This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Srijan Das[1], Tanmay Jain[2], Dominick Reilly[1], Pranav Balaji[3], Soumyajit Karmakar[4], Shyam Marjit[4], Xiang Li[5], Abhijit Das[3], and Michael S. Ryoo[5]
[1]UNC Charlotte, [2] Delhi Technological University, [3] BITS Pilani Hyderabad
[4] Indian Institute of Information Technology Guwahati, [5] Stony Brook University
sdas24@charlotte.edu

Abstract

Vision Transformers (ViTs) have become ubiquitous in computer vision. Despite their success, ViTs lack inductive biases, which can make it difficult to train them with limited data. To address this challenge, prior studies suggest training ViTs with self-supervised learning (SSL) and fine-tuning sequentially. However, we observe that jointly optimizing ViTs for the primary task and a Self-Supervised Auxiliary Task (SSAT) is surprisingly beneficial when the amount of training data is limited. We explore the appropriate SSL tasks that can be optimized alongside the primary task, the training schemes for these tasks, and the data scale at which they can be most effective. Our findings reveal that SSAT is a powerful technique that enables ViTs to leverage the unique characteristics of both the self-supervised and primary tasks, achieving better performance than typical ViTs pre-training with SSL and fine-tuning sequentially. Our experiments, conducted on 10 datasets, demonstrate that SSAT significantly improves ViT performance while reducing carbon footprint. We also confirm the effectiveness of SSAT in the video domain for deepfake detection, showcasing its generalizability. Our code is available at https://github.com/dominickrei/Limited-data-vits.

1. Introduction

Vision Transformers (ViTs) have become a common sight in computer vision owing to their success across various visual tasks, and are now considered a viable alternative to Convolutional Neural Networks (CNNs). Despite this, ViTs are structurally deficient in inductive bias compared to CNNs, which necessitates training them with large-scale datasets to achieve acceptable visual representation, as noted by Dosovitskiy et al. [13]. As a result, when dealing with small-scale datasets, it is essential to utilize a ViT pre-trained on a large-scale dataset such as ImageNet [12] or JFT-300M [52]. However, in domains such as medical datasets, pre-training ViTs on ImageNet or JFT-300M may not result in an optimal model for fine-tuning on those datasets due to a significant domain gap. Thus, the aim of this research is to address the following question: *how can ViTs be trained effectively in domains with limited data*?

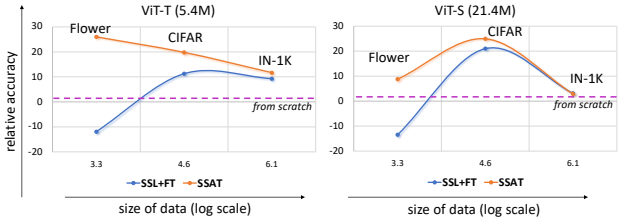

Figure 1. Relative classification accuracy on three datasets with different sizes: (i) Oxford Flower [38] (2K samples), (ii) CIFAR [25] (50K samples), and (iii) ImageNet-1K [12] (IN-1K, 1.2M samples). SSAT consistently outperforms others on all three datasets with two backbones. On the other hand, given the same SSL method, SSL+FT achieves a compromised performance than SSAT, especially on the tiny Oxford Flower dataset (even worse than training from scratch).

Following the introduction of ViTs, second-generation vision transformers have emerged with two different approaches. The first approach is to use a hierarchical structure to introduce inductive bias in ViTs [18, 35, 56]. The second approach involves using hybrid architectures, such as introducing convolutional blocks within ViTs [40, 59]. However, both approaches primarily benefit medium-sized datasets and not small-scale datasets. Several efforts have been made to enhance their locality inductive bias, as reported in literature [15, 28, 29, 34]. Among these methods, SSL has demonstrated exceptional efficacy in training transformers from scratch on small datasets [5, 6, 15, 24, 34, 51, 58]. These methods typically involve sequentially conducting SSL and fine-tuning on the same small dataset to enhance ViT performance.

Meanwhile, another straightforward approach that takes

advantage of SSL is to jointly optimize the self-supervised task along with the primary task like classification or segmentation. We name such SSL tasks as **S**elf-**S**upervised **A**uxiliary **T**ask (**SSAT**). Although SSAT has been explored in the vision community [29, 34, 43] and robotics community [27, 30, 31, 44], there are still many open questions, especially when the size of the dataset is limited.

This paper empirically analyzes the aforementioned joint learning approach with SSAT, as an alternative to sequentially performing SSL and fine-tuning (SSL+FT) on the same dataset. Through an extensive amount of experiments on *ten* image classification datasets of various sizes as well as *two* video classification datasets, surprisingly, we observe that SSAT works significantly better than other baselines like SSL+FT and training from scratch, especially for ViT on small datasets (see Figure 1). Further experiments empirically show that it is most effective when the auxiliary task is image reconstruction from missing pixels among the well known SSL methods we tested. Finally, we perform a detailed model and feature analysis to highlight the unique properties of SSAT-driven models in comparison to other representative baselines. This distinction is particularly notable when comparing with the SSL+FT models which are trained with similar loss functions. We reveal that the advantages of SSAT in a limited-data regime come from better semantic richness, a distinct attention distribution, and an increased capability for feature transformation, which results in higher feature variance.

