Cross-Architectural Knowledge and Attention-Guided Distillation: A Novel Approach with Hybrid Loss Function

MD. Shafat Islam khan
ID-2121517642

Saif Uz Zaman
ID-1931587042

Shadman Ahmed Abeer
ID-2121835642

***Abstract*—Knowledge distillation has emerged as a pivotal approach to transferring knowledge from large, complex teacher models to lightweight student models, speeding up the prediction process while preserving performance. In this study, we propose a novel framework that utilizes a hybrid teacher architecture combining VGG19 and ViT and a student model based on EfficientNet. To enhance the interpretability and reliability of the student model, we applied Attention-guided distillation, emphasizing the alignment of internal attention mechanisms between models.Our approach incorporates multiple loss functions, including Structural Similarity Index Measure (SSIM), cosine similarity, Mean Squared Error (MSE), and Perceptual Loss, to optimize knowledge transfer. Additionally, we introduce a novel hybrid loss function that combines SSIM and cosine similarity, ensuring both structural integrity and vectorial alignment between teacher and student models. This Cross-Architectural Knowledge and Attention-Guided Distillation framework effectively applies the complementary strengths of CNNs and ViTs, demonstrating the potential for reliable and interpretable student models.The results show substantial improvements in student model performance, especially for object detection tasks under resource-constrained environments. This study highlights the potential of applying cross-architectural knowledge distillation with attention-guided mechanisms, setting a new benchmark for future research in knowledge distillation frameworks.**

***Index Terms*—Knowledge Distillation, Attention-Guided Distillation, Hybrid Loss Function,Cross-Architectural Framework**

I. Introduction

The increasing complexity of deep learning models has led to significant improvements in predictive performance across various domains. However, these models often demand substantial computational resources, posing challenges for deployment in real-world, resource-constrained scenarios. Knowledge distillation has emerged as a promising solution. A compact student model learns from a larger, pre-trained teacher model. Despite its success, significant challenges remain, especially in cross-architectural knowledge distillation, where models with different structures and characteristics are involved [1], [2]. Yang et al. [1] highlight that cross-architecture knowledge distillation can effectively leverage the distinct capabilities of convolutional and transformer-based models by aligning their feature representations.

Similarly, Liu et al. [3] demonstrated that incorporating structural regularization in cross-architecture setups could significantly improve model reliability. Through this method, we aim to improve the transfer of knowledge from a hybrid teacher model to a student model, with a focus on capturing both the global attention patterns and fine-grained spatial details, ensuring better performance. We extend traditional knowledge distillation by aligning the internal attention mechanisms of the teacher and student models, ensuring that the student not only replicates predictions but also learns the interpretative attention patterns of the teacher. Shen and Jiang [2] introduced attention-guided answer distillation, which aligns attention mechanisms to enhance student interpretability; this idea motivates our attention-guided approach. Mansourian et al. [4] demonstrated the importance of multi-layer attention mechanisms for fine-grained tasks, supporting our focus on aligning both global and local features.This approach builds on prior works that emphasize hybrid architectures and customized loss functions in advancing efficient knowledge transfer. For example, Li et al. [5] used coordinate attention to strengthen student learning under dual-teacher settings, while Lee et al. [6] demonstrated how combining structural decomposition with knowledge distillation can enhance interpretability in resource-constrained environments.This research focuses on addressing these through a hybrid attention guided distillation framework that combines innovative architectural design along with multiple loss functions and a customized loss function. Our main contributions are -

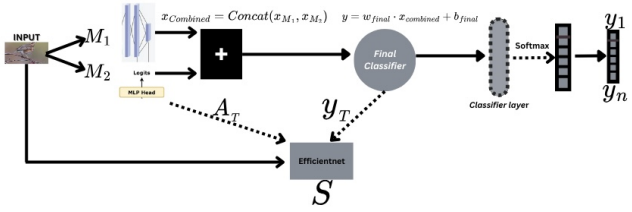

Fig. 1. Hybrid model where M_1 is VGG19 and M_2 is ViT. Their outputs are concatenated and then passed through a final classifier layer followed by a softmax function. The student model S (EfficientNet) learns from the teacher's logits Y_T and attention maps A_T. Various loss functions were used, including a novel hybrid loss function introduced in this work.

- **Hybrid Teacher Architecture**: We introduce a novel teacher model that combines the feature extraction capabilities of VGG19 with the global attention mechanisms

of Vision Transformer (ViT).

