ICCV 2025

Perspective-Aware Teaching: Adapting Knowledge for Heterogeneous Distillation

National Yang Ming Chiao Tung University

Jhe-Hao Lin, Yi Yao, Chan-Feng Hsu,

Hong-Xia Xie, Hong-Han Shuai, Wen-Huang Cheng

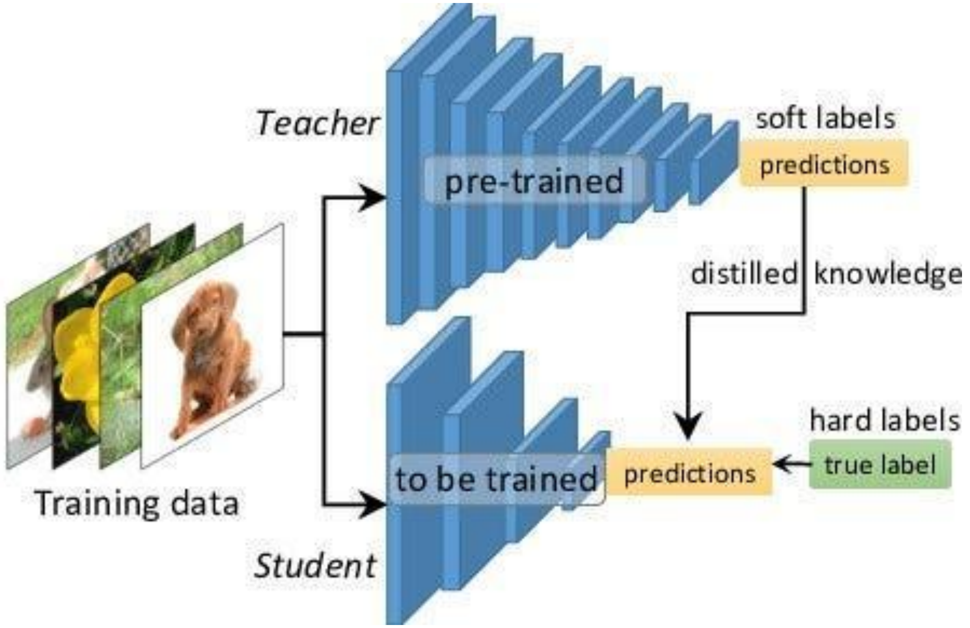
Outline

- Introduction
- Methodology
- Experiments
- Conclusion

Introduction

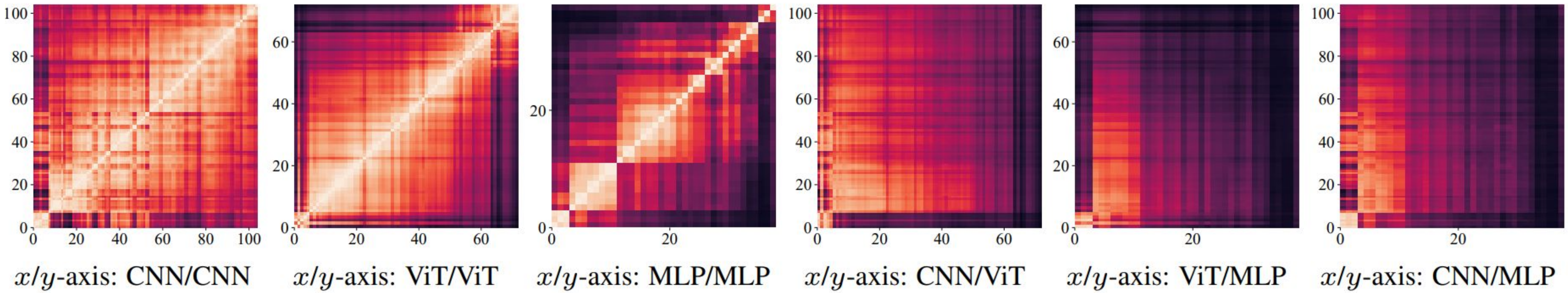
Knowledge Distillation

Knowledge Distillation (KD) aims to develop a lightweight and efficient student model by transferring knowledge from an already trained, larger teacher model.



