Boosting Adversarial Transferability with a Generative Model Perspective

Jongoh Jeong[1], Hunmin Yang[1,2] and Kuk-Jin Yoon[1,†]
[1]Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
[2]Agency for Defense Development (ADD), Republic of Korea
{jeong2, hmyang, kjyoon}@kaist.ac.kr [†]Corresponding author

Abstract

Generative transfer attacks craft adversarial examples by training a perturbation generator on a white-box surrogate and deploying them against unknown black-box targets. While existing generative methods demonstrate effective adversarial transferability and enjoy inference-time efficiency, they overlook the rich, model-shared semantic information in the intermediate generator features, which is key to enhancing transferability. To address this, we propose a self-distilling attack framework via mean teacher that ***effectively exploits these previously under-explored generator features*** *and preserves the semantic structure within the generator with student-teacher generator alignment via EMA updates. We conduct comprehensive evaluations across four metrics —Classification Accuracy, Attack Success Rate, Fooling Rate, and our newly proposed Accidental Correction Rate— to demonstrate consistent gains in both cross-model and cross-domain adversarial transferability.*

1. Introduction

Adversarial examples (AEs) have revealed critical vulnerabilities in deep vision models since [32] first observed that small, human-imperceptible perturbations can cause misclassification. Early white-box attacks—such as FGSM [56] and its iterative variants—exploit gradient information directly, but inherently require full model access. To evaluate real-world robustness, transfer-based (black-box) attacks emerged, leveraging the phenomenon that adversarial perturbations crafted on one model often fool others. Notable developments include Momentum Iterative [6], Diverse Inputs [47], and Translation-Invariant [8] FGSM, which improve transfer success via momentum, input transformations, and scale invariance, respectively. More recently, generative transfer attacks [1–3, 25–27, 29, 35, 45, 51–53, 58] train a dedicated perturbation generator against a surrogate model, enabling fast inference and high transfer rates without per-example optimization.

While these generative approaches achieve efficient inference and strong black-box transferability by aligning surrogate outputs (e.g., intermediate features or logits), they nonetheless neglect the potential of the intermediate features within the generator. In a slightly different perspective from these concurrent works [1, 12, 17, 18, 25–27, 29, 51–53, 58] that primarily focus on manipulating the surrogate features to disrupt the target model-invariant characteristics, the intermediate feature activations of the generator do also display semantically rich contextual cues of the object, such as structures, contours, and textures, that can be used towards adversarial transfer. Nonetheless, current generative model-based transfer attacks underestimate the capacity of the generator, thereby letting the generator features to deteriorate as training progresses, and thus limit the generalization of the generated perturbation to unseen models and domains.

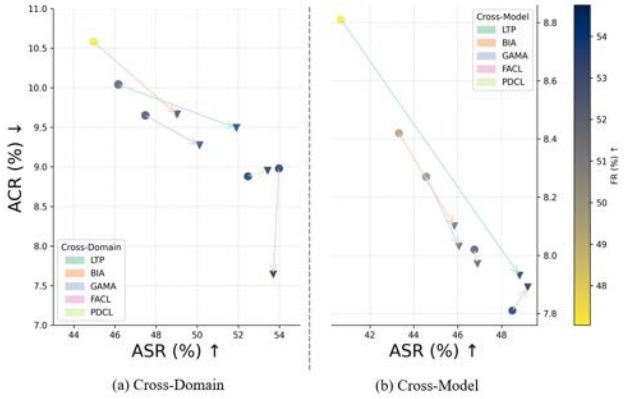

Figure 1. Our self-distilling attack effectively exploits the generative model to craft adversarial examples with enhanced transferability (▼) from the baselines (●) across domains (a) and models (b).

To address this gap, we propose a self-distilling mean teacher framework that exploits under-explored generator feature maps. By maintaining an exponentially averaged "teacher" generator and aligning its intermediate activations with those of the "student" generator, our approach preserves semantic integrity to effectively guide noise generation on these object-centric regions. We perform a fair evaluation

across four metrics to demonstrate consistent improvements in both cross-model and cross-domain settings.

