As Firm As Their Foundations: Can Open-Sourced Foundation Models be Used to Create Adversarial Examples for Downstream Tasks?

Anjun Hu, Jindong Gu, Francesco Pinto, Konstantinos Kamnitsas, Philip Torr
University of Oxford
anjun.hu@eng.ox.ac.uk

Abstract

Foundation models pre-trained on web-scale vision-language data, such as CLIP, are widely used as cornerstones of powerful machine learning systems. While pre-training offers clear advantages for downstream learning, it also endows downstream models with shared adversarial vulnerabilities that can be easily identified through the open-sourced foundation model. In this work, we expose such vulnerabilities in CLIP's downstream models and show that foundation models can serve as a basis for attacking their downstream systems. In particular, we propose a simple yet effective adversarial attack strategy termed Patch Representation Misalignment (PRM). Solely based on open-sourced CLIP vision encoders, this method produces adversaries that simultaneously fool more than 20 downstream models spanning 4 common vision-language tasks (semantic segmentation, object detection, image captioning and visual question-answering). Our findings highlight the concerning safety risks introduced by the extensive usage of public foundational models in the development of downstream systems, calling for extra caution in these scenarios.

1. Introduction

Foundation models that combine both vision and language modalities are becoming increasingly popular, with CLIP [5] standing out as a prime example. CLIP is extensively used by various downstream models, offering comprehensive cross-modality semantics to enhance a wide range of tasks such as open-vocabulary segmentation (OVS), open-vocabulary object detection (OVD), image captioning (IC) and visual question answering (VQA).

While the usage of CLIP yields unprecedented performance improvements across many tasks, in this work, we show that such practice also introduces additional safety risks in downstream systems: since downstream models rely on the semantics acquired during pre-training, adversarial perturbations that can disrupt such pre-trained semantics could significantly hamper these models' performance regardless of their architectural design or intended tasks. This implies that these adversaries tend to exhibit high *transferability* across downstream models, which poses significant safety risks. Such risks are exacerbated by the

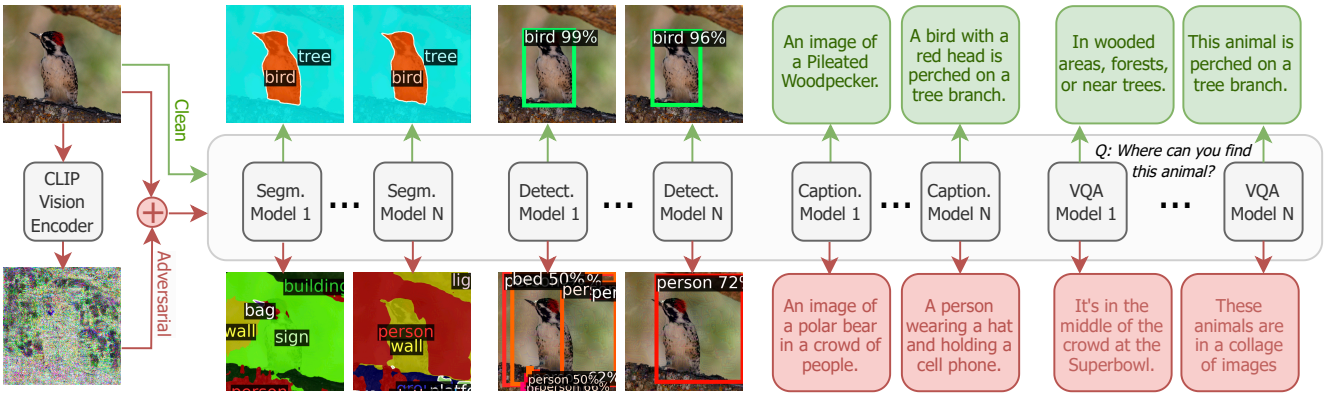

Figure 1. Given a clean image (top left), attackers can leverage the open-sourced CLIP vision encoder to find imperceptible input perturbations (bottom left, magnified by 30×) that distort CLIP's intermediate features. These perturbations are added to the original image to construct an adversarial sample that can simultaneously fool many downstream models intended for various tasks: downstream models that are highly performant on clean samples (top row) suffer significant performance degradation (bottom row) under such attacks.

fact that these highly transferable adversaries can be easily found through the publicly available CLIP model, rendering the downstream systems particularly susceptible to malicious exploitation.

To this end, we demonstrate the extent of such susceptibility by designing an adversarial attack strategy that creates highly transferable cross-task adversaries using only publicly available CLIP vision encoders. Notably, we show that an attacker with access to an off-the-shelf version of the CLIP vision encoder can substantially compromise the integrity of a wide range of downstream target models, including those that use an encoder with a different architecture than the one accessed by the attacker. Our contributions can be summarised as follows:

- We show that adversarial examples built with off-the-shelf CLIP vision encoders can fool a wide range of downstream models in a task-agnostic manner, evidencing the propagation of adversarial vulnerability from open-sourced foundation models to downstream systems as a significant safety threat.
- We introduce Patch Representation Misalignment (PRM), an alarmingly effective adversarial attack strategy which seeks input perturbations that distort CLIP vision encoders' intermediate representations to undermine the predictive capability of downstream systems.
- Comprehensive experiments are conducted to show that this technique induces significant performance degradation in more than 20 target models spanning 4 common vision-language tasks, outperforming competing baselines [2–4] by a significant margin.

2. Methods

Our cross-task attack strategy, Patch Representation Misalignment (PRM), centres around two core ideas: (1) attacking the intermediate features of a strong foundation vision model for cross-task generalisability and (2) using cosine similarity to divert adversarial patch representations from their clean counterparts.

2.1. Preliminaries: Notations and Threat Model.

Given a clean input image sample x, a prediction ground truth y and a target model $\mathcal{M}$ which is under attack, we consider an untargeted adversarial attack which aims at producing an adversarial sample $x' = x + \delta$ by adding an imperceptible perturbation δ to x such that the target model $\mathcal{M}$ would produce an incorrect output prediction when given the adversarial sample x' as input, i.e. $\mathcal{M}(x') \neq y$. To ensure that the attacks remain imperceptible, δ is often constrained by a perturbation budget ϵ such that the adversary stays within the ϵ-neighbourhood of the original input. In this work, we specifically consider L_∞-norm bounded attacks (i.e. $\|x - x'\|_\infty \leq \epsilon$). We assume that the attacker cannot query $\mathcal{M}$ nor access the parameters or gradients of $\mathcal{M}$. The attacker, however, has thorough knowledge about another model $\mathcal{F}$, referred to as a *surrogate* model, on which the perturbations are based. Note that this threat model does

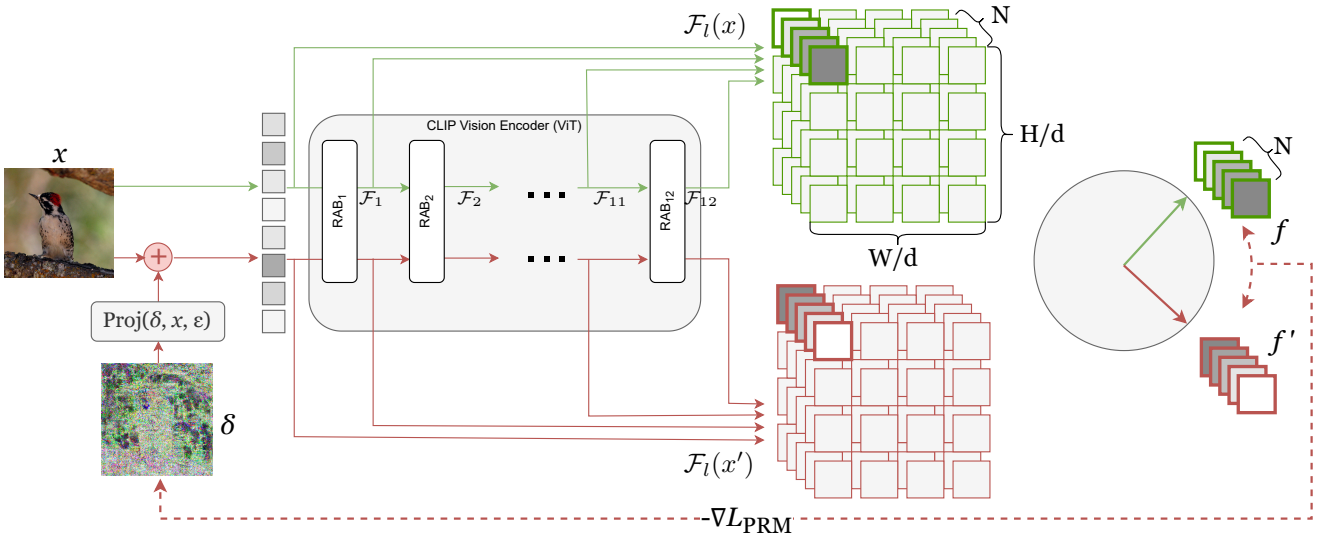

Figure 2. Overview of our attack pipeline. A normal forward pass with clean input is marked in green whereas the forward pass of the adversarial sample is marked in red. Dashed line indicates the flow of loss gradients which are used to update the injected adversarial perturbation. The loss objective minimises the cosine similarity between the adversarial representation of each patch (token) f' and its clean counterpart f along the embedding (ViT) or channel (CNN) dimension of the features. This approach individually diverts each patch representation (indicated by the reversed intensity of the top-left patch representation) to induce semantic distortions in all image regions.

not exclude the possibility of $\mathcal{F}$ and $\mathcal{M}$ having overlapping submodules (e.g. due to shared modules acquired from publicly available foundational models).

2.2. Patch Representation Misalignment (PRM)

As illustrated in Figure 2, our approach uses off-the-shelf CLIP vision encoders as surrogates and creates adversarial examples by iteratively optimising for input perturbations that induce maximal distortion of the surrogate model's internal representations. Given a surrogate vision encoder $\mathcal{F}$ with attackable layers L, we denote intermediate features of a clean sample from encoder layer l as $\mathcal{F}_l(x)$ and those of an adversarial sample as $\mathcal{F}_l(x')$. For transformer-based vision encoders built with a series of residual attention blocks, $\mathcal{F}_l(x), \mathcal{F}_l(x') \in \mathbb{R}^{1+\lceil \frac{HW}{d^2} \rceil \times N}$. Each of them can be expanded as a set of N-dimensional embeddings for individual tokens $\mathcal{F}_l(x) = \{f_l^0, f_l^1, \ldots, f_l^{\lceil \frac{HW}{d^2} \rceil} | f \in \mathbb{R}^N\}$, corresponding to representation of image patches of size d and one global `CLS` token. We treat each token (patch or `CLS`) as an independent sample and minimise the cosine similarity between the adversarial embedding f' of each token and its clean counterpart f, as outlined in Eq. (1). This objective drives adversarial token representations away from their clean counterparts and thereby misaligns them with the correct semantics:

$$\mathcal{L}_{\text{PRM}} = \sum_{l \in L} \sum_{p=0}^{\lceil \frac{HW}{d^2} \rceil} \frac{f_l^p \cdot f_l'^p}{\|f_l^p\| \|f_l'^p\|} \tag{1}$$

It is straightforward to adapt this loss for convolutional vision encoders. In this case, clean and adversarial intermediate features from layer l are $\mathcal{F}_l(x), \mathcal{F}_l(x') \in \mathbb{R}^{\lceil \frac{H}{d} \rceil \times \lceil \frac{W}{d} \rceil \times N}$ where N is the number of filters (channels) and d is a downsampling factor. Rather than viewing the latent features as N *feature maps*, we regard them as a collection of N-dimensional *descriptors* for each spatial element in the features. A similar minimisation of cosine similarity between clean and adversarial descriptors $f \in \mathbb{R}^N$ can be done via Eq. (1) for convolutional vision encoders.

3. Experiments and Evaluation

Target Tasks and Models. To evaluate attack efficacy and transferability, we choose 20 target models across 4 common vision-language understanding tasks (Figure 4): Open-Vocabulary Semantic Segmentation (OVS), Open-Vocabulary Object Detection (OVD), Image Captioning (IC) and Visual Question-Answering (VQA).

Baselines. We assess four attack strategies, including our own, each with two options for surrogate models: (1) Projected Gradient Descent (PGD) [3] with a training loss maximisation objective using two OVS models as surrogates;

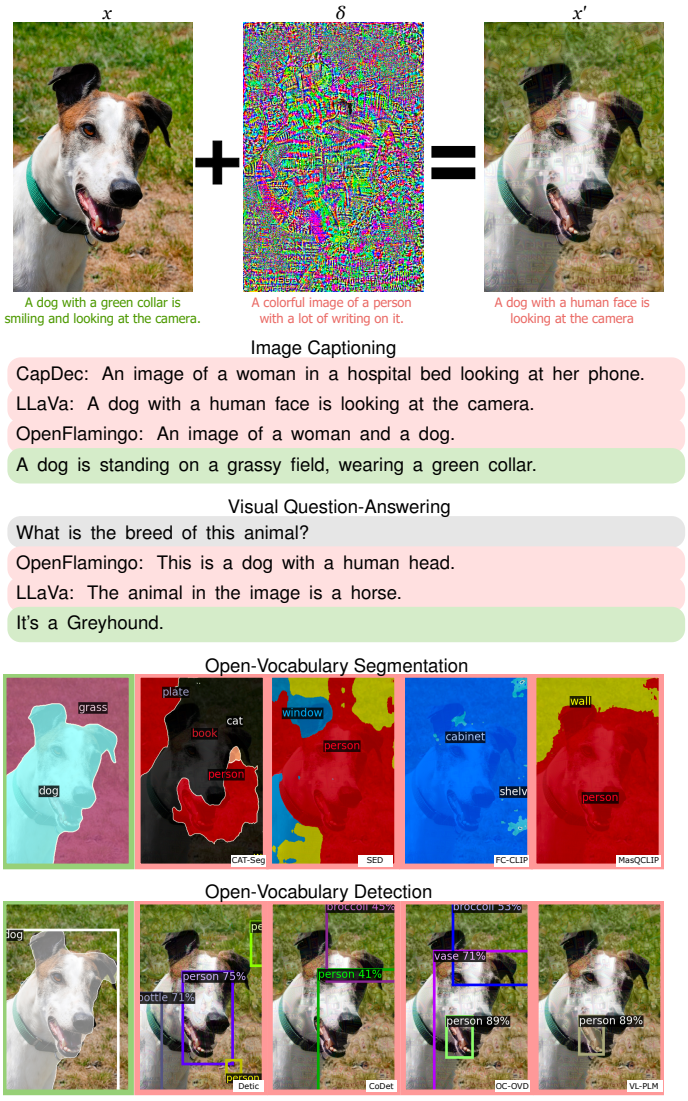

Figure 3. Qualitative results: an adversarial example created by PRM (using ViT-B/16 CLIP vision encoder as a surrogate) can fool a wide range of downstream models across various tasks.

(2) Neural Representation Distortion (NRD) [4], (3) Dispersion Reduction (DR) [2] and (4) PRM (ours) using ViT-B/16 and ConvNeXt-L CLIP vision encoders as surrogates. Further details regarding baseline methods, target tasks, target models, datasets, metrics and implementation can be found in Appendix 6.

Discussions. Exemplified by Figure 3, adversaries created with PRM can effectively deceive a comprehensive collection of victim models with significantly different training paradigms across all four tasks of interest. Figure 4 gives an overview of the relative performance of PRM (red line) and baseline methods, showing that the former outperforms the latter by a significant margin. Despite prior research showing limited transferability of adversarial examples between

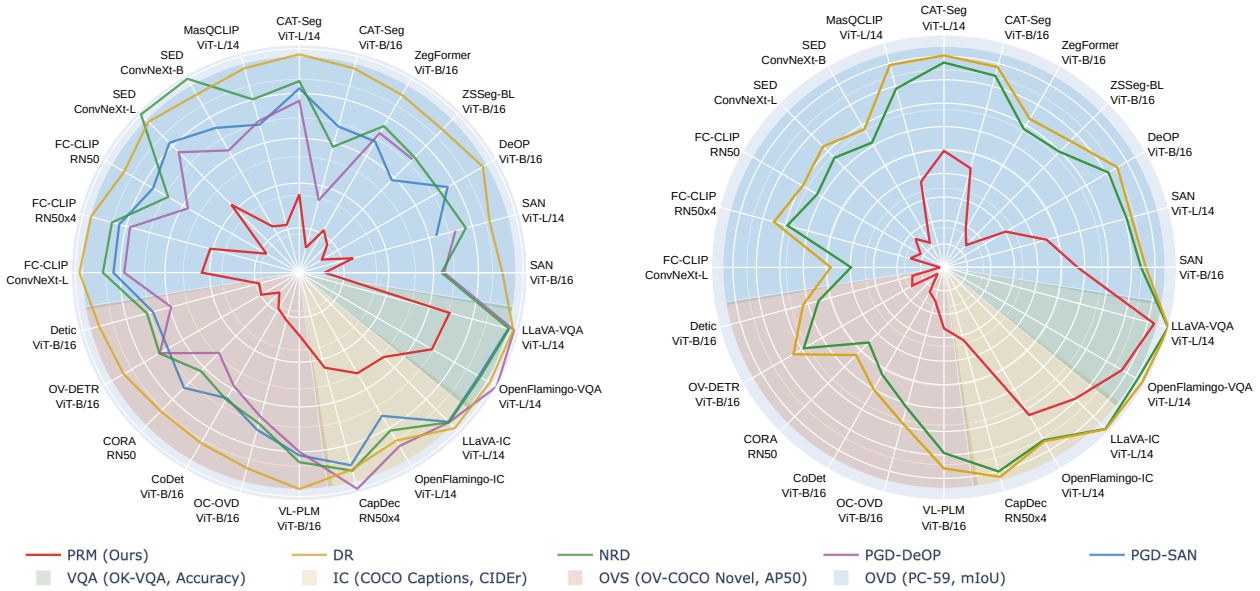

Figure 4. Normalised target model performance (model performance metrics under adversarial attacks divided by metrics on clean samples) of various attack strategies. **Left:** using ViT-B/16 (or task-specific baselines that use ViT-B/16 backbone) as surrogates. **Right:** using ConvNeXt-L as surrogates. PRM (red line) outperforms baseline methods by a significant margin across all tasks with both surrogate choices. The radius of the outer circles represents model performance on clean samples (unitary normalised metrics). Each attack strategy corresponds to a line. Each task is indicated by a differently coloured sector (datasets and metrics used for each task are detailed in the 2$^{\text{nd}}$ line of the legend). Target model names are annotated on the periphery of the circles. White-box scenarios are excluded.

models with significantly different parameters and architectures [1, 6], we notice that substantial transferability can be attained even when a surrogate has a vision encoder architecture that does not match that of the target model. These observations suggest a significant overlap of non-robust features [6] across vision encoders of various architectures, which is likely due to the common vision-language alignment pre-training process. More in-depth discussions regarding our empirical results can be found in Appendix 7.

4. Conclusions

This work aims to raise awareness about a currently overlooked but widespread safety vulnerability introduced by the usage of open-sourced foundation models, such as CLIP, to its downstream applications. We show that the downstream models' reliance on pre-trained features can be exploited by attackers, who may devise extremely effective adversarial attacks using solely open-sourced foundation models. We demonstrate that feature-based attacks such as PRM can significantly compromise the performance of the downstream models independently of the type of task or architecture they adopt. While our investigation focuses on CLIP's downstream systems, the alarming efficacy of PRM warrants additional efforts to determine whether similar phenomena apply to other foundational models. We hope our work will encourage further exploration into the safety implications of using foundation models in downstream learning, as well as inspire the development of effective defence strategies or robust learning approaches.

References

[1] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 888–897, 2018. 4

[2] Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 940–949, 2020. 2, 3

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3

[4] Muzammal Naseer, Salman H Khan, Shafin Rahman, and Fatih Porikli. Task-generalizable adversarial attack based on perceptual metric. *arXiv preprint arXiv:1811.09020*, 2018. 2, 3

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[6] Futa Waseda, Sosuke Nishikawa, Trung-Nghia Le, Huy H Nguyen, and Isao Echizen. Closer look at the transferability of adversarial examples: How they fool different models differently. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1360–1368, 2023. 4

As Firm As Their Foundations: Can Open-Sourced Foundation Models be Used to Create Adversarial Examples for Downstream Tasks?

Supplementary Material

5. Related Works

5.1. Adversarial Transferability.

Adversarial attacks introduce imperceptible perturbations to models' input data with the intention of causing degradation in model performance [65]. Early attack strategies such as [17, 40] operate under the assumption of having white-box access to the target model. More recent works [31, 51] point out that such assumptions about target model accessibility are often unrealistic and subsequently show that it is possible to craft adversarial perturbations that *transfer* across multiple models in a black-box fashion (i.e., without requiring access to the target's parameters or gradients). Such phenomenon of adversarial transferability, being the basis for black-box transfer attacks, has thus gained substantial research interest due to its significant implications for model safety [19, 90]. Crafting transfer-based attacks typically requires access to a surrogate model that shares some similarity with the target model, which could be in terms of model architecture, optimisation objectives, or training data. Having significant differences between the surrogate and the target model in these aspects may result in limited attack transferability [3, 71] though many recent efforts have been made to foster stronger transferability despite these differences [9, 44–46, 53, 55, 64, 69, 75, 78, 81, 86, 88]. In this work, we show that the extensive use of open-sourced foundation models in the development of other systems creates vulnerabilities that can be maliciously leveraged to improve attack transferability among downstream systems that exhibit significant differences in architectures and training paradigms.

5.2. Feature-Based Adversarial Attacks.

Inspired by prior works on cross-task attacks [35, 43, 60] and feature-based attacks [16, 21, 22, 24, 61, 61, 70], we design a simple yet highly effective attack strategy that creates adversarial examples by finding input perturbations that induce maximal distortions in the surrogate model's internal representations. Different from prior works that focus on distorting the magnitude or variance of model features, we introduce a more refined objective that treats each patch independently and emphasises patch-wise directional misalignment between clean and adversarial features. This design serves to induce stronger semantic distortion in the CLIP feature space that densely covers all image regions and thereby produces more destructive adversaries compared to baseline methods.

5.3. Foundation Models as Attack Surrogates.

Recently, there has been a growing focus on investigating and improving the robustness of vision-language foundation models [6, 49, 54, 62, 92, 94]. In particular, several studies have proposed strategies that use vision-language foundation models to devise attacks that transfer to other foundation models [33, 63, 85, 93]. While existing works primarily concern the robustness and adversarial vulnerabilities of foundation models themselves, we focus on the fact that foundation models can serve as a basis for attacking their downstream systems. Our study extends the investigation on foundation model robustness to their downstream systems, which frequently engage in direct interactions with end users and can have a substantial negative impact on real-world applications if their integrity is compromised.

5.4. Safety Implications of Pre-training.

Several existing works have explored the inheritance of certain robustness properties in various transfer learning settings [20, 60, 68, 80, 87]. In particular, how various forms of weaknesses and limitations inherited from pre-training can influence downstream tasks [8, 47] have garnered growing research interest. Building upon this line of investigation, our study represents the first attempt to demonstrate the propagation of adversarial vulnerabilities from foundation models to downstream models, remarking the trade-off between utility and safety when using foundation models in the development of downstream systems.

6. Experimental Setup

6.1. Baselines.

Since there are no well-established and directly comparable baselines that conduct task-generalisable attacks with non-CNN surrogates, we choose three adversarial loss objectives originally developed in other contexts as our baseline methods. We apply minor adaptations to these baselines to fit our problem setting and test each of them on two surrogate choices.

1. **Task-Specific Baseline: Training Loss Maximisation** is the conventional approach taken by the original PGD[40] method which requires a task-specific surrogate and access to ground truths for computing the surrogate loss. We choose two open-vocabulary segmentation models SAN [77] and DeOP [1] as the task-specific surrogate model and use the corresponding negated surrogate training loss (denoted as $-\mathcal{L}_{\text{SAN}}$ or $-\mathcal{L}_{\text{DeOP}}$ in

tables) as the adversarial loss objective.

2. **Cross-Task Baseline 1: Neural Representation Distortion (NRD)** [43] is a cross-task attack strategy based on maximising perceptual dissimilarity between clean and adversarial samples, measured by L_2 distance between VGG-16 conv3.3 features [26]. To make it a comparable baseline applicable to CLIP vision encoders, we simply adapt their objective $\mathcal{L}_{\text{NRD}} = -\sum_{l \in L} \|\mathcal{F}_l(x) - \mathcal{F}_l(x')\|_2$ to ViT and ConvNeXt intermediate features and attack features from all encoder layers (setting L to be all layers).
3. **Cross-Task Baseline 2: Dispersion Reduction (DR)** [35] is a cross-task attack strategy that manipulates adversarial features from one selected CNN layer to become "less dispersed" by minimising their standard deviation. To make it a comparable baseline, we similarly adapt their loss objective $\mathcal{L}_{\text{DR}} = -\sum_{l \in L} \sigma(\mathcal{F}_l(x'))$ to ViT and ConvNeXt intermediate features and attack features from all layers.

6.2. Implementation Details.

We set a perturbation budget of $\epsilon = 8/255$ across all experiments and consider untargeted attacks. All experiments are boosted with random resizing as input augmentation [74, 95] with scaling factors ranging in (0,75, 1.25) and optimized for $N = 250$ (converged) iterations with AdamW [32] optimizer at a learning rate of $\alpha = 0.5/255$. We present our strategy using either ViT-B/16 or ConvNeXt-L CLIP vision encoders as surrogates with the same loss objective in Eq.1. Attackable layers L are set to all 12 Residual Attention Blocks for the former or all ConvNeXtBlocks in the 3 ConvNeXt stages (excluding stem) for the latter. Using two surrogates helps us to (1) validate the efficacy of our loss design on different architectures; (2) examine transferability in scenarios where the vision encoder architecture of the surrogate matches or mismatches that of the target models' vision encoder.

6.3. Tasks, Datasets and Metrics

Target Task 1: Open Vocabulary Semantic Segmentation (OVS). The efficacy of our attack on OVS target models is evaluated using two popular semantic segmentation benchmarks: Pascal Context [42] and COCO-Stuff [5]. COCO-Stuff is a large-scale dataset for semantic segmentation consisting of 118k training images, 5k validation images and 41k testing images with 171 annotated classes. Pascal Context is another semantic segmentation benchmark containing 5k training images and 5k validation images; we use the 59-class label set (PC-59) in our evaluation. We attack eleven SoTA OVS models of which eight are densely supervised by pixel-level annotation (SAN[77], DeOP[1], ZSSeg-Baseline[76], ZegFormer[14], CAT-seg[12], MasQCLIP[79], SED[73], FC-Clip[83]) and three are weakly supervised by image-level labels or text (SegCLIP[36], TCL[7], SimSeg[82]). Target model performance on the two validation sets is measured with the mean of class-wise intersection over union (mIoU) metric. Results are reported in Table 1a for PC-59 and Table 1b for COCO-Stuff.

Target Task 2: Open Vocabulary Object Detection (OVD). For evaluating attack efficacy on OVD target models, we use the OV-COCO [29] dataset which consists of 107761 training images and 4836 validation images. We adopt the widely used base-novel split of class labels, where 48 base classes are used for training and an additional 17 novel classes are introduced during inference. We attack six OVD target models (VL-PLM[89], Detic[91], CORA[72], CoDet[38], object-centric-ovd[39, 58], OV-DETR[84]). Target model performance is measured on the OV-COCO validation set using the AP50 (the average precision at an IoU of 50%) metric. We report model performance on base and novel classes separately in Table 2.

Target Task 3: Image Captioning (IC). We generate adversarially perturbed versions of COCO Captions [11] dataset and evaluate attack efficacy on three victim models (CapDec[50], OpenFlamingo[4], LLaVA[30]) using the 5000-sample test set in Karpathy splits. As shown in Table 3, we evaluate caption quality with several commonly adopted metrics: BLEU score (B@1, B@4) [52], METEOR (M) [13], ROUGE-L (R-L) [28] and CIDEr (C) [67].

Target Task 4: Visual Question Answering (VQA). We generate adversarially perturbed versions of OK-VQA [59] and evaluate attack efficacy on two victim models (OpenFlamingo[4], LLaVA[30]), as shown in Table 3. Target model performance is measured with 5046 validation questions from OK-VQA using the VQA accuracy metric.

7. Discussions

Inheritance of Adversarial Vulnerability. As shown in Table 1-3, our attack transfers to a comprehensive collection of victim models with significantly different training paradigms across all four tasks of interest. Our method (PRM) produces significantly more effective adversaries in comparison to baseline methods with both surrogates, as highlighted by the blue numbers in the tables. This is also clearly illustrated by Figure 4 where the red line (PRM) shows a substantial shrinkage in performance across all target tasks and models. This provides strong empirical evidence that downstream "child" models could inherit adversarial vulnerabilities from "parent" foundation models, rendering them more susceptible to transfer attacks created with the parent models.

Transferability Across Vision Encoder Variants. Despite prior research showing limited transferability of adversarial examples between models with significantly different parameters and architectures [3, 71], we notice that substantial transferability can be attained even when a surrogate has a vision encoder architecture that does not match that of the target model. White cells in the tables indicate such scenarios, exemplified by the last three rows in the first half of Table 1. For some target models, a mismatched surrogate model can even produce stronger attacks than a matched surrogate: ZSSeg-BL[76] and ZegFormer[14] (columns 4-5, Table 1) use ViT-B backbones but adversaries created with ConvNeXt-L cause more harm to them. Notably, none of the VQA and captioning target models share the same backbone as either surrogate yet our attack still induces significant performance degradation, as highlighted by the blue numbers in Table 3. There is also a particularly interesting observation among OVD victim models in Table 2: all OVD target models have either ViT-B/16 or RN50 encoder backbones, yet the most effective attacks for all of them are crafted with ConvNeXt-L encoders. These observations suggest a significant overlap of non-robust features [71] across vision encoders of various architectures, which is likely due to the common vision-language alignment pre-training process.

Transferability Across Tasks. While significant performance degradation is observed on all four tasks of interest, our attack demonstrates particularly strong efficacy among dense predictors (OVS and OVD models in Tables 1-2), where the impact of full-coverage dense semantic distortion induced by our patch-wise cosine similarity minimisation objective is more prominent. Curiously, as exemplified by Fig. 3, OVD target models frequently make false positive predictions of `persons`, which is likely due to biases in the training data [29] where `person` is the highest frequency class. PRM perturbations seem to incur semantically consistent mistakes across downstream models (e.g. most models made mistakes by hallucinating a false positive human in this scene). This is in line with observations made by prior works which show that certain adversarial perturbations can exhibit semantic meaning [23, 66]. Comparing the LLaVA captions of the clean image x, the adversarial sample x' and the perturbation δ, they seem to exhibit an additive relationship in terms of perceived semantic meanings. We also point out that OVD target models' performance on novel classes that are unseen during training ("N" columns in Table 2) suffer more performance degradation (our best attack yields an average adversarial performance equivalent to 29% of the clean performance on base classes and 14% of the clean performance on novel classes). As the models' generalisability to unseen classes relies heavily on the pre-trained cross-modality semantics of CLIP, it is expected that the disruption to pre-trained features impacts their performance on unseen classes to a larger extent. On the other hand, we notice that all strategies presented in this study have a relatively small impact on VQA target models: PRM is the only method that achieves substantial attack efficacy with an average adversarial performance equivalent to 67% of the clean performance while other methods generally cannot degrade VQA accuracy by more than 3%. This may be attributed to the fact that VQA models have a tendency to learn *linguistic shortcuts* [10, 18, 25, 48], a form of bias and spurious correlation in the language modality that reduces their reliance on pre-trained vision features. Developing a stronger attack for tasks that have a greater emphasis on language (e.g. by involving the language modality in attack optimisation [37] or explicitly leveraging contextual hints such as class co-occurrence information [2]) could be an interesting direction for future work.

8. Pseudocode

We use $\mathcal{F}_l(x), \mathcal{F}_l(x')$ to denote intermediate encoder features obtained from layer l. Each $\mathcal{F}$ can be expanded as a set of representations of individual image patches $\mathcal{F}_l(x) = \{f_l^0, f_l^1, \ldots, f_l^{\lceil \frac{HW}{d^2} \rceil}\} = \{f_l^p \in \mathbb{R}^N | p \in [0, \lceil \frac{HW}{d^2} \rceil]\}$, which are traversed through by the inner for loop.

Algorithm 1 Patch Representation Misalignment (PRM)

Require: Surrogate vision encoder $\mathcal{F}$ with attackable layers L, perturbation budget ϵ, clean image x, learning rate α and number of iterations N

```
/* Initialise perturbation and adversary */
δ_0 ~ U(−ϵ, ϵ); x'_0 ← Clip(x + δ_0, 0, 255)
for t ← 0 to N do
    L_PRM = 0
    /* Traverse through all layers */
    for l ∈ L do
        /* Traverse through all patches */
        for (f, f') in (F_l(x), F_l(x'_t)) do
            L_PRM += (f · f')/(||f|| ||f'||)
        end for
        /* Accumulate loss for patches in layer l */
    end for
    /* Accumulate loss for all attackable encoder layers */

    /* Update adversarial perturbation */
    δ_{t+1} ← δ_t − α · Sign(∇_{δ_t} L_PRM)
    /* Project perturbation to ϵ-ball */
    δ_{t+1} ← Clip(δ_{t+1}, −ϵ, ϵ)
    /* Project adversary to valid image range */
    x'_{t+1} ← Clip(x + δ_{t+1}, 0, 255)
end for
```

Table 1. Transfer attack efficacy on OVS target models. Model performance is measured in mean Intersection over Union (**mIoU**) where a lower metric indicates higher attack efficacy. Each row (S) represents a surrogate-loss configuration while each column (T) corresponds to a target model. Vision encoder backbones used by each model are specified below model names. We divide surrogate configurations into two groups (separated by double lines) mirroring the left and right radial plots in Figure 4: the first group uses ViT encoders (or task-specific model with ViT backbones) as surrogates; the second group uses ConvNeXt encoders as surrogates. Light grey cells indicate scenarios where the source and target models have matching vision encoder architectures. Excluding white-box attacks where source and target models are identical, the **best** attack strategy for each target model within each group is highlighted in **blue**.

(a) Pascal Context 59.

S \ T	$\mathcal{L}_{\text{attack}}$	SAN		DeOP	ZSSeg-BL	ZegFormer	CAT-Seg		MasQCLIP
		ViT-B	ViT-L	ViT-B	ViT-B	ViT-B	ViT-B	ViT-L	ViT-L
	Clean	54.07	57.70	48.65	46.92	41.27	57.4	61.97	58.65
SAN	$-\mathcal{L}_{\text{SAN}}$	7.12	36.54	37.21	27.42	27.94	38.84	51.00	40.15
DeOP	$-\mathcal{L}_{\text{DeOP}}$	35.00	41.68	4.57	33.59	29.69	19.25	47.57	41.10
CLIP ViT-B	DR	49.06	50.78	46.07	42.42	37.83	54.30	60.44	55.44
	NRD	34.34	44.45	35.29	34.17	31.20	33.46	53.01	47.00
	PRM	**6.13**	**14.37**	**5.69**	**8.26**	**9.05**	**6.64**	**21.63**	**12.93**
CLIP CNeXt-L	DR	46.57	47.76	41.54	35.77	30.18	50.84	56.00	52.39
	NRD	45.45	46.54	39.39	32.71	28.12	48.62	54.23	46.28
	PRM	**30.81**	**26.18**	**14.83**	**6.31**	**7.64**	**25.15**	**30.79**	**22.20**

S \ T	$\mathcal{L}_{\text{attack}}$	SED		FC-CLIP			SegCLIP	TCL	SimSeg	
		CNeXt-B	CNeXt-L	RN50	RN50×64	CNeXt-L	ViT-B	ViT-B	ViT-S	ViT-B
	Clean	57.67	60.87	51.80	55.15	56.99	24.25	33.9	25.81	26.18
SAN	$-\mathcal{L}_{\text{SAN}}$	42.95	49.94	39.07	45.92	47.32	15.45	22.79	21.66	21.89
DeOP	$-\mathcal{L}_{\text{DeOP}}$	36.27	46.31	29.75	43.17	44.58	17.22	13.61	18.21	19.28
CLIP ViT-B	DR	52.81	58.01	46.77	53.14	55.95	12.64	23.94	24.70	25.71
	NRD	57.51	60.87	35.03	47.77	49.99	15.69	15.59	21.50	22.00
	PRM	**13.71**	**26.11**	**8.79**	**22.71**	**24.80**	**1.83**	**4.87**	**11.22**	**11.32**
CLIP CNeXt-L	DR	39.08	44.39	36.04	41.39	27.55	21.75	21.11	23.93	24.50
	NRD	35.33	40.22	32.33	38.14	22.60	21.46	20.29	23.68	24.19
	PRM	**6.99**	**10.44**	**5.97**	**8.17**	**0.93**	**12.64**	**13.10**	**18.94**	**19.38**

(b) COCO-Stuff.

S \ T	$\mathcal{L}_{\text{attack}}$	SAN		DeOP	ZSSeg-BL	ZegFormer	CAT-Seg	
		ViT-B	ViT-L	ViT-B	ViT-B	ViT-B	ViT-B	ViT-L
	Clean	35.76	43.91	38.24	35.76	34.09	46.17	50.35
SAN	$-\mathcal{L}_{\text{SAN}}$	4.57	30.21	28.20	24.19	22.87	28.93	37.50
DeOP	$-\mathcal{L}_{\text{DeOP}}$	25.94	34.66	2.85	22.23	23.61	16.55	39.52
CLIP ViT-B	DR	38.46	42.24	35.29	32.21	31.22	42.07	48.13
	NRD	26.44	37.80	24.33	26.45	26.39	27.25	44.23
	PRM	**3.89**	**12.85**	**3.48**	**7.06**	**7.47**	**4.23**	**22.44**
CLIP CNeXt-L	DR	37.64	40.82	32.65	27.08	25.61	38.91	46.18
	NRD	36.67	39.65	30.68	24.63	23.81	39.06	45.17
	PRM	**26.88**	**26.08**	**12.10**	**4.56**	**5.09**	**23.34**	**32.09**

S \ T	$\mathcal{L}_{\text{attack}}$	SED		FC-CLIP			SegCLIP	TCL	SimSeg	
		CNext-B	CNeXt-L	RN50	RN50×64	CNeXt-L	ViT-B	ViT-B	ViT-S	ViT-B
	Clean	46.21	49.57	54.85	60.53	63.22	26.04	20.32	27.24	29.73
SAN	$-\mathcal{L}_{\text{SAN}}$	32.50	39.34	42.43	52.14	55.28	13.12	14.09	19.53	21.20
DeOP	$-\mathcal{L}_{\text{DeOP}}$	29.50	37.56	33.29	49.31	52.98	21.57	9.93	25.18	26.39
CLIP ViT-B	DR	41.92	46.46	49.20	57.14	59.93	26.20	20.32	27.13	29.23
	NRD	34.88	42.53	39.43	52.46	55.52	20.14	10.05	24.28	26.23
	PRM	**11.57**	**22.49**	**13.25**	**30.90**	**34.39**	**7.14**	**2.69**	**18.31**	**19.98**
CLIP CNeXt-L	DR	31.55	34.48	35.00	40.40	25.16	24.61	17.70	26.68	28.46
	NRD	27.41	29.51	32.06	37.78	19.60	23.97	17.12	26.21	28.48
	PRM	**4.35**	**5.65**	**5.30**	**7.07**	**0.62**	**18.42**	**11.73**	**23.57**	**25.72**

Table 2. Transfer attack efficacy on OVD target models. Model performance is measured in **AP50**. Metrics of Base (B) and novel (N) classes are reported separately.

S \ T	$\mathcal{L}_{\text{attack}}$	Detic ViT-B/16		OV-DETR ViT-B/16		CORA RN50		CoDet ViT-B/16		OC-OVD ViT-B/16		VL-PLM ViT-B/16	
		B	N	B	N	B	N	B	N	B	N	B	N
	Clean	51.10	27.81	55.67	30.00	36.81	35.49	52.47	30.63	56.58	40.38	56.37	34.07
SAN	$-\mathcal{L}_{\text{SAN}}$	39.42	18.83	42.78	20.34	26.31	25.83	40.85	19.83	44.07	29.35	46.33	27.80
DeOP	$-\mathcal{L}_{\text{DeOP}}$	38.65	16.48	45.00	21.58	19.82	18.00	39.69	17.91	42.66	26.94	46.98	27.25
CLIP ViT-B	DR	47.50	25.74	52.83	27.18	31.94	31.07	48.69	26.92	52.20	36.51	54.37	32.94
	NRD	40.75	19.58	45.67	45.67	23.02	22.06	41.98	20.05	44.39	28.15	49.05	28.84
	PRM	**20.12**	**5.17**	**21.60**	**5.90**	**7.13**	**4.47**	**21.04**	**5.70**	**21.24**	**8.90**	**22.95**	**9.54**
CLIP CNeXt-L	DR	38.89	17.31	46.01	22.24	20.97	18.79	39.87	18.39	43.52	27.74	48.61	29.22
	NRD	36.75	15.41	44.26	20.71	19.03	16.09	37.97	16.16	40.87	24.96	46.93	26.98
	PRM	**15.37**	**3.87**	**17.12**	**4.69**	**3.39**	**1.35**	**15.61**	**3.71**	**15.99**	**5.90**	**22.01**	**8.88**

Table 3. Transfer attack efficacy on image captioning and visual question-answering (VQA) target models. Caption quality is measured with BLEU score (B@1, B@4) [52], METEOR (M) [13], ROUGE-L (R-L) [28] and CIDEr (C) [67]. VQA performance is measured with VQA accuracy. "VB" and "CL" stands for ViT-B/16 and ConvNeXt-L CLIP vision encoders as surrogates.

S \ T	$\mathcal{L}$	CapDec (RN50×4)					OpenFlamingo (ViT-L/14)						LLaVA (ViT-L/14)					
		B@1	B@4	M	R-L	C	B@1	B@4	M	R-L	C	VQA	B@1	B@4	M	R-L	C	VQA
	Clean	68.3	26.8	25.2	51.3	92.4	64.9	24.9	22.7	50.5	87.3	31.9	72.3	28.7	28.7	55.2	106.9	54.1
SAN		64.2	23.6	23.3	48.7	82.3	54.9	18.2	18.8	44.5	64.5	26.7	69.7	28.0	27.6	53.4	100.7	52.3
DeOP		66.6	26.5	24.5	50.4	93.1	56.6	21.0	20.5	47.2	78.1	30.9	70.4	28.0	27.9	54.0	101.1	53.8
V B	DR	66.2	24.9	23.9	49.7	84.1	57.1	21.0	20.5	47.3	75.6	29.8	72.2	28.8	28.4	55.0	105.0	53.6
	NRD	66.2	24.8	23.9	49.8	84.6	56.4	20.0	19.9	46.4	71.1	28.3	71.1	27.7	27.7	54.0	101.1	52.6
	PRM	**50.9**	**13.8**	**16.6**	**39.5**	**40.6**	**51.4**	**14.0**	**15.9**	**39.9**	**45.3**	**20.8**	**57.2**	**16.8**	**20.5**	**43.6**	**57.0**	**37.6**
C L	DR	66.2	24.9	24.1	49.9	85.5	56.9	20.6	20.4	47.0	74.8	29.7	71.9	28.4	28.2	54.7	104.4	53.7
	NRD	65.5	24.6	23.9	49.6	83.3	56.9	20.6	20.5	47.0	74.3	28.9	72.2	28.5	28.2	54.8	104.2	53.6
	PRM	**46.0**	**10.6**	**14.5**	**36.3**	**29.7**	**56.4**	**18.3**	**18.8**	**44.4**	**63.5**	**26.6**	**66.0**	**23.5**	**25.2**	**50.0**	**84.7**	**50.2**

9. Ablation Studies

In this section, we present several ablation studies to validate our adversarial loss design and explore different factors that may affect attack effecacy. All evaluations in the section are done on the COCO-Stuff [5] validation set. We present the performance metrics of a subset of OVS models though we remark that the trend in performance generalises to other tasks and datasets.

9.1. Cosine Similarity-Based Distortion

Could angular distance be a favourable distance metric in the intermediate feature space of vision-language alignment pre-trained encoders? To check this assumption, we repeat our experiments using architecturally identical non-CLIP vision encoders trained with classification objectives as surrogates. As shown in Table 4, while PRM still outperforms baselines, their performance gap shrinks notably (e.g. on selected OVS victim models, the performance gap between PRM and MSE-based NRD shrinks from 23% on CLIP vision encoders to around 11% on classification-trained encoders). This suggests that vision-language alignment pre-training could have played a role in assigning greater semantic importance to directional information over magnitude or variance (in comparison to classification cross entropy-trained alternatives).

To further validate the efficacy of our loss design, we individually examine the two core components of our loss: (1) the patch-wise approach and (2) the cosine similarity-based distortion. As shown in Table 5, the combination of them provides a further performance improvement.

In particular, we attempt an alternative usage of cosine similarity that flattens features before cosine similarity computation (i.e. globally distorting the entire volumetric feature using cosine similarity). While this configuration still outperforms MSE-based and variance-based baselines, it is not as performant as our default patch-wise setup which treats every patch descriptor as an independent sample and computes cosine similarity specifically along the embedding (ViT) or channel (CNN) dimensions of the features. This drop in performance indicates that the patch-wise approach tends to be more effective than the global flattened approach even when both are similarly based on cosine similarity. Moreover, the fact that Global-PRM still outperforms MSE-based alternative (NRD) provide further evidence that cosine-similarity based approach contributes to stronger attack performance.

9.2. Which features to attack

We choose to attack all tokens (or patches) from all vision encoder layers. This contrasts with previous studies that argue single-layer attacks (only distorting features from one

Table 4. Does vision-language alignment pre-training contribute to the emphasis on directional information in the feature space? The fact that cosine-similarity-based PRM has a larger performance gain over NRD (an MSE-based alternative) on alignment-pretrained encoders suggests this is likely a valid assumption.

T / S	$\mathcal{L}_{attack}$	CAT-Seg		SED		FC-CLIP		
		ViT-B	ViT-L	CNeXt-B	CNeXt-L	RN50	RN50×64	CNeXt-L
	Clean	46.17	50.35	46.21	49.57	54.85	60.53	63.22
CLIP ViT-B	NRD	27.25	44.23	34.88	42.53	39.43	52.46	55.52
	PRM	4.23	22.44	11.57	22.49	13.25	30.90	34.39
	Δ	23.02	21.79	23.31	20.04	26.18	21.55	21.13
Classification ViT-B	NRD	35.39	43.45	36.24	42.07	43.13	53.61	56.83
	PRM	24.44	36.92	25.47	35.46	33.57	49.56	52.37
	Δ	10.96	6.53	10.77	6.60	9.56	4.05	4.46

Table 5. To flatten or not? Through this set of ablation studies, we verified that (1) the patch-wise approach tends to be more effective than the global flattened approach when both are similarly based on cosine similarity; (2) cosine-similarity-based distortion (PRM) is more effective than MSE-based alternative (NRD) when both use a similar global (non patch-wise) approach.

T / S	$\mathcal{L}_{attack}$	CAT-Seg		SED		FC-CLIP		
		ViT-B	ViT-L	CNeXt-B	CNeXt-L	RN50	RN50×64	CNeXt-L
	Clean	46.17	50.35	46.21	49.57	54.85	60.53	63.22
CLIP ViT-B	**Patch-PRM**	**4.23**	**22.44**	**11.57**	**22.49**	**13.25**	**30.90**	**34.39**
	Global-PRM	12.07	34.64	21.27	32.90	22.63	41.82	45.38
	Global-NRD	27.25	44.23	34.88	42.53	39.43	52.46	55.52
CLIP CNeXt-L	**Patch-PRM**	**23.34**	**32.09**	**4.35**	**5.65**	**5.30**	**7.07**	**0.62**
	Global-PRM	24.63	33.29	5.22	7.00	6.70	8.72	0.66
	Global-NRD	39.06	45.17	27.41	29.51	32.06	37.78	19.60

Table 6. Which features to Attack? All experiments uses $\mathcal{L}_{\text{PRM}}$

T / S	Layer/Token	CAT-Seg		SED		FC-CLIP		
		ViT-B	ViT-L	CNeXt-B	CNeXt-L	RN50	RN50×64	CNeXt-L
	Clean	46.17	50.35	46.21	49.57	54.85	60.53	63.22
CLIP ViT-B	**AllLayers**	**4.23**	**22.44**	**11.57**	**22.49**	**13.25**	**30.90**	**34.39**
	`CLS`-only	22.24	41.42	29.98	39.54	34.29	49.36	54.08
	MidLayer	10.19	30.77	15.43	28.62	18.42	37.34	41.62
	LastLayer	11.18	34.97	23.76	34.66	26.33	43.72	48.26
CLIP CNeXt-L	**AllLayers**	**23.34**	**32.09**	**4.35**	**5.65**	**5.30**	**7.07**	**0.62**
	MidLayer	38.17	46.00	26.25	38.26	25.07	40.80	47.49
	LastLayer	34.42	41.23	16.94	17.72	18.63	20.15	2.75

surrogate layer) on mid-layer features would be sufficient [35, 43, 61]. To determine whether attacking all features is necessary, we attempt several alternatives that attack subsets of features and present results in Table 6 below.

1. Only attacking `CLS` tokens from all layers (ViT).
2. Only attacking the final layer features (both surrogates).
3. Only attacking mid-layer features (Layer 6 in ViT-B surrogates; Layer 6 of the first ConvNeXt Stage in ConvNeXt-L surrogates).

While attacking the full feature hierarchy results in the strongest attack, we acknowledge that the observation regarding the strong generalizability of mid-layer features remains valid to some extent: mid-layer features have higher generalizability than final-layer features in ViTs. Curiously, the second-best option is the last-layer attack for CNN-based surrogates whereas the mid-layer attack is the second-best option for ViT-based surrogates. Based on observations of the target models, we hypothesise that the improved transferability observed in multi-layer and full-spatial coverage attacks may be attributed to the following reasons:

1. Only attacking the global `CLS` token is insufficient since many downstream models (such as [12] that use CLIP dense features for cost volume computation) tend to prioritise dense patch representations over the global class token. This tendency is particularly notable in tasks involving dense predictions, such as segmentation and object detection.
2. The assumption that distortions in mid-layer features can propagate down the feature hierarchy to affect model decisions [35] may be less applicable due to more intricate modes of interactions between downstream models and pre-trained features. The interactions between the CLIP encoder backbone and other components of the downstream model take many intricate forms, some

Table 7. Attack performance on COCO-Stuff without random rescaling augmentation.

S \ T	$\mathcal{L}_{attack}$	CAT-Seg		SED		FC-CLIP		
		ViT-B	ViT-L	CNeXt-B	CNeXt-L	RN50	RN50×64	CNeXt-L
	Clean	46.17	50.35	46.21	49.57	54.85	60.53	63.22
CLIP ViT-B	DR	44.07	49.20	43.72	48.05	49.33	57.85	60.73
	NRD	43.76	49.21	43.47	47.97	48.57	57.84	60.48
	PRM	**11.18**	**34.97**	**23.76**	**34.66**	**26.33**	**43.72**	**39.04**
CLIP CNeXt-L	DR	44.38	49.47	42.61	46.43	45.82	55.58	4.32
	NRD	43.98	49.47	41.68	46.03	44.60	45.66	2.99
	PRM	**40.78**	**47.67**	**28.70**	**35.35**	**27.35**	**41.82**	**0.00**

Table 8. The effect of alignment pretraining. All experiments uses $\mathcal{L}_{\text{PRM}}$. Adversaries crafted with classification-trained encoders exhibit much lower transferability compared to those crafted with a vision-language alignment-pretrained vision encoder of the same architecture. Further fine-tuning on ImageNet can make alignment-pre-trained off-the-shelf CLIP vision encoders even stronger attack surrogates.

S \ T	CAT-Seg		SED		FC-CLIP		
	ViT-B	ViT-L	CNeXt-B	CNeXt-L	RN50	RN50×64	CNeXt-L
Clean	46.17	50.35	46.21	49.57	54.85	60.53	63.22
LAION V-L Align. ViT-B	**4.23**	**22.44**	**11.57**	**22.49**	**13.25**	**30.90**	**34.39**
ImageNet Class. ViT-B	24.44	36.92	25.47	35.46	33.57	49.56	52.37
LAION V-L Align. CNeXt-L	23.34	32.09	4.35	5.65	5.30	7.07	0.62
ImageNet Class. CNeXt-L	35.71	42.83	19.37	29.32	18.06	31.93	40.40
LAION+ImageNetFT CNeXt-L	**21.32**	**28.72**	**3.82**	**5.14**	**6.43**	**8.52**	**1.56**

downstream models go as far as using CLIP representations as prompt prefixes to query other foundation models [15, 27, 34, 41, 50, 57]. This means that we cannot assume the preservation of feature locality or spatial continuity of a model's internal representations throughout the downstream model's pipeline, which means certain feature distortions may not be propagated to the decision layer.

3. Some downstream models (such as [77]) selectively use certain encoder layers for specific tasks. Consequently, in a downstream model, inputs may not undergo the full forward pass of a pre-trained vision encoder; instead, only specific layers of the vision encoder's intermediate features are extracted for further processing. This implies that attackers cannot assume which intermediate features are important.
4. The hierarchy of multiple latent spaces from various layers of a vision encoder can be interpreted as a self-ensemble [45], which is believed to improve transferability by making the feature and loss landscapes more generalisable.

Therefore, we recommend attacking all tokens from all layers alike to ensure maximum coverage of perturbations, and to increase the likelihood of impacting the features used by the victim model.

9.3. The effect of input scale diversity

In the main experiments, we employ random rescaling augmentation for all competing loss objectives. We ablate this component and show results without input scale augmentation. Our method still outperforms baselins when all are examined without such augmentation. However, while augmentation is usually considered orthogonal to loss objective design, we do notice that input diversity magnifies the performance gap between our method and the baselines.

This technique [74, 95] significantly amplifies attack efficacy for all attacks by fostering scale invariance, which makes them more adept at fooling downstream models that operate at different input resolutions. Indeed, among CLIP's downstream models, some require that the input image be resized to a fixed scale [4] while others can accommodate dynamic scale inputs. Moreover, downstream models that handle dynamic scales but use the ViT-based CLIP vision encoder backbone employ diverse techniques (such as sliding windows [12] and learnable positional embeddings [1]) to address the discrepancy between the original ViT input resolution of the pre-trained backbone and the dynamic test-time resolution. By finding perturbations that effectively disturb model behaviours at various scales, we can expose the alarming efficacy of CLIP feature-based attacks on its downstream models.

9.4. The effect of alignment pretraining.

The fact that attacks crafted with either ViT or ConvNeXt surrogate can transfer to downstream models regardless of the target model's vision encoder architecture (e.g. that use a different vision encoder backbone, as discussed in Sec. 7) prompts us to consider how certain factors in the pre-training process may contribute to shared adversarial vulnerabilities across CLIP's downstream models.

Various factors in the pre-training process (e.g. data, training objective and optimisation techniques) can contribute to the generalisability of an encoder's feature space. To examine how different pretraining protocols affect an encoder's efficacy as an attack surrogate, we perform our attack using architecturally identical vision encoders pre-trained and fine-tuned with different protocols in Table 8. Notably, adversaries crafted with classification-trained encoders exhibit much lower transferability compared to those crafted with a vision-language alignment-pretrained vision encoder of the same architecture, suggesting that the latter encoder serves as a better tool for identifying common non-robust features on which downstream models rely. Further fine-tuning on ImageNet can make alignment-pre-trained off-the-shelf CLIP vision encoders even stronger attack surrogates. We hypothesise that the combination of web-scale vision-language data, the alignment training objective and various training configurations make alignment-trained encoders better feature extractors and hence the basis of more transferable attacks.

10. Prompts used for IC and VQA evaluation

Prompts used for caption generation:
OpenFlamingo:

```
<example> Output:{EXAMPLE} ...
<image> Output:{CAPTION} <|endofchunk|>
```

LLaVA:

```
USER: <image> Please describe the image using one
sentence.  ASSISTANT:
```

Prompts used for VQA:
OpenFlamingo:

```
<image> Question:  {QUESTION} Short answer:
{ANSWER} <|endofchunk|>
```

LLaVA:

```
USER:<image>{QUESTION} Answer the question using
a single word or phrase. ASSISTANT:{ANSWER}
```

References

[1] *Zero-Shot Semantic Segmentation with Decoupled One-Pass Network*, 2023. 1, 2, 7

[2] Abhishek Aich, Calvin-Khang Ta, Akash A Gupta, Chengyu Song, Srikanth Krishnamurthy, M. Salman Asif, and Amit Roy-Chowdhury. GAMA: Generative adversarial multi-object scene attacks. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. 3

[3] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 888–897, 2018. 4, 1, 3

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 2, 7

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 2, 5

[6] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023. 1

[7] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[8] Hao Chen, Bhiksha Raj, Xing Xie, and Jindong Wang. On catastrophic inheritance of large foundation models. *arXiv preprint arXiv:2402.01909*, 2024. 1

[9] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *arXiv preprint arXiv:2305.08192*, 2023. 1

[10] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10800–10809, 2020. 3

[11] 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. 2

[12] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation, 2023. 2, 6, 7

[13] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 2, 5

[14] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. 2022. 2, 3

[15] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3136–3146, 2023. 7

[16] Aditya Ganeshan, Vivek BS, and R Venkatesh Babu. Fda: Feature disruptive attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8069–8079, 2019. 1

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating

the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3

[19] Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*, 2023. 1

[20] Andong Hua, Jindong Gu, Zhiyu Xue, Nicholas Carlini, Eric Wong, and Yao Qin. Initialization matters for adversarial transfer learning. *arXiv preprint arXiv:2312.05716*, 2023. 1

[21] Lifeng Huang, Chengying Gao, and Ning Liu. Defeat: Decoupled feature attack across deep neural networks. *Neural Networks*, 156:13–28, 2022. 1

[22] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2019. 1

[23] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 3

[24] Nathan Inkawhich, Kevin Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *Advances in Neural Information Processing Systems*, 33:20791–20801, 2020. 1

[25] Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016. 3

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 2

[27] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023. 7

[28] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, 2004. 2, 5

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[31] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016. 1

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[33] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models, 2023. 1

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 7

[35] Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 940–949, 2020. 2, 3, 1, 6

[36] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *ICML*, 2023. 2

[37] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[38] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, 2023. 2

[39] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *17th European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 3, 1

[41] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 7

[42] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 2

[43] Muzammal Naseer, Salman H Khan, Shafin Rahman, and Fatih Porikli. Task-generalizable adversarial attack based on perceptual metric. *arXiv preprint arXiv:1811.09020*, 2018. 2, 3, 1, 6

[44] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. On generating transferable targeted perturbations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2021. 1

[45] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving

adversarial transferability of vision transformers. *arXiv preprint arXiv:2106.04169*, 2021. 7

[46] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[47] Laura F Nern, Harsh Raj, Maurice André Georgi, and Yash Sharma. On transfer of adversarial robustness from pretraining to downstream tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[48] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021. 3

[49] David A Noever and Samantha E Miller Noever. Reading isn't believing: Adversarial attacks on multi-modal neurons. *arXiv preprint arXiv:2103.10480*, 2021. 1

[50] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022. 2, 7

[51] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 1

[52] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 2, 5

[53] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018. 1

[54] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023. 1

[55] Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *arXiv preprint arXiv:2210.05968*, 2022. 1

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[57] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023. 7

[58] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *36th Conference on Neural Information Processing Systems (NIPS)*, 2022. 2

[59] Mengye Ren, Ryan Kiros, and Richard Zemel. *Advances in neural information processing systems*, 28, 2015. 2

[60] Shahbaz Rezaei and Xin Liu. A target-agnostic attack on deep models: Exploiting security vulnerabilities of transfer learning. *arXiv preprint arXiv:1904.04334*, 2019. 1

[61] Mathieu Salzmann et al. Learning transferable adversarial perturbations. *Advances in Neural Information Processing Systems*, 34:13950–13962, 2021. 1, 6

[62] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024. 1

[63] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models, 2023. 1

[64] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 31, 2018. 1

[65] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[66] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[67] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2, 5

[68] Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. With great training comes great vulnerability: Practical attacks against transfer learning. In *27th USENIX security symposium (USENIX Security 18)*, pages 1281–1297, 2018. 1

[69] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021. 1

[70] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7639–7648, 2021. 1

[71] Futa Waseda, Sosuke Nishikawa, Trung-Nghia Le, Huy H Nguyen, and Isao Echizen. Closer look at the transferability of adversarial examples: How they fool different models differently. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1360–1368, 2023. 4, 1, 3

[72] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. *ArXiv*, abs/2303.13076, 2023. 2

[73] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation, 2023. 2

[74] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 2, 7

[75] Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14983–14992, 2022. 1

[76] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2, 3

[77] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 7

[78] Xiaojun Xu, Jacky Y Zhang, Evelyn Ma, Hyun Ho Son, Sanmi Koyejo, and Bo Li. Adversarially robust models may not transfer better: Sufficient conditions for domain transferability from the view of regularization. In *International Conference on Machine Learning*, pages 24770–24802. PMLR, 2022. 1

[79] Xin Xu, Tianyi Xiong, Zheng Ding, and Zhuowen Tu. Masqclip for open-vocabulary universal image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 887–898, 2023. 2

[80] Yutaro Yamada and Mayu Otani. Does robustness on imagenet transfer to downstream tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9215–9224, 2022. 1

[81] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Boosting transferability of targeted adversarial examples via hierarchical generative networks. In *European Conference on Computer Vision*, pages 725–742. Springer, 2022. 1

[82] Muyang Yi, Quan Cui, Hao Wu, Cheng Yang, Osamu Yoshie, and Hongtao Lu. A simple framework for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7071–7080, 2023. 2

[83] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *NeurIPS*, 2023. 2

[84] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. 2022. 2

[85] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022. 1

[86] Jianping Zhang, Yizhan Huang, Weibin Wu, and Michael R Lyu. Transferable adversarial attacks on vision transformers with token gradient regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16415–16424, 2023. 1

[87] Ziqi Zhang, Yuanchun Li, Jindong Wang, Bingyan Liu, Ding Li, Yao Guo, Xiangqun Chen, and Yunxin Liu. Remos: reducing defect inheritance in transfer learning via relevant model slicing. In *Proceedings of the 44th International Conference on Software Engineering*, pages 1856–1868, 2022. 1

[88] Anqi Zhao, Tong Chu, Yahao Liu, Wen Li, Jingjing Li, and Lixin Duan. Minimizing maximum model discrepancy for transferable black-box targeted attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8162, 2023. 1

[89] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, pages 159–175. Springer, 2022. 2

[90] Zhengyu Zhao, Hanwei Zhang, Renjue Li, Ronan Sicre, Laurent Amsaleg, Michael Backes, Qi Li, and Chao Shen. Revisiting transferable adversarial image examples: Attack categorization, evaluation guidelines, and new insights. *arXiv preprint arXiv:2310.11850*, 2023. 1

[91] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2

[92] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6311–6320, 2023. 1

[93] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. 1

[94] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. 1

[95] Junhua Zou, Zhisong Pan, Junyang Qiu, Xin Liu, Ting Rui, and Wei Li. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII*, pages 563–579. Springer, 2020. 2, 7