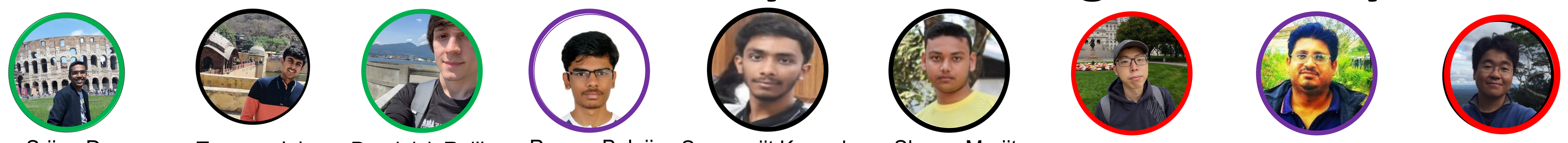
Limited Data, Unlimited Potential: A Study on ViTs Augmented by Masked Autoencoders



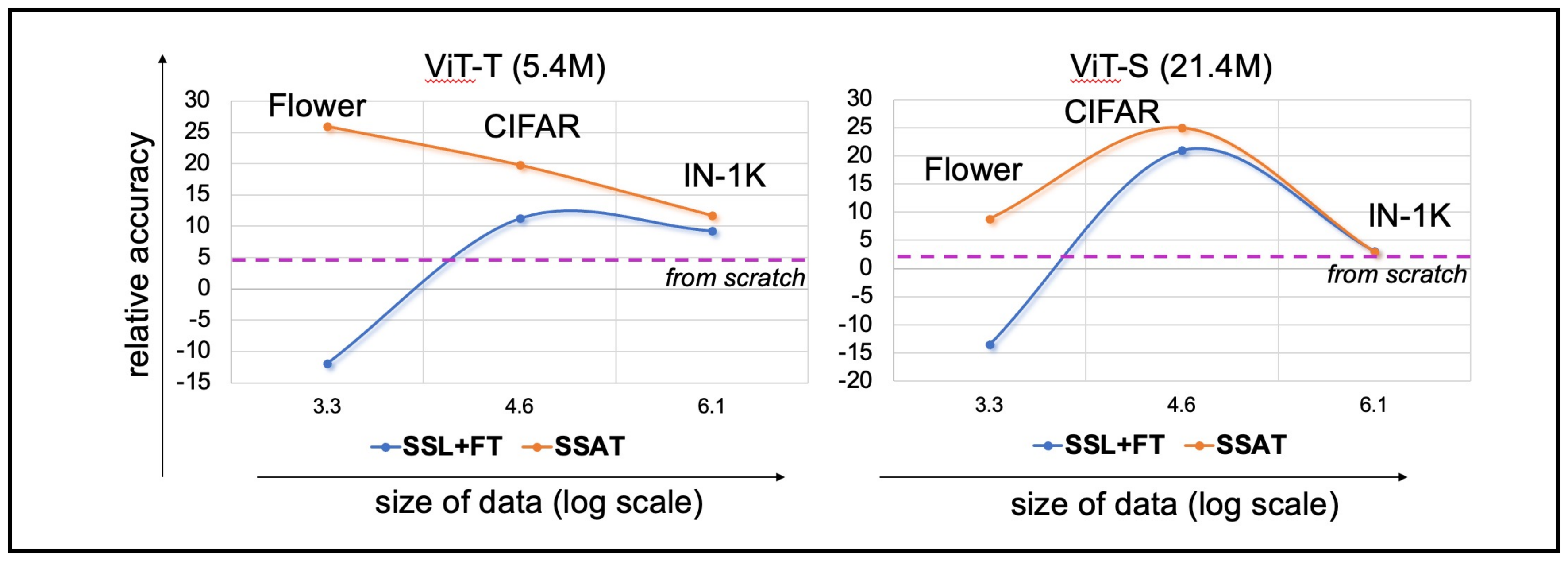
Srijan Das, Tanmay Jain, Dominick Reilly, Pranav Balaji, Soumyajit Karmakar, Shyam Marjit, Xiang Li, Abhijit Das, Michael S. Ryoo