2. Related Work

Vision Transformers. Several vision transformers [1, 2, 8, 13, 36, 45, 47, 56, 60, 64, 66] have been introduced in recent times for a wide range of tasks. However, these models require large-scale pre-training to be effective on different datasets. In an effort to reduce their reliance on extensive training, DeiT [54] introduced extensive data augmentation, regularization, and distillation tokens from convolutions in ViTs. T2T [63], in a similar vein, employed a tokenization technique that flattened overlapping patches and applied a transformer to allow for learning local structural information around a token. Meanwhile, some ViT models [10, 23, 59] have introduced inductive bias into the transformers through the use of convolutional filters. Hierarchical transformers [14, 33, 35, 57] have introduced inductive bias by reducing the number of tokens through patch merging and thus operating at different scales. However, these architectures do not overcome the limitation of ViTs, which require at least a medium-sized dataset for pre-training [40].

Self-supervised Learning. Self-Supervised Learning (SSL) aims to learn visual representations through pretext tasks. Contrastive methods, such as SimCLR [7] and MoCo [21], minimize the distance between differently augmented views of the same image (positive pairs) while maximizing it for dissimilar images (negative pairs). On the other hand, non-contrastive methods like BYOL [17] and DINO [4] only impose minimization between the positive pairs. In contrast, reconstruction based methods [16, 20, 53, 61] have shown to be effective self-supervised learners for various downstream computer vision tasks. In these methods, an encoder operates on a small portion of an image to learn a latent representation, and a decoder decodes the latent representation to reconstruct the original image in the pixel space. These SSL methods are commonly used for large-scale pre-training of ViTs to enhance their effectiveness in various downstream tasks.

ViTs for small datasets. Liu et al. [34] proposed an auxiliary self-supervised task that improves the robustness of ViT training on smaller datasets. The task involves predicting relative distances among tokens and is jointly trained with primary tasks. On the other hand, Li et al. [29] conducted distillation in the hidden layers of ViT from a lightweight CNN-trained model. To address the lack of locality inductive bias, Lee et al. [28] introduced a ViT architecture with shifted patch tokenization and locality self-attention. Gani et al. [15] proposed an SSL+Fine-tuning methodology where the SSL is similar to the pretext task in DINO [4]. These methods eliminate the need for large-scale pre-training and allow ViTs to learn meaningful representations with limited data. In contrast to these methods, we propose SSAT akin to [34], but with an approach that combines the functionality of self-attention and MLPs through image reconstruction.

3. Preliminaries

ViT utilizes a non-overlapping grid of image patches to process a given image X, where each patch is linearly projected into a set of input tokens. ViT consists of a stack of multi-head attention and linear layers as in [13]. The transformer attention layers model the pairwise relationship between the input tokens [55]. For generalizability, We denote the transformer encoder as f. For brevity, we have omitted the parameters of the encoder. In practice, f operates on an augmented version of the input image X to output a discriminative representation $f(A(X))$ where A is the set of image augmentation. This representation is subsequently classified into class labels using a classifier h. A class-wise cross-entropy loss L_{cls} is used to train the transformer encoder.

4. Self-supervised Auxiliary Task (SSAT)

Our objective is to improve the ViT training on the dataset with limited samples. Consequently, we propose to jointly train the primary classification task of ViT alongside a self-supervised auxiliary task (SSAT). The joint optimization of the SSAT and classification task allows the ViT to capture inductive biases from the data without requiring any

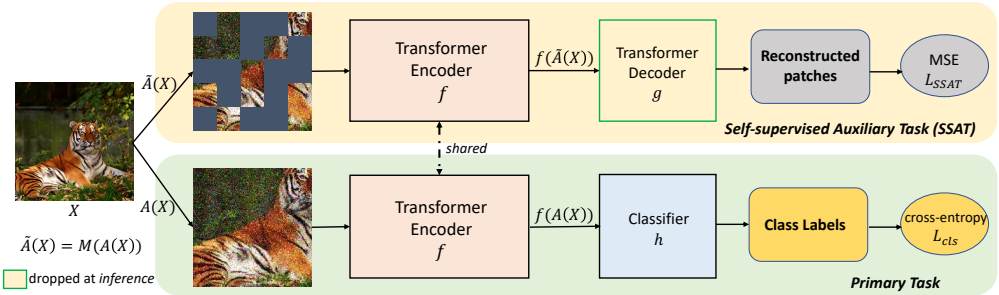

Figure 2. An overview of **ViT training with SSAT**. The input X to the ViT undergoes data augmentations, $A(X)$ and $\tilde{A}(X) = M(A(X))$, using a mask operation M. These augmented inputs are then fed to a Transformer Encoder f, resulting in two latent representations: $f(A(X))$ and $f(\tilde{A}(X))$. These correspond to the masked and full image, respectively. The latent representation of the full image is utilized for the image classification task, while the masked image's representation is used for the image reconstruction task. ViT training involves joint optimization of losses from both these tasks.

additional labels. An overview of our framework is depicted in Figure 2.

