COMPRESSING HEAVY VISION TRANSFORMERS FOR EDGE DEVICES

MSc. Computer Vision

July, 2024

Presented by:

Atanda Abdullahi Adewale

Supervised by:

Walid Brahim

CONTENTS

- **Introduction**
  - Wasoria
  - Problem Statement
  - Why Vision Transformers
    - MHSA
  - Segment Anything Model (SAM)
- **Compression Techniques**
  - Model Prunning and Quantization
  - Knowgdge Distillation
- **Methods**
  - Dataset Design and Preprocessing
  - Teacher Model
    - MobileSAM (TinyVit)
  - Student Models
    - ResNet-18
    - EfficientNet-b0
    - MobileViT-s
  - Model Design
    - Huber Loss
    - Predictor
    - Training Configuration
- **Results**
  - Performance Metrics (mIOU, Params, Flpos)
  - Visualizer and Demo on the Edge
- **Conclusion**
- **References**

WASORIA, FRANCE

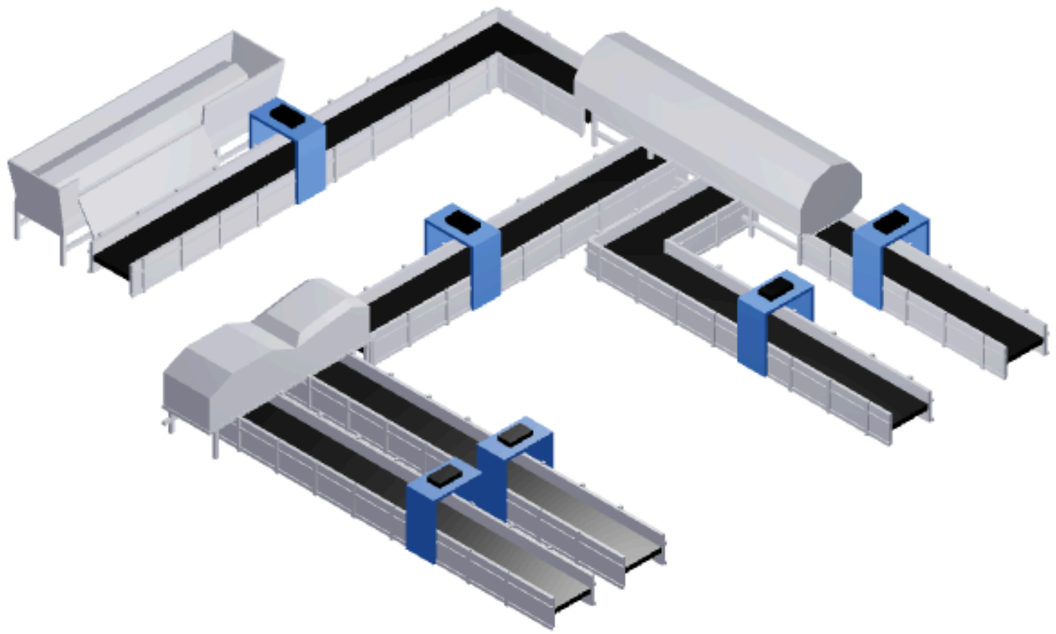
Overview

- Located in Le Creusot, France.
- Waste management and recycling through AI integration.
- Collaborates with local research labs for continuous innovation

Innovative Solutions

- Develops AI Box (CARAC) with controlled light intensity.
- Mounted on conveyor belts for real-time waste sorting
- Instant alerts for sorting issues and quality data gathering.

WASORIA, FRANCE



Overview

- Located in Le Creusot, France.
- Waste management and recycling through AI integration.
- Collaborates with local research labs for continuous innovation

Innovative Solutions

- Develops CARAC AI Box with controlled light intensity.
- Mounted on conveyor belts for real-time waste sorting
- Instant alerts for sorting issues and quality data gathering.

Problem

(a) Texture image

81.4% **Indian elephant**
10.3% indri
8.2% black swan

(b) Content image

71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

(c) Texture-shape cue conflict

63.9% **Indian elephant**
26.4% indri
9.6% black swan

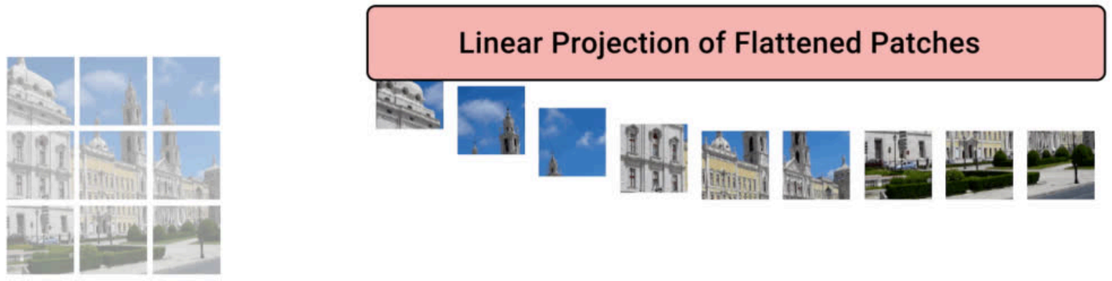
Problem:

- Current methods struggle with distinguishing objects of similar textures.
- Convolutional Neural Networks (CNNs) exhibit texture bias, leading to misclassification.
- Vision Transformers (ViTs) are too heavy for edge devices in low-latency sorting.

Research Objectives:

- Develop ViT/CNN detectors for accurate waste detection.
- Integrate encoder into compressed SAM.
- Distill knowledge from SAM to smaller model.
- Optimize model for edge deployment.

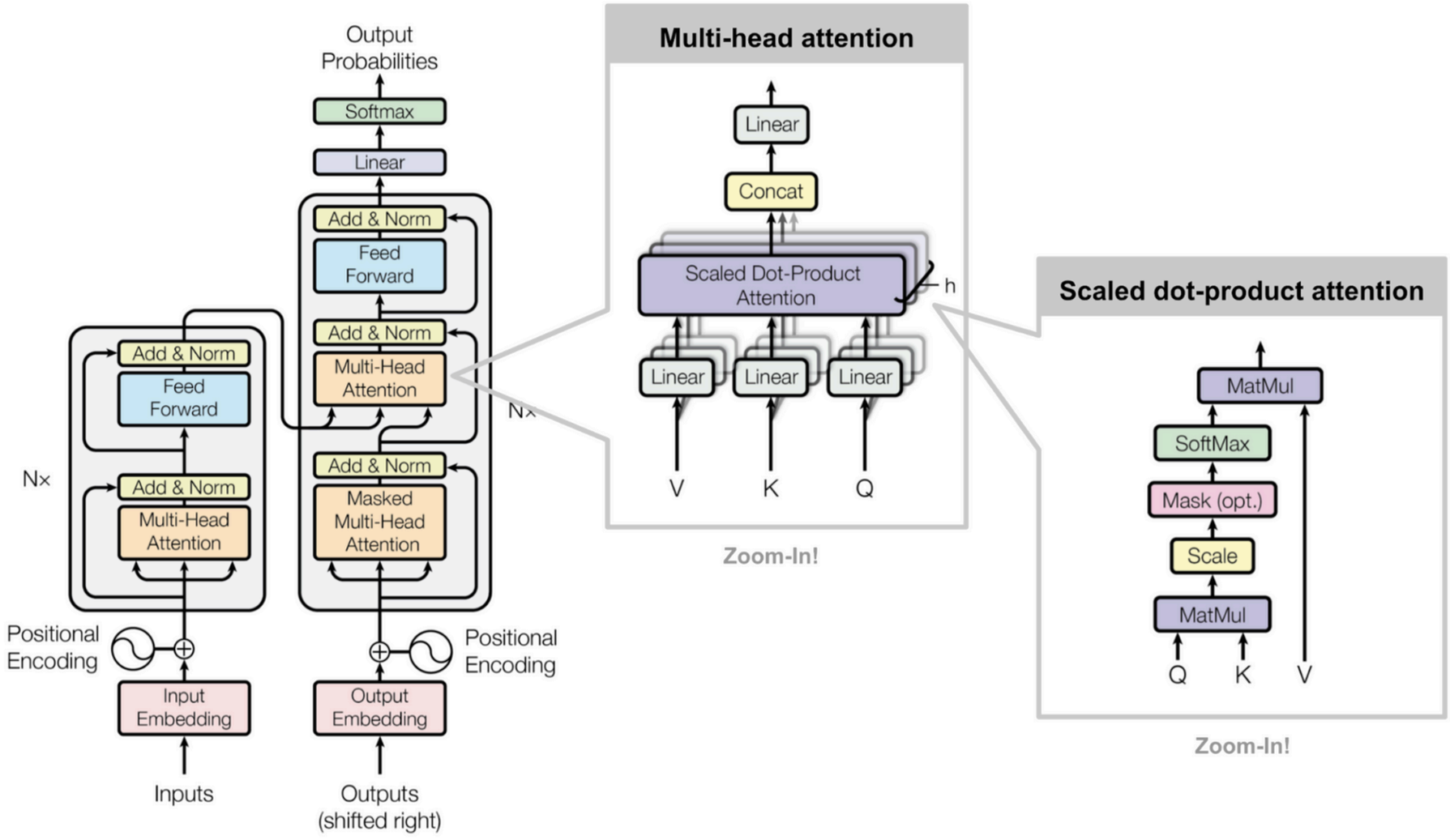
Vision Transformers (ViT)



