# Distillation Dynamics: Towards Understanding Feature-Based Distillation in Vision Transformers

**Huiyuan Tian**[1], **Bonan Xu**[2], **Shijian Li**[1*]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2]Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University
tianhuiyuan@zju.edu.cn, bonan.xu@polyu.edu.hk, shijianli@zju.edu.cn

## Abstract

While feature-based knowledge distillation has proven highly effective for compressing CNNs, these techniques unexpectedly fail when applied to Vision Transformers (ViTs), often performing worse than simple logit-based distillation. We provide the first comprehensive analysis of this phenomenon through a novel analytical framework termed as "distillation dynamics", combining frequency spectrum analysis, information entropy metrics, and activation magnitude tracking. Our investigation reveals that ViTs exhibit a distinctive U-shaped information processing pattern: initial compression followed by expansion. We identify the root cause of negative transfer in feature distillation: a fundamental representational paradigm mismatch between teacher and student models. Through frequency-domain analysis, we show that teacher models employ distributed, high-dimensional encoding strategies in later layers that smaller student models cannot replicate due to limited channel capacity. This mismatch causes late-layer feature alignment to actively harm student performance. Our findings reveal that successful knowledge transfer in ViTs requires moving beyond naive feature mimicry to methods that respect these fundamental representational constraints, providing essential theoretical guidance for designing effective ViTs compression strategies.

**Code** —
https://github.com/thy960112/Distillation-Dynamics
**Extended version** — https://arxiv.org/abs/2511.06848

## 1 Introduction

ViTs (Dosovitskiy et al. 2021; Liu et al. 2021; Han et al. 2022) have revolutionized computer vision, achieving state-of-the-art performance across diverse tasks (Caron et al. 2021; Li et al. 2024b; Bar-Shalom, Bevilacqua, and Maron 2024) by leveraging self-attention mechanisms (Vaswani et al. 2017) to model long-range dependencies. However, their remarkable capabilities come with significant drawbacks: prohibitive computational complexity and excessive data requirements, which severely limit their deployment on resource-constrained devices and in data-scarce scenarios.

Knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015; Park et al. 2019; Tian, Krishnan, and Isola 2020; Zhao et al. 2022; Yang et al. 2023; Li et al. 2024a; Son et al. 2025) offers a promising solution to these challenges by transferring knowledge from large, complex teacher models to smaller, more efficient student models (Choudhary et al. 2020). Originally proposed by Hinton et al. (2015) for compressing neural networks through logit-based distillation, this technique has evolved significantly. During the CNN era, researchers extended distillation beyond output logits to include intermediate feature representations (Romero et al. 2015; Zagoruyko and Komodakis 2017; Heo et al. 2019; Chen et al. 2021; Ji, Heo, and Park 2021; Pham et al. 2024), capturing richer information from teacher models. Through simple convolutions or projection layers, feature-based distillation enabled substantial performance improvements in student CNNs.

Given this success, researchers naturally attempted to apply feature-based distillation methods to Vision Transformers. Surprisingly, however, a striking phenomenon emerged: the simple feature distillation techniques that worked well for CNNs actually degrade performance when applied to ViTs, often performing worse than logit-based distillation (Touvron et al. 2021a; Miles, Elezi, and Deng 2024; Sun et al. 2024b). While recent work such as ViTKD (Yang et al. 2024) has proposed ViT-specific distillation methods and observed similar failures, the fundamental reasons behind this degradation remain poorly understood. This gap in understanding severely limits our ability to design effective distillation strategies for the ViT architectures.

The core challenge lies in the fundamental architectural and representational differences between CNNs and ViTs. While sophisticated feature distillation methods (Hao et al. 2022; Yang et al. 2024; Miles and Mikolajczyk 2024; Fan et al. 2024; Feng et al. 2025) have shown improvements, they often rely on empirical observations rather than principled theoretical guidance (Cheng et al. 2024). What is urgently needed is a deep, mechanistic understanding of how information flows through ViT architectures and why conventional feature distillation fails—insights that can guide the development of more effective distillation methods.

In this work, we conduct a comprehensive theoretical investigation of the ViTs distillation process by introducing novel analytical tools to dissect the information dynamics within these models. We employ frequency spectrum analysis, Shannon entropy analysis through the lens of the In-

*Corresponding author

formation Bottleneck principle (Tishby, Pereira, and Bialek 2000; Tishby and Zaslavsky 2015; Hong et al. 2025), and activation magnitude analysis to reveal the unique information processing patterns of ViTs. Our analysis uncovers fundamental insights about how ViTs process information differently from CNNs, explaining the systematic failure of naive feature distillation approaches.