Heterogeneous KD

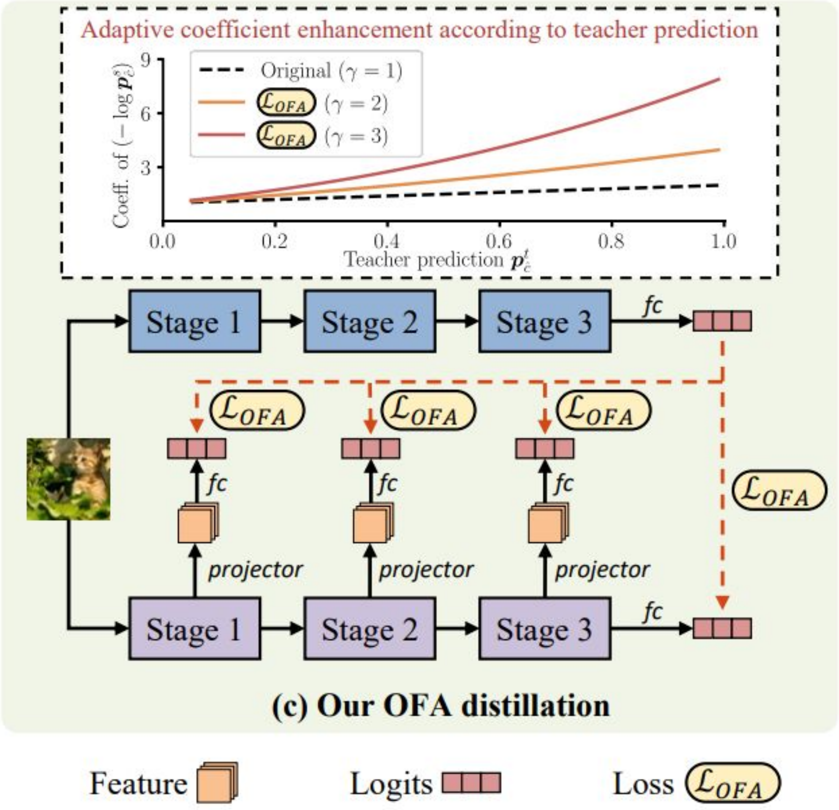
Different architecture exhibit different inductive bias, thus requiring careful and detailed designs to enable the distillation. This often limits the application of the KD to specific teacher-student combinations.



For example, FitNet achieves good results in CNN□CNN format but performs poorly when distilling in a CNN□ViT scenario, only getting 24.06% on CIFAR-100 in ConvNeXt-T□Swin-P, which is lower than basic logit-based method KD by more than 50%.

General Heterogeneous KD

OFA-KD is the first KD framework for universal heterogeneous architecture.



