Boosting Generative Adversarial Transferability with Self-supervised Vision Transformer Features

Shangbo Wu[1] Yu-an Tan[1] Ruinan Ma[1] Wencong Ma[2] Dehua Zhu[1] Yuanzhang Li[2]*
[1]School of Cyberspace Science and Technology, Beijing Institute of Technology
[2]School of Computer Science and Technology, Beijing Institute of Technology
{shangbo.wu, tan2008, ruinan, wencong.ma, zhudehua, popular}@bit.edu.cn

Abstract

The ability of deep neural networks (DNNs) come from extracting and interpreting features from the data provided. By exploiting intermediate features in DNNs instead of relying on hard labels, we craft adversarial perturbation that generalize more effectively, boosting black-box transferability. These features ubiquitously come from supervised learning in previous work. Inspired by the exceptional synergy between self-supervised learning and the Transformer architecture, this paper explores whether exploiting self-supervised Vision Transformer (ViT) representations can improve adversarial transferability. We present **`dSVA`***—a generative dual self-supervised ViT features attack, that exploits both global structural features from contrastive learning (CL) and local textural features from masked image modeling (MIM), the self-supervised learning paradigm duo for ViTs. We design a novel generative training framework that incorporates a generator to create black-box adversarial examples, and strategies to train the generator by exploiting joint features and the attention mechanism of self-supervised ViTs. Our findings show that CL and MIM enable ViTs to attend to distinct feature tendencies, which, when exploited in tandem, boast great adversarial generalizability. By disrupting dual deep features distilled by self-supervised ViTs, we are rewarded with remarkable black-box transferability to models of various architectures that outperform state-of-the-arts. Code available at https://github.com/spencerwooo/dSVA.*

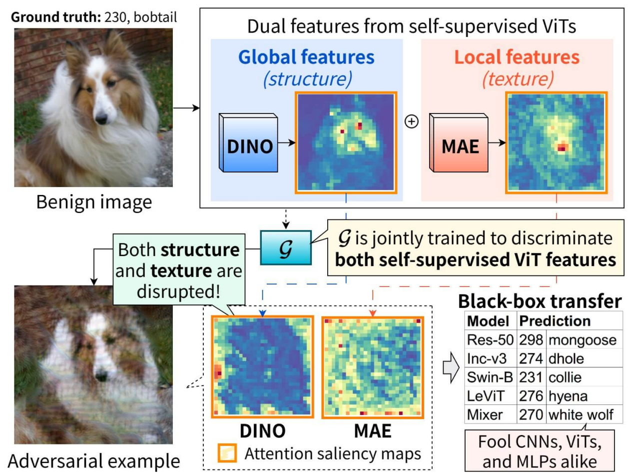

Figure 1. **Demonstration of `dSVA`.** By jointly exploiting deep features of both self-supervised ViTs, i.e., DINO (CL) and MAE (MIM), `dSVA` crafts perturbation that disrupts both structural and textural representations of the image (as visualized in the attention saliency maps), fooling ConvNets, ViTs, and MLPs alike.

1. Introduction

The transferability of adversarial examples enable real-world black-box attacks on DNNs without the adversary's access to their internals. Such attacks require the construction of a local white-box surrogate model. Consequently, their effectiveness relies on the ability to disrupt the shared latent representations, i.e., features, learnt by both models. DNNs learn sample-label correlations over their training process, by identifying the structure and semantic characteristics of the data for classification. These learnt deep features are generalizable enough to essentially serve as the basis that drive downstream tasks such as object detection [5, 47], similarity measurement [18, 72], image super-resolution [35], and style transfer [19]. Prior research has shown that improving transferability is possible by targeting intermediate features of the surrogate model instead of directly attacking hard labels or output gradients [28, 60, 73]. Since deep features of well-trained DNNs are generalizable [69], perturbation designed to disrupt these features are more transferable [29].

The habitual inclusion of label-wise loss in existing work for conducting adversarial attacks acts as a common practice that pushes the surrogate model to be setup with supervised learning. This makes sense for ConvNets where self-supervised learning lags behind supervised. However,

*corresponding author

the advent of ViTs introduced the success of self-supervision in natural language processing to vision [4, 7, 9, 25]. Supervised learning fails to preserve image semantics through human labelling, reducing feature-rich semantic information within images into a single concept represented by a human-assigned category. In contrast, self-supervised ViTs excel at capturing semantics, providing robust positional and semantic relationships throughout model layers, outperforming ConvNets [1]. Driven by the powerful adversarial potentials of self-supervised ViT features, we ask: *How can we fully utilize the rich representations distilled by the harmonious coalition between self-supervision and the Transformer architecture, to boost adversarial transferability?* We attempt to answer this research question in threefold:

(1) Facet-level feature exploitation. ViTs comprise several layers of multi-head self-attention blocks that encode token-wise features. With a goal of extracting adversarially generalizable features, contrary to ConvNets where existing work use the direct output of entire intermediate layers, we propose to extract internal components, i.e. feature facets, of self-attention blocks in ViTs: queries, keys, and values.

(2) Self-attention exploitation. The architectural design of self-attention empowers ViTs to capture semantic context of the image at a high level. We propose, atop the adversarial exploitation of internal facets in ViT blocks, to systematically extract saliency maps from the self-attention mechanism itself, and integrate them into loss optimization as dense semantic guides to identify valuable feature targets.

(3) Joint self-supervision feature discrimination. Two branches of self-supervision paradigms exist for ViTs: contrastive learning (CL) and masked image modeling (MIM). Comparative studies show that CL captures global structural shapes and semantics, while MIM focuses more on local textural details [44]. We hypothesis that, if combined, both aspects will complement each other in generalizability that jointly contribute to enhancing adversarial transferability.

Incorporating all three aspects, we introduce **`dSVA`**—a generative dual self-supervised ViT features attack. We introduce a novel generative training framework, consisting of a generator to craft transferable adversarial perturbation, and discriminative training approaches to jointly exploit the dual intricate features—both structural and textural—distilled by the two types of self-supervised ViTs. We choose the duo: DINO [7] and MAE [25], for CL and MIM respectively. Figure 1 showcases a birds-eye view of `dSVA`.