Our main contributions are as follows:

1. **Novel analytical framework for understanding ViTs representations.** We introduce and systematically apply frequency spectrum analysis along the channel dimension, Shannon entropy analysis, and activation magnitude tracking to characterize the internal dynamics of ViTs.
2. **Discovery of the U-shaped information processing signature.** We reveal that ViT models exhibit a distinctive U-shaped pattern in information entropy across layers, indicating a two-phase processing strategy: initial compression followed by task-specific expansion. This pattern emerges during training and represents a learned, rather than architectural, behavior.
3. **Identification of channel-dimension encoding saturation.** We discover that ViTs fully utilize their channel dimensions for distributed encoding in later layers, revealing that performance degradation in student networks stems not from insufficient parameters, but from fundamental representational capacity constraints that prevent smaller models from replicating the teacher's encoding strategy.
4. **Mechanistic explanation of negative transfer.** Through analysis of distillation evolution, we demonstrate that late-layer feature distillation induces negative transfer by forcing students to adopt representational paradigms incompatible with their limited capacity. We show that information processing patterns evolve from monotonic to U-shaped during training, and that misaligned distillation disrupts this natural progression.

These insights provide the theoretical foundation necessary for designing principled feature distillation methods for Vision Transformers, moving beyond empirical trial-and-error toward theory-guided approaches that respect the fundamental information processing characteristics of these architectures.

# 2 Analytical Methods

To dissect the internal mechanisms of knowledge distillation in ViTs, we devise a multi-faceted analytical framework termed as "distillation dynamics". This framework is designed to characterize how a student model learns to mimic the internal representations of a teacher model by examining three complementary aspects of their activation patterns. First, we propose frequency spectrum analysis to reveal the composition of feature encodings, distinguishing between global, coarse-grained structures and fine-grained, local details. Second, we apply Shannon entropy (Ash 2012) analysis to quantify the information complexity and structural organization within feature maps. Finally, we measure activation magnitudes to further prove our findings by tracking the strength of signal propagation through the network layers.

These methods allow us to triangulate the true nature of ViTs representations: when we observe a U-shaped entropy pattern indicating compression followed by expansion, we can verify this through corresponding changes in frequency spectra (from uniform to low-pass to uniform again) and activation magnitudes (decreasing then increasing). This multi-faceted validation is crucial for establishing that the observed patterns represent fundamental computational strategies rather than artifacts of any single measurement approach.

## 2.1 Frequency Spectrum Analysis

To understand how the network encodes features, we analyze the frequency content of the feature representations at each layer. Such analysis is particularly relevant for ViTs, which have been characterized as behaving like low-pass filters (Park and Kim 2022), a property that can lead to the loss of high-frequency information in deeper layers.

Intermediate activations are extracted from ViT model during forward inference on validation images. Let the activations be denoted as a tensor tensor $\mathbf{A} \in \mathbb{R}^{L\times B\times C\times H\times W}$, where $L$ is the number of layers, $B$ is the batch size, $C$ is the number of channels, and $H \times W$ is the spatial resolution.

To reveal the complexity of feature interactions within the channel dimension, we apply a one-dimensional Fast Fourier Transform (FFT) along the channel axis for each spatial position. This unconventional choice, distinct from a spatial FFT, allows us to assess the structure of the feature space itself. A low-frequency-dominant spectrum suggests that channels are highly correlated, representing smooth variations of a primary feature, whereas a high-frequency spectrum indicates a more complex, decorrelated set of feature detectors. The transform is defined as:

$$\mathbf{F}_{l,b,h,w}[k] = \frac{1}{C}\sum_{c=0}^{C-1} \mathbf{A}_{l,b,c,h,w}\, e^{-j2\pi kc/C}, \qquad (1)$$

where $\mathbf{F} \in \mathbb{C}^{L\times B\times C\times H\times W}$ is the resulting frequency-domain representation, and $j = \sqrt{-1}$ is the imaginary unit. The magnitude spectrum is computed as $|\mathbf{F}|$ and averaged over the batch ($B$), height ($H$), and width ($W$) dimensions to obtain the representative per-layer frequency spectrum:

$$\mathbf{S}_l[k] = \frac{1}{BHW}\sum_{b=1}^{B}\sum_{h=1}^{H}\sum_{w=1}^{W} |\mathbf{F}_{l,b,k,h,w}|. \qquad (2)$$

This yields a per-layer spectral signature $\mathbf{S} \in \mathbb{R}^{L\times C}$, which reveals how feature abstraction and complexity develop throughout the network by showing whether representations are dominated by low- or high-frequency components in the channel dimension.

## 2.2 Information Entropy Analysis

While frequency analysis reveals the composition of learned features, we employ information entropy to quantify their structural complexity. This analysis is explicitly framed

through the lens of the Information Bottleneck (IB) principle (Tishby, Pereira, and Bialek 2000; Tishby and Zaslavsky 2015; Hong et al. 2025). In this context, low entropy signifies a compressed, highly structured representation (a tight bottleneck), whereas high entropy indicates a more uniform or expansive distribution.

The analysis is performed on the activation tensor $\mathbf{A}$. At each spatial position $(h, w)$ for a given layer $l$ and batch item $b$, the vector of channel activations $\mathbf{v}_{l,b,h,w} \in \mathbb{R}^C$ is discretized into $N_b = 100$ bins over the global activation range $[\min(\mathbf{A}), \max(\mathbf{A})]$. This is a standard method for estimating the probability mass function of continuous activations.

