BS-IG: Boundary-Aware Scale-Adaptive Integrated Gradients for Transferable Adversarial Attacks

Shuaiye Lu* Linjiang Zhou† Chao Ma† Xiaochuan Shi† Yuqing Li†
* Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China
† School of Cyber Science and Engineering, Wuhan University, Wuhan, China
Email: {shuaiyelu, linjiang, chaoma, shixiaochuan, liyuqing}@whu.edu.cn

Abstract*—The transferability of adversarial examples remains as a critical challenge, particularly in cross-architecture scenarios (e.g., from CNNs to ViTs). Existing Integrated Gradients (IG)-based attacks attempt to bypass high-curvature regions on the surrogate manifold via macroscopic path planning, yet they neglect the ubiquitous microscopic gradient oscillations caused by local ruggedness. From a geometric perspective, we identify that such oscillations trap perturbations in surrogate-specific local optima, hindering effective transferability. Furthermore, we reveal that directly applying adaptive smoothing to mitigate this noise triggers a *Boundary Collapse* phenomenon, where the sampling region vanishes at data boundaries, leading to optimization failure during the crucial initialization phase. To address these issues, we propose Boundary-aware Scale-adaptive Integrated Gradients (BS-IG). Our method integrates a scale-adaptive sampling mechanism with a minimum sampling region constraint, which effectively smooths high-curvature fluctuations while preventing boundary collapse along the entire integration path. Extensive experiments on ImageNet demonstrate that BS-IG significantly enhances transferability across diverse black-box models, outperforming state-of-the-art IG-based methods by up to 5.9% and other state-of-the-art attacks by an average of 2.42%. Furthermore, when combined with other IG-based methods, the performance can be boosted by 12.94%. *The source code of BS-IG is publicly available at https://github.com/AiShare-WHU/BS-IG.

I. Introduction

Deep Neural Networks (DNNs) remain vulnerable to adversarial examples [1]–[3]. To evaluate real-world robustness without full model access, black-box transfer-based attacks [4]–[6] are often preferred over query-intensive methods [7], [8], leveraging surrogate models to deceive unknown targets.

Existing methods enhance transferability via momentum (e.g., MI-FGSM [9]–[11]) or input transformations (e.g., TIM [12]–[14]), but fundamentally rely on single-point gradients, limiting stability. To address this, Integrated Gradients (IG) [15] has been introduced [4], [5]. By accumulating gradient information along a path, IG yields a highly stable update direction, effectively bypassing surrogate-specific artifacts and significantly boosting cross-model transferability [5], [6].

Nevertheless, existing IG-based methods struggle with cross-architecture transferability [16], [17]. CNNs and Vision Transformers (ViTs) exhibit fundamental processing disparities: CNNs extract local features [18], while ViTs capture global dependencies [16]. This divergence causes severe surrogate overfitting, manifesting geometrically as high-curvature regions on the output manifold [19], [20]. Traversing these rugged regions causes violent gradient oscillations [6], [21], trapping perturbations in surrogate-specific local optima and hindering the extraction of robust, transferable features [20], [22].

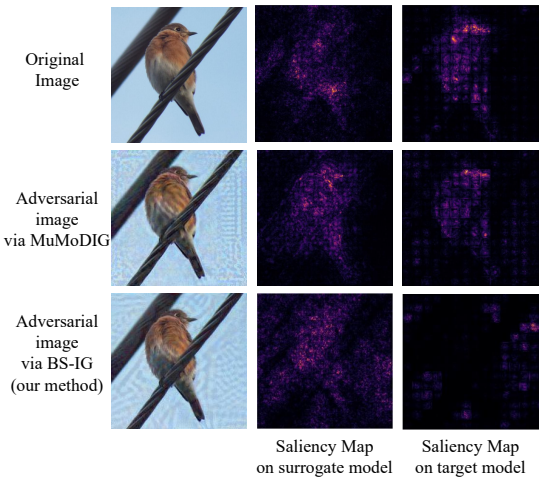


Fig. 1: **Visual comparison of adversarial examples and Integrated Gradients (IG) maps.** The columns display the generated adversarial images, saliency maps on the source model (ResNet-50), and saliency maps on the target model (ViT-B). Unlike MuMoDIG, the proposed BS-IG maintains highly consistent saliency distributions across diverse architectures, effectively shifting model attention to background regions.

To address this challenge, methods like MIG and MuMoDIG [5], [6] attempt to *bypass* high-curvature regions via macroscopic path planning. However, given the inherently rugged nature of the surrogate manifold, gradient estimation remains susceptible to ubiquitous microscopic oscillations [19]. While borrowing AdaptGrad [23] from the field of interpretability offers a theoretical remedy for such noise, its direct application to adversarial generation triggers a critical failure mode.Specifically, to strictly prevent out-of-distribution (OOD) noise, the allowable sampling radius mathematically

Shuaiye Lu and Linjiang Zhou contribute equally to this work. Xiaochuan Shi is the corresponding author of this paper.

converges to zero near the integration endpoints (i.e., data boundaries). We formally term this phenomenon *Boundary Collapse*, which prevents effective smoothing during the crucial initialization phase, causing the attack to rapidly fall into surrogate-specific local optima.

