IRKD: Knowledge-Distilled Vision Transformer via Inter-Resolution Training

Ahmad Jawwad [1] **Salman Ajmal** [2]

Abstract

In this paper, we introduce IRKD, a novel training framework that integrates Knowledge Distillation (KD), Curriculum Learning (CL), and saliency map-based supervision through inter-resolution training. Our method strategically employs curriculum learning by progressively increasing input resolution, starting from simpler, lower-resolution images and advancing to more complex, higher-resolution inputs, thereby facilitating efficient model training. Concurrently, we leverage gradient-based saliency maps to enforce spatial alignment between the student and teacher models, guiding the student's attention toward informative regions within images. This combined methodology aims to ensure effective knowledge transfer from teacher to student models by structuring the learning process in a more efficient and targeted manner.

Introduction

Vision Transformers (ViTs) have revolutionized computer vision, achieving state-of-the-art performance across numerous benchmarks. However, their superior accuracy often comes at the expense of significant computational resources and memory requirements, posing substantial challenges for training. Consequently, there is a critical need to enhance the training pipeline for ViTs without sacrificing their performance.

In this paper, we propose a novel framework that combines Curriculum Learning (CL) with Knowledge Distillation (KD) to enhance the training pipeline. We further show that using saliency maps can improve Curriculum Learning by guiding the training process progressively from simpler to more challenging examples, thereby enhancing convergence speed and robustness (Bengio et al.). Simultaneously, saliency maps enable the student model to focus on the most informative regions of input images, encouraging more effective knowledge transfer (Simonyan et al.). By strategically orchestrating curriculum-based task difficulty and leveraging intermediate teacher checkpoints, our integrated approach aims to significantly reduce the performance gap between teacher and student. Our extensive experiments demonstrate substantial improvements in both efficiency and generalization performance, underscoring the effectiveness of our proposed methodology.The full implementation of our approach is available on Github here.

Related Work

ViTs, while powerful, are computationally intensive due to their large number of parameters and the quadratic complexity of self-attention mechanisms and faster training mechanisms. Techniques such as quantization(Jiao et al.), and Low-Rank Adaptation (LoRa), address these challenges-mainly post-training. Network pruning, and Knowledge Distillation (KD) are also prominent mechanisms.

Knowledge Distillation (KD) is one prominent technique addressing this need by transferring knowledge from a larger, complex "teacher" model to a smaller, compact "student" model (Hinton et al.). Although effective, traditional KD typically employs a one-time transfer of the teacher's final output representations, often leaving a notable performance gap between teacher and student, particularly in deeper architectures such as ViTs (Jiao et al.). Alternative strategies such as pruning (Frankle and Carbin), quantization (Jacob et al.), and low-rank factorization (Kim et al.) have also been explored, each offering distinct trade-offs between efficiency and accuracy.

Irandoust et al. (2022) address the high computational cost of training vision Transformers (ViTs) by proposing an efficient training strategy that enables training a ViT model from scratch using a single GPU in under 24 hours. Their method introduces two key contributions: first, a locality-enhanced feed-forward network that integrates 3×3 depth-wise convolutions into the Transformer's architecture, promoting local spatial interactions within patches; and second, a novel image size-based curriculum learning strategy, which begins training on low-resolution images and gradually increases the image size over time. This curriculum reduces the sequence length in early training stages, thereby lowering computational overhead and enabling faster convergence. Combined, these techniques allow for substantial improvements in training speed and final model accuracy under tight hardware constraints. Compared to baseline models such as DeiT-S and LocalViT-S, their approach demonstrates significant gains in top-1 accuracy on Ima-

Figure 1. CIFAR-10 @ different Resolutions

geNet1k, while maintaining a minimal training footprint. This work highlights the effectiveness of algorithmic innovations, rather than hardware scaling alone, in making ViT training more accessible and environmentally sustainable

Methodology

Curriculum Learning (CL), drawing inspiration from human pedagogy, structures training by presenting samples in a meaningful order, typically from easy to hard. This involves a difficulty measurer to rank samples and a training scheduler to control their introduction. In ViT training, Dinh et al. (2023) demonstrated a CL strategy of progressively increasing input image resolution, which improved accuracy and reduced training time. This approach aligns with CL principles by starting with computationally cheaper, simpler tasks before moving to more complex ones, mitigating the initial high computational load of ViTs.