In our joint optimization framework for SSAT, we have utilized the widely adopted Masked Autoencoder (MAE) approach [19] for reconstructing the missing pixels. Nonetheless, it is worth noting that any SSL method can be integrated into our framework, given its generic nature. Our decision to use MAE was based on its superior performance, as evidenced by our experimental analysis (Table 4).

To the existing ViT frameworks, where the Transformer encoder f and Classifier h process the full image patches $A(X)$ to compute the classification loss L_{cls}. we introduce an augmentation set $\tilde{A} = M(A(X))$, where operation M randomly masks out patches in the input image X. The transformer encoder f also operates on the unmasked tokens, generating latent representation $f(\tilde{A}(x))$ for these tokens. In parallel to the classifier h , SSAT employs a shallow decoder g to reconstruct back the unseen image pixels from the latent representation of the seen tokens $f(\tilde{A}(x))$. Following [19], the decoder takes as input the latent representation of the seen tokens $f(\tilde{A}(x))$ and a learnable masked token. Each token representation at the decoder's output is linearly projected to a vector of pixel values representing a patch. The output $g(f(\tilde{A}(X)))$ is reshaped to form the reconstructed image, thereafter computing the normalized Mean Square Error (MSE) loss L_{SSAT} between the original and reconstructed image. In practice, the MSE is computed only for the masked patches as in [20].

Thus, the entire framework performs a primary task, i.e. *classification* and a self-supervised auxiliary task, i.e. *reconstruction*. This framework can be jointly optimized using a convex combination of the losses from the primary task and SSAT. Thus, the total loss is computed by

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

λ is the loss scaling factor. During inference, the decoder is discarded and the encoder f processes all input patches to generate the classification output only. Our framework supports training of any ViT model and SSAT variants.

5. Experimental Analysis

In this section, we present the superiority of using SSAT while training any vision transformer. Our experiments are based on image and video classification tasks. We use 12 different datasets: (i) 4 small sized datasets: CIFAR-10 [25], CIFAR-100 [25], Oxford Flowers102 [38] (Flowers) and SVHN [37], (ii) 1 medium sized dataset: ImageNet-1K [12] (IN-1K), (iii) 2 medical datasets: Chaoyang [65] and PMNIST [62], (iv) 3 datasets of DomainNet [42]: ClipArt, Infograph, and Sketch, and (v) 2 video datasets for deepfake detection: DFDC [48] and FaceForensics++ [46].

Our experiments for image reconstruction in the context of SSAT generally follow the procedure outlined in [20], unless otherwise stated. In particular, we employ the decoder design from [20] for ViT and utilize the decoder design from ConvMAE [16] and SimMIM [61] for hierarchical encoders such as CVT and swin, respectively. To optimize hyper-parameters for the decoder, we conduct our experiments with the ViT encoder. For augmentation $\tilde{A}$, we use a random masking with 75% masking ratio. Our decoder has a depth of 2 (i.e. 2 transformer layers) and an embedding dimension of 128. We provide ablations on the choice of these hyper-parameters in Appendix C. It is worth noting that our decoder is shallower than that in MAE [20]. The loss scaling factor λ is set to 0.1 for all the datasets.

Our ViT encoders (f) are trained using the training recipe of DeiT [54], unless otherwise specified. The configuration of ViT-T, ViT-S, and ViT-B is identical to the configuration described in [54]. We borrow the network architecture for CVT-13, and Swin from the official code of [59], and [35], respectively. Training is conducted for 100 epochs, unless otherwise specified, using 8 A5000 24GB GPUs for IN-1K and one A5000 24 GB GPU for all other

Table 1. 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 [54]	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 [54]	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 [59]	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 [35]	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 [22]	25.6	91.78	72.80	46.92	96.45

datasets. Additional training details for each dataset can be found in Appendix B.

5.1. Main Results

SSAT on small-sized dataset: In Table 1, we present the classification accuracy on the small-sized datasets with different variants of vision transformers: ViT-T, ViT-S, CVT-13, and Swin-T. In this table, we demonstrate the impact of using SSAT while training the transformers for learning the class labels. All the models have been trained for 100 epochs from scratch. Although the models with SSAT have more training parameters, they have identical operations during inference. SSAT improves the classification accuracy on all the datasets for all the transformer encoders. It is worth noting that ViT-T with 5.4M parameters when trained with SSAT outperforms ViT-S with 21.4M parameters. The highest classification accuracy is achieved with CVT-13 (20M parameters) due to the introduction of convolutions that infuse inductive bias into the transformers. Although convolutions are generally more effective than transformers on small datasets, our experiments demonstrate that the most effective convolutional network (ResNet-50 [22]) for these datasets underperforms most of the transformers when trained with SSAT, except for ViT-T on CIFAR-10 and CIFAR-100 datasets.

SSAT on medium-sized dataset: In Table 2, we present the impact of SSAT on ViTs that were trained on a medium-sized dataset, such as IN-1K [12]. Our results demonstrate that SSAT consistently enhances the classification accuracy of ViTs, even as the number of training samples increases. Notably, this improvement is more pronounced for smaller models, which have 5.4M parameters, than for larger ones. Specifically, we observed a relative performance improvement of 11.8% for ViT-T with SSAT, as compared to only 2.9% for ViT-S+SSAT. These findings suggest that SSAT can be effectively utilized to train lighter transformers that can be deployed on edge devices.

