

# Task-Agnostic Attacks Against Vision Foundation Models

Brian Pulfer<sup>1,\*</sup> Yury Belousov<sup>1,\*</sup> Vitaliy Kinakh<sup>1</sup> Teddy Furon<sup>2</sup> Slava Voloshynovskiy<sup>1</sup>  
<sup>1</sup>University of Geneva <sup>2</sup>University of Rennes, Inria, CNRS, IRISA

## Abstract

*The study of security in machine learning mainly focuses on downstream task-specific attacks, where the adversarial example is obtained by optimizing a loss function specific to the downstream task. At the same time, it has become standard practice for machine learning practitioners to adopt publicly available pre-trained vision foundation models, effectively sharing a common backbone architecture across a multitude of applications such as classification, segmentation, depth estimation, retrieval, question-answering and more. The study of attacks on such foundation models and their impact to multiple downstream tasks remains vastly unexplored. This work proposes a general framework that forges task-agnostic adversarial examples by maximally disrupting the feature representation obtained with foundation models. We extensively evaluate the security of the feature representations obtained by popular vision foundation models by measuring the impact of this attack on multiple downstream tasks and its transferability between models.*

## 1. Introduction

Vision Foundation Models (VFMs) are becoming increasingly popular due to their versatility in handling a wide range of image-based downstream tasks. Once trained, these models can be minimally fine-tuned to perform classification and semantic segmentation [30, 43], object detection [37], depth estimation [65, 66], visual question-answering [8, 56], image captioning [41], retrieval [19], watermarking [20] and more. VFMs are trained with different strategies depending on the Self-Supervised Learning (SSL) framework. Additionally, some VFMs benefit from the bridge between the text and image modalities as offered by CLIP [48] and follow-up works [31, 33, 34].

VFMs have become a crucial component in many advanced systems due to their versatility, performances, and availability as open-sourced models, which on the other hand raises security concerns. Given their ubiquity, it is crucial to assess their robustness and their security against attacks.

\*These authors contributed equally to this work.

This paper defines the **robustness** of VFMs as the system’s ability to produce reliable and consistent outputs for a given downstream task when subjected to standard non-adversarial degradations, such as lossy compression, rota-

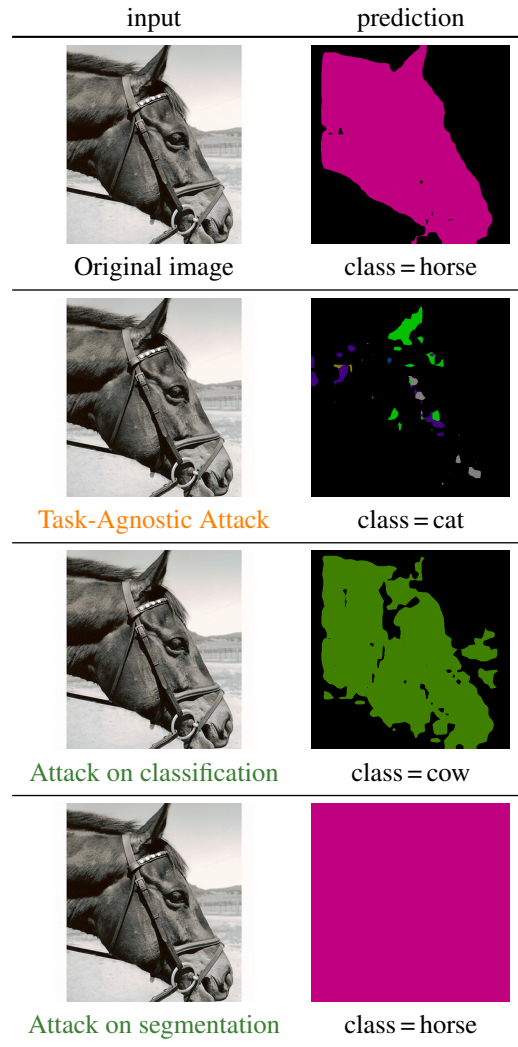


Figure 1. Adversarial example attacks on DinoV2 ViT-S model. The Task-Agnostic Attack deludes both the segmentation and the classification, contrary to attacks specific to a downstream task.

tion, flipping, contrast changes, blurring, etc. The **security** of VFMs, on the other hand, is defined as the system’s resilience in maintaining output integrity for the downstream task when faced with adversarial attacks designed to alter their output. Despite the popularity of foundation models, the security aspect remains under-explored [4].

Traditional adversarial examples are designed with a specific downstream task in mind. We refer to this family of attacks as *task-specific attacks* (TSAs). These exploit the particularities of the downstream task head to create adversarial examples that disrupt the model’s performance on that task. A TSA can be non-targeted, aiming to alter the task label to any incorrect output, or targeted, where the task label is changed to a specific, desired output. While effective, these attacks are limited by their reliance on the characteristics of the downstream tasks.

In contrast, this paper introduces *task-agnostic attacks* (TAAs) which do not consider the downstream task head, but rather only target the foundation model. TAAs can also be divided into untargeted attacks, which aim to move the latent representation far from the original one, and targeted attacks, which aim to move the latent to a specific pre-defined target in the latent space. Fig. 1 illustrates that while TSA perform best for the designated downstream task, transferability to other tasks is limited. In contrast, TAA generates an adversarial sample that effectively deceive the model across a range of applications. This paper investigates the efficiency and transferability of untargeted TAAs on foundation models across various downstream tasks and models.

