UFD-KD: Unified Frequency Decoupled Knowledge Distillation

Sihan Lu[1]
1120212952@bit.edu.cn
Yang Zheng[2]
yang.zheng@ia.ac.cn
Jie Liu[3]
ailj@bupt.edu.cn
Zhenghao Xi[4]
zhenghaoxi@hotmail.com

[1] School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China

[2] Institute of automation
Chinese Academy of Sciences
Beijing, China

[3] The School of Artificial Intelligence
Beijing University of Posts and Telecommunications
Beijing, China

[4] Shanghai University of Engineering Science
Shanghai, China

Abstract

In this paper, we question whether the feature distillation has the same property as logits distillation, that its learning mainly focuses on one or several principal components, easily missing minor but informative features. In principle, the distribution of feature maps is unlikely to the logits intuitively, whose distribution is naturally imbalanced. However, we surprisingly find that the feature distribution becomes extremely imbalanced after applying the discrete cosine transform (DCT), which shows attributes similar to logits distillation. Inspired by this, we propose Unified Frequency Decoupled Knowledge Distillation (UFD-KD), which is designed to address the negative effect of the imbalance of feature distribution. Specifically, UFD-KD applies DCT to spatial and channel dimensions respectively, attending to decouple the features based on the properties of frequency domain. For each dimension, we design a parameterized weight strategy to emphasize the minor and sparse features which used to be out of notice because of the principal features. Considering the orthogonality of the two dimensions, we assign them different weights to balance the overall feature alignment. In the validation experiments on Cifar-100, 11 teacher-student model pairs demonstrates excellent performance, achieving 3.1% accuracy gains for ResNet50→ResNet18. Furthermore, in large-scale validation experiments on ImageNet-1K, UFD maintained competitive performance improvement (71.98% in Swin-T→ResNet18). Additionally, we validate the task transfer capability on ADE20K, achieving 36.89% mIoU for DeepLabv3- ResNet101→DeepLabv3-ResNet18.

*This Work was completed during Sihan Lu's study period in the team of Yang Zheng and Jiu Liu.

1 Introduction

The rapid evolution of deep neural network architectures, such as ResNet and Vision Transformers (ViTs) [3, 18], have driven unprecedented breakthroughs in various computer vision tasks, including classification [8, 11, 14, 19], object detection [9, 22, 33], and semantic segmentation [15, 25, 35]. However, these advancements incur high computational demands: modern vision transformers such as ViT-L suffer from quadratic complexity with high-resolution inputs [32], making them challenging for resource-constrained edge devices. This tension has spurred research into knowledge distillation (KD) [10], a technique aimed at transferring knowledge from complex teacher models to compact student networks to train lightweight deep neural network models.

Research in knowledge distillation has established two fundamental distillation paradigms: logits-based distillation [7, 10, 12, 34] that aligns output probability distributions through KL divergence minimization and feature-based (hint-based) distillation [20, 21, 23, 26] that transfers intermediate representations via layer-wise feature matching. Although logits-based approaches dominated early research for their simplicity and computational efficiency, the emergence of cross-architecture scenarios exposed their limitations in handling structural discrepancies between models [7]. Feature-based methods addressed this by exploiting spatial attention mechanisms and feature projection to help student models imitate the teacher, enabling heterogeneous models [29] (e.g. CNN vs ViT [17]) to share representation subspaces.

A critical challenge across both paradigms is how to align the representation discrepancy resulting from the model size difference and architectural heterogeneity between the teacher network and the student network. In feature distillation, some previous works focus on designing advanced feature alignment mechanisms to alleviate this issue, while methods for logits distillation [7, 16] always consider the neglected information from imbalanced distributions. Specifically, [34] presents the idea of decoupling distillation to enhance the non-salient logits. This observation motivates us to extend decoupling principles from logits to features, which is more friendly to learn fine-grained knowledge in the spatial dimension[12]. According to observations on the frequency domain, the direct components (DC) and alternating components (AC) are extremely imbalanced, shedding light on conducting the decoupling operation in the frequency domain. Unlike Logits-level operating solely in probability space, feature-level contains spatial dimensions (height × width) and channel-wise discriminative relationships. As a result, we explore how to conduct feature decoupling in both the spatial dimension and channel dimension for comprehensive feature alignment.

