SFKD: Spatial–Frequency Joint-Aware Heterogeneous Knowledge Distillation via Multi-Level Wavelet Spectral Interaction

Cuipeng Wang[1,2] and Haipeng Wang[1,2]*

[1] Key Laboratory for Information Science of Electromagnetic Waves, Ministry of Education, Fudan University, Shanghai, China
[2] Discipline and Technology Center of Microwave Vision Intelligent Sensing, Fudan University, Shanghai, China
cpwang23@m.fudan.edu.cn, hpwang@fudan.edu.cn

Abstract. Most existing knowledge distillation methods focus on homogeneous models (*e.g.*, CNN→CNN), thereby overlooking the flexibility and potential of knowledge transfer across heterogeneous models. Due to intrinsic inductive bias discrepancies between heterogeneous models that cause spatial distribution inconsistencies, prior heterogeneous distillation methods often weaken or discard spatial information in heterogeneous representations. However, the spatial information in representations often encodes transferable global structural semantics as well as architecture-specific local details, and therefore should not be directly ignored. To better leverage the spatial information encoded in heterogeneous representations, we propose a Spatial–Frequency Joint-Aware Heterogeneous Knowledge Distillation framework (SFKD). By leveraging the complementary properties of wavelet transform spatial locality and Fourier representations in characterizing global energy distributions, we first apply multi-level discrete wavelet transform to explicitly decouple spatial information. The resulting wavelet sub-bands are further refined by a dual-stream dual-stage refinement module, and finally combined with a Gaussian-filtered frequency loss to selectively capture informative global information. Extensive experiments on multiple benchmark datasets under both homogeneous and heterogeneous models demonstrate the superiority of our method. Code is available at https://github.com/cpcpWang/SFKD.

Keywords: Heterogeneous Distillation · Discrete Wavelet Transform · Fast Fourier Transform

1 Introduction

With the rapid development of sophisticated architectures such as CNNs [8, 22, 32], Transformers [35, 36], and MLP-based models [3, 21, 37], deep learning has achieved remarkable success in computer vision. However, the continuous

* Corresponding author.

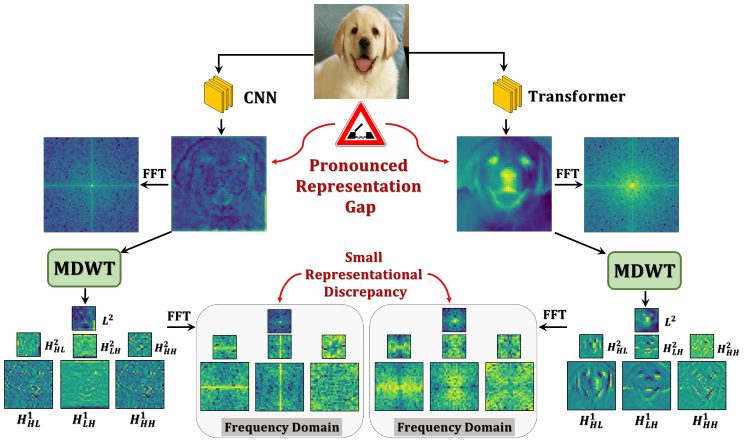


Fig. 1: Illustration of heterogeneous feature maps, multi-level wavelet sub-bands, and corresponding FFT spectral features. Heterogeneous feature maps exhibit substantial spatial discrepancies, while the wavelet sub-bands obtained through MDWT are further transformed by FFT, resulting in spectral features with reduced discrepancies.

increase in model parameters, while improving performance, also incurs substantial computational and memory overhead, posing significant challenges for deployment on resource-constrained edge devices. Knowledge Distillation (KD) has been widely recognized as an effective technique for model compression. The core idea is to transfer knowledge from a powerful teacher model to a compact student model, thereby improving the student's performance without increasing its model size.

Most existing knowledge distillation methods [1, 9, 10, 12, 20, 27, 28, 30, 34, 44] primarily focus on knowledge transfer between homogeneous models. However, merely restricting distillation to homogeneous models introduces several limitations. On the one hand, for a given student model, high-performing homogeneous teacher models suitable for distillation are often scarce, making the selection of an optimal teacher non-trivial. On the other hand, restricting distillation to homogeneous models may fail to exploit the complementary inductive biases inherent in heterogeneous models. Recent studies on hybrid architectures [17, 19] have shown that combining heterogeneous modules, such as CNNs and Transformers, can better leverage complementary strengths and yield superior performance. Consequently, relying solely on homogeneous teachers may prevent the student model from benefiting from such cross-architecture complementary knowledge, resulting in suboptimal distillation performance. For instance, as shown in Tab. 2 and Tab. 3, with our proposed method, distilling a ResNet18 student from a heterogeneous Swin-T teacher achieves better performance (72.54%) than using a homogeneous ResNet34 teacher (72.34%). This observation further highlights the potential advantages of cross-architecture distillation. Motivated by these insights, this work focuses on knowledge distillation across heterogeneous models, aiming to enlarge the teacher selection space and enhance distillation flexibility by leveraging complementary cross-architecture knowledge.

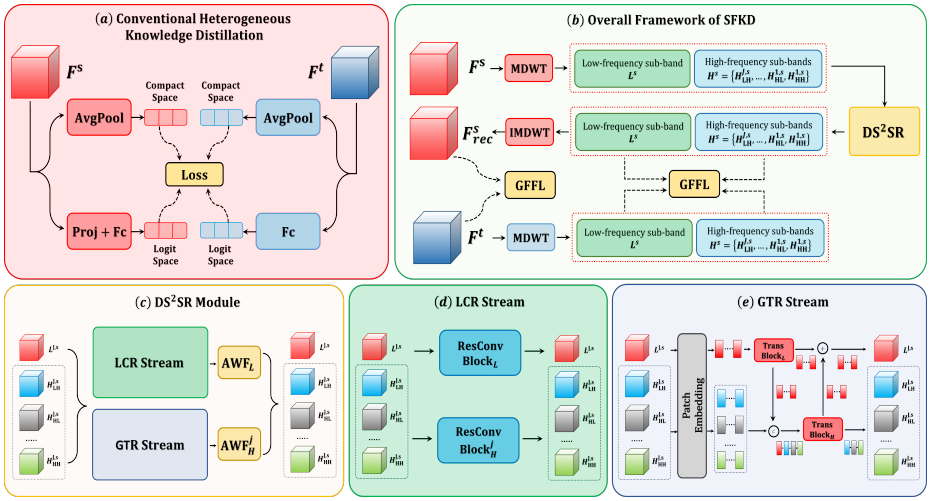


Fig. 2: Comparison between conventional heterogeneous distillation (a) and the proposed SFKD (b). (c) Detailed architecture of the DS^{2}SR module. The LCR stream and GTR stream are illustrated in (d) and (e), respectively. Unlike conventional methods that directly compress or suppress spatial information, SFKD explicitly decouples spatial information via MDWT, further refines wavelet sub-bands with DS^{2}SR, and finally performs frequency-domain alignment using GFFL, thereby enabling more effective exploitation of spatial information in heterogeneous representations.