- **Custom Hybrid Loss Function**: We develop a novel loss function that combines SSIM and cosine similarity to exploit both structural integrity and vectorial alignment between teacher and student attention maps.
- **Attention-Guided Knowledge Distillation with Multiple Loss Function Optimization** - We applied attention-guided distillation between teacher and student models, utilizing multiple loss functions, including Mean Squared Error (MSE), Structural Similarity Index (SSIM), Cosine Similarity, and Perceptual Loss, to enhance the performance and accuracy of the distillation process..
- **Efficient Knowledge Transfer**: We demonstrate the effectiveness of our method on challenging datasets, achieving reliable performance improvements with significant computational efficiency.

This approach not only explores Cross architectural Knowledge and Attention guided distillation but also establishes a reliable methodology for extracting and transferring both local and global representations.Through extensive experiments, we validate the superiority of our approach, setting a new benchmark for efficient knowledge distillation in cross-architectural settings.

II. Literature Review

There is no exact work that aligns perfectly with our research. However, we found some similar works related to our study.

In the first paper, Yufan and co-authors worked in the same domain of Knowledge Distillation [1]. They conducted similar research, using Transformer and CNN features via cross-attention and group-wise linear projectors, paired with a robust multi-view training scheme. They mapped CNN Query, Key, and Value using 3×3 convolutional layers and layered Q, K, and V from the teacher Transformer through the transformer attention space. They then employed a group-wise linear projector to map the student features into the Transformer feature space. Their research utilized both large-scale and small-scale datasets like ImageNet and CIFAR. The results were effective, outperforming 14 state-of-the-art methods on both large and small datasets.

In the second paper, we found similar work on Attention-Guided Answer Distillation [2]. In this study, the authors developed a system where the model reads a passage and answers given questions. To accomplish this, they transferred knowledge from an ensemble model to a single model using attention-guided knowledge distillation. They also performed answer distillation to train the student model. The work primarily utilized MRC datasets and achieved impressive results, with an F1 score improvement of nearly 80%, experiencing only a 0.4% drop in the F1 score. The distillation process sped up the model compared to the ensemble model. However, their research had some drawbacks, such as suboptimal accuracy and a lack of comparative analysis.

In the third paper, Sebastian and co-authors conducted research on neural ranking models with cross-architecture knowledge distillation [3]. They worked on Q/A or text-based recommender systems. Their primary objective was to develop a more efficient model for distillation across various scores of output destinations. They explored BERT, ColBERT, and similar architectures. By transferring knowledge from a BERTCAT teacher model to student models like BERTDOT, ColBERT, and PreTT, they significantly improved re-ranking effectiveness without compromising query latency. On the MSMARCO-DEV dataset, ColBERT's nDCG@10 improved from 0.417 to 0.431, and MRR@10 from 0.357 to 0.370. Similarly, BERTDOT showed gains with nDCG@10 increasing from 0.373 to 0.388 and MRR@10 from 0.316 to 0.330. On the TREC-DL'19 dataset, the ensemble teacher model further improved results. The DistilBERTDOT student achieved nDCG@10 of 0.697 compared to a baseline of 0.626 and MRR@10 of 0.868, showcasing substantial gains. However, the reliance on high computational resources for pre-trained BERTCAT teachers remains a limitation.

Another paper, by Gausia and co-authors, reviewed knowledge distillation in Vision Transformers [4]. This critical review focused on Natural Language Processing, where transformers have revolutionized the field with attention-based encoder-decoder architectures. They evaluated KD techniques for compressing Vision Transformers (ViTs) into resource-efficient models without compromising performance. For instance, the Tiny-ViT model achieved 84.8% Top-1 accuracy with only 21M parameters, comparable to larger models like Swin-B. However, KD methods often require extensive computational resources and pre-trained teacher models, making real-time deployment challenging. The reliance on patch-level mappings and pre-computed logits is also inefficient for large datasets.

In another study, Amir and co-authors proposed the Attention-guided Feature Distillation (AttnFD) method for semantic segmentation [5]. Using the Convolutional Block Attention Module (CBAM), they refined feature maps by incorporating channel and spatial attention. The teacher network's refined features were distilled into the student network using Mean Squared Error (MSE) loss. AttnFD achieved state-of-the-art performance, significantly improving the mean Intersection over Union (mIoU) across multiple datasets like PascalVOC, Cityscapes, COCO, and CamVid, enhancing mIoU by up to 8.95% with lightweight student models like ResNet18 and MobileNet. However, reliance on pre-trained models and computationally demanding methods remain limitations.

