignment.

To tackle **C1**, we introduce **Discrepancy-Aware Teacher Weighting (DATW)**, which assigns sample-wise weights based on teacher confidence and student-teacher prediction discrepancy. This enables dynamic teacher cooperation at the logits level, forming the loss $\mathcal{L}_{\text{SR}}$.

To address **C2**, we propose **Structure Discrepancy-Aware Distillation (SDD)**. Instead of forcing the student to mimic ViT features directly, we introduce a lightweight auxiliary branch $g_{\text{SDB}}(\cdot)$ during training to predict the residual between the ViT and CNN teacher features—capturing architecture-specific knowledge.

Finally, a **Relational Knowledge Distillation (RKD)** loss $\mathcal{L}_{\text{RKD}}$ further preserves high-level relational structure from the ViT teacher. These components work jointly to transfer knowledge adaptively and effectively.

3.3 Discrepancy-Aware Teacher Weighting

A central challenge in dual-teacher distillation lies in determining how a student model should effectively balance and select supervision from two structurally dissimilar teachers To improve the logits-level distillation, we aim to replace the fixed teacher target $\mathbf{z}_{T_{\text{ViT}}}$ in Eq. (1) with a sample-adaptive combination of both teacher logits:

$$\mathbf{z}_{\text{target}}(x) = \omega_{\text{CNN}}(x) \cdot \mathbf{z}_{T_{\text{CNN}}}(x) + \omega_{\text{ViT}}(x) \cdot \mathbf{z}_{T_{\text{ViT}}}(x), \quad (3)$$

where $\omega_{\text{CNN}}(x)$ and $\omega_{\text{ViT}}(x)$ are adaptive weights for each input x. Through an in-depth analysis of our preliminary experiments, we find that the most beneficial knowledge for distillation typically arises when the teacher's predictions exhibit both **1)** high confidence and **2)** large discrepancy from the student's predictions (Detailed analysis is provided in the Appendix). Based on this insight, we propose the **Discrepancy-Aware Teacher Weighting (DATW)** mechanism, which quantifies teacher quality using two metrics and translates them into soft arbitration weights in a differentiable manner.

Measuring Teacher Confidence We measure the confidence of teacher $k \in \{\text{CNN, ViT}\}$ on sample x using the normalized negative entropy of its softmax prediction:

$$\mathcal{C}_k(x) = \frac{1}{\log N} \sum_{i=1}^{N} p_{k,i}(x) \log p_{k,i}(x), \quad (4)$$

where $p_k(x) = \text{Softmax}(\mathbf{z}_k(x)/\tau)$ is the temperature-scaled prediction, and N is the number of classes. Lower entropy implies higher certainty in the teacher's prediction.

Measuring Student–Teacher Discrepancy To assess the potential informational gain from each teacher, we compute the cosine distance between student's and teacher's logits:

$$\mathcal{D}_k(x) = 1 - \frac{\mathbf{z}_{S_{\text{CNN}}}(x) \cdot \mathbf{z}_k(x)}{\|\mathbf{z}_{S_{\text{CNN}}}(x)\| \|\mathbf{z}_k(x)\|}. \quad (5)$$

A larger value of $\mathcal{D}_k(x)$ implies greater disagreement and hence a higher chance for the teacher to provide complementary knowledge.

Combining Confidence and Discrepancy We define a guidance efficacy score $s_k(x)$ by multiplying the above two metrics and normalize the scores across both teachers to obtain the final arbitration weights:

$$\omega_k(x) = \frac{s_k(x)}{\sum_{j \in \{\text{CNN, ViT}\}} s_j(x) + \epsilon}, \quad (6)$$
$$s_k(x) = \mathcal{C}_k(x) \cdot \mathcal{D}_k(x),$$

where ϵ is a small constant for numerical stability. Using the weighted teacher logits $\mathbf{z}_{\text{target}}(x)$, the final logits-level distillation loss becomes:

$$\mathcal{L}_{\text{SR}} = \tau^2 \cdot \mathbb{E}_x \left[\text{KL} \left(\sigma(\mathbf{z}_{\text{target}}(x)/\tau) \,\|\, \sigma(\mathbf{z}_{S_{\text{CNN}}}(x)/\tau) \right) \right]. \quad (7)$$

3.4 Structure Discrepancy-Aware Distillation

Transferring feature-level knowledge from a ViT teacher to a CNN student is challenging due to their divergent architectural priors (Liu et al. 2022a; Hao et al. 2023). ViTs excel at capturing global dependencies through self-attention, while CNNs mainly extract local patterns via convolution. Forcing the student to mimic ViT features holistically may result in poor alignment and limited performance gains due to mismatched representational formats.

