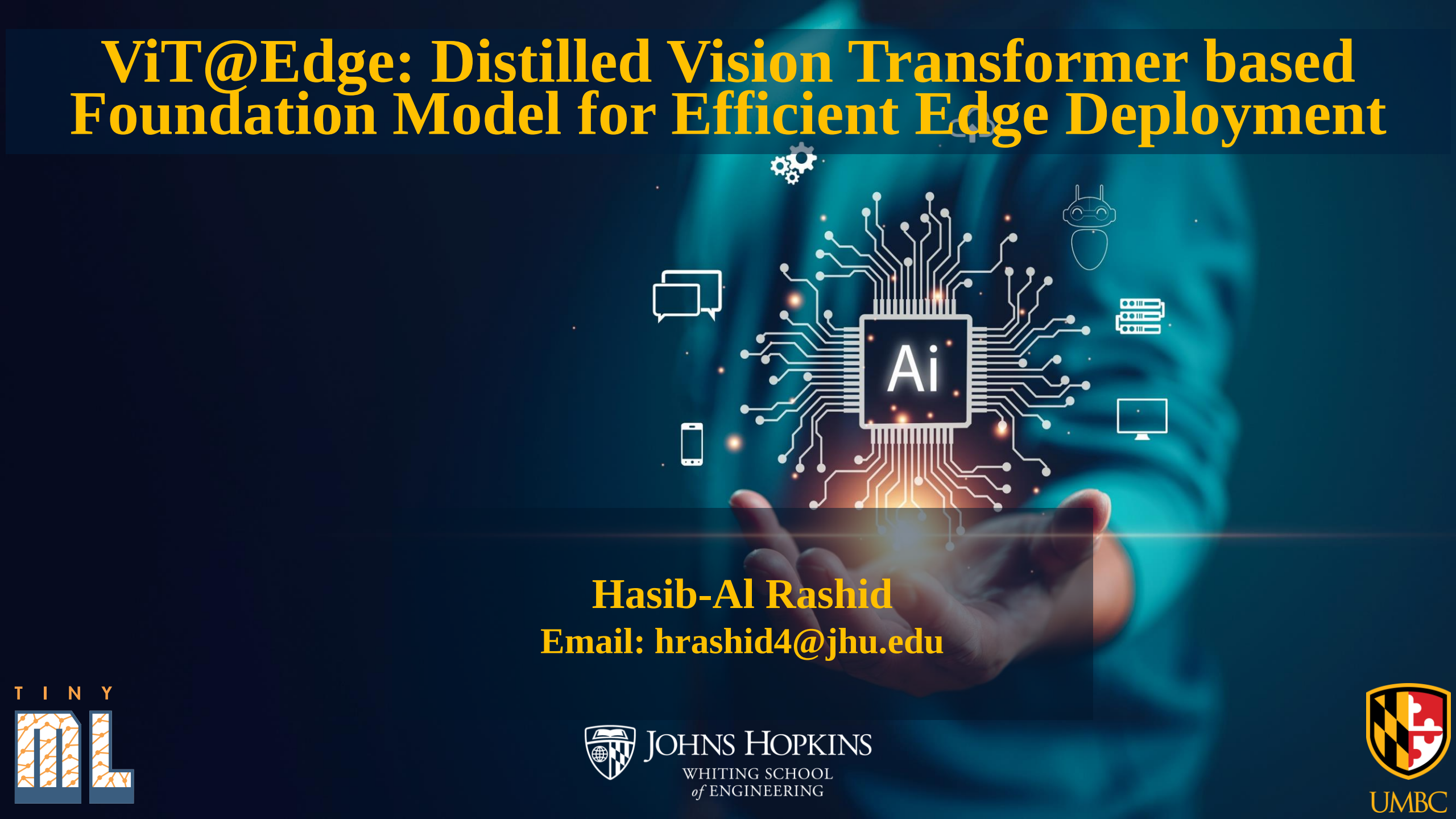
ViT@Edge: Distilled Vision Transformer based Foundation Model for Efficient Edge Deployment



Hasib-Al Rashid
Email: hrashid4@jhu.edu

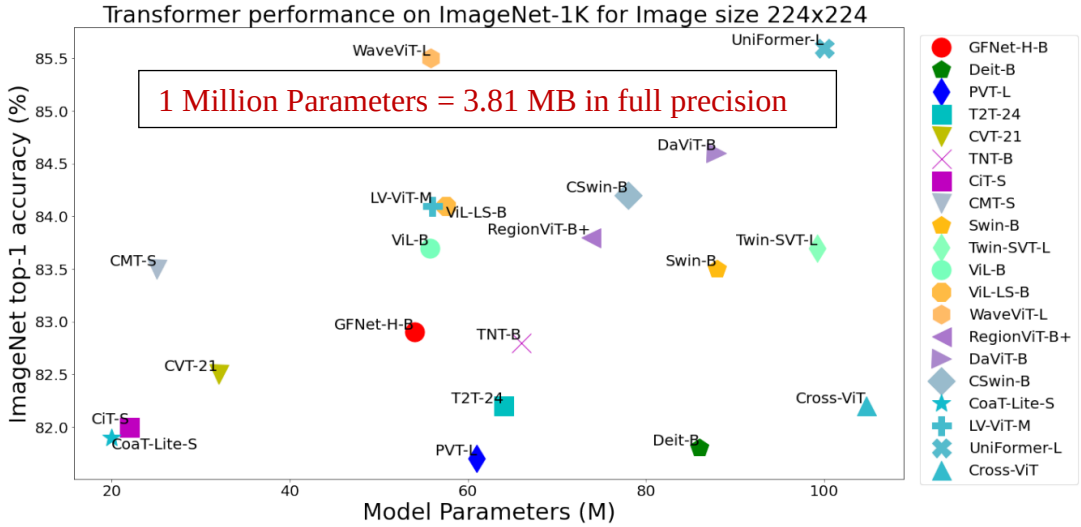
TINY ML

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

UMBC

Motivation

- Vision Transformers (ViTs) are now replacing CNN based foundation models in most of the computer vision tasks.
- Super Resolution is a classical computer vision problem, and it has been studied over decades.
- The aim of the problem is obtaining the high-resolution image from either single or multiple low-resolution images.
- However, none of these methods and many more in the literature can be directly applied and run in mobile devices either because of extremely large number of parameters.



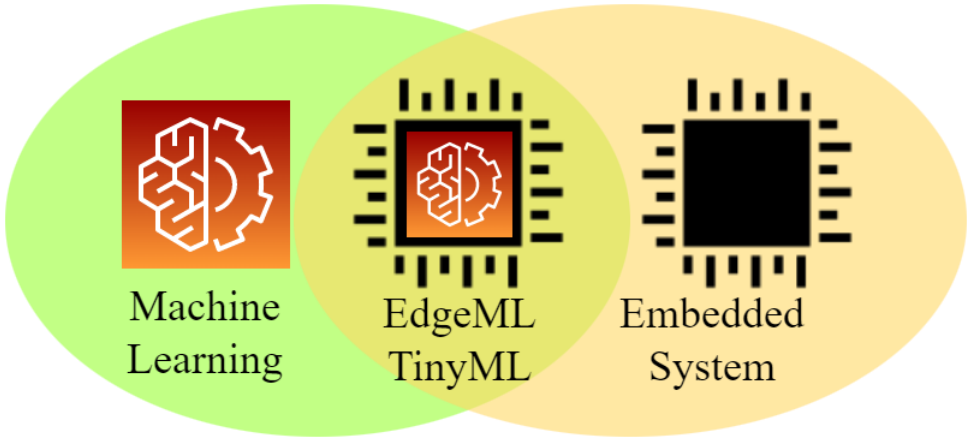
This figure shows the performance of various state-of-the-art vision transformer models across a number of parameters.