Our first contribution is the design of a new untargeted TAA with variants perturbing different tokens of a Vision Transformer (ViT). The second contribution evaluates the pervasion of TAAs in terms of their ability to delude downstream tasks like classification, segmentation, image retrieval, etc. We also compare TAAs and TSAs for the same image distortion budget. Finally, we investigate the transferability of the attacks from one white-box source model to a black-box target model.

We demonstrate that task-agnostic attacks generalize effectively across multiple downstream tasks, model sizes and datasets, revealing their universality and their dangerousness against any self-supervised learning and multi-modal system. These insights emphasize the need for further investigation into the security of foundation models before their widespread deployment in critical applications.

## 2. Related Work

### 2.1. Vision Foundation Models

VFMs provide an unprecedented level of utility and versatility for a wide range of image-based tasks. VFMs differ in their architecture, their size, and their training set, but the

main difference lies in the SSL framework. The training objective is key to create a universal model which feature representations can be applied to a vast amount of downstream tasks.

**MAE** [24] is based solely on the Masked Image Modelling (MIM) objective, where an encoder-decoder architecture learns to reconstruct masked image patches from a few visible patches only. **MSN** [2] combines MIM with Siamese Networks to avoid pixel and token-level reconstructions. A teacher network computes the representation of a view of an image, while a student network computes the representation of another partially masked view. MSN is optimized by learning a student that can match the output of the teacher network. **CAE** [14] also learns to predict masked visual tokens from visible tokens by aligning their representations. **I-JEPA** [3] predicts representations of various blocks of an image given only a context block obtained through a specific masking strategy.

**DiNO** [11] first proposed self-distillation with no labels. The outputs of a teacher and student networks are passed through a softmax, and the objective is for the student to minimize the cross-entropy loss between the two probability distributions. **iBOT** [68] builds on top of DiNO, combining the self-distillation strategy with the MIM objective of BEiT [5] within self-distillation. **DiNOv2** [43] improves on DiNO by using a larger and curated dataset, namely LVD-142M. It relies on an efficient implementation for training at scale and an advanced SSL framework. In particular, the authors combine the DiNO cross-entropy loss with the MIM objective used in iBOT. DiNOv2 also benefits from the Sinkhorn-Knopp batch normalization of SwAV [58]. The largest model, ViT-G, is distilled into smaller models.

Our study considers a set of 18 popular and recent pre-trained VFMs listed in Tab. 1. All models have been pre-trained on the ImageNet dataset [16], except for DiNOv2, which was pre-trained on LVD-142M [43], a superset of ImageNet. We only consider ViTs with patch sizes 14 or 16, since most available VFMs use such a patching strategy.

### 2.2. Adversarial Example Attacks

The seminal works [22, 53] on adversarial examples reveal the paradox that Deep Learning classifiers are robust but not secure. They are robust because their prediction likely remains unchanged when the input image is corrupted by noise addition or JPEG compression. They are not secure because an attacker can craft a low-amplitude perturbation that deludes the classifier. These attacks are typically categorized based on the attacker’s knowledge and access to the target model: white-box attacks and black-box attacks.

In the white-box scenario, the attacker has the complete knowledge of the model’s architecture and parameters, facilitating the generation of adversarial examples. They first define a loss function that may combine the classification

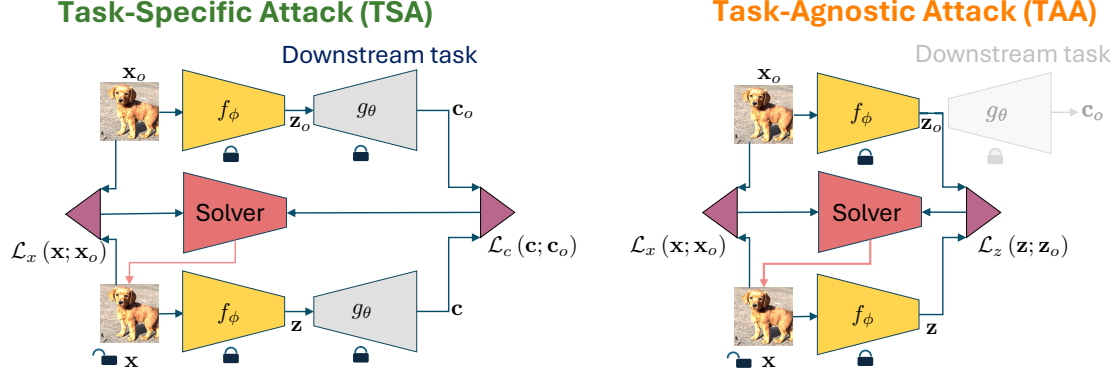


Figure 2. Schematic representation of classic Task-Specific Attack (left) and proposed Task-Agnostic Attack (right).

Table 1. List of the VFMs considered in this work. The table reports the SSL framework, the architecture, the number of parameters, and the dimensionality of the feature space.

