Task-Agnostic Attacks Against Vision Foundation Models

Supplementary Material

1. Ablation study

1.1. Uncentered features in VFMs

We observe that features extracted from foundation models are not inherently centered, making cosine similarity loss unsuitable unless appropriate centering is applied.

In Fig. 1, we present the absolute mean value of each feature dimension for features extracted from the ImageNet validation set using ViT-B and ViT-L models. For each image, the feature vector is constructed by concatenating the class (CLS) token with the average patch token from the final layer. The results in Fig. 1 indicate significant variation in the absolute mean value across feature dimensions for all models. Additionally, for models such as DiNO ViT-B, CAE ViT-B, and MSN ViT-L, the mean feature vector is notably distant from the origin, underscoring the importance of mean-centering as a preprocessing step before computing the loss.

In Tab. 1, we compare performances of TAAs against VFMs with and without the use of mean centering. Mean centering provides strictly better results compared to simply ignoring this processing step.

Table 1. Classification accuracies on the Imagenette dataset for clean samples and TAA samples with and without the use of mean-centering.

Model	Clean	No centering	Centering
DiNO ViT-B	85.0	5.6	**5.4**
DiNOv2 ViT-B	90.5	0.4	**0.2**
MAE ViT-B	71.5	9.2	**6.2**
MSN ViT-B	85.8	10.4	**9.9**

1.2. Impact of VFM layer selection for TAA

We demonstrate in Tab. 2 and Tab. 3 that indeed, more efficient attack is created from the last layer of a VFM. It is not surprising, since the classification and segmentation heads are using the output tokens from the last layer as input. It is interesting to note that attacks carried against feature representations in middle layers perform the worst. We leave an explanation of this phenomenon for future research.

Table 2. Classification accuracy, segmentation mIoU, and cosine similarity between original and adversarial CLS token for the last layer on the PascalVOC dataset for DiNOv2 ViT-S model when attacking **CLS token** with respect to layer from which it taken with PSNR equal to 40 db.

Layer	Classification	Segmentation	CLS cos_sim
No attack	*96.3*	*81.4*	1
1 (first)	64.5	51.5	0.5
2	51.5	41.0	0.4
3	27.1	30.5	0.1
4	32.2	34.0	0.2
5	89.8	64.3	0.7
6	91.1	63.9	0.6
7	82.4	61.2	0.5
8	64.1	44.0	0.3
9	38.5	27.6	0.1
10	19.3	25.2	0.0
11	**6.2**	22.1	-0.0
12 (last)	7.5	**19.1**	**-0.8**

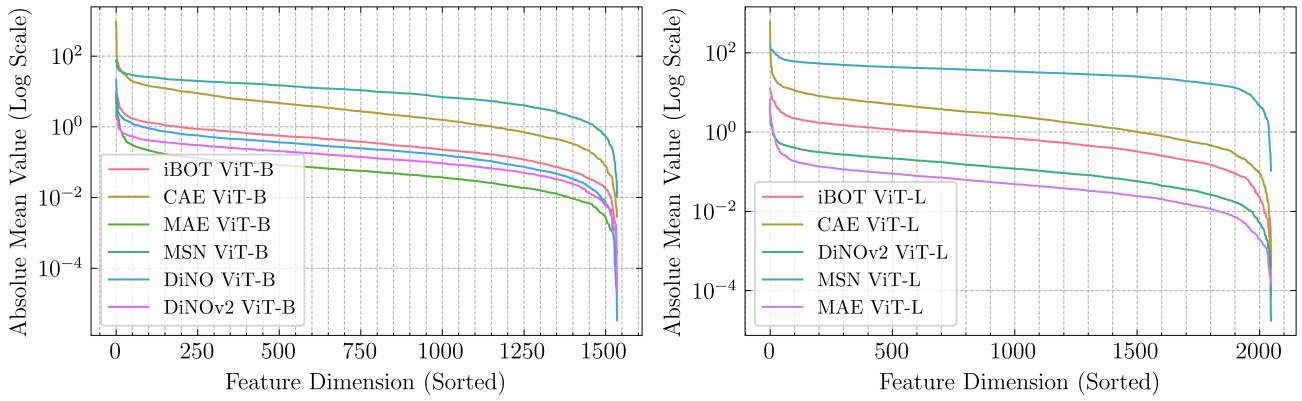
Table 3. Classification accuracy, segmentation mIoU, and cosine similarity between original and adversarial CLS token for the last layer on the PascalVOC dataset for DiNOv2 ViT-S model when attacking **patch tokens** with respect to layer from which it taken with PSNR equal to 40 db.