(c) Our OFA distillation

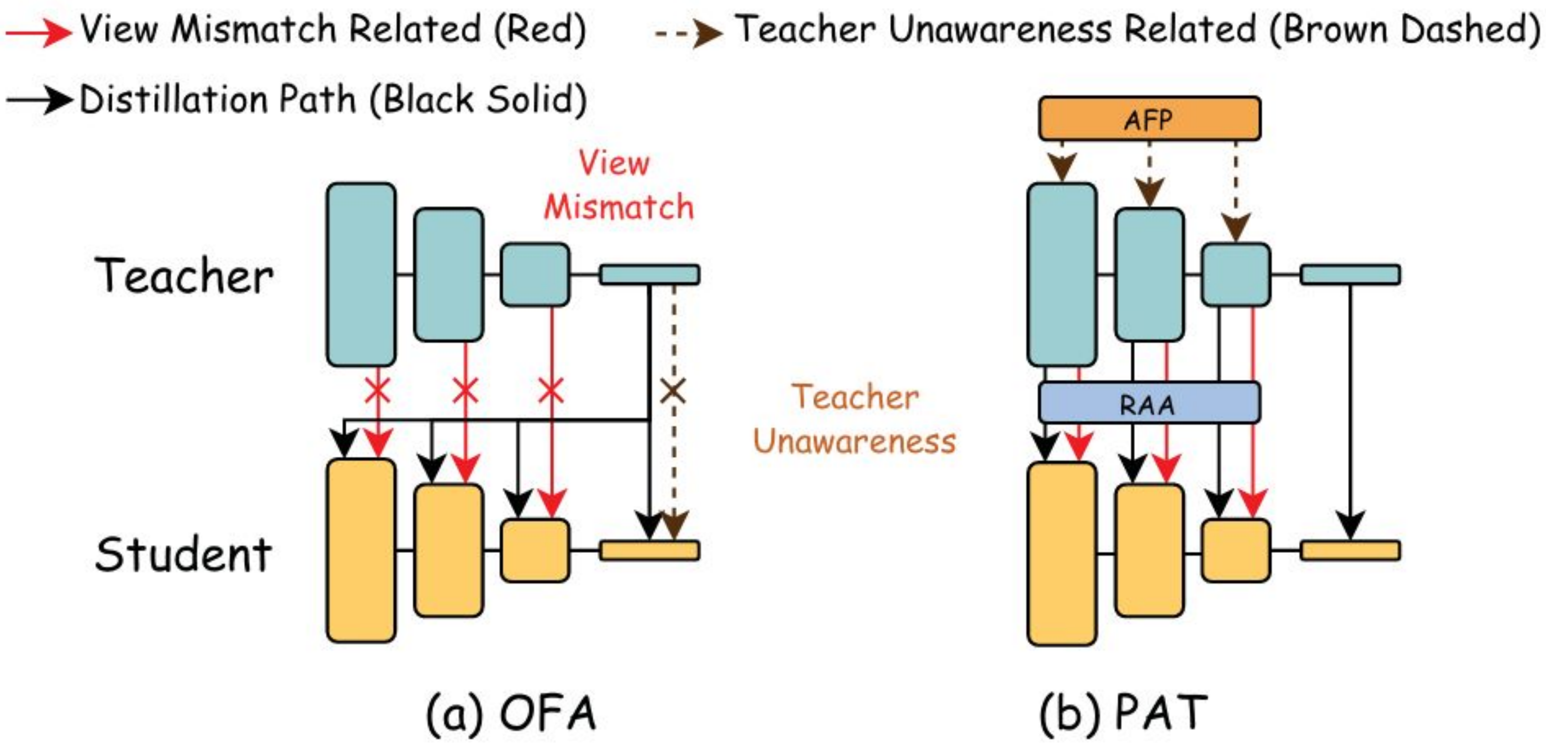
Table 1: KD methods with heterogeneous architectures on ImageNet-1K. The best results are indicated in bold, while the second best results are underlined. †: results achieved by combining with FitNet.

