ON IMPROVING ADVERSARIAL TRANSFERABILITY OF VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision transformers (ViTs) process input images as sequences of patches via self-attention; a radically different architecture than convolutional neural networks (CNNs). This makes it interesting to study the adversarial feature space of ViT models and their transferability. In particular, we observe that adversarial patterns found via conventional adversarial attacks show very *low* black-box transferability even for large ViT models. We show that this phenomenon is only due to the sub-optimal attack procedures that do not leverage the true representation potential of ViTs. A deep ViT is composed of multiple blocks, with a consistent architecture comprising of self-attention and feed-forward layers, where each block is capable of independently producing a class token. Formulating an attack using only the last class token (conventional approach) does not directly leverage the discriminative information stored in the earlier tokens, leading to poor adversarial transferability of ViTs. Using the compositional nature of ViT models, we enhance transferability of existing attacks by introducing two novel strategies specific to the architecture of ViT models. *(i) Self-Ensemble:* We propose a method to find multiple discriminative pathways by dissecting a single ViT model into an ensemble of networks. This allows explicitly utilizing class-specific information at each ViT block. *(ii) Token Refinement:* We then propose to refine the tokens to further enhance the discriminative capacity at each block of ViT. Our token refinement systematically combines the class tokens with structural information preserved within the patch tokens. An adversarial attack when applied to such refined tokens within the ensemble of classifiers found in a single vision transformer has significantly higher transferability and thereby brings out the true generalization potential of the ViT's adversarial space. Our code will be publicly released.

1 INTRODUCTION

Transformers compose a family of neural network architectures based on the self-attention mechanism, originally applied in natural language processing tasks achieving state-of-the-art performance (Vaswani et al., 2017; Devlin et al., 2018; Brown et al., 2020). The transformer design has been subsequently adopted for vision tasks (Dosovitskiy et al., 2020), giving rise to a number of successful vision transformer (ViT) models (Touvron et al., 2020; Yuan et al., 2021; Khan et al., 2021). Due to the lack of explicit inductive biases in their design, ViTs are inherently different from convolutional neural networks (CNNs) that encode biases e.g., spatial connectivity and translation equivariance. ViTs process an image as a sequence of patches which are refined through a series of self-attention mechanisms (transformer blocks), allowing the network to learn relationships between any individual parts of the input image. Such processing allows wide receptive fields which can model global context as opposed to the limited receptive fields of CNNs. These significant differences between ViTs and CNNs give rise to a range of intriguing characteristics unique to ViTs (Caron et al., 2021; Tuli et al., 2021; Mao et al., 2021; Paul & Chen, 2021; Naseer et al., 2021b).

Adversarial attacks pose a major hindrance to the successful deployment of deep neural networks in real-world applications. Recent success of ViTs means that adversarial properties of ViT models also become an important research topic. A few recent works explore adversarial robustness of ViTs (Shao et al., 2021; Mahmood et al., 2021; Bhojanapalli et al., 2021) in different attack settings. Surprisingly, these works show that large ViT models exhibit lower transferability in black-box attack setting, despite their higher parameter capacity, stronger performance on clean images, and

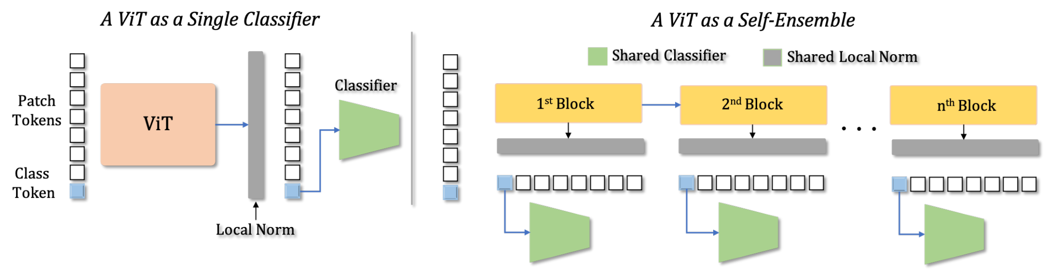


Figure 1: *Left:* Conventional adversarial attacks view ViT as a single classifier and maximize the prediction loss (e.g., cross entropy) to fool the model based on the last classification token only. This leads to sub-optimal results as class tokens in previous ViT blocks only indirectly influence adversarial perturbations. In contrast, our approach (*right*) effectively utilizes the underlying ViT architecture to create a self-ensemble using class tokens produced by all blocks within ViT to design the adversarial attack. Our self-ensemble enables to use hierarchical discriminative information learned by all class tokens. Consequently, an attack based on our self-ensemble generates transferable adversaries that generalize well across different model types and vision tasks.

better generalization (Shao et al., 2021; Mahmood et al., 2021). This finding seems to indicate that as ViT performance improves, its adversarial feature space gets weaker. In this work, we investigate whether the weak transferability of adversarial patterns from high-performing ViT models, as reported in recent works (Shao et al., 2021; Mahmood et al., 2021; Bhojanapalli et al., 2021), is a result of weak features or a weak attack. To this end, we introduce a highly transferable attack approach that augments the current adversarial attacks and increase their transferability from ViTs to the unknown models. Our proposed transferable attack leverages two key concepts, multiple discriminative pathways and token refinement, which exploit unique characteristics of ViT models.

Our approach is motivated by the modular nature of ViTs (Touvron et al., 2020; Yuan et al., 2021; Mao et al., 2021): they process a sequence of input image patches repeatedly using multiple multi-headed self-attention layers (transformer blocks) (Vaswani et al., 2017). We refer to the representation of patches at each transformer block as *patch tokens*. An additional randomly initialized vector (*class token*[1]) is also appended to the set of patch tokens along the network depth to distill discriminative information across patches. The collective set of tokens is passed through the multiple transformer blocks followed by passing of the class token through a linear classifier (head) which is used to make the final prediction. The class token interacts with the patch tokens within each block and is trained gradually across the blocks until it is finally utilized by the linear classifier head to obtain class-specific logit values. The class token can be viewed as extracting information useful for the final prediction from the set of patch tokens at each block. Given the role of the class token in ViT models, we observe that class tokens can be extracted from the output of each block and each such token can be used to obtain a class-specific logit output using the final classifier of a pretrained model. This leads us to the proposed *self-ensemble* of models within a single transformer (Fig. 1). We show that attacking such a self-ensemble (Sec. 3) containing multiple discriminative pathways significantly improves adversarial transferability from ViT models, and in particular from the large ViTs.

