Improving Integrated Gradient-based Transferable Adversarial Examples by Refining the Integration Path

Yuchen Ren[1], Zhengyu Zhao[1]*, Chenhao Lin[1], Bo Yang[2], Lu Zhou[3], Zhe Liu[4], Chao Shen[1]

[1]Xi'an Jiaotong University
[2]Information Engineering University
[3]Nanjing University of Aeronautics and Astronautics
[4]Zhejiang Lab
ryc98@stu.xjtu.edu.cn, {zhengyu.zhao, linchenhao}@xjtu.edu.cn,
yangbo_hn@163.com, {lu.zhou, zhe.liu}@nuaa.edu.cn, chaoshen@mail.xjtu.edu.cn

Abstract

Transferable adversarial examples are known to cause threats in practical, black-box attack scenarios. A notable approach to improving transferability is using integrated gradients (IG), originally developed for model interpretability. In this paper, we find that existing IG-based attacks have limited transferability due to their naive adoption of IG in model interpretability. To address this limitation, we focus on the IG integration path and refine it in three aspects: multiplicity, monotonicity, and diversity, supported by theoretical analyses. We propose the Multiple Monotonic Diversified Integrated Gradients (MuMoDIG) attack, which can generate highly transferable adversarial examples on different CNN and ViT models and defenses. Experiments validate that MuMoDIG outperforms the latest IG-based attack by up to 37.3% and other state-of-the-art attacks by 8.4%. In general, our study reveals that migrating established techniques to improve transferability may require non-trivial efforts.

Appx. & Code — https://github.com/RYC-98/MuMoDIG

Introduction

Deep learning networks (DNNs) are known to be vulnerable to adversarial attacks, which slightly alter input examples to cause model prediction errors (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015). Adversarial attacks can be divided into white-box and black-box attacks. White-box attacks assume that all information about the target model is transparent. Conversely, black-box attacks (Wang and He 2021; Brendel, Rauber, and Bethge 2018; Ren et al. 2024) are more challenging due to the lack of model details. In particular, transferable black-box attacks (Liu et al. 2017) even assume no access to the target model's output, and they just rely on the transferability of adversarial examples generated on known surrogate models. To assess and then improve the robustness of DNN models in practical scenarios, various transferable attacks have been proposed (Long et al. 2022; Wu et al. 2020; Zhao, Liu, and Larson 2021).

Recently, several transferable attacks (Huang and Kong 2022; Ma et al. 2023) have migrated the idea of integrated gradients (IG) (Sundararajan, Taly, and Yan 2017) from the task of DNN interpretability to transferable attacks. Originally, IG interprets a DNN model by attributing its predictions to its input features along a given path. In the context of transferable attacks, the accumulation effect of IG leads to a stable update direction of adversarial gradients and substantially enhances transferability. In particular, existing IG-based attacks naively adopt IG implementations that rely on a single path.

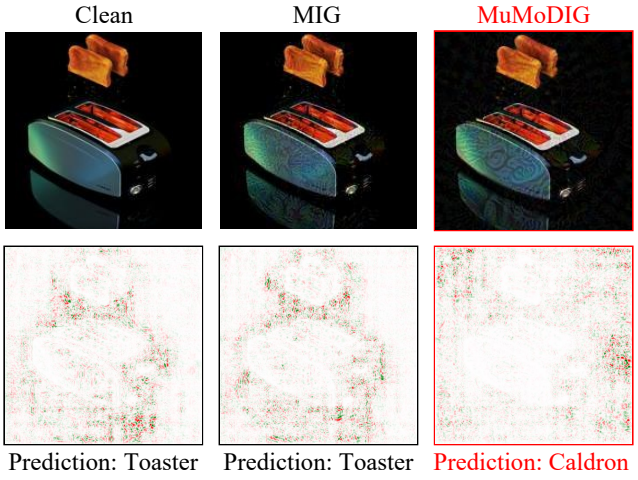

Figure 1: Model attribution results for adversarial examples generated by our MuMoDIG vs. MIG (Ma et al. 2023) on the target model PiT-T. MuMoDIG concentrates more on the background, showcasing its superiority in disrupting the model prediction. Here, RN-18 is the surrogate model.

In this paper, we demonstrate that, however, such a naive adoption has largely limited the transferability of IG-based attacks, since the calculated gradients may be easily affected by the high curvature position of the output manifold (Kapishnikov et al. 2021). To address this limitation, we propose to refine the integration path in three key aspects: multiplicity, monotonicity, and diversity. First, we conduct theoretical analyses to show that directly adopting arbitrary multiple paths even harms the transferability due to the conflict

*Corresponding Author

between integration paths and gradients. Instead, we should ensure the monotonicity of the paths, which is especially achieved by a newly proposed Lower Bound Quantization (LBQ) method. Furthermore, based on our finding that the cosine similarity of interpolated points along the path is too high to avoid overfitting, we propose to diversify the paths based on composite random transformation.

Following the above steps, we finally propose the Multiple Monotonic Diversified Integrated Gradients (MuMoDIG) attack, which can generate highly transferable adversarial examples. As illustrated in Figure 1, the model attribution of adversarial examples generated by our MuMoDIG focuses more on the image background than the foreground objects. This explains the stronger disruption of our adversarial examples on the target model's predictions than the latest IG-based attack, MIG (Ma et al. 2023).

In sum, we make the following main contributions:

- We demonstrate that existing IG-based attacks suffer from limited transferability since they naively adopt IG with a single integration path. We address this limitation by refining the integration path in three key aspects: multiplicity, monotonicity, and diversity.
- We propose a new transferable attack called the Multiple Monotonic Diversified Integrated Gradients (MuMoDIG), and we conduct all-sided analyses to support its design. In particular, to enforce the monotonicity, we propose a Lower Bound Quantization (LBQ) method.
- We validate the superiority of MuMoDIG through extensive experiments on ImageNet against various CNN and ViT models and popular defenses. The results show that MuMoDIG outperforms the latest IG-based attack by up to 37.3% and other advanced attacks by 8.4%.