In this work, we propose Boundary-aware Scale-adaptive Integrated Gradients (BS-IG). Our core strategy leverages the adaptive sampling mechanism of AdaptGrad while introducing a minimum sampling region constraint. This design not only effectively smooths high-curvature surfaces on the rugged manifold but also fundamentally resolves the boundary collapse issue, ensuring sufficient exploration range even at the integration endpoints. Consequently, BS-IG accurately captures truly robust shared features. As illustrated in Figure 1, compared to MuMoDIG [6], the adversarial examples generated by BS-IG exhibit highly consistent IG distributions across both the source (ResNet-50 [24]) and target (ViT-B [25]) models. This confirms that our method effectively misleads both models into simultaneously shifting their attention from the primary object to irrelevant background regions, underscoring the critical importance of generating stable and consistent gradient directions.

Our main contributions are summarized as follows:

- **Geometric Analysis:** We provide a geometric analysis of the high-curvature bottleneck hindering cross-architecture transferability. We identify the absence of microscopic smoothing in existing IG variants and reveal the *Boundary Collapse* issue arising from the direct application of adaptive sampling, which traps the algorithm in local optima.
- **Method Proposal:** We propose the BS-IG algorithm. By integrating a scale-adaptive sampling mechanism with a minimum sampling region constraint, our method fundamentally addresses the microscopic gradient oscillations induced by high curvature, achieving effective smoothing and rectification along the entire integration path.
- **Extensive Experiments:** We conduct extensive experiments on the ImageNet dataset against various CNNs, ViTs, and defensive models. The results demonstrate that BS-IG outperforms the latest IG-based SOTA methods (e.g. MuMoDIG [6]) by up to **5.9%** and other state-of-the-art attacks by **2.42%**. Furthermore, when combined with IG-based methods, the performance can be further boosted by up to **12.94%**.

II. Related Work

In this section, we review the literature relevant to our approach. First, we analyze the **theoretical foundations** of transferability, identifying the high-curvature bottleneck on the surrogate manifold as the primary obstacle. Next, we categorize existing **transfer-based attacks**, focusing on the evolution and limitations of Integrated Gradients (IG)-based methods. Finally, we discuss **gradient smoothing** techniques from interpretability and highlight the critical *Boundary Collapse* failure mode arising from the direct application of adaptive sampling.

A. Theoretical Foundations of Transferability

The mechanism of adversarial transferability remains a fundamental challenge in the black-box setting. Ilyas et al. [22] proposed that adversarial examples exploit non-robust features—highly predictive but brittle patterns derived from the data distribution. However, in cross-architecture scenarios, transferability is severely hindered by the fundamental disparity in image processing mechanisms. Mahmood et al. [16] revealed that CNNs prioritize local textures, whereas ViTs rely on global shape information. This difference in feature preferences leads to surrogate overfitting, creating a significant domain gap.

From a geometric perspective, the robustness of a model is intrinsic to the geometry of its output manifold. Moosavi-Dezfooli et al. [19] demonstrated that high-curvature regions in the loss landscape directly correlate with vulnerability to small perturbations. Furthermore,Dombrowski et al. [20] showed that in such rugged manifolds, the gradient directions are unstable and can be easily manipulated. These works collectively identify high-curvature regions on the surrogate manifold as the primary bottleneck for effective gradient estimation and transferability.

B. Transfer-based Attacks

Transfer-based attacks primarily improve attack performance through two main strategies: gradient optimization and input transformation. In the realm of gradient optimization, MI-FGSM [9] introduces a momentum term into FGSM [1] to stabilize the update direction. Building on this, VMI-FGSM [11] further mitigates gradient variance via a variance tuning mechanism, while the recent GI-FGSM [10] employs a global momentum initialization strategy to refine the optimization trajectory. Complementary to these approaches are input transformation methods, which aim to disrupt the specific feature dependencies of the surrogate model. Specifically, TIM [12] convolves gradients with a Gaussian kernel to simulate translation invariance; SIA [13] boosts sample diversity by splitting the image into blocks and applying various transformations to each block; and BSR [14] expands the input space by randomly shuffling and rotating image blocks.

Recently, methods based on Integrated Gradients (IG) [15] have been introduced to the adversarial domain to address the limitations of single-point estimation. TAIG [4] pioneered the application of IG in adversarial scenarios, leveraging its path attribution properties to generate perturbations. Subsequently, MIG [5] incorporated a momentum mechanism during the path accumulation process to stabilize updates, while MuMoDIG [6] adopted a multi-monotonic path strategy to improve the robustness of attack. However, as highlighted by Guided IG [21], the integration path inevitably traverses high-curvature regions of the model manifold, causing gradients to oscillate violently under noise interference.