Going one step further, we study if the class information extracted from different intermediate ViT blocks (of the self-ensemble) can be enhanced to improve adversarial transferability. To this end, we introduce a novel *token refinement* module directed at enhancing these multiple discriminative pathways. The token refinement module strives to refine the information contained in the output of each transformer block (within a single ViT model) and aligns the class tokens produced by the intermediate blocks with the final classifier in order to maximize the discriminative power of intermediate blocks. Our token refinement exploits the structural information stored in the patch tokens and fuses it with the class token to maximize the discriminative performance of each block.

Both the *refined tokens* and *self-ensemble* ideas are combined to design an adversarial attack that is shown to significantly boost the transferability of adversarial examples, thereby bringing out the true generalization of ViTs' adversarial space. Through our extensive experimentation, we empirically demonstrate favorable transfer rates across different model families (convolutional and transformer) as well as different vision tasks (classification, detection and segmentation).

[1]Average of patch tokens can serve as a class token in our approach for ViT designs that do not use an explicit class token such as Swin transformer (Liu et al., 2021) or MLP-Mixer (Tolstikhin et al., 2021)

2 Background and Related Work

Adversarial Attack Modeling: Adversarial attack methods can be broadly categorized into two categories, *white-box* attacks and *black-box* attacks. While the white-box attack setting provides the attacker full access to the parameters of the target model, the black-box setting prevents the attacker from accessing the target model and is therefore a harder setting to study adversarial transferability.

White-box Attack: Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) and Projected Gradient Descent (PGD) (Madry et al., 2018) are two initially proposed white-box attack methods. FGSM corrupts the clean image sample by taking a single step within a small distance (perturbation budget ϵ) along the objective function's gradient direction. PGD corrupts the clean sample for multiple steps with a smaller step size, projecting the generated adversarial example onto the ϵ-sphere around the clean sample after each step. Other state-of-the-art white-box attack methods include Jacobian-based saliency map attack (Papernot et al., 2016), Sparse attack (Modas et al., 2019), One-pixel attack (Su et al., 2019), Carlini and Wagner optimization (Carlini & Wagner, 2017), Elastic-net (Chen et al., 2018), Diversified sampling (Tashiro et al., 2020), and more recently Auto-attack (Croce & Hein, 2020b). We apply white-box attacks on surrogate models to find perturbations that are then transferred to black-box target models.

Black-box Attack and Transferability: Black-box attacks generally involve attacking a source model to craft adversarial signals which are then applied on the target models. While gradient estimation methods that estimate the gradients of the target model using black-box optimization methods such as Finite Differences (FD) (Chen et al., 2017; Bhagoji et al., 2018) or Natural Evolution Strategies (NES) (Ilyas et al., 2018; Jiang et al., 2019) exist, these methods are dependent on multiple queries to the target model which is not practical in most real-world scenarios. In the case of adversarial signal generation using source models, it is possible to directly adopt white-box methods. In our work, we adopt FGSM and PGD in such a manner. Methods like (Dong et al., 2018) incorporate a momentum term into the gradient to boost the transferability of existing white-box attacks, building attacks named MIM. In similar spirit, different directions are explored in literature to boost transferability of adversarial examples; *a) Enhanced Momentum:* Lin *et al.* (Lin et al., 2019) and Wang *et al.* (Wang & He, 2021) improve momentum by using Nesterov momentum and variance tuning respectively during attack iterations, *b) Augmentations:* Xie *et al.* (Xie et al., 2019) showed that applying differentiable stochastic transformations can bring diversity to the gradients and improve transferability of the existing attacks, *c) Exploiting Features:* Multiple suggestions are proposed in the literature to leverage the feature space for adversarial attack as well. For example, Zhou *et al.* (Zhou et al., 2018) incorporate the feature distortion loss during optimization. Similarly, (Inkawhich et al., 2020b;a; Huang et al., 2019) also exploit intermediate layers to enhance transferability. However, combining the intermediate feature response with final classification loss is non-trivial as it might require optimization to find the best performing layers (Inkawhich et al., 2020b;a), and *d) Generative Approach:* Orthogonal to iterative attacks, generative methods (Poursaeed et al., 2018; Naseer et al., 2019; 2021a) train an autoencoder against the white-box model. In particular, Naseer *et al.* show that transferability of an adversarial generator can be increased with relativistic cross-entropy (Naseer et al., 2019) and augmentations (Naseer et al., 2021a). Ours is the first work to address limited transferability of ViT models.

The Role of Network Architecture: Recent works exploit architectural characteristics of networks to improve the transferability of attacks. While Wu et al. (2020) exploit skip connections of models like ResNets and DenseNets to improve black-box attacks, Guo et al. (2020) build on similar ideas focused on the linearity of models.

Our work similarly focuses on unique architectural characteristics of ViT models to generate more transferable adversarial perturbations with the existing white-box attacks.