Source: Patro, Badri N., and Vijay Srinivas Agneeswaran.
"Efficiency 360: Efficient vision transformers." arXiv preprint arXiv:2302.08374 (2023).

A representation of the Super Resolved images from LapSRN models (Source: http://vllab.ucmerced.edu/wlai24/LapSRN/)

Motivation

- Deploying ViT-based foundation models for CV tasks on the edge devices are creating buzz.
- Edge Machine Learning and Tiny Machine Learning are transforming AI application on low-power devices, prioritizing efficiency.
- Our target is to deploy ViT-based Foundation Models at edge, bridging the gap between power and performance with our groundbreaking approaches.



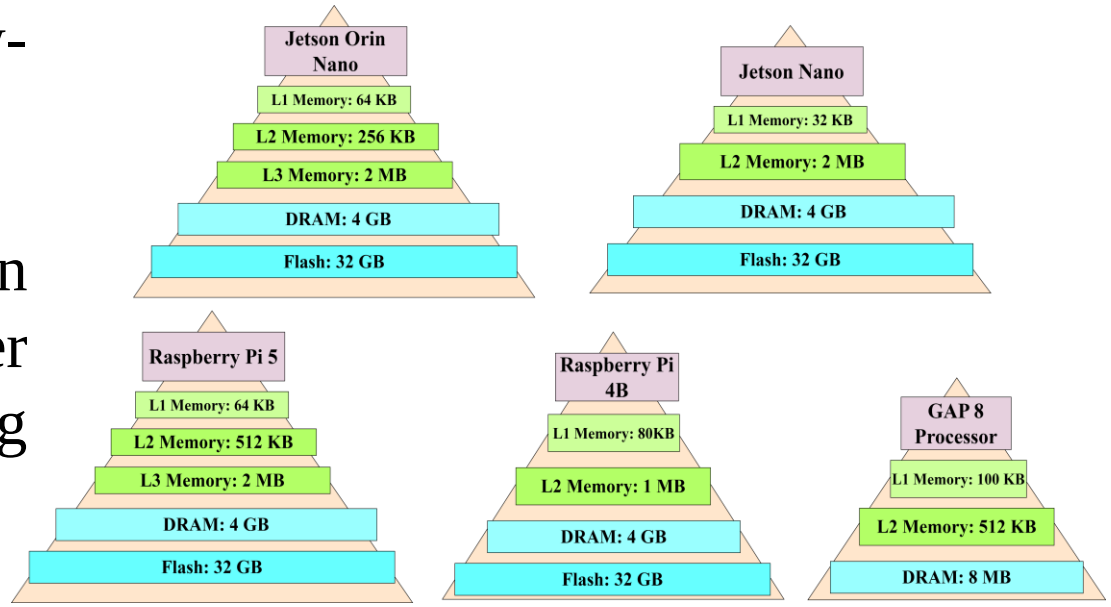
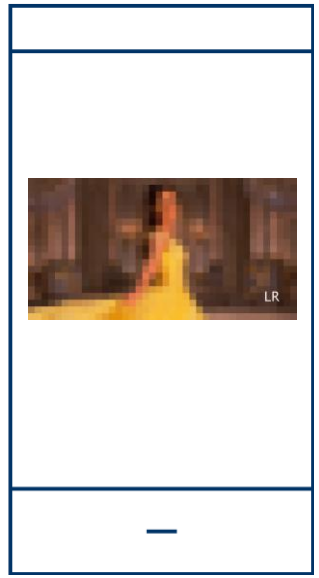


Fig: Memory Hierarchies of Different Off-the-Shelf Edge Devices

Problem Formulation

- Objective:
  - Develop a ViT-based solution for Image Super Resolution Task
  - Running on mobile/edge devices



Evaluation Metrics

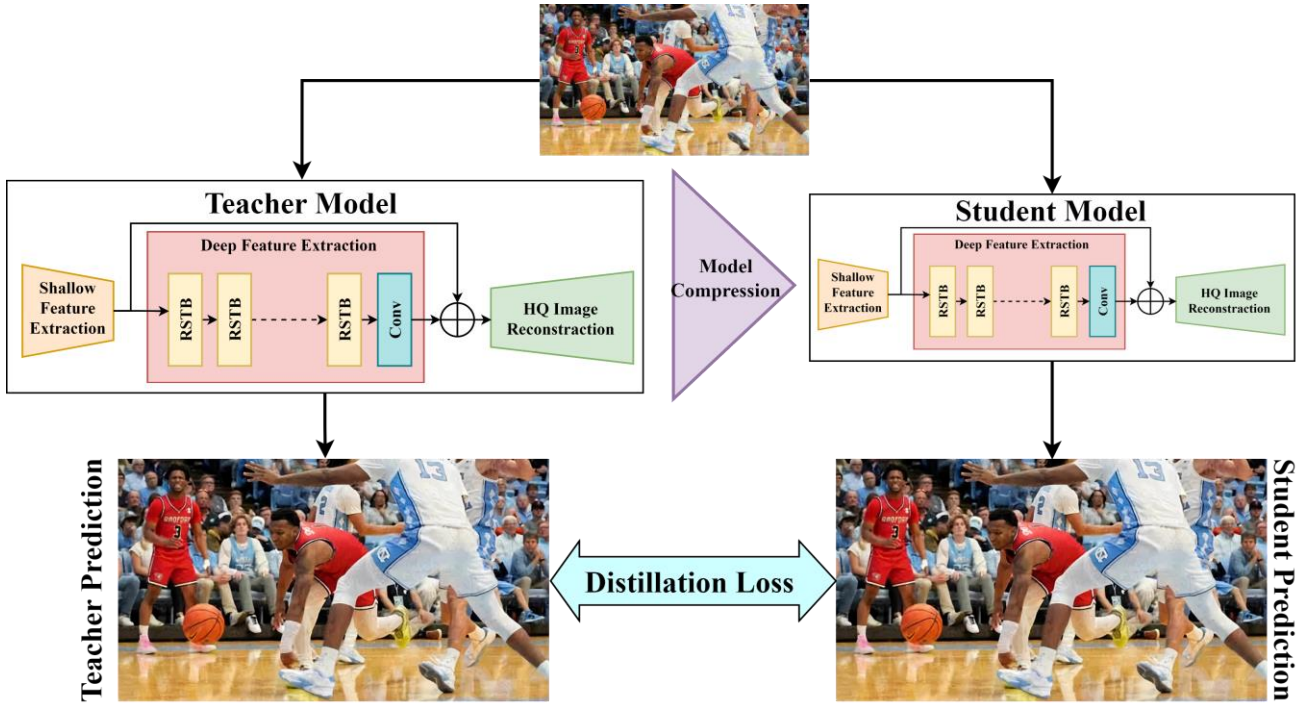
1. **PSNR:** Peak Signal to Noise Ratio is the most common technique used to determine the quality of results. It can be calculated directly from the MSE using the formula below, where L is the maximum pixel value possible (255 for an 8-bit image).