Leveraging the powerful latent representations distilled by self-supervised ViTs, `dSVA` achieves outstanding adversarial effectiveness. We show an example in Fig. 1 of `dSVA` successfully disrupting both structural features from DINO (CL) and textural representations from MAE (MIM) (visualized in the attention maps), enabling impressive transferability towards black-box models of distinct architectures. Our experiments demonstrate `dSVA`'s outstanding transferability to models across ViTs, ConvNets, and MLPs alike, and its ability to evade defenses, surpassing various state-of-the-arts.

To conclude, we summarize our contributions as follows.

- We present **`dSVA`**, a generative adversarial attack, that crafts highly transferable black-box adversarial examples by exploiting dual self-supervised ViT features.
- We first aim at, instead of attacking the direct output of intermediate layers, targeting the internal facets of the self-attention blocks in ViTs, namely, the queries, keys, and values, to take advantage of the Transformer architecture and extract generalizable and transferable features.
- We further introduce a method to exploit the self-attention mechanism itself by extracting saliency maps from the self-attention maps of ViTs, acting as guides for important feature targets, providing, in essence, a regularization scheme that enable boosted adversarial generalizability.
- We finally propose to jointly exploit the two self-supervised learning schemes—CL and MIM—to craft perturbation that attend to and disrupt both global structural shapes and local textural details from within the image.

2. Related Work

Generative adversarial attacks is initially introduced in Poursaeed et al. [45] to address both sample-agnostic and sample-specific adversarial perturbation. This approach paved the way for generative methods in creating unrestricted perturbations [52] and utilizing GANs [64]. The generative strategy has further proven to be beneficial for transferability, where Naseer et al. [41] developed CDA for cross-domain attacks, Nakka and Salzmann [40] incorporated mid-level features, and Zhang et al. [71] presented BIA for generating cross-domain perturbation with only knowledge from ImageNet. We follow this foundational generative approach in our work. Other studies refine the generator to improve *targeted* attack effectiveness [17, 61, 68] or introduce *outside knowledge* from foundation models trained on web-scale datasets [67]. We do not consider them as our competitors.

Self-supervised learning has enjoyed its remarkable success in natural language processing, particularly with wide applications in modern language models [14, 46]. In vision tasks, although several self-supervised techniques have been developed for ConvNets [6, 22, 24], it is with ViTs that the self-supervised learning strategy, through both CL [7–9, 43] and MIM [2, 4, 25, 66], has truly excelled. Self-supervised ViTs have shown to encode rich features that carry incredible capabilities out-of-the-box, often surpassing comparable methods that require additional supervised fine-tuning [1, 16, 49]. In this work, we propose to jointly exploit the dual aspects of features provided in CL and MIM for crafting generalizable adversarial perturbation with superior transferability. Note that we choose to use DINO [7] instead of DINOv2 [43] for fair comparison, as DINOv2 is trained on a far larger dataset than vanilla ImageNet.

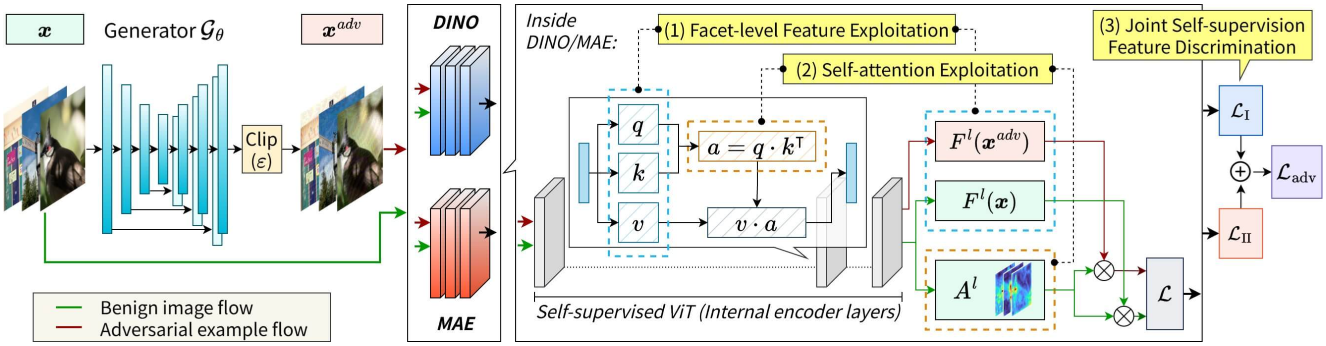

Figure 2. **The dSVA Training Framework.** Sample x is fed through $\mathcal{G}$ to create adversarial example x^{adv}, which are then both fed into the self-supervised models DINO and MAE, to extract deep representations and attention saliency maps from both global structural and local textural feature aspects. The feature discriminative loss is derived from both ViTs, which jointly form the adversarial loss $\mathcal{L}_\text{adv}$.

3. Methodology

3.1. Threat Model

We consider the standard ℓ_∞ threat model. Given a DNN classifier $\mathcal{F}(\cdot) : x \in \mathbb{R}^m \mapsto y$ where x is a benign sample and y denotes its ground truth label. The adversary aims to create an adversarial example $x^{adv} = x + \delta$, with perturbation δ restricted by an ℓ_p-ball (ℓ_∞ in our case), such that $\mathcal{F}(x^{adv}) \neq y$. We incorporate a generator $\mathcal{G}_\theta$ to craft x^{adv} by discriminating the latent intermediate features of the self-supervised ViTs as

$$\theta^* \leftarrow \arg\max_{\theta} \mathcal{D}\left(F(x), F(x^{adv})\right), \; s.t. \; \|\delta\|_\infty \leq \varepsilon, \quad (1)$$

where $x^{adv} = \mathcal{G}_\theta(x)$, $F(\cdot)$ extracts the self-supervised ViT features from an image, and $\mathcal{D}(\cdot,\cdot)$ measures the feature distance. We now present our proposed dSVA for the training of the adversarial generator $\mathcal{G}_\theta$.

3.2. Facet-level Feature Exploitation

Previous arts have highlighted the strong transferability potential of feature-space adversarial perturbation, but they focus on *supervised ConvNets*. In this work, we first explore the rich features offered by the harmonic combination of self-supervision and the Transformer architecture.