Robustness of ViTs: Adversarial attacks on ViT models are relatively unexplored. Shao et al. (2021) and Bhojanapalli et al. (2021) investigate adversarial attacks and robustness of ViT models studying various white-box and black-box attack techniques. The transferability of perturbations from ViT models is thoroughly explored in (Mahmood et al., 2021) and they conclude that ViT models do not transfer well to CNNs, whereas we propose a methodology to solve this shortcoming. Moreover, Mahmood et al. (2021) explores the idea of an ensemble of CNN and ViT models to improve the transferability of attacks. Our proposed ensemble approach explores a different direction by converting a single ViT model into a collection of models (self-ensemble) to improve attack

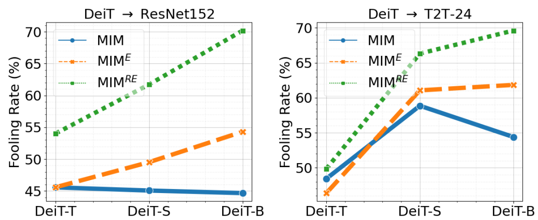


Figure 2: Adversarial examples for ViTs have only moderate transferability. In fact transferabililty (%) of MIM (Dong et al., 2018) perturbations to target models goes down as the source model size increases such as from DeiT-T (Touvron et al., 2020) (5M parameters) to DeiT-B (Touvron et al., 2020) (86M parameters). However, the performance of the attack improves significantly when applied on our proposed ensemble of classifiers found within a ViT (MIM^E & MIM^{RE}).

transferability. In essence, our proposed method can be integrated with existing attack approaches to take full advantage of the ViTs' learned features and generate transferable adversaries.

3 Enhancing Adversarial Transferability of ViTs

Preliminaries: Given a clean input image sample $\boldsymbol{x}$ with a label y, a source ViT model $\mathcal{F}$ and a target model $\mathcal{M}$ which is under-attack, the goal of an adversarial attack is generating an adversarial signal, $\boldsymbol{x}'$, using the information encoded within $\mathcal{F}$, which can potentially change the target network's prediction ($\mathcal{M}(\boldsymbol{x}')_{argmax} \neq y$). A set of boundary conditions are also imposed on the adversarial signal to control the level of distortion in relation to the original sample, i.e., $\|\boldsymbol{x} - \boldsymbol{x}'\|_p < \epsilon$, for a small perturbation budget ϵ and a p-norm, often set to infinity norm (ℓ_∞).

Motivation: The recent findings (Shao et al., 2021; Mahmood et al., 2021) demonstrate low black-box transferability of ViTs despite their higher parametric complexity and better feature generalization. Motivated by this behaviour, we set-up a simple experiment of our own to study the adversarial transferability of ViTs (see Fig. 2). We note that transferability of adversarial examples found via momentum iterative fast gradient sign method (Dong et al., 2018) (MIM) at $\ell_\infty \leq 16$ on DeiT (Touvron et al., 2020) does not increase with model capacity. In fact, adversarial transferability from DeiT base model (DeiT-B) on ResNet152 and large vision transformer (ViT-L (Dosovitskiy et al., 2020)) is lower than DeiT tiny model (DeiT-T). This is besides the fact that DeiT-B has richer representations and around 17$\times$ more parameters than DeiT-T. We investigate if this behavior is inherent to ViTs or merely due to a sub-optimal attack mechanism. To this end, we exploit unique architectural characteristics of ViTs to first find an ensemble of networks within a single pretrained ViT model (self-ensemble, right Fig. 1). The class token produced by each self-attention block is processed by the final local norm and classification MLP-head to refine class-specific information (Fig. 2). In other words, our MIM^E and MIM^{RE} variants attack class information stored in the class tokens produced by *all* the self-attention blocks within the model and optimize for the adversarial example (Sec. 3.1 and 3.2). Exploring the adversarial space of such multiple discriminative pathways in a self-ensemble generates highly transferable adversarial examples, as we show next.

3.1 Self-Ensemble: Discriminative Pathways of Vision Transformer

A ViT model (Dosovitskiy et al., 2020; Touvron et al., 2020), $\mathcal{F}$, with n transformer blocks can be defined as $\mathcal{F} = (f_1 \circ f_2 \circ f_3 \circ \dots f_n) \circ g$, where f_i represents a single ViT block comprising of multi-head self-attention and feed-forward layers and g is the final classification head. To avoid notation clutter, we assume that g consists of the final local norm and MLP-head (Touvron et al., 2020; Dosovitskiy et al., 2020). Self-attention layer within the vision transformer model takes a sequence of m image patches as input and outputs the processed patches. We will refer to the representations associated with the sequence of image patches as patch tokens, $P_t \in \mathbb{R}^{m \times d}$ (where d is the dimensionality of each patch representation). Attention in ViT layers is driven by minimizing the empirical risk during training. In the case of classification, patch tokens are further appended with the class token ($Q_t \in \mathbb{R}^{1 \times d}$). These patch and class tokens are refined across multiple blocks (f_i) and attention in these layers is guided such that the most discriminative information from patch tokens is preserved within the class token. The final class token is then projected to the number of classes by the classifier, g. Due to the availability of class token at each transformer block, we can create an ensemble of classifiers by learning a shared classification head at each block along the ViT hierarchy. This provides us an ensemble of n classifiers from a single ViT, termed as the *self-ensemble*:

$$\mathcal{F}_k = \left(\prod_{i=1}^{k} f_i \right) \circ g, \quad \text{where } k = 1, 2, \dots, n. \tag{1}$$

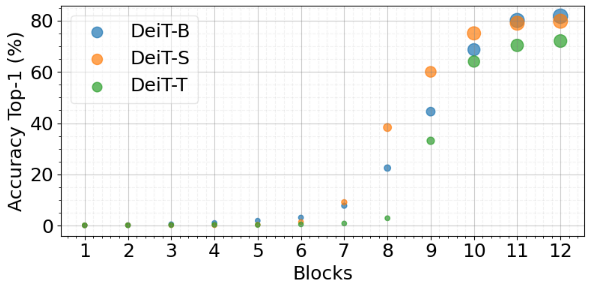


Figure 3: Distribution of discriminative information across blocks of DeiT models. Note how multiple intermediate blocks contain features with considerable discriminative information as measured by top-1 accuracy on the ImageNet val set. These are standard models pretrained on ImageNet with no further training. Each block (x-axis) corresponds to a classifier $\mathcal{F}_k$ as defined in Equation 1.