Framework	Architecture	# Pars. (M)	Dim.
DiNOv2	ViT-S/14	21	384
DiNO MSN	ViT-S/16		
CAE DiNOv2	ViT-B/14	85	768
DiNO iBOT MAE MSN	ViT-B/16		
CAE DiNOv2	ViT-L/14		
iBOT MAE MSN	ViT-L/16		
I-JEPA MAE	ViT-H/14	630	1'280
DiNOv2	ViT-G/14	1'136	1'536
I-JEPA	ViT-G/16	1'011	1'408

loss of the input w.r.t. the ground truth class and another most likely class together with the distortion w.r.t. the original image. The attack relies on the gradient of this loss w.r.t. the input computed by backpropagation through the model. The typical representatives of white-box attacks include PGD [38], targeting a high Attack Success Rate (ASR) within a given distortion budget, and DeepFool [42] or CW [9] minimizing the distortion of a successful adversarial example. Recent efforts in this area aim at improving the speed of the attack [7, 45]. The black-box scenario limits the attacker's access to the output of the model, without

knowledge of its internals. This category includes methods such as RayS [12], SurFree [39], and CGBA [51].

Adversarial examples delude not only classifiers, but also semantic segmentation and object detection [61], depth estimation [64, 67], visual question answering [52], captioning [62], or retrieval [57] models. Note that each work proposes an attack specific to the targeted application. The core process is always the gradient computation but the definition of the adversarial loss is driven by the application.

Since a vast majority of foundation models are open-source, we investigate white-box scenarios with higher emphasis than black-box. At the same time, unlike traditional attacks, our method aims to be universal across applications.

### 2.3. Attacks against Vision Foundation Models

It was shown that complementing classifiers with self-supervision during training improves robustness [25]. Another study [15] confirms that many VFMs (BarlowTwins, BYOL, SimCLR, SimSiam, SwAV, and DiNO) inherit from this fact a greater robustness against image corruptions. Robustness is here understood in the statistical point of view, i.e., how the performance smoothly degrades as the distribution shift increases. This does not encompass adversarial examples crafted with the sole goal of damaging downstream performance.

On the contrary, DINO self-supervision does not improve the security against adversarial example attacks compared to traditional supervision in classification tasks [50]. To patch this vulnerability, some works propose to combine self-supervision and adversarial training [25, 29].

All these works do not study the intrinsic vulnerability of the VFM per se, but once used in a classification task. One exception is the report [28] which designs an attack such that adversarial examples are classified as Out-Of-Distribution samples. Therefore, whatever the downstream task, the system refuses to process them.

Our paper studies the security of VFM without assuming any downstream task.

### 3. Definition of adversarial attacks

This section pinpoints the differences between TSAs and TAAs as depicted in Fig. 2.

#### 3.1. Classical Task-Specific Attacks

Let us consider a model composed by a VFM  $f_\phi$  with a downstream task head  $g_\theta$ . For a given input  $\mathbf{x}_o$ , let  $\mathbf{c}_o = g_\theta(f_\phi(\mathbf{x}_o))$ . The attacker defines two losses:  $\mathcal{L}_x(\mathbf{x}; \mathbf{x}_o)$  measuring the perceptual distance between an input  $\mathbf{x}$  and  $\mathbf{x}_o$ , and  $\mathcal{L}_c(\mathbf{c}; \mathbf{c}_o)$  gauging how the task performed on  $\mathbf{x}$  differs from the result  $\mathbf{c}_o$ . For instance, for the classification task, the result  $\mathbf{c}_o$  is a predicted class and  $\mathcal{L}_c(\mathbf{c}; \mathbf{c}_o) = p(\mathbf{c}_o|\mathbf{x}) - \max_{\mathbf{c} \neq \mathbf{c}_o} p(\mathbf{c}|\mathbf{x})$ . Under a distortion budget constraint, the attack looks for the minimizer  $\mathbf{x}_a$  of  $\mathcal{L}_c(\mathbf{c}; \mathbf{c}_o)$  over  $\{\mathbf{x} : \mathcal{L}_x(\mathbf{x}; \mathbf{x}_o) \leq \epsilon\}$ , with  $\mathbf{c} = g_\theta(f_\phi(\mathbf{x}))$ . An alternative is to solve the dual problem, i.e. the attack looks for the minimizer  $\mathbf{x}_a$  of  $\mathcal{L}_x(\mathbf{x}; \mathbf{x}_o)$  under the constraint  $\mathcal{L}_c(\mathbf{c}; \mathbf{c}_o) \leq \tau_c$ , with  $\mathbf{c} = g_\theta(f_\phi(\mathbf{x}))$ .

The adversarial example  $\mathbf{x}_a$  crucially depends on the downstream task via functions  $g_\theta$  and  $\mathcal{L}_c$ . This implies that the attacker knows the downstream task head (white-box) and that  $\mathbf{x}_a$  may not be adversarial for different tasks.

#### 3.2. Proposed Task-Agnostic Attack

In contrast to TSAs, we target a new type of task-agnostic attack when the downstream head or even the downstream task are unknown. The objective is to forge *pervasive* adversarial examples, i.e. jeopardizing a large spectrum of downstream tasks. The proposed TAA aims at maximally perturbing the features obtained from the VFM backbone.