To achieve this goal, we propose Unified Frequency Decoupling Knowledge Distillation (UFD-KD), a simple and effective framework that establishes a comprehensive frequency decoupling mechanism for knowledge transfer. To be specific, UFD transfers features from spatial and channel domain to frequency domain through DCT respectively, and decouple according to DC and AC mask strategy to form a comprehensive decouple distillation paradigm, which contains two core designs:

- **Spectral Decoupling via Dual-Dimensional DCT**: UFD-KD innovatively applies DCT in spatial (H×W) and channel (C) dimensions, which isolates structural contours (DC) from textural details (AC) in spatial domain, while decoupling category semantics (DC) and instance-specific patterns (AC) in channel domain.
- **Dimensional Spectral Balance Mechanism**: UFD defines α-hyperparameters [$\alpha_{\mathrm{dc}}^{\mathrm{spatial}}$, $\alpha_{\mathrm{ac}}^{\mathrm{spatial}}$, $\alpha_{\mathrm{dc}}^{\mathrm{channel}}$, $\alpha_{\mathrm{ac}}^{\mathrm{channel}}$] to regulate DC/AC transfer ratios to emphasize alternating components that are usually overshadowed by direct components for detail enhancement.

Integrating these designs, UFD resolves the spectral entanglement in single-domain decomposition, achieving comprehensive 3D feature decoupling across H/W/C axes and balanced dimensional alignment through spectral allocation.

To validate UFD's effectiveness, we perform systematic experiments on several benchmarks and different tasks, including image classification and semantic segmentation. On Cifar-100 [13], we broadly validate the effectiveness of our method, achieving state-of-the-art performance across 11 network pairs. Furthermore, we evaluate on the large classification benchmark ImageNet-1K [2], achieving a Top-1 accuracy of 71.98% in Swin-T→ResNet18 distillation, outperforming OFA by a margin of 0.13%. Lastly, to test UFD applicability in other vision tasks, we conduct the experiments on ADE-20K [37, 38], achieving 36.89% mIoU on ResNet101→ResNet18 network pairs with DeepLab-v3 [1] segmentation framework. Our comprehensive evaluation demonstrates UFD's superiority.

2 Related Works

This section briefly introduces the advanced methodology that inspired our work.

Originating from Hinton's seminal work in 2014 [10], which built upon earlier CNN compression attempts, knowledge distillation utilizes soft labels to transfer "dark knowledge," enabling students to mimic the decision logic of complex teachers. Early studies focused on model compression by adjusting the output distribution entropy via temperature parameters.

Decoupled Knowledge Distillation (DKD) pioneered logits-level semantic disentanglement [34] by separating target-class (TCKD) and non-target-class (NCKD) knowledge components. Its improvement in performance demonstrates the effectiveness of decoupling in mitigating student overfitting to teacher biases.

Despite DKD's success in logits-level decoupling, its inherent limitation in accessing intermediate feature hierarchies (e.g.entangled edge-texture semantics in shallow layers) motivated the shift toward feature distillation methods [20, 21, 23, 26]. These approaches focus on aligning intermediate representations through spatial attention mechanisms and relational graph modeling. However, conventional feature distillation predominantly operates in the spatial domain, making it susceptible to input degradations like noise and blur. This limitation exposes the need for robust representation learning, catalyzing frequency-domain explorations.

Rooted in the Frequency Principle [36], which states that neural networks prioritize learning low-frequency components encoding global structures (e.g.object shapes) over high-frequency local details (e.g.surface textures) [30], we leverage frequency decomposition for feature disentanglement. ScaleKD further bridges spatial and frequency domains via Discrete Cosine Transform (DCT) [5], revealing the complementary role of spectral alignment in spatial feature matching.

Although existing feature-based distillation methods have achieved remarkable performance, only a few of them primarily focus on decoupling objects [6] or decoupling the foreground and background [31] within single domains. Notably, none addresses spatial and channel-wise disentanglement simultaneously.