Teacher	Student	From Scratch T.	From Scratch S.	hint-based FitNet	hint-based CC	hint-based RKD	hint-based CRD	Logits-based KD	Logits-based DKD	Logits-based DIST	Logits-based OFA
CNN-based students											
DeiT-T	ResNet18	72.17	69.75	70.44	69.77	69.47	69.25	70.22	69.39	70.64	**71.34**
Swin-T	ResNet18	81.38	69.75	71.18	70.07	68.89	69.09	71.14	71.10	70.91	**71.85**
Mixer-B/16	ResNet18	76.62	69.75	70.78	70.05	69.46	68.40	70.89	69.89	70.66	**71.38**
DeiT-T	MobileNetV2	72.17	68.87	70.95	70.69	69.72	69.60	70.87	70.14	71.08	**71.39**
Swin-T	MobileNetV2	81.38	68.87	71.75	70.69	67.52	69.58	72.05	71.71	71.76	**72.32**
Mixer-B/16	MobileNetV2	76.62	68.87	71.59	70.79	69.86	68.89	71.92	70.93	71.74	**72.12**
ViT-based students											
ResNet50	DeiT-T	80.38	72.17	75.84	72.56	72.06	68.53	75.10	75.60†	75.13†	**76.55†**
ConvNeXt-T	DeiT-T	82.05	72.17	70.45	73.12	71.47	69.18	74.00	73.95	74.07	**74.41**
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**
ResNet50	Swin-N	80.38	75.53	78.33	76.05	75.90	73.90	77.58	78.23†	77.95†	**78.64†**
ConvNeXt-T	Swin-N	82.05	75.53	74.81	75.79	75.48	74.15	77.15	77.00	77.25	**77.50**
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**
MLP-based students											
ResNet50	ResMLP-S12	80.38	76.65	78.13	76.21	75.45	73.23	77.41	78.23†	77.71†	**78.53†**
ConvNeXt-T	ResMLP-S12	82.05	76.65	74.69	75.79	75.28	73.57	76.84	77.23	77.24	**77.53**
Swin-T	ResMLP-S12	81.38	76.65	76.48	76.15	75.10	73.40	76.67	76.99	77.25	**77.31**

By converting student features into a logits space, OFA allows alignment across architectures by removing architecture-specific details. However, the intermediate feature with rich information is abandoned.