We note that the multiple classifiers thus formed hold significant discriminative information. This is validated by studying the classification performance of each classifier (Eq. 1) in terms of top-1 (%) accuracy on ImageNet validation set, as demonstrated in Fig. 3. Note that multiple intermediate layers perform well on the task, especially towards the end of the ViT processing hierarchy.

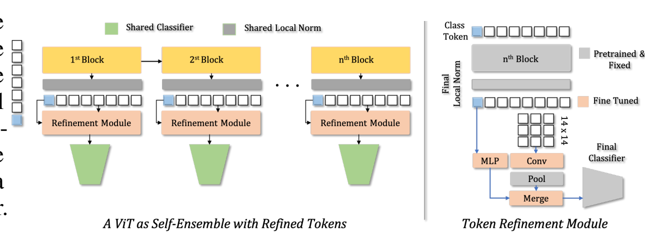
For an input image $\boldsymbol{x}$ with label y, an adversarial attack can now be optimized for the ViT's self-ensemble by maximizing the loss at each ViT block. However, we observe that initial blocks (1-6) for all considered DeiT models do not contain useful discriminative information as their classification accuracy is almost zero (Fig. 3). During the training of ViT models (Touvron et al., 2020; Yuan et al., 2021; Mao et al., 2021), parameters are updated based on the last class token only, which means that the intermediate tokens are not directly aligned with the final classification head, g in our self-ensemble approach (Fig. 3) leading to a moderate classification performance. To resolve this, we introduce a token refinement strategy to align the class tokens with the final classifier, g, and boost their discriminative ability, which in turn helps improve attack transferability. We explain the refinement process below.

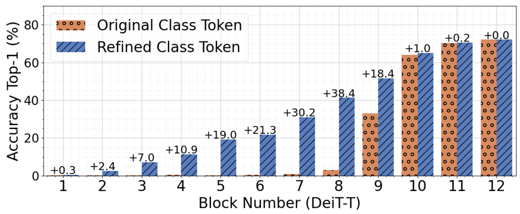
3.2 Token Refinement

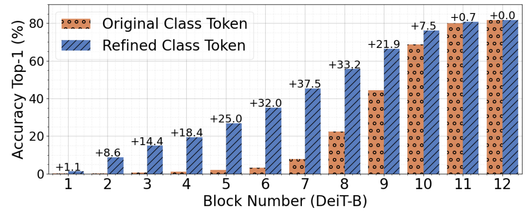
As mentioned above, the multiple discriminative pathways within a ViT give rise to an ensemble of classifiers (Eq. 1). However, the class token produced by each attention layer is being processed by the final classifier, g. This puts an upper bound on classification accuracy for each token which is lower than or equal to the accuracy of the final class token. Our objective is to push the accuracy of the class tokens in intermediate blocks towards the upper bound as defined by the last token. For this purpose, we introduced a token refinement module to fine-tune the class tokens.

Our proposed token refinement module is illustrated in Fig. 4. It acts as an intermediate layer inserted between the outputs of each block (after the shared norm layer) and the shared classifier head. Revisiting our baseline ensemble method (Fig. 1), we note that the shared classifier head contains weights directly trained only on the outputs of the last transformer block. While the class tokens of previous layers may be indirectly optimized to align with the final classifier, there exists a potential for misalignment of these features with the classifier: the pretrained classifier (containing weights compatible with the last layer class token) may not extract all the useful information from the previous layers. Our proposed module aims to solve this misalignment by refining the class tokens in a way such that the shared (pretrained) classifier head is able to extract all discriminative information contained within the class tokens of each block. Moreover, intermediate patch tokens may contain additional information that is not at all utilized by the class tokens of those blocks, which would also be addressed by our proposed block. Therefore, we extract both patch tokens and the class token from each block and process them for refinement, as explained next.

Figure 4: Recent ViTs process 196 image patches, leading to 196 patch tokens. We rearranged these to create a 14x14 feature grid which is processed by a convolutional block to extract structural information, followed by average pooling to create a single patch token. Class token is refined via a MLP layer before feeding to the classifier. Both tokens are subsequently merged.







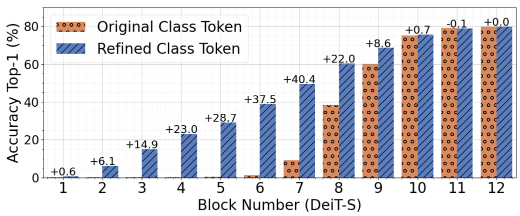


Figure 5: **Self-Ensemble for DeiT (Touvron et al., 2020):** We measure the top-1 accuracy on ImageNet using the class-token of each block and compare to our refined tokens. These results show that fine-tuning helps align tokens from intermediate blocks with the final classifier enhancing their classification performance. Thus token refinement leads to strengthened discriminative pathways allowing more transferable adversaries.

−Patch Token Refinement: One of the inputs to the token refinement module is the set of patch tokens output from each block. We first rearrange these patch tokens to regain their spatial relationships. The aim of this component within the refinement module is to extract information relevant to spatial structure contained within the intermediate patch tokens. We believe that significant discriminative information is contained within these patches. The obtained rearranged patch tokens are passed through a convolution block (standard ResNet block containing a skip connection) to obtain a spatially aware feature map, which is then average pooled to obtain a single feature vector (of same dimension as the class token). This feature vector is expected to extract all spatial information from patch tokens.

−Class Token Refinement: By refining the class tokens of each block, we aim to remove any misalignment between the existing class tokens and the shared (pretrained) classifier head. Also, given how the class token does not contain a spatial structure, we simply use a linear layer to refine it. We hypothesize that refined class token at each block would be much more aligned with the shared classifier head allowing it to extract all discriminative information contained within those tokens.

−Merging Patch and Class Token: We obtain the refined class token and the patch feature vector (refined output of patch tokens) and sum them together to obtain a merged token. While we tested multiple approaches for merging, simply summing them proved sufficient.