Hao et al. [7] use CKA to analyze intermediate representations across heterogeneous architectures and reveal significant discrepancies between them. Such discrepancies primarily arise from intrinsic inductive bias differences across heterogeneous models [26,29]. CNNs [8,22,32] exhibit strong locality and translation invariance due to convolutional operations, whereas Transformers [3,21,37] and MLP-based models [35, 36] rely on patchification and long-range dependency modeling, leading to distinct structural properties.

Bridging this representational gap is crucial for effective knowledge distillation between heterogeneous models. Prior studies make significant efforts in this direction. However, we observe that, due to the pronounced spatial distribution discrepancies between heterogeneous representations, as illustrated in Fig. 1, several approaches often mitigate this gap by reducing explicit spatial dependencies, thereby limiting the effective utilization of spatial structures. For example, Hao et al. [7] maps intermediate features to the logit space to alleviate representational differences, Wu et al. [41] align representations in a compact embedding space, and Li et al. [16] introduces a hybrid bridging model with a structure-agnostic loss. However, spatial information in intermediate representations inherently encode both transferable global semantics and architecture-specific local details. Excessively discarding or suppressing such spatial information may limit the effective utilization of complementary cross-architecture knowledge.

To better exploit the spatial information encoded in representations of heterogeneous architectures, we propose a Spatial–Frequency Joint-Aware Heterogeneous Knowledge Distillation framework (**SFKD**). Specifically, we first apply

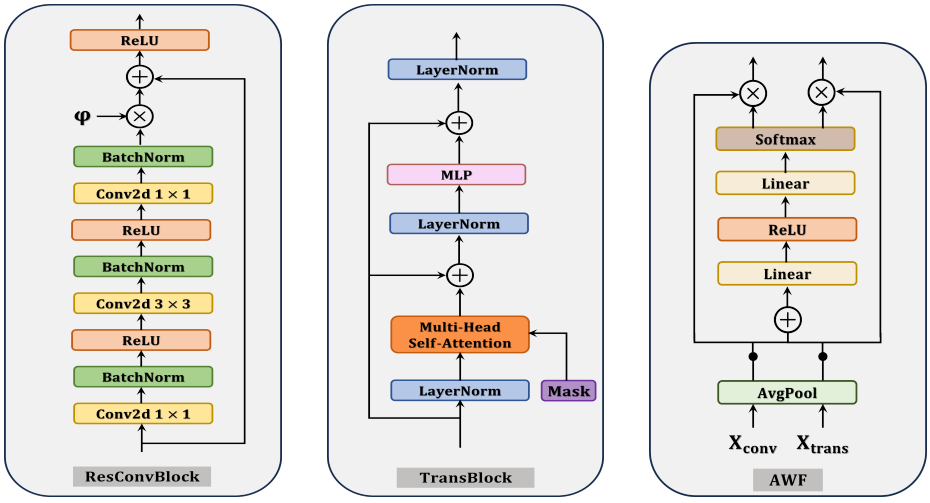
multi-level discrete wavelet transform to explicitly decouple teacher and student representations in the spatial domain. In the resulting wavelet spectrum, the low-frequency(LF) sub-band primarily captures global structural semantics, while the high-frequency(HF) sub-bands encode local details such as edges and textures. Due to the relatively limited representational capacity of the student network, its wavelet sub-bands often exhibit greater structural instability and distribution discrepancies compared to those of the teacher, which hinders effective feature alignment. To address this issue, we introduce a novel dual-stream dual-stage refinement module to further refine the student wavelet sub-bands, enhancing their structural consistency. Considering the spatial discrepancies between heterogeneous representations, directly aligning sub-bands in the spatial domain may be unstable. In contrast, the Fourier domain can characterize global energy distribution and overall structural patterns. Therefore, we design a Gaussian-filtered frequency loss to selectively emphasize the dominant information in each wavelet sub-band, facilitating effective heterogeneous knowledge transfer. Wavelet decomposition provides spatial locality, whereas Fourier transformation offers global spectral modeling. By integrating these complementary properties, SFKD establishes a spatial–frequency joint-aware distillation framework that effectively extracts and emphasizes informative spatial information in heterogeneous representations. Extensive experiments on multiple benchmark datasets demonstrate that SFKD consistently outperforms existing methods and achieves state-of-the-art or competitive performance across diverse settings.

In summary, the main contributions of our work are as follows:

- We provide a systematic analysis of existing heterogeneous distillation methods and identify their limited exploitation of spatial information in intermediate representations. To address this issue, we propose to integrate multi-level wavelet transform with Fourier-domain modeling to decouple, filter, and effectively transfer informative spatial information.
- We propose a Spatial–Frequency Joint-Aware Heterogeneous Knowledge Distillation framework (SFKD). The framework applies multi-level discrete wavelet transform to explicitly decouple spatial information in heterogeneous representations, incorporates a dual-stream dual-stage refinement module to enhance student wavelet sub-bands, and then employs a Gaussian-filtered frequency loss to facilitate efficient heterogeneous knowledge transfer.
- Extensive experiments on CIFAR-100 and ImageNet-1K under both heterogeneous and homogeneous settings demonstrate the superiority of SFKD in improving student performance, while also validating the effectiveness of decoupling and selectively exploiting spatial information in representations.

2 Related Work

We discuss **Related Work** and defer a concentrated account to the Appendix (see Sec. A).



(a) ResConvBlock **(b)** TransBlock **(c)** AWF

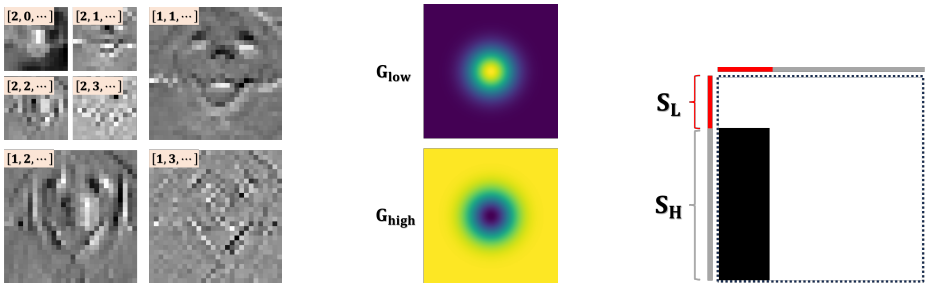
Fig. 3: The architectural details of the ResConvBlock, TransBlock, and Adaptive Weighted Fusion (AWF) modules.