Related Work

Transferable Attacks

Transferable attacks have been extensively studied due to their practicality. Momentum Iterative Method (MIM) (Dong et al. 2018) incorporates a momentum term in each iteration. Diversity Input Method (DIM) employs padding and resizing input transformations. Skip Gradient Method (SGM) (Wu et al. 2020) scales the gradients passing through the residual modules, reducing the overfitting to the surrogate model. Pay No Attention and Patch Out (PNAPO) (Wei et al. 2022) truncates gradients from the attention branch and randomly masks the input patch to mitigate overfitting. See Zhao et al. (2023) for a detailed review of five typical categories of transferable attacks.

A common practice to further improve transferability is by combining different operations, such as multiple image transformations and gradient modifications. Specifically, Spectrum Simulation Attack (SSA) (Long et al. 2022) combines random Gaussian noise and spectrum masking. Gradient Relevance Attack (GRA) (Zhu et al. 2023) adjusts gradient direction with the aid of noisy inputs, utilizing a gradient relevance framework and decay indicator. Structure Invariant Attack (SIA) (Wang, Zhang, and Zhang 2023) applies a library of ten different image transformations to various local regions of the input.

Integrated Gradients

Gradient-based attribution methods excel in model interpretability, producing visual results aligned with human vision. Integrated Gradients (IG) (Sundararajan, Taly, and Yan 2017) attribute model's prediction to the input along a straight line from a given black image baseline to the input. Blur Integrated Gradients (BlurIG) (Xu, Venugopalan, and Sundararajan 2020) considers the path generated by a sequence of blurred image baselines. Guided Integrated Gradients (GIG) (Kapishnikov et al. 2021) posits that the high curvature of a DNN's output manifold results in larger gradients, which markedly influence the final attribution values. To address this problem, GIG employs an adaptive path method. Important Direction Gradient Integration (IDGI) (Yang, Wang, and Bilgic 2023) further improves the consistency with human visual experience, separating the original gradients into noisy gradients and important gradients, and using only the important gradients for attribution.

Recently, several studies have introduced IG to transferable attacks (Huang and Kong 2022; Ma et al. 2023). The latest MIG (Ma et al. 2023) achieves high transferability by incorporating a momentum term. Our work follows this research direction and specifically addresses the problem of existing IG-based attacks by refining their integration paths.

Methodology

Integrated Gradients in Transferable Attacks

Given a clean image x with channel C, height H, width W, true label y, transferable attacks optimize an adversarial perturbation δ' bounded by ϵ on a surrogate model f. This process can be formulated as:

$$\delta' = \arg\max_{\delta} L(f(x+\delta), y), s.t.\ \|\delta\|_\infty \leq \epsilon, \quad (1)$$

where $f(x)$ denotes the softmax output, and L is the loss function, typically the cross-entropy loss. ϵ is commonly an L_p norm and also other percetual measures are explored (Zhao, Liu, and Larson 2020, 2023; Chen et al. 2024).

Integrated gradients (IG) is a tensor with the same dimension as the input image. Given a surrogate model f, the i-th element of IG during generating the adversarial perturbation can be obtained by:

$$IG(x_t)_i = ((x_t)_i - b_i) \cdot \int_0^1 \frac{\partial f(b + \alpha \cdot (x_t - b))}{\partial (x_t)_i} d\alpha, \quad (2)$$

where $x_t = x + \delta_t$ is the input image at the t-th iteration, the baseline b is typically a black image, and $((x_t) - b)$ is a straight integration path. Eq. (2) can be approximated by:

$$IG(x_t)_i \approx \frac{((x_t)_i - b_i)}{N_I} \cdot \sum_{k=0}^{N_I - 1} \frac{\partial f(b + \frac{k+\lambda}{N_I} \cdot (x_t - b))}{\partial (x_t)_i}, \quad (3)$$

where N_I is the number of interpolation points, and the position factor $\lambda \in [0, 1]$ is introduced to control the position of interpolated points in each segment of the straight path.

In the following, we present our transferable attack method that refines the existing IG-based method by gradually achieving multiple, monotonic, and diversified integration paths, as illustrated in Figure 2.

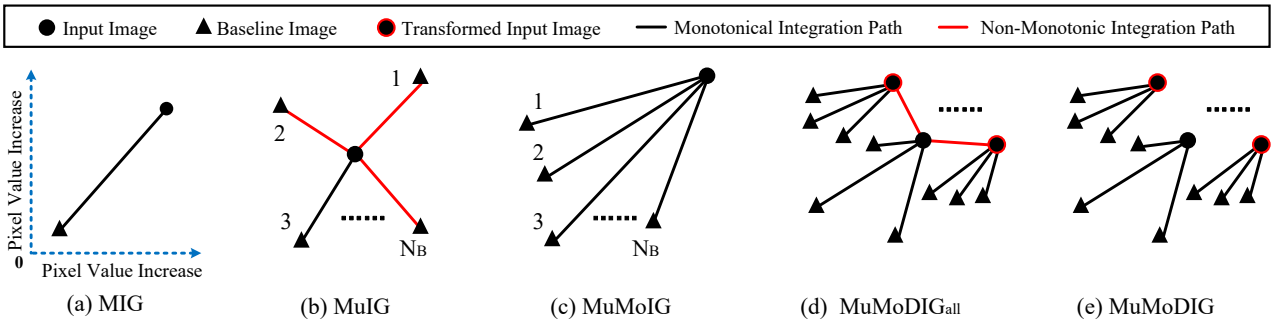

Figure 2: (a) MIG (Ma et al. 2023) adopts a single integration path with a black image baseline. Our (b) MuIG adopts multiple integration paths with arbitrary baselines, with (c) MuMoIG further enforcing monotonicity, (d) MuMoDIG$_{\text{all}}$ diversifying paths and keeping all without enforcing their monotonicity, and (e) MuMoDIG removing non-monotonic diversified paths.

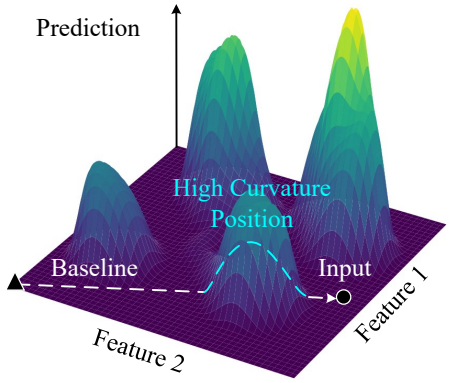

Figure 3: The influence of the output manifold's high curvature position towards the single integration path.

Attack with Multiple Monotonic Integration Paths

Existing work on model interpretability demonstrates that IG with only one integration path may accumulate unstable gradients due to the high curvature position of the output manifold (Kapishnikov et al. 2021), as illustrated in Figure 3. We suppose this problem would also limit the direct use of IG in transferable attacks. Therefore, we propose the Multiple Integrated Gradients (MuIG), assuming that the baseline b follows a certain distribution rather than a fixed black image as before (Huang and Kong 2022; Ma et al. 2023). To this end, the i-th element of MuIG becomes:

$$MuIG(x_t)_i = \mathrm{E}_b(IG(x_t)_i), \tag{4}$$

which adds the expectation operation E_b to Eq. 2.

When utilizing multiple integration paths to interpret DNNs, there are numerous options for baseline selection (Sturmfels, Lundberg, and Lee 2020), such as solid color images, noisy images, and blurred images (Xu, Venugopalan, and Sundararajan 2020). However, our exploratory experiments show that using arbitrary baselines severely harms transferability as the number of baselines increases. This is because in this case, the sign of the integration path $(x_t - b)$ does not necessarily align with the sign of the gradient $\frac{\partial f(x_t+\alpha\cdot(x_t-b))}{\partial x_t}$. As a result, when the signs of corresponding elements between the two are completely opposite, the gradient direction will be most severely disrupted, leading to opposite update direction and deteriorating transferability.

To avoid such a conflict between the integration path and gradient, we define the Monotonic Integration Path and give a proposition (see proof in Appendix) as follows:

Definition 1 (Monotonic Integration Path) *Consider an integration path consisting of a sequence of interpolated points $(x_0, ..., x_{N_I-1})$. For any interpolated point x_k, if the following conditions hold: 1) $\forall s < k$, $(x_s)_i \leq (x_k)_i$; 2) $\forall m > k$, $(x_m)_i \geq (x_k)_i$, where $0 \leq s < k < m \leq N_I - 1$ and $0 \leq i \leq C \cdot H \cdot W$, then this path is called a Monotonic Integration Path.*

Proposition 1 *The integration path should be a Monotonic Integration Path when using integrated gradients to generate adversarial examples in transferable attacks.*

Definition 1 ensures that the elements of earlier interpolated points in the sequence are always less than or equal to the corresponding elements of the later interpolated points. Proposition 1 engages that the signs of the elements in the gradients during the adversarial example generation process are not altered by the integration path, since all elements of $(x_t - b)$ are positive. Definition 1 also explains the enhanced transferability by using a black image baseline, as in the existing method, MIG (Ma et al. 2023), which naturally forms a monotonic integration path. In contrast, using integration paths constructed by arbitrary baselines, such as noisy images, does not adhere to the above definition.

Inspired by the Randomized Quantization method (Wu et al. 2023), we propose a Lower Bound Quantization (LBQ) method to generate baselines that conform to Proposition 1, *i.e.*, all elements of the baseline $LBQ(x_t)$ are less than or equal to x_t. Specifically, LBQ is implemented as follows:

1. Convert each channel of x_t into a one-dimensional vector, with the elements of each channel's vector sorted in ascending order of value.
2. Randomly select $N_R - 1$ $(2 \leq N_R)$ divisions to split each vector into N_R regions and replace all elements in each region with the minimum value of that region.
3. Convert the above-processed vectors of each channel back to the original dimensions of each channel.

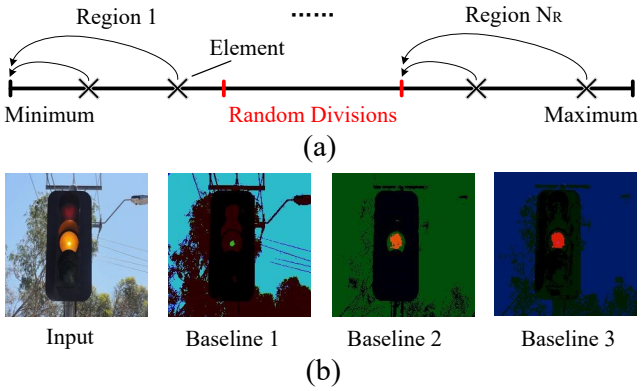

Figure 4: (a) Lower Bound Quantization (LBQ) quantizes all elements in each region to their minimum value, resulting in (b) baseline images that enforce monotonic paths.

Figure 4 illustrates LBQ and visualizes some examples of the generated image baselines. We term MuIG with the above monotonic integration paths as Multiple Monotonic Integrated Gradients (MuMoIG).

Attack with Diversified Integration Paths

After ensuring the monotonicity of multiple integration paths, we further look at the properties of interpolated points along the path. Specifically, we compute the cosine similarity of interpolated points. Figure 5 shows that gradients calculated at a sequence of interpolated points lack diversity, as the cosine similarity between the gradients at adjacent positions is very high. Accumulating such similar gradients using Eq. 3 cannot effectively reduce the overfitting to the surrogate model f, because these gradients provide limited information in the high-dimensional space.

To address this limitation, we follow previous work (Zhang et al. 2024) to apply input transformations to reduce the cosine similarity of gradients as follows:

$$MuMoDIG_{\text{all}}(x_t)_i = \mathrm{E}_T(\mathrm{E}_{b|T}(IG_{\text{all}}(x_t)_i)), \quad (5)$$

where T denotes the input transformations, and all diversified paths are kept no matter if they satisfy monotonicity. This forms the Multiple Monotonic Diversified Integrated Gradients$_{\text{all}}$ (MuMoDIG$_{\text{all}}$). To ensure sufficient gradient diversity, the transformation typically contains composite operations, from which one operation is randomly selected each time. Here, we follow the common practice of using two simple transformations: the resizing and padding (RP) (Xie et al. 2019) and the affine transformation (AF), selected with equal probability.

Furthermore, it is easy to notice that the integration paths from transformed inputs to the input cannot be guaranteed to be monotonic, violating our Proposition 1. Therefore, we discard such non-monotonic paths, and Eq. 5 becomes:

$$MuMoDIG(x_t)_i = \mathrm{E}_T(\mathrm{E}_{b|T}(IG_{\text{mo}}(x_t)_i)), \quad (6)$$

which can be approximated with Monte Carlo Sampling by:

$$MuMoDIG(x_t)_i \approx I \cdot \sum_{p=0}^{N_T} \sum_{q=0}^{N_B-1} (IG_{\text{mo}}(x_t)_{p,q})_i, \quad (7)$$

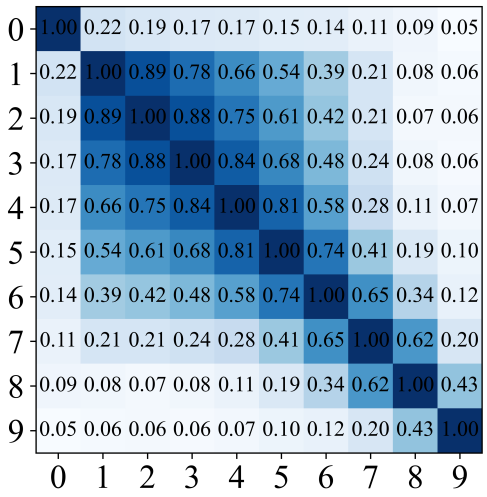

Figure 5: The cosine similarity calculated between the gradients at 10 interpolated points along a straight path.

where $I = \frac{1}{(N_T+1)\cdot N_B}$, with N_T as the number of sampled transformations and N_B as the number of sampled baselines. Here, a fixed identity transformation is also applied to the input image, except for the N_T extra transformations. After all the above steps, we achieve the ultimate version of our method, Multiple Monotonic Diversified Integrated Gradients (MuMoDIG).

Similar to MIG (Ma et al. 2023), our MuMoDIG is also integrated with the momentum term (Dong et al. 2018) to further boost transferability. In this case, the above I can be omitted as it cancels out during gradient normalization. We also find that using $-\log f(x)$ instead of $-f(x)$ for the loss function yields a slightly better performance for both our and existing IG-based attacks.

IG for Interpretability vs. for Transferabability

Here, we shed more light on the fundamental differences between using IG in model interpretability and transferable attacks. In general, IG in model interpretability focuses on providing better visual explanations by prioritizing the overall magnitude of the product between gradients and the integration path. Differently, transferable attacks aim to disrupt predictions, making the sign of the gradient direction more crucial than the product's magnitude.

Specific weaknesses of directly using advanced IG techniques, such as BlurIG, GIG, and IDGI, in transferable attacks include: 1) BlurIG employs a sequence of blurred images as baselines, encountering a similar issue as discussed in MuIG: they fail to form a monotonic integration path; 2) GIG and IDGI, although with black images as baselines, face the same issue as MuMoDIG$_{\text{all}}$: while the straight line connecting their starting and ending points forms a monotonic integration path, the actual integration paths in between cannot guarantee monotonicity.

Attack	RN-18	RN-101	RNX-50	DN-121	ViT-B	PiT-B	Vis-S	Swin-T	AT	HGD	NRP	Bit	JPEG	Mean
MIG	100.0*	55.8	62.3	87.5	20.6	33.6	45.8	49.8	45.8	46.3	42.5	60.2	59.0	50.8
MuIG ($N_B = 6, N_I = 6$)	18.6*	6.6	9.3	14.1	3.5	5.7	6.8	9.1	40.7	5.8	23.9	13.3	23.8	13.6
MuMoIG ($N_B = 1, N_I = 6$)	100.0*	59.2	64.3	91.4	23.1	35.3	49.3	51.7	45.4	50.0	42.7	60.8	62.6	53.0
MuMoIG ($N_B = 6, N_I = 6$)	100.0*	64.4	70.5	94.3	25.9	36.2	53.4	55.4	46.6	54.4	45.0	64.5	66.1	56.4
MuMoDIG$_{\text{all}}$	99.8*	70.6	74.0	94.6	33.6	45.5	61.6	61.0	47.6	64.4	47.8	68.4	69.9	61.6
MuMoDIG	**100.0***	**85.3**	**86.9**	**97.8**	**43.5**	**57.0**	**73.6**	**75.1**	**51.9**	**89.7**	**60.7**	**85.2**	**80.8**	**73.9**

Table 1: Attack success rates (%) of gradually refining MIG (Ma et al. 2023) to form our ultimate attack, MuMoDIG. The surrogate model is RN-18, and the "Mean" column excludes the white-box results marked with *.

Surrogate	Attack	RN-18	RN-101	RNX-50	DN-121	ViT-B	PiT-B	Vis-S	Swin-T	AT	HGD	NRP	Bit	JPEG	Mean
RN-18	MIG	**100.0***	55.8	62.3	87.5	20.6	33.6	45.8	49.8	45.8	46.3	42.5	60.2	59.0	50.8
	GRA	**100.0***	58.2	64.0	90.9	25.3	33.3	47.2	56.9	50.0	51.5	49.6	64.0	64.1	54.6
	SSA	**100.0***	58.3	62.8	90.9	25.0	35.9	47.2	53.0	47.1	50.8	48.3	63.7	66.7	54.1
	SIA	**100.0***	81.2	84.6	**98.2**	36.9	50.4	70.1	71.4	47.6	70.0	49.8	75.5	70.4	67.2
	MuMoDIG	**100.0***	**85.3**	**86.9**	97.8	**43.5**	**57.0**	**73.6**	**75.1**	**51.9**	**89.7**	**60.7**	**85.2**	**80.8**	**73.9**
DN-121	MIG	85.1	65.6	69.7	**100.0***	30.1	40.9	55.9	54.9	44.6	58.2	45.3	64.2	61.1	57.5
	GRA	95.3	82.5	85.0	**100.0***	46.9	58.5	70.7	71.5	**53.5**	80.6	**64.6**	78.6	77.9	72.5
	SSA	94.8	83.0	83.8	**100.0***	43.5	56.1	71.1	70.4	47.6	77.8	60.4	77.1	77.3	70.7
	SIA	**97.2**	86.7	90.1	**100.0***	43.5	59.6	77.4	73.4	47.5	80.9	53.8	79.6	72.8	72.1
	MuMoDIG	**97.2**	**88.7**	**90.2**	**100.0***	**49.1**	**63.3**	**78.8**	**75.5**	47.8	**87.9**	57.8	**82.4**	**78.2**	**75.0**
MN-v3	MIG	69.2	30.0	35.8	60.0	20.1	26.7	35.0	39.8	46.8	46.3	34.3	45.0	48.7	41.4
	GRA	75.0	32.6	38.7	61.5	21.4	27.8	35.0	43.2	50.8	51.5	38.9	49.3	54.5	44.6
	SSA	71.6	28.1	35.9	61.4	17.6	24.4	31.7	38.4	49.1	50.8	37.2	46.3	53.9	42.0
	SIA	84.4	43.4	48.8	75.9	26.9	37.2	48.4	52.2	50.3	70.0	40.6	57.9	56.9	53.3
	MuMoDIG	**88.6**	**52.4**	**57.0**	**80.9**	**36.4**	**46.8**	**57.2**	**62.4**	**51.7**	**89.7**	**47.3**	**66.2**	**65.4**	**61.7**
PiT-T	MIG	70.0	38.6	44.3	64.5	32.2	43.5	46.0	56.6	45.9	35.7	37.8	54.3	55.4	48.1
	GRA	86.9	61.5	66.2	82.2	57.1	68.2	68.9	77.8	52.6	58.2	57.9	74.3	72.0	68.0
	SSA	84.2	55.7	61.2	77.8	46.3	60.2	62.4	72.0	49.4	53.6	51.3	68.9	71.1	62.6
	SIA	**93.7**	71.9	76.7	88.8	64.2	77.2	79.8	84.3	51.8	64.2	55.0	80.7	74.6	74.1
	MuMoDIG	92.5	**74.8**	**79.0**	**90.5**	**69.5**	**80.3**	**80.9**	**86.9**	**53.5**	**73.9**	**59.3**	**83.4**	**78.4**	**77.1**
DeiT-T	MIG	66.4	36.0	41.6	61.6	56.2	39.6	46.3	68.6	45.3	31.3	39.8	56.3	57.0	49.7
	GRA	81.2	52.6	59.5	77.1	77.0	60.5	65.8	83.6	50.9	49.7	56.2	73.2	72.5	66.1
	SSA	79.0	48.2	54.2	73.0	70.9	52.6	58.9	78.5	49.3	45.6	50.2	68.5	71.3	61.6
	SIA	90.4	70.0	75.8	87.8	83.3	79.8	83.5	90.2	52.0	64.8	58.3	82.8	77.2	76.6
	MuMoDIG	**91.6**	**73.8**	**77.5**	**89.3**	**83.8**	**82.4**	**84.8**	**90.9**	**53.8**	**73.4**	**63.2**	**84.2**	**80.8**	**79.2**
Swin-T	MIG	47.2	25.5	32.4	42.1	23.7	33.5	41.9	98.8*	42.9	19.9	28.7	45.7	45.2	35.7
	GRA	**83.0**	69.9	74.4	80.8	**73.7**	81.3	84.4	98.6*	**51.8**	66.1	**69.1**	80.5	**78.4**	**74.5**
	SSA	80.4	66.2	70.2	79.8	64.3	75.1	82.4	**98.9***	47.5	61.9	62.5	78.2	77.7	70.5
	SIA	79.3	67.1	72.7	80.8	59.0	79.5	85.2	98.7*	45.3	54.9	48.2	77.0	68.1	68.1
	MuMoDIG	82.8	**73.4**	**76.9**	**84.3**	67.6	**82.7**	**85.2**	98.3*	46.9	**66.3**	53.5	**82.0**	73.9	73.0

Table 2: Attack success rates (%) of our MuMoDIG vs. state-of-the-art transformation-based attacks. For defenses, AT uses RN-50, HGD uses its default setting, and NRP, JPEG, and Bit results are averaged on eight target models. The results with underlined in the "Mean" columns are the second best, and the "Mean" column excludes the white-box results marked with *.

Experiments

Experimental Settings

Dataset and attack baselines. Following many previous works (Wang and He 2021; Zhu et al. 2023; Long et al. 2022), 1k images from the ILSVRC2012 (Russakovsky et al. 2015) validation set are adopted in our experiments. We compare our MuMoDIG to the latest IG-based attack, MIG (Ma et al. 2023), as well as state-of-the-art transferable attacks with composite operations, such as SSA (Long et al. 2022), GRA (Zhu et al. 2023), and SIA (Wang, Zhang, and Zhang 2023). All of them are equipped with multiple input transformations or special gradient modifications based on at least one transformation. Note that Path-Augmented Method (PAM) (Zhang et al. 2023) and Neuron Attribution-based Attack (NAA) (Zhang et al. 2022) are path-related but not IG-based attacks, and detailed results in the Appendix show the superiority of our MuMoDIG over them.

Parameters. Following the common practice, for all attacks, we set the maximum attack iterations as $K = 10$, the maximum perturbation bound $\epsilon = 16$, the step size $\alpha = 1.6$, the decay factor $\mu = 1.0$ in the momentum. We set the position factor $\lambda = 0.65$ and the region number $N_R = 2$ in LBQ. For a fair comparison, we set the number of total auxiliary inputs $N = 6$ at each iteration for all attacks. Specifically, for our

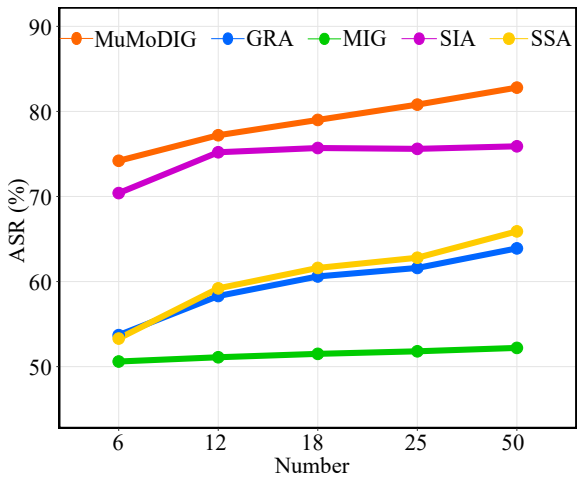

Figure 6: Average success rates (%) when adjusting the number of total auxiliary inputs N. The surrogate model is RN-18, and target models are the remaining seven models.

Surrogate	Attack	CNNs	ViTs
RN-18	SGM	61.0	36.4
	MuMoDIG	90.0	62.3
	MuMoDIG+SGM	**91.1**	**66.3**
PiT-T	PNAPO	58.7	55.5
	MuMoDIG	81.4	79.4
	MuMoDIG+PNAPO	**83.4**	**82.0**

Table 3: Attack success rates (%) of integrating MuMoDIG with other attacks averaged on CNNs and ViTs.

MuMoDIG, we set $N_T = 6$, $N_B = 1$, and $N_I = 1$ such that $N = N_T \cdot N_B \cdot N_I = 6$. All experiments are conducted on an RTX 4060 GPU with 8GB of VRAM. Generating one image (RN-18) for MIG, GRA, SSA, SIA, and our MuMoDIG costs: 0.21s, 0.24s, 0.19s, 0.20s, and 0.28s, respectively.

Models and defenses. Surrogate models contain three CNNs, *i.e.*, RN-18 (He et al. 2016), DN-121 (Huang et al. 2017) and MN-v3 (Howard et al. 2017), and three ViTs, *i.e.*, PiT-T (Heo et al. 2021), Deit-T (Touvron et al. 2021) and Swin-T (Liu et al. 2021). Target models contain four CNNs, *i.e.*, RN-18, RN-101, RNX-50 (Xie et al. 2017), DN-121, and four ViTs, *i.e.*, ViT-B (Dosovitskiy et al. 2020), PiT-B, Vis-S (Chen et al. 2021), and Swin-T.

Defenses include adversarial training (AT) (Tramèr et al. 2018), high-level representation guided denoiser (HGD) (Liao et al. 2018), neural representation purifier (NRP) (Naseer et al. 2020), Bit depth reduction (BDR) (Xu, Evans, and Qi 2018), and JPEG compression (Guo et al. 2018). We also use MuMoDIG to attack the online Baidu Cloud API.

Attack Results

MuMoDIG vs. other IG-based attacks. We first compare our MuMoDIG and its intermediate versions with the latest IG-based attack, MIG. Table 1 shows the results, which fully support our claims in the above section. Specifically,

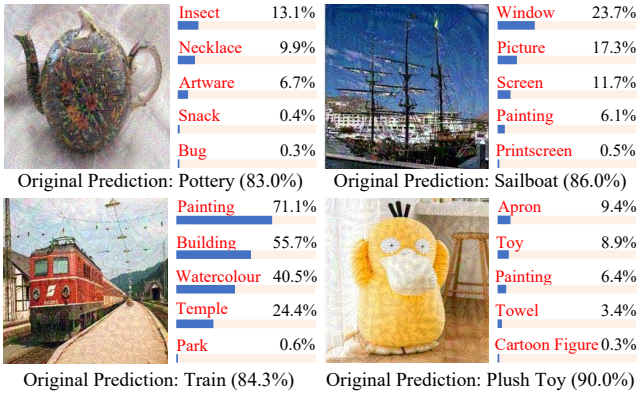

Figure 7: Attack examples of our MuMoDIG on Baidu Cloud API. RN-18 is the surrogate model. Results on 100 images show that our MuMoDIG achieves a high ASR of 91.0%, compared to 74.0% for GRA and 87.0% for SIA.

MuIG performs the worst, indicating that using arbitrary multiple baselines even harms the transferability. MuMoIG ($N_B = 1, N_I = 6$) and MuMoIG ($N_B = 6, N_I = 6$) exhibit stronger performance than MIG, highlighting the effectiveness of multiple integration paths. Lastly, MuMoDIG outperforms MuMoDIG$_{\text{all}}$, confirming that removing the non-monotonic paths from the diversified paths is reasonable.

MuMoDIG vs. state-of-the-art transformation-based attacks. Table 2 shows our MuMoDIG consistently outperforms other advanced attacks in almost all cases. For example, when generating adversarial examples on MN-v3 and attacking Swin-T, our MuMoDIG achieves an attack success rate of 62.4%, which is 10.2% higher than the best of the other attacks. In addition, when attacking models with defenses, MuMoDIG still outperforms others in most cases. Figure 6 further shows that the superiority of MuMoDIG is consistent across varied N. Note that SIA (Wang, Zhang, and Zhang 2023) applies different transformations to specific image blocks, while our composite random transformation applies fewer transformations to the entire image.

MuMoDIG integrated with other attacks. In addition to the above transformation-based attacks, we integrate our MuMoDIG with another typical type of transferable attack that refines the surrogate model. Two such attacks are involved, with SGM (Wu et al. 2020) targeting CNNs with residual blocks and PNAPO (Wei et al. 2022) targeting ViTs. From Table 3, our MuMoDIG can further boost the transferability when being integrated with these two attacks.

MuMoDIG on attacking a real-world system. To test the practical usefulness of MuMoDIG, we finally consider a challenging setting with the real-world visual system, the Baidu Cloud API. As illustrated in Figure 7, for all tested examples, adding the perturbations causes the target system to make incorrect predictions, despite the original predictions being correct with high confidence.

Ablation Studies

We analyze the influence of four important parameters in MuMoDIG on the attack performance. The surrogate is RN-

Type/Number	1	6	12	18	25	50
N_I	**59.2**	66.8	67.0	67.1	66.9	67.9
N_B	**59.2**	69.8	71.7	72.6	73.1	75.0
N_T	**59.2**	**74.2**	**77.2**	**79.0**	**80.8**	**82.8**

Table 4: Ablation studies about the interpolation point number N_I, baseline number N_B, or sampled transformation number N_T. Results are averaged on seven target models. When one number is varied, the others are fixed as 1.

Type	CNNs	ViTs
$N_T = 1, N_B = 1, N_I = 6$	84.7	53.3
$N_T = 1, N_B = 2, N_I = 3$	85.8	56.5
$N_T = 1, N_B = 3, N_I = 2$	87.3	57.0
$N_T = 1, N_B = 6, N_I = 1$	86.6	57.2
$N_T = 2, N_B = 1, N_I = 3$	88.1	57.8
$N_T = 2, N_B = 3, N_I = 1$	88.4	58.8
$N_T = 3, N_B = 1, N_I = 2$	89.0	59.9
$N_T = 3, N_B = 2, N_I = 1$	89.4	61.1
$N_T = 6, N_B = 1, N_I = 1$	**90.0**	**62.3**

Table 5: Ablation studies about combinations of the sampled transformation number N_T, baseline number N_B, and interpolation point number N_I, with $N = N_T \cdot N_B \cdot N_I = 6$.

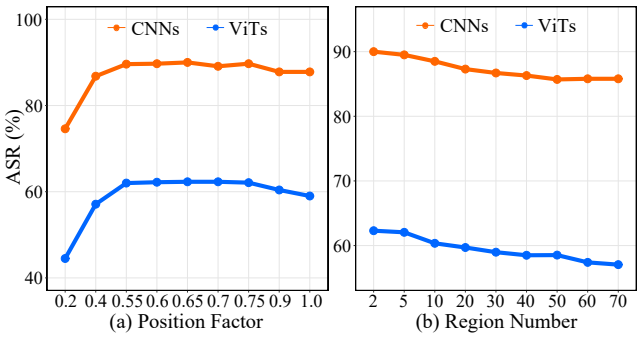

Figure 8: Ablation studies of (a) position factor λ and (b) region number N_R.

Transformation	CNNs	ViTs
Affine Transformation (AF)	82.1	53.1
Resizing&Padding (RP)	85.4	57.1
AF+Blur	83.1	51.2
RP+Blur	85.6	54.9
AF+Noise	85.6	55.5
RP+Noise	87.5	58.4
AF+RP (MuMoDIG)	**90.0**	**62.3**

Table 6: Ablation studies of the input transformation T.

18, and the target CNNs are RN-101, DN-121, and RNX-50, and ViTs are ViT-B, PiT-B, Visformer-S, and Swin-T.

The number of total auxiliary inputs N is calculated by multiplying N_I, N_B, and N_T. Table 4 shows that increasing the number of either of these three types of auxiliary inputs can enhance the transferability. Specifically, increasing N_T is the most effective, while increasing N_I is the least effective. This is because the pixel variations of interpolated points are less diverse, as they only reflect pixel scaling, which is more susceptible to overfitting to the surrogate model. Results in Table 5 further support it with different combinations of auxiliary inputs.

The position factor λ determines the position of the interpolated points in each interval along the straight path. As can be seen from Figure 8 (a), the attack performance is not sensitive to this λ and a moderate $\lambda = 0.65$ leads to a satisfactory attack performance.

The region number N_R is a critical parameter in LBQ. A large N_R indicates that the produced baseline is close to the input, whereas a small N_R means that the generated baseline is close to a black image baseline. Figure 8 (b) demonstrates that transferability diminishes as N_R increases. This is because a larger N_R makes the produced baseline closely approximate the input. Therefore, the integration paths become shorter, causing interpolated points along each path to become more similar. This similarity leads to greater gradient homogeneity, increasing the risk of overfitting to the surrogate model and consequently reducing transferability. Note that $N_R = 1$ means the black image baseline.

The input transformation T may also influence the attack performance. As can be seen from Table 6, removing either of the existing transformations or replacing it with another, common transformation, such as blur or noise, will decrease the transferability.

Conclusion and Outlook

In this paper, we have improved the transferability of integrated gradients (IG)-based attacks by refining their integration paths in three aspects: multiplicity, monotonicity, and diversity. Concretely, we propose the Multiple Monotonic Diversified Integrated Gradients (MuMoDIG) attack, which can craft highly transferable adversarial examples on various models and defenses. In particular, the design of MuMoDIG is supported by our theoretical analyses of the fundamental differences in using IG for model interpretation and transferable attacks. For future work, we would continue to advance the use of integrated gradients in transferable attacks from the perspectives of baseline generation and more reasonable integration paths. Additionally, we would further promote the fusion of other interpretability methods with transferable attacks for more explainable model evaluations.

Acknowledgments

This research is supported by the National Key Research and Development Program of China (2023YFE0209800), the National Natural Science Foundation of China (62406240, T2341003, 62376210, 62161160337, 62132011, U2441240, U21B2018, U20A20177, 62206217, U24B20185), the Shaanxi Province Key Industry Innovation Program (2023-ZDLGY-38).

References

Brendel, W.; Rauber, J.; and Bethge, M. 2018. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *Proceedings of the International Conference on Learning Representations*.

Chen, Z.; Li, B.; Wu, S.; Jiang, K.; Ding, S.; and Zhang, W. 2024. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36.

Chen, Z.; Xie, L.; Niu, J.; Liu, X.; Wei, L.; and Tian, Q. 2021. Visformer: The Vision-friendly Transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, 569–578.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting Adversarial Attacks with Momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9185–9193.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*.

Guo, C.; Rana, M.; Cisse, M.; and van der Maaten, L. 2018. Countering Adversarial Images using Input Transformations. In *Proceedings of the International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; and Oh, S. J. 2021. Rethinking Spatial Dimensions of Vision Transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 11916–11925.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2261–2269.

Huang, Y.; and Kong, A. W.-K. 2022. Transferable Adversarial Attack based on Integrated Gradients. In *Proceedings of the International Conference on Learning Representations*.

Kapishnikov, A.; Venugopalan, S.; Avci, B.; Wedin, B.; Terry, M.; and Bolukbasi, T. 2021. Guided Integrated Gradients: an Adaptive Path Method for Removing Noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5048–5056.

Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; and Zhu, J. 2018. Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1778–1787.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2017. Delving into Transferable Adversarial Examples and Black-box Attacks. In *International Conference on Learning Representations (ICLR)*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE International Conference on Computer Vision*, 9992–10002.

Long, Y.; Zhang, Q.; Zeng, B.; Gao, L.; Liu, X.; Zhang, J.; and Song, J. 2022. Frequency Domain Model Augmentation for Adversarial Attack. In *Proceedings of the European Conference on Computer Vision*, 549–566.

Ma, W.; Li, Y.; Jia, X.; and Xu, W. 2023. Transferable Adversarial Attack for Both Vision Transformers and Convolutional Networks via Momentum Integrated Gradients. In *Proceedings of the IEEE International Conference on Computer Vision*, 4607–4616.

Naseer, M.; Khan, S.; Hayat, M.; Khan, F. S.; and Porikli, F. 2020. A Self-supervised Approach for Adversarial Robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 259–268.

Ren, Y.; Zhu, H.; Liu, C.; and Li, C. 2024. Efficient polar coordinates attack with adaptive activation strategy. *Expert Systems with Applications*, 249: 123850.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3): 211–252.

Sturmfels, P.; Lundberg, S.; and Lee, S.-I. 2020. Visualizing the Impact of Feature Attribution Baselines. *Distill*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the International Conference on Machine Learning*, volume 70, 3319–3328.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning*, 10347–10357.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *Proceedings of the International Conference on Learning Representations*.

Wang, X.; and He, K. 2021. Enhancing the Transferability of Adversarial Attacks through Variance Tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1924–1933.

Wang, X.; Zhang, Z.; and Zhang, J. 2023. Structure Invariant Transformation for better Adversarial Transferability. In *Proceedings of the IEEE International Conference on Computer Vision*, 4584–4596.

Wei, Z.; Chen, J.; Goldblum, M.; Wu, Z.; Goldstein, T.; and Jiang, Y.-G. 2022. Towards Transferable Adversarial Attacks on Vision Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2668–2676.

Wu, D.; Wang, Y.; Xia, S.-T.; Bailey, J.; and Ma, X. 2020. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. In *Proceedings of the International Conference on Learning Representations*.

Wu, H.; Lei, C.; Sun, X.; Wang, P.-S.; Chen, Q.; Cheng, K.-T.; Lin, S.; and Wu, Z. 2023. Randomized Quantization: A Generic Augmentation for Data Agnostic Self-supervised Learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 16305–16316.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving Transferability of Adversarial Examples With Input Diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Xie, S.; Girshick, R.; Doll¨¢r, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5987–5995.

Xu, S.; Venugopalan, S.; and Sundararajan, M. 2020. Attribution in Scale and Space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9677–9686.

Xu, W.; Evans, D.; and Qi, Y. 2018. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Proceedings of the Network and Distributed System Security Symposium*.

Yang, R.; Wang, B.; and Bilgic, M. 2023. IDGI: A Framework to Eliminate Explanation Noise from Integrated Gradients. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 23725–23734.

Zhang, J.; Huang, J.-t.; Wang, W.; Li, Y.; Wu, W.; Wang, X.; Su, Y.; and Lyu, M. R. 2023. Improving the transferability of adversarial samples by path-augmented method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8173–8182.

Zhang, J.; Wu, W.; Huang, J.-t.; Huang, Y.; Wang, W.; Su, Y.; and Lyu, M. R. 2022. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14993–15002.

Zhang, Y.; Hu, S.; Zhang, L.; Shi, J.; Li, M.; Liu, X.; Wan, W.; and Jin, H. 2024. Why Does Little Robustness Help? A Further Step Towards Understanding Adversarial Transferability. In *Proceedings of the IEEE Symposium on Security and Privacy*, 14–14.

Zhao, Z.; Liu, Z.; and Larson, M. 2020. Towards Large Yet Imperceptible Adversarial Image Perturbations With Perceptual Color Distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1036–1045.

Zhao, Z.; Liu, Z.; and Larson, M. 2021. On Success and Simplicity: A Second Look at Transferable Targeted Attacks. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, 6115–6128.

Zhao, Z.; Liu, Z.; and Larson, M. 2023. Adversarial Image Color Transformations in Explicit Color Filter Space. *IEEE Transactions on Information Forensics and Security*, 18: 3185–3197.

Zhao, Z.; Zhang, H.; Li, R.; Sicre, R.; Amsaleg, L.; Backes, M.; Li, Q.; and Shen, C. 2023. Revisiting transferable adversarial image examples: Attack categorization, evaluation guidelines, and new insights. *arXiv preprint arXiv:2310.11850*.

Zhu, H.; Ren, Y.; Sui, X.; Yang, L.; and Jiang, W. 2023. Boosting Adversarial Transferability via Gradient Relevance Attack. In *Proceedings of the IEEE International Conference on Computer Vision*, 4718–4727.