Don and co-authors introduced Coordinate Attention Guided Dual-Teacher Adaptive Knowledge Distillation (CAG-DAKD) to enhance image classification by distilling knowledge from two teacher networks to a lightweight student network [6]. However, their research lacked statistical data.

Ebrahim and co-authors conducted research on hybrid attention model knowledge distillation using feature decomposition for glucose forecasting [7]. They proposed GlucoNet, a hybrid model combining LSTM and Transformer. Using Variational Mode Decomposition (VMD), they managed non-linear and non-stationary glucose data patterns. GlucoNet achieved 60%

RMSE improvement and reduced parameters by 21%, outperforming state-of-the-art methods with improvements in RMSE and MAE metrics by 51% and 57%, respectively. However, the reliance on high-quality multimodal input data poses challenges for model predictions and deployment on edge devices.

Zhimeng and co-authors proposed cross-architecture knowledge distillation for efficient monocular depth estimation [8]. They developed a convolutional-based Monocular Depth Estimation (MDE) model, distilling knowledge from transformer models using a ghost decoder and attentive distillation loss on KITTI and NYU Depth V2 datasets. Their model achieved competitive performance with significant reductions in FLOPs, parameters, and latency. For instance, DisDepth-B0 achieved similar performance with less than 7.5% of FLOPs compared to BTS. However, the approach's specificity and novelty limit its general applicability.

Nitay and co-authors explored knowledge distillation for Natural Language Generation (NLG) with pseudo-target training [9]. They examined modeling decisions and implemented improvements based on performance evaluations. Using KD with Joint-Teaching and pseudo-targets, they achieved 75% of the teacher-student performance gap closure for tasks like summarization and question generation. However, the model's reliance on pseudo-targets and large teacher models like GPT-4 makes it computationally expensive.

Sangchul and Heeyoul proposed Self-Knowledge Distillation (SKD) to improve NLP tasks like language modeling and neural machine translation [10]. Their experiments on the Penn Treebank dataset reduced the negative log-likelihood from 101.40 to 99.38 and increased BLEU scores from 9.01 to 9.87. However, SKD requires high-quality word embeddings and substantial computational resources.

Finally, Siqi and co-authors implemented knowledge distillation for BERT models in the healthcare sector [11]. Using Patient Knowledge Distillation, they trained students from multiple intermediate teacher layers via PKD-Last and PKD-Skip strategies. They achieved a minimal performance drop (2.3% on SST and 1.4% on QNLI) compared to the 12-layer BERT teacher model, with a speed-up of 1.94x. However, their model's sensitivity to training data volume limits its application.

Weisong and co-authors utilized cross-architecture knowledge distillation for face recognition [12]. They transferred knowledge from a Transformer teacher to a CNN student using Adaptable Prompting Teacher (APT) and Unified Receptive Fields Mapping. On IJB-C, they achieved TPR@FPR=1e-4 of 94.40%, outperforming DarkRank's 93.06%. However, their reliance on Transformer models limits applicability in resource-constrained environments like mobile devices.

III. Methodology

A. Datasets

For our experiments we have worked with the following datasets:-

- Caltech101- 101 classes, 9,146 images, 300x200 pixels
- Caltech256- 256 classes, 30,607 images
- Dtd- 47 classes, 5,640 images
- Cifar10- 10 classes, 60,000 images, 32x32 pixels
- Cifar100- 100 classes, 60,000 images, 32x32 pixels
- Coil100- 100 classes, 72,000 images, 128x128 pixels
- MIT Indoor Scenes- 67 classes, 15,000 images
- Dermnet- 23 classes, 23,000 images
- TinyImageNet200- 200 classes, 64x64 pixels

B. Problem Formulation

Knowledge distillation (KD) is a training paradigm where a smaller, more efficient *student model* learns to mimic the behavior of a larger, more complex *teacher model*. In this work, we consider a hybrid teacher model T that integrates the complementary strengths of a convolutional neural network (VGG19) and a vision transformer (ViT). The student model S is based on EfficientNet, chosen for its computational efficiency and strong performance on image classification tasks.