Learning via Feature Residuals To mitigate this, we propose a Structure Discrepancy-Aware Distillation (SDD) mechanism, which reframes the learning target as the feature residual between a ViT and a structurally aligned CNN teacher. This residual emphasizes global cues encoded by the ViT but missing from the CNN, offering a more focused and complementary supervision signal for the student. Formally, the residual is defined as:

$$f_{\text{res}} = \phi_{\text{ViT}}(f_{T_{\text{ViT}}}) - \phi_{\text{CNN}}(f_{T_{\text{CNN}}}), \quad (8)$$

where $\phi(\cdot)$ is an optional projection function for aligning feature dimensions. This design leverages the CNN teacher as a reference anchor, allowing the student to learn the ViT's distinct structural cues without being overwhelmed by direct, incompatible supervision.

To localize supervision to the most informative regions, we propose a Dual-Masked Sparse Distillation strategy,

combining two spatial masks: First, the residual-based mask $\mathcal{M}_{\text{res}}$ identifies regions where the ViT and CNN teachers differ significantly, by thresholding the residual map f_{res} at its ζ-quantile:

$$\mathcal{M}_{\text{res}}(i,j) = \begin{cases} 1, & \text{if } f_{\text{res}}(i,j) > Q_\zeta \\ 0, & \text{otherwise} \end{cases}. \tag{9}$$

Second, to suppress irrelevant background noise and enhance semantic concentration, we define an activation-based mask $\mathcal{M}_{\text{act}}$ that highlights semantically relevant regions based on high average activations from the CNN teacher:

$$\mathcal{M}_{\text{act}}(i,j) = \begin{cases} 1, & \text{if } \frac{1}{C}\sum_{c=1}^{C} f_{T_{\text{CNN}}}^{(c)}(i,j) > Q_\rho \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $f_{T_{\text{CNN}}}^{(c)}(i,j)$ denotes the activation value at spatial location (i,j) in channel c, and C is the total number of channels. We apply supervision only over the intersection $\mathcal{M}_{\text{res}} \cap \mathcal{M}_{\text{act}}$, focusing learning on structurally distinctive and semantically salient regions.

Auxiliary Learning with Structure Discrepancy Branch (SDB) To avoid entangling this learning with the student's core prediction path, we introduce a lightweight auxiliary module called the Structure Discrepancy Branch (SDB) g_{SDB}. This branch operates during training only and is detached at inference. It takes the student's backbone features $f_{S_{\text{CNN}}}$ as input and predicts the ViT-CNN residual:

$$\hat{f}_{\text{res}} = g_{\text{SDB}}(f_{S_{\text{CNN}}}) = \text{SE}\big(\text{NL}(f_{S_{\text{CNN}}})\big). \tag{11}$$

Here, $\text{NL}(\cdot)$ is a non-local block that models global spatial dependencies, mimicking ViT-style attention, while $\text{SE}(\cdot)$ is a squeeze-and-excitation module that recalibrates channel-wise responses to enhance feature expressiveness. Isolating these components in the SDB allows the student to learn architecture-specific global information without disrupting its primary learning pathway.

The final distillation loss is computed as the masked MSE between the predicted and ground-truth residuals:

$$\mathcal{L}_{SD} = \|(g_{\text{SDB}}(f_{S_{\text{CNN}}}) - f_{\text{res}}) \odot \mathcal{M}_{\text{res}} \odot \mathcal{M}_{\text{act}}\|_2^2. \tag{12}$$

By modeling ViT–CNN structural differences explicitly and restricting supervision to the most informative regions, SDB offers a principled, efficient pathway for the student to selectively absorb global inductive knowledge that would otherwise be difficult to transfer across architectures.

3.5 Preserving Structural Knowledge with RKD

As previously mentioned, directly aligning intermediate features between ViT and CNN models is often suboptimal due to their inherently different architectural priors and feature semantics. To alleviate this, we incorporate Relational Knowledge Distillation (RKD) (Park et al. 2019) to transfer architecture-agnostic structural knowledge.

RKD bypasses absolute feature matching and instead focuses on preserving the pairwise relational structure among samples in the feature space. This approach captures higher-order inter-sample dependencies, which are more transferable across heterogeneous architectures.