Contributions. We summarize our contributions as follows: (1) For the first time in generative model-based transfer attacks, we leverage the semantic information embedded in the perturbation generator, which is *overlooked by previous works*, as a useful hint to enhance adversarial transferability, (2) We introduce a self-distilling mean teacher framework to better preserve and align the semantic structure of the object during generator training, which is empirically observed to be degraded over training iterations, and (3) We comprehensively evaluate the effectiveness of our attack using four evaluation metrics, including our proposed Accidental Correction Rate (ACR), which shows consistent behavior with the other metrics.

2. Background

2.1. Transferable Adversarial Attacks

Transferable adversarial attacks are a significant area of research in machine learning, which exploits the phenomenon where adversarial examples generated for one model can also trigger mis-prediction on other models, even if they are trained on different datasets or architectures. In a recent decade, numerous iterative [5, 7, 16, 20, 21, 24, 37, 38, 40, 44, 48] methods have exploited this transferable nature of adversarial examples to enable transfer attacks on unknown models. However, high computational costs for iterative optimization and limited transferability to target models that are significantly distinct in architecture from the source led to the development of highly transferable and inference-time cost-efficient generative model-based attacks [1, 25, 27, 29, 36, 51–53, 58].

2.2. Self-Knowledge Distillation

Self-knowledge distillation [13, 19, 19, 50, 54, 55, 57], where a model teaches itself, has been known to improve generalization and robustness without external teachers. [10, 59] first exploited iterative self- and peer-distillation. The Mean Teacher [33] approach, adapted from semi-supervised learning, then constructs the teacher as an EMA of student weights, using temporal ensembling to enforce consistency in predictions or feature maps and to smooth high-frequency noise under label-free supervision. In this approach, the network can be better calibrated and more robust to domain variations, thanks to enhanced domain invariance in the representations, which is key to black-box generalization.

3. Self-Distillation via Mean Teacher

Preliminaries. The framework for generative model-based transfer attacks comprises an adversarial perturbation generator $\mathcal{G}_\theta(\cdot)$, producing unconstrained adversarial examples x^{adv} from benign inputs x, which are then projected via a perturbation projector $\mathcal{P}(\cdot)$ to satisfy $||\mathcal{P}(x^{adv}) - x||_\infty \leq \epsilon$. To train $\mathcal{G}_\theta(\cdot)$ in a label-free manner in the untargeted attack, we employ a white-box surrogate model to provide an adversarial supervisory signal for generator updates via backpropagation. The adversarial loss uses mid-layer surrogate features $\mathcal{F}_k(\cdot)$, which contain model-shared characteristics that are crucial for adversarial transferability.

Algorithm 1 Self-distilling Perturbation Generator Training

Input: Generator $\mathcal{G}_\theta(\cdot)$, training dataset $\mathcal{D}_{train}$, a frozen surrogate model $\mathcal{F}_k(\cdot)$, perturbation projector $\mathcal{P}(\cdot)$
Ensure: Randomly initialize student $\mathcal{G}_\theta(\cdot)$, and initialize a mean teacher $\mathcal{G}_{\theta'}(\cdot)$ with student weights θ
repeat
  Randomly sample a mini-batch x_i from train dataset $\mathcal{D}_{\text{train}}$
  Acquire student generator features s.t. $\mathbf{g}_{i,s} \leftarrow \mathcal{G}_\theta^{\text{enc}}(x_i)$
  Acquire teacher generator features s.t. $\mathbf{g}_{i,t} \leftarrow \mathcal{G}_{\theta'}^{\text{enc}}(x_i)$
  Generate unbounded adversarial examples, s.t. $x_i^{\text{adv}} \leftarrow \mathcal{G}_\theta^{\text{dec}}(\mathbf{g}_{i,s})$
  Project x_i^{adv} within the perturbation budget ϵ, s.t. $||\mathcal{P}(x_i^{\text{adv}}) - x_i||_\infty \leq \epsilon$
  Forward pass x_i and x_i^{adv} through surrogate $\mathcal{F}_k(\cdot)$ to acquire $f_i^{\text{benign}}, f_i^{\text{adv}}$
  Compute loss using $f_i^{\text{benign}}, f_i^{\text{adv}}, \mathbf{g}_{i,s}, \mathbf{g}_{i,t}$: $\mathcal{L} = \mathcal{L}_{\text{adv}} + \lambda_{\text{distil}} \cdot \mathcal{L}_{\text{distill}}$ ▷ Eqs. 1,2
  Update student $\mathcal{G}_\theta(\cdot)$ gradients via backpropagation
  EMA update teacher with student, s.t. $\theta \mapsto \theta'$
until $\mathcal{G}_\theta(\cdot)$ converges.
return $\mathcal{G}_{\theta'}(\cdot)$