Saliency maps, a visual explanation technique, highlight input regions most influential to a model's decision, enhancing transparency. Methods like Grad-CAM generate these maps. Beyond interpretability, saliency maps can guide training. Del Amor et al. used attention maps (via Grad-CAM) in inter-resolution KD, training a student model on low-resolution images to match the teacher's attention on high-resolution images, thereby improving performance by transferring information about discriminative regions.

However, the integration of KD with CL remains largely unexplored. CL can provide a structured learning path for the student, starting with simpler instances where the teacher's knowledge is more easily absorbed, then gradually moving to complex ones. This could improve both learning efficiency and the student's final performance by making the distillation process more manageable .

Datasets and Implementation Details

Two experiments are conducted on the CIFAR-10 dataset, comprising 60,000 color images of resolution 32×32, evenly distributed across 10 object classes. The standard split of 50,000 training and 10,000 test images is used, with 10% of the training data held out for validation. As should in Figure 1, we scale the dataset to 12 * 12 and move up to 32 in increments of 4. We see as a finer, more detailed , image corresponding to greater features that the model has to learn. Upsampling images from 32 by 32 was adding noise which was not generating the type of dataset that we needed. A fixed random seed is applied to ensure reproducibility of the train-validation split. We employ the ViT-Small architecture as the teacher model across all experiments.

For our Teacher model, we employ a Vision Transformer (ViT Small) architecture. Input images are normalized using CIFAR-10-specific mean and standard deviation values. In addition to standard augmentations like random cropping and horizontal flipping, we enhance generalization using AutoAugment with reduced intensity, as well as advanced augmentation strategies such as CutMix and MixUp to further diversify the training data and improve teacher performance. While these methods provided a strong foundation, we acknowledge that the teacher model's performance could potentially be further improved through more extensive hyperparameter tuning. We consider this a valuable direction for future work, as it may yield additional gains in overall model accuracy and robustness.

Teacher Training from Scratch

To avoid the mismatch between high-resolution pretraining and our low-resolution curriculum, we trained the teacher ViT entirely from scratch. Pretrained checkpoints are typically learned on 224×224 images with a 16×16 patch size, which proved suboptimal when fine-tuning on our target resolutions (12×12 - 32×32) and smaller patch sizes (2×2 or 4×4). By initializing weights randomly and training the teacher on 32×32 inputs—using exactly the same patch size as in the student curriculum—we ensured that its learned representations were directly compatible with downstream

stages.

Resolution Consistency and Knowledge Transfer

Although the teacher was only exposed to 32×32 images during training, it retained strong performance down to 20×20 resolution. We therefore align both teacher and student to process images at the same resolution at each curriculum stage. Even when the teacher has never seen lower-resolution data during its own training, its stable accuracy in the 20–32 pixel range provides a robust starting point for the student. This resolution-matched distillation allows the student to "stand on the teacher's shoulders," learning first on simpler, coarser inputs before progressing to finer detail. As Table 1 demonstrates, this strategy yields substantial accuracy gains over curriculum learning or distillation alone.

For our Student model, images are normalized using CIFAR-10-specific per-channel means and standard deviations. During training, data augmentation is applied using random cropping with 4-pixel padding and horizontal flipping. Validation and test images undergo normalization only, without augmentation.

Optimizer Selection

When using a resolution-based curriculum, we found that re-initializing the optimizer at each new resolution consistently yielded better student performance than carrying forward a single optimizer state. Intuitively, each resolution stage represents a different "task" in terms of token count, feature granularity, and convergence speed—so it makes sense to start its optimization with a fresh set of hyperparameters and zeroed momentum buffers rather than inherit stale gradients and learning-rate history from the previous stage.

In contrast, when training without curriculum learning (i.e. on the full 32×32 inputs from the outset), we only instantiate the optimizer once at the beginning of training. In that setting the model's objective, batch sizes, and data distribution remain static, so carrying momentum and adaptive learning-rate state throughout seemed natural.

We therefore adopt the following protocol in our experiments:

- **Curriculum Learning (CL)**: At the start of each resolution stage r, we re-instantiate an AdamW optimizer with identical hyperparameters (learning rate 3×10^{-4}, weight decay 1×10^{-4}), zeroing all accumulated moments and adopting fresh parameter groups for decay vs. no-decay.
- **Non-Curriculum (Single-Stage)**: We initialize AdamW once on 32×32 inputs and train end-to-end, preserving optimizer state across all epochs.