3 Method

3.1 Overall Framework

As illustrated in Fig. 2, the proposed SFKD framework primarily consists of three core components: the Multi-level Discrete Wavelet Transform (**MDWT**) module, the Dual-Stream Dual-Stage Spectral Refinement (**DS^{2}SR**) module, and the Gaussian-Filtered Frequency Loss (**GFFL**). It should be noted that the student features $\mathbf{F}^s$ mentioned below refer to the representations that have been projected by a projector and aligned with the teacher features $\mathbf{F}^t$. For simplicity, this projection process is omitted in Fig. 2 and the following descriptions.

Compared with existing heterogeneous distillation methods [7, 16, 41] that typically weaken or discard spatial information in intermediate representations, we instead explicitly model such spatial characteristics in a structured manner. Specifically, we first apply a MDWT to both teacher and student representations, facilitating an explicit spatial decoupling of intermediate representations into LF and HF sub-band components. Considering the relatively limited representation capacity of the student network, its wavelet-decomposed sub-bands tend to exhibit greater structural instability and representational discrepancies compared to those of the teacher. Such discrepancies hinder effective alignment between student and teacher representations. To mitigate these discrepancies, we propose a novel dual-stream dual-stage spectral refinement (DS^{2}SR) module that adaptively refines the student's wavelet sub-bands, recovering structural details and global semantic consistency degraded in the student network. In Sec. 4.3, we validate the necessity of sub-band refinement and the effectiveness of the proposed DS^{2}SR module through ablation studies. Moreover, motivated



(a) 4D positional encoding **(b)** Gaussian filters **(c)** Attention mask

Fig. 4: Illustration of the proposed 4D positional encoding, Gaussian low-pass and high-pass filtering masks, and the designed attention mask. In Fig. 4c, the red lines denote low-frequency tokens, while the gray lines denote high-frequency tokens.

by the efficacy of Fourier spectra representations in characterizing stable characterization of global energy distribution and structural information, we propose a Gaussian-Filtered Frequency Loss (GFFL). By incorporating Gaussian-filtered mask, GFFL selectively emphasizes informative frequency components within each sub-band, thereby facilitating effective knowledge transfer and enabling joint spatial–frequency awareness in the distillation process.

3.2 Multi-level Discrete Wavelet Transform (MDWT)

Wavelet transform is a general signal decomposition technique that explicitly decouples low-frequency and high-frequency components within representations. In this work, we adopt the Haar wavelet transform [6] as the underlying decomposition operator. Given an input feature tensor $\mathbf{F} \in \mathbb{R}^{B\times C\times H\times W}$, we apply the discrete wavelet transform operator $\mathrm{DWT}(\cdot)$ independently to each sample in the batch to obtain one low-frequency sub-band $\mathbf{L}^1 \in \mathbb{R}^{B\times C\times \frac{H}{2}\times \frac{W}{2}}$ and three high-frequency sub-bands $\mathcal{H}^1 = \left\{\mathbf{H}^1_{LH}, \mathbf{H}^1_{HL}, \mathbf{H}^1_{HH}\right\} \in \mathbb{R}^{B\times C\times \frac{H}{2}\times \frac{W}{2}}$, corresponding to vertical, horizontal, and diagonal frequency components, respectively.

$$(\mathbf{L}^1, \mathcal{H}^1) = \mathrm{DWT}(\mathbf{F}) \tag{1}$$

By recursively applying the discrete wavelet transform to the low-frequency sub-band for J levels, we obtain the J-th level wavelet decomposition, which consists of a low-frequency sub-band $\mathbf{L}^J \in \mathbb{R}^{B\times C\times \frac{H}{2^J}\times \frac{W}{2^J}}$ and a high-frequency sub-band set $\mathcal{H}^J \in \mathbb{R}^{B\times C\times \frac{H}{2^J}\times \frac{W}{2^J}}$.

After performing a J-level discrete wavelet transform on the input feature map $\mathbf{F}$, we obtain the complete set of multi-level sub-bands, denoted as $\mathbf{X}$,

$$\mathbf{X} = \left\{\mathbf{L}^J, \mathcal{H}^J, \mathcal{H}^{J-1}, \ldots, \mathcal{H}^1\right\}. \tag{2}$$

We collectively denote these sub-bands as the J-level wavelet spectrum of $\mathbf{F}$, defined as:

$$\mathbf{X} = \mathrm{MDWT}(\mathbf{F}, J). \tag{3}$$

Since the discrete wavelet transform is inherently invertible, the original feature map $\mathbf{F}$ can be reconstructed through the corresponding J-level inverse DWT, defined as:

$$\mathbf{F} = \mathrm{IMDWT}(\mathbf{X}, J). \tag{4}$$

3.3 Dual-Stream Dual-Stage Spectral Refinement(DS^{2}SR)

Although the discrete wavelet transform explicitly decouples low- and high-frequency components, it does not inherently eliminate representational noise and structural instability. Moreover, since practical wavelet filters are implemented with finite-length kernels, they cannot achieve ideal band-limited separation [23, 33]. As a result, spectral leakage or residual frequency overlap across sub-bands may arise during multi-level decomposition [38, 39]. Consequently, several recent works [4, 5, 13, 18, 24, 40, 42] introduce additional refinement mechanisms after wavelet decomposition to enhance sub-band stability and discriminative capability. Inspired by these pioneering works and considering the inherent representation discrepancies arising from the divergent inductive biases of heterogeneous models, we propose a novel dual-stream dual-stage spectral refinement (DS^{2}SR) module tailored for heterogeneous knowledge distillation.

Convolutional and self-attention modules possess distinct and highly complementary modeling capabilities [26, 29]. Convolutional layers encode strong locality inductive bias, enabling the extraction of fine-grained patterns such as edges and textures, whereas self-attention mechanisms model long-range dependencies through multi-head attention, promoting globally coherent structural representations. To jointly leverage the complementary modeling capabilities of convolution and self-attention, while enabling architecture-agnostic refinement across heterogeneous representations, we design an novel and adaptive DS^{2}SR module to effectively refine wavelet sub-bands across heterogeneous architectures. The proposed DS^{2}SR module consists of two functionally distinct yet complementary branches: a Local Convolutional Refinement (**LCR**) stream and a Global Transformer Refinement (**GTR**) stream. The LCR stream consists of convolutional layers and focuses on refining fine-grained local structures within each wavelet sub-band. In contrast, the GTR stream leverages self-attention to capture long-range dependencies, thereby enhancing global structural coherence and contextual modeling. Each branch independently refines the wavelet sub-bands. As illustrated in Fig. 1, considering the distinct statistical properties and structural patterns of low-frequency and high-frequency components, we separately feed the refined low-frequency sub-band and the refined high-frequency sub-band set from both branches into distinct Adaptive Weighted Fusion (**AWF**) modules for frequency-specific integration. The detailed architectures of the LCR stream, the GTR stream, and the AWF module are introduced below.