Given a foundation model  $f_\phi : \mathcal{X} \rightarrow \mathcal{Z}$ , the features of VFM are computed as  $\mathbf{z} = f_\phi(\mathbf{x}) \in \mathcal{Z}$ . For a ViT,  $\mathbf{z}$  may represent any aggregating function of the output class and patch tokens. We compute latent embeddings  $\mathbf{z}_j = f_\phi(\mathbf{x}_j)$  for some training samples  $\{\mathbf{x}_j\}_{j=1}^{N_T}$ , which we use to compute an empirical mean  $\boldsymbol{\mu} = \frac{1}{N_T} \sum_{k=1}^{N_T} \mathbf{z}_k$  which can be used to center features extracted with the VFM as  $\tilde{\mathbf{z}} = \mathbf{z} - \boldsymbol{\mu}$ .

For a given input  $\mathbf{x}_o$ , we denote  $\mathbf{z}_o = f_\phi(\mathbf{x}_o) \in \mathcal{Z}$ . The loss function is now defined in the feature space  $\mathcal{Z}$  as:

$$\mathcal{L}_z(\mathbf{z}; \mathbf{z}_o) = \text{cos\_sim}(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}_o). \quad (1)$$

We notice that features extracted from some foundation models are not centered around the origin, making the mean-centering process necessary to compute meaningful cosine similarities. We report results without this pre-processing step in supplementary material.

Again, two alternatives are 1) minimize  $\mathcal{L}_z(\mathbf{z}; \mathbf{z}_o)$  under the distortion budget constraint  $\mathcal{L}_x(\mathbf{x}; \mathbf{x}_o) \leq \epsilon$ , or 2) minimize the distortion  $\mathcal{L}_x(\mathbf{x}; \mathbf{x}_o)$  under an objective constraint  $\mathcal{L}_z(\mathbf{z}; \mathbf{z}_o) < \tau_z$ . We prefer the first option as it eases the fair comparison of the attacks by monitoring the Attack

Success Rate for a given distortion budget  $\epsilon$ . Under a distortion budget constraint, the attack looks for the minimizer  $\mathbf{x}_a$  of  $\mathcal{L}_x(\mathbf{x}; \mathbf{x}_o)$  under the constraint  $\mathcal{L}_z(\mathbf{z}; \mathbf{z}_o) < \tau_z$ . Algorithm 1 summarizes our method.

---

#### Algorithm 1 Task-Agnostic Attack

---

```

1: Input:  $\mathbf{x}_o$ : original image,  $\{\mathbf{x}_j\}_{j=1}^{N_T}$ : training images,
    $f_\phi$ : backbone model,  $\epsilon_\infty$ : maximum per-pixel perturbation,  $\alpha$ : step size,  $N$  number of steps;
   { // Compute mean vector from set of images for centering,  $sg$ : stop gradient }
2:  $\{\mathbf{z}_j\}_{j=1}^{N_T} \leftarrow \{sg(f_\phi(\mathbf{x}_j))\}_{j=1}^{N_T}$ 
3:  $\boldsymbol{\mu} \leftarrow \frac{1}{N_T} \sum_{k=1}^{N_T} \mathbf{z}_k$ 
   { // initialize attack and target }
4:  $\mathbf{x}_a \leftarrow \mathbf{x}_o$ 
5:  $\tilde{\mathbf{z}}_o \leftarrow sg(f_\phi(\mathbf{x}_o)) - \boldsymbol{\mu}$ 
   { // Run PGD in feature space using cos-sim }
6: for  $t = 0, \dots, N - 1$  do
7:    $\tilde{\mathbf{z}} \leftarrow f_\phi(\mathbf{x}_a) - \boldsymbol{\mu}$ 
8:    $\mathcal{L}_z \leftarrow \text{cos\_sim}(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}_o)$ 
9:    $\mathbf{x}_a \leftarrow \mathbf{x}_a - \alpha \times \nabla_{\mathbf{x}_a} \mathcal{L}_z$ 
10:   $\mathbf{x}_a \xleftarrow{\text{constraints}} \mathbf{x}_a$  { // impose constraints via  $\epsilon_\infty$  }
11: end for
12: Return: Attacked image  $\mathbf{x}_a$ 

```

---

**Implementation** We use Projected Gradient Descent (PGD) [38] as our solver. We set the step size  $\alpha = 0.0004$ , the total number of optimization steps to 50, and  $\epsilon_\infty = \frac{8}{255}$ . The adversarial tensors are then clipped and quantized to get real adversarial images, which are stored in *png* format and re-loaded for evaluation. This post-processing can result in slightly higher or lower  $\mathcal{L}_x(\mathbf{x}; \mathbf{x}_o)$  with respect to the target  $\tau_z$ . To carry out TAAs, we developed a highly customizable PyTorch [44] code which comes with public pre-trained SSL models used in this work. Source code is available at <https://github.com/BrianPulfer/fsaa>.

### 4. Experimental setup

**Models** Tab. 1 lists the models of the experimental setup. For each model, we use its default normalization to convert images to tensors.

**Datasets and metrics for downstream tasks** We use popular datasets and metrics for each downstream task:

- Classification: Accuracy on PascalVOC [18] or ImageNet [26]
- Segmentation: Mean Intersection over Union (mIoU) on PascalVOC [18]
- Visual Question Answering: Accuracy on VQA v2 [23]

- Image Captioning: BLEU-4, METEOR, ROUGE-L and CIDEr scores on COCOCaptions [13]
- Image Retrieval: Mean Average Precision (mAP) on the Revisited Oxford buildings dataset [47].