Does SSAT promotes overfitting? In Table 2, we also analyse the robustness of ViTs to natural corruptions. Given that we recommend the use of SSAT to enhance representation learning in transformer training, it is reasonable to question whether this approach can lead to overfitting on small training samples. To address this concern, we evaluate the performance of our trained models on perturbed versions of the data, specifically, CIFAR-100-p and IN-1K-p, which are obtained by applying random perspective transformations to images following [50]. Our results demonstrate that ViTs trained with SSAT exhibit greater robustness to these natural corruptions compared to the baseline ViTs. We observe notable improvements in performance for tiny ViTs, as evidenced by the results for ViT-T+SSAT in Table 2, as well as for smaller datasets such as CIFAR-100-p.

Comparison of SSAT with SSL+FT: In Table 3, we present the superiority of joint training of the SSL loss with the classification loss over the two-step sequential training approach, where the model is first trained with SSL and then fine-tuned (FT) for classification. Our empirical analysis is conducted on ViT-T, where we compare the performance of ViT trained from scratch and ViT + SSAT, which are trained for 100 epochs. To establish baselines for our SSL+FT model, we conducted experiments using four different training protocols: (1) 50 epochs of SSL training followed by 50 epochs of fine-tuning, (2) 50 epochs of SSL training followed by 100 epochs of fine-tuning, (2) 100 epochs of SSL training followed by 50 epochs of fine-tuning, and (4) 100 epochs of SSL training followed by 100 epochs of fine-tuning. Additionally, we quantify the carbon emission of the models trained using different methods with the help of a tool provided by [26]. Note that the GFLOPs, training time, and Kg CO_2 eq. are specified for the model trained on IN-1K for better generalizability. Our empirical results show that all models incorporating SSL outperform those trained from scratch, highlighting the importance of self-supervised learning when training transformers on small datasets. Moreover, even when requiring an additional 4 hours of training time and resulting in approximately 0.6 Kg CO_2 equivalent of additional carbon emissions, our SSAT models demonstrate superior performance compared to the SSL+FT model (50 epoch

Table 2. Top-1 classification accuracy (%) on ImageNet-1K (IN-1K), perturbed CIFAR-100 (CIFAR-100-*p*), and perturbed ImageNet-1K (IN1K-*p*)

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**

Table 3. Top-1 accuracy and efficiency of ViT-T trained from scratch, with SSL+FT, and with SSAT. We provide the GFLOPs, training time (GPU hours), and CO_2 emissions (kg eq) for IN-1K.

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
(1) SSL+FT	0.43+1.26	85.33	60.43	70.09	55	5.46
(2) SSL+FT	0.43+1.26	86.48	63.28	71.1	82	8.15
(3) SSL+FT	0.43+1.26	85.3	60.3	70.5	74	7.35
(4) 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

Table 4. Top-1 accuracy of existing SSL strategies used as SSAT. MAE as the SSAT achieves the best result on both CIFAR-10 and CIFAR-100.

SSAT (SSL)	CIFAR-10	CIFAR-100
SimCLR [7]	55.21	36.49
DINO [4]	80.07	60.6
MAE [20]	**91.65**	**69.64**

SSL + 100 epoch FT). Although accuracy improves when SSL+FT models are trained on CIFAR-10 and CIFAR-100 for 104 GPU hours, our SSAT approach remains superior, requiring 26 GPU hours less training time and burning approximately 2.8 Kg CO_2 equivalent. However, the SSL+FT model outperforms SSAT when a large amount of training data is available.

Appropriate SSL for joint training: Table 4 presents a comparison of the performance of the SSAT approach, implemented with different SSL strategies, namely, contrastive (SimCLR [7]), non-contrastive (DINO [4]), and reconstruction based (MAE [20]), on the ViT model. Our analysis reveals that the use of SimCLR results in a decrease in the ViT's performance, which can be attributed to the conflicting losses that arise while optimizing the cross-entropy loss to learn class labels and the contrastive loss. However, DINO and MAE both enhance the ViT's performance when jointly trained with cross-entropy. Notably, the improvement observed with MAE is more significant than that with DINO. The superior performance of MAE can be attributed to the centering and sharpening technique employed in DINO, which impedes the learning of class labels while only facilitating the SSL. On the other hand, as mentioned in [41], MAE encourages MLPs in ViTs to be more representative. While the cross-entropy loss primarily contributes more to the self-attention blocks. Thus, SSAT implemented with reconstruction based SSL harmonizes the impact of both tasks, thus improving the ViT's learning capabilities.