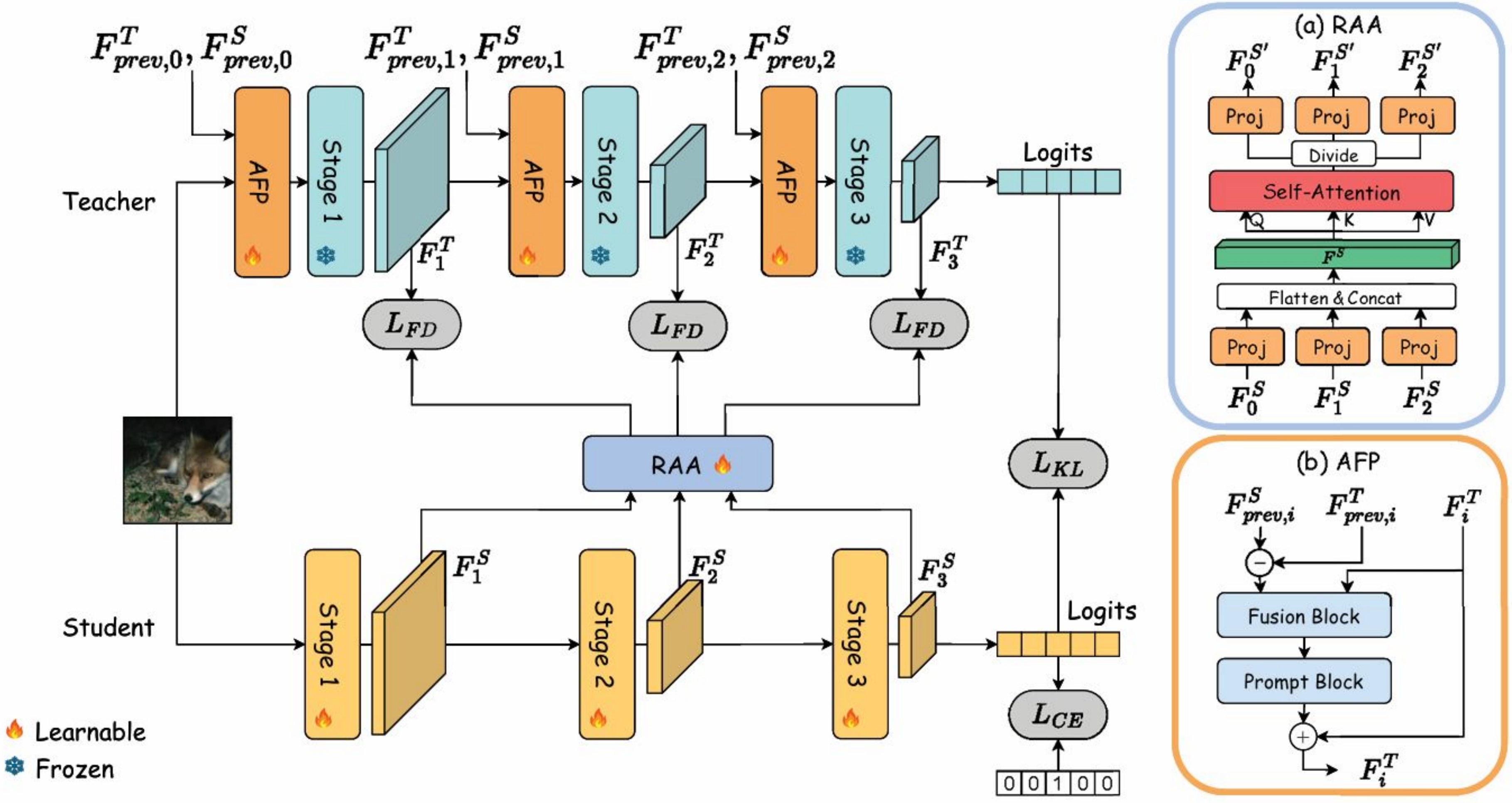
Methodology

PAT Framework



- View mismatch issue, the perspective is different between the teacher and student models due to their distinct architectural receptive fields
- Teacher unawareness problem, where the teacher model fails to adjust the feature based on the student model's learning process.