UNC Charlotte, DTU, BITS Pilani Hyderabad, IIIT Guwahati, Stony Brook University

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

Motivation

- Training transformers require **large-scale data**
- What about small-scale data distribution?
  - Use pre-trained models
- What about domains with Limited data?
  - For example – **Medical data**

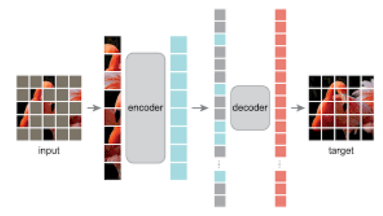


Relative classification accuracy on three datasets with different sizes: (i) *Oxford Flower* (**2K** samples), (ii) *CIFAR* (**50K** samples), and (iii) *ImageNet-1K* (IN-1K, **1.2M** samples). **Self-supervised Auxiliary Task** (SSAT) consistently outperforms others on all three datasets with two backbones. On the other hand, given the same Self-supervised Learning (SSL) method, SSL+ Fine-tuning (FT) achieves a compromised performance than SSAT, especially on the tiny Oxford Flower dataset (even worse than training from scratch).

Self-supervised Auxiliary Task (SSAT)

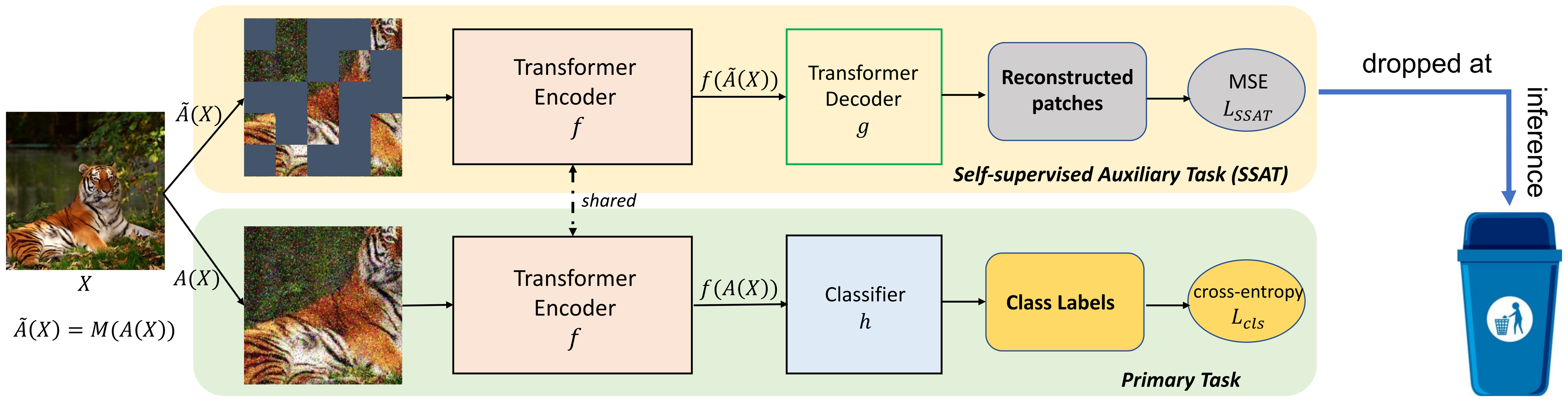
Straightforward implementation

Which SSL task?



Training SSAT

- Inflate the input batch of images with two set of Augmentations
- Two streams -> **Primary** + **Auxiliary**
- **Primary** stream – Performing classification of the image
- **Auxiliary** Stream – Reconstructing the input image from 25% of the input
- $Total\ Loss = \lambda Loss_{Primary} + (1-\lambda)Loss_{Auxiliary}$



Experiments

SSAT improves ViT performance while reducing carbon footprints

SSAT trained with outperforms on 12 Computer Vision tasks

Top-1 classification accuracy (%) of different ViT variants with and without SSAT on **CIFAR-10**, **CIFAR-100**, **Flowers102**, and **SVHN** datasets. All models were trained for 100 epochs.

