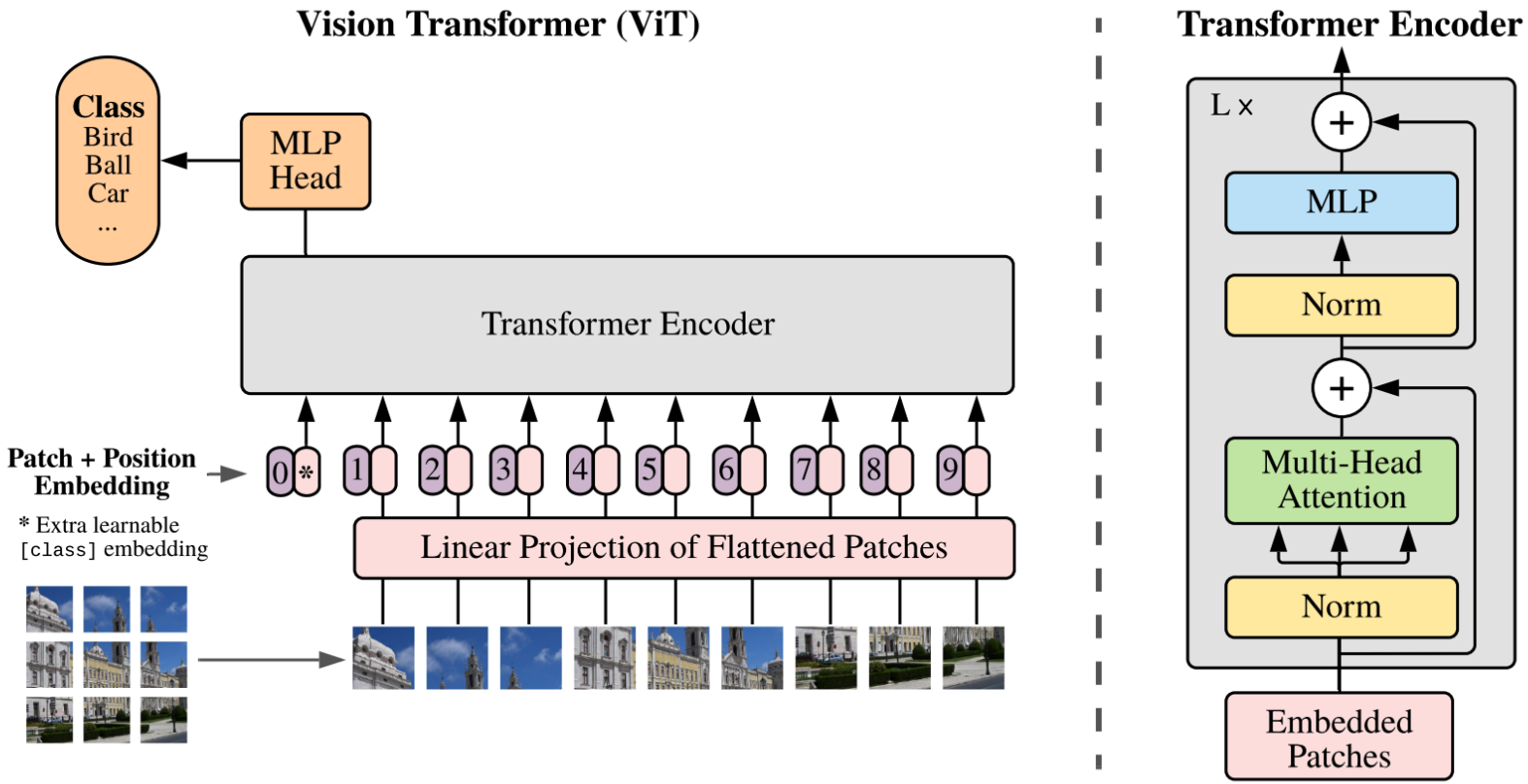
The University of Texas at Austin
Chandra Department of Electrical and Computer Engineering
Cockrell School of Engineering

Energy Aware Computing Research Group

SupMAE: Supervised Masked Autoencoders Are Efficient Vision Learners

Feng Liang, Yangguang Li, Diana Marculescu
The University of Texas at Austin
dianam@utexas.edu
enyac.org