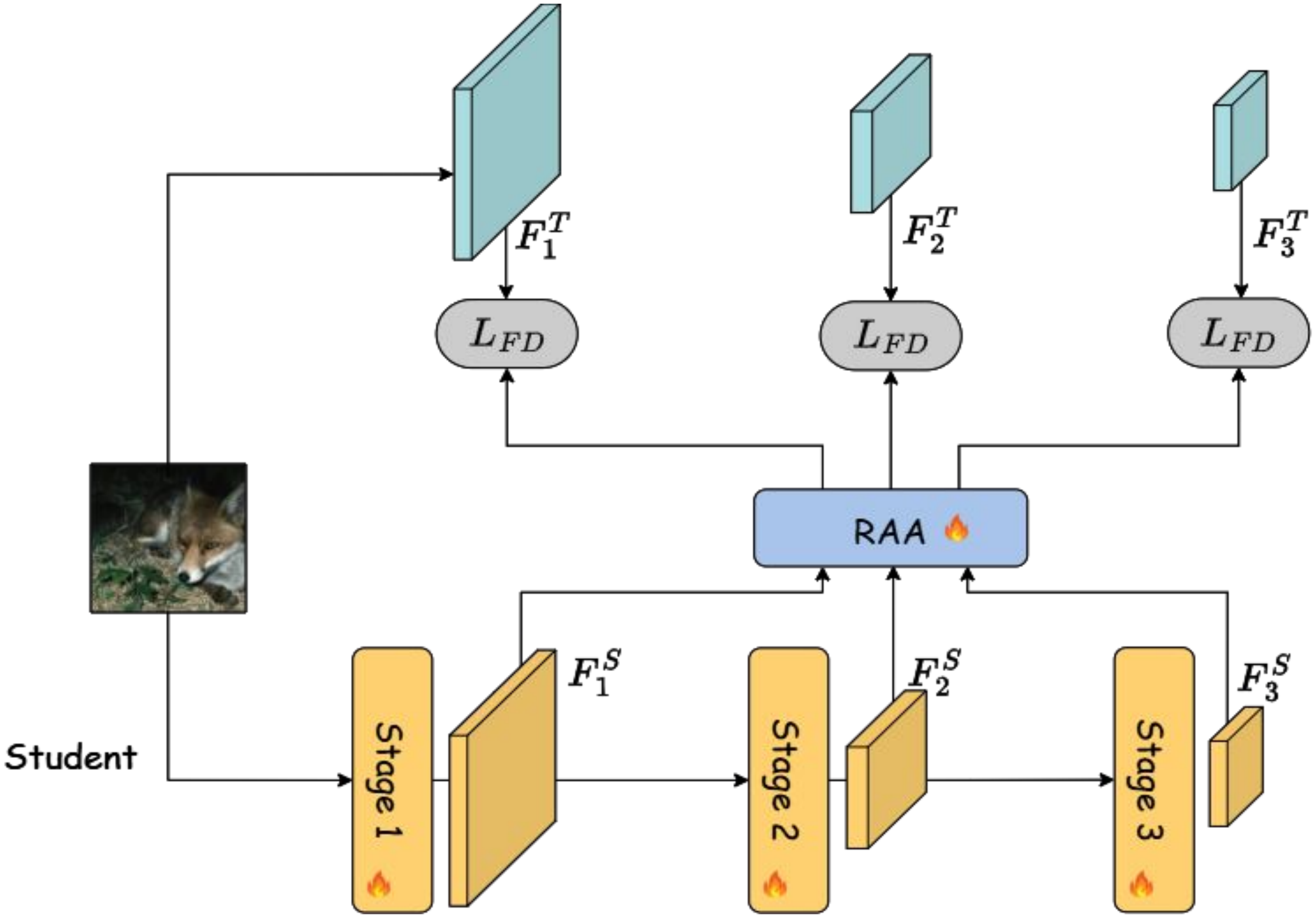
PAT Framework



We introduce a Perspective-Aware Teaching (PAT) framework to enable distillation in feature space across diverse architecture, which consist of two key components, RAA and AFP.

Region-Aware Attention

Mitigate "View Mismatch" via RAA modules.