Irrespective of training strategy, ViTs process images in the same manner. The input image is divided into n non-overlapping patches $\{p_i\}$ ($i \in [1, n]$) and linearly projected onto a D-dimensional latent space. Positional embeddings and the `[CLS]` token are added thereafter, forming a set of tokens to be fed through L layers of transformer encoders. Each encoder block comprises alternating layers of multi-head self-attention (MSA) and MLP blocks, with LayerNorm (LN) applied before each block. We denote the output token sequence at layer l as $T^l = \{t_0^l, t_1^l, \cdots, t_n^l\}$.

If we were to follow previous practice, we would directly use intermediate encoder layer outputs, i.e., tokens, as the feature representation. In contrast, the Transformer architecture encodes features within MSA blocks that offer better generalizability. At each layer l, the MSA block encodes tokens from the previous layer T^{l-1} into *queries*, *keys*, and *values*, i.e., $q_i^l = w_q^l \cdot t_i^{l-1}$, $k_i^l = w_k^l \cdot t_i^{l-1}$, and $v_i^l = w_v^l \cdot t_i^{l-1}$ (with w^l being the weights), which are fused back into T^l. Therefore, each image patch p_i corresponds to a set of *deep features* at the facet-level, namely $\{q_i^l, k_i^l, v_i^l, t_i^l\}$, with each representing its internal query, key, value, and the final output as a fused token at layer l. In ViTs, the *query* is the part of input the model is focusing on, whereas the *key* is then compared with the *query* to determine the attention. They are then aggregated into the *value* vector for feature concatenation. Facets *key* and *query* are directly associated with the input, inherently providing high quality, less noisy features that favor generalizability. We later empirically investigate the impact of facet selection to adversarial effectiveness.

As in Fig. 2, to train $\mathcal{G}_\theta$ for perturbation generation, dSVA is designed to deviate the latent representations of a benign image and its generated adversarial example, that is, to minimize the cosine similarity between the deep features extracted. In this way, the crafted perturbation would be able to *neutralize* critical decisive low-level features within the sample, thereby misleading black-box DNNs. Hence, the discriminative loss at this stage is formulated as

$$\theta^* \leftarrow \arg\min_{\theta} \mathcal{D}_\text{cos}\left(F^l\left(x\right), F^l\left(x^{adv}\right)\right), \quad (2)$$

where $F^l(\cdot)$ gives one of q^l, k^l, v^l, t^l as the target facet-level feature extracted at layer l within the ViT $\mathcal{F}(\cdot)$. At inference time, the trained generator $\mathcal{G}_{\theta^*}$ crafts adversarial example x^{adv} within perturbation budget as

$$x^{adv} = \text{clip}(\mathcal{G}_{\theta^*}(x), \varepsilon). \quad (3)$$

3.3. Self-attention Exploitation

Caron et al. [7] revealed that the attention heads of self-supervised ViTs attend to salient foreground regions in an

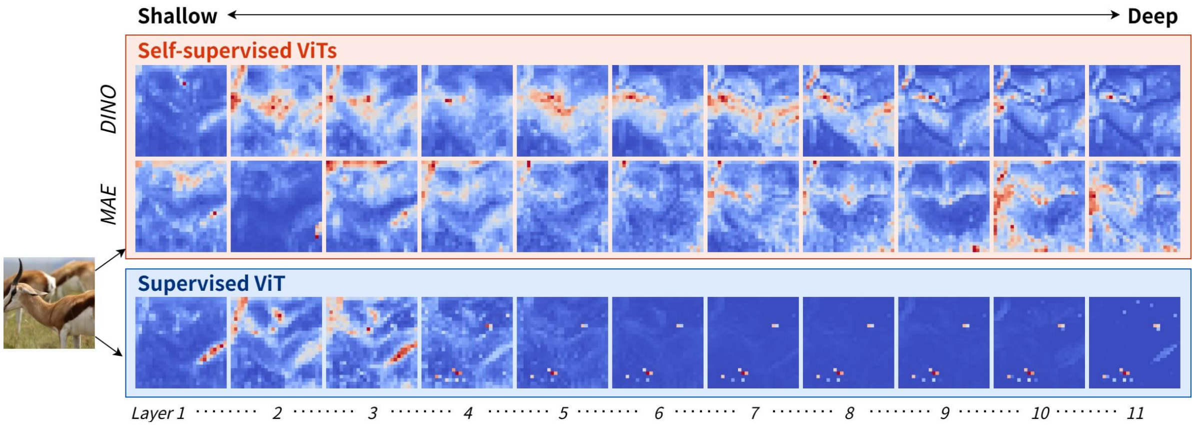

Figure 3. **Attention saliency maps.** We visualize the attention saliency maps derived from both self-supervised ViTs DINO (first row) and MAE (second row), and the supervised ViT (third row). From left to right, layer depth increase from shallow to deep (from 1 to 11).

image, and Amir et al. [1] further demonstrated that these encoded features represent powerful learnt common ground across images. As such, we propose an incremental regularization to leverage saliency maps derived from the self-attention mechanism of pretrained ViTs as *feature landmarks*, so as to offer additional guidance to target more impactful features during optimization in dSVA.

We first extract the self-attention maps for benign sample $\boldsymbol{x}$ at layer l, i.e., the attention weights associated with each head of each token attending to every other token, denoted as A^l. Next, we select the attention weights from the [CLS] token to all other tokens across all heads as

$$A^l_{\texttt{[CLS]}} = A^l[:, :, 0, 1:]. \tag{4}$$

The attention saliency map S^l at layer l is calculated as the mean attention from the [CLS] token to all other tokens over each attention head as

$$S^l = \frac{1}{H} \sum_{h=1}^{H} A^l_{\texttt{[CLS]}}[h], \tag{5}$$

where H is the number of attention heads. Thus, S^l serves as a *feature landmark guidance* for targeting intermediate features, regularizing the global semantic knowledge learnt by the generator. We apply a scaling factor of γ to S^l for loss optimization. Building on Eq. (2), loss function $\mathcal{L}$ at this stage is thus formulated as

$$\mathcal{L} = \arg\min_{\boldsymbol{\theta}} \mathcal{D}_{\cos} \left(F^l \left(\boldsymbol{x} \right) \odot \left(\gamma \cdot S^l \right), F^l \left(\boldsymbol{x}^{adv} \right) \odot \left(\gamma \cdot S^l \right) \right). \tag{6}$$

Shown in Fig. 3 is the attention saliency maps extracted from ViTs with self-supervision (red background) vs. supervision (blue background), as well as the variance of saliency maps with increasing layer depth from left to right. Compared to the saliency maps extracted from a supervised ViT, those from the self-supervised ViTs DINO (first row) and MAE (second row) are less noisy and capture various levels of global and local semantics, respectively. From shallow to deep layers, the self-supervised representations favor less spatial information and more textural information, whereas the supervised ViT's representations collapse into homogeneous primitive patterns. These visualizations showcase the powerful representations offered only by the self-attention of *self-supervised ViTs*, acting as feature landmarks to be integrated in dSVA for transferability boosts.