1) Teacher Model: Hybrid Architecture: The teacher model T processes an input image $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W represent the number of channels, height, and width of the image, respectively. The hybrid teacher consists of two branches:

- **VGG19 Branch:** Outputs a feature vector $\mathbf{f}_{\text{VGG}} \in \mathbb{R}^d$, where $d = \text{num\_classes}$.
- **ViT Branch:** Outputs a feature vector $\mathbf{f}_{\text{ViT}} \in \mathbb{R}^d$, where $d = \text{num\_classes}$.

The feature outputs are concatenated to form a combined representation:

$$\mathbf{f}_{\text{combined}} = \text{Concat}(\mathbf{f}_{\text{VGG}}, \mathbf{f}_{\text{ViT}}) \in \mathbb{R}^{2d}.$$

The final logits are computed by a fully connected layer:

$$\mathbf{z}_T = \mathbf{W}_{\text{final}}\mathbf{f}_{\text{combined}} + \mathbf{b}_{\text{final}}, \quad \mathbf{z}_T \in \mathbb{R}^d,$$

where $\mathbf{W}_{\text{final}} \in \mathbb{R}^{d\times 2d}$ and $\mathbf{b}_{\text{final}} \in \mathbb{R}^d$ are the learnable parameters of the final classifier.

2) Student Model: The student model S is a lightweight EfficientNet variant that maps the same input $\mathbf{x}$ to logits $\mathbf{z}_S \in \mathbb{R}^d$ through a series of convolutional operations.

3) Knowledge Distillation Objective: The goal of KD is to transfer the knowledge of the teacher model T to the student model S by minimizing a combined loss function:

$$\mathcal{L}_{\text{KD}} = (1-\alpha)\mathcal{L}_{\text{CE}}(\mathbf{z}_S, \mathbf{y}) + \alpha\tau^2\mathcal{L}_{\text{KL}}(\mathbf{z}_S, \mathbf{z}_T),$$

where:

- $\mathcal{L}_{\text{CE}}(\mathbf{z}_S, \mathbf{y})$ is the cross-entropy loss between the student's predictions and the ground truth labels $\mathbf{y}$.
- $\mathcal{L}_{\text{KL}}(\mathbf{z}_S, \mathbf{z}_T)$ is the Kullback-Leibler divergence between the student and teacher logits:

$$\mathcal{L}_{\text{KL}} = \sum_{i=1}^{d} q_T^{(i)} \log \frac{q_T^{(i)}}{q_S^{(i)}},$$

where $q_T^{(i)}$ and $q_S^{(i)}$ are the softened probabilities for class i computed as:

$$q_T^{(i)} = \frac{\exp\left(z_T^{(i)}/\tau\right)}{\sum_{j=1}^{d} \exp\left(z_T^{(j)}/\tau\right)}, \quad q_S^{(i)} = \frac{\exp\left(z_S^{(i)}/\tau\right)}{\sum_{j=1}^{d} \exp\left(z_S^{(j)}/\tau\right)}.$$

- $\tau > 0$ is the temperature parameter that controls the smoothness of the softened probability distribution.
- $\alpha \in [0, 1]$ is the trade-off parameter that balances the two loss terms.

4) Problem Objective: The optimization objective for the student model is to minimize $\mathcal{L}_{\text{KD}}$, thereby ensuring that the student learns from both the teacher's predictions and the ground truth labels. Formally:

$$\min_{\theta_S} \mathcal{L}_{\text{KD}}(\theta_S),$$

where θ_S are the trainable parameters of the student model.

C. Proposed Methodology: Attention-Guided Knowledge Distillation

We propose an **Attention-Guided Knowledge Distillation (AGKD)** framework to improve the effectiveness of knowledge distillation by focusing on aligning attention maps between the teacher and student models. The hybrid teacher model combines VGG19 and ViT-B16, leveraging their complementary strengths in spatial and patch-level attention mechanisms. The student model, EfficientNet-B0, is trained using various loss functions to optimize both output and internal feature representations.

1) Teacher Model: Hybrid Architecture: The teacher model is a hybrid of VGG19 and ViT-B16, designed to combine spatial feature extraction and patch-based self-attention mechanisms:

- **VGG19 Branch:** The convolutional layers of VGG19 generate spatial feature maps $\mathbf{A}_{\text{VGG}} \in \mathbb{R}^{C_T \times H_T \times W_T}$, where C_T, H_T, and W_T represent the number of channels, height, and width of the feature map.
- **ViT-B16 Branch:** The ViT-B16 model computes attention weights over patches, producing attention maps $\mathbf{A}_{\text{ViT}} \in \mathbb{R}^{N \times N}$, where N is the number of image patches.