### Metrics for attacks

**Absolute efficiency:** We measure the drop in performance with the metrics and datasets mentioned above.

**Relative efficiency:** We define the *relative efficiency*  $\eta$  of a TAA w.r.t. a downstream task as a percentage, where 0% indicates no impact on task accuracy, and 100% reflects an impact equivalent to the respective TSA. Formally,

$$\eta = 100 \times \frac{\text{perf}(\text{TAA}) - \text{perf}(\text{No attack})}{\text{perf}(\text{TSA}) - \text{perf}(\text{No attack})}, \quad (2)$$

where perf is a performance metric of the downstream task.

**Image quality:** For a fair comparison of the attacks, we set a target Peak Signal-to-Noise Ratio PSNR =  $10 \log_{10} \frac{255^2}{\text{MSE}}$  with  $\text{MSE} = \frac{1}{3HW} \|\mathbf{x}_o - \mathbf{x}_a\|^2$ , where  $\mathbf{x} \in [0, 1]^{H \times W \times 3}$ . In experiments we set PSNR = 40 dB.

## 5. Experimental results on VFMs

### 5.1. Robustness vs. Security

Foundation models enable the *generalisation* of the downstream heads for many applications. Systems built on foundation models demonstrate remarkable performance across various downstream tasks under clean conditions. They also enable great *robustness* against common editing transformation such as flipping, blurring or JPEG compression. The accuracy smoothly degrades as the shift between distributions of the training and testing data gets larger. This motivates the analysis of the security level which would have been useless if not robust first and foremost. Downstream heads are learned on the PascalVOC training dataset for *classification* and *segmentation* each while the model backbone DiNOv2 ViT-S is kept frozen. Tab. 2 gauges the robustness against common image processing and the security against Task-Specific Attacks over images from the validation dataset and PSNR = 40 dB. For these two applications, the global system is clearly robust yet not secure against a TSA.

### 5.2. TAA vs. TSA

With the identical setup as previously described, Tab. 4 provides a comparison of TAAs and TSAs. The conclusions are clear: TSAs present a more effective attack approach compared to TAAs when examining the performance impacts on the targeted applications. Yet, TSAs exhibit limited pervasiveness. For example, the TSA directed at segmentation reduces the accuracy of the classification task only by around half. In contrast, TAAs negatively impact both tasks equally.

Table 2. Classification and segmentation performance for a set of distortions (*robustness*) and TSA (*security*).

Transform	Classification	Segmentation
<i>clean performance</i>	96.3	81.4
horizontal flip	95.9	81.3
vertical flip	87.9	61.0
wiener filter size = 21	95.7	76.6
blur kernel_size = 21	94.5	80.3
jpeg quality = 50	95.8	81.2
grayscale	96.3	81.3
rotation by 90°	88.5	65.7
resize to 98 × 98	88.5	61.3
brightness factor = 2	95.3	81.2
contrast factor = 2	95.8	80.9
hue factor = 0.5	95.9	81.1
Task-Specific Attack	0.0	11.8

Tab. 4 shows that TAAs can reach a minimum relative efficiency of 80%. Notably, TAAs that compromise patch tokens demonstrate near equivalent efficiency as TSAs.

### 5.3. Extending TAAs to other downstream tasks

**Image retrieval:** We measure the efficacy of TAAs for the image retrieval task on the revisited Oxford dataset [47] using DiNOv2 backbones and a target PSNR of 40 dB. Tab. 3 reports the mean average precision. Interestingly, image retrieval capabilities do not improve beyond a ViT-L model, yet, the adversarial robustness to TAAs does improve with a larger ViT-G architecture. Still, all model performances are severely degraded.

Table 3. mAP on the image retrieval with DiNOv2 on the R-Oxford dataset. PSNR is set to 40 dB.

Model	Difficulty	Clean↑	TAA↓
DiNOv2 ViT-S	Easy	82.1	0.6
	Medium	67.3	1.1
	Hard	41.4	0.5
DiNOv2 ViT-B	Easy	85.7	1.6
	Medium	71.7	2.0
	Hard	49.4	0.8
DiNOv2 ViT-L	Easy	87.9	1.0
	Medium	74.4	1.9
	Hard	53.1	1.3
DiNOv2 ViT-G	Easy	85.2	7.4
	Medium	72.9	7.5
	Hard	52.3	4.2

Table 4. Absolute and relative efficiency (2) of attacks against classification and segmentation by two DiNOv2 models on PascalVOC validation set. TAAs are more pervasive across tasks while TSAs are more harmful w.r.t. the targeted task. Target PSNR = 40 dB. The arrow indicates the direction of success for attacker.

Backbone	Attack	Type	Classification abs↓ (rel↑)	Segmentation abs↓ (rel↑)
ViT-S	No attack		96.3 (0%)	81.4 (0%)
	Class token	TAA	7.9 (92%)	19.2 (86%)
	Patch tokens	TAA	<b>0.1 (100%)</b>	<b>11.6 (97%)</b>
	Class+patch tokens	TAA	2.0 (98%)	13.3 (94%)
	Classification	TSA	<b>0.0 (100%)</b>	19.8 (85%)
	Segmentation	TSA	40.6 (58%)	<b>9.2 (100%)</b>
ViT-B	No attack		97.0 (0%)	80.8 (0%)
	Class token	TAA	11.8 (88%)	23.5 (80%)
	Patch tokens	TAA	<b>0.0 (100%)</b>	<b>7.5 (102%)</b>
	Class + patch tokens	TAA	2.1 (98%)	11.8 (96%)
	Classification	TSA	<b>0.0 (100%)</b>	14.1 (93%)
	Segmentation	TSA	43.9 (55%)	<b>8.9 (100%)</b>



Figure 3. Our TAA deludes Segment-Anything-Model [30]. Original (left) and adversarial (right, PSNR = 40 dB) images.

Table 5. Accuracy in zero-shot classification on ImagenetV2 for CLIP model with PSNR = 40 dB.

	No attack ↑	TAA ↓
Top-1	55.95	0.03
Top-5	83.39	0.21

**Zero-shot segmentation:** In Fig. 3 the TAA targets the patch tokens of the ViT-H encoder in the Segment Anything

Model (SAM) [30], without relying on the prompt encoder and mask decoder. It shows segmentation masks before and after the attack with the same query (the green star point or the bounding box).

**Zero-shot classification:** Recent Vision-Language Models (VLM) [6, 31, 32, 36, 59, 60, 63] combine a pre-trained VFM as an image encoder together with a pre-trained language model, the latter usually involving more parameters and being computationally more demanding than the former. TAAs on the VFM compromise performances of VLMs without requiring computationally expensive evaluation of the language model.

Tab. 5 reports how vulnerable is *zero-shot classification* based on CLIP [49]. Our results are consistent with previous TSAs designed in [21, 40]. However, we show that this lack of security holds true for TAA as well.

**Image captioning and VQA:** We run TAAs against PaliGemma [6] to measure the drop in performance for *image captioning* over the COCO validation set [13] and *visual question answering* over the VQAv2 dataset [1, 23]. In both cases, we simply aim at maximally perturbing the feature representation obtained with SigLip, and do not make use of the much larger Gemma [54, 55] language model.

Tab. 6 shows quantitative results illustrated with captioning examples in Fig. 4. For both downstream tasks, a severe drop in performances occurs even with lower levels of PSNR. Particularly, we record a higher drop in accuracy for VQAv2 for the less trivial ”other” category of questions, and a dramatic drop in CIDEr on the captioning dataset.



Figure 4. Examples of regular (left) and adversarial (right) captions obtained with TAAs attacking the VFM of PaliGemma. PSNR is 40 dB.

Table 6. TAAs against the SigLip VFM of the PaliGemma VLM. Impact on captioning and question answering. For all metrics, lower values mean more harmful attack. Drop in performance is measured for increasingly stronger perturbations.

Attack PSNR	Captioning COCO				Question answering VQAv2		
	BLEU-4 ↓	METEOR ↓	ROUGE-L ↓	CIDEr ↓	number ↓	yes/no ↓	other ↓
<i>No attack</i>	29.6	30.3	59.0	131.4	72.5	95.9	76.9
45 dB	5.8	17.3	32.2	33.0	50.6	83.4	53.4
40 dB	3.8	13.6	27.2	16.4	38.7	76.0	41.6
35 dB	1.9	9.7	22.3	3.6	25.6	67.5	28.5

## 6. Results on transferability between models

So far, TAAs targeted a given VFM in a white-box scenario. This section extends the attack surface to the black-box scenario. In the adversarial example literature, transferability attacks are sometimes efficient to delude an unknown classifier. Our setup takes the same spirit. The attacker mounts the TAA against this white-box *source* model in the hope that this adversarial example also deludes the application based on the black-box *target* VFM.

### 6.1. Transferability across VFM backbones

For all backbones, a downstream head is trained for classification and semantic segmentation. For classification, we train linear heads on the ImageNet train set, but we evaluate attacks on the smaller Imagenette dataset [26]. For semantic segmentation, we concatenate the CLS token to every patch token along the feature dimension and use bi-cubic interpolation followed by a 2-dimensional convolution to obtain

predictions for the original image size. Again, TSAs use the trained heads whereas TAAs do not.

Fig. 5 presents the relative efficiency as defined in Eq. (2) for TAA compared to TSA, averaged across the two proposed tasks. TAA demonstrates near equivalent efficiency to TSA, affirming its effectiveness even without downstream task knowledge. TAAs are more successful against smaller models, and we attribute this to the lower dimensionality of the feature space. Both TSA and TAA exhibit moderate transferability across different models, with DINO and MSN being the most susceptible to transfer effects. However, this transferability is rare and lacks a clear pattern that would explain why it occurs specifically between these models. Furthermore, transferability tends to be more likely when both source and target models belong to the same pre-training framework but differ in size. Even within these model families, however, transfer is not guaranteed and remains inconsistent. Different solvers and perturbation budget are reported in supplementary material.

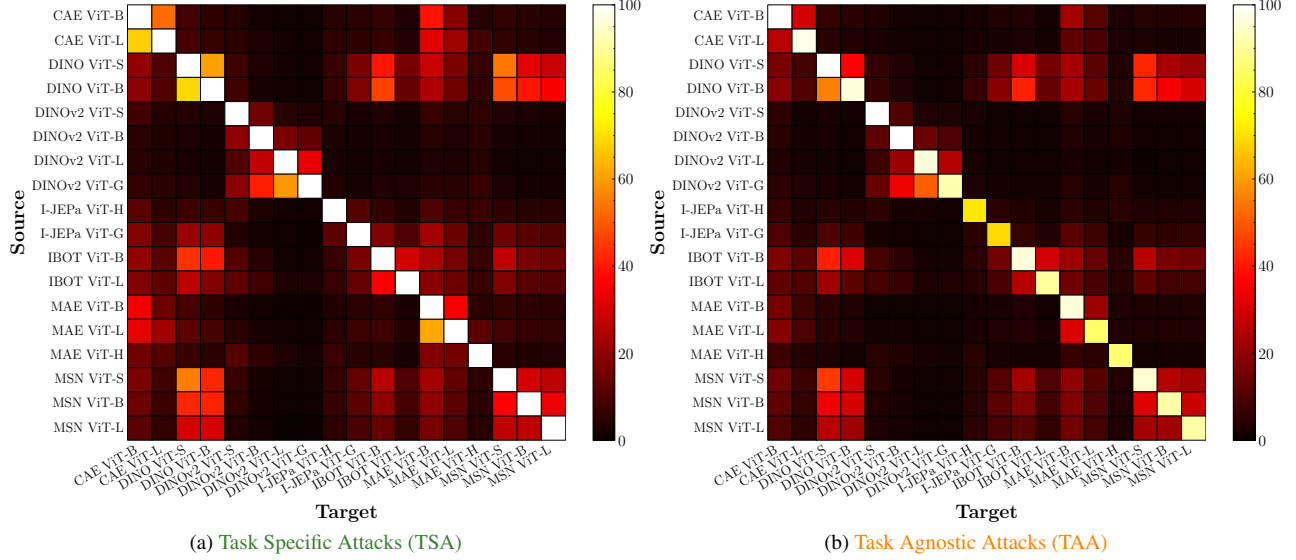


Figure 5. Comparison of relative efficiency (2) of TAAAs (right) with respect to TSAs (left) averaged over classification and semantic segmentation tasks. TAAAs perform comparably to TSAs across models and tasks.

## 6.2. Transferability to fine-tuned VFMs

While it is rare for practitioners to pre-train a VFM, it is common practice nowadays to fine-tune an existing open-source VFM for a specific task. Parameter efficient fine-tuning techniques such as LoRA and variants [17, 27, 35] have become increasingly popular, as they usually represent a more convenient solution compared to full fine-tuning. Learning more about the degree of transferability of TSAs and TAAAs crafted on the publicly available foundation models to their fine-tuned versions is crucial to better understand and design safer machine learning systems. To this end, we fully fine-tune and LoRA fine-tune three VFMs on classification for ImageNet. LoRA fine-tuning creates query and value matrices adapters in self-attention layers with rank  $r = 8$ , and enforces a dropout of 0.1 during training.

Tab. 7 highlights classification accuracies against TSAs and TAAAs using the original VFM as the source model to attack the fine-tuned target model. A LoRA-based fine-tuning is not sufficient to defend against either TSAs or TAAAs, however, the effectiveness of attacks against fully fine-tuned VFMs and downstream heads is greatly reduced.

## 7. Conclusion

This work investigates the robustness and security of the currently most popular pre-trained SSL vision foundation models. It introduces a family of adversarial example attacks that is task-agnostic and exploits publicly available foundation models to craft adversarial attacks that maximally perturb their feature representation.

Firstly, we find that a TAA performs comparably to its

Table 7. Classification accuracy over ImageNet for fine-tuned models and transferable Task-Specific / Task-Agnostic Attacks.

Attack	Backbone	No FT	LoRA FT	Full FT
TAA	DiNOv2	0.2	10.4	92.4
TAA	MAE	6.2	15.6	87.3
TAA	MSN	9.9	14.0	84.5
TSA	DiNOv2	4.4	5.0	92.0
TSA	MAE	5.9	6.1	77.1
TSA	MSN	5.1	6.0	75.7

TSA counterpart, while transferring better to other tasks. Secondly, we show that TAAAs are capable of disrupting classification, semantic segmentation, zero-shot classification, image captioning, visual question answering, and image retrieval systems. Finally, when it comes to transferability across different or fully fine-tuned models, we observe low transferability for both TSAs and TAAAs, however, we find that fine-tuning through low rank adapters does not protect from attacks crafted on the non-tuned publicly available model.

This research aims to enhance AI security by identifying vulnerabilities, promoting defenses that protect models in open-source environments: (i) **Awareness for Accelerated Defense.** By sharing these findings, we aim to drive proactive research on defenses, reducing potential risks; (ii) **Limited Risk Scope.** Our attack lacks resilience to transformations (e.g., resizing, flipping), limiting its real-world impact; (iii) **Potential Defensive Strategies.** Adversarial security can be achieved through adversarial training [25, 29, 46] or could be incorporated ad hoc [10].