Superiority of SSAT over Large-scale pre-training: In situations where training samples are limited and the data distribution differs from that of natural images, large-scale pretraining can be challenging. The main obstacle is the lack of data that accurately represents the downstream data distribution. Consequently, we conducted experiments using ViTs on medical and domain adaptation datasets (Tables 5 and 6) where data is scarce. In Table 5, we demonstrate how SSAT significantly enhances the classification performance of ViT-T on the Chaoyang and PMNIST datasets. The resulting model not only surpasses a comparable ViT model that was pre-trained on ImageNet [12], but also outperforms its larger ViT-S model when trained without SSAT. We observed similar trends of improvement on three datasets from DomainNet [42] in Table 6. It is worth mentioning that our CVT model, when trained using SSAT, outperforms $\mathcal{L}_{drloc}$ [34], which is another state-of-the-art self-supervised loss designed to enhance transformer performance on small datasets.

Table 5. Top-1 accuracy on medical image datasets. All models are trained for 100 epochs.

	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**

Table 6. Top-1 accuracy on DomainNet datasets. All models are trained for 100 epochs

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
+$\mathcal{L}_{drloc}$ [34]	60.64	20.05	57.56
+SSAT	**60.66**	**21.27**	**57.71**

Loss scaling factor: In Figure 3 we perform an empirical analysis to determine the optimal value for the loss scaling factor λ. Our analysis focused on CIFAR datasets show that the choice of $\lambda = 0.1$ is an optimal choice when SSAT positively impacts the primary classification task.

Extended training: In this experiment, we extend the training schedules of both the scratch and SSAT model as illustrated in Figure 4. Our findings indicate that the performance enhancement of our SSAT model, relative to the ViT

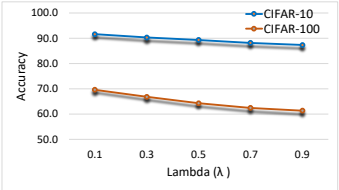

Figure 3. Ablation for loss scaling factor λ.

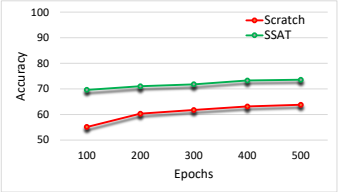

Figure 4. ViT (scratch) vs. ViT+SSAT for longer epochs on CIFAR-100.

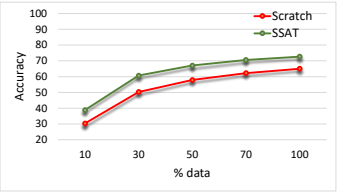

Figure 5. ViT (scratch) vs. ViT+ SSAT for different subset of IN-1K.

baseline, remains consistent throughout the entire training period. These results suggest that the improvement in the SSAT model's performance is not due to a faster convergence rate, but rather to superior optimization capabilities.

Training for different subsets of IN-1K: Figure 5 presents our analysis of the performance of the ViT baseline and SSAT model for varying training sample sizes, specifically on subsets of IN-1K. Our results demonstrate that the performance enhancement of the SSAT model, relative to the baseline model, is consistent across all subsets (i.e., different sizes of the training data). These findings substantiate that models with low training parameters, such as ViT-T, can benefit from SSAT at all scales of training data.

5.2. Diagnosis of features learned by SSAT

In this section, we differentiate the properties of ViTs learned from scratch, SSL+FT, and our SSAT method. We investigate the learned ViT properties by analyzing their attention weights, token representation, feature transformation, and loss landscape. We answer the following key questions:

How are the attention weights distributed? The objective of this experiment is to examine the mean attention weights received from other tokens in a sample in the data distribution. As outlined in [55], the sum of all values in a column of an $n \times n$ self-attention matrix, where n denotes the number of tokens, represents the aggregated attention associated with a token. Figure 6 displays the attention weight distribution across the n tokens for various ViT blocks on both Flower (top row) and CIFAR-100 (bottom row) datasets. The attention weights are uniformly distributed in the first transformer block of the scratch model on both datasets, implying an equal focus on all image regions. However, this distribution changes slightly in the deeper layers. Intriguingly, SSL+FT and SSAT models display sharply peaked attention distributions in the initial and middle transformer layers, but the distributions do not necessarily align with each other. Specifically, in the first transformer block, the attention weight distributions of SSL+FT and SSAT models complement each other, indicating that lower-level features learned by these models are complementary. Moreover, the SSL+FT models exhibit sharp peaks in the final layers, whereas the peaks in SSAT models have a lower magnitude, possibly because the latter model has a better inductive bias. Therefore, although both models are trained on the same set of losses, they use different mechanisms to learn attention weights that differ in the initial layers, and the attention weights learned by the SSAT model are smoother in the final layers, indicating a better inductive bias of the model.

What is the quality of the learned tokens? In this study, we investigated the average distance between tokens within a sample across different transformer blocks. Our analysis involves plotting the average Euclidean distance between tokens in images from the Flower and CIFAR-100 datasets at the output of the transformer layers, as shown in Figure 7. Our results indicate that the scratch model yields a lower inter-token distance than the other models, implying homogeneous token representation. We also observe that SSL+FT models yield higher inter-token distances than SSAT models at the middle transformer layers, but this distance diminishes as we go deeper into the ViTs. Consequently, the SSL+FT models suffer from homogeneous token representation, which adversely affects the ViT training, leading to sub-optimal classification accuracy. In contrast, the inter-token distance of SSAT models increases with ViT depth, indicating that the token representations are discriminative and are semantically rich.