From this discretization, we estimate a probability mass function $p_n$ for each bin $n$:

$$p_n = \frac{h_n}{C}, \quad n = 1, \ldots, N_b, \tag{3}$$

where $h_n$ is the count of activations falling into bin $n$. The Shannon entropy is then computed by summing over bins with non-zero probabilities:

$$E_{l,b,h,w} = -\sum_{n:p_n>0} p_n \log_2 p_n. \tag{4}$$

This procedure yields an entropy map $\mathbf{E} \in \mathbb{R}^{L\times B\times H\times W}$. We then compute the average entropy for each layer $\bar{E}_l$ by averaging across the batch and spatial dimensions:

$$\bar{E}_l = \frac{1}{BHW} \sum_{b=1}^{B} \sum_{h=1}^{H} \sum_{w=1}^{W} E_{l,b,h,w}. \tag{5}$$

By comparing the layer-wise entropy profiles $\{\bar{E}_l\}_{l=1}^{L}$ of the student and teacher models, we can quantitatively assess whether the student successfully emulates the teacher's information compression and expansion strategy. This provides a powerful tool for understanding how well the core representational dynamics are transferred.

### 2.3 Activation Magnitude Analysis

Finally, to complement our analysis of feature content (frequency) and structure (entropy), we examine the signal propagation strength throughout the network. This is accomplished by measuring the mean activation magnitude at each layer, which serves as a proxy for how the network amplifies or attenuates information as it flows through successive layers.

Using the same activation tensor $\mathbf{A}$, the mean activation magnitude for each layer $M_l$ is defined as the mean absolute value of its activation:

$$M_l = \frac{1}{BCHW} \sum_{b=1}^{B} \sum_{n=1}^{C} \sum_{d=1}^{H} \sum_{d=1}^{W} |\mathbf{A}_{l,b,c,h,w}|, \tag{6}$$

where the analysis spans all $L$ layers. Plotting these per-layer magnitudes $M_l$ against the layer index reveals the macroscopic information flow within the network.

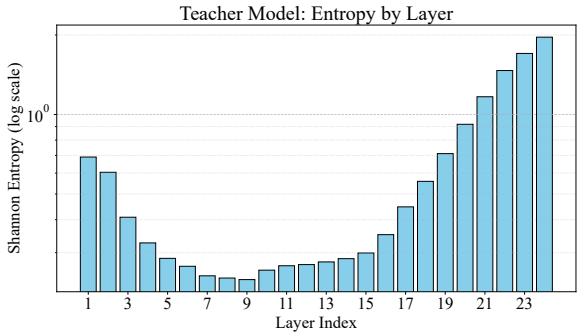



Figure 1: Layer-wise Shannon entropy of the CaiT-S24 teacher model exhibits a characteristic U-shaped profile. Entropy decreases from layers 1-9 (compression phase), then increases through layer 24 (expansion phase).

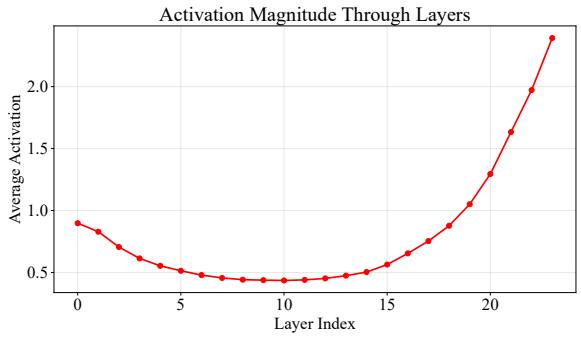



Figure 2: U-shaped profile of average activation magnitude through network layers.

### 2.4 Distillation Methods for Validation

To validate our analytical insights, we implement two feature-based distillation methods. First, SpectralKD (Tian et al. 2025) aligns frequency spectra by applying 2D FFT to spatial dimensions of feature maps. After channel alignment via adaptive pooling when $C_s \neq C_t$, we compute the frequency alignment loss:

$$\mathcal{L}_{\text{Freq}} = \text{MSE}(\mathcal{F}_{\text{stack}}(\mathbf{A}_s), \mathcal{F}_{\text{stack}}(\mathbf{A}_t)), \tag{7}$$

where $\mathcal{F}_{\text{stack}}$ denotes the 2D RFFT followed by stacking real and imaginary components.

Second, ProjectorKD inspired by FitNet (Romero et al. 2015) uses a learnable projector to match feature dimensions directly in spatial domain:

$$\mathcal{L}_{\text{Proj}} = \text{MSE}(\text{Projector}(\mathbf{A}_s), \mathbf{A}_t). \tag{8}$$

Both methods combine feature losses with standard KD loss: $\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{KD}} + \beta\mathcal{L}_{\text{Feature}}$, where $\beta$ weights the feature component.

## 3 Analyses of Representational Dynamics

This section initiates our empirical investigation by first establishing a detailed characterization of the teacher model's

internal information processing dynamics. By dissecting the layer-wise behavior of CaiT-S24 teacher (Touvron et al. 2021b), we define a precise information processing signature that a student model should ideally learn to replicate.