3.4. Joint Self-supervision Feature Discrimination

Recall that two branches of self-supervision strategies currently stand for ViTs: CL and MIM. Reflected in both learnt latent representations and self-attention favoritism, CL better captures global long-range shape-wise features by learning globally projected representations to discriminate each other, while MIM focuses more on local textural details as it is a generative task that predicts masked regions. We hypothesis that features derived from CL and MIM would complement each other from an adversarial perspective. Therefore, we propose to jointly exploit both feature aspects in dSVA to disrupt structure-biased and texture-biased image features, thereby enhancing adversarial transferability.

To this end, we jointly train generator $\mathcal{G}_{\boldsymbol{\theta}}$ against both CL and MIM ViTs, i.e., DINO and MAE. The final loss function $\mathcal{L}_{\text{adv}}$ is thus formulated as

$$\mathcal{L}_{\text{adv}} = \lambda \cdot \mathcal{L}_{\text{I}} + (1 - \lambda) \cdot \mathcal{L}_{\text{II}}, \tag{7}$$

where $\mathcal{L}_{\text{I}}$ and $\mathcal{L}_{\text{II}}$ are derived as in Eq. (6) from DINO and MAE, respectively. Doing so, dSVA is able to craft highly transferable perturbation that targets both structural and textural image features, greatly boosting transferability across various black-box models with diverse architectures.

Attack	VGG-16	Res-50	Den-121	Eff-B0	Inc-v3	Inc-v4	Swin-B	MaxViT	PiT-B	Visformer	LeViT	Mixer
CDA (VGG-19)	**99.31**	69.23	59.19	76.38	52.94	61.96	16.53	14.63	9.48	32.40	29.79	23.02
CDA (Res-152)	92.98	88.88	87.02	75.32	63.85	74.97	11.82	7.78	5.86	39.03	35.85	22.78
CDA (Den-169)	92.98	87.63	**97.03**	90.96	67.59	78.94	26.88	22.41	20.98	69.67	65.11	52.01
BIA (VGG-19)	97.58	74.32	84.93	77.77	66.63	76.96	19.35	15.25	12.46	34.68	35.96	27.53
BIA (Res-152)	94.94	**92.52**	86.47	65.11	62.46	81.37	22.18	17.32	11.40	45.55	29.15	29.60
BIA (Den-169)	93.67	86.07	95.49	81.17	75.40	71.78	17.36	9.44	10.65	32.71	44.47	38.98
CDA (ViT-B/16)	92.75	74.32	90.10	87.23	81.82	82.25	**62.13**	33.09	**59.74**	78.05	85.20	80.63
BIA (ViT-B/16)	52.93	21.83	33.77	32.13	31.55	34.62	8.89	5.50	6.39	17.81	27.34	40.68
MI (ViT-B/16)	52.59	32.33	47.85	52.34	38.07	35.61	49.69	31.02	42.92	47.31	43.51	65.16
PNA (ViT-B/16)	46.49	33.99	42.68	50.64	37.97	36.05	50.84	35.68	46.96	51.04	51.49	74.30
TGR (ViT-B/16)	54.89	35.14	51.60	57.02	37.54	40.35	51.15	34.02	45.26	50.72	46.38	79.78
ATT (ViT-B/16)	60.41	40.85	56.55	64.47	43.32	44.43	59.10	40.15	51.12	58.80	56.02	82.52
dSVA (DINO)	86.54	57.59	83.17	88.51	78.50	78.61	33.05	21.27	35.04	72.67	67.41	78.81
dSVA (MAE)	94.36	78.07	86.36	84.04	77.75	79.71	47.38	31.85	33.55	63.25	64.32	56.64
dSVA (Joint)	96.78	81.70	94.83	**95.32**	**89.73**	**91.73**	59.83	**41.29**	50.48	**81.37**	**85.21**	**85.38**

Table 1. **Comparison of black-box transferability.** We showcase the black-box fooling rate (%) of dSVA and compared baseline attacks, against target black-box models with various architectures, including a total of 6 ConvNets, 5 ViTs, and an MLP-Mixer.

4. Experiments

4.1. Experimental Settings

Datasets. The training set of ImageNet with over 1.28 million samples is used for training the generator. Following work that focus on transferability, the dataset from *NeurIPS 2017 Adversarial Learning* [34], comprising 1000 images from the ImageNet validation set, is used for evaluation.

Implementation details. ViT-B/16 architectures with default stride $s = 16$ is chosen for both the self-supervised DINO and MAE, and the normal supervised variant. Pretrained weights on ImageNet are sourced from their original implementations. Following baseline methods [41, 71], we use the same ResNet generator for $\mathcal{G}_\theta$. It is trained with the Adam optimizer with learning rate $\eta = 2{\times}10^{-4}$ over a single epoch. Scaling factor of attention saliency map $\gamma = 100$. We report results of dSVA trained with (1) DINO only, (2) MAE only, and (3) both DINO and MAE (Joint). *(dSVA collapses to SVA when only one self-supervised ViT is used, but we stick to the name of dSVA to avoid ambiguity.)*

Parameters. For both DINO and MAE, we choose features extracted at the penultimate layer $l = 10$. We select the *key* facet of DINO and the *query* facet of MAE to exploit. The joint training parameter of dSVA is set as $\lambda = 0.5$. The rationale and empirical evaluations supporting these selections are presented in Secs. 4.4 and 4.5.

Metric. We employ the fooling rate, i.e., the ratio of the adversarial examples which successfully fool the target model among all generated samples, as the evaluation metric.