Specifically, for each sample pair (i,j) in a mini-batch, we compute their Euclidean distances in both teacher and student feature spaces, denoted as $d_T(i,j)$ and $d_S(i,j)$, and normalize by the batch-wise mean distances $\bar{d}_T$, $\bar{d}_S$. The RKD loss is then defined as:

$$\mathcal{L}_{\text{RKD}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \text{SmoothL1}\left(\frac{d_S(i,j)}{\bar{d}_S}, \frac{d_T(i,j)}{\bar{d}_T}\right) \tag{13}$$

where $\mathcal{P}$ is the set of all distinct sample pairs in the batch. This encourages the student to preserve the geometric structure induced by the teacher, without requiring explicit feature alignment.

3.6 Training and Inference Details

Training Objective The student model is trained end-to-end by minimizing a comprehensive objective function that integrates the standard supervised loss with our proposed multi-level distillation losses. The final objective, $\mathcal{L}_{\text{Total}}$, is formulated as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{SR}} + \beta\mathcal{L}_{\text{SD}} + \gamma\mathcal{L}_{\text{RKD}} \tag{14}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss on the ground-truth labels, while α, β, and γ are hyperparameters that balance the contributions of our proposed distillation components.

Inference Efficiency During inference, the student model operates independently without requiring access to either teacher models or auxiliary modules. In particular, the auxiliary structure discrepancy branch (SDB), which is introduced only during training to guide feature learning via $\mathcal{L}_{SD}$, is entirely discarded after training. As a result, the student retains its original architecture, with no additional parameters or computational overhead introduced at test time.

4 Experiments

4.1 Datasets and Compared Methods

We evaluate our method on four widely used video action recognition benchmarks: **HMDB51** (Kuehne et al. 2011), **EPIC-KITCHENS-100** (Damen et al. 2020), **Kinetics-400** (Kay et al. 2017), and **Something-Something V2** (Goyal et al. 2017). Detailed descriptions of all datasets are provided in the Appendix.

We focus on three lightweight and efficient CNN architectures as student models: **MoViNet** (Kondratyuk et al. 2021), **X3D** (Feichtenhofer 2020), and **SlowFast** (Feichtenhofer et al. 2019b). The ViT teacher is **MViTv2** (Li et al. 2022b), and the CNN teacher shares the same architecture as the student. The compared state-of-the-art video knowledge distillation methods include: **ATKD** (Zagoruyko and Komodakis 2016), **MGD** (Yang et al. 2022), **DKD** (Zhao et al. 2022), and **PixelKD** (Guo et al. 2024), covering diverse paradigms of feature-based distillation and logits-based distillation.

4.2 Implementation Details

All models are initialized with standard ImageNet and Kinetics-400 pre-trained weights. We train using the SGD optimizer with a batch size of 64, an initial learning rate of 0.01, and a MultiStepLR scheduler. The input clip length

Method		Only ViT Teacher				ViT Teacher+CNN Teacher	
ViT Teacher	Student w/o distillstion	ATKD	MGD	DKD	PixelKD	CNN Teacher	**Ours**
MViTv2-S (76.93)	MoViNet–A2 (71.76)	72.88	72.48	75.16	70.20	MoViNet-A2 (71.76)	**76.41**
	SlowFast-R50 (69.28)	68.17	67.84	69.28	72.03	SlowFast-R50 (69.28)	**75.23**
	X3D-S (72.42)	64.64	70.72	73.66	75.16	X3D-M (69.80)	**77.06**

Table 1: State-of-the-art comparison on HMDB51. Our dual-teacher method consistently outperforms existing distillation methods. Notably, our distilled X3D-S student achieves **77.06%** accuracy, surpassing its powerful MViTv2-S teacher (76.93%) while achieving a **96%** reduction in FLOPs and an **89%** reduction in parameters (3.76M vs. 34.5M).

Method				Only ViT Teacher				ViT Teacher+CNN Teacher	
Task	Student Model	ViT Teacher	Student	ATKD	MGD	DKD	PixelKD	CNN Teacher	**Ours**
Verb	MoViNet-A2	64.88	62.03	60.19	62.58	62.30	62.38	63.99	**62.75**
Noun		47.18	44.65	43.20	43.61	43.80	43.68	47.71	**46.50**
Verb	SlowFast-R50	64.88	63.99	63.62	64.38	64.55	63.50	63.99	**65.83**
Noun		47.18	47.71	48.76	48.48	48.96	48.87	47.71	**49.80**

Table 2: State-of-the-art comparison on EPIC-Kitchens. We adopt SlowFast-R50 as the CNN Teacher model for comparisons.