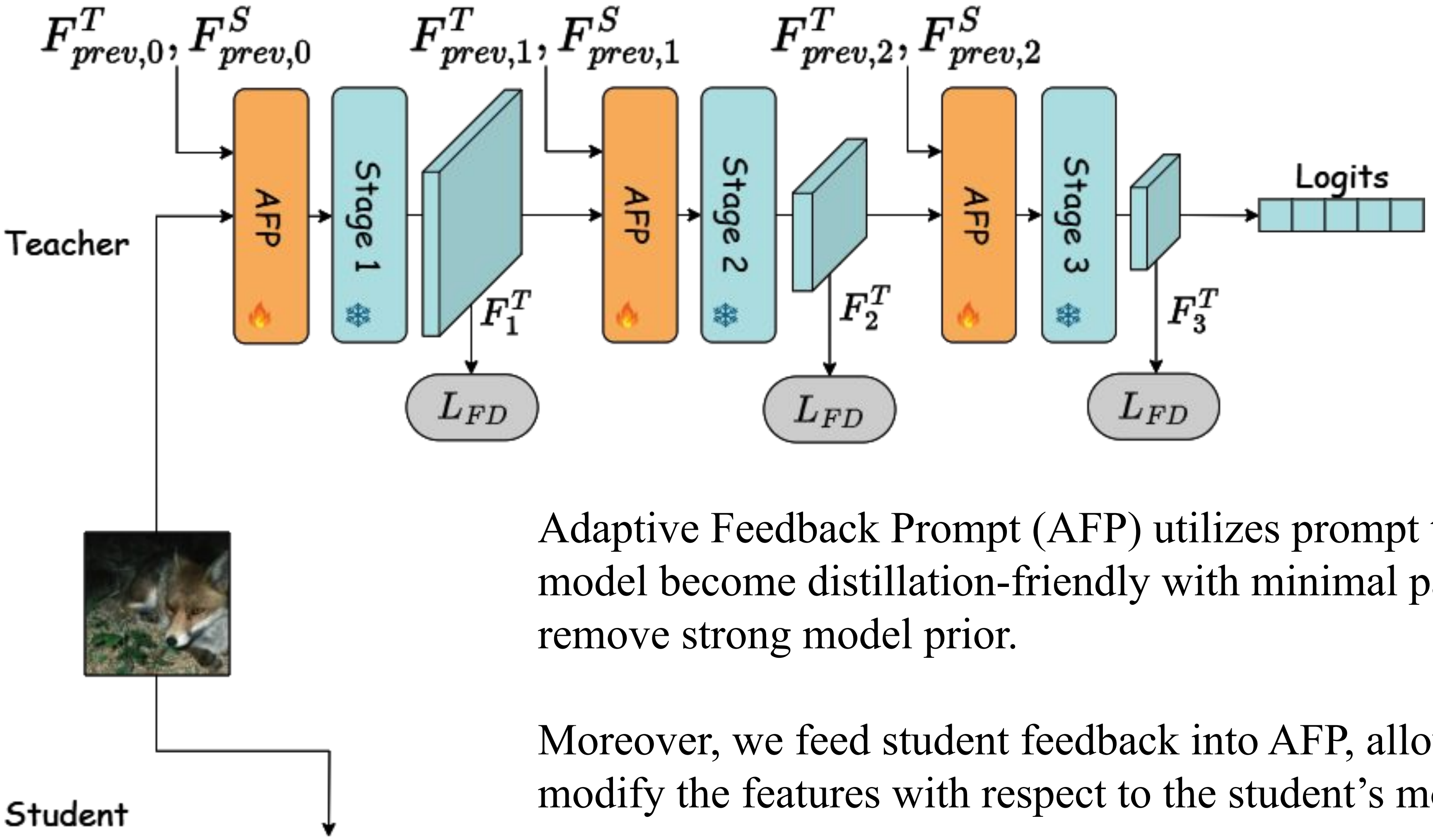


Region-Aware Attention (RAA) utilizes attention to enable student model to learn how to blend features from its various regions among stages to integrate a one with similar view as the teacher's.

After RAA, the traditional stage-wise feature matching can now be performed as the blended features have a similar view as the teacher's.

Adaptive Feedback Prompt

Mitigate "Teacher Unawareness" via AFP modules.



Adaptive Feedback Prompt (AFP) utilizes prompt tuning to make teacher model become distillation-friendly with minimal parameters, aiming to remove strong model prior.

Moreover, we feed student feedback into AFP, allowing teacher model to modify the features with respect to the student's model learning process.

Experiments

Image Classification (CIFAR-100)

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

Image Classification (ImageNet-1K)

AIMM LAB
NYCU & NTU

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

Our PAT achieves competitive results with the previous logits-based SOTA OFA, and further improves the performance of feature-based methods.

Object Detection

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

Our PAT outperforms previous method. This underscores that by mitigating view mismatch and teacher unawareness issues, the feature-mimicking technique can effectively leverage the abundant intermediate features for improved performance across classification and downstream tasks.

Conclusion

Conclusion

- **Perspective-Aware Teaching (PAT)**
  - Addressing view mismatch problem via Region-Aware Attention (RAA)
    - Let the student model learn how to reblend features from different patches and stages to achieve a similar perspective with the corresponding teacher features via the attention mechanism.
  - Solving teacher unawareness problem via Adaptive Feedback Prompt (AFP)
    - Allow the teacher model to remove distillation unfriendly feature and dynamically adapt its features in response to the student model's feedback via prompt tuning methods.
- **Corresponding results on CIFAR-100, ImageNet-1K, and COCO demonstrate the effectiveness of the proposed generic heterogeneous KD method PAT.**
  - Achieve SOTA on CIFAR-100, ImageNet-1K, and COCO


Thank you
for
listening!