Vision transformers (ViTs) as a CV engine

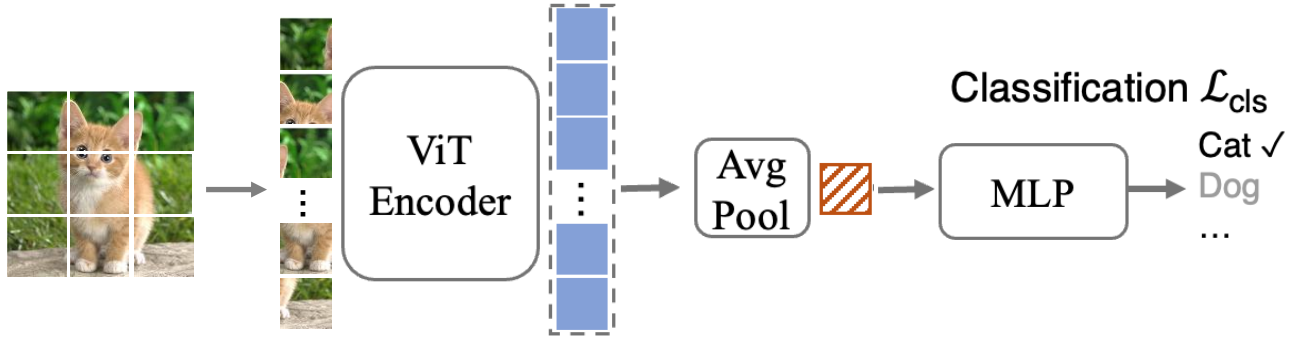


ViT [A. Dosovitskiy *et. al.*]

Vision transformers (ViTs) have emerged as *the* new architecture for computer vision

ViTs are hard to train: Can we combine best of both worlds?

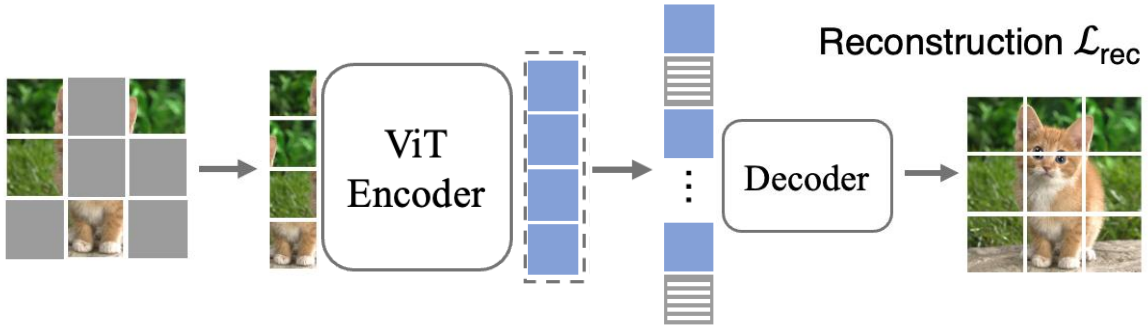
Supervised training



DeiT [H. Touvron *et. al.*]