Type	Method	Acc (%)
Baseline	w/o distill	73.06
Only ViT Teacher	ATKD	72.06
	MGD	73.05
	DKD	73.38
	PixelKD	73.14
ViT Teacher+CNN Teacher	**Ours**	**74.16**

Table 3: State-of-the-art comparison on Kinetics-400. ViT Teacher: MViTv2-S, CNN Teacher/Student: I3D-R50.

Student	w/o distill.	Only ViT Teacher				**Ours**
		ATKD	MGD	DKD	PixelKD	
MoViNet–A2	56.27	56.81	56.55	57.02	57.11	**58.00**
SlowFast-R50	61.70	61.80	61.61	62.73	62.71	**62.89**
X3D-S	59.19	58.87	57.43	59.18	59.93	**60.85**

Table 4: Comparisons on Something-Something V2. ViT teacher: MViTv2-S (68.10%), CNN teacher shares the same architecture as the student.

is set to 16 frames, and both teachers remain fixed during distillation. Note that MoViNet's original training settings are not fully available, so exact reproduction is infeasible. However, all methods are trained under the *same* configuration, ensuring fair comparison and that the gains are method-driven. More training details are provided in the Appendix.

4.3 Comparisons with State-of-the-arts

Results on HMDB51 Table 1 shows that our framework outperforms all existing methods across lightweight CNN students. For SlowFast-R50, it improves accuracy by 5.95% over the baseline and 3.2% over the previous best (75.23% vs. 72.03%). MoViNet-A2 and X3D-S also improve to 76.41% and 77.06%. We also compare two CNN teacher settings: using the same architecture as the student (e.g., MoViNet-A2, SlowFast-R50) and a larger variant (e.g., X3D-M), both yielding clear gains. Notably, the distilled X3D-S even surpasses its ViT teacher (77.06% vs. 76.93%), highlighting the strong generalization of our framework.

Results on EPIC-KITCHENS-100 As shown in Table 2, our dual-teacher framework outperforms all methods across both verb and noun classification. With MoViNet-A2, it achieves 46.50% noun accuracy, surpassing the best single-teacher result by 2.70%. On SlowFast-R50, the student reaches 65.83% verb and 49.80% noun accuracy, outperforming all baselines and even the ViT teacher. These results show that combining ViT and CNN teachers enables more effective knowledge transfer, benefiting from their complementary global-context and architectural alignment.

Results on Kinetics-400 Table 3 presents the results of various single-teacher distillation methods using MViT-S to distill I3D-R50. Our dual-teacher framework outperforms all compared feature-, logit-, and hybrid-based approaches. The relatively modest improvement on Kinetics-400 can be attributed to its data characteristics—most videos are short, centered, and less reliant on global context modeling, which aligns well with the strengths of CNNs. Nonetheless, our method still brings consistent improvements, demonstrating strong generality and robustness.

Results on Something-Something V2 As shown in Table 4, across all student models, our framework improves performance over the baselines. The lightweight MoViNet-A2 reaches 58.00% (+1.73%), SlowFast-R50 achieves 62.89% (+1.19%), and X3D-S attains 60.85% (+1.66%). These results confirm that our dual-teacher distillation effectively enhances representation learning and temporal understanding on fine-grained, temporally complex videos.

4.4 Ablation Studies

Effectiveness of Discrepancy-Aware Teacher Weighting (DATW) Table 5 compares DATW with three baseline strategies: **1) Uniform**, which assigns equal weight to

Method	Accuracy (%)
MViTv2-S (ViT Teacher)	76.93
SlowFast-R50 (CNN Teacher)	69.28
SlowFast-R50 (CNN Student)	69.28
Average	71.63
Teacher Confidence: $\mathcal{C}_k(x)$	73.86
Stu–Tea Discrepancy: $\mathcal{D}_k(x)$	73.20
DATW: $\mathcal{C}_k(x) \cdot \mathcal{D}_k(x)$	**73.99**

Table 5: Ablation study for our Discrepancy-Aware Teacher Weighting (DATW) mechanism. Our full method, which combines both Teacher Confidence $\mathcal{C}_k(x)$ and Student-Teacher Discrepancy $\mathcal{D}_k(x)$, significantly outperforms other strategies. (*The SDB module was deactivated for this study.*)