We analyze its intermediate activations generated from a forward pass on 32 randomly selected samples from the ImageNet (Deng et al. 2009) validation dataset. We employ the Shannon entropy and activation magnitude analyses detailed in Section 2 to probe the model's layer-wise dynamics. Our findings reveal a consistent and highly structured two-phase processing pattern, which we identify as a form of representational bottleneck. The overall structure reveals an active, two-stage computational strategy: the network first builds a compressed, abstract model of the input (compression phase) and then queries this model to construct an answer for the specific task (expansion phase).

### 3.1 The U-Shaped Entropy Profile

Our analysis begins with the layer-wise Shannon entropy defined in Section 2.2. As depicted in Figure 1, the CaiT-S24 teacher model exhibits a distinct U-shaped entropy profile across its 24 layers. The entropy initially decreases from the input layer, reaches a nadir around layers 8-9, and then steadily increases toward the final layer. Our observation of a U-shaped curve in ViTs points to a more complex, multi-stage information processing strategy that goes beyond simple, progressive feature stabilization.

This distinctive profile can be interpreted as direct empirical evidence for the IB principle operating within ViTs. The IB framework posits that an effective model learns by first compressing input data to retain only task-relevant information, and then using this compressed representation for prediction (Hong et al. 2025). The observed U-shaped entropy curve provides clear evidence of such a two-phase strategy:

1. **Compression Phase (Layers 1 to 9):** The initial, descending arm of the U-shape aligns with the compression phase of the IB. In these early layers, the model processes high-dimensional image patch embeddings. The decreasing entropy signifies that the activation distributions are becoming more concentrated, structured, and less random as the network filters out redundant visual details and noise. This process extracts and organizes essential, low-level features, which form the building blocks of the model's representation.
2. **Refinement and Expansion Phase (Layers 9 to 24):** The subsequent, ascending arm of the curve corresponds to a refinement or expansion phase. Having formed a compact, abstract representation at the entropy nadir, the model begins to expand the feature space to construct more complex, high-level semantic concepts necessary for the final classification task. This involves combining the compressed features in myriad ways, leading to a more uniform and higher-entropy distribution as the model prepares the representation for the classifier.

We further present the U-shape as a consistent operational signature of the tested ViT models, as similar structures are also observable in other architectures like vanilla ViT trained by supervised learning (Dosovitskiy et al. 2021) or self-supervised learning like MAE (He et al. 2022) (see extended version for more details). This U-shaped entropy curve marks the functional transition point where the model switches from general-purpose feature extraction to task-specific information aggregation.

### 3.2 The U-Shaped Signal Propagation

To further substantiate our finding of a two-phase model, we analyze the mean activation magnitude at each layer, as detailed in Section 2.3. The resulting curve, shown in Figure 2, strikingly mirrors the U-shaped entropy profile. The activation magnitude initially decreases, bottoms out in the middle layers, and then rises in the latter half of the network. This provides strong corroborating evidence for our findings.

This parallel U-shaped dynamic can be interpreted as follows:

1. **Phase 1 (Signal Attenuation):** The initial decrease in activation magnitude suggests that during the compression phase, the model actively attenuates signals corresponding to irrelevant or redundant visual information. This functions as a form of dynamic, input-dependent feature pruning, allowing the model to focus its computational resources on more salient regions or tokens.
2. **Phase 2 (Signal Amplification):** The subsequent increase in magnitude indicates that during the refinement phase, the model amplifies the signals of the most discriminative features. This targeted amplification ensures that the information most critical for the final decision is given the greatest weight.

The signal amplification in the later layers can be connected to the recently identified phenomenon of massive activations (Sun et al. 2024a; Owen et al. 2025) in Transformer models. These researches have shown that a small subset of neurons in both LLMs and ViTs can exhibit exceptionally large activation magnitudes. The emergence pattern of these massive activations which appears abruptly after the initial layers and persisting until the final few layers is highly consistent with the U-shaped magnitude profile we observe.

The similar U-shaped pattern in Shannon entropy and activation magnitude together constitute the teacher's unique information processing signature. This signature reveals a sophisticated, dynamic process: the model first navigates a representational bottleneck by compressing and attenuating irrelevant information, and then expands and amplifies task-specific information for the final prediction.

## 4 Results and Analyses

This section evaluates the efficacy of transferring these internal representations to a smaller DeiT-Tiny student model. Our experimental results reveal a series of counterintuitive phenomena, most notably the general failure of feature-based distillation to outperform a simple logits-only baseline. By dissecting these results through our proposed analytical framework, we find a fundamental representational mismatch between the teacher and student, particularly in the later layers. We further analyze the observed negative transfer in a frequency-domain, offering a new perspective on the challenges of knowledge distillation for ViTs.