−Training: Given a ViT model containing k transformer blocks, we plugin k instances of our token refinement module to the output of each block as illustrated in Figure 4. We obtain the pretrained model, freeze all existing weights, and train only the k token refinement modules for only a single epoch on ImageNet training set. We used SGD optimizer with learning rate set to 0.001. Training finishes in less than one day on a single GPU-V100 even for a large ViT model such as DeiT-B (Touvron et al., 2020).

As expected, the trained token refinement module leads to increased discriminability of the class tokens, which we illustrate in Figure 5. Note how this leads to significant boosting of discriminative power especially in the earlier blocks, solving the misalignment problem. We build on this enhanced discriminability of the ensemble members towards better transferability, as explained next.

3.3 Adversarial Transfer

Our modifications to ViT models with respect to multiple discriminative pathways and token refinement are exploited in relation to adversarial transfer. We consider black-box attack perturbations are generated using a source (surrogate) ViT model. The source model is only pretrained on ImageNet, modified according to our proposed approach and is subsequently fine-tuned to update only the token refinement module for a single epoch. We experiment with multiple white-box attacks, generating the adversarial examples using a joint loss over the outputs of each block. The transferability of adversarial examples is tested on a range of CNN and ViT models. Given input sample $\boldsymbol{x}$ and its label y, the adversarial object for our self-ensemble (Eq. 1) for the untargeted attack is defined as,

$$\max_{\boldsymbol{x}'} \sum_{i=1}^{k} [\![\mathcal{F}_k(\boldsymbol{x}')_{argmax} \neq y]\!], \quad \text{s.t. } \|\boldsymbol{x}-\boldsymbol{x}'\|_p \leq \epsilon, \;\; k \in \{1, 2, \dots, n\} \tag{2}$$

Source (↓)	Attack	Convolutional			Transformers				
		Res152	WRN	DN201	ViT-L	T2T-24	TnT	ViT-S	T2T-7
Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014)									
VGG19$_{bn}$	FGSM	28.56	33.92	33.22	13.18	10.78	12.96	25.08	29.90
MNAS	FGSM	39.82	40.10	44.34	16.60	22.56	25.82	34.10	48.96
Deit-T	FGSM	37.10	38.86	42.40	44.38	35.42	50.58	73.32	57.62
	FGSM^E	39.60	41.42	45.58	48.34	35.08	51.00	80.74	62.82
	FGSM^{RE}	39.82$_{(+2.7)}$	41.26$_{(+2.4)}$	46.06$_{(+3.7)}$	46.76$_{(+2.4)}$	32.68$_{(-2.7)}$	48.00$_{(-2.6)}$	80.10$_{(+6.8)}$	63.90$_{(+6.3)}$
Deit-S	FGSM	31.04	33.58	36.28	36.40	33.4	41.00	58.78	43.48
	FGSM^E	38.38	41.06	46.00	47.20	39.00	51.44	78.90	56.70
	FGSM^{RE}	43.86$_{(+12.8)}$	46.26$_{(+12.7)}$	51.88$_{(+15.6)}$	47.92$_{(+11.5)}$	39.86$_{(+6.5)}$	55.7$_{(+14.7)}$	82.00$_{(+23.2)}$	66.20$_{(+22.7)}$
Deit-B	FGSM	31.58	33.86	34.96	30.50	27.84	33.08	50.24	40.50
	FGSM^E	41.46	43.02	47.12	42.28	35.40	46.22	73.04	57.32
	FGSM^{RE}	45.74$_{(+14.2)}$	48.46$_{(+14.6)}$	52.64$_{(+17.7)}$	41.68$_{(+11.2)}$	36.60$_{(+8.8)}$	49.60$_{(+16.5)}$	74.40$_{(+24.2)}$	40.50$_{(+0.0)}$
Projected Gradient Decent (PGD) (Madry et al., 2018)									
VGG19$_{bn}$	PGD	28.56	33.92	33.22	5.94	10.78	12.96	13.08	29.90
MNAS	PGD	36.28	36.22	40.20	8.04	18.04	21.16	19.60	41.70
Deit-T	PGD	23.98	24.16	26.76	35.70	21.54	44.24	86.86	53.74
	PGD^E	24.58	25.46	28.38	39.84	21.86	45.08	88.44	53.80
	PGD^{RE}	34.64$_{(+10.7)}$	37.62$_{(+13.5)}$	40.56$_{(+13.8)}$	58.60$_{(+22.9)}$	26.58$_{(+5.0)}$	55.52$_{(+11.3)}$	96.34$_{(+9.5)}$	66.68$_{(+12.9)}$
Deit-S	PGD	24.96	26.38	30.38	37.84	33.46	60.62	84.38	52.76
	PGD^E	27.72	29.54	32.90	44.30	35.40	64.76	89.82	52.76
	PGD^{RE}	38.92$_{(+14.0)}$	42.84$_{(+16.5)}$	46.82$_{(+16.4)}$	60.86$_{(+23.0)}$	40.30$_{(+6.8)}$	76.10$_{(+15.5)}$	97.32$_{(+12.9)}$	71.54$_{(+18.8)}$
Deit-B	PGD	25.56	27.90	30.24	34.08	31.98	52.76	69.82	39.80
	PGD^E	32.84	35.40	38.66	43.56	37.82	64.20	82.32	51.68
	PGD^{RE}	49.10$_{(+23.5)}$	53.38$_{(+25.5)}$	56.96$_{(+26.7)}$	56.90$_{(+22.8)}$	45.70$_{(+13.7)}$	79.56$_{(+26.8)}$	94.10$_{(+24.3)}$	74.78$_{(+35.0)}$

Table 1: **Fool rate (%)** on 5k ImageNet val. adversarial samples at $\epsilon \leq 16$. Perturbations generated from our proposed self-ensemble with refined tokens from a vision transformer have significantly higher success rate.

where $[\![\cdot]\!]$ is an indicator function. In the case of target attack, the attacker optimizes the above objective towards a specific target class instead of an arbitrary misclassification.