Method	Accuracy (%)
Baseline (w/o SDB)	75.23
SDB (SENet only)	76.14
SDB (Non-Local Block only)	76.54
SDB (Non-Local + SENet)	**77.06**

Table 6: Ablation study of the Structure Discrepancy Branch (SDB) on HMDB51. The Non-local Block provides a larger gain due to its global modeling capability.

all samples and serves as a static baseline; **2) Teacher Confidence-only**, which uses only the teacher's confidence score $\mathcal{C}_k(x)$ for weighting; and **3) Student-Teacher Discrepancy-only**, which relies on the student–teacher prediction gap $\mathcal{D}_k(x)$. The results show that both **Confidence-only** and **Discrepancy-only** outperform the **Uniform** baseline, indicating that dynamic, sample-specific weighting improves knowledge transfer. Our full DATW, combining confidence and discrepancy, achieves the best result of **73.99%**, confirming that jointly modeling reliability and novelty provides the most effective supervision.

Effectiveness of the Structure Discrepancy Branch (SDB) To evaluate the effectiveness of different instantiations of SDB, we compare three variants: **1) SDB with SE module**, **2) SDB with Non-local block**, and **3) SDB with both**. As shown in Table 6, all variants surpass the baseline (75.23%), confirming the benefit of using ViT-CNN feature residuals as supervision. The Non-local block achieves 76.54% (+1.31%), outperforming the SE module (76.14%, +0.91%) due to better long-range modeling. Combining both gives the highest result of 77.06% (+1.83%), showing their complementary benefits. This further validates the SDB design and the use of ViT-specific global cues.

Generalization Study with Different Teachers. To assess the generalization capability of our framework, we experiment with four teacher combinations on HMDB51, keeping the student model fixed as X3D-S. Specifically, we experiment with two ViT teachers (Swin-B and MViTv2-S) and two CNN teachers (R(2+1)D and X3D-M), covering both heterogeneous and homogeneous architectures. As shown in Table 7, our framework consistently improves the student's performance across all teacher combinations. The highest accuracy of **77.06%** is achieved when the CNN teacher (X3D-M) matches the student's architecture, highlighting the advantage of structural alignment in distillation.

ViT Teacher	CNN Teacher	CNN Student	Acc (%)
–	–	X3D-S	72.42
Swin-B	R(2+1)D	X3D-S	75.36
	X3D-M	X3D-S	76.80
MViTv2-S	R(2+1)D	X3D-S	74.77
	X3D-M	X3D-S	**77.06**

Table 7: Results with different teacher combinations on HMDB51. Greatest gain is achieved when the CNN teacher is structurally homogeneous with the student (X3D-M with X3D-S), highlighting the benefit of architectural alignment.

Arch.	Model	FLOPs (G)	Params (M)	Acc (%)
ViT	MViTv2-S	64.45	34.50	76.93
	UniFormerv2-B	96.74	49.81	73.59
	TimeSformer	180	86.11	72.81
	Swin-B	282	88.05	75.42
CNN	I3D-R50	33.27	27.33	71.05
	R(2+1)D	213	63.58	74.12
	SlowFast-R50	36.10	34.40	69.28
	SlowFast-R50 (Ours)	36.10	34.40	**75.23**
	MoViNet-A2	3.85	4.80	71.76
	MoViNet-A2 (Ours)	3.85	4.80	**76.41**
	X3D-S	2.45	3.76	72.42
	X3D-S (Ours)	2.45	3.76	**77.06**

Table 8: Comparison of model complexity and accuracy on HMDB51. Our method establishes a superior performance-efficiency trade-off, enabling lightweight students to achieve high accuracy while maintaining computational efficiency.

Comparison of Model Complexity and Performance. Table 8 shows that, with our framework, X3D-S improves by 4.64% to **77.06%** without any inference overhead—surpassing not only other lightweight models but also its ViT teacher (i.e., MViTv2-S). MoViNet-A2 and SlowFast-R50 show similar gains. These results confirm that our approach can effectively overcome the performance bottleneck of lightweight CNNs, enabling them to rival or exceed larger ViT models at lower cost.