Local Convolutional Refinement (LCR) Stream. Due to the distinct statistical distributions and structural characteristics of low- and high-frequency wavelet sub-bands, we apply two independent refinement stages to the low-frequency sub-band and the high-frequency sub-band set to enhance fine-grained

local structural modeling. The detailed architecture of the ResConvBlock is illustrated in Fig. 3a. ResConvBlock integrates convolutional operations with different strides and a weighted residual connection mechanism. By fusing the convolutionally refined features with the original input through weighted residual aggregation, it enhances local structural representations while preserving shallow feature information, thereby avoiding over-smoothing and structural degradation.

Specifically, after applying MDWT, the wavelet spectrum is organized into the low-frequency sub-band $\mathbf{L}^J$ and and the high-frequency sub-band set $\left\{\mathcal{H}^j\right\}_{j=1}^J$.

For the low-frequency sub-band $\mathbf{L}^J$, We feed it into a dedicated residual convolutional block (ResConvBlock_L) for refinement, yielding the refined low-frequency sub-band denoted as:

$$\mathbf{L}_{\text{conv}}^J = \text{ResConvBlock}_L\left(\mathbf{L}^J\right). \tag{5}$$

For the high-frequency sub-band set $\{\mathcal{H}^j\}_{j=1}^J$, each level is independently processed by a corresponding residual convolutional block (ResConvBlock_H^j), yielding:

$$\mathcal{H}_{\text{conv}}^j = \text{ResConvBlock}_H^j\left(\mathcal{H}^j\right), \quad j \in \{1, \ldots, J\}. \tag{6}$$

The final convolutional refinement output for the high-frequency components is denoted as $\mathcal{H}_{\text{conv}} = \{\mathcal{H}_{\text{conv}}^j\}_{j=1}^J$.

Global Transformer Refinement (GTR) Stream. While the LCR stream effectively captures fine-grained local structures, its strong locality inductive bias and limited receptive field restrict its capacity to model long-range dependencies and global contextual information. To overcome this limitation, we introduce a Transformer-based global refinement stream that leverages self-attention to aggregate global context and capture long-range structural dependencies in the wavelet sub-bands.

Before feeding the wavelet sub-bands into the TransBlock for refinement, we first perform tokenization for each level of the wavelet spectrum. Specifically, the spectrum is separated into the low-frequency sub-band $\mathbf{L}^J$ and and the multi-level high-frequency sub-band set $\left\{\mathcal{H}^j\right\}_{j=1}^J$. For each level j, the corresponding high-frequency sub-band set $\mathcal{H}^j$ is embedded through a convolutional projection layer. For the low-frequency sub-band $\mathbf{L}^J$, a dedicated convolutional projection layer is employed. The resulting tokens are denoted as $\mathbf{S}_L^J$ and $\mathcal{S}_H = \left\{\mathbf{S}_H^j\right\}_{j=1}^J$, respectively.

$$\mathbf{S}_L^J = \text{Conv2d}_L\left(\mathbf{L}^J\right) \tag{7}$$

$$\mathbf{S}_H^j = \text{Conv2d}_H^j\left(\mathcal{H}^j\right), \quad j = 1, \ldots, J \tag{8}$$

After convolutional embedding of the wavelet sub-bands, we further introduce positional encoding to preserve structural and hierarchical information. In standard Vision Transformers [3], each image patch is assigned a 2D absolute

positional encoding. However, such 2D absolute spatial encoding alone is insufficient to distinguish between different wavelet levels and sub-band types within each level. To address this limitation, we extend the standard 2D positional encoding by incorporating a wavelet-level index $j \in \{1, \ldots, J\}$ and a sub-band category indicator $d \in \{0, 1, 2, 3\}$, where $d = 0$ represents the low-frequency sub-band $\mathbf{L}$, while $d \in \{1, 2, 3\}$ correspond to distinct high-frequency orientations. Consequently, as illustrated in Fig. 4a, each token is assigned a 4D positional index $[j, d, \mathrm{Pos}_h, \mathrm{Pos}_w]$, where Pos_h and Pos_w denote the spatial coordinates. The resulting 4D positional index is encoded using the standard sinusoidal frequency-based positional encoding adopted in Vision Transformers [3], and subsequently added to the corresponding patch embedding to preserve both spatial layout and wavelet-domain hierarchical information.

Considering the distinct statistical distributions of low- and high-frequency wavelet sub-bands, as well as the presence of residual high-frequency components within the low-frequency sub-band [38, 39], existing approaches that independently refine the two components during global modeling [13, 24, 40] may limit the low-frequency representation from fully exploiting the structural cues encoded in the high-frequency sub-bands for effective suppression. Motivated by this observation, we design a dual-stage Transformer-based interactive refinement module. While modeling the low-frequency sub-band and the high-frequency sub-band set separately, we introduce a masked Transformer block to enable cross-frequency interaction, allowing the high-frequency components to guide the suppression of residual high-frequency noise in the low-frequency sub-band. The detailed architecture of the TransBlock is illustrated in Fig. 3b.

In the first stage, $\mathbf{S}_L^J$ is fed into a $\mathrm{TransBlock}_L$ without any attention mask to perform preliminary refinement on its smooth structural components. The resulting low-frequency tokens are denoted as:

$$\bar{\mathbf{S}}_L^J = \mathrm{TransBlock}_L\left(\mathbf{S}_L^J\right). \tag{9}$$

In the second stage, as illustrated in Fig. 2, the input consists of both the high-frequency token set $\mathcal{S}_H = \{\mathbf{S}_H^j\}_{j=1}^J$ and the refined low-frequency tokens $\bar{\mathbf{S}}_L^J$. To enable cross-frequency interaction, we introduce an attention mask within the $\mathrm{TransBlock}_H$, as illustrated in Fig. 4c. Specifically, high-frequency tokens are visible to low-frequency tokens so that high-frequency structural cues can guide the refinement of the low-frequency component. In contrast, low-frequency tokens are invisible to high-frequency tokens to avoid reverse interference. After the second stage, we obtain the further refined low-frequency tokens $\tilde{\mathbf{S}}_L^J$ and high-frequency token set $\tilde{\mathcal{S}}_H$:

$$\tilde{\mathbf{S}}_L^J,\ \tilde{\mathcal{S}}_H = \mathrm{TransBlock}_H\left(\left[\bar{\mathbf{S}}_L^J, \mathcal{S}_H\right], \mathbf{M}\right) \tag{10}$$

where $[\cdot, \cdot]$ denotes concatenation along the token dimension.