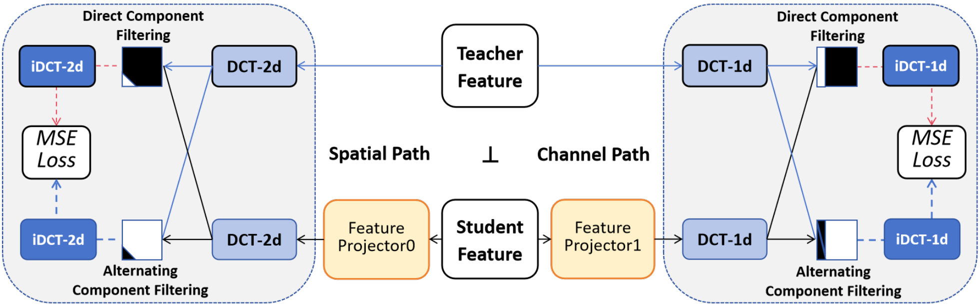


Figure 1: Unified Frequency Decoupling(UFD-KD):UFD-KD achieves architecture-agnostic alignment through spatial frequency mapping and channel frequency normalization, proposing a duel-dimension distillation framework to bridge the feature mismatch (in form of spectral discrepancy) observed in 3.1.

3 Methodology

3.1 Spectral Discrepancy Analysis

In UFD, we apply Discrete Cosine Transform(DCT) to transform feature from space domain to frequency domain. The DCT of a feature map $F \in \mathbb{R}^{H\times W\times C}$ can be formulated as:

$$\mathcal{F}_{dct}(u,v,c) = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} \Lambda(u,v) F(i,j,c) \cos\left(\frac{(2i+1)u\pi}{2H}\right) \cos\left(\frac{(2j+1)v\pi}{2W}\right). \quad (1)$$

where $\Lambda(u,v)$ denotes normalization coefficients, u,v are the frequency indices for horizontal (u) and vertical (v) components in the frequency domain, and i,j are spatial indices for height (i) and width (j) in the input feature.

To better look into the mismatch problem across models, we analyze spectral energy allocation across ViT and CNN architectures through block-wise decomposition: (1) DC/AC ratios in spatial (H×W) vs. channel (C) dimensions reveal transformer and CNN divergences; (2)Jensen-Shannon divergence of AC components quantifies cross-architecture mismatch.

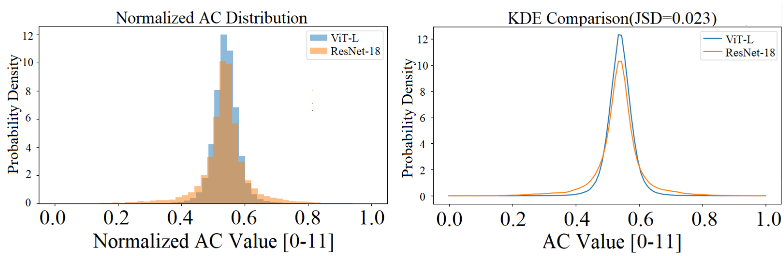


Figure 2: Model Spectral Divergence [(Left):Vit-Large, (Right):ResNet-18]: The results reveal spectral mismatches among different models, particularly in low-frequency structural preservation versus high-frequency detail encoding.

Through DCT, we identify a pattern that ViT/Swin exhibits over 70% DC energy concentration, while ResNet18 shows over 60% AC energy in spatial domain (H×W). This systematic

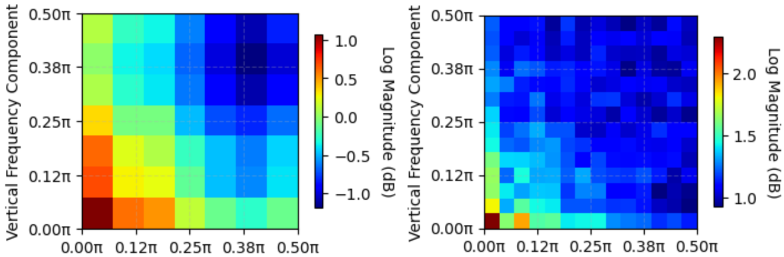


Figure 3: Distribution Heatmaps of the models: [(left):ResNet-18 heatmap, (right):ViT-L heatmap].