Preserving semantic integrity. With the recent works seeking to center their perturbations around salient objects [4, 14, 43], or manipulate either input data space [39, 41] or intermediate-level perturbations from the surrogate [12, 17, 18, 25], object-focused feature-level divergence are crucial for generating adversarial noise that is transferable across black-box models. In the generative attack framework likewise, we propose to explore the capacity of the generator in the context of crafting transferable AEs to induce more effective and transferable noise by preserving the semantic structure in the early intermediate generative features.

In pursuit of this goal, we employ Mean Teacher [33] (with $\eta = 0.999$) to first build a noise-reduced reference to the student generator features, encapsulating smoothed semantics of an object without high-frequency noise. Then, we strictly enforce the semantic structure consistency between the early intermediate features of the teacher and the student via our distillation loss formulated in Eq. 1. We base our approach on our empirical observations of the trained generators from [26, 29, 58] where the early intermediate features are relatively well preserved in the residual learning stage, a stage in which adversarial noise is primarily generated. By guiding the generated noise to lie on the semantic structure

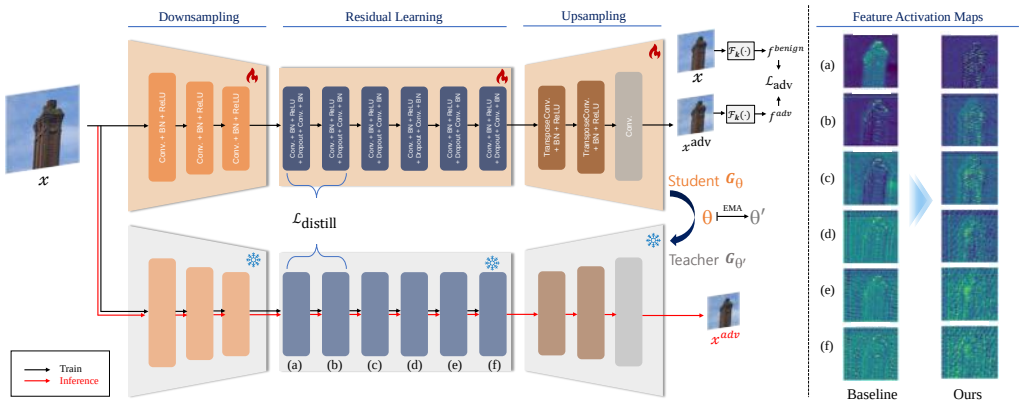

Figure 2. **Overview of our self-distilling attack via mean teacher (SDA).** Given an input image, x, the teacher generator $\mathcal{G}_{\theta'}$ maintains a smoothed copy of the student $\mathcal{G}_\theta$, which is optimized by $\mathcal{L}_{adv}$ using the surrogate features and $\mathcal{L}_{\text{distill}}$ using the intermediate features from the student and the teacher (*left*), thereby further guiding noise to be generated around the semantic structure of the object as the perturbation generator training progresses in the residual learning stage (*right*).

that is at least coarsely maintained by the mean teacher in the early intermediate features, our method induces the added noise to be generated acutely along the object-salient regions in a more structured manner in the residual learning stage.

$$\mathcal{L}_{\text{distill}} = \sum_{j \in L} \max\left(0,\ \tau - \frac{\mathbf{g}_{i,t}^{j\top}\, \mathbf{g}_{i,s}^{j}}{\|\mathbf{g}_{i,t}^{j}\|_2\, \|\mathbf{g}_{i,s}^{j}\|_2}\right), \qquad (1)$$

where $L, \tau, \mathbf{g}_{i,s}, \mathbf{g}_{i,t}$ represent the set of indices of intermediate residual block to distill, similarity threshold (0.6), student and teacher generator features at layer j, respectively. We set the intermediate layer set $L = \{1, 2\}$ for distilling the first and second residual block features, as shown in blue in Fig. 2. In sum, the total loss objective then becomes:

$$\mathcal{L} = \mathcal{L}_{\text{adv}} + \lambda_{\text{distill}} \cdot \mathcal{L}_{\text{distill}}, \qquad (2)$$

where λ_{distill} is a weight term for $\mathcal{L}_{\text{distill}}$, and $\mathcal{L}_{adv}$ is the surrogate loss. We use $\mathcal{L}_{\text{adv}} = \texttt{cos\_sim}(\mathcal{F}_k(x), \mathcal{F}_k(x^{adv}))$ at mid-layer $k = 16$ of VGG-16 surrogate for the baseline [58].

4. Experiments

Attack settings. We evaluate adversarial transfer under two black-box protocols. In the cross-model setting, perturbations are crafted on surrogate models trained with the same data distribution (ImageNet-1K [28]) and then tested on unseen target architectures. In the cross-domain setting, adversarial examples are to generalize across domain shifts without access to any target-distribution samples.

Datasets. We train the perturbation generator using data from ImageNet-1K [28]. Following [1, 26, 27, 29, 51–53, 58], we train the generator on the ImageNet-1K [28] containing 1.2 M natural images and evaluate on CUB-200-2011 [34], Stanford Cars [15], FGVC Aircraft [22] for fine-grained cross-domain settings, and various ImageNet-1K pre-trained model architectures for cross-model.

Victim models. For cross-model evaluation, we employ ImageNet-1K pre-trained classification models of various network architectures with their publicly available model weights. We source the pre-trained models from TorchVision [23] and Timm [42] libraries. Compared to previous approaches [29, 52, 53] demonstrating cross-model architecture transferability, we expand the evaluation to a wider scope of target model architectures for enhanced architecture-agnostic transferability.

Baselines. We compare our attacks against the state-of-the-art baselines that rely on the same ResNet generator to craft adversarial examples, *i.e.* LTP [29], BIA [58], GAMA [1], FACL-Attack [51, 52], and PDCL-Attack [53].

Against robust models. We also test our method on attacking robust models, *i.e.* adversarially trained models with Inception-V3 [16], ViT [9] and ConvNeXt [31], and robust input processing methods such as JPEG (75%) [11], bit reduction (BDR; 4-bit) [49] and randomization (R&P) [46] in Table 3.

Evaluation metrics. We evaluate the transferability of adversarial attacks across model architecture and domain shifts. Specifically, we comprehensively assess the attack effectiveness on top-1 accuracy (Acc.) and Fooling Rate (FR) on *all* test data, and Attack Success Rate (ASR) on all test samples originally classified *correctly* by the target model. Additionally, we newly propose to assess our attack by the Accidental Correction Rate (ACR) on all test samples originally classified *incorrectly* by the target model.

Main results. We evaluate our method in both cross-model (Table 1) and cross-domain (Table 2) scenarios, achieving notable transferability gains of 1.98%p and 3.97%p in accuracy, respectively, with consistent improvements across all other metrics. The ACR shows the smallest relative gain due to the inherently low prevalence of accidentally corrected

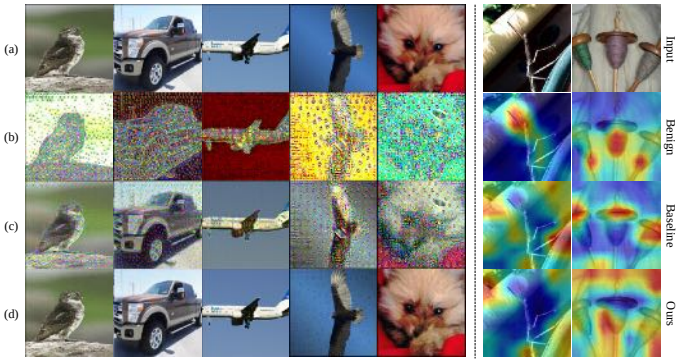

Figure 3. **Qualitative results.** Our self-distilling mean teacher successfully focuses perturbations particularly on the semantically meaningful regions, thereby fooling the victim classifier. *Left:* benign input image (a), generated perturbation (normalized for visual purposes; b), unbounded adversarial image (benign with perturbation; c), and $\ell_\infty \leq 10$-bounded adversarial image (d), on CUB-200-2011 [34], Stanford Cars [15], FGVC Aircraft [22], and ImageNet-1K [28] domains. *Right:* We emphasize that our method induces Grad-CAM [30] to focus on *drastically different regions* in our adversarial examples compared to both the benign image and the adversarial examples crafted by the baseline [58]. Moreover, our approach *noticeably reduces the high activation regions* observed in the benign and baseline cases, enhancing the transferability of our adversarial perturbations.

Table 1. **Quantitative results in the cross-model setting.** We trained our perturbation generator against VGG-16 surrogate on ImageNet-1K domain, and evaluated on black-box models, given a perturbation budget of $\ell_\infty \leq 10$. We report the black-box average denoted as *Bb. Avg*, with better results in **boldface**. *Please zoom in to see the results for each method.*

		ConvNets																Transformers												Mixer			Bb.
Method	**Metric**	vgg16	vgg19	r50	r152	d121	d169	inc-v3	wrn50	regnety	mobile	mnas	squeeze	shuffle	efficient	convnext-b	resnext-b	vit-b	vit-l	swin-b	t16	deit-b	pvt	max	vit-b	beit-b	efficient	vit	hrnet	mobilevit	cait-s24	davit-b	mlp-b	mlp-l	conv-b	conv-l	**Avg.**
Clean	Acc. (%) ↓	70.15	70.95	74.60	77.33	74.22	75.74	76.19	77.29	77.95	69.96	66.5	61.96	69.64	67.91	82.12	76.64	77.24	77.56	79.78	82.27	65.99	83.07	83.95	69.97	63.65	77.26	81.96	81.96	72.26	69.77	78.04	79.7	75.08
Baseline [58]	Acc. (%) ↓	1.56	3.60	25.36	42.98	26.97	32.35	41.20	33.31	31.30	10.04	34.30	10.48	30.14	60.06	48.04	29.83	69.31	70.97	45.68	74.60	29.86	62.51	78.39	55.35	26.23	30.42	73.00	55.94	61.12	57.72	58.21	57.39	45.44
	ASR (%) ↑	98.02	95.42	67.98	46.92	65.65	59.49	49.05	59.20	61.89	86.65	51.41	83.09	58.30	27.33	43.41	62.94	13.50	11.84	44.58	11.77	57.38	26.51	9.20	24.91	61.54	62.54	13.27	34.00	22.60	23.21	28.39	30.97	43.32
	FR (%) ↑	98.26	96.03	72.13	52.75	70.53	64.71	55.38	64.08	66.25	88.83	59.93	86.60	65.84	34.51	48.16	67.69	21.03	19.26	50.09	17.71	65.65	31.84	13.63	35.13	67.85	67.18	18.89	39.89	31.95	36.72	35.46	37.48	49.57
	ACR (%) ↓	0.58	1.20	5.75	8.48	5.64	6.83	9.86	7.82	7.23	2.35	5.95	2.33	6.01	12.27	8.41	6.11	10.98	11.56	7.25	11.37	5.10	8.65	13.48	9.36	4.83	6.52	10.63	10.26	11.62	13.70	10.57	11.68	8.42
w/ Ours	Acc. (%) ↓	1.59	3.05	23.12	40.98	25.66	32.35	37.44	33.24	29.71	9.18	31.02	7.63	26.01	59.66	45.34	28.89	68.64	70.72	33.69	73.26	30.97	60.92	77.99	54.27	23.70	28.89	72.80	51.97	59.62	56.78	53.89	55.44	**43.46**
	ASR (%) ↑	98.00	96.17	70.81	49.38	67.29	59.44	53.60	59.34	63.85	87.81	56.08	87.89	64.08	27.90	46.55	64.16	14.34	12.21	58.88	13.19	56.13	28.50	9.73	26.50	65.21	64.52	13.54	38.63	24.66	24.42	33.66	33.34	**45.85**
	FR (%) ↑	98.26	96.7	74.70	55.03	72.01	64.65	59.58	64.17	67.98	89.77	63.70	90.29	70.59	35.07	50.92	68.75	21.89	19.49	63.32	19.29	64.61	33.88	14.25	36.84	71.17	68.97	19.14	44.28	33.89	37.84	40.51	39.71	**51.88**
	ACR (%) ↓	0.63	1.13	5.30	8.12	5.36	6.74	8.77	8.02	6.96	2.16	5.41	1.94	5.34	12.50	7.72	6.07	10.90	11.69	4.37	10.37	5.93	8.99	13.78	9.46	4.29	6.50	10.75	9.24	11.78	13.37	9.65	11.40	**8.10**

Table 2. **Quantitative results in the cross-domain setting.** We compare the attack performance on fine-grained domains: CUB-200, Stanford Cars, and FGVC Aircraft (better results in **boldface**).

		CUB-200-2011			Stanford Cars			FGVC Aircraft			
Method	**Metric**	**res50**	**se-net**	**se-res101**	**res50**	**se-net**	**se-res101**	**res50**	**se-net**	**se-res101**	**Avg.**
Clean	Acc. (%) ↓	87.35	86.61	86.56	94.35	93.66	92.97	92.23	92.08	91.90	90.85
Baseline [58]	Acc. (%) ↓	32.74	52.99	58.04	39.61	69.90	70.17	28.92	60.31	46.92	51.07
	ASR (%) ↑	63.16	40.54	34.69	58.94	26.47	26.28	69.13	35.37	50.05	44.96
	FR (%) ↑	66.00	45.24	39.54	59.87	28.58	28.52	70.93	38.64	52.60	47.76
	ACR (%) ↓	4.36	10.73	10.54	10.39	15.73	15.75	6.11	12.08	9.52	10.58
w/ Ours	Acc. (%) ↓	35.92	49.48	58.32	27.62	66.30	65.09	20.10	56.80	44.31	**47.10**
	ASR (%) ↑	59.59	44.77	34.27	71.13	29.99	31.10	78.61	39.21	52.50	**49.02**
	FR (%) ↑	62.67	49.09	39.30	71.76	32.10	33.14	79.75	42.30	54.79	**51.66**
	ACR (%) ↓	4.90	11.65	10.54	7.14	14.04	14.69	4.96	10.57	8.46	**9.66**

Table 3. Attack performance with our method against defenses and input processing methods (better results in **boldface**).

Method	**Metric**	**Adv.Inc-V3**	**Adv.ViT**	**Adv.ConvNeXt**	**JPEG**	**BDR**	**R&P**	**Avg.**
Clean	Acc. ↓	76.33	48.82	58.44	74.68	74.68	76.58	68.26
Baseline [58]	Acc. (%) ↓	68.54	45.64	53.88	63.49	47.82	44.78	54.03
	ASR (%) ↑	14.95	11.72	10.26	20.24	40.76	44.59	23.75
	FR (%) ↑	24.02	25.48	19.40	28.09	48.06	51.60	32.78
	ACR (%) ↓	15.30	4.96	3.46	11.45	11.30	10.56	9.51
w/ Ours	Acc. (%) ↓	67.92	45.33	53.62	60.83	44.07	39.01	**51.80**
	ASR (%) ↑	15.75	11.95	10.65	23.74	45.37	51.63	**26.52**
	FR (%) ↑	24.83	25.31	19.60	31.61	52.22	57.86	**35.28**
	ACR (%) ↓	15.23	4.57	3.38	11.48	10.29	9.08	**9.01**

samples, limiting room for improvement. Against the robust defenses (Table 3), our approach uniformly surpasses the baseline. Qualitative comparisons (Fig. 3) and quantitative perceptual evaluations (Table 4) further confirm that our adversarial examples exhibit enhanced imperceptibility, fulfilling a key requirement for practical attack efficacy.

Table 4. **Comparison of accuracy and image perceptual quality of AEs**, with $\ell_\infty \leq 10$. Our method further improves the black-box accuracy across both domains and models, and the generated adversarial examples improve pixel-level similarity by a slight margin in parentheses while maintaining the structural integrity (–).

Method	**Cross-domain Acc. ↓**	**Cross-model Acc. ↓**	**PSNR ↑**	**SSIM ↑**	**MS-SSIM ↑**
LTP [29]	49.91	47.40	29.11	0.76	0.94
w/ Ours	**44.51**	**41.23**	29.26 (+0.15)	0.77 (+0.01)	0.95 (+0.01)
BIA [58]	51.07	45.44	28.08	0.75	0.94
w/ Ours	**47.10**	**43.46**	28.76 (+0.68)	0.75 (–)	0.94 (–)
GAMA [1]	48.56	44.53	28.62	0.74	0.94
w/ Ours	**46.09**	**43.35**	28.69 (+0.07)	0.74 (–)	0.94 (–)
FACL [52]	44.05	41.20	28.61	0.74	0.93
w/ Ours	**41.78**	**41.01**	28.67 (+0.05)	0.74 (–)	0.93 (–)
PDCL [53]	43.91	42.81	28.68	0.74	0.94
w/ Ours	**43.06**	**42.69**	28.70 (+0.02)	0.74 (–)	0.94 (–)

5. Conclusion

In this paper, we have proposed a self-distillation attack framework that leverages intermediate generator representations in mean teacher to significantly improve cross-model and domain transferability. By moving beyond the conventional surrogate space manipulation approaches, we hope our novel transferability-enhancing perspective using the generator sheds further light on advancing the attack.

Acknowledgements

This work was supported by the Technology Innovation Program (1415187329, 20024355, Development of autonomous driving connectivity technology based on sensor-infrastructure cooperation) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea).

References

[1] Abhishek Aich, Calvin-Khang Ta, Akash Gupta, Chengyu Song, Srikanth Krishnamurthy, Salman Asif, and Amit Roy-Chowdhury. Gama: Generative adversarial multi-object scene attacks. *Advances in Neural Information Processing Systems*, 35:36914–36930, 2022. 1, 2, 3, 4

[2] Shumeet Baluja and Ian Fischer. Adversarial transformation networks: Learning to generate adversarial examples. *arXiv preprint arXiv:1703.09387*, 2017.