| Method | Layer(s) | $\beta$ | Top-1 Acc (%) |
|---|---|---|---|
| SoftKD | N/A | N/A | 76.99 |
| SoftKD | N/A | N/A | 78.07 (500 epochs) |
| SpectralKD | F 2, L 6 | 0.2 | 77.07 |
| SpectralKD | F 2, L 1 | 0.2 | 77.06 |
| SpectralKD | F 1, L 1 | 0.2 | 77.08 |
| SpectralKD | F 1 | 0.2 | 77.00 |
| SpectralKD | L 1 | 0.2 | 76.83 |
| SpectralKD | L 1 | 0.2 | 77.59 (500 epochs) |
| SpectralKD | L 1 | 0.1 | 76.48 |
| SpectralKD | L 8 | 0.2 | 76.69 |
| ProjectorKD | F 1 | 0.2 | 76.86 |
| ProjectorKD | F 1 | 0.2 | 76.72 |
| ProjectorKD | F 1, L 1 | 0.2 | 76.80 |

Table 1: Top-1 Accuracy on ImageNet for a DeiT-Tiny student distilled from a CaiT-S24 teacher. The SoftKD baseline uses only logits-based distillation. All feature-based methods (SpectralKD, ProjectorKD) are combined with logits-based distillation. "F X" and "L Y" refer to aligning the first X and last Y layers of the student with corresponding layers from the teacher. $\beta$ is the weight of feature-based distillation loss. All models are trained with 300 epochs except those noted with 500 epochs.

### 4.1 Empirical Evaluation of Feature Distillation

To systematically investigate the transfer of intermediate representations, we conduct a series of distillation experiments using CaiT-S24 model as the teacher and a smaller DeiT-Tiny model as the student. We evaluate three primary distillation schemes: a baseline using only logit-based distillation (SoftKD), our proposed non-parametric frequency alignment method (SpectralKD), and a parametric projector-based feature distillation method (ProjectorKD). For the feature-based methods, we explore distilling knowledge from various layers or combinations of layers. The ImageNet Top-1 accuracy are presented in Table 1.

A puzzle observation in Table 1 is that, contrary to the prevailing assumption that intermediate features provide a richer and more informative supervisory signal, the majority of our feature distillation experiments fail to surpass the performance of the simple SoftKD baseline ($76.99\%$) with distillation temperature set to 1. The only configurations that yield a marginal improvement are those that incorporate knowledge from the teacher's early layers. The best model ($77.08\%$) uses SpectralKD to align the first and last layers of teacher and student models.

This outcome is consistent across both the non-parametric SpectralKD and the parametric ProjectorKD. Specifically, attempts to distill knowledge from the teacher's final layers prove to be actively detrimental. For instance, using ProjectorKD on the final layer results in a Top-1 accuracy of $76.72\%$, a drop of $0.27\%$ compared to the SoftKD baseline. Similarly, SpectralKD applied to the last layer yields $76.83\%$, and the last eight layers results in $76.69\%$. This consistent performance drop across different methods points toward a systemic issue with transferring knowledge from the teacher's late-stage representations. The phenomenon has also been observed by previous works like FitNet (Romero et al. 2015) and ViTKD (Yang et al. 2024).

These observations strengthen our hypothesis: since two methodologically distinct approaches point to the same conclusion, the problem likely lies not in the mechanism of transfer, but in the intrinsic nature of the knowledge being transferred from different stages of the teacher model.

### 4.2 Negative Transfer from Late Layers

The consistent performance degradation observed when distilling from the teacher's later layers reveals a phenomenon of negative transfer. Our experiments underscore the severity of this issue through a surprising result. Initially, we hypothesize that negative transfer might occur because the feature distillation loss overpowers the primary classification loss. To test this, we reduce the feature-distillation weight $\beta$ for SpectralKD on the last layer from $0.2$ to $0.1$. Conventional wisdom would predict that this weaker, less restrictive guidance should alleviate the negative effects. However, we observe the opposite: performance drops even further, from $76.83\%$ to $76.48\%$ (Table 1).

We also explore whether extend training may improve feature map distillation performance, reasoning that feature maps contain rich information requiring more time to learn effectively. After $500$ training epochs, simple SoftKD achieved $78.07\%$ accuracy while SpectralKD with last-layer distillation reached only $77.59\%$, widening the performance gap despite the longer training period.

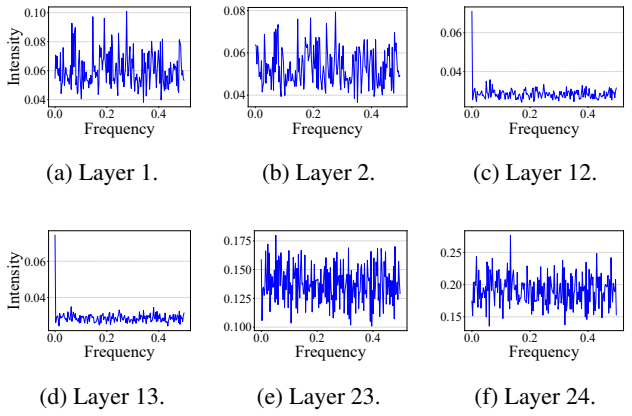
This counterintuitive outcome strongly suggests that the problem with late-layer distillation is not one of magnitude but of direction. The guidance from the teacher's final layers fundamentally misdirects the student's learning trajectory, providing a supervisory signal that conflicts with the student's own optimization path.