4 Experiments

We conduct thorough experimentation on a range of standard attack methods to establish the performance boosts obtained through our proposed transferability approach. We create ℓ_∞ adversarial attacks with $\epsilon \leq 16$ and observe their transferability by using the following protocols:

Source (white-box) models: We mainly study three vision transformers from DeiT (Touvron et al., 2020) family due to their data efficiency. Specifically, the source models are Deit-T, DeiT-S, and DeiT-B (with 5, 22, and 86 million parameters, respectively). They are trained without CNN distillation. Adversarial examples are created on these models using an existing white-box attack (e.g., FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2018) and MIM (Dong et al., 2018)) and then transferred to the black-box target models.

Target (black-box) models: We test on black-box transferability across several vision tasks including classification, detection and segmentation. We consider convolutional networks ResNet152 (Res152) (He et al., 2016), Wide-ResNet-50-2 (WRN) (Zagoruyko & Komodakis, 2016), DenseNet201 (DN201) (Huang et al., 2017) and other ViT models including Token-to-Token transformer (T2T) (Yuan et al., 2021), Transformer in Transformer (TnT) (Mao et al., 2021), DINO (Caron et al., 2021), and Detection Transformer (DETR) (Carion et al., 2020) as the black-box target models.

Datasets: We use ImageNet training set to fine tune our proposed token refinement modules. For evaluating robustness, we selected 5k samples from ImageNet validation set such that 5 random samples from each class that are correctly classified by ResNet50 and ViT small (ViT-S) (Dosovitskiy et al., 2020) are present. In addition, we conduct experiments on COCO (Lin et al., 2014) (5k

Source (↓)	Attack	Convolutional			Transformers				
		Res152	WRN	DN201	ViT-L	T2T-24	TnT	ViT-S	T2T-7
Momemtum Iterative Fast Gradient Sign Method (MIM) (Dong et al., 2018)									
$VGG19_{bn}$	MIM	46.98	54.04	57.32	12.80	21.84	25.72	28.44	47.74
MNAS	MIM	54.34	55.40	64.06	18.88	34.54	38.70	40.58	60.02
Deit-T	MIM	45.56	47.86	53.26	63.84	48.44	72.52	96.44	77.66
	MIM^E	45.58	47.98	54.50	67.16	46.38	71.02	97.74	78.02
	MIM^{RE}	$54.02_{(+8.5)}$	$58.48_{(+10.6)}$	$63.00_{(+9.7)}$	$79.12_{(+15.3)}$	$49.86_{(+1.4)}$	$77.80_{(+5.3)}$	$99.14_{(+2.7)}$	$85.50_{(+7.8)}$
Deit-S	MIM	45.06	47.90	52.66	63.38	58.86	79.56	94.22	68.00
	MIM^E	49.52	52.98	58.40	71.78	61.06	84.42	98.12	74.58
	MIM^{RE}	$61.72_{(+16.7)}$	$65.10_{(+17.2)}$	$71.74_{(+19.1)}$	$84.30_{(+20.9)}$	$66.32_{(+7.5)}$	$92.02_{(+12.5)}$	$99.42_{(+5.2)}$	$89.08_{(+21.1)}$
Deit-B	MIM	44.66	47.98	52.14	57.48	54.40	70.84	69.82	59.34
	MIM^E	54.30	58.34	63.32	70.42	61.84	82.80	82.32	73.66
	MIM^{RE}	$70.18_{(+25.5)}$	$74.08_{(+26.1)}$	$79.12_{(+27.0)}$	$81.28_{(+23.8)}$	$69.6_{(+15.2)}$	$92.20_{(+21.4)}$	$94.10_{(+24.3)}$	$89.72_{(+30.4)}$
MIM with Input Diversity (DIM) (Xie et al., 2019)									
$VGG19_{bn}$	DIM	62.08	68.30	73.48	16.86	30.16	34.70	35.42	58.62
MNAS	DIM	62.08	68.30	73.48	25.06	42.92	47.24	52.74	71.98
Deit-T	DIM	68.30	70.06	77.18	62.00	70.16	82.68	89.16	72.68
	DIM^E	70.06	69.84	78.00	66.38	72.30	85.98	93.72	90.78
	DIM^{RE}	$70.78_{(+2.5)}$	$70.78_{(+0.7)}$	$78.40_{(+1.2)}$	$67.58_{(+5.6)}$	$68.56_{(-1.6)}$	$84.18_{(+1.5)}$	$93.36_{(+4.2)}$	$78.40_{(+5.7)}$
Deit-S	DIM	62.12	63.42	67.30	62.62	73.84	79.50	82.32	67.3
	DIM^E	74.44	75.34	80.14	76.22	84.10	91.92	94.92	88.42
	DIM^{RE}	$81.30_{(+19.18)}$	$82.64_{(+19.22)}$	$86.98_{(+19.68)}$	$78.88_{(+16.3)}$	$85.26_{(+11.4)}$	$93.22_{(+13.7)}$	$96.56_{(+14.2)}$	$86.98_{(+19.7)}$
Deit-B	DIM	59.14	60.64	64.44	61.38	69.54	73.96	76.32	64.44
	DIM^E	78.36	80.28	83.70	79.06	85.10	91.84	94.38	86.96
	DIM^{RE}	$84.92_{(+25.8)}$	$86.36_{(+25.7)}$	$89.24_{(+24.8)}$	$78.90_{(+17.5)}$	$84.00_{(+14.5)}$	$92.28_{(+18.3)}$	$95.26_{(+18.9)}$	$89.24_{(+24.8)}$

Table 2: **Fool rate (%)** on 5k ImageNet val. adversarial samples at $\epsilon \leq 16$. Perturbations generated from our proposed self-ensemble with refined tokens from a vision transformer have significantly higher success rate.

images) and PASCAL-VOC12 (Everingham et al., 2012) (around 1.2k images) validation set to test the cross-task transferability.