Training time*	ImageNet acc.
91.5 hours	81.8
✓	✗

Self-supervised pre-training



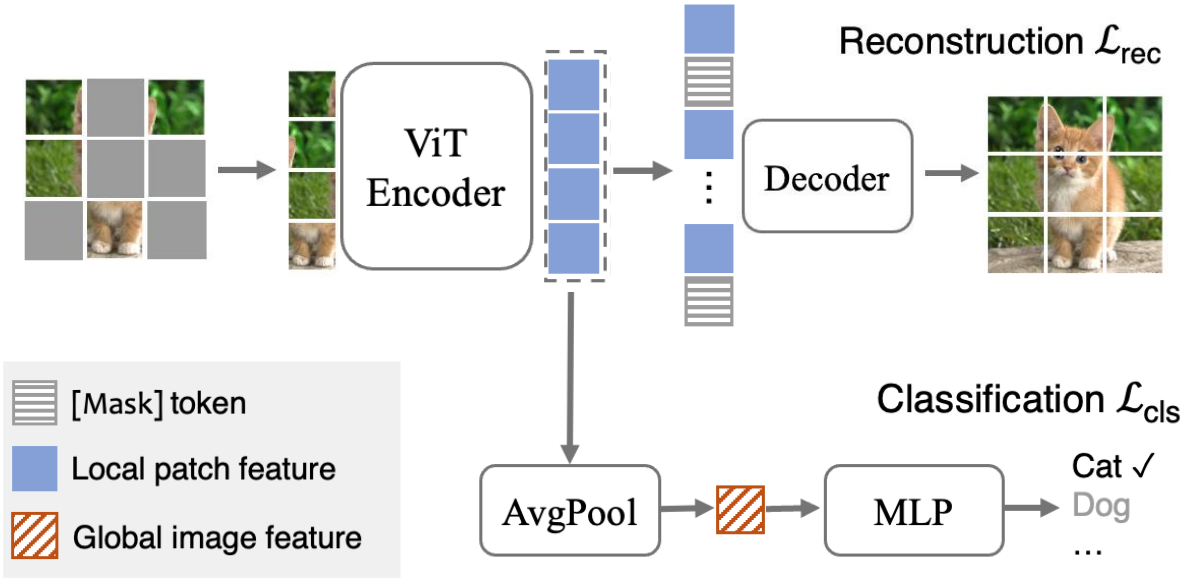
Masked AutoEncoders [K. He *et. al.*]

Training time*	ImageNet acc.+
394 hours	83.6
✗	✓

* Time is measure on 8 A5000 GPUs

+ Accuracy is obtained after supervised fine-tuning on ImageNet

SupMAE achieves the best of both worlds



The proposed SupMAE extends MAE by adding a supervised classification branch

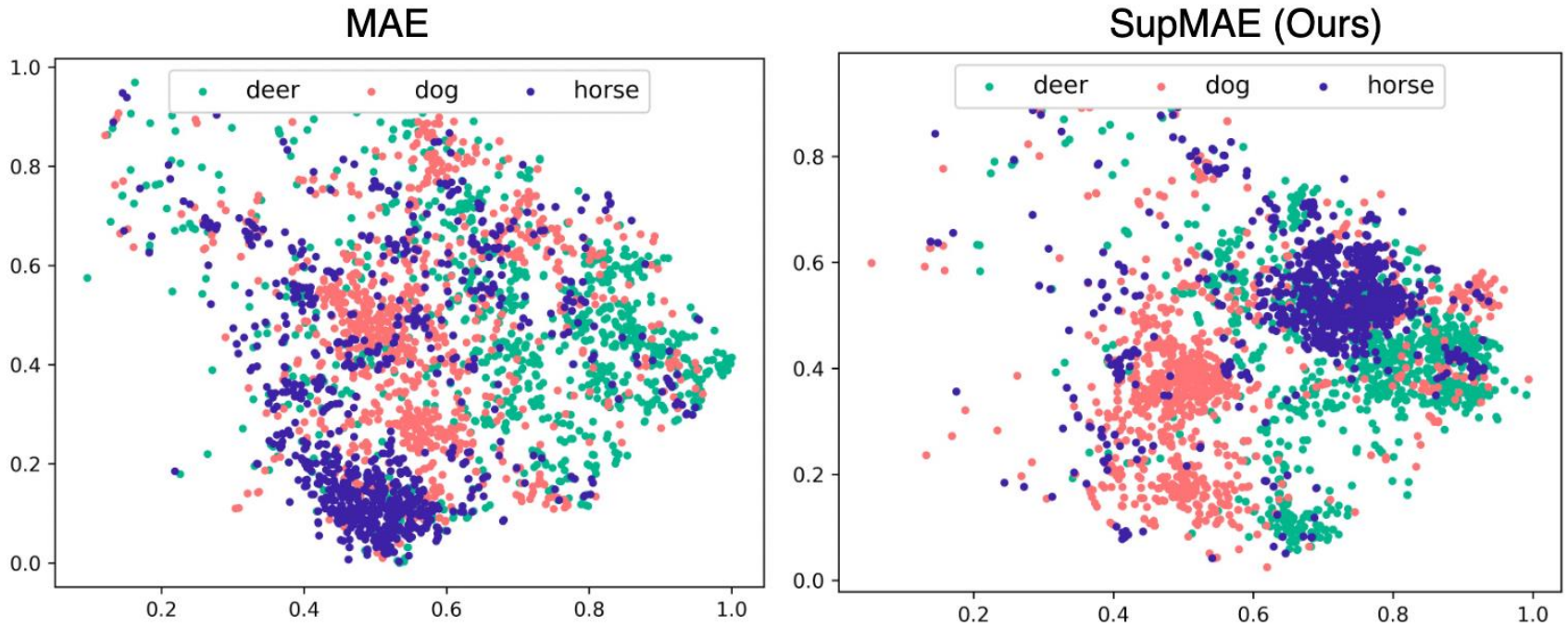
- **Reconstruction loss:** learn middle-level features
- **Classification loss:** learn global features