Layer	Classification	Segmentation	CLS cos_sim
No attack	*96.3*	*81.4*	1
1 (first)	49.9	38.8	0.4
2	56.6	45.4	0.5
3	18.7	21.3	0.1
4	43.4	42.9	0.3
5	72.6	50.1	0.5
6	60.7	42.6	0.4
7	38.5	32.5	0.2
8	9.8	27.3	0.0
9	5.8	22.4	0.0
10	6.3	23.1	0.0
11	2.0	14.1	-0.0
12 (last)	**0.1**	**11.5**	**-0.3**

2. Extra experimental results

2.1. Transferability across models

We report the transferability of attacks across 18 models for Targeted Adversarial Attacks (TAAs) and Transferable Surrogate Attacks (TSAs) on both classification and segmentation tasks in Fig. 3 and Fig. 4, respectively. From these figures, we observe that while TAAs generally underperform compared to TSAs for their respective tasks, their performance remains comparable. Notably, for some model



(a) ViT-B models.

(b) ViT-L models.

Figure 1. Absolute mean value per dimension for ViT-B (left) and ViT-L (right) VFMs. Feature coordinates are sorted in decreasing order of absolute mean value for features extracted from the ImageNet validation set.

families, such as I-JEPA and MAE, TAAs significantly underperform relative to TSAs. However, in specific cases, such as classification with DiNOv2, TAAs demonstrate superior performance over TSAs.

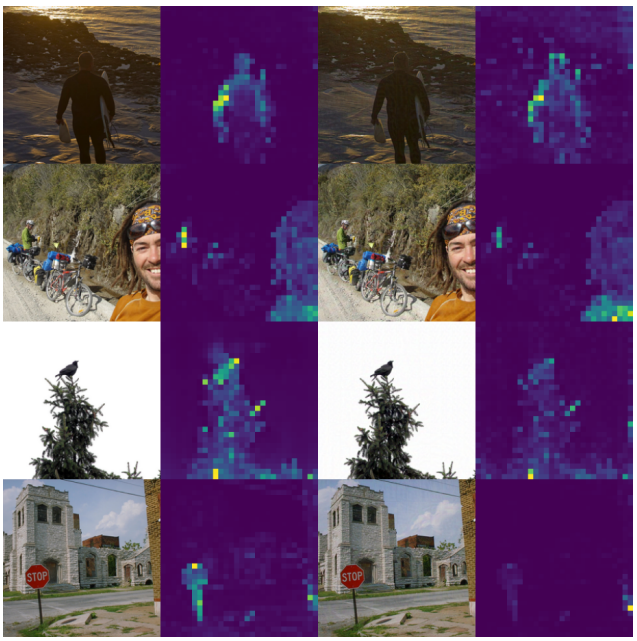


Figure 2. Corruption of attention masks of DiNO ViT-B using MSN ViT-B as a surrogate on the COCO2017 dataset. Original images and attention masks are in the first and second columns. Relative adversarial results are in the third and fourth columns. The target PSNR is 40dB.

2.2. Qualitative results for captioning and VQA

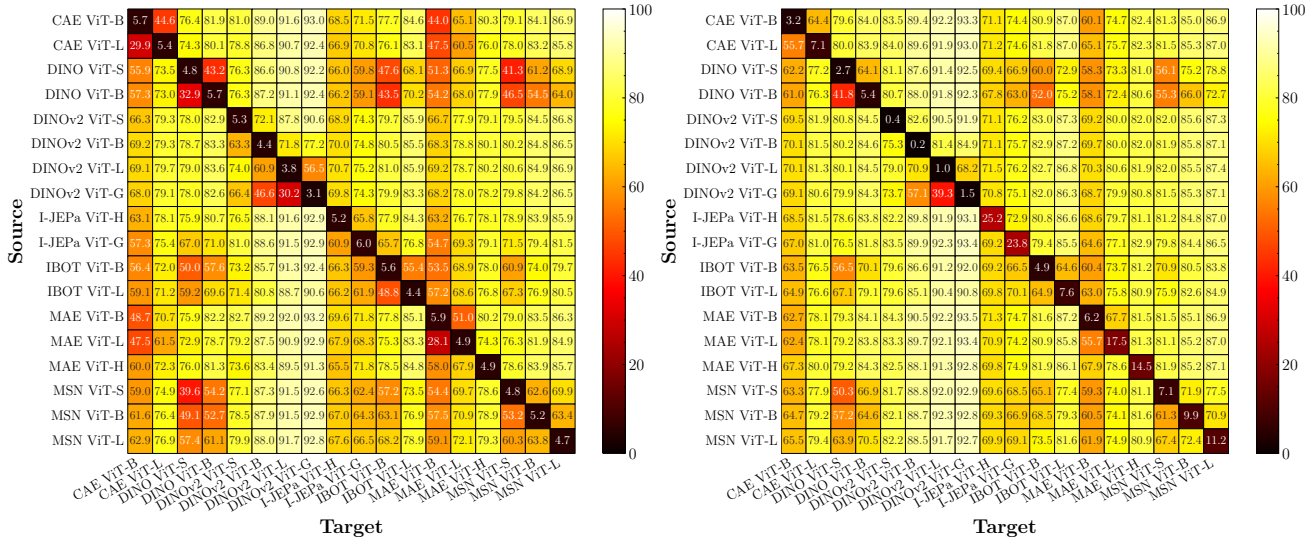
In Fig. 5, we report a more exhaustive collection of captions and answers obtained with Paligemma for task-agnostic adversarial images on the COCO dataset.