## References

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 6
- [2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *Computer Vision – ECCV 2022*, pages 456–473, Cham, 2022. Springer Nature Switzerland. 2
- [3] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023. 2
- [4] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning, 2023. 2
- [5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. 2
- [6] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024. 6
- [7] Benoit Bonnet, Teddy Furon, and Patrick Bas. Generating Adversarial Images in Quantized Domains. *IEEE Transactions on Information Forensics and Security*, 2022. 3
- [8] Rui Cao and Jing Jiang. Modularized zero-shot vqa with pre-trained models. *arXiv preprint arXiv:2305.17369*, 2023. 1
- [9] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. 3
- [10] Nicholas Carlini, Florian Tramèr, Krishnamurthy Dvijotham, Leslie Rice, Mingjie Sun, and Zico Kolter. (certified!!) adversarial robustness for free! *International Conference on Learning Representations (ICLR)*, 2023. 8
- [11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 2
- [12] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1739–1747, 2020. 3
- [13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5, 6
- [14] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning, 2022. 2
- [15] Prakash Chandra Chhipa, Johan Rodahl Holmgren, Kanjar De, Rajkumar Saini, and Marcus Liwicki. Can self-supervised representation learning methods withstand distribution shifts and corruptions?, 2023. 3
- [16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2
- [17] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024. 8
- [18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>. 4
- [19] Pierre Fernandez, Matthijs Douze, Hervé Jégou, and Teddy Furon. Active image indexing. *arXiv preprint arXiv:2210.10620*, 2022. 1
- [20] Pierre Fernandez, Alexandre Sablayrolles, Teddy Furon, Hervé Jégou, and Matthijs Douze. Watermarking images in self-supervised latent spaces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3054–3058. IEEE, 2022. 1
- [21] Stanislav Fort. Adversarial examples for the openai clip in its zero-shot classification regime and their semantic generalization, 2021. 6
- [22] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2
- [23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Evaluating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4, 6
- [24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 2
- [25] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve

- model robustness and uncertainty. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3, 8
- [26] Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, 2019. 4, 7
- [27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 8
- [28] Nathan Inkawich, Gwendolyn McDonald, and Ryan Luley. Adversarial attacks on foundational vision models. *arXiv preprint arXiv:2308.14597*, 2023. 3
- [29] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020. 3, 8
- [30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 6
- [31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 6
- [32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 6
- [33] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1
- [34] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023. 1
- [35] Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. Relora: High-rank training through low-rank updates. In *The Twelfth International Conference on Learning Representations*, 2023. 8
- [36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 6
- [37] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 1
- [38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 3, 4
- [39] Thibault Maho, Teddy Furon, and Erwan Le Merrer. Surfree: a fast surrogate-free black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10430–10439, 2021. 3
- [40] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022. 6
- [41] Ron Mokady, Amir Hertz, and Amit H Bermano. Clip-cap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 1
- [42] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 3
- [43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1, 2
- [44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 4
- [45] Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast minimum-norm adversarial attacks through adaptive norm constraints. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 20052–20062, 2021. 3
- [46] Brian Pulfer, Yury Belousov, and Slava Voloshynovskiy. Robustness tokens: Towards adversarial robustness of transformers. In *European Conference on Computer Vision*, pages 110–127. Springer, 2024. 8
- [47] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *CVPR*, 2018. 5
- [48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1
- [49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervi-

- sion. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6
- [50] Javier Rando, Nasib Naimi, Thomas Baumann, and Max Mathys. Exploring adversarial attacks and defenses in vision transformers trained with DINO. In *New Frontiers in Adversarial Machine Learning Workshop, ICML, 2022*. 3
- [51] Md Farhamdur Reza, Ali Rahmati, Tianfu Wu, and Huaiyu Dai. CGBA: curvature-aware geometric black-box attack. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 124–133. IEEE, 2023. 3
- [52] Vasu Sharma, Ankita Kalra, and Labhesh Patel. Attend and attack: Attention guided adversarial attacks on visual question answering models. In *Secure Machine Learning Workshop, NeurIPS, 2018*. 3
- [53] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014*. 2
- [54] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanov, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. 6
- [55] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbinson, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshov, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshtir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. 6
- [56] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjointing large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022. 1
- [57] Giorgos Tolias, Filip Radenovic, and Ondrej Chum. Targeted mismatch adversarial attack: Query with a flower to retrieve

- the tower. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3
- [58] Ali Varamesh, Ali Diba, Tinne Tuytelaars, and Luc Van Gool. Self-supervised ranking for representation learning, 2020. 2
- [59] Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. Locca: Visual pretraining with location-aware captioners. *arXiv preprint arXiv:2403.19596*, 2024. 6
- [60] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks (2023). URL <https://arxiv.org/abs/2311.06242>. 6
- [61] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017. 3
- [62] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. Exact adversarial attack to image captioning via structured output learning with latent variables. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4135–4144. 3
- [63] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 6
- [64] Koichiro Yamanaka, Ryutaroh Matsumoto, Keita Takahashi, and Toshiaki Fujii. Adversarial patch attacks on monocular depth estimation networks. *IEEE Access*, 8:179094–179104, 2020. 3
- [65] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 1
- [66] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1
- [67] Junhao Zheng, Chenhao Lin, Jiahao Sun, Zhengyu Zhao, Qian Li, and Chao Shen. Physical 3d adversarial attacks against monocular depth estimation in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24452–24461, 2024. 3
- [68] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022. 2