Experiment 1: Custom Student Model

Experiment 1 involves deploying the proposed training pipeline on a custom-designed student model with approximately 600,000 trainable parameters. This model was intentionally selected due to its relatively small size, allowing for rapid prototyping and iteration. The primary objective of this stage is to validate the core functionality and stability of the training framework before scaling to more complex architectures.

By utilizing a lightweight model, we efficiently identified any implementation issues, assessed the potential behavior of the knowledge distillation and curriculum learning components in isolation, and establish a basic baseline for performance under constrained computational resources. This step also provides preliminary insights into how the student model responds to guidance from the teacher model under a curriculum-based training regime, thereby informing subsequent experiments with more expressive models.

Experiment 2: ViT-Tiny as Student Model

Experiment 2 transitions to a more advanced architecture by using ViT-Tiny as the student model. ViT-Tiny offers a significantly more expressive capacity compared to the custom model, enabling a more realistic evaluation of the proposed training strategies in settings closer to real-world applications. The use of ViT-Tiny allows us to examine the scalability and generalization of the distillation and curriculum learning techniques to transformer-based architectures.

During this experiment, we observed that a patch size of 4 yielded better results than a patch size of 2 in preliminary runs which was a little shocking as ViTs perform better with greater number of patches, however we realised that 2x2 patch size was too granular leading to low information content (4-pixels only) and model being exposed to more noise.

Experiment 1

Experiment 1.1: Curriculum Learning with Progressive Resolution

To examine the impact of curriculum learning on low-capacity transformer models, we trained a lightweight Vision Transformer (ViT) student model using a resolution-based curriculum learning strategy. The model was trained from scratch on the CIFAR-10 dataset, progressing through six input resolution stages: 12, 16, 20, 24, 28, and 32 pixels. Each stage lasted for 10 training epochs, with model performance monitored through early stopping and checkpointing. This progressive increase acted as a curriculum in resolution that increased the task complexity with a higher resolution.

Curriculum-based training led to a substantial performance

Experiment Setup	Custom Student (PS 2)	ViT Tiny-S (PS 2)	ViT Tiny-S (PS 4)
Fixed-Resolution (CE Only)	66.85	–	74.00
CL	74.89	–	79.67
KD	76.70	82.28	81.35
KD + CL	77.12	79.92	83.29
KD + CL + Saliency Maps	**78.72**	**81.67**	**84.29**
Teacher Model (ViT-Small)	82.95	82.95	86.00

Table 1. Top-1 classification accuracy (%) of ViT student under different training regimes. PS represents Patch Size for both student and teacher

improvement, with the student model achieving a top-1 test accuracy of **74.89%**, compared to 66.85% under fixed-resolution.

An explanation for this improvement lies in the input complexity. During the early stages of training, low-resolution inputs contain reduced spatial detail, effectively simplifying the learning task and reducing the risk of overfitting to high-frequency noise or irrelevant patterns. This encourages the model to focus on coarse, generalizable features essential for foundational representation learning. As training progresses and the input resolution increases, the model is gradually exposed to more intricate visual information, enabling it to refine its early representations into more discriminative and class-specific features. This allows for a faster training regime in the earlier phase.

In fixed-resolution training (Figure 2), we observe a clear gap between training and validation accuracy, especially in later epochs. While training accuracy continues to rise, validation accuracy quickly plateaus and fluctuates, indicating overfitting and poor generalization.

In contrast, curriculum training (Figure 3) shows a more balanced progression. Both training and validation accuracies improve steadily, with noticeable gains at each stage transition. The smaller gap between the two curves suggests better generalization. The staged increase in input resolution allows the model to first learn robust, general features on simpler data before refining its understanding on more complex inputs, leading to more stable training and improved final performance.

A resolution-wise performance analysis (Figure 4) further illustrates this dynamic. At lower resolutions (12px to 20px), the model shows consistent improvements in validation accuracy across epochs, suggesting that the simplified input facilitates the learning of high-level abstractions. These conditions help mitigate overfitting and encourage the formation of robust, transferable features. In contrast, at higher resolutions (24px to 32px), while training accuracy continues to rise, validation accuracy exhibits increased variability. This reflects the growing challenge of generalization as the task complexity escalates, underscoring the need for gradual exposure to high-resolution inputs.