For the final transformer refinement outputs, the refined low-frequency sub-band is obtained by summing the two-stage results followed by reshaping:

$$\mathbf{L}_{\mathrm{trans}}^J = \mathrm{Reshape}\left(\bar{\mathbf{S}}_L^J + \tilde{\mathbf{S}}_L^J\right), \tag{11}$$

while the refined high-frequency sub-band set is obtained by directly reshaping:

$$\mathcal{H}_{\text{trans}} = \text{Reshape}\left(\tilde{\mathcal{S}}_H\right). \tag{12}$$

Adaptive Weighted Fusion (AWF). Given the complementary modeling capabilities of the LCR and GTR streams, we perform selective adaptive fusion on the wavelet sub-bands generated by the two streams. Furthermore, considering the distinct statistical distributions of low- and high-frequency wavelet sub-bands, we apply adaptive fusion separately to each component. The detailed architecture of the Adaptive Weighted Fusion (AWF) module is illustrated in Fig. 3c.The fused refined wavelet sub-bands are formulated as:

$$\mathbf{L}^J_{\text{ref}} = \text{AWF}_L\left(\mathbf{L}^J_{\text{conv}}, \mathbf{L}^J_{\text{trans}}\right), \tag{13}$$

$$\mathcal{H}^j_{\text{ref}} = \text{AWF}^j_H\left(\mathcal{H}^j_{\text{conv}}, \mathcal{H}^j_{\text{trans}}\right), \quad j \in \{1, \ldots, J\}, \tag{14}$$

3.4 Gaussian-Filtered Frequency Loss (GFFL)

In the early sections, we employ MDWT to explicitly decouple the intermediate representations of both teacher and student networks into low-frequency and high-frequency components. Specifically, we obtain the teacher sub-bands $\mathbf{L}^{J,t}$ and $\mathcal{H}^t$, as well as the student sub-bands. Considering the representational capacity gap between teacher and student networks, the student wavelet sub-bands typically exhibit greater structural instability. Therefore, we further refine the student sub-bands using the proposed DS^{2}SR, yielding the refined low-frequency sub-band $\mathbf{L}^{J,s}_{\text{ref}}$ and high-frequency sub-band set $\mathcal{H}^s_{\text{ref}}$.

Although wavelet transform effectively captures spatial locality, it does not fully exploit global frequency-domain characteristics. Moreover, due to intrinsic inductive bias discrepancies across heterogeneous architectures, not all spatial information in the wavelet sub-bands is equally transferable. To better leverage frequency-domain properties and selectively emphasize informative components, we propose a Gaussian-Filtered Frequency Loss (GFFL), enabling spatial–frequency joint-aware distillation.

Specifically, for both the teacher wavelet sub-bands $\mathbf{L}^{J,t}, \mathcal{H}^t$ and the refined student sub-bands $\mathbf{L}^{J,s}_{\text{ref}}, \mathcal{H}^s_{\text{ref}}$, we first apply the Fast Fourier Transform (FFT). FFT is an efficient algorithm for computing the discrete Fourier transform (DFT) [2], reducing computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \log N)$. The 2D Fourier transform of a sub-band feature $\mathbf{X}$ is defined as:

$$\mathcal{F}(\mathbf{X})(u, v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \mathbf{X}(x, y) \exp\left(-j2\pi\left(\frac{ux}{H} + \frac{vy}{W}\right)\right), \tag{15}$$

where (u, v) denote frequency coordinates.

As localized spatial structures are explicitly modeled through wavelet decomposition, the frequency-domain component is accordingly designed to emphasize

global structural information and spectral energy distribution. Considering that the amplitude spectrum of Fourier representations effectively encodes global structural information and energy distribution [14, 25], we extract the amplitude spectrum from the Fourier-transformed features. Given a feature map $\mathbf{F}$, its amplitude spectrum is computed as:

$$|\mathcal{F}(\mathbf{F})| = \sqrt{\mathrm{Re}(\mathcal{F}(\mathbf{F}))^2 + \mathrm{Im}(\mathcal{F}(\mathbf{F}))^2}, \tag{16}$$

where $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote the real and imaginary components, respectively.

As different wavelet sub-bands focus on distinct frequency ranges, we employ Gaussian filters to selectively emphasize the dominant frequency components of each sub-band. Compared with ideal truncation-based filters, Gaussian filters provide smooth transitions in the frequency domain, thereby mitigating boundary artifacts and ringing effects caused by abrupt spectral truncation. The Gaussian low-pass and high-pass filters are defined as:

$$G_{\mathrm{low}}(u,v) = \exp\left(-\frac{u^2+v^2}{2\sigma^2}\right), \qquad G_{\mathrm{high}}(u,v) = 1 - G_{\mathrm{low}}(u,v). \tag{17}$$

where σ denotes a bandwidth parameter that flexibly controls the emphasis on low-frequency or high-frequency components.

The specific Gaussian filtering masks are illustrated in Fig. 4b. For the low-frequency sub-band, the dominant information is primarily concentrated in the low-frequency region (i.e., the central area of the spectrum after FFT and frequency shifting). Therefore, a Gaussian low-pass filtering mask is applied. In contrast, for the high-frequency sub-band set, the dominant components are mainly distributed in the high-frequency regions (i.e., the peripheral area of the shifted spectrum). Accordingly, a Gaussian high-pass filtering mask is used.

Given a student sub-band $\mathbf{X}^s$ and its corresponding teacher sub-band $\mathbf{X}^t$ (either low-frequency or high-frequency), we first compute their Fourier amplitude spectra and apply Gaussian filtering, followed by an InfoNCE loss to align the filtered spectral representations. The unified formulation of the Gaussian-Filtered Frequency Loss (GFFL) is given by:

$$\mathcal{L}_{\mathrm{GFFL}}\left(\mathbf{X}^s, \mathbf{X}^t; G\right) = \mathcal{L}_{\mathrm{InfoNCE}}\left(G \odot |\mathcal{F}(\mathbf{X}^s)|, G \odot \left|\mathcal{F}(\mathbf{X}^t)\right|\right), \tag{18}$$

Inspired by FCFD [20], to enhance task-oriented alignment, we further reconstruct the refined student wavelet sub-bands via IMDWT to obtain a recovered student representation $\mathbf{F}^s_{\mathrm{rec}}$. The $\mathbf{F}^s_{\mathrm{rec}}$ is then aligned with the original teacher representation $\mathbf{F}^t$ using a GFFL without frequency filtering. In addition, the $\mathbf{F}^s_{\mathrm{rec}}$ is fed into a newly initialized classifier to produce logits $\mathbf{Z}^s_{\mathrm{rec}}$, which are supervised by the teacher logits through a Kullback–Leibler (KL) divergence loss. The overall training objective is formulated as:

$$\begin{aligned}\mathcal{L}_{\mathrm{total}} = \lambda_1 \mathcal{L}_{\mathrm{GFFL}}\left(\mathbf{L}^{J,s}_{\mathrm{ref}}, \mathbf{L}^{J,t}; G_{\mathrm{low}}\right) + \lambda_2 \mathcal{L}_{\mathrm{GFFL}}\left(\mathcal{H}^s_{\mathrm{ref}}, \mathcal{H}^t; G_{\mathrm{high}}\right) \\ + \lambda_3 \mathcal{L}_{\mathrm{GFFL}}\left(\mathbf{F}^s_{\mathrm{rec}}, \mathbf{F}^t\right) + \lambda_4 \mathcal{L}_{\mathrm{KL}}\left(\mathbf{Z}^s_{\mathrm{rec}}, \mathbf{Z}^t\right).\end{aligned} \tag{19}$$

Table 1: Comparison of Heterogeneous Distillation Results on CIFAR-100. The best and second-best results are emphasized in **bold** and underlined cases.