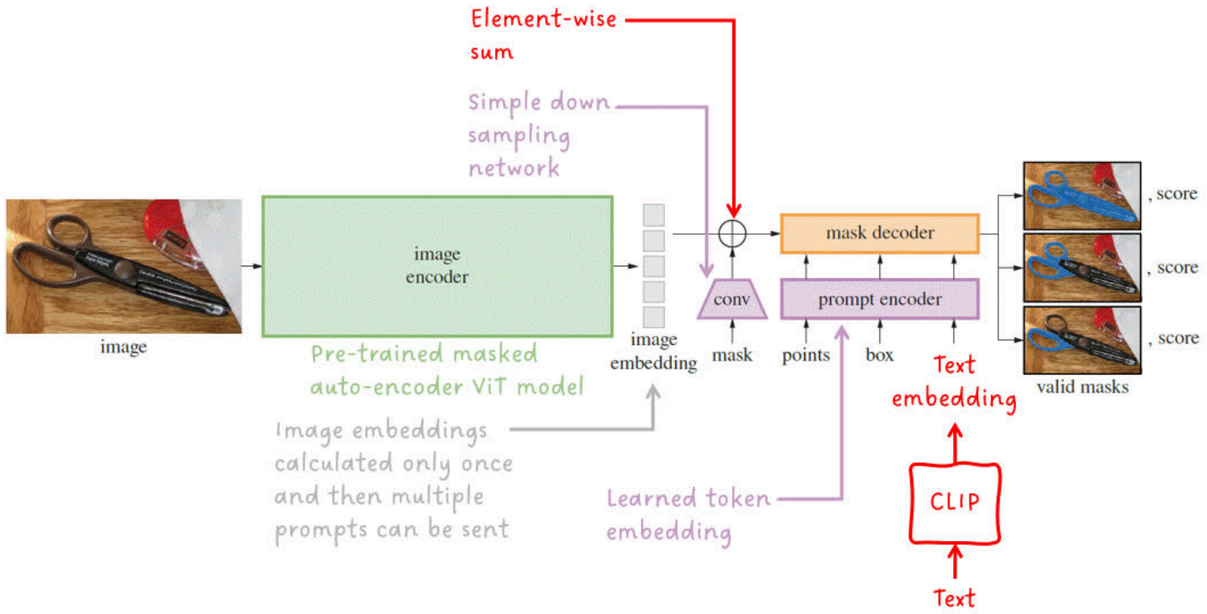
Vision Transformer ViT Architecture

- Split an image into fixed-size patches.
- Flatten the image patches.
- Create lower-dimensional linear embeddings from the patches.
- Include positional embeddings.
- Feed the sequence as input to a state-of-the-art (SOTA) transformer encoder.
- Pre-train the ViT model with image labels on a large dataset.
- Fine-tune on a downstream dataset for image classification

Multi-head Self Attention

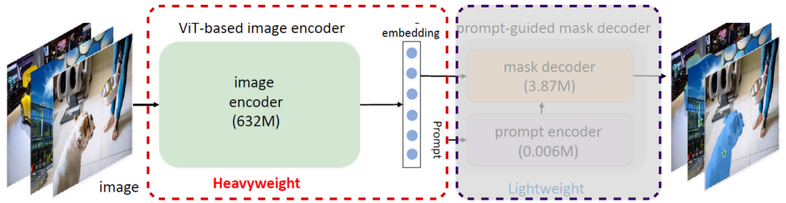


Segment Anything Model (SAM)



Decoupled approach:

We want to know a lot about a little

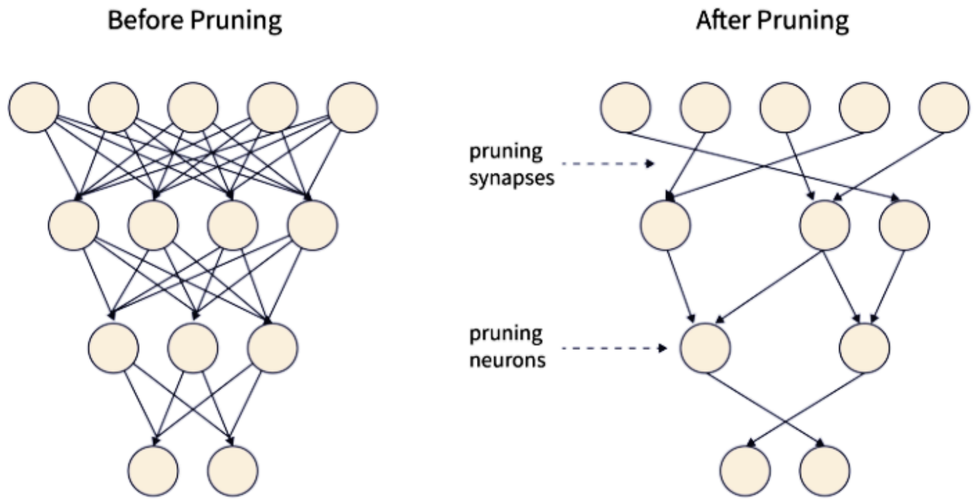


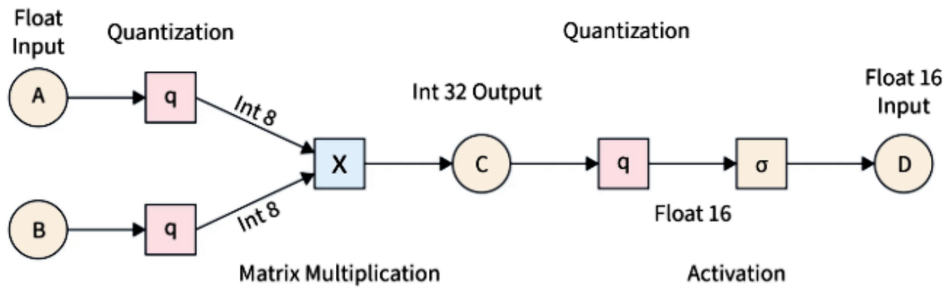
Vision foundation model for semantic image segmentation