2.3. Qualitative results for classification and segmentation

We present examples of outputs from the DiNOv2 ViT-S model, where linear layers were trained on top for classification and segmentation tasks, after applying various adversarial attacks targeting a PSNR of 40 dB, as shown in Tab. 5. While attacks based on the corresponding downstream loss yield the most effective results for the specific task, they exhibit poorer transferability to other downstream tasks. For instance, the third row illustrates an almost unchanged predicted segmentation mask when the attack is performed in the classification space, whereas an attack in the segmentation space fails to flip classification predictions in 3 out of 6 cases (rows 1, 2, and 5). Additionally, in classification attacks, the predicted segmentation contours remain nearly unaffected, with only the predicted class being altered (rows 1, 4, 5, and 6). Conversely, TAA effectively produce adversarial samples that significantly degrade model performance across multiple tasks.

2.4. Analysing Self-attention

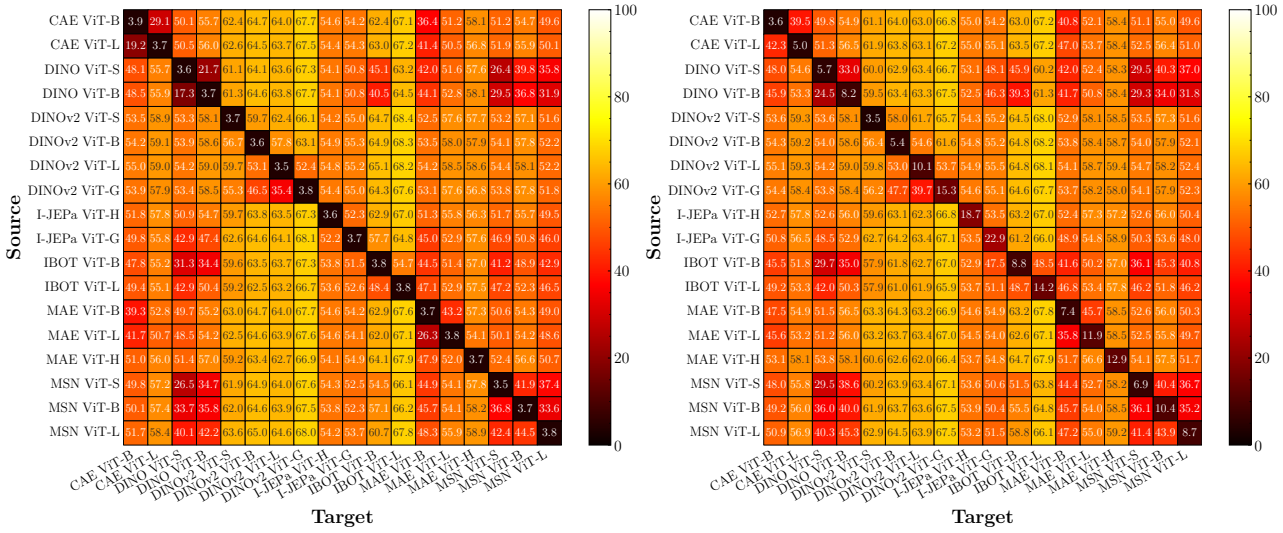
In Fig. 2 we qualitatively visualize self-attention in DiNO on images from the MS-COCO2017 dataset. We show images before and after corruption in the black-box setting, using an MSN ViT-B model as the surrogate and a target PSNR of 40dB. Original images and attention masks are in the first and second columns respectively, while adversarial results are in the third and fourth columns. With this qualitative study, it emerges that slightly different images can indeed yield different self-attention maps with respect to the `[CLS]` token. We suspect that these manipulations can turn out to be harmful to segmentation and detection models based on features extracted with DiNO.