Overall, the staged resolution progression serves as an implicit form of regularization and capacity scheduling (Irandourst, Saghar et al.) . Rather than confronting the model with full-resolution inputs from the outset—an approach that can lead to suboptimal convergence or overfitting—curriculum learning ensures a smoother optimization trajectory. By aligning input complexity with the model's representational maturity, it promotes convergence to flatter minima, which are typically associated with better generalization.

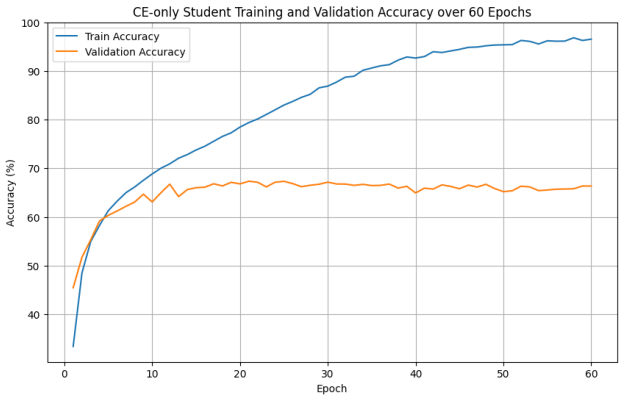

Figure 2. Training and validation accuracy over epochs for ViT student model trained with fixed-resolution (cross-entropy only).

Experiment 1.2: Knowledge Distillation

In the second stage, a high-capacity transformer model is trained from scratch to serve as a teacher. Tiny-S is trained using the knowledge distillation framework introduced in DeiT, where a dedicated distillation token is employed to guide the student through soft supervision from a strong teacher model (Touvron et al., 2021). The teacher model, trained for 135 epochs, achieves a top-1 accuracy of 82.95%, while the student model reaches a top-1 accuracy of 76.70%. For comparison, the same student architecture trained from scratch using only cross-entropy loss achieved a top-1 accu-

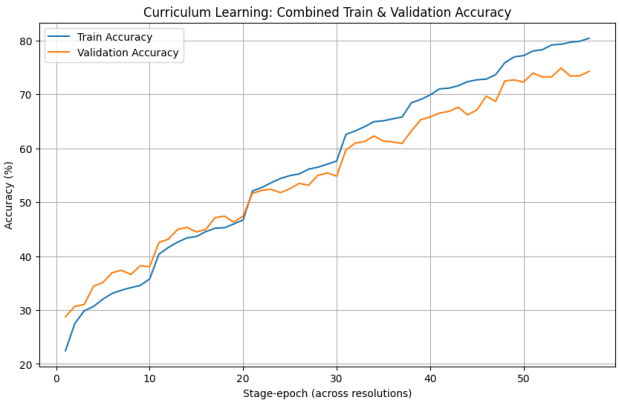

Figure 3. Training and validation accuracy over epochs for ViT student model trained with progressive resolution curriculum learning.

racy of 66.85%, highlighting a substantial 9.85 percentage point improvement attributable to the use of knowledge distillation.

Experiment 1.3: Knowledge Distillation with Cirriculum Learning

Building upon the previous knowledge distillation (KD) setup, we further investigated the impact of incorporating curriculum learning into the distillation process. The student model was trained using the distillation framework, but with an added resolution-based curriculum learning (CL) strategy. Specifically, training was structured into six progressive stages of increasing image resolution: 12, 16, 20, 24, 28, and 32 pixels, with each stage spanning 10 epochs. This approach aimed to align the complexity of the visual input with the model's learning capacity over time, allowing the student to first focus on coarse, low-resolution features before gradually refining its representations on more detailed inputs. The final student model, trained with both KD and curriculum learning, achieved a top-1 test accuracy of 77.12%.The results suggest that curriculum learning contributes additional training stability and regularization, enabling the student to better align with the teacher's outputs during early learning phases. Moreover, the resolution-based progression helps mitigate the optimization challenges commonly faced in KD, particularly when student models are prone to underfitting complex high-resolution data early in training. These findings indicate that the integration of curriculum learning into the KD framework offers a principled, low-overhead enhancement to student model performance and may generalize well to other distillation scenarios.