Training time*	ImageNet acc.+
125.9 hours	83.6
✓	✓

* Time is measure on 8 A5000 GPUs

+ Accuracy is obtained after supervised fine-tuning on ImageNet

SupMAE learns better global features than MAE



t-SNE visualization of pre-trained checkpoints^

SupMAE's features can be better clustered into true categories, revealing that better global features are learnt with proposed supervised branch

^ MAE / SupMAE is pre-trained on ImageNet. We select three categories in CIFAR-10 validation set for t-SNE visualization.

Comparison with supervised and self-supervised methods

Table 1: **Comparison with supervised and self-supervised pre-training methods** All methods are using ViT-B/16 model. Besides the number of pre-training (PT) and fine-tuning (FT) epoch, we further estimate the wall-clock time for PT and FT, benchmarked on 8 A5000 GPUs. The normalized cost is relative to SupMAE. SupMAE shows a great efficiency and can achieve the same accuracy as MAE using only 30% compute.

method	PT epochs	PT cost (Hours)	FT epochs	FT cost (Hours)	Total cost (Hours)	Normalized cost	Top1 acc.
Supervised pre-training methods.							
ViT (Dosovitskiy *et al.* 2020)	-	-	-	-	-	-	77.9
DeiT (Touvron *et al.* 2021)	300	91.5	-	-	91.5	0.73×	81.8
Naive supervised (He *et al.* 2021)	300	90	-	-	90	0.71×	82.3
SupMAE(Ours)	400	95.9	100	30	125.9	1×	**83.6**

- Compared with other supervised methods, SupMAE achieves better performance
- Compared with self-supervised methods, SupMAE achieves comparable performance with much less compute *e.g.,* 30% of MAE