Method	# params. (M)	CIFAR-10	CIFAR-100	Flowers102	SVHN
ViT-T (Touvron et al., ICML 21)	5.4	79.47	55.11	45.41	92.04
+SSAT	5.8	**91.65** (+12.18)	**69.64** (+14.53)	**57.2** (+11.79)	**97.52** (+5.48)
ViT-S (Touvron et al., ICML 21)	21.4	79.93	54.08	56.17	94.45
+SSAT	21.8	**94.05** (+14.12)	**73.37** (+19.29)	**61.15** (+4.98)	**97.87** (+3.42)
CVT-13 (Wu et al., ICCV 21)	20	89.02	73.50	54.29	91.47
+SSAT	20.3	**95.93** (+6.91)	**75.16** (+1.66)	**68.82** (+14.53)	**97** (+5.53)
Swin-T (Liu et al., CVPR 21)	29	59.47	53.28	34.51	71.60
+SSAT	29.3	**83.12** (+23.65)	**60.68** (+7.4)	**54.72** (+20.21)	**85.83** (+14.23)
ResNet-50 (He et al., CVPR 16)	25.6	91.78	72.80	46.92	96.45

Does SSAT promotes overfitting?

Method	IN-1K	CIFAR-100-*p*	IN-1K-*p*
ViT-T	65.0	25.1	48.3
+SSAT	**72.7**	**37.6**	**59.6**
ViT-S	74.2	22.5	62.7
+SSAT	**76.4**	**43.9**	**64.5**

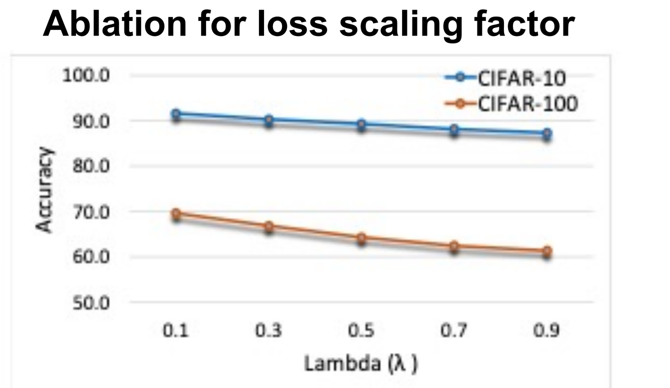
Comparison of SSAT with SSL+FT

Method	GFLOPs	CIFAR-10	CIFAR-100	IN-1K	Train time	Kg CO_2 eq.
Scratch	1.26	79.47	55.11	65.0	60	5.96
SSL+FT	0.43+1.26	85.33	60.43	70.09	55	5.46
SSL+FT	0.43+1.26	86.48	63.28	71.1	82	8.15
SSL+FT	0.43+1.26	88.72	67.53	**74.07**	104	10.33
Ours	1.67	**91.65**	**69.64**	72.69	78	7.55

Appropriate SSL for SSAT

SSAT (SSL)	CIFAR-10	CIFAR-100
SimCLR (Chen et al., ICLR 20)	55.21	36.49
DINO (Caron et al., CVPR 20)	80.07	60.6
MAE (He et al., CVPR 20)	**91.65**	**69.64**

Ablation for loss scaling factor

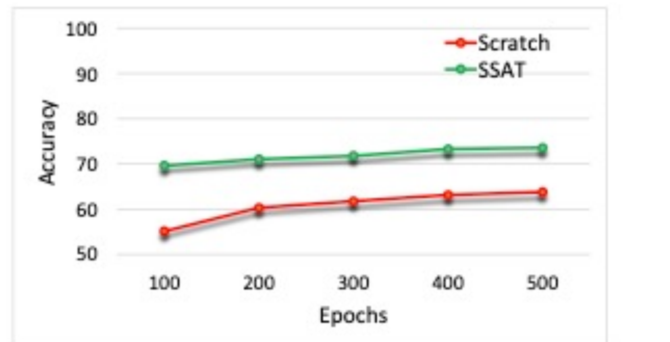


Performance of SSAT based ViT on Medical and DomainNet datasets

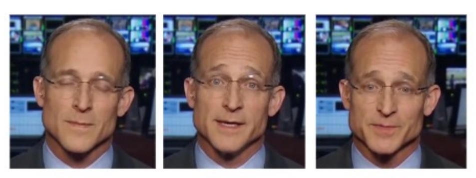
	Method	Chaoyang	PMNIST
ViT-T	Scratch	77.37	90.22
	IN-1K pretrained + FT	78.78	91.99
	Scratch + **SSAT**	**82.52**	**93.11**
ViT-S	Scratch	80.04	91.19
	IN-1K pretrained + FT	80.18	92.63
	Scratch + **SSAT**	**81.25**	**93.27**

Method	ClipArt	Infograph	Sketch
ViT-T	29.66	11.77	18.95
+SSAT	47.95	16.37	46.22
CVT-13	60.34	19.39	56.98
$+\mathcal{L}_{drloc}$ (Liu et al., NeurIPS 21)	60.64	20.05	57.56
+SSAT	**60.66**	**21.27**	**57.71**

ViT vs. ViT+SSAT for longer training epochs



SSAT for Video DeepFake Detection

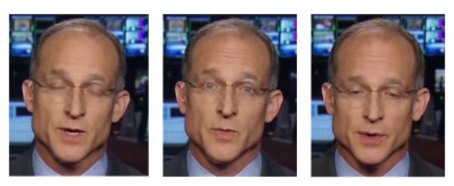


Source

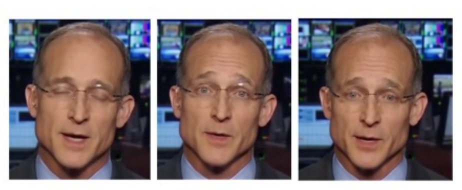
Manipulations

Deepfakes

FaceSwap



Face2Face



NeuralTextures

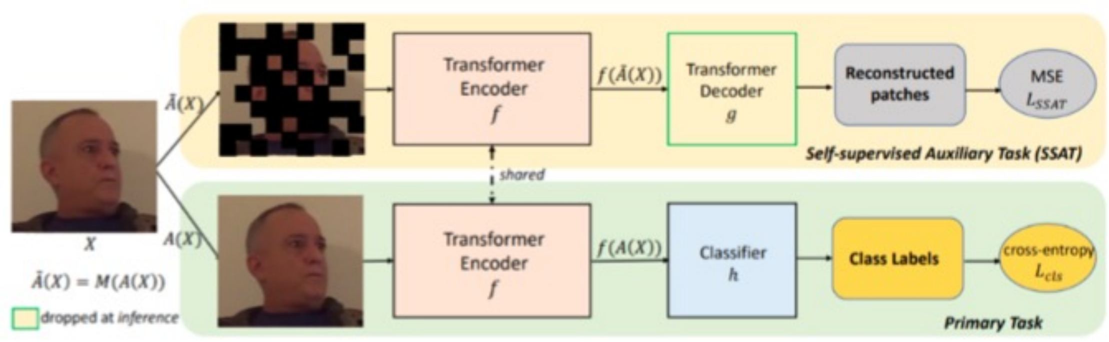
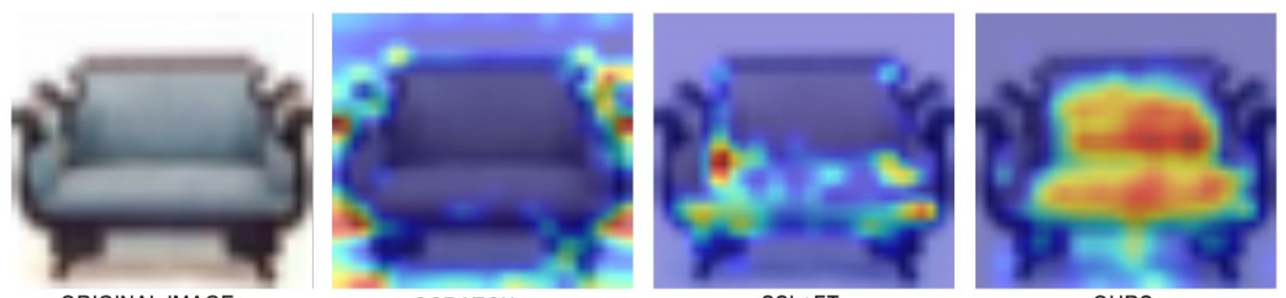
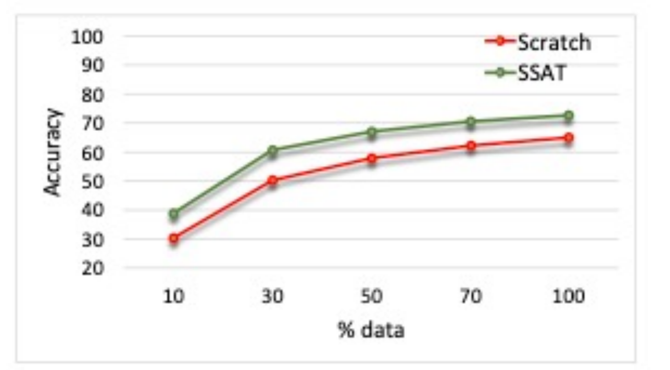


Figure 11. Mask Autoencoder as a Self Supervised Auxiliary Task for deepfake detection.

GradCAM visualizations



ViT vs. ViT+SSAT for different fraction of data.



Cross training evaluation and zero-shot transfer results of DeepFake detection on FaceForensics++ with SSAT.

Method	cross-training evaluation				zero-shot transfer			
	Deepfakes	Face2Face	FaceSwap	NeuralTextures	Deepfakes	Face2Face	FaceSwap	NeuralTextures
Scratch	84.48	79.21	56.63	**82.08**	-	-	-	-
Cross-efficient-vit (Coccomini et al., ICIAP 22)	82.67	69.89	79.93	64.87	-	-	-	-
DFDC winner (Seferbekov, DFDC 20 winner)	96.43	73.93	86.07	58.57	88.57	57.50	80.36	54.64
VideoMAE SSL (0.95)	82.67	64.16	58.42	63.44	86.28	49.82	69.18	51.97
VideoMAE SSL (0.75)	78.34	65.59	57.35	61.65	82.67	48.39	65.23	51.97
VideoMAE (0.95) + SSAT	92.42	79.21	89.61	81.36	**92.42**	**61.65**	**92.83**	**62.37**
VideoMAE (0.75) + SSAT	**96.75**	**80.65**	**91.40**	72.76	87.73	60.57	88.17	61.65

State-of-the-art comparison

Method	# enc. params.	epochs	CIFAR-10	CIFAR-100
CVT-13+$\mathcal{L}_{drloc}$ (Liu et al., NeurIPS 21)	20M	100	90.30	74.51
CVT-13+ SSAT			**95.93**	**75.16**
ViT (scratch)	2.8M	300	93.58	73.81
SL-ViT (Lee et al., WACV 23)			94.53	76.92
ViT[†] (SSL+FT) (Gani et al., BMVC 22)			94.2	76.08
ViT + SSAT			**95.1**	**77.8**
DeiT-Ti+$\mathcal{L}_{guidance}$ (Li et al., ECCV 22)	6M	300	-	78.15
DeiT-Ti+$\mathcal{L}_{guidance}$ + SSAT			-	**79.46**

Conclusion

- We presented a very simple method of using **SSL** to train **ViTs** on domains with Limited data
- **SSAT** – Jointly optimize the **primary task** with SSL as an **auxiliary task**
- Effectiveness validated on **10 image classification** datasets + **2 video** datasets
- If you plan to use ViTs on a small training distribution, consider using **SSAT**!