How are representations transformed? The aim of our experiment is to showcase the variation in feature map evolution between ViTs that are trained using different mechanisms. We conducted feature variance measurements across the ViT layers on Flower and CIFAR-100 datasets, and the results are presented in Figure 8. Our analysis confirms the findings of previous studies that the feature variance across ViTs trained from scratch remains constant. However, we observed that the SSL+FT models exhibit an increase in feature variance until a certain layer, after which the rate of an increase either decreases (in Flower dataset) or begins to fall (in CIFAR-100 dataset). Conversely, the feature variance in our SSAT models accumulates with each ViT layer and tends to increase as the depth increases. Consequently, as we go deeper in the SSAT models, the feature map uncertainty decreases, which facilitates optimization through ensembling and stabilizing the transformed feature maps [39].

Why is SSAT better than SSL+FT? In this study, we investigated the loss landscapes of ViT models trained using different training mechanisms. We follow [40] to display the Eigenvalue Spectral Density of Hessian for the different ViT models trained (see Figure 9). Our results indicate

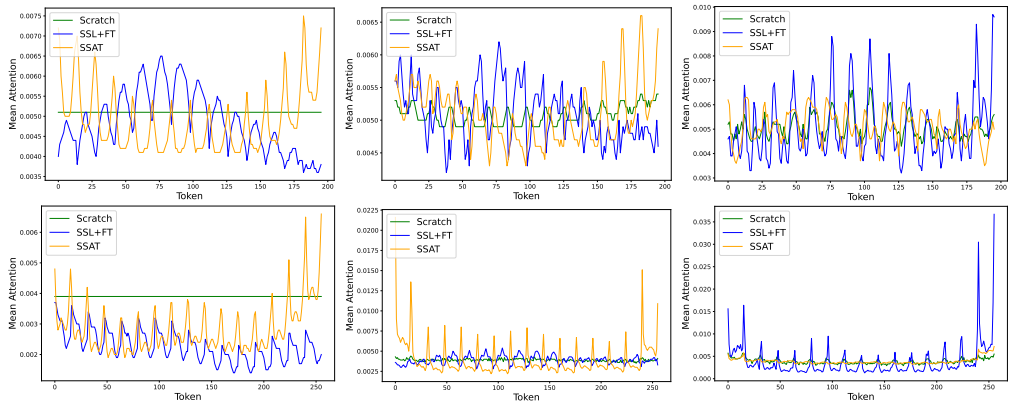

Figure 6. The **distribution of attention weights** across the n tokens for different ViT-T blocks on two datasets: Oxford Flower (top row) and CIFAR-100 (bottom row). The first, second, and third columns correspond to the attention distributions of the first, sixth, and twelfth ViT-T blocks, respectively.

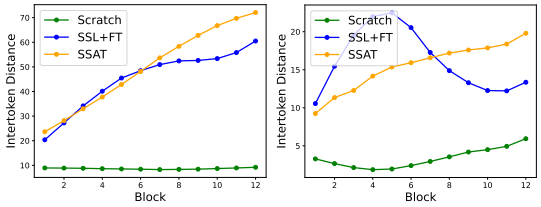

Figure 7. **Average Euclidean Inter-token Distance** of ViTs trained from scratch, using SSL+FT and using SSAT, for two different datasets: Oxford Flowers (on the left) and CIFAR-100 (on the right).

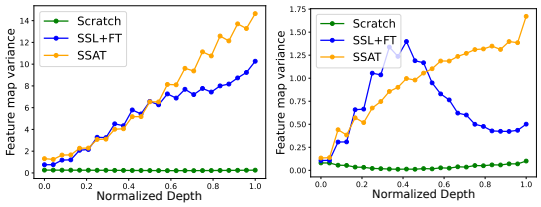

Figure 8. **Feature Map Variance** of ViTs trained from scratch, using SSL+FT and using SSAT, for two different datasets: Oxford Flowers (on the left) and CIFAR-100 (on the right).

that the scratch ViT model exhibits a wide range of negative Hessian eigenvalues, implying non-convex loss landscapes. Interestingly, the number of negative Hessian eigenvalues is slightly higher in the SSL+FT ViT model than in the scratch model (9622 vs 9667). However, the lower magnitude of some of the negative Hessian eigenvalues in the SSL+FT model makes their qualitative visualization difficult. In contrast, SSAT reduces the number of negative Hessian eigenvalues by 12% in comparison to the SSL+FT model. This finding suggests that the SSL approach convexifies losses and suppresses negative eigenvalues in the small data regime. Additionally, the SSAT ViT model reduces the average magnitude of negative Hessian eigenvalues by 70% compared to the SSL+FT models. Therefore, SSAT effectively reduces the magnitude of large Hessian eigenvalues and enhances the ViTs' ability to learn better representations.

Table 7. Comparison of our SSAT to existing state-of-the-art approaches on small datasets. † indicates that [15] is replicated with 300 epochs. Results of [29] is not reported on CIFAR-10.

