Appendix

A. Dataset Description

This paper presents a comprehensive evaluation of our ViT models on 10 different image datasets, comprising prominent computer vision benchmarks such as ImageNet-1K [12] (IN-1K), CIFAR-10 and CIFAR-100 [25], Oxford Flowers102 [37], and SVHN [36]. In addition, we include three datasets namely ClipArt, Infograph, and Sketch from DomainNet [41], a widely adopted benchmark for domain adaptation tasks. Moreover, we explore the performance of our approach on two medical image domain datasets: Chaoyang [63] and PneumoniaMNIST [60]. The dataset size, sample resolution, and the number of classes are further elaborated in Table 9. Note that the accuracies reported for CIFAR in Figure 1 of the main paper is an average of the classification accuracy of CIFAR-10 and CIFAR-100.

Table 9. Details of image classification datasets (sample size, resolution, and number of classes) evaluated in our experiments.

Dataset	**Train Size**	**Test Size**	**Dimensions**	**# Classes**
CIFAR-10	50,000	10,000	32×32	10
CIFAR-100	50,000	10,000	32×32	100
Flowers102	2,040	6,149	224×224	102
SVHN	73,257	26,032	32×32	10
ImageNet-1K	1,281,167	100,000	224×224	1000
ClipArt	33,525	14,604	224×224	345
Infograph	36,023	15,582		
Sketch	50,416	21,850		
Chaoyang	4,021	2,139	512×512	4
PMNIST	5,232	624	28×28	2

Table 10. Ablation of decoder depth.

Decoder Depth	**Accuracy**	
	CIFAR-10	**CIFAR-100**
1	91.59	68.41
2	**91.65**	**69.64**
4	90.88	67.46
8	90.59	67.78
12	91.08	66.94

B. Training Configurations

We follow the configurations introduced in MAE [20]. A comprehensive set of training configurations for all datasets used in this study is provided in Table 14 for reference. During training two parameters, **image and patch sizes** vary depending upon the datasets and the rest of the parameters are the same across all the datasets.

Table 11. Ablation of decoder embedding dimension.

Decoder Dimension	**Accuracy**	
	CIFAR-10	**CIFAR-100**
64	89.20	67.11
128	**91.65**	**69.64**
256	91.64	66.98
512	90.53	66.19

Table 12. Ablation of Decoder Heads.

Decoder Heads	Accuracy	
	CIFAR-10	CIFAR-100
1	91.54	69.44
2	92.44	69.28
4	**92.49**	68.59
8	92.09	69.52
16	91.65	**69.64**

Table 13. Statistics of video datasets generated by different manipulating techniques available in Faceforensics++

Split	**DeepFake**	**Face2Face**	**FaceSwap**	**NeuralTextures**	**Original**	**Total**
Train	720	720	720	720	720	3600
Val	140	140	140	140	140	700
Test	140	140	140	140	140	700
Total	1000	1000	1000	1000	1000	5000

Swin and ConvMAE training configurations: We have adopted the training pipeline from [31] and [16] for Swin [34] and ConvMAE [16] respectively. For each of them, we have combined their reconstruction based self-supervised learning (SSL) and fine-tuning in a joint learning framework, keeping the training configurations same. Note that UM-MAE [31] with its secondary masking strategy, is an efficient version of SimMIM [59] allowing the reconstruction based SSL in hierarchical transformers like Swin [34] and PVT [54].

C. ViT Decoder for Reconstruction based SSL

In contrast to MAE [20], this paper employs a reconstruction-based SSL approach with class-wise supervision. Consequently, we explore the effect of different design choices of the ViT decoder, which can impact the SSL training while simultaneously optimizing cross-entropy in Self-supervised Auxiliary Task (SSAT). To this end, we conduct experiments that involve modifying three decoder attributes: **depth**, **dimension**, and **attention heads**. We evaluate the resulting impact on the model's top-1 accuracy using two datasets: CIFAR-10 and CIFAR-100.

Decoder Depth: In this study, we investigated the impact of decoder depth on model performance, as shown in Table 10. During the experiments, we maintained a fixed decoder dimension of 128, decoder heads of 16, and a value of λ equal to 0.1. Our findings demonstrate that the optimal results for both datasets were obtained at a decoder depth of 2.