### 4.3 Representational Mismatch in Frequency Domain

To uncover the mechanistic origin of this pathological guidance, we employ the frequency spectrum analysis detailed in Section 2.1. By examining the spectral content of the teacher model's activations along the channel dimension at different layers, we can reveal the encoded patterns of the teacher model in the frequency domain. Figure 3 presents the frequency spectra for representative layers of the CaiT-S24 teacher model.

The spectral signatures reveal a distinct three-phase evolution that aligns perfectly with the U-shaped information processing profile identified in Section 3:

1. **Phase 1: Early Layers.** In the initial layers, the frequency spectra are relatively uniform and noisy, with no dominant pattern. This corresponds to the high-entropy, high-magnitude beginning of the U-shaped curves. At this stage, the model processes raw patch embeddings, and the representations remain generic, containing a broad mixture of low- and high-frequency information across channels.




(a) Layer 1. (b) Layer 2. (c) Layer 12.

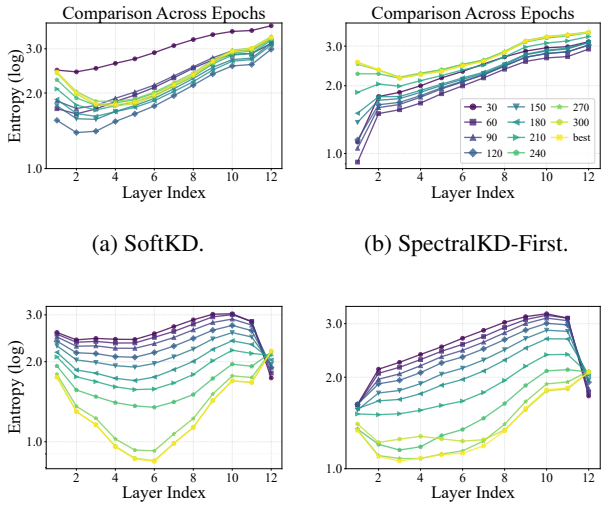
(d) Layer 13. (e) Layer 23. (f) Layer 24.

Figure 3: Frequency spectra of channel-wise feature representations across layers in CaiT-S24. Early layers (a-b) show uniform, noisy spectra; middle layers (c-d) exhibit low-pass filter characteristics corresponding to the representational bottleneck; late layers (e-f) return to uniform spectra with higher energy, indicating distributed high-dimensional encoding. This three-phase evolution aligns with the U-shaped information processing profile in ViTs.

2. **Phase 2: Middle Layers.** Around the middle of the network, the spectra exhibit a pronounced decay from low-frequency to high-frequency components, resembling a low-pass filter response. This spectral pattern coincides with the nadir of the U-shaped entropy curve, further confirming the representational bottleneck in ViTs. Here, the model has compressed the input by filtering out high-frequency noise while retaining structured, abstract representations of the most salient information.
3. **Phase 3: Late Layers.** In the final layers, the spectrum becomes more uniform again, but with higher overall energy compared to early layers. This corresponds to the "expansion phase" of the U-shaped entropy curve.

The spectral pattern in Phase 3 provides critical evidence for explaining negative transfer. The return to a uniform, high-energy spectrum does not represent a regression to the noisy state of early layers. Instead, it signifies that the teacher model is performing complex, high-dimensional feature expansion, distributing and entangling semantic information across the entire channel space in an intricate manner. This sophisticated computational strategy depends intrinsically on the teacher's high representational capacity and its massive channel dimension.

Interestingly, CNN architectures like ResNet (He et al. 2016) exhibit different spectral patterns. Their final stages maintain spectral encoding characteristics similar to Phase 2 (middle layers) rather than the distributed encoding of Phase 3. This suggests that CNNs underutilize their channel dimension capacity, which explains why smaller student models can successfully learn CNN teacher features through distillation. Further details on CNN spectral characteristics are provided in the extended version.




(a) SoftKD. (b) SpectralKD-First.

(c) SpectralKD-Last. (d) SpectralKD-Both.

Figure 4: Distillation Evolution.

This analysis reframes the problem of late-layer distillation in ViTs. The issue extends beyond a simple quantitative "capacity gap" in parameter count or layer width. It represents a fundamental representational paradigm mismatch. The teacher's late layers operate under a distributed, high-dimensional encoding paradigm, while the student model, with its severely limited channel dimension, is architecturally incapable of replicating this approach. Instead, the student is forced to operate within a more "compact, feature-centric" paradigm, where information must be encoded efficiently within its constrained channel space. Consequently, this knowledge becomes effectively non-transferable to smaller students through direct mimicry.

## 5 Distillation Evolution

To deepen our understanding of the representational mismatch identified in Section 4, we examine the temporal evolution of the student's layer-wise Shannon entropy during training. This analysis reveals how distillation strategies shape the development of the student's information processing signature over epochs. By comparing the entropy profiles at every 30 epochs (up to 300), we uncover how feature alignment alters the emergence of the U-shaped structure, explaining the observed negative transfer from late layers.