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(I(i) - \hat{I}(i))^2,$$

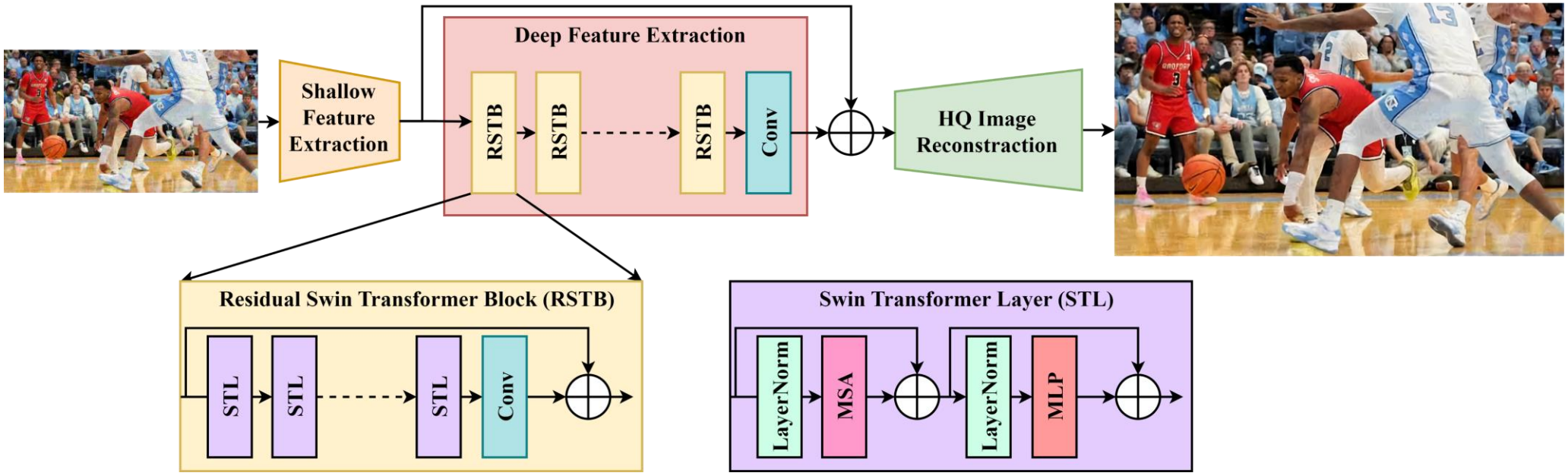
$$\mathrm{PSNR} = 10 \cdot \log_{10}(\frac{L^2}{\mathrm{MSE}}).$$

Proposed ViT@Edge

- **Contributions**:
  - ➢ Pretraining ViT-based model from scratch
  - ➢ Model Compression with
    - Quantization and
    - Pruning/ Sparsification
    - Knowledge Distillation

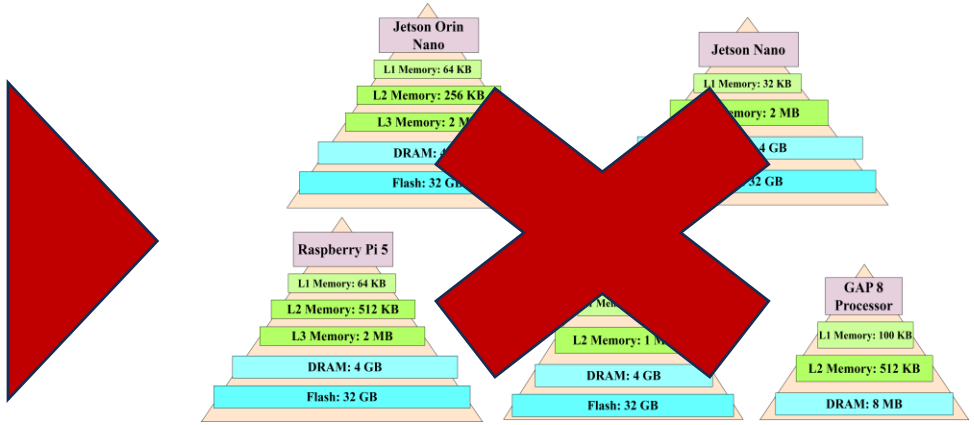


Pretraining ViT-based Model



SwinIR [1] based Pretrained Model

- PSNR: 38 dB
- Parameters: 11.8 M
- Memory: 47.2 MB

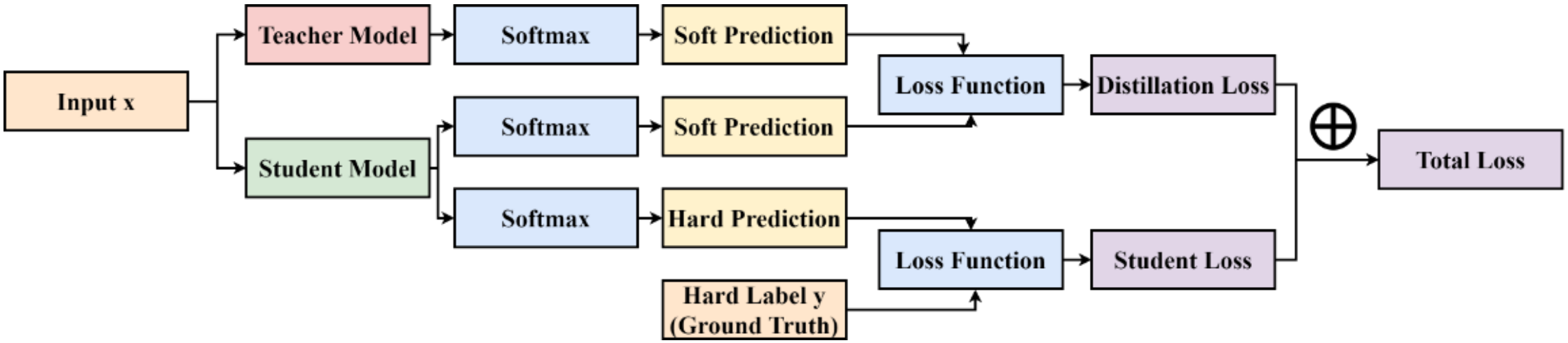


[1] Liang, Jingyun, et al. "Swinir: Image restoration using swin transformer." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