It is to be noted that traditional regularization via hyperparameter tuning (e.g., dropout rates, weight decay) can also improve generalization, however, it often involves a time-consuming and computationally intensive search for optimal values. In contrast, curriculum learning offers a more straightforward and efficient alternative. Acting as an implicit regularizer, it reduces training time while still enhancing model performance and learning, making it a practical and effective strategy for training compact models.

Experiment 1.4: Curriculum Learning with Knowledge Distillation and Saliency Supervision

To build upon the existing KD + CL framework, we integrated saliency map supervision to guide the student model's spatial attention throughout training. This method leverages gradient-based saliency maps, inspired by Simonyan et al., computed at selected transformer layers (1st, 2nd , last). The goal was to enforce spatial alignment between the student and teacher models early on by minimizing the mean squared error between their normalized saliency maps. Training proceeded in a resolution-based curriculum, where input sizes increased progressively from 12 to 32 pixels across six stages. At each stage, the total loss combined the standard cross-entropy and KD losses from the base model with an additional saliency loss term:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \alpha \cdot \mathcal{L}_{KD} + \gamma \cdot \mathcal{L}_{saliency} \quad (1)$$

The saliency supervision was implemented by registering hooks on intermediate transformer layers to extract features during forward passes, followed by computing attention maps. Visualizations of these saliency maps (Figure 5) illustrate how student attention evolves under this guidance. At lower resolutions (e.g., 12px), initial student maps appear noisy and weakly aligned with the teacher's focus. However, by the end of each stage, especially at epoch 10, the student learns to focus on more semantically relevant regions. This effect becomes more pronounced at intermediate resolutions like 20px, where the student's saliency maps exhibit more structured spatial emphasis that closely resembles the teacher's patterns. At the final stage (32px), the alignment is notably strong, with the student consistently capturing the same object-centric regions as the teacher, indicating successful spatial transfer. These observations confirm that saliency supervision under CL enhances the spatial inductive bias of the student model. Quantitatively, augmenting the base KD + CL model with saliency supervision improved test accuracy to **79.37%**, demonstrating the effectiveness of spatial alignment in boosting generalization performance.

These results demonstrate the synergistic benefits of combining curriculum learning, knowledge distillation, and saliency-based supervision. The use of visual attention alignment helps compact models learn more effective representations under constrained capacity and resolution regimes.