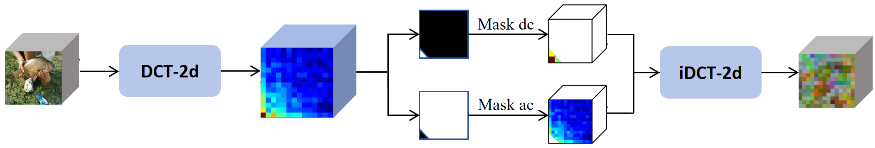


Figure 4: Flowchart of frequency filtering strategy

spectral discrepancy (the visualization of ViT-Large and ResNet18 in Figure 2 and 3 as an example) motivates our gradient conflict analysis and frequency decoupling design.

This distinction between ViT-L and ResNet18 reflects the mismatch during distillation due to competing low/high frequency alignment objectives, indicating difference between DC and AC. Based on this, we establish the framework of frequency-domain feature decoupling.

3.2 UFD-KD Framework

3.2.1 Feature Decoupling

The proposed UFD-KD method establishes frequency-decoupled alignment through Discrete Cosine Transform (DCT), achieving independent optimization of DC and AC components:

$$\frac{\partial \mathcal{L}_{\text{UFD-KD}}}{\partial \hat{\mathcal{F}}_{\text{dc}}} = 2\alpha_{\text{dc}}(\hat{\mathcal{F}}_{\text{dc}} - \mathcal{F}_{\text{dc}}), \quad \frac{\partial \mathcal{L}_{\text{UFD-KD}}}{\partial \hat{\mathcal{F}}_{\text{ac}}} = 2\alpha_{\text{ac}}(\hat{\mathcal{F}}_{\text{ac}} - \mathcal{F}_{\text{ac}}). \tag{2}$$

Based on the property of frequency domain, we propose a frequency-aware masking strategy to decouple structural (DC) and detail (AC) components after DCT transform. Given transformed features $\mathcal{F}_{dct}$ via DCT, the decoupling is achieved by:

$$\mathcal{F}_{dc} = \mathcal{F}_{dct} \odot M_{dc}, \quad \mathcal{F}_{ac} = \mathcal{F}_{dct} \odot (1 - M_{dc}), \tag{3}$$

where $M_{dc} \in \{0,1\}^{H \times W \times C}$ is a binary mask that preserves only the DC coefficients (u = 0, v = 0) in the spatial frequency domain, eliminating cross-term interference by isolating the DC / AC optimization pathways, as shown in Figure 4.

3.2.2 Dual-Domain Decomposition

We apply the basic idea of feature decoupling to both spatial and channel dimensions to fully improve the learning process.

- Spatial Decomposition (2D DCT): Processes spatial correlations while preserving channel independence:

$$\mathcal{F}_{spatial}(u,v,c) = \frac{2}{\sqrt{HW}} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} F(i,j,c) \cos\left[\frac{(2i+1)u\pi}{2H}\right] \cos\left[\frac{(2j+1)v\pi}{2W}\right]. \quad (4)$$

The decoupling strategy of spatial dimension is designed as follows to extract direct components and alternating components:

$$\begin{cases} \mathcal{F}_{dc-spatial} = \mathcal{F}_{spatial} \odot M_{dc}\ , \quad M_{dc}[0,0,c] = 1 \\ \mathcal{F}_{ac-spatial} = \mathcal{F}_{spatial} \odot (1 - M_{dc})\,. \end{cases} \quad (5)$$

- Channel Decomposition (1D DCT): Models cross-channel relationships independently per spatial location:

$$\mathcal{F}_{channel}(i,j,k) = \sqrt{\frac{2}{C}} \sum_{c=0}^{C-1} F(i,j,c) \cos\left[\frac{(2c+1)k\pi}{2C}\right]. \quad (6)$$

The decoupling strategy of channel dimension is designed as follows:

$$\begin{cases} \mathcal{F}_{dc-channel} = \mathcal{F}_{channel} \odot M_{dc}\ , \quad M_{dc}[h,w,0] = 1 \\ \mathcal{F}_{ac-channel} = \mathcal{F}_{channel} \odot (1 - M_{dc})\,. \end{cases} \quad (7)$$