5 Conclusion

This work has tackled two core limitations of cross-architecture knowledge distillation (CAKD) for video action recognition: the over-reliance on ViT supervision and the neglect of structurally aligned CNN teachers. We have proposed a dual-teacher framework that integrates the complementary strengths of both architectures through adaptive teacher weighting and structure discrepancy-aware distillation. These strategies jointly address the key challenges of supervision balancing and heterogeneous knowledge transfer. Extensive experiments show that our method consistently enhances lightweight CNN students, outperforming both teachers and existing distillation approaches. While we have demonstrated strong results with fixed teachers, future work could explore joint optimization or dynamic teacher adaptation to further improve flexibility and effectiveness.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant Nos. 62202311 and 62376099), the Natural Science Foundation of Guangdong Province (Grant No. 2024A1515010989), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515011512), the Key Scientific Research Project of the Department of Education of Guangdong Province (Grant No. 2024ZDZX3012), the Shenzhen Science and Technology Foundation (Grant No. JCYJ20250604173210013) and the Excellent Science and Technology Creative Talent Training Program of Shenzhen Municipality (Grant No. RCBS20221008093224017).

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *Icml*, volume 2, 4.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, D.; Mei, J.-P.; Zhang, H.; Wang, C.; Feng, Y.; and Chen, C. 2022. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11933–11942.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5008–5017.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; and Wray, M. 2020. The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines. *arXiv preprint arXiv:2005.00343*.

Feichtenhofer, C. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 203–213.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019a. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019b. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; Hoppe, F.; Thurau, C.; Bax, I.; and Memisevic, R. 2017. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Guo, G.; Zhang, D.; Han, L.; Liu, N.; Cheng, M.-M.; and Han, J. 2024. Pixel Distillation: Cost-Flexible Distillation Across Image Sizes and Heterogeneous Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9536–9550.

Hao, Z.; Guo, J.; Han, K.; Tang, Y.; Hu, H.; Wang, Y.; and Xu, C. 2023. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Advances in Neural Information Processing Systems*, 36: 79570–79582.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kondratyuk, D.; Yuan, L.; Li, Y.; Zhang, L.; Tan, M.; Brown, M.; and Gong, B. 2021. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16020–16030.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*, 2556–2563. IEEE.

Li, K.; Wang, Y.; Peng, G.; Song, G.; Liu, Y.; Li, H.; and Qiao, Y. 2022a. UniFormer: Unified Transformer for Efficient Spatial-Temporal Representation Learning. In *International Conference on Learning Representations*.

Li, T.; Liu, L.; Liu, K.; Wang, X.; Zhou, B.; Yang, H.; and Lu, K. 2025. Adaptive Dual Guidance Knowledge Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 18457–18465.

Li, Y.; Wu, C.-Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022b. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4804–4814.

Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2023. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1504–1512.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7083–7093.

Liu, Y.; Cao, J.; Li, B.; Hu, W.; Ding, J.; and Li, L. 2022a. Cross-architecture knowledge distillation. In *Proceedings of the Asian conference on computer vision*, 3396–3411.

Liu, Y.; Wu, Z.; Lu, Z.; Nie, C.; Wen, G.; Zhu, Y.; and Zhu, X. 2025. Noisy node classification by bi-level optimization based multi-teacher distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19033–19040.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022b. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022c. Video Swin Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3202–3211.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5191–5198.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3967–3976.

Peng, B.; Jin, X.; Liu, J.; Li, D.; Wu, Y.; Liu, Y.; Zhou, S.; and Zhang, Z. 2019. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5007–5016.

Sun, S.; Ren, W.; Li, J.; Wang, R.; and Cao, X. 2024. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15731–15740.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, G.; Zhao, P.; Shi, Y.; Zhao, C.; and Yang, S. 2024. Generative Model-Based Feature Knowledge Distillation for Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15474–15482.

Wang, R.; Chen, D.; Wu, Z.; Chen, Y.; Dai, X.; Liu, M.; Yuan, L.; and Jiang, Y.-G. 2023. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6312–6322.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-Local Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, Z.; Li, Z.; Shao, M.; Shi, D.; Yuan, Z.; and Yuan, C. 2022. Masked generative distillation. In *European conference on computer vision*, 53–69. Springer.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4133–4141.

Yu, Y.-C.; Huang, C.-P.; Chen, J.-J.; Chang, K.-P.; Lai, Y.-H.; Yang, F.-E.; and Wang, Y.-C. F. 2024. Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models. In *European Conference on Computer Vision*, 219–236. Springer.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4320–4328.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11953–11962.

Zhao, W.; Zhu, X.; He, Z.; Zhang, X.-Y.; and Lei, Z. 2023. Cross-Architecture Distillation for Face Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8076–8085.