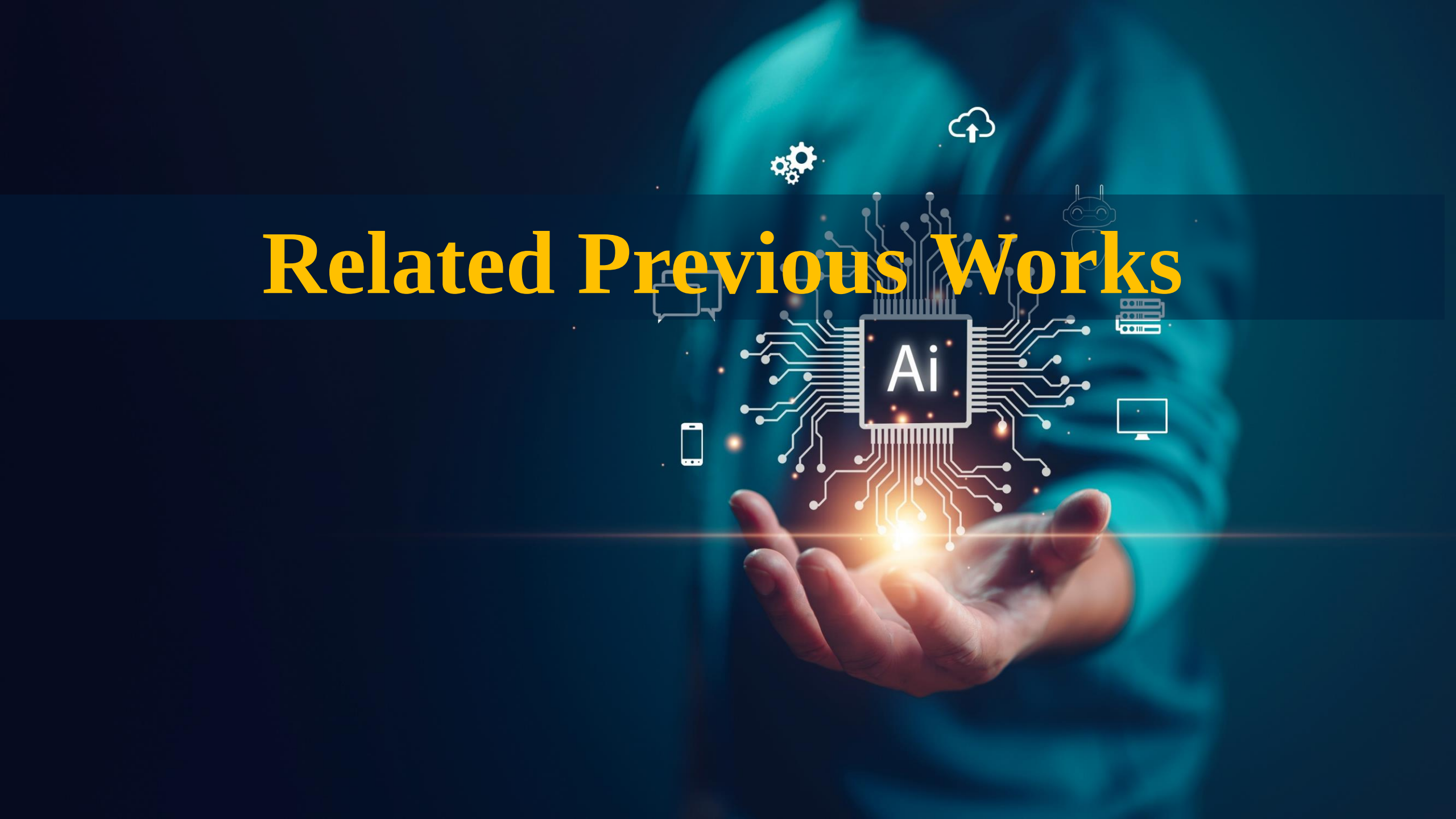
Model Compression Techniques

- Neural Network Model Compression and Computation Reduction
  - Quantization
  - Pruning
  - Structured Sparsity
  - Knowledge Distillation

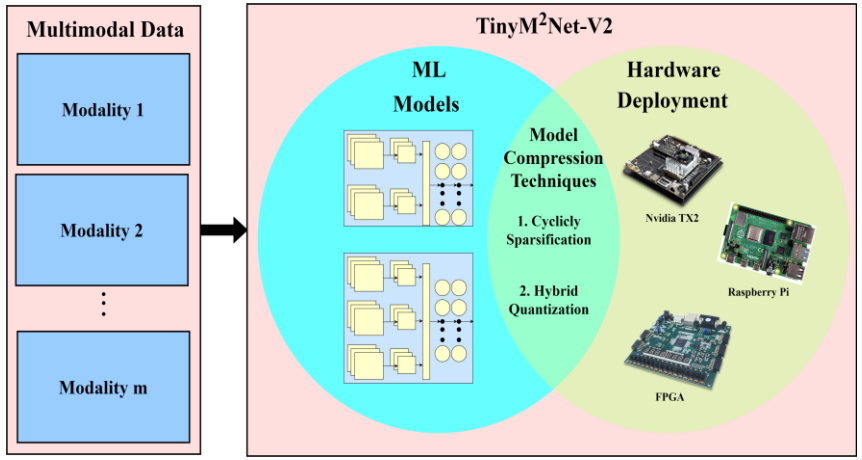
The main idea is to minimize the divergence between the teacher's and student's probability distributions.

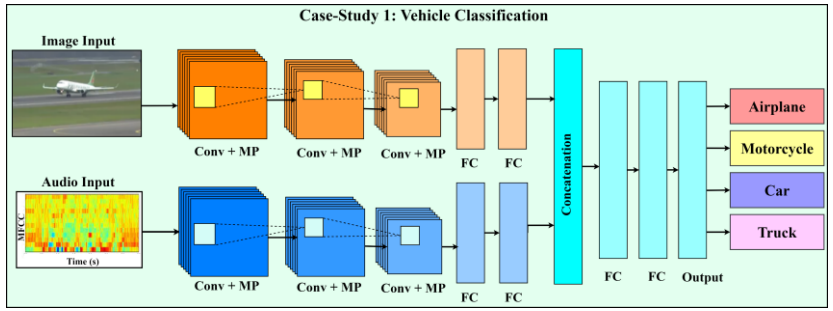


Related Previous Works

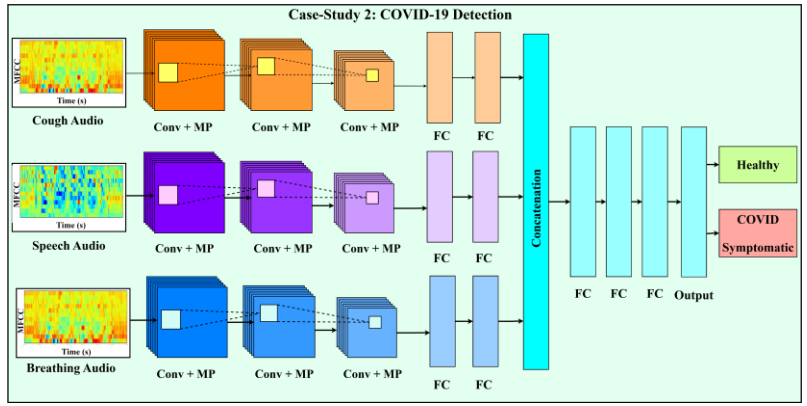


TinyM^{2}Net-V2: A Compact Low Power Software Hardware Architecture for Multimodal Deep Neural Networks (ACM TECS 2023)





Compression Techniques Used:
- Cyclic Sparsely Connected (CSC) Layers
- Hybrid (Mixed-Precision) Quantization