The separation of spatial (2D) and channel (1D) DCT operations is grounded in their orthogonal basis functions. Let $\phi_u(i) = \cos\left(\frac{(2i+1)u\pi}{2H}\right)$ *and* $\psi_k[c] = \cos\left(\frac{(2c+1)k\pi}{2C}\right)$ denote spatial and channel basis functions respectively. Their orthogonality satisfies the following:

$$\sum_{i=0}^{H-1} \phi_{u_1}(i)\phi_{u_2}(i) = \delta(u_1 - u_2), \quad \sum_{c=0}^{C-1} \psi_{k_1}(c)\psi_{k_2}(c) = \delta(k_1 - k_2). \quad (8)$$

The orthogonal decomposition ensures spatial frequencies (u, v) preserve geometric transformations while channel frequencies (k) maintain cross-feature interactions independent of spatial locations, enabling independent optimization pathways. The joint spatial-channel representation forms a complete feature basis that diminishes information loss during dual-domain transformation, theoretically enabling reconstruction of original features.

3.2.3 Loss Formulation

The distillation loss function of UFD-KD integrates dual-domain frequency alignment through spatial and channel components:

- **Spatial Domain Alignment**:
  Aligns DC and AC components in the spatial dimension (H×W) to enforce structural contour prior transfer via channel averaging constraints:

$$\mathcal{L}_{\mathrm{dc}}^{\mathrm{spatial}} = \frac{1}{C} \sum_{c=1}^{C} \left\| \mathcal{F}_S^{\mathrm{dc}}(0,0,c) - \mathcal{F}_T^{\mathrm{dc}}(0,0,c) \right\|_2^2 \quad (9)$$

$$\mathcal{L}_{\mathrm{ac}}^{\mathrm{spatial}} = \frac{1}{C(H+W-1)} \sum_{c=1}^{C} \sum_{u+v>0} \left(\mathcal{F}_S^{\mathrm{ac}}(u,v,c) - \mathcal{F}_T^{\mathrm{ac}}(u,v,c)\right)^2 \quad (10)$$

where Equation (9) aligns structural contours through DC components ($\mathcal{F}_{Model}^{\mathrm{dc}}(0,0,c)$), while Equation (10) preserves cross-layer detail consistency via AC constraints.

- **Channel Domain Alignment**:
  Constrains DC responses in the channel dimension (C) to stabilize cross-feature discriminative relationships, while constrains AC enforces spatially averaged sparse constraints on channel AC components to eliminate redundant feature coupling:

$$\mathcal{L}_{\text{dc}}^{\text{channel}} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left\| \mathcal{F}_S^{\text{dc}}(i,j,0) - \mathcal{F}_T^{\text{dc}}(i,j,0) \right\|_2^2 \tag{11}$$

$$\mathcal{L}_{\text{ac}}^{\text{channel}} = \frac{1}{HW(C-1)} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k>0}^{C} (\mathcal{F}_S^{\text{ac}}(i,j,k) - \mathcal{F}_T^{\text{ac}}(i,j,k))^2 \tag{12}$$

The unified loss integrating these components of UFD-KD module and original KD through weighting parameters is defined as follows:

$$\begin{aligned} \mathcal{L}_{UFD} = \mathcal{L}_{ori} + (\alpha_{\text{dc}}^{\text{sp}} \mathcal{L}_{\text{dc}}^{\text{spatial}} + \alpha_{\text{ac}}^{\text{sp}} \mathcal{L}_{\text{ac}}^{\text{spatial}})_{\text{Spatial Loss}} \\ + (\alpha_{\text{dc}}^{\text{ch}} \mathcal{L}_{\text{dc}}^{\text{channel}} + \alpha_{\text{ac}}^{\text{ch}} \mathcal{L}_{\text{ac}}^{\text{channel}})_{\text{Channel Loss}} + \mathcal{L}_{kd} \end{aligned} \tag{13}$$

where [$\alpha_{\text{dc}}^{\text{spatial}}$, $\alpha_{\text{ac}}^{\text{spatial}}$, $\alpha_{\text{dc}}^{\text{channel}}$, $\alpha_{\text{ac}}^{\text{channel}}$] is predefined param, $\mathcal{L}_{ori}$ is the cross-entropy loss, while $\mathcal{L}_{kd}$ is the vanilla logits-based KD loss which is widely used in previous research in knowledge distillation.