(a) Task Specific Attacks (TSA)

(b) Task Agnostic Attacks (TAA)

Figure 3. Absolute classification accuracy on the ImageNette dataset for TSAs (left) and TAAs (right)



(a) Task Specific Attacks (TSA)

(b) Task Agnostic Attacks (TAA)

Figure 4. Absolute segmentation mIoU on the Pascal-VOC dataset for TSAs (left) and TAAs (right).

2.5. Transferability across models

We observe that the transferability of attacks across models is limited when we set the PSNR to 40dB. In Tab. 4, we study the transferability of TSA and TAA for classification under different solvers, namely PGD [2], SIA [3] and MI-FGSM [1], setting a lower PSNR value of 32dB.

The results in Tab. 4 suggest that great improvements in transferability can be obtained with higher distortion levels and with advanced optimization algorithms. We leave a more thorough study of the transferability of TAAs across models for future work.

Table 4. Classification accuracy of the targeted models with attacks using model iBOT-B as a source model. PSNR= 32dB.

Attack	iBOT-B	MSN-B	MAE-B	DiNOv2-B
PGD - TAA	**0.1**	75.1	49.5	85.5
PGD - TSA	5.5	63.3	44.3	83.1
MI-FGSM - TAA	5.1	50.5	38.3	80.0
MI-FGSM - TSA	5.5	42.8	33.1	69.4
SIA - TAA	0.5	**17.9**	**13.5**	**39.2**
SIA - TSA	2.5	20.7	17.6	42.1