Attacks. Generative attack baselines include BIA [71] and CDA [41]. We use VGG-19 [50], ResNet-152 (Res-152) [23], and DenseNet-169 (Den-169) [27] as their surrogates with the same perturbation budget of $\varepsilon = 10$ to follow their setups. We also compare against BIA and CDA trained on supervised ViT-B/16. We additionally include evaluations against gradient-based attacks, including the classic MI-FGSM (MI) [15], and 3 other state-of-the-art attacks designed for ViTs (PNA [62], TGR [70], ATT [38]). *(In Tab. 1 and Tab. 2, MI-FGSM is abbreviated as MI so as to avoid confusion with MIM—masked image modeling.)*

4.2. Transferability to Black-box Models

We first evaluate black-box transferability within the ImageNet domain. For attack targets, we choose 3 ConvNets with the same structure as the surrogates of the compared methods to follow baseline settings (VGG-16, ResNet-50 (Res-50), DenseNet-121 (Den-121)). We add 3 ConvNets with a different structure (EfficientNet-B0 (Eff-B0) [55], Inception-v3 (Inc-v3) [53], Inception-v4 (Inc-v4) [54]), 5 ViTs (Swin-B [36], MaxViT-T [58], PiT-B [26], VisFormer-S [11], LeViT-128 [21]), and an MLP Mixer (Mixer-B/16) [56]. We report the results in Tab. 1.

Across all models, dSVA consistently achieves exceptional transferability, outperforming baselines methods. As expected, BIA and CDA with surrogates VGG-19, Res-152, and Den-169 slightly outperforms dSVA on specific target models VGG-16, Res-50, and Den-121, as they share the same structure. Nevertheless, the transferability of dSVA (Joint) surpasses all compared attacks on the remaining models, particularly non-ConvNets. Even when using a *supervised ViT surrogate*, competing attacks fail to match dSVA's performance, including state-of-the-art attacks that are tailored for ViTs. Only CDA with a supervised ViT matches dSVA in 2 cases (Swin-B and PiT-B). Our results show that (1) without our proposed exploitation schemes in dSVA, existing feature-level attacks simply cannot take full advantage of the Transformer architecture, and (2) dSVA (Joint) outperforms its single model variants by 13.70% on aver-

Attack	Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v4$_{ens4}$	IncRes-v2$_{ens}$	IncRes-v2$_{adv}$	Eff-b0$_{ap}$
CDA (VGG-19)	25.05	16.36	9.78	10.73	34.90	67.39
CDA (Res-152)	43.01	38.60	28.88	29.27	61.89	73.91
CDA (Den-169)	53.44	41.11	27.08	24.58	66.00	83.33
BIA (VGG-19)	39.57	28.35	21.24	17.60	62.19	79.71
BIA (Res-152)	32.26	27.15	19.89	17.50	63.29	70.29
BIA (Den-169)	55.91	43.40	37.64	30.52	59.08	86.23
CDA (ViT-B/16)	65.91	53.98	50.67	38.54	71.11	86.23
BIA (ViT-B/16)	22.80	15.38	12.02	10.83	24.97	52.17
MI (ViT-B/16)	26.67	22.46	21.91	18.85	26.98	55.07
PNA (ViT-B/16)	27.63	22.90	22.70	19.79	29.69	55.07
TGR (ViT-B/16)	30.22	25.85	24.83	21.67	29.89	67.39
ATT (ViT-B/16)	40.43	36.21	33.03	29.79	41.52	75.36
dSVA (DINO)	66.13	54.09	49.33	43.85	75.03	**89.96**
dSVA (MAE)	50.11	32.39	28.88	23.85	66.70	76.09
dSVA (Joint)	**79.03**	**68.16**	**62.70**	**52.50**	**88.06**	89.13

Table 2. **Comparison of transferability against models with defenses.** We report the black-box fooling rate (%) of dSVA and compared baseline attacks in defenses evasion, on various models with adversarial training enabled within ImageNet.

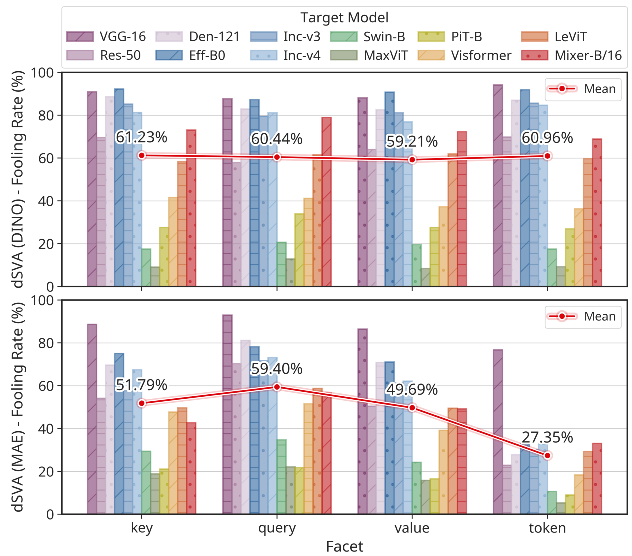

Figure 4. **Impact of the choice of facet.** We evaluate the transferability of dSVA (DINO) and dSVA (MAE) that exploit feature facets at layer 10 of *query*, *key*, *value*, and *token*, respectively.

age, underscoring the importance of our joint exploit of the complementary structural and textural features from the self-supervised strategy duo.

4.3. Transferability to Defense Models

Next, we validate our approach against defenses, an aspect previously unexplored in the context of generative attacks. We follow previous setups [38, 62, 70] and use 6 robust black-box models on ImageNet to evaluate defense evasion, namely Inc-v3$_{adv}$, IncRes-v2$_{adv}$ [33], Inc-v3$_{ens3}$, Inc-v4$_{ens4}$, IncRes-v2$_{ens}$ [57], and EfficientNet-B0 with AdvProp [65] (Eff-b0$_{ap}$). Shown in Tab. 2, we once again observe that dSVA shows superior performance across all adversarially trained models, with dSVA (Joint) achieving transferability that exceeds all compared attacks by an average margin of 32.98%. We contend that while adversarial training enhances DNN robustness by developing more resilient features, they ultimately need to use these same essential features for classification. By fully exploiting self-supervised ViT representations, decisive elements of the sample are destroyed at a more generalized level, allowing dSVA to evade these defenses. We provide additional results against state-of-the-art defenses and robust ViTs in Appendix C.