We focus on four configurations: logits-only distillation (SoftKD), frequency alignment on the first layer (SpectralKD-First), on the last layer (SpectralKD-Last), and on both the first and last layers (SpectralKD-Both). Entropy values were computed as described in Section 2.2 on DeiT-Tiny students distilled from the CaiT-S24 teacher.

### 5.1 Dynamics under Logits-Only Distillation

Under SoftKD (Figure 4a), the student's entropy profile in early training exhibits a increasing trend across layers. As

training advances, however, the profile transforms. By epoch 120, entropy in the middle layers begins to dip, and by the final best checkpoint, a pronounced U-shaped structure emerges, with a minimum around layers 4 to 5. This gradual development suggests that the U-shaped signature is not an architectural artifact of ViTs but a learned behavior.

The teacher's soft labels alone, suffice to guide the student toward discovering this two-phase process: initial compression followed by expansion.

### 5.2 Dynamics with Late-Layer Alignment

Incorporating frequency alignment on the last layer (SpectralKD-Last, Figure 4c) introduces distinct perturbations to the student's developmental trajectory. In early epochs (30 to 90), the overall entropy is elevated compared to SoftKD, but the last layer starts notably lower. This initial suppression suggests that mimicking the teacher's late-layer spectral structure imposes a compressive constraint, forcing the student to adopt a more structured (lower-entropy) representation prematurely in what should be its expansion phase.

As training progresses, the last-layer entropy gradually increases, while early-layer entropy decreases. This upward drift in the last layer implies a conflict: the feature alignment loss pulls the student's representation toward the teacher's distributed, high-dimensional encoding, but the student's limited capacity and the primary classification objective push for higher entropy to enable flexible feature recombination. This dynamic highlights negative transfer as a tug-of-war, where late-layer constraints hinder the student's natural progression toward the teacher's full signature.

### 5.3 Early-Layer Alignment and Global Effects

Aligning the first layer (Figure 4b) yields a different trajectory. Early training shows lower entropy in initial layers, rising steadily to layer 12, which is a monotonic increase without a clear dip. Over epochs, entropy in early layers rises substantially, but the profile remains largely increasing even in the best model, lacking the teacher's U-shaped bottleneck. This suggests that early-layer alignment enforces a more uniform, high-frequency-dominant structure from the outset (aligning with the teacher's noisy Phase 1 spectra), accelerating feature extraction but preventing the compression phase from fully developing. The absence of a mid-network minimum may indicate over-specialization to low-level details, yet performance remains comparable to SoftKD ($77.00\%$), implying that early guidance complements logits without disrupting later expansion.

Combining first- and last-layer alignment (Figure 4d) reveals compounded effects. The profile starts with moderate entropy and evolves to a shallow increase, with minimal U-shape. Strikingly, introducing late-layer alignment causes early-layer entropy to decrease over training. This global propagation, where last-layer constraints retroactively suppress early-layer entropy, demonstrates that feature distillation has global impacts. It contradicts the expected U-shape's high initial entropy, suggesting that dual constraints over-regularize both phases, flattening the bottleneck. Despite this, the marginal performance gain ($77.08\%$) indicates a delicate balance where complementary alignments may mitigate some representational mismatches.

### 5.4 Implications and Insights

These dynamics underscore that feature distillation is not a static process of copying knowledge but a dynamic process of guiding the student's developmental trajectory. It fundamentally reshapes representational development, acting as a form of curriculum learning where a mismatched curriculum can hinder performance.

A new observation is the "entropy rebound" in late layers under last-layer alignment: the student's attempt to increase entropy despite constraints points to an adaptive resilience, where the classification loss counters over-compression from the distillation loss. This could explain why reducing the distillation weight $\beta$ worsens performance (Section 4.2). A weaker alignment allows a partial rebound toward the student's preferred high-entropy state but ultimately disrupts the learned equilibrium, leading to a poorer solution.

Furthermore, the global effects of local alignment suggest the presence of emergent, highly coupled inter-layer dependencies in ViTs, where spectral constraints can propagate bidirectionally through the network during training.

## 6 Conclusion

This work provides a mechanistic explanation for the systemic failure of feature-based distillation in ViTs. Proposing a novel analytical framework termed as "distillation dynamics" that combines frequency, entropy, and activation analysis, we discover that ViTs learn a characteristic "U-shaped" information processing signature, defined by an initial compression phase followed by a task-specific expansion phase. The failure of feature-based distillation stems from a fundamental representational paradigm mismatch between teacher and student models. In their later layers, large teacher models employ a sophisticated, distributed encoding strategy that entangles information across their entire high-dimensional channel space. Smaller student models, with their limited channel capacity, are architecturally incapable of replicating this strategy.

Forcing a student to mimic this late-layer representation provides a conflicting and actively harmful supervisory signal. Our analysis of the distillation process over time shows that this misguided mimicry disrupts the student's natural learning trajectory, preventing it from developing its own effective U-shaped processing pattern. In contrast, aligning only the earlier, compressive-phase layers proves more compatible and less detrimental.