Teacher	Student	From Scratch		Logits-based				Feature-based					
		T.	S.	KD [10]	DKD [44]	DIST [12]	OFA [7]	FitNet [30]	CC [28]	RKD [27]	CRD [34]	FBT [16]	**SFKD**
Swin-T	ResNet18	89.26	74.01	78.74	80.26	77.75	80.54	78.87	74.19	74.11	77.63	81.61	**83.26**
ViT-S	ResNet18	92.04	74.01	77.26	78.10	76.49	80.15	77.71	74.26	73.72	76.60	81.93	**82.10**
Mixer-B/16	ResNet18	87.29	74.01	77.79	78.67	76.36	79.39	77.15	74.26	73.75	76.42	81.90	**81.98**
Swin-T	MobileNetV2	89.26	73.68	74.68	71.07	72.89	80.98	74.28	71.19	69.00	79.80	81.28	**82.03**
ViT-S	MobileNetV2	92.04	73.68	72.77	69.80	72.54	78.45	73.54	70.67	68.46	78.14	**82.10**	79.97
Mixer-B/16	MobileNetV2	87.29	73.68	73.33	70.20	73.26	78.78	73.78	70.73	68.95	78.15	80.83	**81.17**
ConvNeXt-T	DeiT-T	88.41	68.00	72.99	74.60	73.55	75.76	60.78	68.01	69.79	65.94	79.57	**83.91**
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	74.40	**81.14**
ConvNeXt-T	Swin-P	88.41	72.63	76.44	76.80	76.41	78.32	24.06	72.63	71.73	67.09	80.73	**82.16**
Mixer-B/16	Swin-P	87.29	72.63	75.93	76.39	75.85	76.65	75.20	73.32	70.82	67.03	78.44	**81.60**
ConvNeXt-T	ResMLP-S12	88.41	66.56	72.25	73.22	71.93	75.21	45.47	67.70	65.82	63.35	78.03	**85.08**
Swin-T	ResMLP-S12	87.29	66.56	71.89	72.82	11.05	73.58	63.12	68.37	64.66	61.72	77.20	**84.01**
Average Improvements				↑3.12	↑3.16	↓2.31	↑6.19	↓5.21	↓0.33	↓1.39	↓0.02	↑8.38	**↑10.92**

4 Experiments

4.1 Implementation Details

Detailed descriptions of the **Models**, **Datasets**, **Baselines**, and **Training Protocols** used in our experiments are provided in Appendix (see Sec. B).

4.2 Main Results

We conduct extensive experiments on a wide range of teacher–student pairs. Compared with existing methods, our approach achieves state-of-the-art or competitive performance on both CIFAR-100 and ImageNet-1K. Across heterogeneous teacher–student pairs, SFKD improves the average Top-1 accuracy by 10.92% on CIFAR-100 and 2.51% on ImageNet-1K.

Results on CIFAR-100. To evaluate the effectiveness of SFKD on CIFAR-100, we conduct experiments on 12 heterogeneous teacher–student pairs. As shown in Tab. 1, for feature-based methods, except for FBT [16], which is specifically designed for heterogeneous distillation, most methods fail to achieve consistent improvements in heterogeneous models, and in some cases even degrade student performance. This observation aligns with our analysis in Sec. 1, where significant representational discrepancies between heterogeneous models make direct feature alignment challenging. In particular, when the student model is Transformer-based, methods that perform feature projection followed by pixel-wise MSE alignment (*e.g.*, FitNet [30]) tend to exhibit highly degraded performance. For example, in the ConvNeXt-T → Swin-P setting, FitNet achieves only 24.06% accuracy. For logit-based methods originally designed for homogeneous distillation, the performance gains in heterogeneous settings are generally limited. FBT [16] and OFA [7], which are specifically designed for heterogeneous distillation, mitigate discrepancies by compressing representations into a compact embedding space or mapping intermediate features to the logit space, respectively. In contrast, SFKD explicitly decouples and selectively emphasizes

Table 2: Comparison of Heterogeneous Distillation Results on ImageNet-1K. The best and second-best results are emphasized in **bold** and underlined cases.

Teacher	Student	From Scratch		Logits-based				Feature-based					
		T.	S.	KD [10]	DKD [44]	DIST [12]	OFA [7]	FitNet [30]	CC [28]	RKD [27]	CRD [34]	FBT [16]	**SFKD**
Deit-T	ResNet18	72.17	69.75	70.22	69.39	70.64	71.01	70.44	69.77	69.47	69.25	71.22	**71.24**
Swin-T	ResNet18	81.38	69.75	71.14	71.10	70.91	71.76	71.18	70.07	68.89	69.09	72.21	**72.54**
Mixer-B/16	ResNet18	76.62	69.75	70.89	69.89	70.66	71.38	70.78	70.05	69.46	68.40	71.44	**71.71**
Deit-T	MobileNetV2	72.17	68.87	70.87	70.14	71.08	71.39	70.95	70.69	69.72	69.60	71.78	**71.83**
Swin-T	MobileNetV2	81.38	68.87	72.05	71.71	71.76	72.32	71.75	70.69	67.52	69.58	72.54	**72.67**
Mixer-B/16	MobileNetV2	76.62	68.87	71.92	70.93	71.74	72.12	71.59	70.79	69.86	68.89	72.31	**72.43**
ResNet50	Deit-T	80.38	72.17	75.10	75.60	75.13	75.73	75.84	72.56	72.06	68.53	75.64	**75.89**
ConvNeXt-T	Deit-T	82.05	72.17	74.00	73.95	74.07	74.41	70.45	73.12	71.47	69.18	75.26	**75.29**
Mixer-B/16	Deit-T	76.62	72.17	74.16	72.82	74.22	74.46	74.38	72.82	72.24	68.23	**75.00**	74.69
ResNet50	Swin-N	80.38	75.53	77.58	76.24	77.29	77.76	76.83	76.05	75.90	73.90	77.79	**78.28**
ConvNeXt-T	Swin-N	82.05	75.53	77.15	77.00	77.25	77.50	74.81	75.79	75.48	74.15	77.73	**77.83**
Mixer-B/16	Swin-N	76.62	75.53	76.26	75.03	76.54	76.63	76.17	75.81	75.52	73.38	**76.87**	76.67
ConvNeXt-T	ResMLP-S12	82.05	76.65	76.87	77.23	77.24	77.26	74.69	75.79	75.28	73.57	77.33	**78.35**
Swin-T	ResMLP-S12	81.38	76.65	76.67	76.99	77.25	77.31	76.48	76.15	75.10	73.40	77.42	**77.88**
Average Improvements				↑1.61	↑1.05	↑1.65	↑2.05	↑1.00	↑0.56	↓0.30	↓1.65	↑2.31	**↑2.51**

spatial information through spatial–frequency joint-aware modeling. Compared with the strongest baseline FBT, our method achieves an additional average improvement of 2.54% on CIFAR-100.