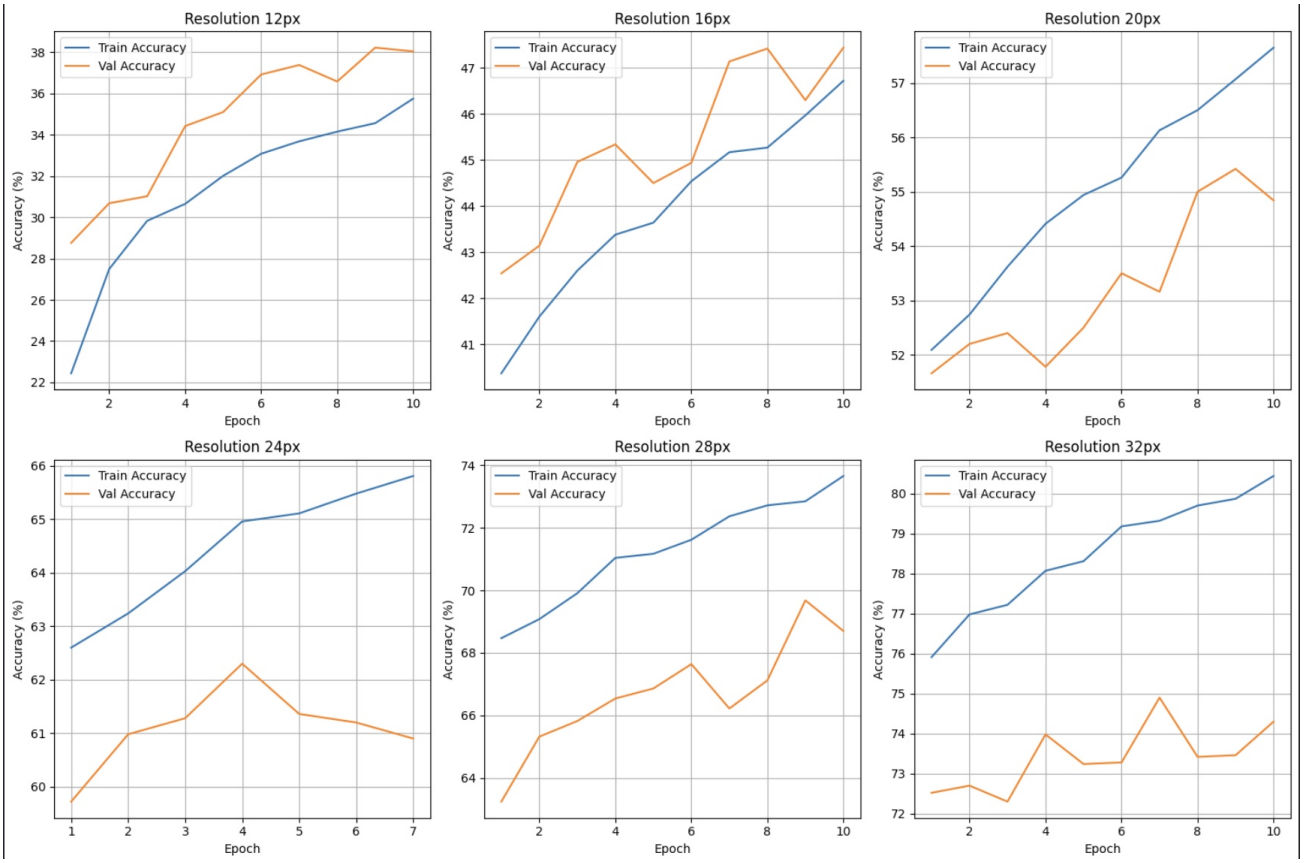

Figure 4. Training and validation accuracy over epochs for ViT student model trained with progressive resolution curriculum learning per resolution

Experiment 2

In the following experiments we use two ViT-Tiny models (patch size 2 and patch size 4). We notice a considerable improvement in the overall performance with the latter patch size.

Experiment 2.1: Curriculum Learning with Progressive Resolution

We use ViT-Tiny as the student model and experiment with a patch size of 4 only for this stage. The CE-only baseline achieves an accuracy of 73.5%, which improves to 79.8% under curriculum learning—a notable gain of 5.67 percentage points. A key insight from these results is the effectiveness of using larger patch sizes in this setting. Larger patches introduce an inductive bias toward capturing global image structure, which is especially beneficial in classification tasks like CIFAR-10. With fewer tokens per image, the model can more efficiently learn high-level, spatially coherent features—particularly valuable during the early curriculum stages when input resolution is low.

Additionally, the reduced number of tokens simplifies self-attention computations and eases the representational burden on the model. This leads to faster convergence and improved learning dynamics, which is especially important for lightweight transformer architectures that may otherwise struggle with small-patch inputs due to their fragmented nature. In this context, larger patch sizes offer a more favorable balance between spatial granularity and computational efficiency, ultimately contributing to better generalization and overall performance.

Experiment 2.2: Knowledge Distillation with ViT-Small Teacher

We now explore KD baselines for both ViT-Tiny models - patch size 2 and patch size 4. We see a baseline accuracy of 82.28% and 81.35% respectively. We will use these as the baselines to check the improvement of the IRKD curriculum that we introduce.