SA1B dataset : 1.1 billion masks from approximately 11 million images

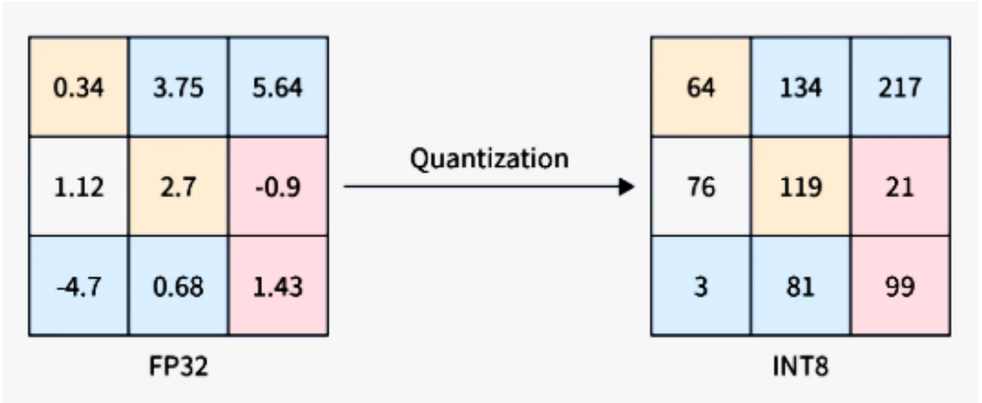
SAM knows a lot about a lot

Model Prunning and Quantization

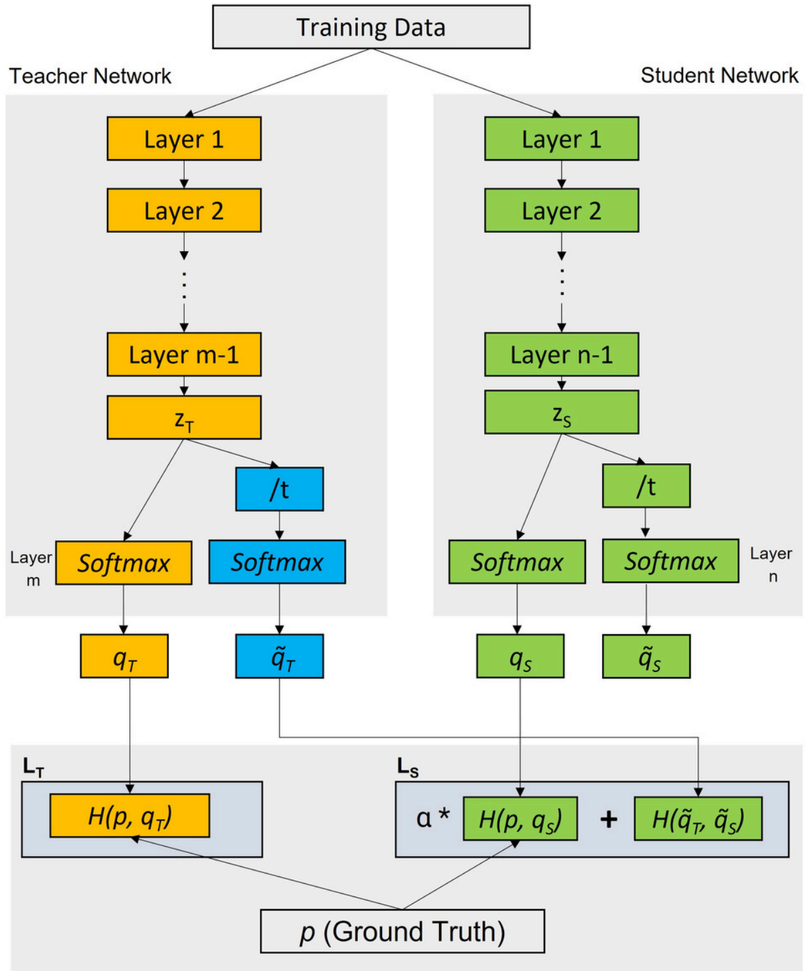


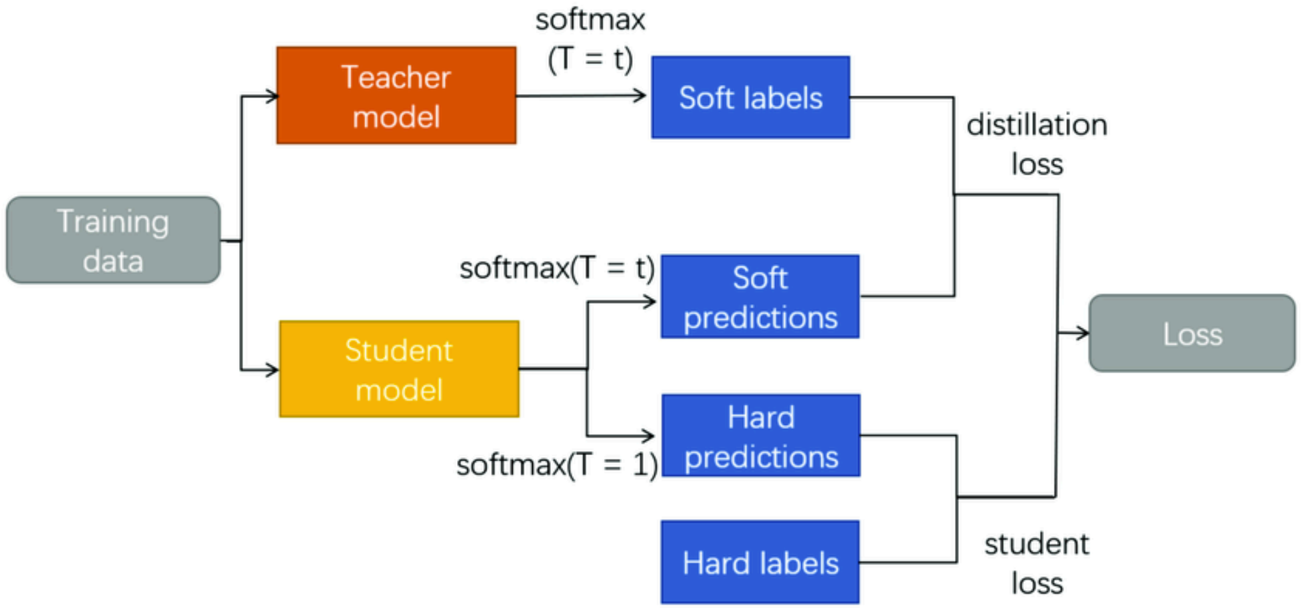


Method	Description	Benefits	Trade-offs
Post-Training Quantization	Lower precision after training	Simple, immediate reduction	Accuracy drop, esp. small models
Quantization-Aware Training	Train with quantization	Better accuracy	Complex, longer training
Unstructured Pruning	Removes individual weights	Large size reduction	Irregular memory access
Structured Pruning	Removes neurons, filters, heads	Hardware-friendly	Higher accuracy loss

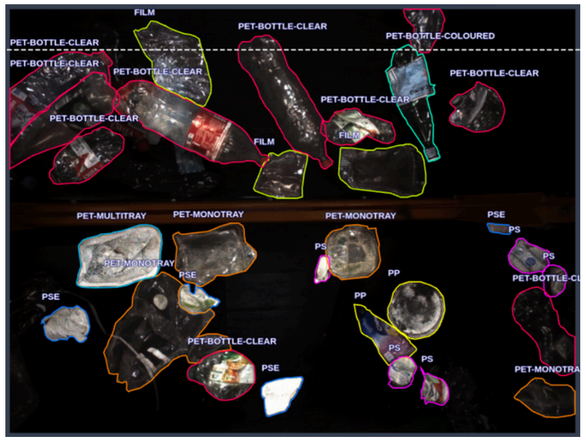


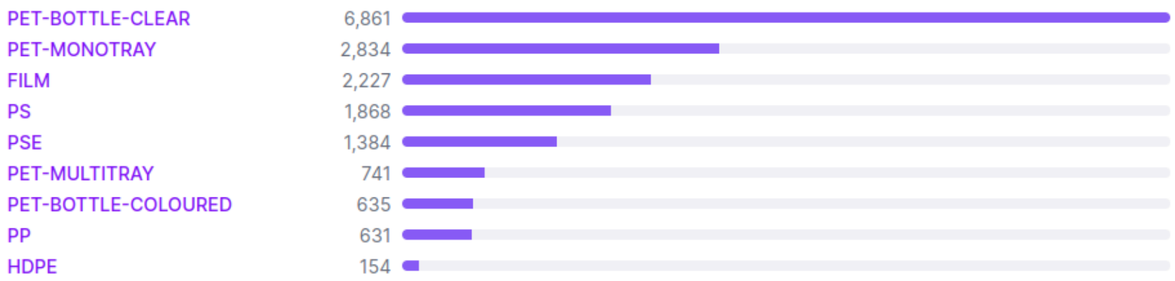
Knowledge Distillation





WASORIA 5K Dataset





- Obtained from a waste sorting center in France; designed for instance segmentation tasks.
- COCO Format: (image metadata, categories, bounding boxes, segmentations).
- 5080 images with 18,810 annotated instances across eight waste material classes.
- Data Split: Training set: 4523 images; validation set: 558 images.
- Annotated with Roboflow and Annolid, with detailed polygon annotations.
- Images auto-oriented and resized to 640x640 pixels.
- Augmentations: horizontal flipping, cropping, and rotation.

MobileSAM

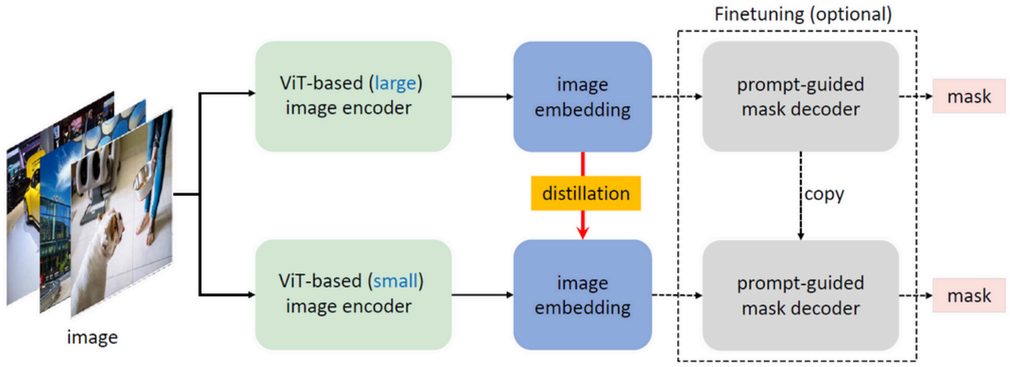
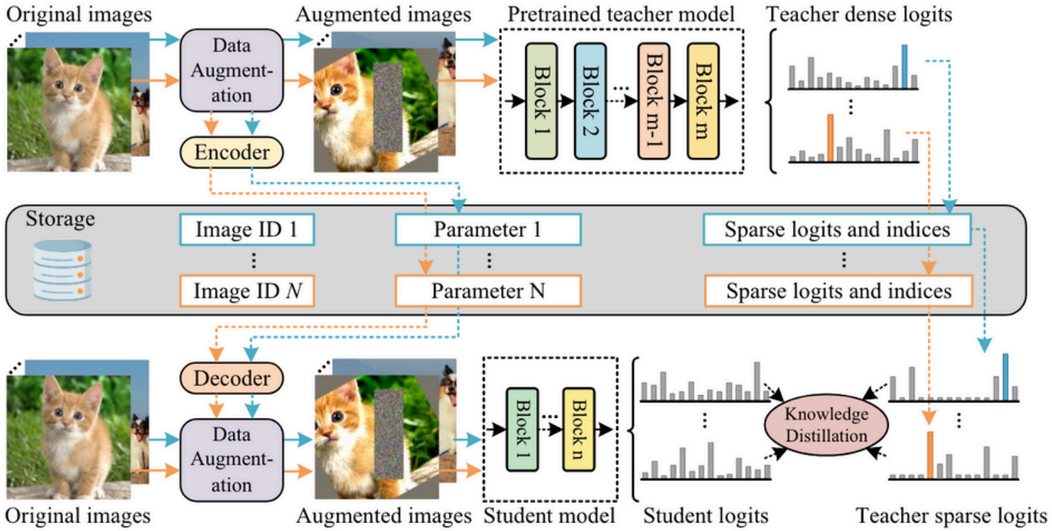


Figure 3: Decoupled distillation for SAM.

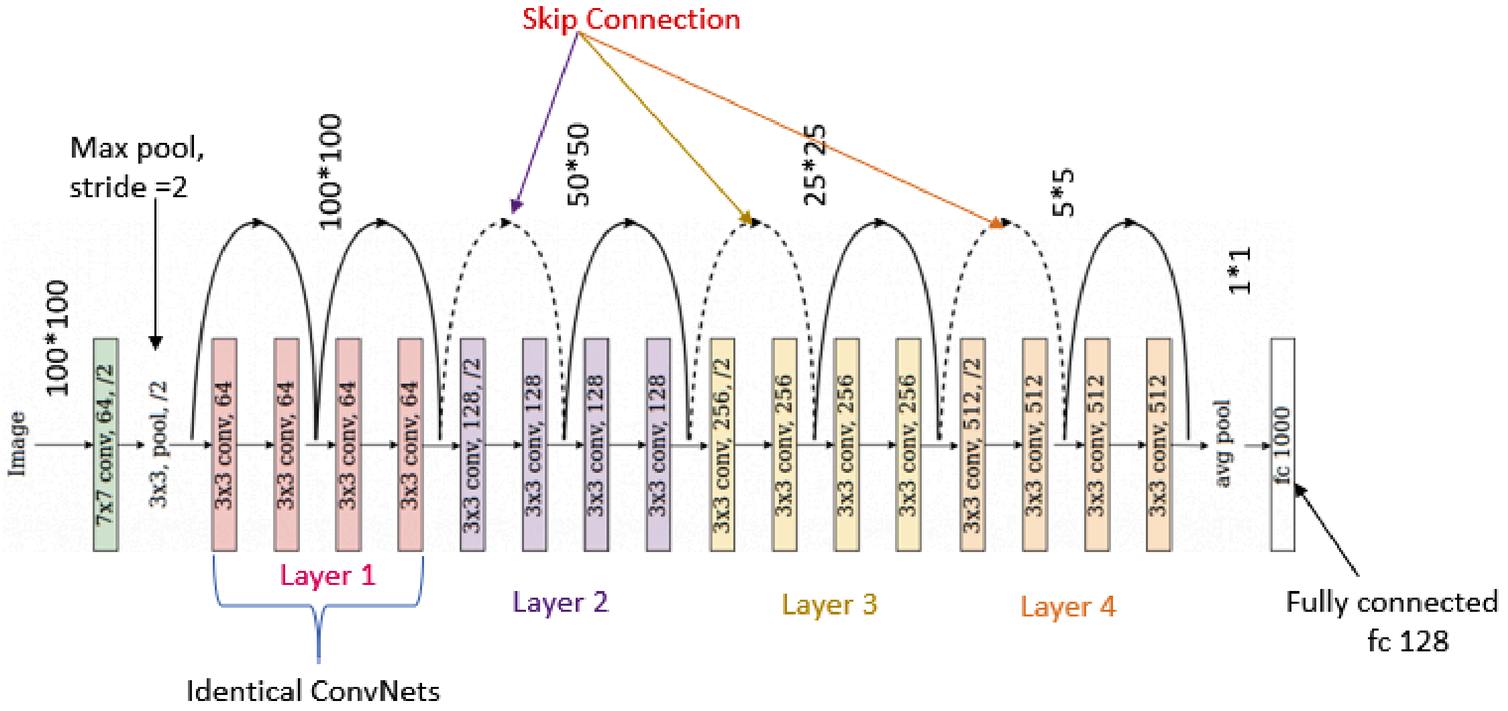
- Replace the heavy (ViT-H) image encoder in SAM with a lightweight TinyViT encoder
- Maintains the same processing pipeline as the original SAM
- Total parameter count of 9.66 million
- Inference speed is 60 times faster than SAM

TinyViT

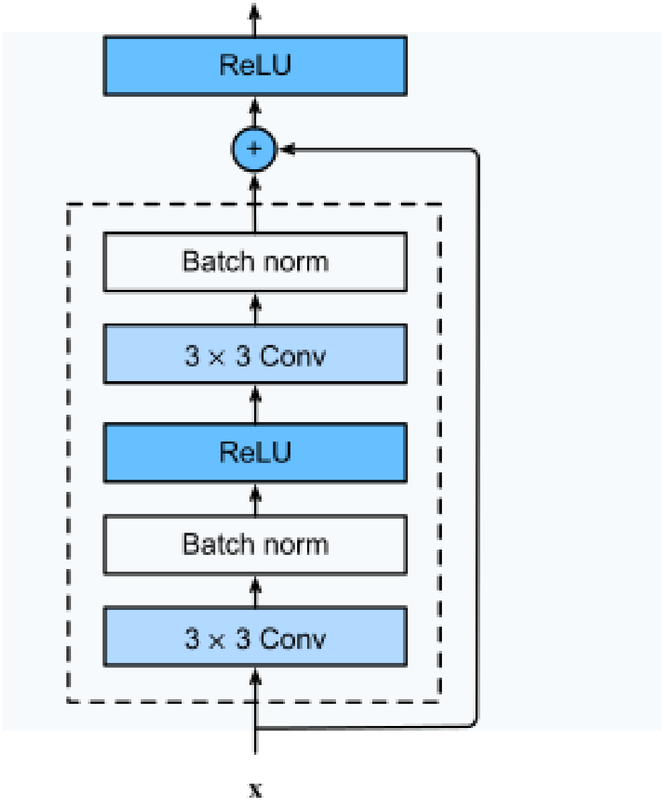


- Efficient training with precomputed data and soft labels.
- Stores sparse soft labels.
- Encodes augmentation parameters into a single parameter.
- Uses MBConvs in early stages, transformer blocks with window attention in later stages.
- Includes attention biases, depthwise convolutions, residual connections, and GELU activation functions.

ResNet-18

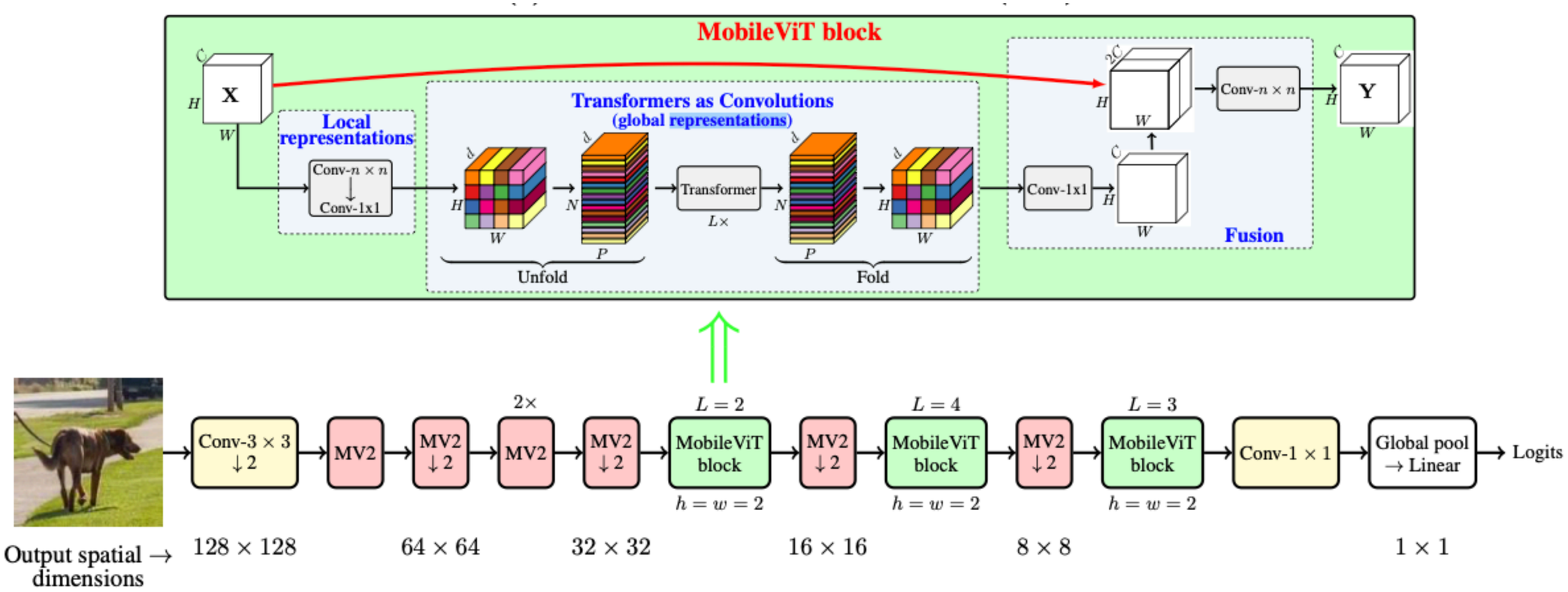


ResNet-18 Architecture



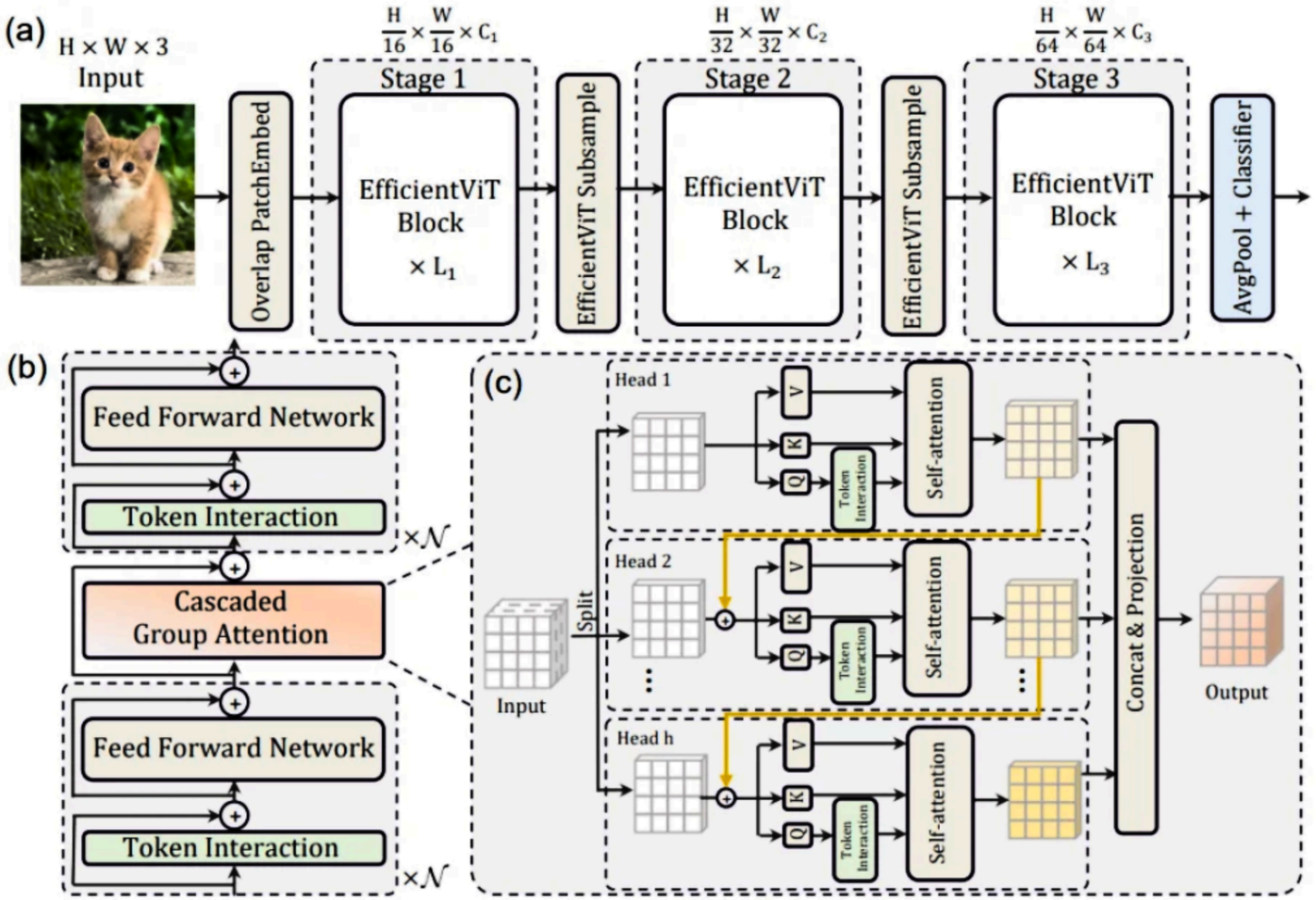
Fruit 360 Input Image size= 100*100 px

MobileViT-s



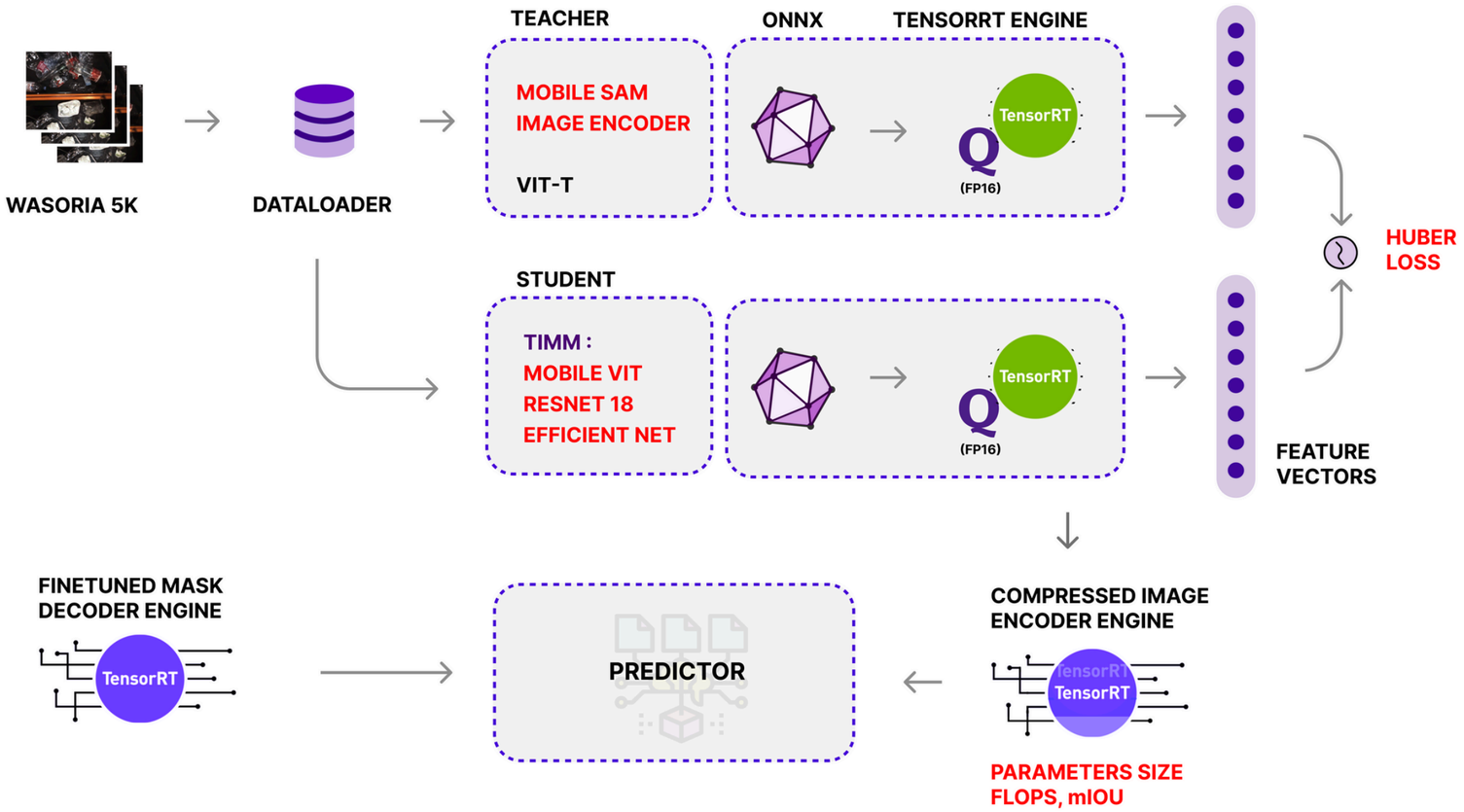
- input X feature maps, unfolded into non-overlapping patches of size P.for transformers, fold patches back to their original spatial dimensions
- X**Y=Conv-1x1(X+G)** :) The addition X+G combines the original and global features before applying the final 1x1 convolution.

EfficientVit



- Implements group cascade attention for efficient spatial dependency capture.
- Utilizes a sandwich layout with interleaved attention and convolution layers.
- Offers various size variants for different deployment scenarios.
- Optimizes local and global feature extraction for improved performance

Framework outline



- NVIDIA TensorRT optimizes and accelerates deep learning inference on NVIDIA GPUs for reduced latency and increased throughput.
- ONNX provides an open-source format for model interoperability between different deep learning frameworks.

Huber Loss

Mean Square Error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

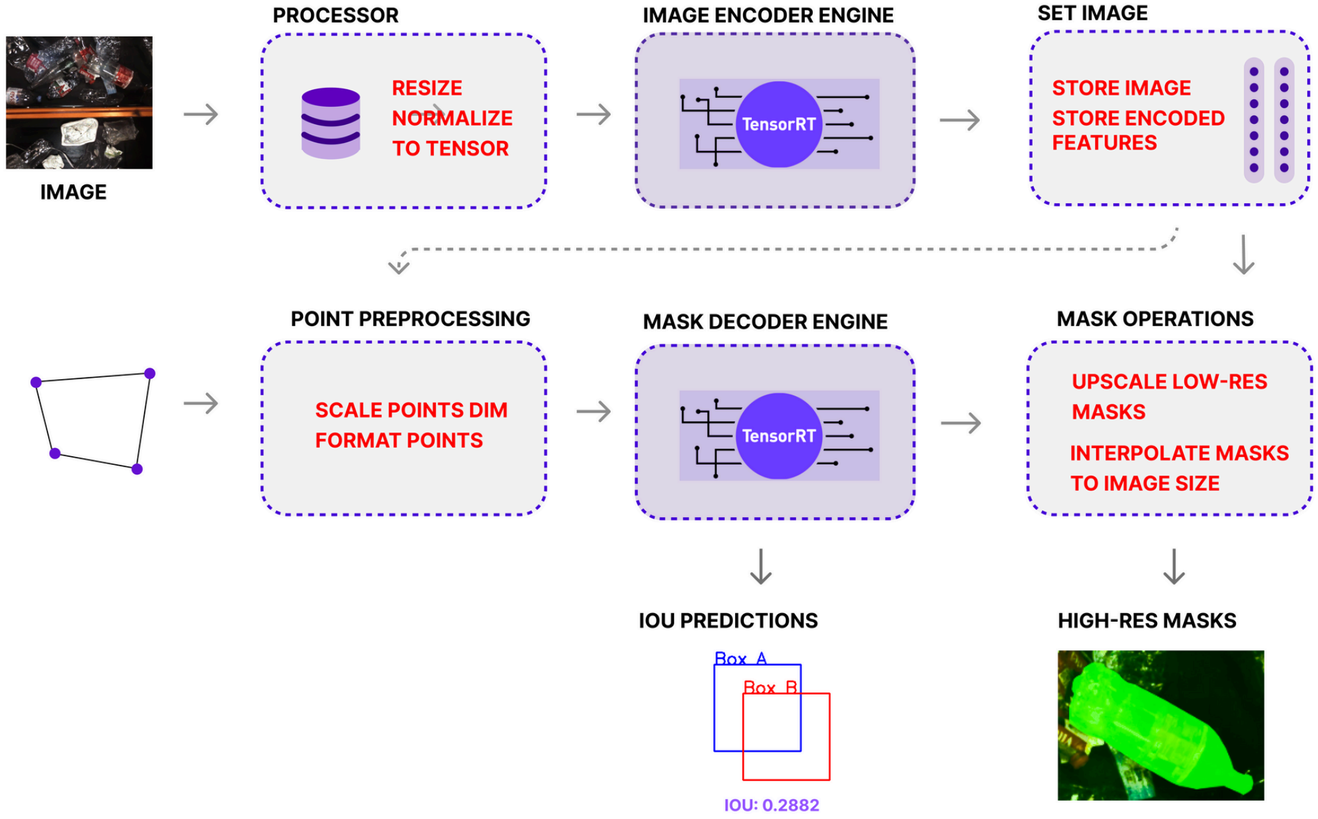
$$L_\delta = \begin{cases} \frac{1}{2}\left(y - f(x)\right)^2, & if\ |y - f(x)| \le \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2, & otherwise \end{cases}$$

Quadratic

Linear

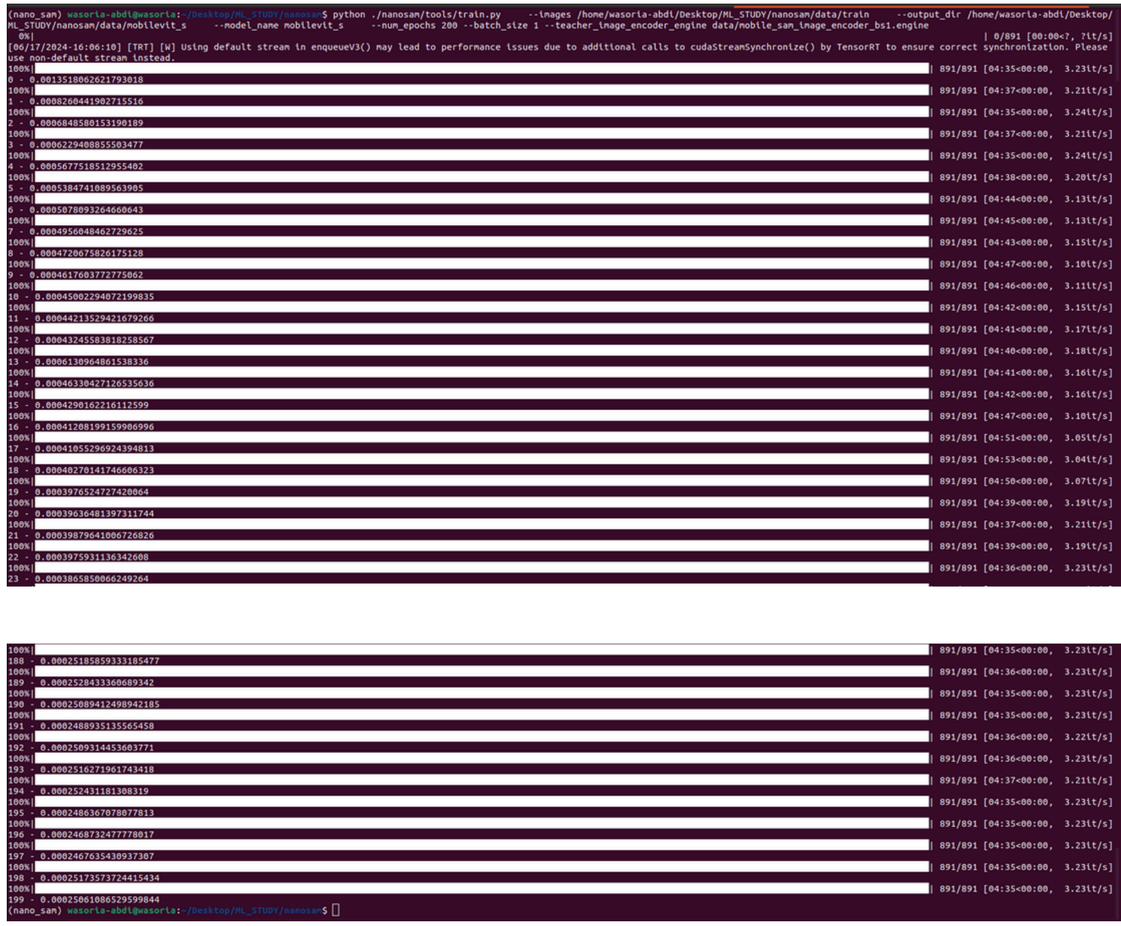
- **δ:** The threshold parameter that controls the point where Huber Loss transitions from quadratic to linear.
- loss function that is less sensitive to outliers than Mean Squared Error (MSE).
- **Quadratic for Small Errors:** When the error is small (I y-f(x) I ≤δ), the loss behaves like MSE.
- **Linear for Large Errors:** When the error is large (I y-f(x) I >δ), the loss behaves like MAE
- Improves features alignment between complex teacher and student models by penalizing large errors less severely

Predictor


PROCESSOR
IMAGE ENCODER ENGINE
SET IMAGE
IMAGE
RESIZE
NORMALIZE
TO TENSOR
TensorRT
STORE IMAGE
STORE ENCODED FEATURES
POINT PREPROCESSING
MASK DECODER ENGINE
MASK OPERATIONS
SCALE POINTS DIM
FORMAT POINTS
TensorRT
UPSCALE LOW-RES MASKS
INTERPOLATE MASKS TO IMAGE SIZE
IOU PREDICTIONS
HIGH-RES MASKS
Box_A
Box_B
IOU: 0.2882

Training configuration

Configuration Parameter	Description	Value
images	Path to the image dataset for distillation	data/coco/train2017
output_dir	Directory to store checkpoints and visualizations	data/wasoria_5k/train
model_name	Name of the student model architecture	resnet18, mobile_vit ...
student_size	Size of the image to feed to the student model	1024
num_images	Limit on the number of images per epoch	None
num_epochs	Number of training epochs	200
batch_size	Number of images per batch	16
num_workers	Number of data loader workers	8
learning_rate	Learning rate for the optimizer	3e-4
distillation loss	Loss function used for distillation	huber
weight decay	Weight decay (L2 regularization)	1e-5
data_augmentation	Data augmentation techniques in training	random crop, flip
checkpoint_path	Path to save and load model checkpoints	data/model/checkpoint.pth
log_file	Path to save training logs	data/model/log.txt



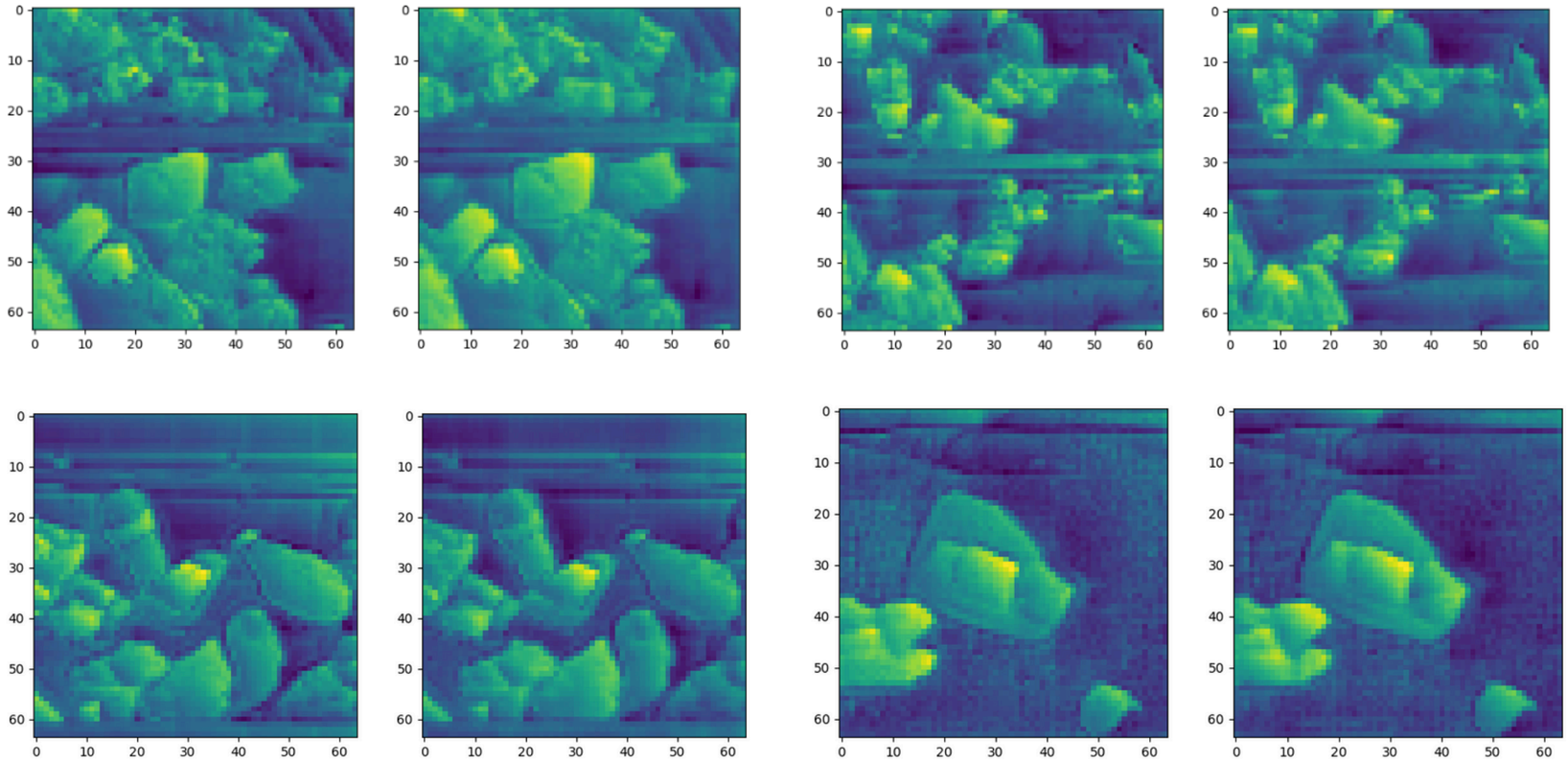
Performance metrics

Model	Default		After Distillation		
	Parameter Size	FLOPs	mIOU all	Parameter Size	FLOPs
MobileSAM	>5M	>10G	-	-	-
ResNet18	11.6M	1.8G	0.73	15.83 M	0.08G
MobileViT-s	9.88M	2.0G	0.75	5.6M	0.10G
EfficientViT-b0	5.3M	0.8G	0.68	4.45M	0.11G

MobileViT-s emerged as the best all-around student model with substantial reductions in parameter size (5.6M) and FLOPs (100 Mega FLOPs), maintaining a high mIOU(). Ideal balance between performance and efficiency.

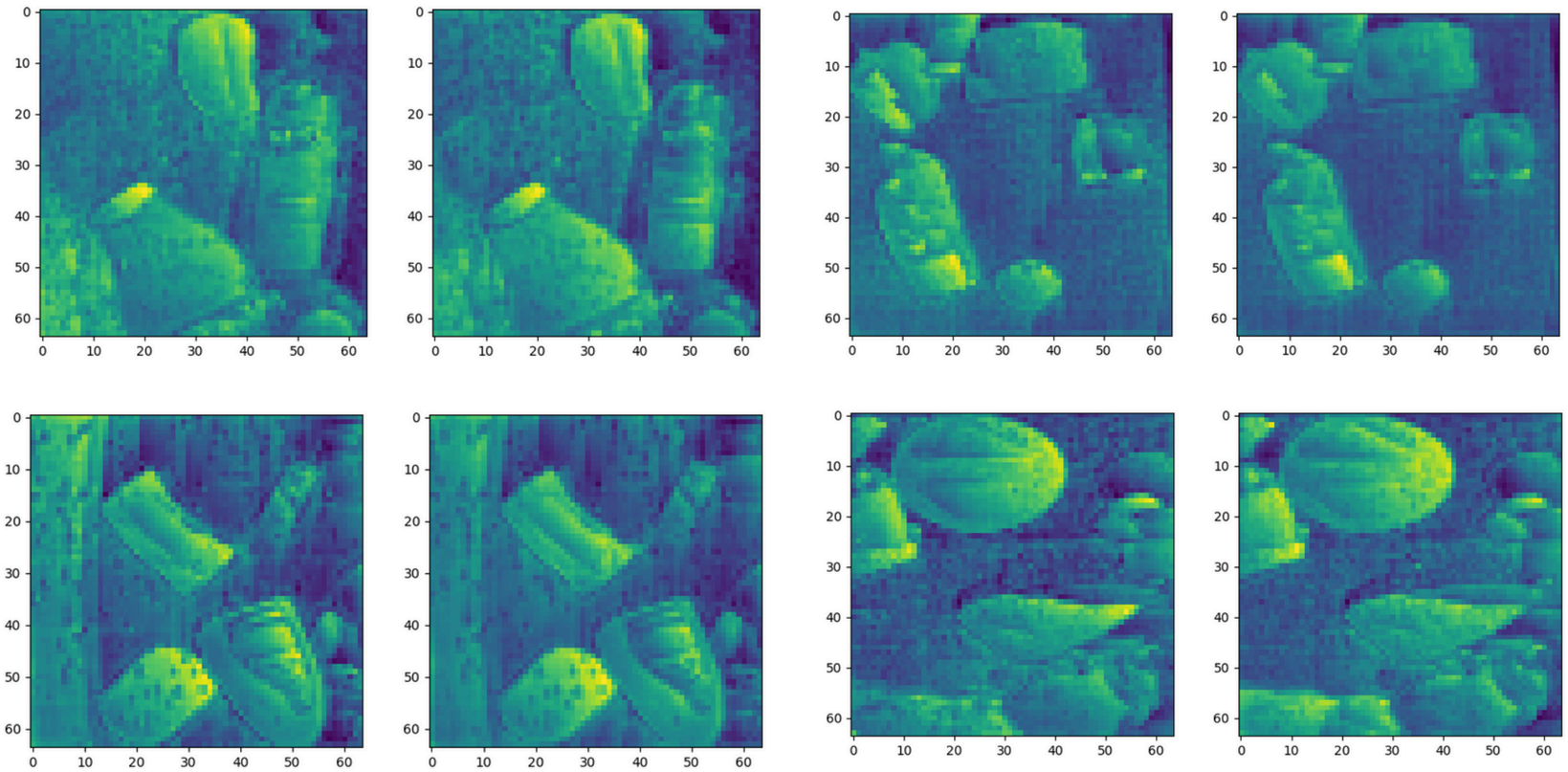
```
(nano_sam) wasoria-abdi@wasoria:~/Desktop/ML_STUDY/nan
/home/wasoria-abdi/Desktop/ML_STUDY/nanosam/data/resne
ONNX Model Analysis
------------------------------------------------
Resnet18
Total Parameters: 15.83M
FLOPs: 0.08G
(nano_sam) wasoria-abdi@wasoria:~/Desktop/ML_STUDY/nan
/home/wasoria-abdi/Desktop/ML_STUDY/nanosam/data/model
onnx
ONNX Model Analysis
------------------------------------------------
Mobilevit S Image Encoder
Total Parameters: 9.88M
FLOPs: 0.10G
(nano_sam) wasoria-abdi@wasoria:~/Desktop/ML_STUDY/nan
/home/wasoria-abdi/Desktop/ML_STUDY/nanosam/data/model
ONNX Model Analysis
------------------------------------------------
Efficientvit B0
Total Parameters: 4.45M
FLOPs: 0.11G
```

MobileViT-s



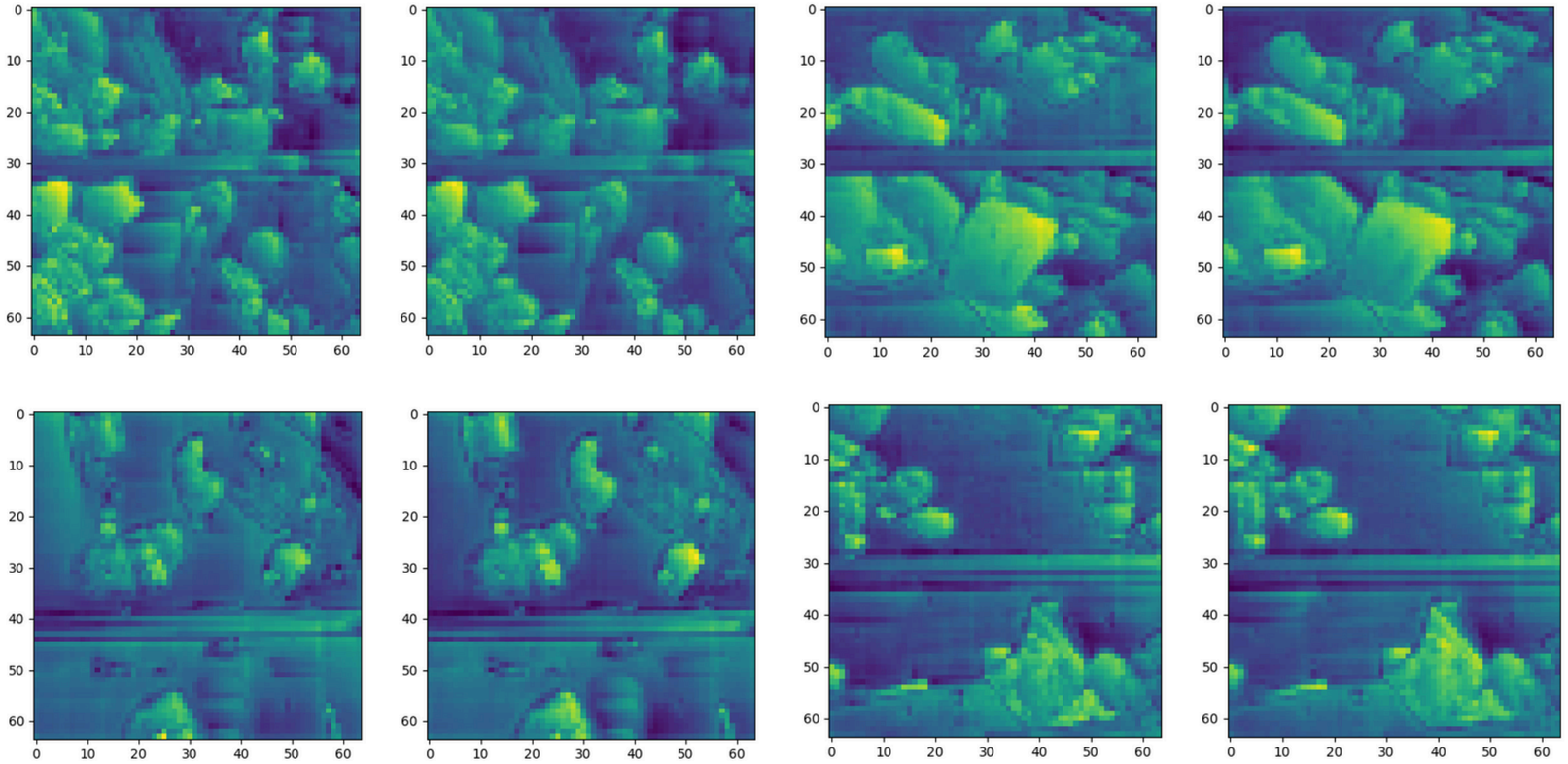
Left: Teacher (MobileSAM), Right: Student (MobileViT-s) showing successful knowledge transfer

ResNet-18



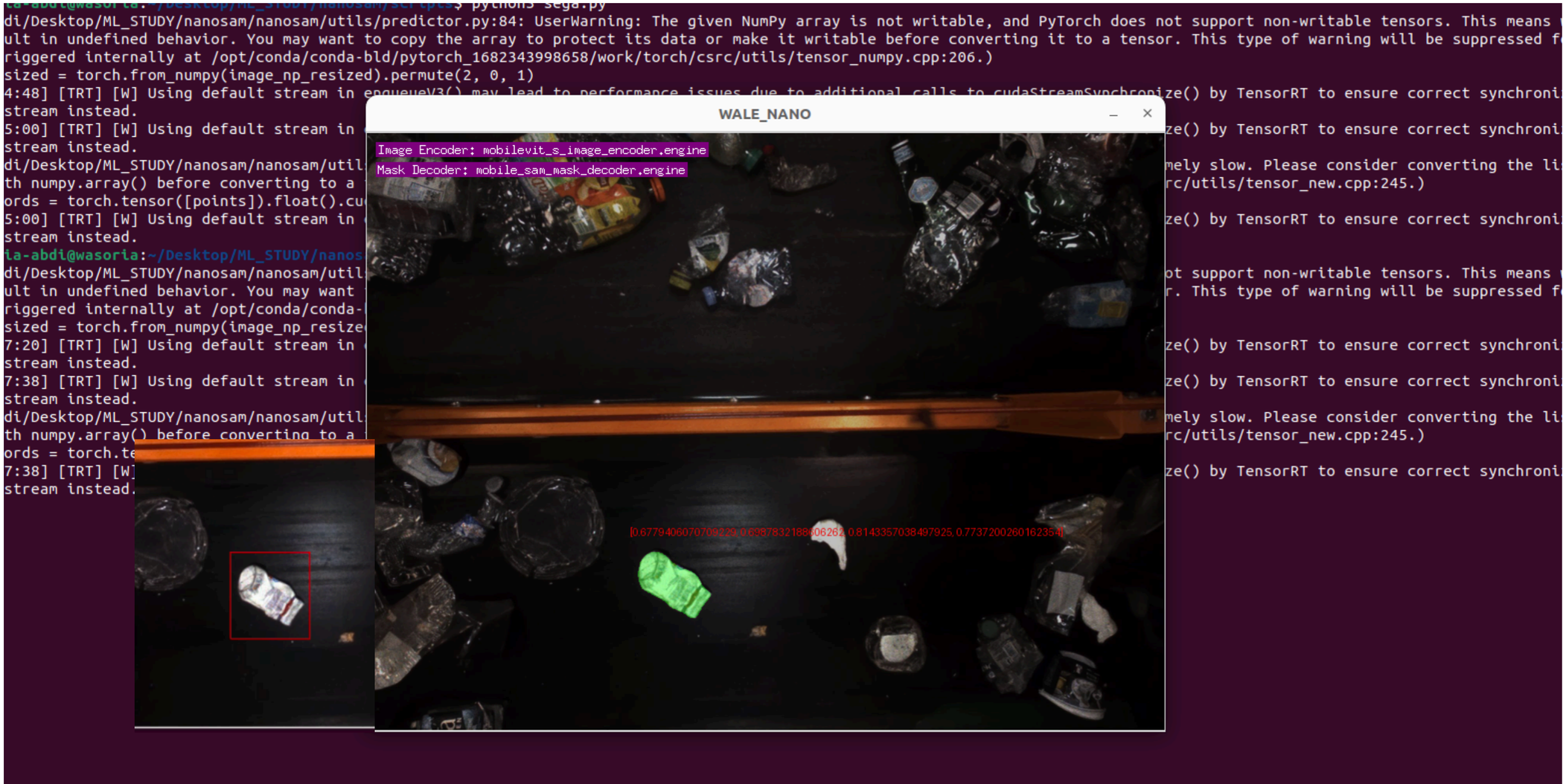
Left: Teacher (MobileSAM), Right: Student (Resnet-18) showing successful knowledge transfer

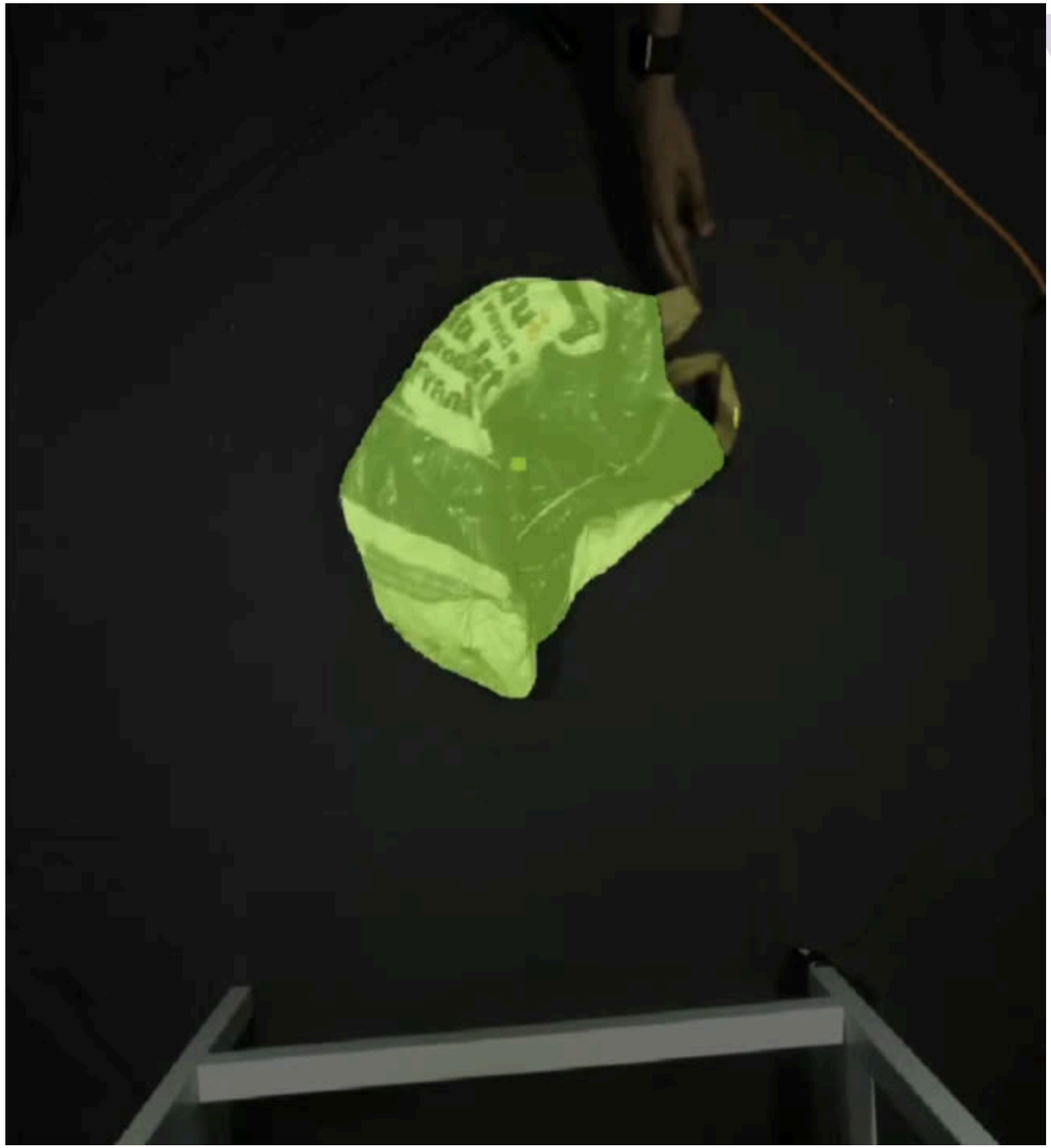
EfficientViT-b0



Left: Teacher (MobileSAM), Right: Student (EfficientVit-b0) showing successful knowledge transfer

Mask Visualizer





References

1. I**mageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness**.
Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019).

2. **Tinyvit: Scaling down vision transformers for efficient transfer learning.**
Microsoft Research Blog, 2022.
https://www.microsoft.com/en-us/research/blog/inyvit-scaling-down-vision-transformers-for-efficient-transfer-learning/

3. **Onnx: Open neural network exchange**. ONNX Official Website, 2023. URL https://onnx.ai/.

4. **Ketan Doshi. Transformers explained visually** - multi-head attention, deep dive. Ketan Doshi Blog, 2023. URL https://ketanhdoshi.github.io/ transformers-explained-visually.

5. **Adversarially robust distillation,** Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein In Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.

6. **Cross-modal vision transformer for efficient multimodal learning:** Yuhang Gong, Wenhao Jiang, Tiejun Huang, and Yonghong Tian. Xmodalvit: . IEEE Transactions on Multimedia, 24:2373– 2383, 2022.

7. **Knowledge distillation: A survey**, Jianping Gou et al. . International Journal of. Computer Vision, 129(6):1789–1819, 2021.

8. **Robust estimation of a location parameter.** Peter J. Huber. The Annals of Mathematical Statistics, 35(1):73–101, 1964.

10. **Efficientvit: Memory efficient vision transformer with cascaded group attention:** Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. . Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. URL https://arxiv.org/abs/2305.07027.

11. **Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer** Manoj Kumar et al. . arXiv preprint arXiv:2110.02178, 2022.

12. **NVIDIA. Tensorrt: High-performance deep learning inference.** NVIDIA Developer Blog, 2023. URL https://developer.nvidia.com/tensorrt.

13. **Segment anything model (sam)** Meta AI Research. Meta AI Blog, 2023. URL https://ai. facebook.com/blog/segment-anything-model-sam/.

14. **Mobilenetv2: Inverted residuals and linear bottlenecks.** Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4510–4520, 2018a. URL https://arxiv.org/abs/1801.04381.

15. **Mobilesam: Lightweight segment anything model,** Kyunghee University Research Team. arXiv preprint arXiv:2304.09867, 2023a.

16. **Exploring mobilevit: Architecture and performance**.Hugo Touvron et al. E Medium, 2022. URL https://medium.com/@hugotouvron/ exploring-mobilevit-architecture and-performance-9e7f5b0f354a.

17. **Tinyvit: Fast pretraining distillation for small vision transformers,** Bin Xiao et al. arXiv preprint arXiv:2207.10633, 2022.

18. **Your classifier is secretly an energy-based model and you should treat it like one**. Jianmin Zhang et al.. International Conference on Learning Representations (ICLR), 2020.

Thank You