Evaluation Metrics: We report fooling rate (percentage of samples for which the predicted label is flipped after adding adversarial perturbations) to evaluate classification. In the case of object detection, we report the decrease in mean average precision (mAP) and for automatic segmentation, we use the popular Jaccard Index. Given the pixel masks for the prediction and the ground-truth, it calculates the ratio between the pixels belonging to intersection and the union of both masks.

Baseline Attacks: We show consistent improvements for single step fast gradient sign method (FGSM) (Goodfellow et al., 2014) as well as iterative attacks including PGD (Madry et al., 2018), MIM (Dong et al., 2018) and input diversity (transformation to the inputs) (DIM) (Xie et al., 2019) attacks. Iterative attacks ran for 10 iterations and we set transformation probability for DIM to default 0.7 (Xie et al., 2019). Our approach is not limited to specific attack settings, but existing attacks can simply be adopted to our self-ensemble ViTs with refined tokens (refer appendices A, B C, D, E, and F for extensive experimental analysis with more datasets and attacks).

4.1 Classification

In this section, we discuss the experimental results on adversarial transferability across black-box classification models. For a given attack method *'Attack'*, we refer *'Attack^E'* and *'Attack^{RE}'* as self-ensemble and self-ensemble with refined tokens, respectively, which are the two variants of our approach. We observe that adversarial transferability from ViT models to CNNs is only moderate for conventional attacks (Tables 1 & 2). For example, perturbations found via iterative attacks from DeiT-B to Res152 has even lower transfer than $VGG19_{bn}$. However, the same attacks when applied using our proposed ensemble strategy (Eq. 1) with refined tokens consistently showed improved transferability to other convolutional as well as transformer based models. Strength of our method is also evident by blockwise fooling rate in white-box setting (Fig. 6). It is noteworthy how MIM

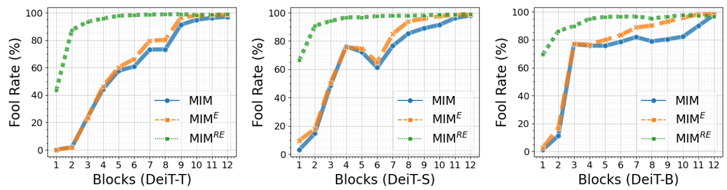


Figure 6: **Ablative Study**: Fooling rate of intermediate layers under MIM (white-box) attack using our self-ensemble approach. We obtain favorable improvements for our method.

Source (→)	DeiT-T		DeiT-S		DeiT-B	
`No Attack`	`MIM`	`MIM`^{RE}	`MIM`	`MIM`^{RE}	`MIM`	`MIM`^{RE}
38.5	24.0	19.7	23.7	19.0	22.9	16.9
	`DIM`	`DIM`^{RE}	`DIM`	`DIM`^{RE}	`DIM`	`DIM`^{RE}
	20.5	13.7	20.3	12.0	19.9	11.1

Table 3: **Cross-Task Transferability** (*classification→detection*) Object Detector DETR (Carion et al., 2020) is fooled. mAP at [0.5:0.95] IOU on COCO val set. Our self-ensemble approach with refined token (RE) significantly improves cross-task transferability. (*lower the better*)

Source (→)	DeiT-T		DeiT-S		DeiT-B	
`No Attack`	`MIM`	`MIM`^{RE}	`MIM`	`MIM`^{RE}	`MIM`	`MIM`^{RE}
42.7	32.5	31.6	32.5	31.0	32.6	30.6
	`DIM`	`DIM`^{RE}	`DIM`	`DIM`^{RE}	`DIM`	`DIM`^{RE}
	31.9	31.4	31.7	31.3	32.0	31.0

Table 4: **Cross-Task Transferability** (*classification→segmentation*) DINO (Caron et al., 2021) is fooled. Jaccard index metric is used to evaluate segmentation performance. Lower is better. Best adversarial transfer results are achieved using our method. (*lower the better*)

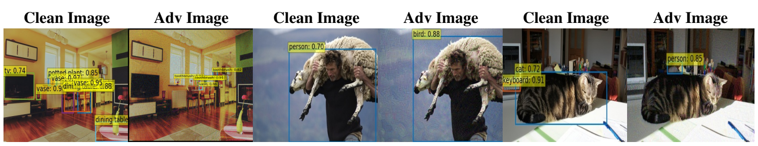


Figure 7: Visualization of DETR failure cases for our proposed DIM^{RE} attack generated from DeiT-S source model. *(best viewed in zoom)*

fails to fool the initial blocks of ViT, while our approach allows the attack to be as effective in the intermediate blocks as for the last class token. This ultimately allows us to fully exploit ViT's adversarial space leading to high transfer rates for adversarial perturbations (refer appendices A, B C, D, E, and F for extensive experimental analysis with more datasets and attacks).

4.2 Cross-Task Transferability

Self-attention is the core component of transformer architecture regardless of the task; classification (Dosovitskiy et al., 2020; Touvron et al., 2020; Yuan et al., 2021; Mao et al., 2021), object detection (Carion et al., 2020), or unsupervised segmentation (Caron et al., 2021). We explore the effectiveness of our proposed method on two additional tasks: object detection (DETR) (Carion et al., 2020) and segmentation (DINO) (Caron et al., 2021). We select these methods considering the use of transformer modules employing the self-attention mechanism within their architectures. While the task of object detection contains multiple labels per image and involves bounding box regression, the unsupervised model DINO is trained in a self-supervised manner with no traditional image-level labels. Moreover, DINO uses attention maps of a ViT model to generate pixel-level segmentations, which means adversaries must disrupt the entire attention mechanism to degrade its performance.

We generate adversarial signals on source models with a classification objective using their initial predictions as the label. In evaluating attacks on detection and segmentation tasks at their optimal setting, the source ViT need to process images of different sizes (e.g., over 896×896 pix for DETR).