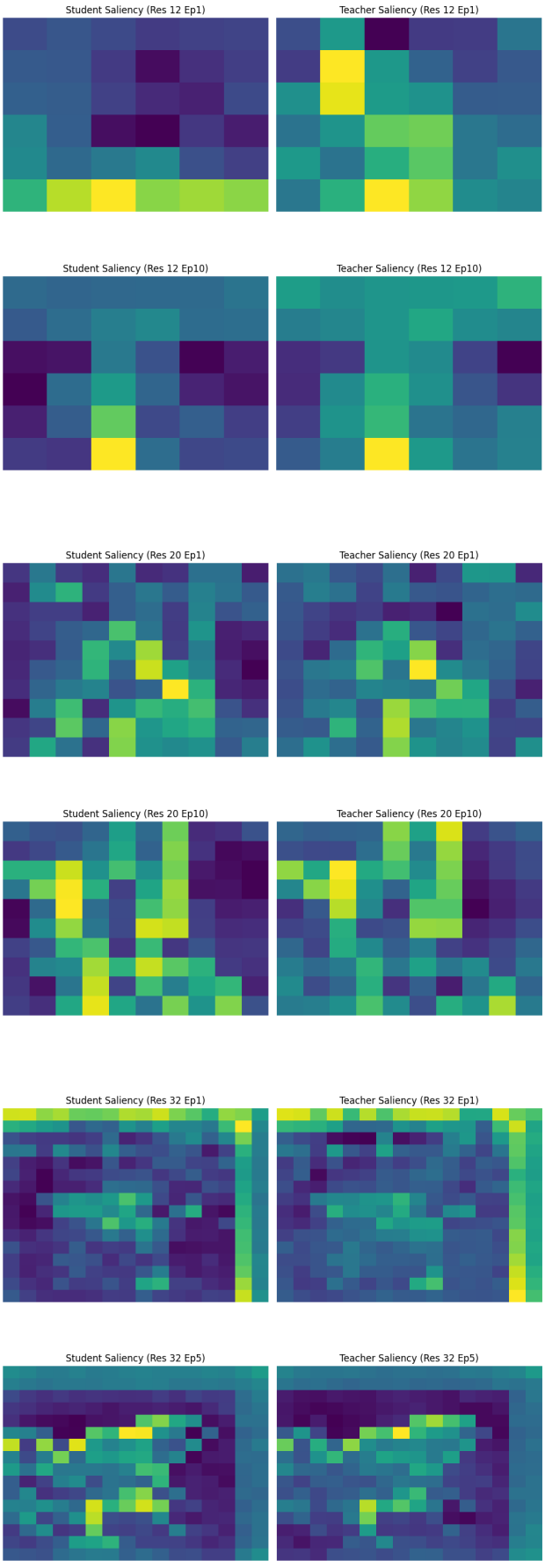

Figure 5. Saliency maps across different resolutions and epochs.

Experiment 2.3: Knowledge Distillation with Curriculum Learning

By applying the six-stage, resolution-based curriculum while varying the ViT's patch size, we observe that a finer 2×2 patch partition yields a respectable 79.92 % top-1 accuracy, whereas a coarser 4×4 patch configuration further boosts performance to 83.29 %. The larger patches effectively simplify early representations—complementing the curriculum's low-resolution focus—allowing the student to capture salient structures before refining subtle details, while still benefiting from the teacher's guidance via distillation.

Experiment 2.4: Curriculum Learning with Knowledge Distillation and Saliency Supervision

We finally combined Knowledge Distillation (KD) and Curriculum Learning (CL) framework with saliency map supervision to guide the student model's spatial attention across all 12 layers. Saliency maps were computed across all the transformer layers and used to align the spatial focus of the student with that of the teacher. The total training loss integrated cross-entropy, KD, and saliency components, as defined in Equation 1. Experiments were conducted using a ViT-Tiny student model with a fine-grained patch size of 2 reaching an accuracy of **81.67%** while student model with patch size 4 achieved an accuracy of **83.29**.

Discussion

The experimental results for Tiny-ViT with patch size 4 demonstrate a consistent advantage of incorporating resolution-based curriculum learning into the training of compact Vision Transformers. The progressive increase in input complexity enabled the models to develop robust low-level feature representations before encountering high-resolution inputs, which in turn facilitated better generalization. This was evident in both the custom student and ViT-Tiny P-4 architectures, where curriculum learning alone outperformed fixed-resolution training. Notably, in the ViT-Tiny model with a 4×4 patch configuration, curriculum learning led to an improvement from 74.00% to 79.67%, underscoring its role not merely as a training heuristic, but as a structural enhancement to the learning process.

When curriculum learning was coupled with knowledge distillation the Tiny-S with patch-4 exhibited a higher top-1 accuracy of **83.29%**. It is to be noted that we did not see a similar increase in the Tiny-Vit P-2 model's accuracy rather it plateaued slightly.