Method	# enc. params.	epochs	CIFAR-10	CIFAR-100
CVT-13+$\mathcal{L}_{drloc}$ [34]	20M	100	90.30	74.51
CVT-13+ SSAT			**95.93**	**75.16**
ViT (scratch)	2.8M	300	93.58	73.81
SL-ViT [28]			94.53	76.92
ViT† (SSL+FT) [15]			94.2	76.08
ViT + SSAT			**95.1**	**77.8**
DeiT-Ti+$\mathcal{L}_{guidance}$ [29]	6M	300	-	78.15
DeiT-Ti+$\mathcal{L}_{guidance}$ + SSAT			-	**79.46**

5.3. Comparison with the state-of-the-art

Table 7 presents a comparison of SSAT with state-of-the-art (SOTA) methods. To ensure a fair evaluation, we implemented SSAT with the ViT encoder used in the respective methods. We find that MAE as SSAT outperforms Drloc [34] which takes predicting relative distance between patches as SSAT. This shows that the choice of SSAT plays a crucial role in the effective training of ViTs. Moreover, we find that SSAT outperforms SL-ViT [28] and [15] when trained for an equal number of epochs. This indicates that SSAT, without any architectural modifications, can surpass SOTA methods through its joint training strategy. Additionally, we trained a ViT with SSAT and feature-level distillation from a light-weight CNN as described in [29]. The improvement over the baseline [29], which involves training ViT with feature-level distillation only, demonstrates the complementary nature of the representations learned by ViT when trained with SSAT.

In Figure 10, we present the Grad-CAM visualizations [49]. SSL+FT (3rd col) focuses on few specific pixelwise regions, while our method (4th col) focuses on areas corresponding to the entire primary object. We also provide the attention visualization of ViTs trained using different strategies in Appendix E (see Figure 12).

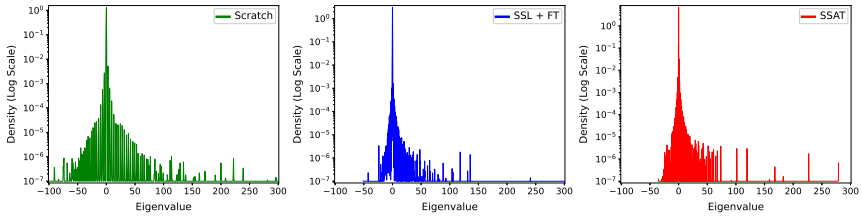

Figure 9. **Hessian max eigenvalues spectra** of ViTs trained from scratch (on the left), SSL + FT (in the middle), and SSAT (on the right).

Table 8. Cross training evaluation and zero-shot transfer results of DeepFake detection on FaceForensics++ with SSAT. [9] is trained on both DFDC and FaceForensics++, thus zero-shot transfer results have not been provided.

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 [9]	82.67	69.89	79.93	64.87	-	-	-	-
DFDC winner [48]	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

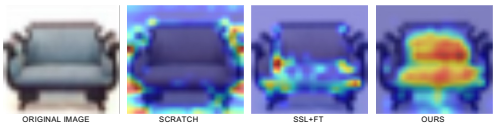

Figure 10. **GradCAM visualizations** of our SSAT model and the representative baselines.

5.4. Performance of SSAT in video domain

We have also assessed the efficacy of the SSAT within the video domain for the task of deepfake detection. In this experiment, the model's generalization capabilities for deepfake detection is validated, as presented in Table 8. For video encoding, we employ ViT as in [53]. Our VideoMAE + SSAT model is a direct extension of the MAE+SSAT model designed for image data; the only modification lies in the choice of encoder. The primary task involves binary classification to distinguish between real and manipulated videos. Notably, we experimented with two masking ratios, 0.75 and 0.95, during the training of VideoMAE + SSAT.

To assess model generalizability, we conducted cross-manipulation training based on the FaceForensics++ dataset [11]. We trained the model using videos generated by all possible combinations of three manipulation techniques (Deepfakes, Face2Face, FaceSwap, and NeuralTextures) plus original videos, and then evaluated its performance on the videos generated by the remaining technique. This approach simulates real-world scenarios where multiple manipulation techniques might be encountered post-training. All models, except the scratch model, are pre-trained on the DFDC dataset [48] before being evaluated on the FaceForensics++ dataset. To enable a fair comparison with VideoMAE + SSAT, the baseline VideoMAE SSL models are first pretrained and fine-tuned on DFDC, and are subsequently employed for deepfake classification task. The evaluation involved both (1) cross-dataset fine-tuning on FaceForensics++ and (2) zero-shot transfer assessment where pre-trained models are evaluated on FaceForensics++ without additional training.

Our findings, as detailed in Table 8, reveal that the VideoMAE+SSAT models demonstrate a superior generalized capability than the other baselines to distinguish between real and manipulated videos. Note that the scratch model outperforms all models on detecting videos generated using NeuralTextures without any pretraining but it is not suitable for zeros-shot transfer. Interestingly, the VideoMAE models exhibit complementary behavior when subjected to different masking ratios, which warrants a future investigation. More details including implementation and training details of these experiments can be found in Appendix D.