Decoder Embedding Dimension: This section investigates the influence of the decoder embedding dimensions on model performance, as presented in Table 11. Throughout these experiments, we maintained a constant value of λ at 0.1, a decoder depth of 2, and 16 decoder heads. Our results indicate that the optimal performance was achieved with a decoder dimension of 128.

Decoder Heads: Table 12 presents the outcomes of the ablation study performed to evaluate the impact of the number of heads on the ViT's performance. The hyperparameters, namely $\lambda = 0.1$, $decoder\_depth = 2$, and $decoder\_dimension = 128$, are fixed to their optimal values from the prior experiments. The experimental findings indicate that retaining 4 heads for CIFAR-10 and 8 heads for CIFAR-100 resulted in the highest performance levels. To ensure generalizability across our experiments, we fixed the number of decoder heads to 16.

D. Details of deepfake detection experiments

In this section, we elaborate the cross-training manipulation and zero-shot transfer experimental details for deepfake detection.

D.1. Datasets

We employ two publicly available popular dataset on Deepfakes.

FaceForensics++: The FaceForensics++ dataset [44] is a large-scale benchmark dataset for face manipulation detection, which is created to help develop automated tools that can detect deepfakes and other forms of facial manipulation. The dataset consists of more than 1,000 high-quality videos with a total of over 500,000 frames, which were generated using various manipulation techniques such as facial reenactment, face swapping, and deepfake generation.

The videos in the dataset are divided into four categories, each corresponding to a different manipulation technique: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. Deepfakes use machine learning algorithms to generate realistic-looking fake videos, while Face2Face and FaceSwap involve manipulating the facial expressions and identity of a person in a video. NeuralTextures uses a different approach by altering the texture of a face to make it appear different. The dataset includes both real and manipulated videos, with each manipulation technique applied to multiple individuals. The statistics of different manipulating techniques available in faceforensics++ is provided in Table 13.

DFDC: The Deepfake Detection Challenge (DFDC) dataset [46] is a large-scale benchmark dataset for deepfake detection. The dataset consists of more than 100,000 videos generated using various facial modification algorithms. The DFDC dataset consists of two versions: a preview dataset with 5k videos featuring two facial modification algorithms and a full dataset with 124k videos featuring eight facial modification algorithms. The DFDC dataset is the largest currently and publicly available face swap video dataset, with around 120,000 total clips sourced from 3,426 paid actors. The videos are produced using several Deepfake, GAN-based, and non-learning methods. The official DFDC train, validation and test splits are also designed to simulate real-world performance, with the validation set consisting of a manipulation technique not present in the train set, and the test set containing much more challenging augmentations and perturbations.

D.2. Methodology

VideoMAE [51]: VideoMAE is a self-supervised video pre-training method that extends masked autoencoders (MAE) to videos. VideoMAE performs the task of masked video modelling for video pre-training. It employs an extremely high masking ratio (90%-95%) and tube masking strategy to create a challenging task for self-supervised video pre-training. The temporally redundant video content enables a higher masking ratio than that of images. This is partially ascribed to the challenging task of video reconstruction to enforce high-level structure learning.

SSAT: In this experiment, we use the same backbone as in the original work [51] and we use rigorous augmentations as used by the winners of the DFDC Challenge [46] in our experimental setting. For training VideoMAE along with SSAT on DFDC, we extend our image based framework to videos (as illustrated in Figure 11) and jointly optimize the primary deepfake classification loss L_{cls} and the auxiliary video reconstruction loss L_{SSAT} as

$$L = \lambda * L_{cls} + (1 - \lambda) * L_{SSAT} \tag{2}$$

where $\lambda = 0.1$ is the loss scaling factor.

D.3. Implementation details

While training VideoMAE+SSAT models follow the training recipe of [51], we have incorporated specific modifications tailored for deepfake detection.