Results on ImageNet-1K. To further validate the generalization capability of SFKD, we conduct experiments on the larger-scale ImageNet-1K dataset using 14 heterogeneous teacher–student pairs, as shown in Tab. 2. Compared with CIFAR-100, ImageNet-1K provides substantially more training data. As discussed in the ViT [3], Transformers typically require larger-scale training to compensate for weaker convolutional inductive biases, and thus benefit more from increased data volume. Consequently, feature-based distillation methods tend to exhibit more stable improvements on ImageNet-1K than on CIFAR-100, especially when the student is Transformer-based. For example, in the ResNet-50 → DeiT-T setting, FitNet [30] improves the DeiT-T accuracy to 75.84% Nevertheless, their performance gains remain limited compared with approaches specifically designed for heterogeneous distillation, such as FBT [16] and OFA [7]. Moreover, their performance can be unstable and may even degrade the student model in certain cases. For instance, in the ConvNeXt-T → DeiT-T pair, FitNet results in lower accuracy than the baseline without distillation. On ImageNet-1K, SFKD consistently achieves the best or competitive performance across heterogeneous settings, further demonstrating the effectiveness and generality of explicitly decoupling and selectively exploiting spatial information in heterogeneous representations. Moreover, compared with OFA, which utilizes features from four stages for distillation, our method relies solely on the final representation, similar to FBT, while still achieving comparable or superior performance.

Results on Homogeneous Distillation. To further verify the generality and effectiveness of decoupling and selectively exploiting spatial information in representations through spatial–frequency joint awareness in SFKD, we also conduct experiments under homogeneous distillation settings, as shown in Tab. 3. We evaluate two commonly used homogeneous teacher–student pairs: ResNet34 → ResNet18 and ResNet50 → MobileNet. As shown in the Tab. 3,

Table 3: Comparison of Homogeneous Distillation Results on ImageNet-1K. The best and second-best results are emphasized in **bold** and underlined cases.

	T.	S.	AT [43]	OFD [9]	CRD [34]	Review [1]	DKD [44]	DIST [12]	FCFD [20]	OFA [7]	FBT [16]	SFKD
ResNet34 ResNet18	73.31	69.75	70.69	70.81	71.17	71.61	71.70	72.07	72.24	72.10	72.29	**72.34**
ResNet50 MobileNet	76.61	68.58	69.56	71.25	71.37	72.56	72.05	73.24	73.37	73.28	73.45	**73.81**

Table 4: Ablation studies on loss components and DS^{2}SR module. The best and second-best results are highlighted in **bold** and underlined, respectively.

(a) Ablation study on the effectiveness of each loss component.

$\mathcal{L}_1^{LF}$	$\mathcal{L}_2^{HF}$	$\mathcal{L}_3^{Rec}$	$\mathcal{L}_4^{KL}$	Acc
×	✓	✓	✓	72.19
✓	×	✓	✓	72.32
✓	✓	×	✓	72.31
✓	✓	✓	×	72.46
✓	✓	✓	✓	**72.54**

(b) Ablation study on the necessity of wavelet sub-band refinement and the effectiveness of the DS^{2}SR module.

LCR	GTR	Swin-T ResNet18	ResNet50 Swin-N
×	×	71.24	76.83
✓	×	72.19	78.24
×	✓	72.46	77.83
✓	✓	**72.54**	**78.28**

SFKD achieves the best performance in both settings, outperforming existing homogeneous and heterogeneous distillation approaches. These results indicate that even when the representational discrepancy between teacher and student is relatively small, the proposed strategy of decoupling and selectively exploiting spatial information in representations remains effective for improving distillation performance.

4.3 Ablation Study

Necessity of Wavelet Sub-band Refinement and Effectiveness of the DS^{2}SR Module. As discussed in Sec. 3.3, wavelet decomposition cannot achieve ideal band-limited separation [23, 33], which may lead to spectral leakage or residual frequency overlap between sub-bands [38, 39]. To address this issue, and considering the distinct inductive biases across heterogeneous models, we design a novel and universal DS^{2}SR module to further refine the wavelet sub-bands. To evaluate the effectiveness of each refinement stream, we conduct experiments on ImageNet-1K using two heterogeneous teacher–student pairs: Swin-T → ResNet18 and ResNet50 → Swin-N. The results are summarized in Tab. 4b. When no refinement stream is introduced, the distillation performance remains limited for both CNN-based and Transformer-based students. This observation, as discussed in Sec. 3.1, further confirms that the weaker representation capability of student models leads to more unstable representations that require further refinement. After introducing either refinement stream, the performance of both student models improves noticeably. Further analysis reveals that the CNN student (ResNet18) is more sensitive to the GTR stream, achieving an additional improvement of 0.27% compared with using only the LCR stream. In contrast, the Transformer student (Swin-N) benefits more from the LCR stream, yielding an additional improvement of 0.41% compared with using only the GTR stream.

We believe this observation is consistent with the analysis in FBT [16], indicating that introducing streams with complementary inductive biases can better facilitate the student model in capturing knowledge from heterogeneous models. Finally, by adaptively combining the two streams, the DS^{2}SR module achieves the best performance of both student models.