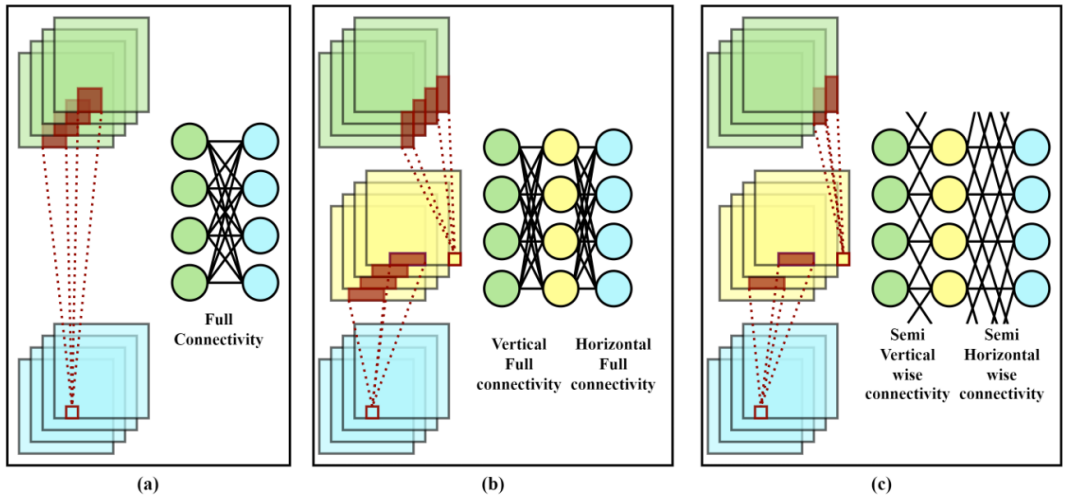
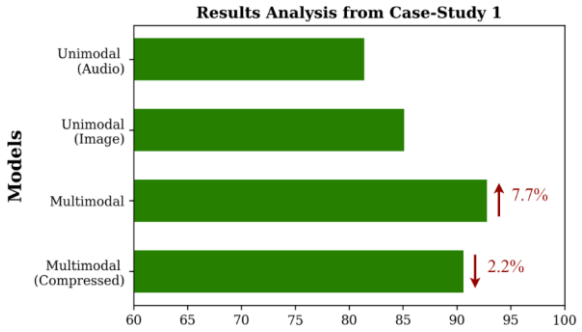
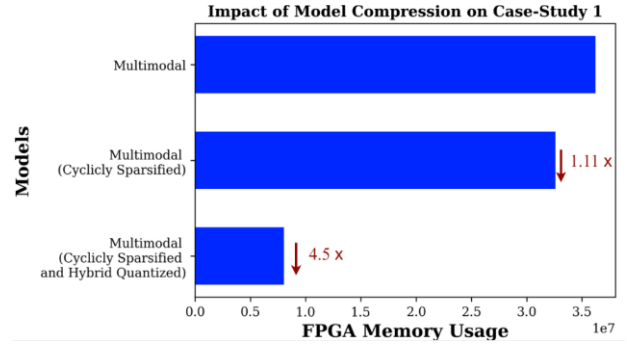
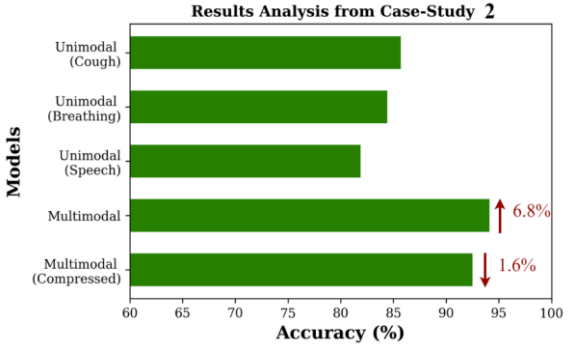
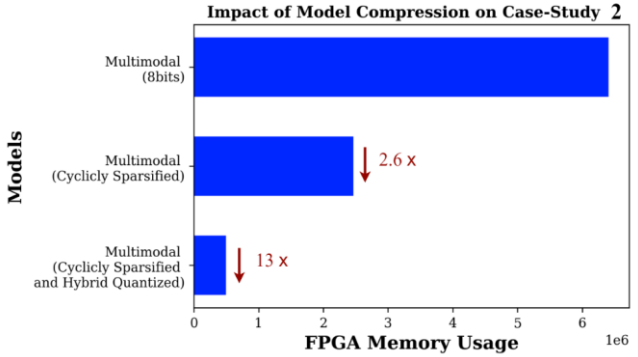


Figure: Key idea in compact schemes of CSC architecture illustrated with a typical multi-layer architecture and an equivalent structured graphs. (a) Baseline CNN with fully connectivity (b) Low-Rank Expansion (c) Cyclic Sparsely Connected (CSC) CNN