Fake class weight: Assigns weight w to the class representing *fake* in the weighted cross entropy loss. This was used since the training set is very imbalanced (82% fake - 18% real).

$$L_{CE} = -(wt_{real} \log p_{real} + (1 - w)t_{fake} \log p_{fake}) \tag{3}$$

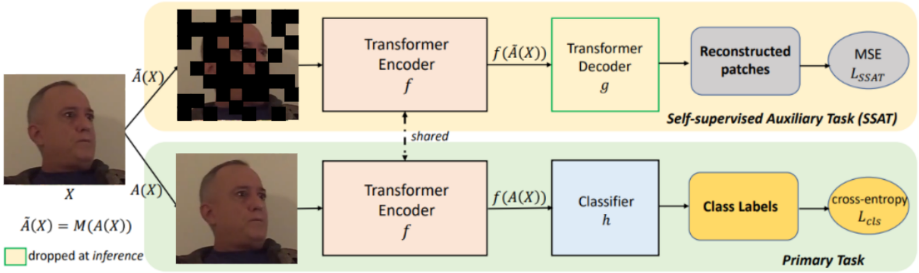

Figure 11. Mask Autoencoder as a Self Supervised Auxiliary Task for deepfake detection.

Equation 3: Weighted Cross-Entropy Loss. w is the weight of the *real* class while p_{real} and p_{fake} are the predicted probabilities, and t_{real} and t_{fake} are the ground truth indicator variables.

Augmentations: The choice of augmentations has a profound impact on validation performance. The set of augmentations that work best are Image Compression, Gaussian Noise, Gaussian Blur, Horizontal Flip, Brightness Contrast, FancyPCA, Hue Saturation, Greyscale and shift-scale-rotate, all available in the Albumentations library [3] and used in the DFDC challenge's winning solution by Selim Seferbekov [46]. Other augmentations like Reversal, Random up / down sampling and heavy Gaussian Noise seem to have a detrimental effect, possibly because they do not generalize to the validation set. Meanwhile, having no augmentations also decreases the generalizability.

Testing: During testing, predictions are obtained by averaging the results from all 16-frame segments across the entire video.

E. Attention Visualization

In Figure 12, we illustrate attention visualization for a few sample images drawn from the Flower and ImageNet datasets. Our analysis of the visualization highlights that the ViT trained with SSAT generates attention maps that emphasize the primary object class to a greater extent than the attention maps computed by ViTs trained from scratch and trained with SSL+FT. These findings indicate that the ViT trained with SSAT exhibits higher efficacy in image classification.