4.4. Analysis on the Impact of Relevant Parameters

We now turn our focus to the deciding *parameters* within dSVA, that is, (1) the feature facet (*query*, *key*, *value*, or the entire layer's output: *token*), (2) the feature layer l, and (3) λ, for the joint variant of dSVA. *(In Figs. 4 to 6 of Sec. 4.4, the bold red line represents the mean transferability of the evaluated variant of dSVA, aggregated over observations against all target black-box models.)*

The choice of facet $\{q, k, v, t\}$. We first evaluate the performances of dSVA (DINO) and dSVA (MAE) with respect to the exploited facets. We report the black-box transferability of them in Fig. 4. We first note that the variants that directly exploit the *token* facet, i.e, the entire intermediate layer output, always lags behind, especially in the case of dSVA (MAE). These findings underline the efficacy of our proposed facet-level exploit to capitalize on the adversarial potential of the features distilled by the Transformer architecture. For MAE, the *query* directly serves as the input with masked patches, which is intuitively more crucial for its reconstruction task. The *key* facet in this case only provides additional *context* of the current masked modelling session. This aligns with our observation that dSVA (MAE) performs best with the *query* facet. For DINO, the student network generates one view of the image as the *query*, while the teacher uses another as the *key*. The teacher, acting as a guide, would provide a better contrastive signal. Our results, although not as pronounced as the MAE variant, show that dSVA (DINO) performs best when exploiting the *key* facet.

The choice of layer l. Next, we investigate the impact of layer l. We report dSVA (DINO) and dSVA (MAE)'s transferability that exploit layer l from 1 to 11 in Fig. 5. We notice that the transferability of dSVA tends to increase as layer deepens. We reason that as the layers deepen, both self-supervised strategies manage to encode richer and more generalizable semantic information, benefiting adversarial transferability. Notably, transferability of both variants drops at the final 11th layer. This is expected as the final layer of Transformer-based models is often optimized for specific

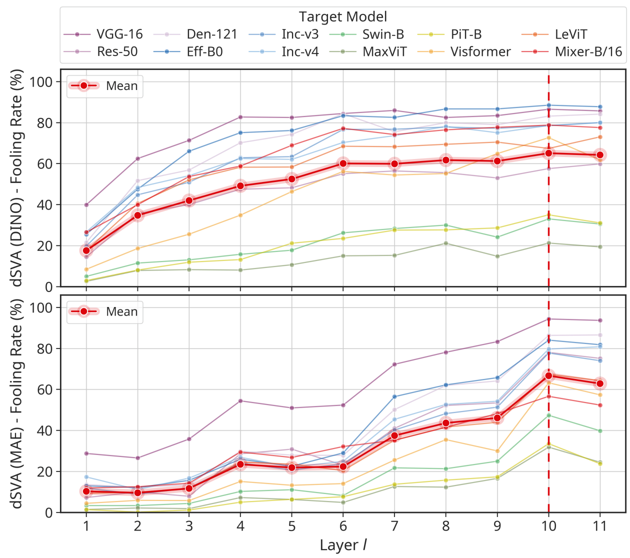

Figure 5. **Impact of the choice of layer l.** We evaluate the transferability of dSVA (DINO) and dSVA (MAE) with layer l from 1 to 11 (from left to right).

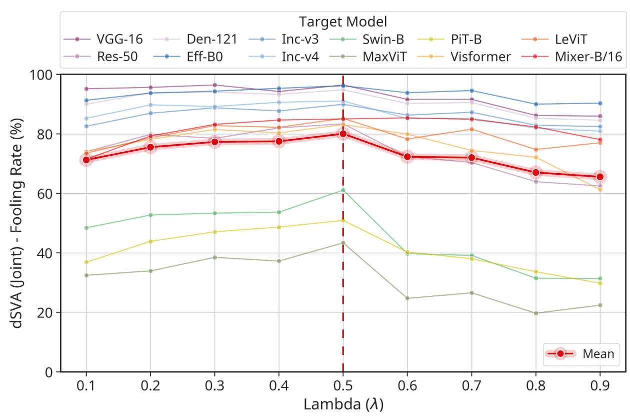

Figure 6. **Impact of the choice of λ.** We evaluate the transferability of dSVA (Joint) with default parameters employed except for λ. λ is applied from 0 to 1 with a step size of 0.1.

training setups, which results in significant reduction in generalizability [14]. In terms of vision tasks, ViTs have also shown to maintain spatial and positional information in all but the last layer [20, 30]. We choose the penultimate layer of $l = 10$ of both DINO and MAE in dSVA.

The choice of joint training parameter λ. We finally explore the key factor of dSVA (Joint), that is, the balance between feature disruption for DINO (CL) and MAE (MIM), which is controlled by λ as described in Eq. (7). The transferability of dSVA (Joint) with λ in $(0, 1)$ with a step size of 0.1 is reported in Fig. 6. We observe two interesting trends. First, as the dual aspects of features are more incorporated into dSVA (as λ approaches the midpoint), adversarial effectiveness increases. This behavior substantiates our hypothesis that the features provided by CL and MIM complement each other under an adversarial context, where both global and local relationships are to be destroyed, highlighting the importance of our proposed joint feature disruption. In addition, as λ decreases from 0.9 to 0.5, that is, as the aspect of MIM features increase while CL features decrease, adversarial effectiveness show a tendency to rise. We argue that the while CL provided structures are crucial for shape/object distinction from a human standpoint, to craft generalized perturbation for fooling DNNs, textural details distilled by MIM ought to be more purposefully considered, as DNNs favor these fine-grained details. $\lambda = 0.5$ yields the best performances in our setup, but given the similarity of the trends for λ in $[0.3, 0.5]$, we suggest that the optimal λ may vary depending on the specific task or dataset.

4.5. Visualizing Facet-level Feature Disruption

In Fig. 7, we visualize how self-supervised ViT features are more meaningful than supervised ones, and how some ViT feature facets are more crucial than others. We conduct PCA on DINO, MAE, and supervised ViT-B/16's features on all facets. We notice once again that the self-supervised features are richer and less noisy than the supervised ones. We find that, for both DINO and MAE, the *value* and *token* facets are noisier than the *query* and *key* facets. For DINO, its *key* facet shows more distinct shapes and objects, whereas for MAE, its *query* facet shows less noisy textured details. These observations align with our parameter selections. We also show how dSVA's adversarial perturbation equally destroys meaningful semantics within the image, underscoring our approach's effectiveness in feature disruption.