The attention maps from VGG19 and ViT are resized to the same resolution and combined using an element-wise addition or concatenation operation to form the teacher attention map:

$$\mathbf{A}_T = \text{Combine}(\mathbf{A}_{\text{VGG}}, \mathbf{A}_{\text{ViT}}),$$

which serves as the reference for training the student model.

2) Student Model: EfficientNet-B0: The student model is a lightweight and efficient EfficientNet-B0, which generates its own attention maps $\mathbf{A}_S \in \mathbb{R}^{C_S \times H_S \times W_S}$ from intermediate layers. These attention maps are resized using bilinear interpolation to match the teacher's attention map $\mathbf{A}_T$, enabling direct comparison and alignment during training.

3) Loss Functions: The training objective of AGKD includes multiple loss components to optimize the student model:

1) **Logits-Based Distillation Loss:** This loss aligns the softened logits of the teacher and student using Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{\text{logits}} = KL(p_T \| p_S),$$

where p_T and p_S are the softened probability distributions of the teacher and student, computed using:

$$p(x) = \frac{\exp(z(x)/T)}{\sum_j \exp(z_j(x)/T)},$$

with $z(x)$ as the logits and T as the temperature parameter.

2) **Attention Alignment Loss:** This loss minimizes the discrepancy between teacher and student attention maps:

$$\mathcal{L}_{\text{attention}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{A}_T^i - \mathbf{A}_S^i\|_2^2,$$

where N is the number of layers for which attention maps are computed.

To improve performance, alternative loss functions were evaluated:

- **Structural Similarity Index (SSIM) Loss:**

$$\mathcal{L}_{\text{SSIM}} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)},$$

where μ_x, μ_y are the means, σ_x^2, σ_y^2 are the variances, σ_{xy} is the covariance, and C_1, C_2 are small constants.
- **Cosine Similarity Loss:**

$$\mathcal{L}_{\text{cosine}} = 1 - \frac{\sum_{i=1}^{d} A_T^i A_S^i}{\|A_T\|_2 \|A_S\|_2},$$

where d is the dimensionality of the attention vectors.
- **Perceptual Loss:**

$$\mathcal{L}_{\text{perceptual}} = \frac{1}{N} \sum_{i=1}^{N} \|\phi_l(x) - \phi_l(y)\|_2^2,$$

where $\phi_l(x)$ and $\phi_l(y)$ are features extracted from a pre-trained network layer l.
- **Hybrid Loss (SSIM + Cosine Similarity):**

$$\mathcal{L}_{\text{hybrid}} = \alpha\mathcal{L}_{\text{SSIM}} + \beta\mathcal{L}_{\text{cosine}},$$

where α and β balance the contribution of each term.

4) Total Loss Function: The overall loss function is defined as:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{logits}} + \beta\mathcal{L}_{\text{attention}}$$

where α, β are weighting factors to balance the contributions of each loss term.

TABLE I
RESULTS OF TEACHER MODEL AND STUDENT MODEL ON SOFT LABEL LOGITS BASED LOSS DISTILLATION FOR SELECTED DATASETS

Dataset	Teacher Model				Student Model			
	Train Loss	Val Loss	Train Acc.	Val Acc.	Train Loss	Val Loss	Train Acc.	Val Acc.
Caltech101	0.44	0.25	88%	92%	0.18	0.15	93%	96%
Caltech256	0.55	0.55	87%	85%	0.20	0.18	96%	97%
DTD	0.46	0.21	85%	93%	1.80	2.24	45%	38%
CIFAR100	0.35	0.68	89%	79%	0.34	0.90	89%	74%
COIL100	0.08	0.09	98%	97%	0.13	0.17	96%	95%
MIT Indoor Scenes	0.56	0.71	83%	79%	0.60	0.75	81%	75%
Dermnet	1.02	1.62	67%	56%	2.00	2.19	39%	34%
CIFAR10	0.11	0.18	96%	94%	0.16	0.26	94%	92%
TinyImageNet200	1.30	0.66	68%	82%	2.73	2.57	35%	52%