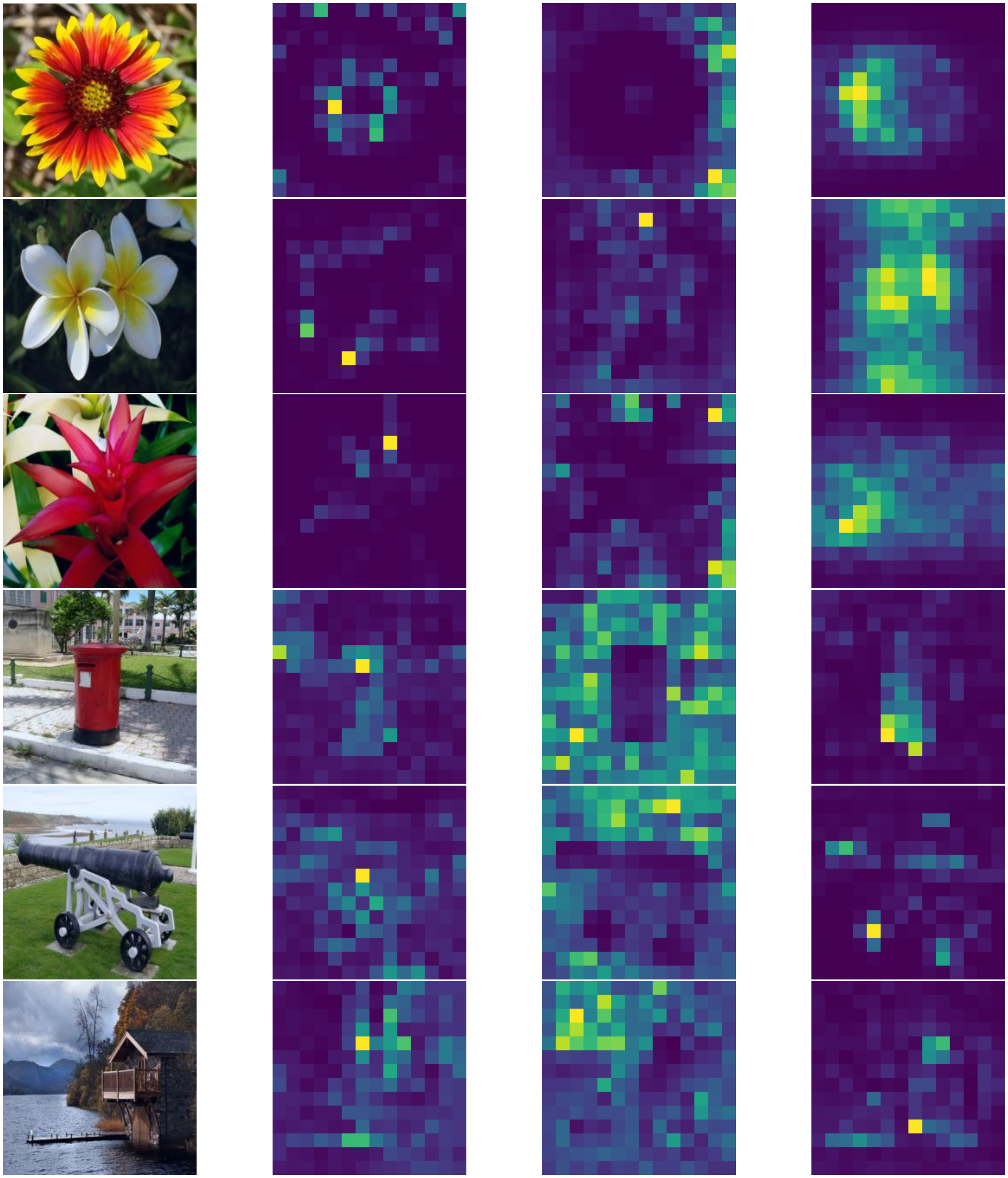

Figure 12. Attention visualization of six images, three from the Oxford Flowers-102 dataset (top 3 rows) and three from the ImageNet dataset (bottom 3 rows). The attention heatmaps in the second, third, and fourth columns correspond to models trained from scratch using ViT, models trained using SSL+FT, and models trained using SSAT, respectively.

Table 14. Our ViT training settings across different datasets.

Input Size	PMNIST	28×28
	CIFAR10, CIFAR100, SVHN	32×32
	Flowers, ImageNet-1K ClipArt, Infograph, Sketch	224×224
	Chaoyang	512×512
Patch Size	PMNIST, CIFAR10, CIFAR100, SVHN	2×2
	Flowers, ImageNet-1K ClipArt, Infograph, Sketch	16×16
	Chaoyang	32×32
Batch Size	64	
Optimizer	AdamW	
Optimizer Epsilon	1e-08	
Momentum	$\beta_1 = 0.9,\ \beta_2 = 0.999$	
layer-wise lr decay	0.75	
Weight Decay	0.05	
Gradient Clip	None	
Learning Rate Schedule	Cosine	
Learning Rate	1e-03	
Warmup LR	1e-06	
Min LR	1e-6	
Epochs	100	
Warmup Epochs	5	
Decay Rate	0.05	
Drop Path	0.01	
Lambda (λ)	0.1	
Masking Ratio	0.75	
Random Resized Crop Scale, Ratio	(0.08, 1.0), (0.75, 1.3333)	
Interpolation	bicubic	
Random Horizontal Flip Probability	0.5	
Rand Augment	n = 2	
Random Erasing Probability, Mode and Count	0.25, Pixel, (1, 1)	
Color Jittering	None	
Auto-augmentation	rand-m9-mstd0.5-inc1	
Mixup	True	
Cutmix	False	
Mixup, Cutmix Probability	1, 0	
Mixup Switch Probability	0.5	
Mixup Mode	Batch	
Label Smoothing	0.1	