SupMAE is more training efficient

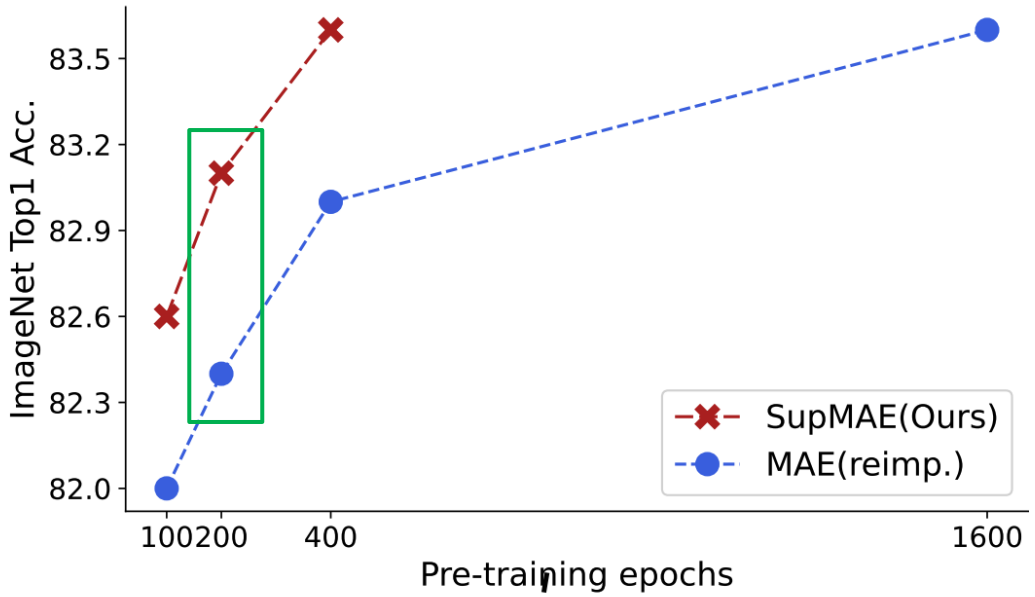


Figure 2: **Performance of different pre-training epochs** Comparison between MAE and SupMAE when pre-trained for different epochs. SupMAE is efficient and shows a much faster convergence speed.

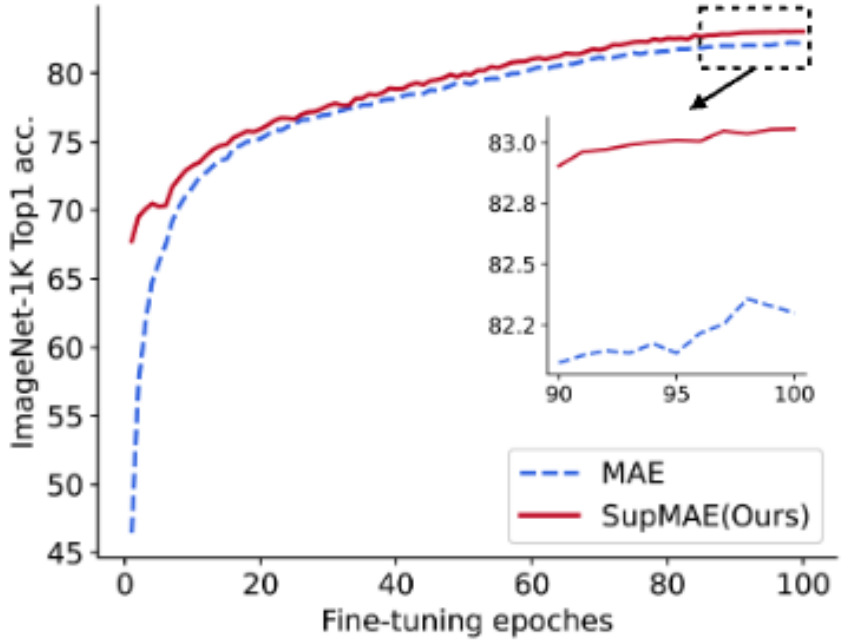


Figure 3: Comparison between MAE and SupMAE when fine-tuned for 100 epochs on ImageNet-1K. The model architecture is ViT-B/16. Both MAE and SupMAE are pre-trained for 200 epochs. Our SupMAE brings a much better initialization point than its MAE counterpart.

- Compared with MAE, SupMAE shows better training efficiency

SupMAE model shows better robustness

ImageNet Sketch (Wang et al.)

ImageNet-R (Hendrycks et al.)

ImageNet-A (Hendrycks et al.)

Table 2: **Robustness evaluation on robustness benchmark.** All methods use the same ViT-B/16 architecture. The metric is top-1 accuracy, except for IN-Corruption (Hendrycks and Dietterich 2019) which uses mean corruption error. We test the same SupMAE model as in Tabel 1 on 4 ImageNet variants *without* any specialized fine-tuning. The score is measured by the averaging metric across four variants (we use '100 - error' for the IN-Corruption performance metric). DeiT results are reproduced using the official checkpoint. Our SupMAE model shows better robustness on the benchmark.

dataset	MAE	DeiT	SupMAE(Ours)
IN-Corruption ↓	51.7	**47.4**	48.1
IN-Adversarial	**35.9**	27.9	35.5
IN-Rendition	48.3	45.3	**51.0**
IN-Sketch	34.5	32.0	**36.0**
Score	41.8	39.5	**43.6**

- Compared with MAE, SupMAE shows better robustness

SupMAE learns more transferable features

Table 3: **Few-shot transfer learning.** All methods use the same ViT-B/16 architecture. We report the linear probing and fine-tuning averaged scores on 20 image classification datasets. X-shot denotes the number of labeled images per category used during transfer learning. Our SupMAE significantly outperforms its MAE counterpart. MAE and MoCo-v3 results are from Li *et al.* (2022a).