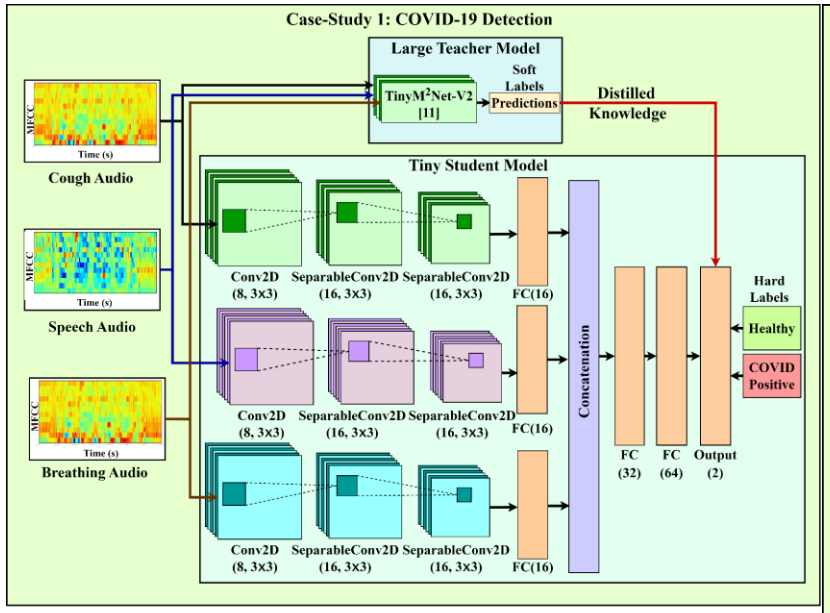




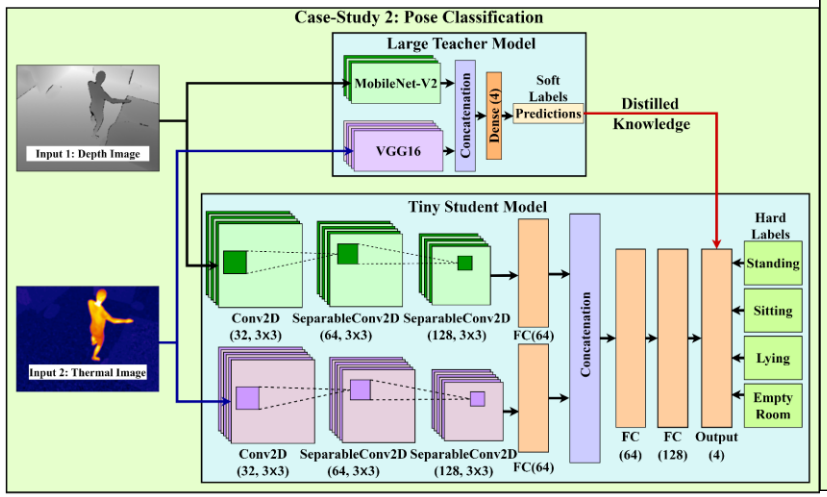




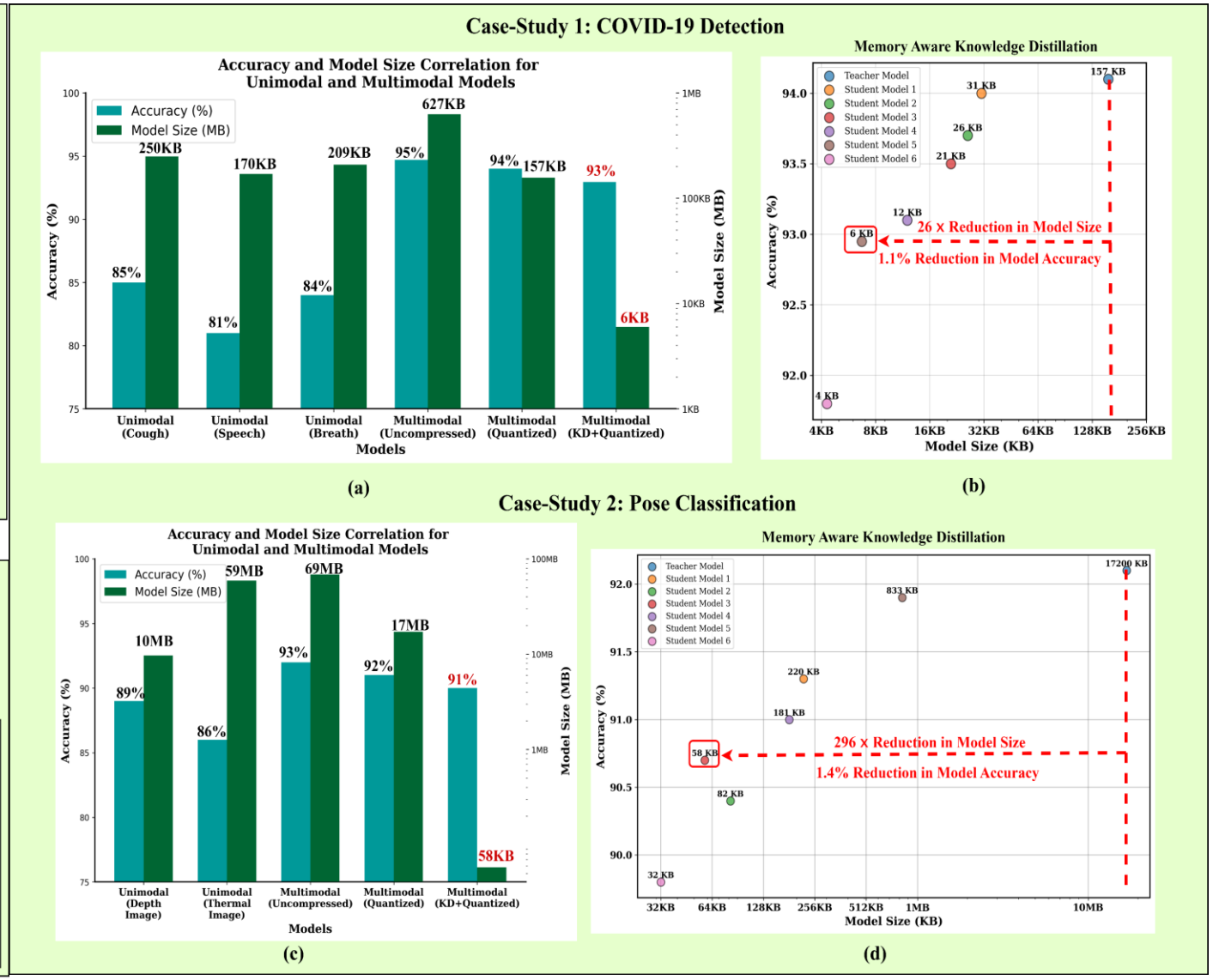
TinyM²Net-V3: Memory-Aware Compressed Multimodal Deep Neural Networks for Sustainable Edge Deployment (AAAI-SAI 2024)



(a)



(b)

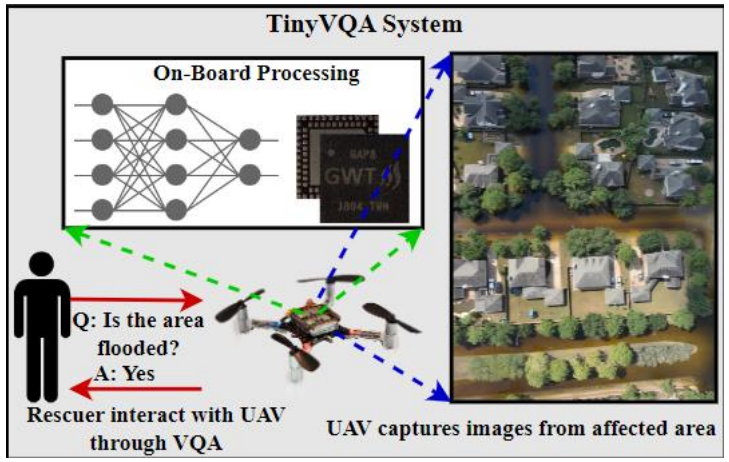


(a) (b) (c) (d)

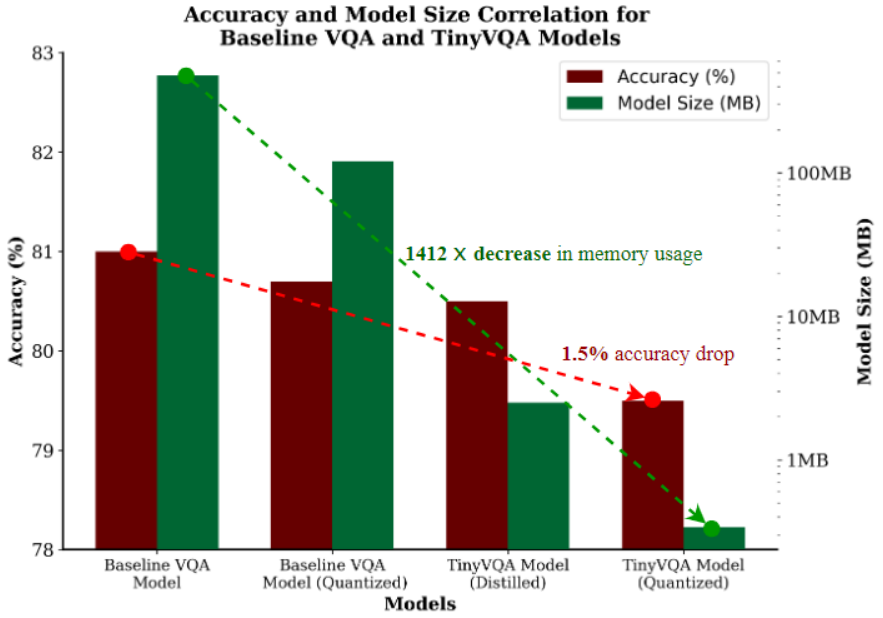
Compression Techniques Used:

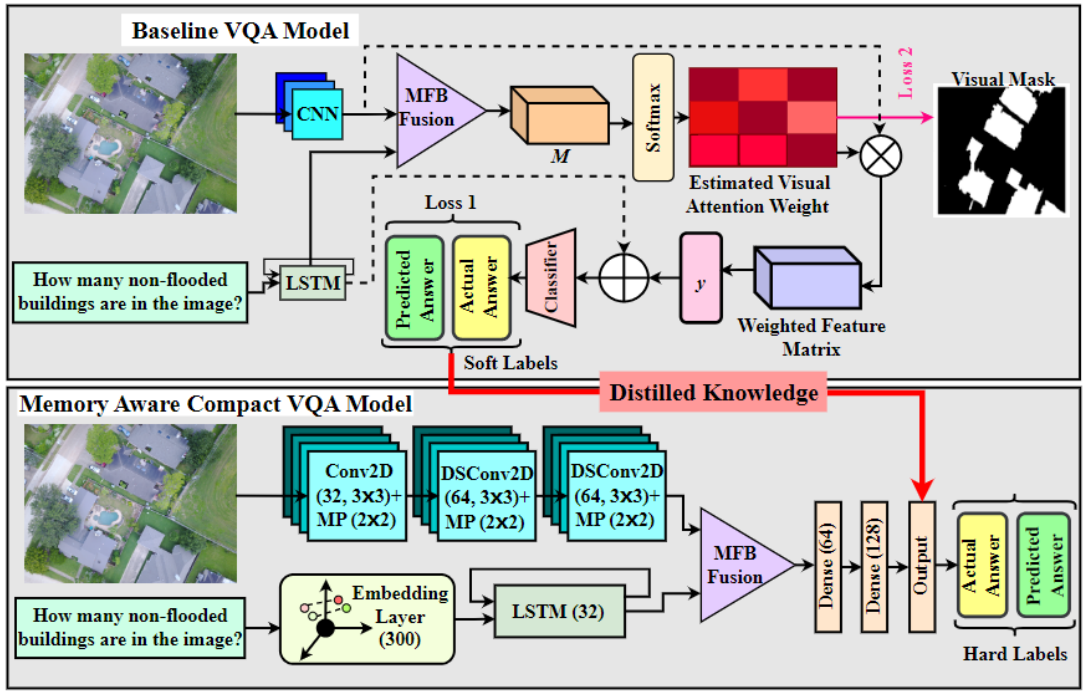
- Knowledge Distillation
- Quantization

TinyVQA: Compact Multimodal Deep Neural Network for Visual Question Answering on Resource-Constrained Hardware (tinyML 2024, accepted)



Compression Techniques Used:
- Knowledge Distillation
- Quantization





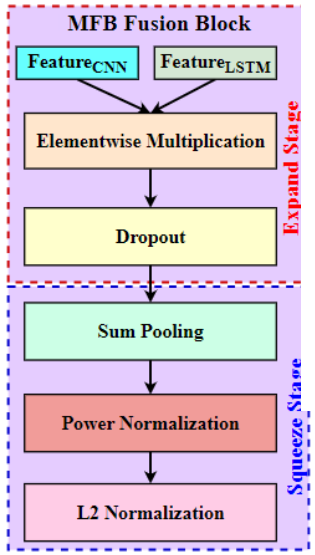


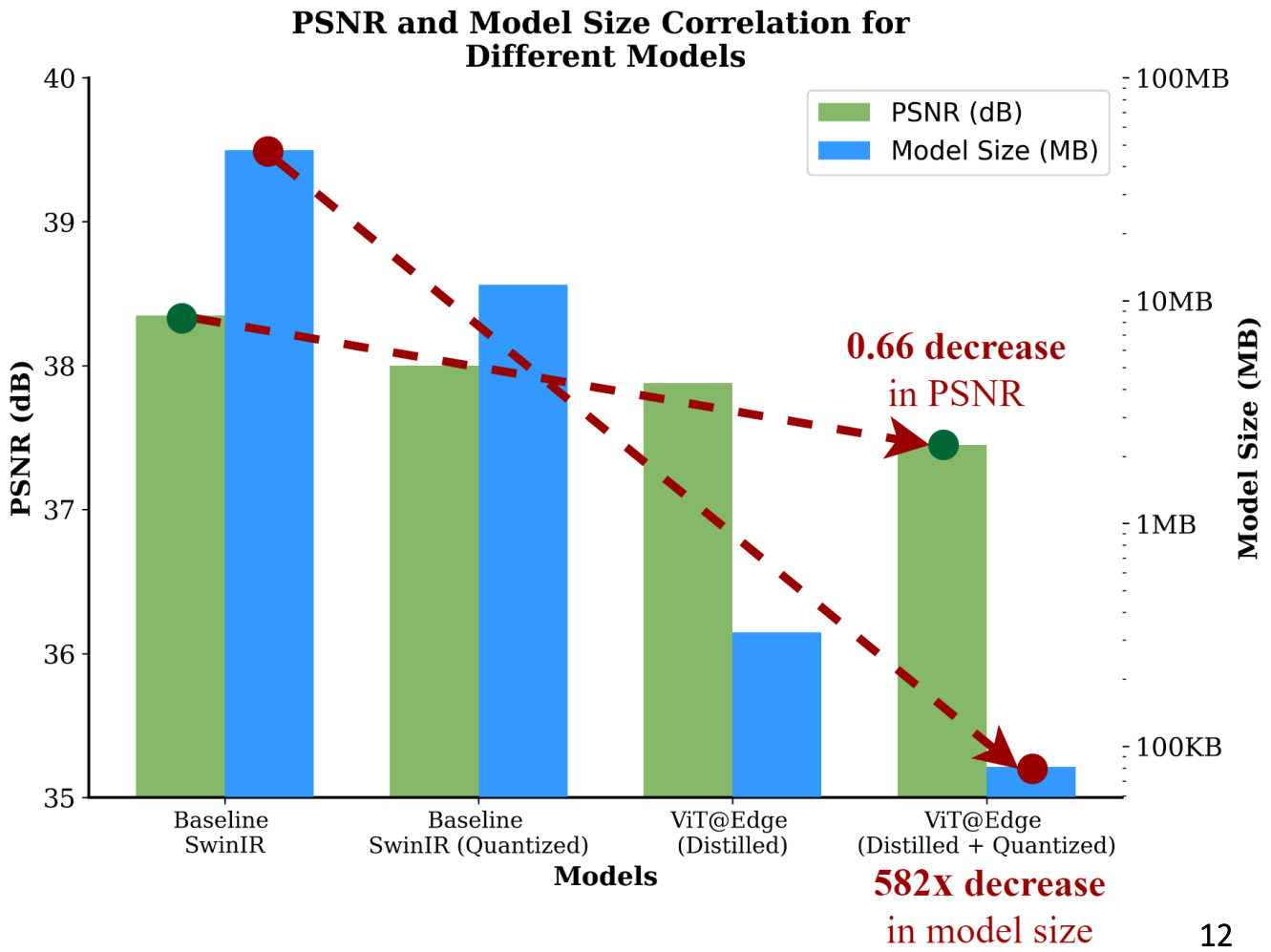
Figure 3: Accuracy and Model Size Correlation for Baseline VQA and TinyVQA for FloodNet [19] dataset. Baseline model achieved 81% accuracy with 479 MB model size whereas final TinyVQA model achieved 79.5% accuracy with 339 KB model size.