4.6. Ablation Study

We finally conduct an ablation study on two factors: (1) self-supervision, and (2) self-attention exploitation. We report the transferability of dSVA with supervised ViT, DINO, MAE, and Joint variants, both w/ and w/o attention saliency map regularization applied, in Fig. 8. For single model variants, we aggregate the results over all facets. For dSVA (Joint), we aggregate the observations over $\lambda \in [0.3, 0.5]$.

Self-supervision. When comparing variants of dSVA with self-supervised features to the supervised variant under identical conditions, even single model variants, including dSVA (DINO) and dSVA (MAE), outperform the supervised version across all models. We once again showcase that the synergy between self-supervision and the Transformer architecture, the central motivation of our work, pushes generative adversarial effectiveness to a new level, heightening the capability of our proposed approach.

Self-attention exploitation. We first observe that the self-attention of supervised ViTs actually impair adversarial effectiveness when applied as a regularization. As previously

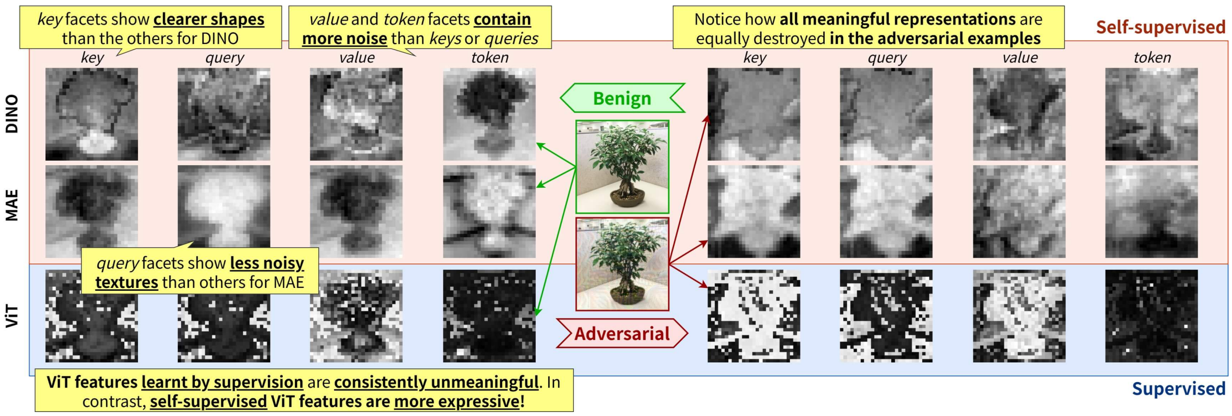

Figure 7. **Visualization of feature disruption.** We present PCA visualizations of the features extracted from all facets of DINO, MAE, and supervised ViT-B/16. Features of benign images are shown on the left, and adversarial examples crafted by `dSVA` (Joint) on the right.

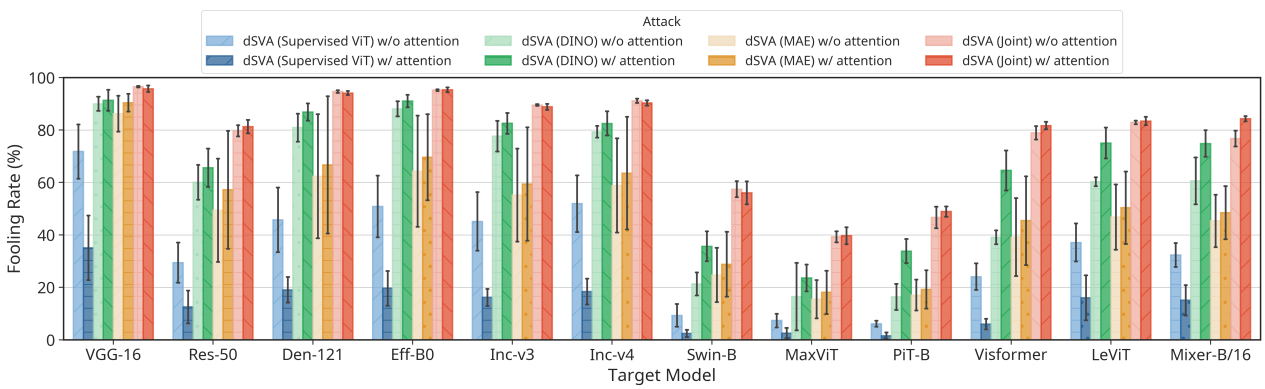

Figure 8. **Ablation study.** We present comparisons of the transferability of `dSVA` with supervised ViT, DINO, MAE, and Joint variants, with and without our proposed attention regularization applied, respectively. Results are aggregated over multiple observations.

shown, attention saliency maps extracted from the supervised ViT fail to match its self-supervised counterparts when acting as a regularization for feature landmark guidance. `dSVA` with self-supervised ViTs DINO and MAE consistently perform better when self-attention is also exploited. While `dSVA` (Joint) outperforms all single model variants, its transferability occasionally slightly degrades when attention regularization is applied, particularly when transferability is already high. We find that `dSVA` (Joint) works best with attention regularization active when targeting stronger or more sophisticated models.

4.7. Cross-domain Transferability

Our major competitors BIA and CDA show strong cross-domain transferability with only ImageNet domain knowledge. We provide additional comparisons under cross-domain settings in Appendix B. Results show that `dSVA` still maintains superior transferability to both coarse and fine-grained classification domains in most cases, offering boosts of approximately 6% on average.

5. Conclusion

We present a novel generative adversarial attack, `dSVA`, that successfully exploits deep intermediate features distilled through the self-supervised learning of ViTs. By aiming at facet-level feature representations, `dSVA` takes full advantage of the ViT's internal architecture. With self-attention regularization, `dSVA` vigilantly focuses on salient feature targets that are valuable for exploitation. Through our joint disruption of both structural and textural representations distilled by the self-supervised learning duo—CL and MIM—`dSVA` crafts remarkably generalizable perturbation, achieving state-of-the-art transferability. We demonstrate, through extensive experiments, the superior adversarial transferability of `dSVA` to various black-box DNNs of distinct architectures. *Our research strongly indicates that effective adversarial exploitation of ViTs, especially feature-wise, is very much muted by the use of surrogate models constrained by supervised learning.* We believe this work encourages further exploration of the robustness implications of DNNs within a self-supervised learning context.