TABLE II
PERFORMANCE OF THE STUDENT MODEL WITH DIFFERENT ATTENTION LOSS FUNCTIONS ON IMAGENET200 FOR AGKD

Loss Function	Train Loss	Val Loss	Train Accuracy (%)	Val Accuracy (%)
MSE	2.87	1.95	37.06	54.18
Cosine Similarity	0.983	1.45	74.97	69.62
SSIM	1.24	1.32	68.62	69.32
Perceptual Loss	0.85	1.58	78.24	69.23
Hybrid (SSIM + Cosine Similarity)	1.24	1.29	68.90	69.37

5) Workflow Summary: The proposed AGKD framework consists of the following steps:

- **Attention Map Extraction:** Extract attention maps $\mathbf{A}_{\text{VGG}}$ and $\mathbf{A}_{\text{ViT}}$ from the teacher model. Combine them into a unified attention map $\mathbf{A}_T$.
- **Student Attention Resizing:** Extract attention maps $\mathbf{A}_S$ from the student model and resize them to match the dimensions of $\mathbf{A}_T$.
- **Loss Computation:** Compute the total loss $\mathcal{L}_{\text{total}}$ as a weighted combination of logit-based and attention-based losses.
- **Model Update:** Update the student model parameters θ_S using backpropagation based on the computed loss.

IV. EXPERIMENTS

A. Setup

Hardware and Software Environment: The experiments were conducted on cloud GPU of Google Colab equipped with an NVIDIA A100 GPU. The models were implemented using PyTorch 2.0.1 and the Torchvision library 0.15.2, with Python 3.10.6 as the programming language. Models were trained using the Adam optimizer with a learning rate of (0.01) and a cosine annealing learning rate scheduler. The training process used a batch size of 128 and ran for 15-20 epochs.

Evaluation Metrics: The performance of the models was evaluated using the following metrics:

- **Loss**: Training loss and validation loss were recorded for each experiment.
- **Accuracy**: Training and validation accuracy were computed for all datasets.

These metrics provided a comprehensive evaluation of the models' ability to distill knowledge effectively.

B. Conclusion

This study presents a novel multi-cross architecture knowledge distillation framework, introducing a hybrid loss function for improved performance. By combining SSIM and Cosine Similarity, our approach enhances both spatial and semantic fidelity. Results across diverse datasets validate the generalizability and effectiveness of this methodology The combination of SSIM and Cosine Similarity strikes a balance between preserving spatial features and aligning semantic representations MSE alone achieved 54.18 percent validation accuracy on Tiny ImageNet, while the hybrid loss surpassed 69 percent. SSIM's focus on structural fidelity complements Cosine Similarity's feature alignment, making the hybrid loss highly effective.

REFERENCES

[1] Yufan et al., *Knowledge Distillation with Transformers and CNNs*, 2023.
[2] Authors et al., *Attention-Guided Answer Distillation*, 2023.
[3] Sebastian et al., *Cross-Architecture Knowledge Distillation for Neural Ranking Models*, 2022.
[4] Gausia et al., *A Review on Knowledge Distillation in Vision Transformers*, 2023.
[5] Amir et al., *Attention-Guided Feature Distillation for Semantic Segmentation*, 2023.
[6] Don et al., *Coordinate Attention Guided Dual-Teacher Adaptive Knowledge Distillation*, 2022.
[7] Ebrahim et al., *Hybrid Attention Model Knowledge Distillation for Glucose Forecasting*, 2023.
[8] Zhimeng et al., *Efficient Monocular Depth Estimation via Cross-Architecture Knowledge Distillation*, 2023.
[9] Nitay et al., *Knowledge Distillation for Natural Language Generation with Pseudo-Target Training*, 2023.
[10] Sangchul et al., *Self-Knowledge Distillation for NLP Tasks*, 2022.
[11] Siqi et al., *Patient Knowledge Distillation for Healthcare Applications*, 2023.
[12] Weisong et al., *Cross-Architecture Knowledge Distillation for Face Recognition*, 2022.
[13] Y. Liu, J. Cao, B. Li, W. Hu, J. Ding, and L. Li, *Cross-Architecture Knowledge Distillation*, in Proceedings of the Asian Conference on Computer Vision (ACCV), 2022, pp. 3396–3411.

[14] P. Parmar and B. T. Morris, *Multimodal Knowledge Distillation for Audio-Visual Models*, IEEE Transactions on Multimedia, vol. 23, pp. 2890–2901, 2021. doi:

[15] H. Pham, Z. Dai, Q. Xie, M.-T. Luong, and Q. V. Le, *Teacher-Student Optimization for Robust Deep Learning*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7439–7448. doi: