PGA: Prior-free Generative Attack for Practical No-box Scenario

Supplementary Material

A. Additional Experiments

A.1. Experiments on Broader Source Domains

While the main paper utilizes standard datasets such as ImageNet [10] and MS COCO [5], we further evaluate generalizability under the Practical No-box Scenario (PNS) by exploring source domains with extreme distribution shifts or limited semantic diversity. Specifically, we employ **Paintings**[1], representing a significant *style shift*, and **Oxford 102 Flowers** [8] (abbreviated as Flowers), which presents a *narrow semantic distribution*. Following the protocol in the main paper, we utilize 4,000 unlabeled samples from each domain to train the surrogate and generator from scratch. As shown in Tab. 9, our method demonstrates remarkable resilience to these shifts, consistently outperforming peer methods across all fourteen victim models. On Paintings, our method surpasses the nearest competitor (GAPF [11]) by over **6%** and **9%** in AVG_c and AVG_v, respectively. On Flowers, where GAPF suffers severe degradation due to overfitting, our method maintains robust performance, exceeding the second-best by approximately **10%** and **9%**. These results confirm the strong transferability of our method, demonstrating that transferable attacks can be launched using only limited unlabeled samples.

A.2. Experiments on Fewer Available Samples

In the previous evaluation, our experiments utilized 4,000 unlabeled samples from the source domain. To further assess performance under conditions of greater data scarcity, we restrict the available data from the ImageNet domain to **only 2,000 unlabeled samples** for crafting the attacks. To ensure a fair comparison, we adjust the training protocol by doubling the training epochs for all methods to compensate for the dataset size being halved. This adjustment ensures that the total number of training iterations remains consistent with previous experiments, thereby guaranteeing that the surrogates for all methods can converge effectively and stably. As shown in Tab. 10, the performance degradation is not significant despite the 50% data reduction, suggesting that 2,000 samples still provide sufficient semantic information when explored thoroughly. Notably, our method maintains high transferability, outperforming the nearest competitor by over **13%** on CNNs and **11%** on ViTs / MLPs. These results further confirm that the superiority of our method does not rely on data volume.

A.3. Experiments on Surrogate Learning Strategies

In the main paper, we adopt SimSiam [3] as the default surrogate training framework for all competing generative attacks. To verify the rigor of this setting, we conduct experiments to evaluate different mainstream self-supervised learning frameworks under the PNS. Specifically, we employ SimCLR [2], VICReg [1], and SimSiam [3], alongside our proposed surrogate learning strategy, to train the surrogates. To ensure a fair comparison, BIA [13] is consistently used for generator training across all settings. As shown in Tab. 11, SimSiam outperforms other self-supervised methods, confirming that our baseline selection is fair and reasonable. Furthermore, our method achieves significant performance gains over the others, validating the effectiveness of our Curriculum-Guided Micro-Robust Optimization.

[1] https://kaggle.com/competitions/painter-by-numbers

Table 9. Additional results in ASR (%) on **broader source domains**. The evaluation includes fourteen victim models from CNN, ViT, and MLP architectures on ImageNet. The perturbation budget is $\epsilon = 16$, and only 4,000 unlabeled samples are used for the attacks. The AVG_c and AVG_v denote the average ASR on CNNs and ViTs / MLPs, respectively. The best results are shown in **bold**.

Dataset	Method	CNNs										ViTs / MLPs					
		Res-101	VGG-16	Den-121	Inc-v3	Inc-v4	WR-50	Mob-v2	SE-101	PNA	AVG_c	Mlp-B	ViT-B	Swin-B	Deit3	Beit-B	AVG_v
Paintings	CDA [7]	30.53	70.58	56.83	54.03	49.62	34.21	59.53	20.97	47.50	47.09	36.90	33.41	11.12	31.22	23.02	27.13
	GAPF [11]	41.10	82.60	68.97	67.39	58.75	47.73	67.65	25.20	47.84	56.36	41.59	37.38	13.50	36.11	30.06	31.73
	BIA [13]	37.12	76.81	67.87	62.09	55.65	43.60	66.37	24.45	52.24	54.02	46.03	37.81	12.16	35.68	28.82	32.10
	BIA+DA [13]	36.60	77.46	67.55	61.34	55.81	43.20	66.84	24.28	52.63	53.97	46.30	38.22	12.46	35.75	28.87	32.32
	BIA+RN [13]	37.02	76.59	67.47	60.22	53.97	42.98	63.46	22.68	50.29	52.74	43.17	37.04	11.80	34.37	26.90	30.66
	FACL [12]	38.40	77.09	68.08	62.09	56.29	44.08	66.74	24.63	52.11	54.39	46.14	38.29	12.75	35.81	28.28	32.25
	Ours	**45.16**	**83.95**	**76.37**	**70.50**	**64.88**	**50.94**	**77.67**	**28.14**	**64.17**	**62.42**	**57.66**	**50.02**	**15.06**	**47.55**	**38.26**	**41.71**
Flowers	CDA [7]	30.97	66.49	53.33	49.28	48.51	35.44	60.28	23.43	44.92	45.85	32.92	27.65	13.55	32.22	23.27	25.92
	GAPF [11]	32.84	70.47	51.75	55.70	49.22	34.04	54.95	22.88	42.71	46.06	31.53	28.30	12.23	29.84	21.92	24.76
	BIA [13]	39.92	73.79	61.67	61.00	56.54	43.74	68.66	26.54	54.87	54.08	43.96	36.08	15.01	38.80	29.78	32.73
	BIA+DA [13]	40.78	73.29	63.08	61.02	57.54	44.51	69.63	25.65	54.92	54.49	45.54	38.87	14.47	39.37	30.13	33.68
	BIA+RN [13]	39.92	73.77	64.02	59.64	56.62	45.89	69.34	24.43	53.82	54.16	46.54	39.67	14.71	39.42	31.40	34.35
	FACL [12]	39.82	73.75	63.86	60.38	56.49	45.17	68.96	25.08	54.42	54.21	45.50	38.74	14.94	38.68	30.42	33.66
	Ours	**53.86**	**83.78**	**73.99**	**66.95**	**59.54**	**59.43**	**84.80**	**33.72**	**63.35**	**64.38**	**60.00**	**52.65**	**20.42**	**47.02**	**38.52**	**43.72**

Table 10. Additional results in ASR (%) on **fewer available samples**. The evaluation includes fourteen victim models from CNN, ViT, and MLP architectures on the ImageNet domain. The perturbation budget is $\epsilon = 16$, and **only 2,000 unlabeled samples** are used for the attacks. The AVG_c and AVG_v denote the average ASR on CNNs and ViTs / MLPs, respectively. The best results are shown in **bold**.

Method	CNNs										ViTs / MLPs					
	Res-101	VGG-16	Den-121	Inc-v3	Inc-v4	WR-50	Mob-v2	SE-101	PNA	AVG_c	Mlp-B	ViT-B	Swin-B	Deit3	Beit-B	AVG_v
CDA [7]	30.78	69.62	60.06	52.79	47.97	34.78	55.44	20.14	47.66	46.58	38.10	34.51	9.37	31.79	25.43	27.84
GAPF [11]	39.01	82.43	68.51	62.95	54.54	45.57	65.69	23.36	46.81	54.32	40.59	34.80	13.48	34.87	29.41	30.63
BIA [13]	35.20	78.08	68.99	61.02	54.62	42.00	62.08	22.93	49.76	52.74	43.08	35.77	10.65	33.77	26.62	29.98
BIA+DA [13]	34.64	78.59	67.12	59.70	54.48	41.16	62.00	21.84	50.58	52.23	42.95	35.98	10.95	34.53	27.18	30.32
BIA+RN [13]	34.93	79.01	68.54	60.94	51.84	42.34	62.18	21.81	49.21	52.31	42.64	35.48	11.38	34.22	26.69	30.08
FACL [12]	34.29	78.61	68.02	60.02	53.77	41.93	61.98	21.71	49.88	52.25	42.40	35.33	10.45	33.84	26.32	29.67
Ours	**44.54**	**93.99**	**87.92**	**75.04**	**71.53**	**51.16**	**86.49**	**28.56**	**69.12**	**67.59**	**56.09**	**51.88**	**16.41**	**48.59**	**39.57**	**42.51**

Table 11. Additional results in ASR (%) under **different surrogate learning strategies**. The generator is trained using BIA [13] for all methods to ensure fairness. Both AVG_c and AVG_v are defined as in Tab. 9. The best results are shown in **bold**.

Method	SimCLR [2]	VICReg [1]	SimSiam [3]	Ours
AVG_c	51.78	53.50	54.18	**65.44**
AVG_v	30.11	31.71	32.95	**43.31**

Table 12. Additional ablation studies on the selection of the **intermediate layer j** of the surrogate during generator training. Both AVG_c and AVG_v are defined as in Tab. 9.

Layer	Maxpool	Layer-1	Layer-2	Layer-3	Layer-4
AVG_c	39.00	58.51	**71.39**	66.17	65.35
AVG_v	21.61	33.31	**47.50**	43.12	42.40

Table 13. Additional ablation studies on **loss weights α and γ** during the generator training. Here, we vary one parameter within $[0.25, 1.75]$ while fixing the other at the default value of 1.0. Both AVG_c and AVG_v are defined as in Tab. 9.

Parameter	α	γ	AVG_c	AVG_v
Varying α	0.25	1.0	67.05	44.52
	0.5	1.0	68.04	45.16
	0.75	1.0	69.53	46.69
	1.25	1.0	70.17	47.16
	1.5	1.0	69.33	46.11
	1.75	1.0	69.78	45.95
Varying γ	1.0	0.25	68.52	44.58
	1.0	0.5	69.69	45.55
	1.0	0.75	70.51	46.46
	1.0	1.25	69.48	46.74
	1.0	1.5	69.33	46.11
	1.0	1.75	68.62	44.84
Default	**1.0**	**1.0**	**71.39**	**47.50**

B. Additional Ablation Studies

In this section, we conduct additional ablation studies to provide further explanations. All experiments are trained on ImageNet, and the experimental settings remain identical to those in Tab. 1 of the main paper.

B.1. Ablation Studies on Attack Layer j

We conduct additional ablation studies to evaluate the impact of the selected intermediate layer j of the surrogate on transferability during the generator training phase (refer to Eq. (10) and Eq. (11) in the main paper). Specifically, we evaluate the performance when targeting the 'Maxpool', 'Layer-1', 'Layer-2', 'Layer-3', and 'Layer-4' of the surrogate (ResNet-18). As shown in Tab. 12, our method achieves the optimal attack performance when 'Layer-2' is selected, yielding the highest AVG_c and AVG_v. Consistent with previous research [11, 13], perturbing intermediate layer features results in better transferability.

B.2. Ablation Studies on Loss Weights α and γ

To evaluate the sensitivity of the hyperparameters α and γ in the objective function (Eq. (14) in the main paper), we conduct additional ablation studies by varying one parameter within the range $[0.25, 1.75]$ while keeping the other fixed at the default value of 1.0. As shown in Tab. 13, we observe that the performance remains stable and robust to variations in α and γ around the value of 1.0. Furthermore, the default configuration ($\alpha = 1.0, \gamma = 1.0$) achieves the optimal balance between regional fine-grained guidance and spatial consistency, yielding the highest transferability. This confirms that the selected hyperparameters are reasonable and not overly sensitive. Such stability demonstrates that the effectiveness of the method stems from the intrinsic design of the loss function rather than precise parameter tuning, thereby validating the robustness of our approach.

C. Qualitative Analysis

C.1. Visualization of Adversarial Perturbations

To provide a more intuitive understanding of why our method achieves superior transferability, we visualize the adversarial perturbations generated by our method and compare them with existing generative attacks. As shown in Fig. 7, existing methods often suffer from structural degradation under the Practical No-box Scenario (PNS). Specifically, the perturbations produced by these methods tend to be **coarse-grained** or **spatially fragmented**. Due to

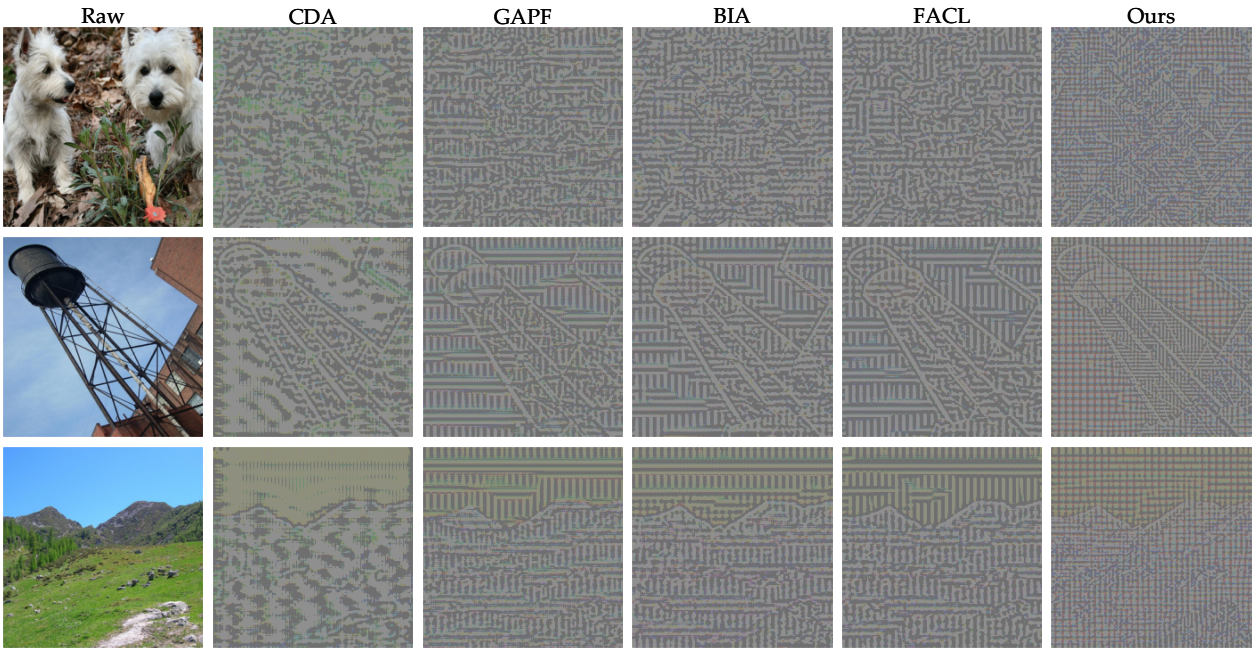

Figure 7. Visualization of adversarial perturbations. The perturbation budget is set to $\epsilon = 16$. Compared with existing generative attacks that often yield coarse-grained patterns, our method generates significantly more fine-grained and spatially coherent perturbations.

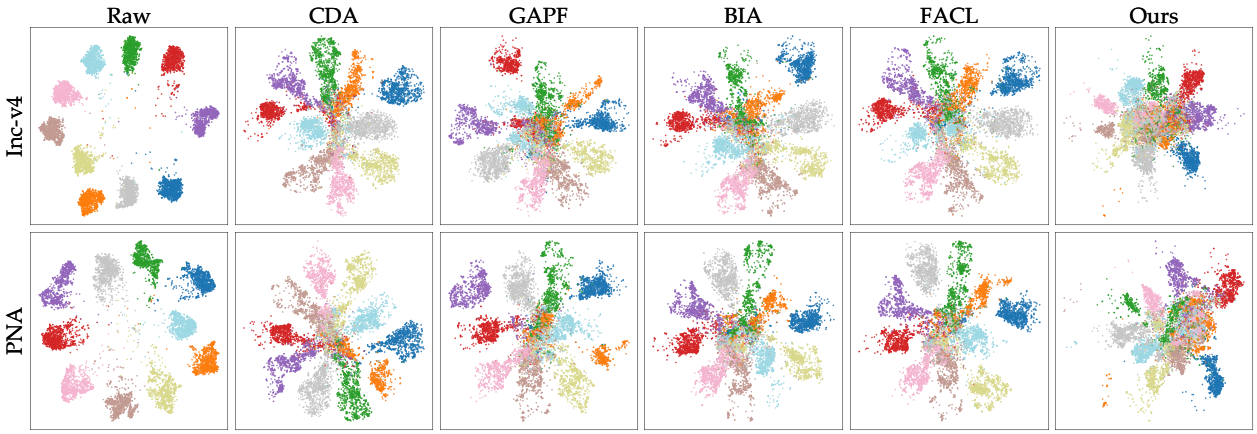

Figure 8. Visualization of feature distributions using t-SNE. We visualize the feature representations of 10 randomly selected classes from ImageNet on two victim models (Inc-v4 and PNA). In contrast to the distinct clusters formed by clean images, our method induces severe feature entanglement, effectively collapsing the decision boundaries between different categories.

the lack of abundant prior information and supervision, their generators often fall into local optima, producing repetitive or scattered noise patterns that fail to effectively cover the semantic regions of the image. In contrast, our method generates perturbations that are visibly more **fine-grained** and **spatially coherent**. Our generator is guided to produce dense and structurally consistent noise that spans the entire image. This qualitative comparison demonstrates that our method successfully mitigates the issue of overfitting under limited data, thereby leading to the significant improvement in transferability observed in our quantitative experiments.

C.2. Visualization of t-SNE Feature Distributions

To further investigate the impact of our attack on the internal representations of victim models, we randomly select 10 classes from the ImageNet, utilizing 1,000 images per class, and visualize their feature distributions via t-SNE. As shown in Fig. 8, in the "Raw" column, the features of clean images form distinct and well-separated clusters, indicating that the victim models classify them correctly. While existing methods induce varying degrees of disturbance, the original class structures remain largely identifiable. In con-

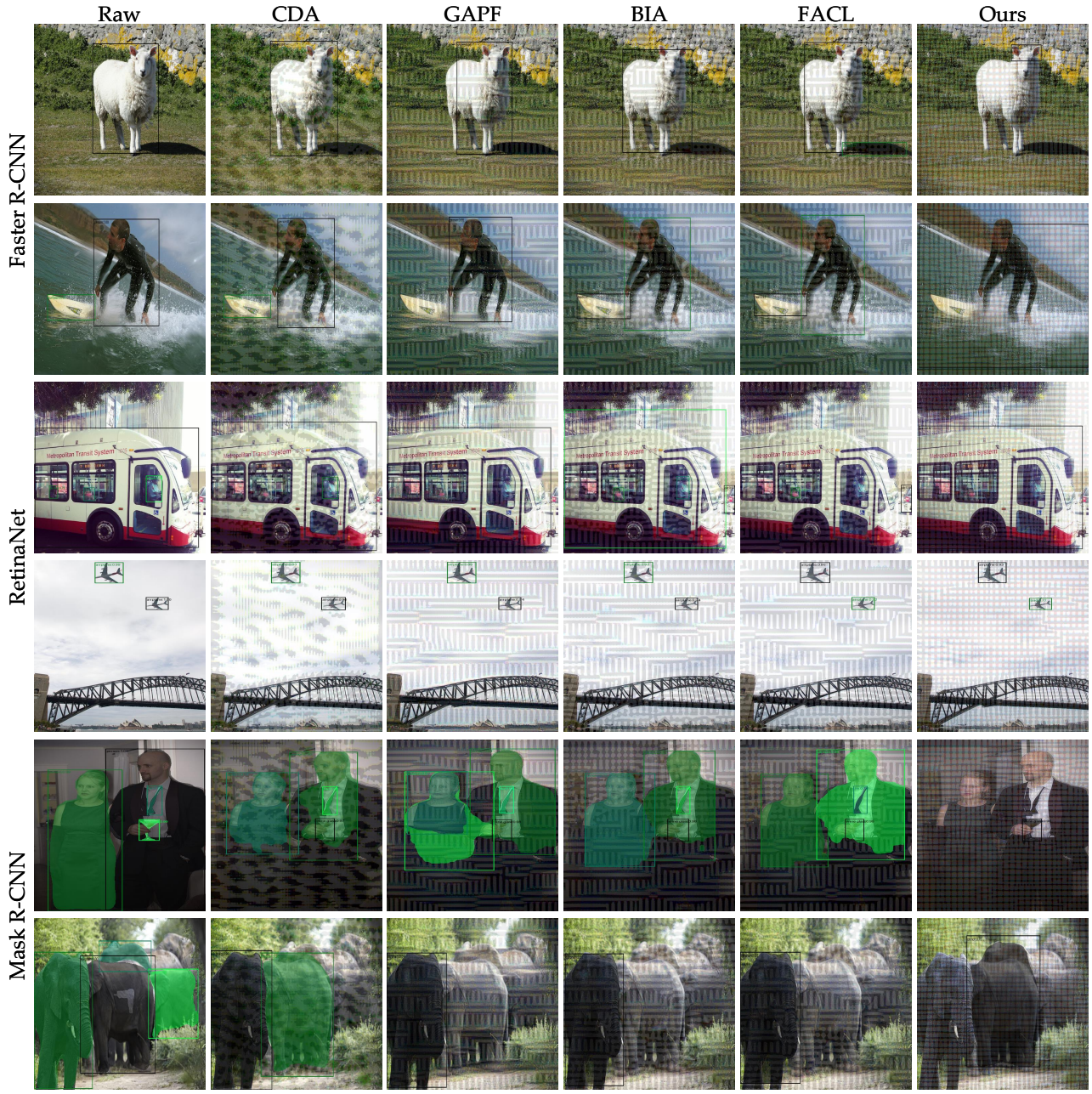

Figure 9. Visualization of cross-task attack results on object detection and instance segmentation. Our method effectively suppresses targets or induces severe misdetections, while existing generative attacks often fail to disrupt the predictions.

trast, the adversarial examples generated by our method result in severe feature entanglement and confusion, thereby yielding superior attack performance and transferability.

C.3. Visualization of Cross-Task Transferability

To comprehensively demonstrate the transferability of our approach, we visualize results across both object detection and instance segmentation tasks. We evaluate the attacks against three representative victim models: Faster R-CNN [9], RetinaNet [6], and Mask R-CNN [4]. As shown in Fig. 9, while the baseline methods degrade detection and segmentation precision to a certain degree, they fail to suppress a significant number of targets. In stark contrast, our method achieves notably superior attack efficacy, manifesting as severe misdetections or the complete suppression of object instances. This visual evidence corroborates that our approach disrupts fundamental semantic features, thereby facilitating better cross-task transferability.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, pages 1–12, 2022. 1, 2

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 1, 2

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 1, 2

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 4

[5] 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 *ECCV*, pages 740–755, 2014. 1

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 4

[7] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. In *NeurIPS*, pages 1–11, 2019. 1, 2

[8] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729, 2008. 1

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 1–9, 2015. 4

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1

[11] Mathieu Salzmann et al. Learning transferable adversarial perturbations. In *NeurIPS*, pages 13950–13962, 2021. 1, 2

[12] Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Facl-attack: frequency-aware contrastive learning for transferable adversarial attacks. In *AAAI*, pages 6494–6502, 2024. 1, 2

[13] Qilong Zhang, Xiaodan Li, YueFeng Chen, Jingkuan Song, Lianli Gao, Yuan He, and Hui Xue. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. In *ICLR*, pages 1–12, 2022. 1, 2