Ackowledgements

This work was supported by the National Natural Science Foundation of China (no. U2336201).

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. On the effectiveness of vit features as local semantic descriptors. In *ECCV Workshops*, pages 39–55. Springer, 2022. 2, 4

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pages 15619–15629. IEEE, 2023. 2

[3] Yatong Bai, Mo Zhou, Vishal M. Patel, and Somayeh Sojoudi. Mixednuts: Training-free accuracy-robustness balance via nonlinearly mixed classifiers. *Trans. Mach. Learn. Res.*, 2024, 2024. 13

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *ICLR*, 2022. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 1

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640. IEEE, 2021. 2, 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9620–9629. IEEE, 2021. 2

[10] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *CVPR*, pages 5157–5166. Computer Vision Foundation / IEEE, 2019. 12

[11] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *ICCV*, pages 569–578. IEEE, 2021. 5

[12] Adam Coates, A. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, 2011. 12

[13] Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. In *SaTML*, pages 225–253. IEEE, 2023. 13

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics, 2019. 2, 7

[15] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, pages 9185–9193. Computer Vision Foundation / IEEE Computer Society, 2018. 5

[16] Paul Engstler, Luke Melas-Kyriazi, Christian Rupprecht, and Iro Laina. Understanding self-supervised features for learning unsupervised instance segmentation. *CoRR*, abs/2311.14665, 2023. 2

[17] Weiwei Feng, Nanqing Xu, Tianzhu Zhang, and Yongdong Zhang. Dynamic generative targeted attacks with pattern injection. In *CVPR*, pages 16404–16414. IEEE, 2023. 2

[18] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023. 1

[19] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423. IEEE Computer Society, 2016. 1

[20] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? A visual exploration. *CoRR*, abs/2212.06727, 2022. 7

[21] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *ICCV*, pages 12239–12249. IEEE, 2021. 5

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, R émi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *NeurIPS*, 2020. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 5

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020. 2

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 15979–15988. IEEE, 2022. 2

[26] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *ICCV*, pages 11916–11925. IEEE, 2021. 5

[27] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269. IEEE Computer Society, 2017. 5

[28] Qian Huang, Isay Katsman, Zeqi Gu, Horace He, Serge J. Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *ICCV*, pages 4732–4741. IEEE, 2019. 1

[29] Nathan Inkawhich, Wei Wen, Hai (Helen) Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *CVPR*, pages 7066–7074. Computer Vision Foundation / IEEE, 2019. 1

[30] Samy Jelassi, Michael E. Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. In *NeurIPS*, 2022. 7

[31] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, pages 554–561. IEEE Computer Society, 2013. 12

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *University of Toronto*, 2009. 12

[33] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017. 6

[34] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 195–231. Springer, 2018. 5

[35] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 105–114. IEEE Computer Society, 2017. 1

[36] 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 *ICCV*, pages 9992–10002. IEEE, 2021. 5

[37] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 12

[38] Di Ming, Peng Ren, Yunlong Wang, and Xin Feng. Boosting the transferability of adversarial attack on vision transformer with adaptive token tuning. In *NeurIPS*, 2024. 5, 6

[39] Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. In *NeurIPS*, 2022. 13

[40] Krishna Kanth Nakka and Mathieu Salzmann. Learning transferable adversarial perturbations. In *NeurIPS*, pages 13950–13962, 2021. 2

[41] Muzammal Naseer, Salman H. Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. In *NeurIPS*, pages 12885–12895, 2019. 2, 5

[42] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 4. Granada, 2011. 12

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 2

[44] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *ICLR*, 2023. 2

[45] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge J. Belongie. Generative adversarial perturbations. In *CVPR*, pages 4422–4431. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[46] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[47] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 1

[48] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *NeurIPS*, 2020. 13

[49] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *ICLR*, 2023. 2

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[51] Naman Deep Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. In *NeurIPS*, 2023. 13

[52] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *NeurIPS*, pages 8322–8333, 2018. 2

[53] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826. IEEE Computer Society, 2016. 5

[54] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pages 4278–4284. AAAI Press, 2017. 5

[55] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114. PMLR, 2019. 5

[56] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In *NeurIPS*, pages 24261–24272, 2021. 5

[57] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR*, 2018. 6

[58] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan C. Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV*, pages 459–479. Springer, 2022. 5

[59] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 12

[60] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *ICCV*, pages 7619–7628. IEEE, 2021. 1

[61] Zhibo Wang, Hongshan Yang, Yunhe Feng, Peng Sun, Hengchang Guo, Zhifei Zhang, and Kui Ren. Towards transferable targeted adversarial examples. In *CVPR*, pages 20534–20543. IEEE, 2023. 2

[62] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. In *AAAI*, pages 2668–2676. AAAI Press, 2022. 5, 6

[63] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *ICLR*, 2020. 13

[64] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *IJCAI*, pages 3905–3911. ijcai.org, 2018. 2

[65] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. In *CVPR*, pages 816–825. Computer Vision Foundation / IEEE, 2020. 6

[66] 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 *CVPR*, pages 9643–9653. IEEE, 2022. 2

[67] Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Prompt-driven contrastive learning for transferable adversarial attacks. In *ECCV*, pages 36–53. Springer, 2024. 2

[68] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Boosting transferability of targeted adversarial examples via hierarchical generative networks. In *ECCV*, pages 725–742. Springer, 2022. 2

[69] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328, 2014. 1

[70] Jianping Zhang, Yizhan Huang, Weibin Wu, and Michael R. Lyu. Transferable adversarial attacks on vision transformers with token gradient regularization. In *CVPR*, pages 16415–16424. IEEE, 2023. 5, 6

[71] Qilong Zhang, Xiaodan Li, Yuefeng Chen, Jingkuan Song, Lianli Gao, Yuan He, and Hui Xue. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. In *ICLR*, 2022. 2, 5, 12

[72] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 1

[73] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *ECCV*, pages 471–486. Springer, 2018. 1