Effectiveness of Each Loss Components. As shown in Eq. (19), the proposed objective consists of four loss terms, denoted as $\mathcal{L}_1^{LF}$, $\mathcal{L}_2^{HF}$, $\mathcal{L}_3^{Rec}$, and $\mathcal{L}_4^{KL}$, respectively. To investigate the contribution of each loss component, we conduct an ablation study on the Swin-T→ResNet18 teacher-student pair over ImageNet-1K. The results are summarized in Tab. 4a. The results indicate that removing $\mathcal{L}_1^{LF}$ causes the largest performance degradation, demonstrating that the low-frequency sub-band, which mainly preserves the global structural information of representations, plays critical role in knowledge transfer. Removing $\mathcal{L}_2^{HF}$ also results in a noticeable performance drop, indicating that the high-frequency sub-band set provides complementary local structural details that are likewise essential for effective distillation. Both $\mathcal{L}_1^{LF}$ and $\mathcal{L}_2^{HF}$ perform alignment at the wavelet sub-band level. Specifically, Gaussian low- and high-pass filters are introduced to selectively emphasize the dominant frequency components of different sub-bands, while the objective in Eq. (18) is essentially an InfoNCE-style isotropic contrastive objective. However, as demonstrated in FCFD [20], neural networks utilize intermediate features in an anisotropic manner. Consequently, performing only sub-band level alignment with an isotropic objective may still introduce structural discrepancies between $\mathbf{F}^s_{\text{rec}}$ and $\mathbf{F}^t$. To alleviate this issue, $\mathcal{L}_3^{Rec}$ is introduced to further constrain the reconstructed feature $\mathbf{F}^s_{\text{rec}}$ after IMDWT, encouraging overall structural consistency with $\mathbf{F}^t$. In addition, $\mathcal{L}_4^{KL}$ provides task-oriented supervision to ensure that the transferred knowledge remains beneficial for the downstream classification objective. As shown in Table 4a, removing either $\mathcal{L}_3^{Rec}$ or $\mathcal{L}_4^{KL}$ consistently degrades the student performance, further validating the effectiveness of all four loss components.

Extended Experiments. We provide additional experiments in Appendix (see Sec. C) to further evaluate SFKD in a more comprehensive manner.

5 Conclusion and Future Work

We discuss **Conclusion and Future Work** and defer a concentrated account to Appendix (see Sec. D).

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62271153.

References

1. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5008–5017 (2021) 2, 14, 1, 4

2. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex fourier series. Mathematics of computation **19**(90), 297–301 (1965) 10
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 1, 3, 8, 9, 13
4. Du, P., Li, H., Xu, H., Jeon, P.B., Lee, D., Ji, D., Yang, R., Zhu, F.: Diffusion transformer meets multi-level wavelet spectrum for single image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19700–19710 (2025) 7, 2
5. Finder, S.E., Amoyal, R., Treister, E., Freifeld, O.: Wavelet convolutions for large receptive fields. In: European conference on computer vision. pp. 363–380. Springer (2024) 7, 2
6. Haar, A.: Zur theorie der orthogonalen funktionensysteme. Georg-August-Universitat, Gottingen. (1909) 6
7. Hao, Z., Guo, J., Han, K., Tang, Y., Hu, H., Wang, Y., Xu, C.: One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. Advances in Neural Information Processing Systems **36**, 79570–79582 (2023) 3, 5, 12, 13, 14, 1, 4, 7
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1, 3
9. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1921–1930 (2019) 2, 14, 1, 4
10. Hinton, G.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 2, 12, 13, 1, 4
11. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017) 3
12. Huang, T., You, S., Wang, F., Qian, C., Xu, C.: Knowledge distillation from a stronger teacher. Advances in Neural Information Processing Systems **35**, 33716–33727 (2022) 2, 12, 13, 14, 1, 4
13. Huang, Y., Huang, J., Liu, J., Yan, M., Dong, Y., Lv, J., Chen, C., Chen, S.: Wavedm: Wavelet-based diffusion models for image restoration. IEEE Transactions on Multimedia **26**, 7058–7073 (2024) 7, 9
14. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13919–13929 (2021) 11, 3
15. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 3
16. Li, G., Wang, Q., Yan, K., Ding, S., Gao, Y., Xia, G.S.: Fuse before transfer: Knowledge fusion for heterogeneous distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3445–3454 (2025) 3, 5, 12, 13, 14, 15, 1, 2, 4
17. Li, J., Hassani, A., Walton, S., Shi, H.: Convmlp: Hierarchical convolutional mlps for vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6307–6316 (2023) 2, 1
18. Li, J., Cheng, B., Chen, Y., Gao, G., Shi, J., Zeng, T.: Ewt: Efficient wavelet-transformer for single image denoising. Neural Networks **177**, 106378 (2024) 7, 2

19. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unifying convolution and self-attention for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12581–12600 (2023) 2, 1
20. Liu, D., Kan, M., Shan, S., Chen, X.: Function-consistent feature distillation. arXiv preprint arXiv:2304.11832 (2023) 2, 11, 14, 15, 1, 4
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 1, 3
22. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022) 1, 3
23. Mallat, S.: A wavelet tour of signal processing. Elsevier (1999) 7, 14
24. Miao, Y., Deng, J., Han, J.: Waveface: Authentic face restoration with efficient frequency recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6583–6592 (2024) 7, 9
25. Oppenheim, A.V., Lim, J.S.: The importance of phase in signals. Proceedings of the IEEE **69**(5), 529–541 (2005) 11, 3, 6
26. Park, N., Kim, S.: How do vision transformers work? arXiv preprint arXiv:2202.06709 (2022) 3, 7
27. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019) 2, 12, 13, 1, 4
28. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5007–5016 (2019) 2, 12, 13, 1, 4
29. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? Advances in neural information processing systems **34**, 12116–12128 (2021) 3, 7
30. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014) 2, 12, 13, 1, 4, 7
31. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015) 3
32. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018) 1, 3
33. Strang, G., Nguyen, T.: Wavelets and filter banks. SIAM (1996) 7, 14
34. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019) 2, 12, 13, 14, 1, 4
35. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in neural information processing systems **34**, 24261–24272 (2021) 1, 3
36. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., et al.: Resmlp: Feedforward networks for image classification with data-efficient training. IEEE transactions on pattern analysis and machine intelligence **45**(4), 5314–5321 (2022) 1, 3
37. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021) 1, 3

38. Unser, M.A.: Ten good reasons for using spline wavelets. In: Wavelet applications in signal and image processing V. vol. 3169, pp. 422–431. SPIE (1997) 7, 9, 14, 6
39. Vetterli, M., Kovacevic, J.: Wavelets and subband coding, vol. 87. Prentice Hall PTR Englewood Cliffs, NJ (1995) 7, 9, 14, 6
40. Wang, Q., Li, Z., Zhang, S., Chi, N., Dai, Q.: Wavefusion: A novel wavelet vision transformer with saliency-guided enhancement for multimodal image fusion. IEEE Transactions on Circuits and Systems for Video Technology (2025) 7, 9
41. Wu, H., Xiao, L., Zhang, X., Miao, Y.: Aligning in a compact space: Contrastive knowledge distillation between heterogeneous architectures. arXiv preprint arXiv:2405.18524 (2024) 3, 5, 1
42. Yao, T., Pan, Y., Li, Y., Ngo, C.W., Mei, T.: Wave-vit: Unifying wavelet and transformers for visual representation learning. In: European conference on computer vision. pp. 328–345. Springer (2022) 7, 2
43. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016) 14, 1, 4
44. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11953–11962 (2022) 2, 12, 13, 14, 1, 4