6. Conclusion

The main focus of this paper was on the use of self-supervised learning (SSL) to effectively train ViTs on domains with limited data. We demonstrate that by jointly optimizing the primary task of a ViT encoder with SSL as an auxiliary task, we can achieve discriminative representations for the primary task. This simple and easy-to-implement method called SSAT outperforms the traditional approach of sequentially training with SSL followed by fine-tuning on the same data. Our joint training framework learns features that are different from those learned by the dissociated framework, even when using the same losses. These results highlight the potential of SSAT as an effective training strategy with a lower carbon footprint. We anticipate that SSAT will become a standard norm for training vision transformers on small datasets.

Acknowledgments

We thank the members of the Charlotte Machine Learning Lab at UNC Charlotte for valuable discussion. This work is supported by the National Science Foundation (IIS-2245652).

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. VIVIT: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 2

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Int. Conf. on Mach. Learn.*, July 2021. 2

[3] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 14

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 5

[5] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1

[6] Richard J Chen and Rahul G Krishnan. Self-supervised vision transformers learn visual concepts in histopathology. *arXiv preprint arXiv:2203.00585*, 2022. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 5

[8] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv preprint arXiv:2107.06278*, 2021. 2

[9] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pages 219–229. Springer, 2022. 8

[10] Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael Ryoo, and Francois Bremond. MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection. In *CVPR*, 2022. 2

[11] Abhijit Das, Srijan Das, and Antitza Dantcheva. Demystifying attention mechanisms for deepfake detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7. IEEE, 2021. 8

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1, 3, 4, 5, 12

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1, 2

[14] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 2

[15] Hanan Gani, Muzammal Naseer, and Mohammad Yaqub. How to train vision transformer on small-scale datasets? *arXiv preprint arXiv:2210.07240*, 2022. 1, 2, 7

[16] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 2, 3, 12

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[18] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer, 2021. 1

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 3

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3, 5, 12

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[23] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[24] Saarthak Kapse, Srijan Das, and Prateek Prasanna. Cd-net: Histopathology representation learning using pyramidal context-detail network, 2022. 1

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 3, 12

[26] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019. 4

[27] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *Int. Conf. on Mach. Learn.*, pages 5639–5650. PMLR, 2020. 2

[28] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *CoRR*, abs/2112.13492, 2021. 1, 2, 7

[29] Kehan Li, Runyi Yu, Zhennan Wang, Li Yuan, Guoli Song, and Jie Chen. Locality guidance for improving vision transformers on tiny datasets. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 110–127. Springer, 2022. 1, 2, 7

[30] Xiang Li, Varun Belagali, Jinghuan Shang, and Michael S Ryoo. Crossway diffusion: Improving diffusion-based visuomotor policy via self-supervised learning. *arXiv preprint arXiv:2307.01849*, 2023. 2

[31] Xiang Li, Jinghuan Shang, Srijan Das, and Michael S Ryoo. Does self-supervised learning really improve reinforcement learning from pixels? In *Advances in Neural Information Processing Systems*, 2022. 2

[32] Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. *arXiv:2205.10063*, 2022. 12

[33] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 2

[34] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco De Nadai. Efficient training of visual transformers with small datasets. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 5, 7

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, pages 10012–10022, 2021. 1, 2, 3, 4, 12

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 2

[37] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 3, 12

[38] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 1, 3, 12

[39] Namuk Park and Songkuk Kim. Blurs behave like ensembles: Spatial smoothings to improve accuracy, uncertainty, and robustness. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17390–17419. PMLR, 17–23 Jul 2022. 6

[40] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 1, 2, 6

[41] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *International Conference on Learning Representations*. 5

[42] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 3, 5, 12

[43] AJ Piergiovanni, Anelia Angelova, and Michael S. Ryoo. Evolving losses for unsupervised video representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[44] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023. 2

[45] Kanchana Ranasinghe, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Michael S Ryoo. Self-supervised video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2874–2884, 2022. 2

[46] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019. 3, 13

[47] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: Adaptive Space-Time Tokenization for Videos. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[48] Selim Seferbekov. Dfdc 1st place solution, 2020. 3, 8, 13, 14

[49] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 7

[50] Jinghuan Shang, Srijan Das, and Michael S Ryoo. Learning viewpoint-agnostic visual representations by recovering tokens in 3d space. In *Advances in Neural Information Processing Systems*, 2022. 4

[51] Thomas Stegmüller, Antoine Spahr, Behzad Bozorgtabar, and Jean-Philippe Thiran. Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification. *arXiv preprint arXiv:2202.07570*, 2022. 1

[52] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, pages 843–852. IEEE Computer Society, 2017. 1

[53] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 2, 8, 13

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers: Distillation through attention. In *Int. Conf. on Mach. Learn.*, volume 139, pages 10347–10357, July 2021. 2, 3, 4

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2, 6

[56] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVTv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021. 1, 2, 12

[57] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 2

[58] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–195. Springer, 2021. 1

[59] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 1, 2, 3, 4

[60] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 2

[61] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2, 3, 12

[62] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 3, 12

[63] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 2

[64] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021. 2

[65] Chuang Zhu, Wenkai Chen, Ting Peng, Ying Wang, and Mulan Jin. Hard sample aware noise robust learning for histopathology image classification. *IEEE Transactions on Medical Imaging*, 41(4):881–894, 2021. 3, 12

[66] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2