Expected ViT@Edge Results

Dataset

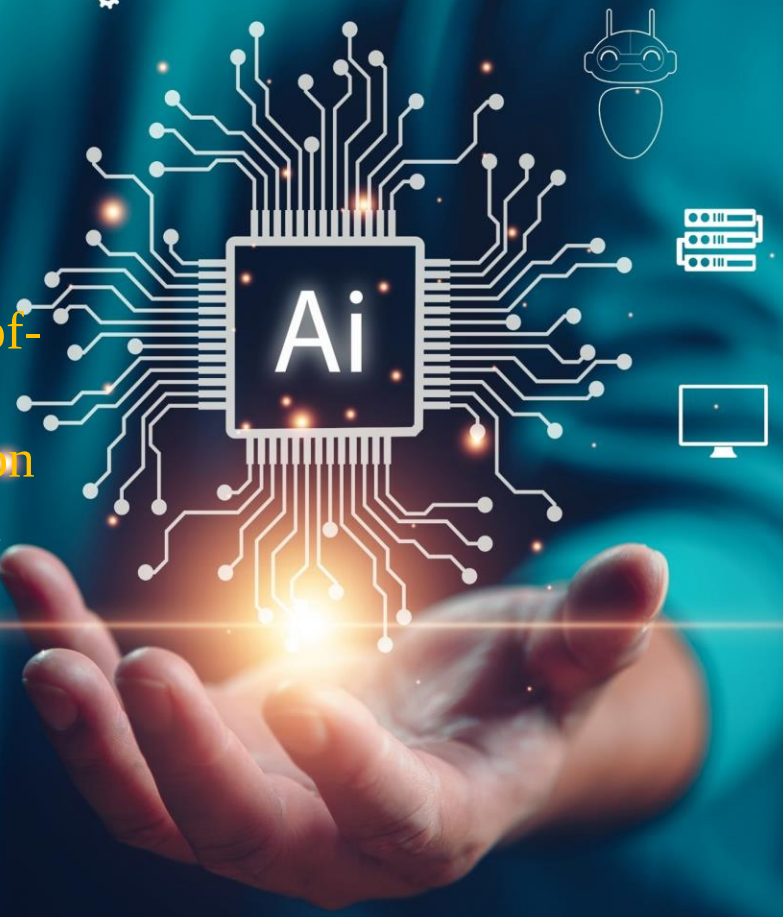
- The DIV2K dataset is a collection of high-quality images specifically designed for the development and evaluation of image processing algorithms
- The images in the DIV2K dataset have a resolution of approximately 2K (2048 pixels on the longer side).
- Training set with 800 images, a validation set with 100 images, and a testing set with 100 images.

Compression Techniques Used in ViT@Edge:
- Knowledge Distillation
- Low-bit Quantization



Conclusion

- ViT@Edge develops compressed computer vision models for edge deployment
- Knowledge distillation and Quantization is used to compress the model
- Further compression is possible if we can use state-of-the-art sparsification techniques
- For our experiments, we got 582 X memory reduction at the cost of very minimal performance degradation

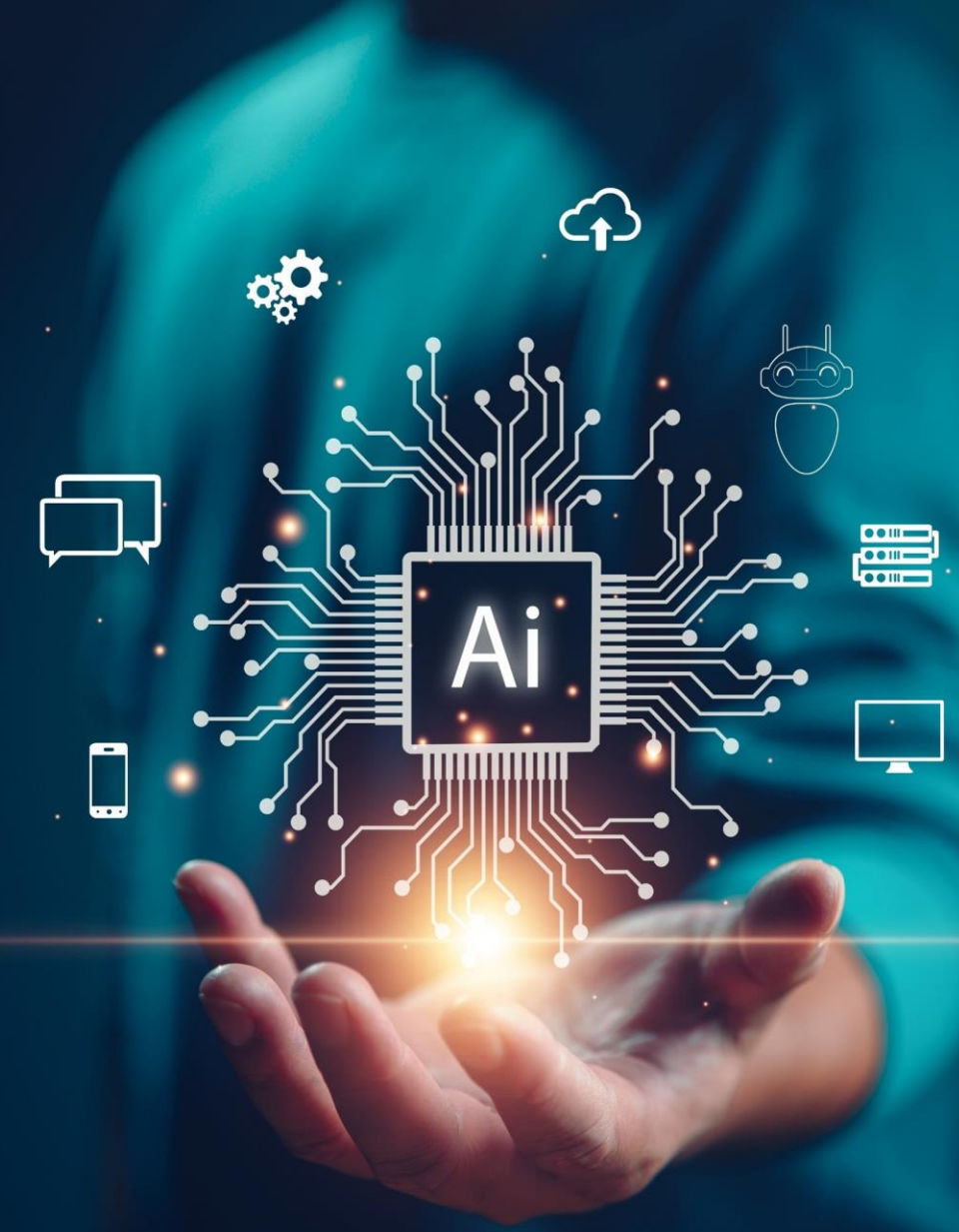


TINY ML

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

UMBC

thank you!



Contact: hrashid4@jhu.edu

TINY ML

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

UMBC