Pre-training Settings		20 Image Classification Datasets		
Checkpoint	Method	5-shot	20-shot	50-shot
Linear Probing				
MAE	Self-Sup.	33.37 ± 1.98	48.03 ± 2.70	58.26 ± 0.84
MoCo-v3	Self-Sup.	**50.17** ± **3.43**	**61.99** ± **2.51**	**69.71** ± **1.03**
SupMAE(Ours)	Sup.	47.97 ± 0.44	60.86 ± 0.31	66.68 ± 0.47
Fine-tuning				
MAE	Self-Sup.	36.10 ± 3.25	54.13 ± 3.86	65.86 ± 2.42
MoCo-v3	Self-Sup.	39.30 ± 3.84	58.75 ± 5.55	70.33 ± 1.64
SupMAE(Ours)	Sup.	**46.76** ± **0.12**	**64.61** ± **0.82**	**71.71** ± **0.66**

Table 4: **Transferring to semantic segmentation on ADE20K** All methods use UperNet with ViT-B/16 backbone. For a fair comparison with supervised methods, we use a fine-tuned model for MAE and SupMAE. Naive supervised results are from He *et al.* (2021). MAE results are reproduced using the official fine-tuned checkpoint.

method	mIoU	aAcc	mAcc
Naive supervised	47.4	-	-
MAE	48.6	**82.8**	59.4
SupMAE (ours)	**49.0**	82.7	**60.2**

- SupMAE shows better transfer learning performance compared to other supervised or self-supervised methods

Ablation Study

Table 5: **SupMAE ablation experiments** All experiments are using ViT-B/16 on ImageNet-1K. We report fine-tuning (ft) and linear probing (lin) accuracy (%). If not specified, the default is: the loss ratios of reconstruction (rec) and classification (cls) are 1 and 0.01, global pooling feature is used for classification, the decoder has depth 8, the data augmentation is random resized cropping, the masking ratio is 75%, and the pre-training length is 200 epochs. Default settings are marked in gray.

(a) **Pre-training objectives**. Reconstruction and classfication supports each other.

rec	cls	ft	lin
✓		82.4	58.0
	✓	79.9	59.9
✓	✓	**83.1**	**70.1**

(b) **`Class` token**. Global pooling feature performs better than the additional `class` token.

case	ft	lin
cls token	79.1	65.8
global pool	**83.1**	**70.1**

(c) **Data augmentation**. Our SupMAE works with minimal data augmentation like MAE.

data aug	ft	lin
randcrop	**83.1**	70.1
randcrop,cjit	83.0	**70.3**

(d) **Loss ratio**. Small classification loss ratio works best.

cls ratio	ft	lin
0.02	82.9	**70.2**
0.01	**83.1**	70.1
0.005	**83.1**	69.8
0.002	82.8	68.8

(e) **Decoder depth**. SupMAE works well with a light decoder, *i.e.*, an one-layer transformer decoder.

blocks	ft	lin
1	**83.1**	65.7
4	**83.1**	68.2
8	**83.1**	**70.1**

(f) **MLP layers**. An appropriate number of layers should be set for the classification head.

mlp layers	ft	lin
1	83.0	**72.5**
2	**83.1**	70.1
3	82.9	69.5

SupMAE: more training efficient, with SOTA accuracy

- SupMAE extends MAE to a fully-supervised setting by adding a supervised classification branch, thereby enabling MAE to effectively learn global features from golden labels

- Through experiments, we demonstrate that not only is SupMAE more training efficient but also it learns more robust and transferable features

- Training cost is 4x less for similar performance

The University of Texas at Austin
Chandra Department of Electrical and Computer Engineering
Cockrell School of Engineering

Energy Aware Computing Research Group

Check our code & models

Thank you!