4 Experiment

4.1 Experiment Setup

We validate UFD-KD through three axes: (1) Image classification (CIFAR-100 [13]/ImageNet-1K [2]), (2) Cross-task adaptation (ADE20K[37, 38] semantic segmentation), and (3) ablation analysis. Datasets & Training configurations are as follows:

Image Classification. For CIFAR-100, we utilize 11 heterogeneous teacher-student pairs spanning CNNs(ResNet-18 [8]/MobileNetV2 [24]), ViTs(DeiT-T [28]/Swin-P [18]), and MLP-Mixers [27], with all input images resized to 224×224 through bicubic interpolation. Training employs AdamW optimizer with cosine learning rate decay (initial lr=0.001, weight decay=0.05) and standard augmentations (random crop, horizontal flip, color jitter). Baselines: FitNet [23], CC [21], RKD [20], CRD [26], KD [10], DKD [34], DIST [12], OFA [7].

For ImageNet1K, we adopt SGD with momentum (0.9), stepwise lr(0.1) scheduling (10× drops at 30/60/90 epochs), and label smoothing (0.1) to prevent overfittingand for 2 pairs(ResNet34/Swin-T->ResNet18), with image resolution of 224×224. Baselines: CC [21], RKD [20], CRD [26], KD [10], DKD [34], DIST [12], OFA [7], NORM [16].

Semantic Segmentation. We adopt DeeplabV3 segmentation framework [1], ResNet-18 [8] and Mobilenetv2 [24] for student backbones, Resnet-101 for teacher backbone on ADE20K [40K iterations, bs=24, lr=0.02, optimized by SGD with momentum of 0.9]. Baseline: Af-DCD [4].

4.2 Main Results

4.2.1 Results on Image Classification

Cifar-100. As shown in Table 1, for cross-architecture distillation, it attains 83.13% accuracy in Swin-T→ResNet18 transfers (vs. OFA's 80.54%) and 80.87% in ViT-S→MobileNetV2 (vs.

Teacher		From Scratch		Hint-based				Logits-based				
		T.	S.	FitNet	CC	RKD	CRD	KD	DKD	DIST	OFA	UFD
					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>	**83.13**
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>	**81.99**
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>	**80.74**
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>	**83.03**
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>	**80.87**
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>	**79.77**
					ViT-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>	**78.83**
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>	**75.15**
ConvNeXt-T	Swin-P	88.41	72.63	24.06	72.63	71.73	67.09	76.44	76.80	76.41	<u>78.32</u>	**80.23**
Mixer-B/16	Swin-P	87.29	72.63	75.20	73.32	70.82	67.03	75.93	76.39	75.85	<u>78.93</u>	**79.63**
					Homogeneous students							
ResNet50	ResNet18	87.02	74.01	–	–	–	–	77.11	78.52	76.34	–	**82.14**

Table 1: Results on CIFAR-100. Aligning with OFA [7], we compared 11 teacher-student pairs of three architecture types. Evaluation metric: Top-1 accuracy (%). **Bold** denotes our UFD-KD results, underlined values indicate best baseline results (the data are basically from OFA's experiments [7], and "–" marks untested configurations.

Teacher		Student		Methods								
Model	Acc	Model	Acc	CC	RKD	CRD	KD	DKD	DIST	OFA	NORM	UFD
Swin-T	81.38	ResNet18	69.75	70.07	68.89	69.09	71.14	71.10	70.91	<u>71.85</u>	–	**71.98**
ResNet-34	73.31	ResNet18	69.75	69.96	71.34	71.17	70.68	–	71.89	–	<u>72.14</u>	**72.22**

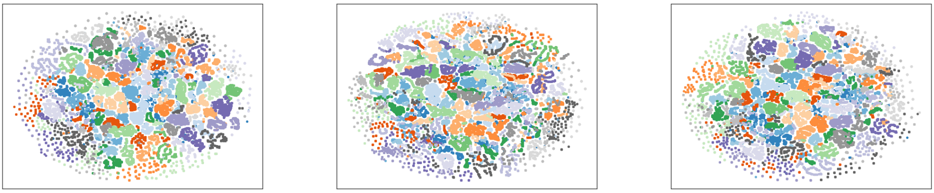
Table 2: Results on ImageNet-1K. Accuracy (%) comparison between teacher-student pairs. 'NORM [16]' and 'OFA [7]' are the SOTA method for each pair respectively, **bold** highlights our UFD-KD results, "–" indicates unreported configurations.