The integration of saliency supervision into the KD + CL pipeline further refined the student model's learning, particularly in terms of spatial focus. Saliency-guided training encouraged alignment between the student and teacher in terms of attention distribution. The improvement in accu-

racy—from 83.29% to 84.29% with saliency maps—may appear incremental, but it reflects a deeper alignment in feature space, enabling the student to prioritize semantically relevant regions even with fewer parameters. The qualitative evolution of saliency maps across training stages illustrated this effect clearly: initial maps were diffuse, while later stages exhibited strong convergence toward the teacher's attention regions. This progression validates the hypothesis that attention supervision, when synchronized with resolution-based curriculum learning, enhances the spatial inductive bias of compact ViTs.

0.1. Future Improvements in Experimental Setup

Moving forward, we plan to refine our training protocol in two key ways:

- **Optimizer State Continuity:** Rather than re-instantiating AdamW at each resolution stage, we will carry forward its moment estimates and learning-rate history across the curriculum. This could reduce stage-to-stage interference and allow smoother convergence, as the optimizer will preserve useful gradient information instead of restarting from scratch.
- **Enhanced Teacher for Tiny (PS=2):** To disentangle the impact of patch granularity from teacher underfitting, we will improve the ViT-Tiny teacher trained with patch size 2 by:
  1. Pretraining on a higher resolution (e.g. 64×64) before fine-tuning on CIFAR-10 32x32 (apply reverse inter-resolution),
  2. Introducing additional regularization (e.g. stronger weight decay, stochastic depth)

  These enhancements will yield a more robust results and clarify whether the observed performance gap arises from excessively fine patches or insufficient teacher capacity. It will further allow us to confirm our hypothesis.

Conclusion

The experimental findings underscore that thoughtfully designed training strategies can unlock substantial performance gains in compact Vision Transformers, without the need for increasing model size or computational overhead. By integrating curriculum learning with knowledge distillation and saliency-guided supervision, the proposed IRKD framework achieves better accuracy than traditional fixed-resolution or standalone distillation baselines. Notably, it does so with improved training efficiency—converging faster and consuming fewer computational resources by leveraging low-resolution inputs in early stages and optimizing attention through spatial alignment. This makes IRKD not only an effective approach in terms of accuracy but also a practical solution for real-world deployment where training time and resource usage are critical considerations.

References

Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations (ICLR), 2021.

Wang, Chaofei, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. "Efficient Knowledge Distillation from Model Checkpoints." arXiv, 12 Oct. 2022, arXiv:2210.06458v1.

Irandoust, Saghar, et al. "Training a Vision Transformer from Scratch in Less Than 24 Hours with 1 GPU." arXiv, 9 Nov. 2022, arXiv:2211.05187v1.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." arXiv, 16 Nov. 2014, arXiv:1312.6034v2.

Touvron, Hugo, et al. "Training data-efficient image transformers distillation through attention." *International Conference on Machine Learning*, 2021, pp. 10347–10357.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network." arXiv preprint arXiv:1503.02531, 2015.

Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." arXiv preprint arXiv:2106.09685, 2021, .

Del Amor, Rocío, et al. "Attention to Detail: Inter-Resolution Knowledge Distillation." arXiv preprint arXiv:2401.06010, 2024.

Habib, Gousia, et al. "Knowledge Distillation in Vision Transformers: A Critical Review." arXiv preprint arXiv:2302.02108, 2023,.

Jacob, Benoit, et al. "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2704–2713

Papa, Lorenzo, et al. "A Survey on Efficient Vision Transformers: Algorithms, Techniques, and Performance Benchmarking." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, Dec. 2024, pp. 7682–7700. .

Li, Zheng, et al. "Curriculum Temperature for Knowledge Distillation." ArXiv.org, 24 Dec. 2022,.

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston. "Curriculum Learning." Proceedings of the

26th Annual International Conference on Machine Learning (ICML), 2009

Panigrahi, Abhishek, et al. “Progressive Distillation Induces an Implicit Curriculum.” ArXiv.org, 2024, . Accessed 7 May 2025

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 and cifar-10 (canadian institute for advanced research), 2009. URL http://www.cs.toronto.edu/ kriz/cifar.html. MIT License.

Gupta, Shivam, and Sushrut Karmalkar. “Efficient Knowledge Distillation via Curriculum Extraction.” ArXiv.org, 2025, www.arxiv.org/abs/2503.17494. Accessed 7 May 2025.