To cater for this, we process images in parts (refer appendix G) which allows generation of stronger adversaries. The performance degradation of DETR and DINO on generated adversaries are summarised in Tables 3 & 4. For DETR, we obtain clear improvements. In the more robust DINO model, our transferability increases well with the source model capacity as compared to the baseline.

5 Conclusion

We identify a key shortcoming in current state-of-the-art approaches for adversarial transferability of vision transformers (ViTs) and show the potential for much strong attack mechanisms that would exploit the architectural characteristics of ViTs. Our proposed novel approach involving multiple discriminative pathways and token refinement is able to fill in these gaps, achieving significant performance boosts when applied over a range of state-of-the-art attack methods.

Reproducibility Statement: Our method simply augments the existing attack approaches and we used open source implementations. We highlight the steps to reproduce all of the results presented in our paper, **`a) Attacks:`** We used open source implementation of patchwise attack (Gao et al., 2020) and Auto-Attack (Croce & Hein, 2020b) (refer Appendix B) with default setting. Wherever necessary, we clearly mention attack parameters, e.g., iterations for PGD (Madry et al., 2018), MIM (Dong et al., 2018) and DIM (Xie et al., 2019) in section 4 (Experiments: Baseline Attack). Similarly, transformation probability for DIM is set to the default value provided by the corresponding authors that is 0.7, **`b) Refined Tokens:`** We fine tuned class tokens for the pretrained source models (used to create perturbations) using open source code base (`https://github.com/pytorch/examples/tree/master/imagenet`). We provided training details for fine tuning in section 3.2. Further, we will publicly release all the models with refined tokens, **`c) Cross-Task Attack Implementation:`** We provided details in section 4.2 and pseudo code in appendix G for cross-task transferability (from classification to segmentation and detection), and **`d) Dataset:`** We describe the procedure of selecting subset (5k) samples from ImageNet val. set in section 4. We will also release indices of these samples to reproduce the results.

Ethics Statement: Since our work focuses on improving adversarial attacks on models, in the short run our work can assist various parties with malicious intents of disrupting real-world deployed deep learning systems dependent on ViTs. However, irrespective of our work, the possibility of such threats emerging exists. We believe that in the long run, works such as ours will support further research on building more robust deep learning models that can withstand the kind of attacks we propose, thus negating the short term risks. Furthermore, a majority of the models used are pre-trained on ImageNet (ILSVRC'12). We also conduct our evaluations using this dataset. The version of ImageNet used contains multiple biases that portray unreasonable social stereotypes. The data contained is mostly limited to the Western world, and encodes multiple gender / ethnicity stereotypes Yang et al. (2020) while also posing privacy risks due to unblurred human faces. In future, we hope to use the more recent version of ImageNet Yang et al. (2021) which could address some of these issues.

REFERENCES

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 2020. 14, 15

Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *ECCV*, 2018. 3

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. *ArXiv*, abs/2103.14586, 2021. 1, 2, 3

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020. 7, 9

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *S&P*, 2017. 3

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 1, 7, 9

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *AISec*, 2017. 3

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *AAAI*, 2018. 3

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pp. 2196–2205. PMLR, 2020a. 14, 15

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020b. 3, 10, 14, 15, 16

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 3, 4, 7, 8, 10, 14, 17, 18, 19

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 4, 7, 9

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012. 8

Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Hengtao Shen. Patch-wise attack for fooling deep neural network. In *European Conference on Computer Vision*, 2020. 10, 14, 15

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2014. 3, 7, 8

Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. In *Advances in Neural Information Processing Systems*, 2020. 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 7

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017. 7

Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4733–4742, 2019. 3

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018. 3

Nathan Inkawhich, Kevin Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=rJxAo2VYwr. 3

Nathan Inkawhich, Kevin J Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *arXiv preprint arXiv:2004.14861*, 2020b. 3

Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. Black-box adversarial attacks on video recognition models. In *ACM MM*, 2019. 3

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021. 1

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 14

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019. 3

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014. 7

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. *arXiv preprint arXiv:2103.14030*, 2021. 2, 14, 16

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 3, 7, 8, 10, 14, 16, 17

Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. *ArXiv*, abs/2104.02610, 2021. 1, 2, 3, 4

Yuxin Mao, Jing Zhang, Zhexiong Wan, Yuchao Dai, Aixuan Li, Yunqiu Lv, Xinyu Tian, Deng-Ping Fan, and Nick Barnes. Transformer transforms salient object detection and camouflaged object detection. *arXiv preprint arXiv:2104.10127*, 2021. 1, 2, 5, 7, 9

Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Sparsefool: a few pixels make a big difference. In *CVPR*, 2019. 3

Muzammal Naseer, Salman H Khan, Harris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 2019. 3

Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. On generating transferable targeted perturbations. *arXiv preprint arXiv:2103.14641*, 2021a. 3, 14

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers, 2021b. 1

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*, pp. 722–729. IEEE, 2008. 14

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *EuroS&P*, 2016. 3

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners, 2021. 1

Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4422–4431, 2018. 3

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *ArXiv*, abs/2103.15670, 2021. 1, 2, 3, 4

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. In *IEEE Transactions on Evolutionary Computation*. IEEE, 2019. 3

Yusuke Tashiro, Yang Song, and Stefano Ermon. Output diversified initialization for adversarial attacks. *arXiv preprint arXiv:2003.06878*, 2020. 3

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 2

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1, 2, 4, 5, 6, 7, 9, 14, 15

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision?, 2021. 1

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 1, 2

Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. *arXiv preprint arXiv:2103.15571*, 2021. 3

Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *ICLR*, 2020. 3

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019. 3, 8, 10, 14, 17, 18, 19

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *faacct*, pp. 547–558, 2020. 10

Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. *arXiv preprint arXiv:2103.06191*, 2021. 10

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 1, 2, 5, 7, 9, 15

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 7

Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 452–467, 2018. 3