CRD's 78.14%), demonstrating robust capability in bridging architectural discrepancies.UFD-KD maintains consistent superiority through frequency-domain decoupling, outperforming the second-best baselines by 2.55% on average. The framework particularly benefits lightweight models—MobileNetV2 students gain 1.25-3.07% accuracy improvements. Homogeneous distillation also validates its universality, with ResNet50→ResNet18 achieving 82.14% .

ImageNet-1K. Preliminary results demonstrate UFD's effectiveness in large-scale cross-architecture distillation. As shown in Table 2, for Swin-T→ResNet18 pairs, UFD achieves competitive performance (71.98% accuracy) against eight baselines. In ResNet-34→ResNet18 distillation, UFD maintains performance consistency despite teacher complexity variations, while conventional methods like CC and DIST fail to generalize across all scenarios (RKD: 69.96% vs. KD: 71.17%).

4.2.2 Results on Semantic Segmentation

ADE20K. The semantic segmentation experiments in Table 3 demonstrate UFD-KD's cross-task effectiveness. The framework achieves consistent improvements across architectures: ResNet18 student attains 36.89 mIoU (+3.58% over baseline) and MobileNetV2 reaches 36.24 mIoU (+1.92%), outperforming Af-DCD by 0.68% and 0.27% respectively. The visualization comparison is shown in Figure 5



(a) (b) (c)

Figure 5: The t-SNE visualizations of original and UFD-trained ResNet18 on the first 50 ADE20K classes which demonstrate enhanced feature discriminability and better intra-class compactness. (a): teacher model, (b): Original ResNet18, (c): UFD-trained ResNet18.

Teacher	Student	From Scratch		KD		Af-DCD		UFD-KD	
		T. (mIoU)	S. (mIoU)	pixAcc	**mIoU**	pixAcc	**mIoU**	pixAcc	**mIoU**
ResNet-101	ResNet18	42.70	33.91	76.32	34.88	77.43	<u>36.21</u>	77.73	**36.89**
ResNet-101	MobileNetV2	42.70	34.32	76.74	34.92	77.28	<u>35.97</u>	77.62	**36.24**

Table 3: Semantic Segmentation Results on ADE20K. Evaluated by pixel accuracy (pixAcc%) and mean IoU (mIoU%). The baseline (From Scratch) demonstrates the original teacher-student performance gap. Af-DCD [4] represents the state-of-the-art method, while **bold** hightlights our UFD-KD results. Input: 512×512, iteration: 40k.

4.3 Ablation Study

Ablation experiments are conducted on Cifar-100 to test the effect of each component in UFD-KD. Key findings of Table 4 demonstrate:

Multi-module Synergy. Full combination of DC/AC in both spatial and channel dimensions achieves optimal performance. For ResNet18, this configuration boosts Top-1 accuracy by 1.80% compared to baseline (82.34% vs 80.54%), indicating that the complementary nature of DC and AC creates general enhancement. Combined channel and spatial attention modules enhance performance synergistically, as evidenced by MobileNetV2's +1.13% gain with full spatial-channel modules versus +0.48% with spatial-only.

Architecture Sensitivity. ViT-based students show greater variance. DeiT-T with full modules gains 3.09% improvement, but improper combinations (e.g.Swin-P with spatial-only) degrade performance (-1.45%). This aligns with findings that transformer architectures exhibit higher parameter sensitivity, requiring balanced spectral allocation across network depth.

5 Conclusion

The proposed Unified Frequency Decoupled Knowledge Distillation (UFD-KD) establishes a new paradigm for knowledge transfer through fundamental rethinking of feature representation and feature decoupling in spectral domains, achieving: (1) cross-dimensional efficiency gain via spatial-channel coordinated modulation with frequency-aware gates; (2) better cross-architecture compatibility via DCT-based dual-domain normalization unifying heterogeneous representations. Future work will focus on architecture-aware dynamic weight learning networks with parameter-efficient adaptation modules, where meta-learned controllers generate layer-specific balancing coefficients (α) based on architecture pairs.