[3] Shumeet Baluja and Ian Fischer. Learning to attack: Adversarial transformation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1

[4] Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2022. 2

[5] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 2

[6] 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. 1

[7] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 2

[8] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3

[10] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pages 1607–1616. PMLR, 2018. 2

[11] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 3

[12] 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, 2

[13] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6567–6576, 2021. 2

[14] Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Diverse generative perturbations on attention space for transferable adversarial attacks. In *2022 IEEE international conference on image processing (ICIP)*, pages 281–285. IEEE, 2022. 2

[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 3, 4

[16] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2, 3

[17] Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Improving adversarial transferability via intermediate-level perturbation decay. *Advances in Neural Information Processing Systems*, 36:32900–32912, 2023. 1, 2

[18] Qizhang Li, Yiwen Guo, and Wangmeng Zuo. Enhancing visual adversarial transferability via affine transformation of intermediate-level perturbations. *Pattern Recognition Letters*, 191:51–57, 2025. 1, 2

[19] Zheng Li, Xiang Li, Lingfeng Yang, Renjie Song, Jian Yang, and Zhigeng Pan. Dual teachers for self-knowledge distillation. *Pattern Recognition*, 151:110422, 2024. 2

[20] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019. 2

[21] 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

[22] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *ArXiv*, abs/1306.5151, 2013. 3, 4

[23] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1485–1488, 2010. 3

[24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 2

[25] Krishna Kanth Nakka and Alexandre Alahi. Nat: Learning to attack neurons for enhanced adversarial transferability. In

2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 7593–7604. IEEE, 2025. 1, 2

[26] 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. 2, 3

[27] 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, 2, 3

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3, 4

[29] Mathieu Salzmann et al. Learning transferable adversarial perturbations. *Advances in Neural Information Processing Systems*, 34:13950–13962, 2021. 1, 2, 3, 4

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 4

[31] Naman D Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. In *NeurIPS*, 2023. 3

[32] 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

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2

[34] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. 3, 4

[35] Chaoyue Wang, Chang Xu, Chaohui Wang, and Dacheng Tao. Perceptual adversarial networks for image-to-image transformation. *IEEE Transactions on Image Processing*, 27 (8):4066–4079, 2018. 1

[36] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022. 2

[37] 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. 2

[38] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16158–16167, 2021. 2

[39] 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. 2

[40] 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. 2

[41] Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Enhancing the self-universality for transferable targeted attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12281–12290, 2023. 2

[42] Ross Wightman. PyTorch Image Models. 3

[43] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1161–1170, 2020. 2

[44] Weibin Wu, Yuxin Su, Michael R Lyu, and Irwin King. Improving the transferability of adversarial samples with adversarial transformations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9024–9033, 2021. 2

[45] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018. 1

[46] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017. 3

[47] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. In *Computer Vision and Pattern Recognition*. IEEE, 2019. 1

[48] 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

[49] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017. 3

[50] Chuanguang Yang, Zhulin An, Helong Zhou, Linhang Cai, Xiang Zhi, Jiwen Wu, Yongjun Xu, and Qian Zhang. Mixskd: Self-knowledge distillation from mixup for image recognition. In *European Conference on Computer Vision*, 2022. 2

[51] Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Facl-attack: Frequency-aware contrastive learning for transferable adversarial attacks. 2023. 1, 2, 3

[52] Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Facl-attack: Frequency-aware contrastive learning for transferable adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6494–6502, 2024. 3, 4

[53] Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Prompt-driven contrastive learning for transferable adversarial attacks. In *European Conference on Computer Vision*, pages 36–53. Springer, 2024. 1, 2, 3, 4

[54] Xiaotong Yu, Shiding Sun, and Yingjie Tian. Self-distillation and self-supervision for partial label learning. *Pattern Recognition*, 146:110016, 2024. 2

[55] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. Regularizing class-wise predictions via self-knowledge distillation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[56] Chaoning Zhang, Adil Karjauv, Philipp Benz, Soomin Ham, Gyusang Cho, Chan-Hyun Youn, and In So Kweon. Is fgsm optimal or necessary for l∞ adversarial attack? In *Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV)*. Computer Vision Foundation (CVF), IEEE Computer Society, 2021. 1

[57] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2021. 2

[58] 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. *arXiv preprint arXiv:2201.11528*, 2022. 1, 2, 3, 4

[59] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018. 2