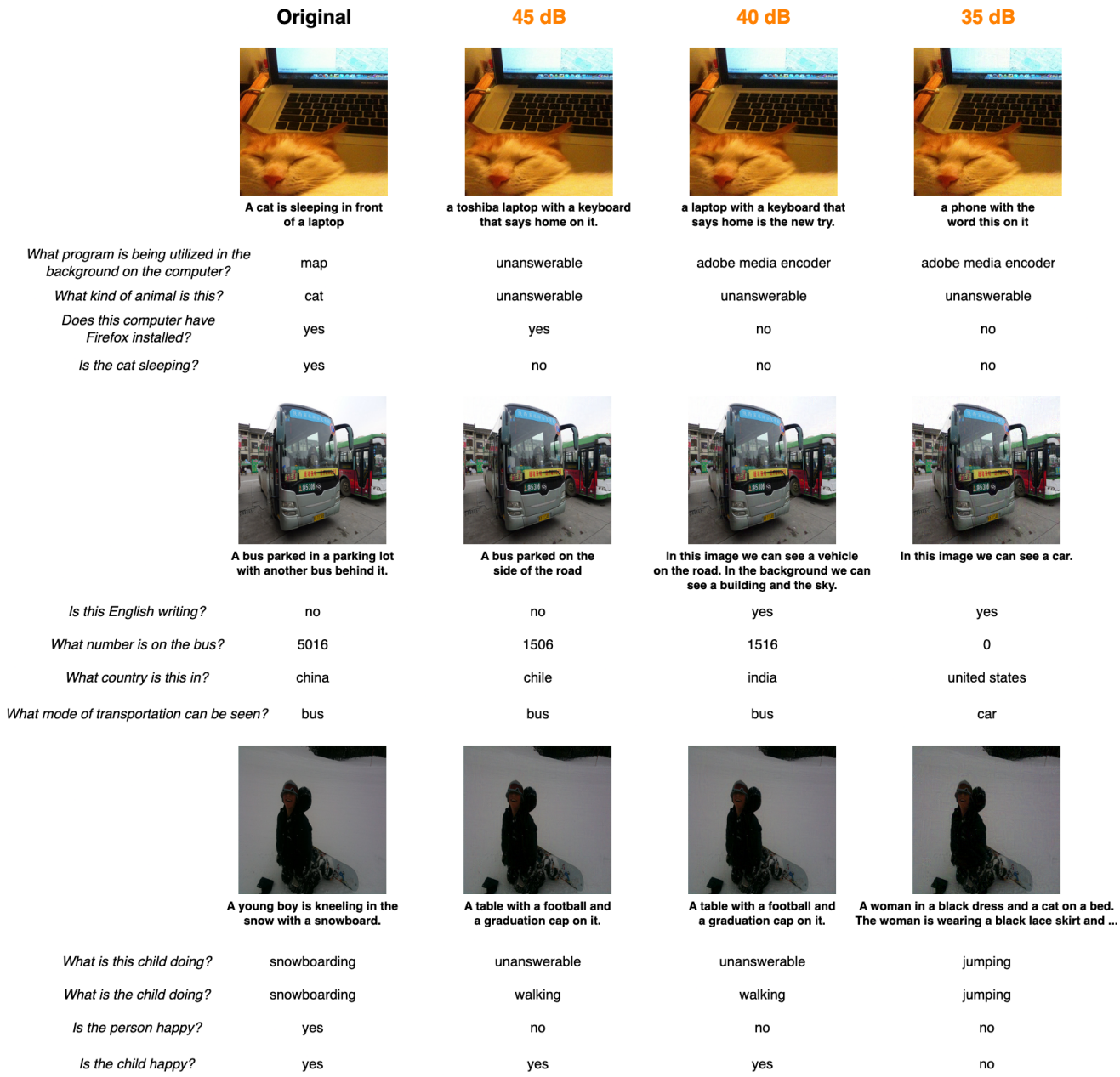
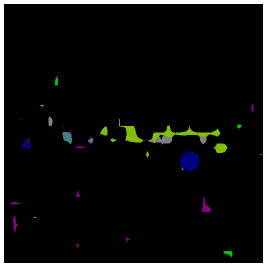
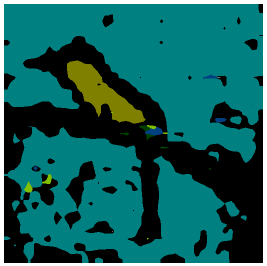
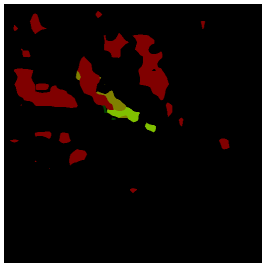
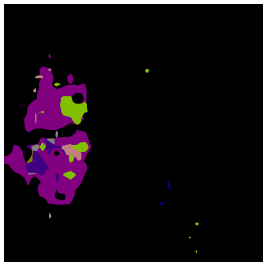
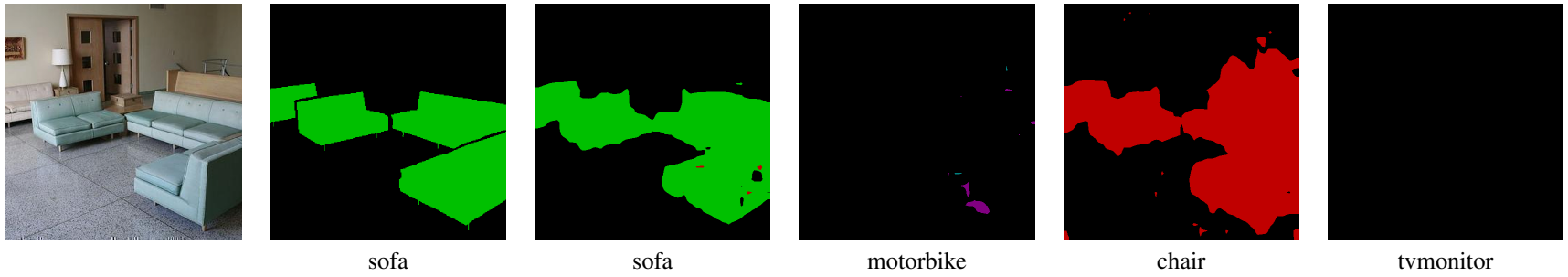


Figure 5. Example of captions and answers for images in the COCO dataset with TAAs using Paligemma.

Table 5. Adversarial attacks for segmentation and classification downstream tasks for DiNOv2 ViT-S model with a target PSNR of 40 dB.

Clean image	**GT label**	**Clean prediction**	**TAA patch tokens**	**TSA classification**	**TSA segmentation**
	 bus	 bus	 bottle	 train	 bus
	 bird	 bird	 dog	 aeroplane	 bird
	 person	 person	 bottle	 cat	 bottle


sofa
sofa
motorbike
chair
tvmonitor
horse
horse
cat
cow
horse
train
train
bottle
bus
person

References

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 3

[2] 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

[3] Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4607–4619, 2023. 3