Teacher	Student	Spatial		Channel		Top-1	Δ	Teacher	Student	Spatial		Channel		Top-1	Δ
		DC	AC	DC	AC					DC	AC	DC	AC		
CNN-based students															
Swin-T	ResNet18	✓	✓	✓	✓	82.34	+1.80	Swin-T	MobileNetV2	✓	✓	✓	✓	82.11	+1.13
		✓		✓		81.14	+0.60			✓		✓		80.44	-0.54
		✓	✓			82.03	+1.49			✓	✓			81.46	+0.48
						80.54	–							80.98	–
ViT-based students															
ConvNeXt-T	DeiT-T	✓	✓	✓	✓	78.85	+3.09	ConvNeXt-T	Swin-P	✓	✓	✓	✓	80.33	+2.01
		✓		✓		74.64	-1.12			✓		✓		76.87	-1.45
		✓	✓			77.65	+1.89			✓	✓			79.72	+1.40
						75.76	–							78.32	–
Homogeneous students															
ResNet-50	ResNet18	✓	✓	✓	✓	82.84	+3.03								
		✓		✓		81.33	+1.52								
		✓	✓			82.15	+2.34								
						79.81	–								

Table 4: UFD-KD Ablation Analysis on CIFAR-100. Three architecture groups showing Top-1 accuracy (%) with spatial/channel DC/AC module combinations. Checkmarks (✓) indicate activated modules. Δ indicates improvement over OFA baseline [7]. Baseline results marked "–".

Acknowledgements

This work has been supported by the National Key R&D Program of China under Grant NO. 2024YFF0907200.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[4] Jiawei Fan, Chao Li, Xiaolong Liu, Meina Song, and Anbang Yao. Augmentation-free dense contrastive knowledge distillation for efficient semantic segmentation. In *NeurIPS*, 2023.

[5] Jiawei Fan, Chao Li, Xiaolong Liu, and Anbang Yao. Scalekd: Strong vision transformers could be excellent teachers. In *NeurIPS*, 2024.

[6] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *CVPR*, 2021.

[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 *NeurIPS*, 2023.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2018.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[11] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. In *CVPR*, 2019.

[12] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *NeurIPS*, 2022.

[13] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[15] Jin Liu, Yi Pan, Min Li, Ziyue Chen, Lu Tang, Chengqian Lu, and Jianxin Wang. Applications of deep learning to mri images: A survey. *Big Data Mining and Analytics*, (1):1–18, 2018.

[16] Xiaolong Liu, Lujun Li, Chao Li, and Anbang Yao. Norm: Knowledge distillation via n-to-one representation matching. In *ICLR*, 2023.

[17] Yufan Liu, Jiajiong Cao, Bing Li, Weiming Hu, Jingting Ding, and Liang Li. Cross-architecture knowledge distillation. In *ACCV*, 2022.

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[19] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 2018.

[20] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019.

[21] Baoyun Peng, Xiao Jin, Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Yu Liu, Dongsheng Li, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *ICCV*, 2019.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2016.

[23] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.

[24] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

[25] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2016.

[26] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2022.

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

[28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention. In *ICML*, 2021.

[29] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation, 2022.

[30] Zhi-Qin John Xu, Yaoyu Zhang, and Tao Luo. Overview frequency principle/spectral bias in deep learning. *Communications on Applied Mathematics and Computation*, pages 1–38, 2024.

[31] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *CVPR*, 2022.

[32] Zhengqing Yuan, Rong Zhou, Hongyi Wang, Lifang He, Yanfang Ye, and Lichao Sun. Vit-1.58b: Mobile vision transformers in the 1-bit era, 2024.

[33] Taolin Zhang, Sunan He, Dai Tao, Bin Chen, Zhi Wang, and Shu-Tao Xia. Vision-language pre-training with object contrastive learning for 3d scene understanding. In *AAAI*, 2023.

[34] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022.

[35] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[36] Zhi-Qin John Xu Zhi-Qin John Xu, Yaoyu Zhang Yaoyu Zhang, Tao Luo Tao Luo, Yanyang Xiao Yanyang Xiao, and Zheng Ma Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, (5):1746–1767, 2020. ISSN 1815-2406.

[37] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

[38] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019.