These findings demonstrate that effective ViTs compression requires moving beyond naive feature mimicry. Promising directions include phase-specific distillation, which applies feature alignment only to the compatible early-to-middle layers of the teacher, and transformation-based approaches that explicitly translate the teacher's complex representations into a format the student can digest. This work provides the theoretical foundation for designing principled and effective compression strategies tailored to the unique information dynamics of ViTs.

## Acknowledgments

This work was supported by the STI 2030 Major Projects (grant 2021ZD0200403) and by the Zhejiang Provincial Natural Science Foundation of China (grant LD24F030002).

## References

Ash, R. B. 2012. *Information theory*. Courier Corporation.

Bar-Shalom, G.; Bevilacqua, B.; and Maron, H. 2024. Subgraphormer: Unifying Subgraph GNNs and Graph Transformers via Graph Products. In *Forty-first International Conference on Machine Learning*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021. Distilling Knowledge via Knowledge Review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5008–5017.

Cheng, X.; Cheng, L.; Peng, Z.; Xu, Y.; Han, T.; and Zhang, Q. 2024. Layerwise Change of Knowledge in Neural Networks. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 8038–8059. PMLR.

Choudhary, T.; Mishra, V.; Goswami, A.; and Sarangapani, J. 2020. A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53: 5113–5155.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fan, J.; Li, C.; Liu, X.; and Yao, A. 2024. ScaleKD: Strong Vision Transformers Could Be Excellent Teachers. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 63290–63315. Curran Associates, Inc.

Feng, Q.; Li, W.; Lin, T.; and Chen, X. 2025. Align-KD: Distilling Cross-Modal Alignment Knowledge for Mobile Vision-Language Large Model Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4178–4188.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

Hao, Z.; Guo, J.; Jia, D.; Han, K.; Tang, Y.; Zhang, C.; Hu, H.; and Wang, Y. 2022. Learning Efficient Vision Transformers via Fine-Grained Manifold Distillation. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 9164–9175. Curran Associates, Inc.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979–15988.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; and Choi, J. Y. 2019. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1921–1930.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Hong, J.-H.; Kim, H.-J.; Jeon, K.-S.; and Lee, S.-W. 2025. Comprehensive Information Bottleneck for Unveiling Universal Attribution to Interpret Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 25166–25175.

Ji, M.; Heo, B.; and Park, S. 2021. Show, Attend and Distill: Knowledge Distillation via Attention-based Feature Matching. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9): 7945–7952.

Li, L.; Bao, Y.; Dong, P.; Yang, C.; Li, A.; Luo, W.; Liu, Q.; Xue, W.; and Guo, Y. 2024a. DetKDS: Knowledge Distillation Search for Object Detectors. In *Forty-first International Conference on Machine Learning*.

Li, X.; Ding, H.; Yuan, H.; Zhang, W.; Pang, J.; Cheng, G.; Chen, K.; Liu, Z.; and Loy, C. C. 2024b. Transformer-based visual segmentation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Miles, R.; Elezi, I.; and Deng, J. 2024. VkD: Improving Knowledge Distillation using Orthogonal Projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15720–15730.

Miles, R.; and Mikolajczyk, K. 2024. Understanding the Role of the Projector in Knowledge Distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5): 4233–4241.

Owen, L.; Chowdhury, N. R.; Kumar, A.; and Güra, F. 2025. A Refined Analysis of Massive Activations in LLMs. arXiv:2503.22329.

Park, N.; and Kim, S. 2022. How Do Vision Transformers Work? In *International Conference on Learning Representations*.

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

Pham, C.; Nguyen, V.-A.; Le, T.; Phung, D.; Carneiro, G.; and Do, T.-T. 2024. Frequency attention for knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2277–2286.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. arXiv:1412.6550.

Son, S.; Ryu, J.; Lee, N.; and Lee, J. 2025. The Role of Masking for Efficient Supervised Knowledge Distillation of Vision Transformers. In *European Conference on Computer Vision*, 379–396. Springer.

Sun, M.; Chen, X.; Kolter, J. Z.; and Liu, Z. 2024a. Massive Activations in Large Language Models. In *First Conference on Language Modeling*.

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

Tian, H.; Xu, B.; Li, S.; and Pan, G. 2025. SpectralKD: A Unified Framework for Interpreting and Distilling Vision Transformers via Spectral Analysis. arXiv:2412.19055.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. arXiv:physics/0004057.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, 1–5.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jegou, H. 2021a. Training data-efficient image transformers & distillation through attention. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 10347–10357. PMLR.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021b. Going Deeper With Image Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 32–42.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Yang, Z.; Li, Z.; Zeng, A.; Li, Z.; Yuan, C.; and Li, Y. 2024. ViTKD: Feature-based Knowledge Distillation for Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1379–1388.

Yang, Z.; Zeng, A.; Li, Z.; Zhang, T.; Yuan, C.; and Li, Y. 2023. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17185–17194.

Zagoruyko, S.; and Komodakis, N. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *International Conference on Learning Representations*.

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