C. Gradient Smoothing

In the realm of model interpretability, SmoothGrad [26] was proposed to eliminate irrelevant random noise. By averaging

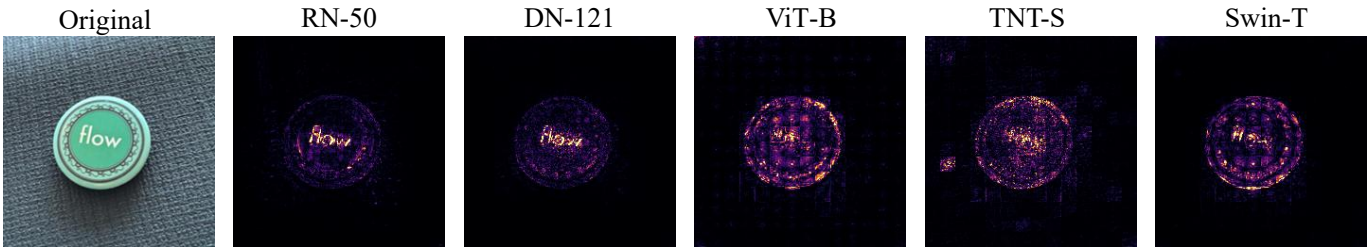


Fig. 2: **Visualization of gradient distributions across diverse architectures.** The figure displays the original image alongside Integrated Gradients (IG) maps computed on representative CNNs (RN-50, DN-121) and ViTs (ViT-B, TNT-S, Swin-T). The corresponding ground-truth class probabilities predicted by each model are 98.5%, 97.2%, 99.1%, 96.8%, and 98.9%, respectively. The distinct discrepancies in salient regions across different models, despite their high predictive confidence, empirically highlight the root cause of poor cross-architecture transferability.

gradients over Gaussian-perturbed inputs, it yields sharper saliency maps. Similarly, NoiseGrad [27] estimates the gradient distribution by injecting noise into model parameters, further enhancing the reliability and fidelity of interpretation results.

Inspired by these approaches, we attempt to introduce AdaptGrad [23], an adaptive variant from the interpretability domain, for gradient denoising in adversarial attacks. However, directly transferring this method to adversarial example generation encounters the *Boundary Collapse* problem. Specifically, to strictly satisfy distributional constraints, the sampling radius of AdaptGrad mathematically converges to zero near data boundaries. This phenomenon causes the attack to lose necessary exploration capability during the critical initialization phase, thereby rapidly falling into local optima.

III. Methodology

In this section, we first introduce the preliminaries and the motivation for employing adaptive gradient sampling, specifically analyzing the geometric cause of the *Boundary Collapse* phenomenon. Subsequently, we provide a detailed description of the proposed **BS-IG** method. Finally, we present the complete algorithmic framework to demonstrate how our method effectively prevents gradient vanishing caused by sampling collapse, thereby enhancing the transferability of adversarial examples.

A. Preliminaries

Adversarial Attack Problem. Given a clean image x with channel C, height H, width W, and its true label y, transfer-based attacks aim to optimize an adversarial perturbation δ bounded by ϵ on a surrogate model f. This process can be formulated as:

$$\delta^* = \underset{\|\delta\|_\infty \leq \epsilon}{\arg\max}\, L(f(x+\delta), y), \tag{1}$$

where $f(x)$ denotes the softmax output, and L is the loss function, typically the cross-entropy loss. $\|\cdot\|_\infty$ denotes the L_∞ norm constraint.

Integrated Gradients (IG). Integrated Gradients is an axiomatic interpretability method. A key property of IG is *Implementation Invariance*: if two networks are functionally equivalent (i.e., they produce the exact same output for every given input), their feature attributions are identical regardless of implementation differences. This implies that IG captures intrinsic model geometry rather than architecture-specific details, making it highly suitable for transfer attacks.

Formally, given a surrogate model f, the i-th element of IG for an input x relative to a baseline b (typically a black image) is defined as:

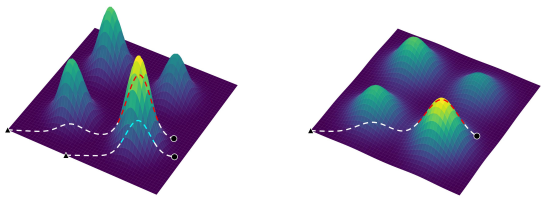
$$IG(x)_i = (x_i - b_i) \cdot \int_0^1 \frac{\partial f(b + \alpha \cdot (x-b))}{\partial x_i} d\alpha, \tag{2}$$

where the integral computes the accumulated gradients along the straight path from baseline b to input x. In practice, Equation 2 is approximated via a Riemann sum with m interpolation steps:

$$IG(x)_i \approx \frac{x_i - b_i}{m} \sum_{k=0}^{m-1} \frac{\partial f\left(b + \frac{k}{m} \cdot (x-b)\right)}{\partial x_i} \tag{3}$$

Recent methods leverage this integrated information to smooth out local gradient noise and identify robust critical regions for attack generation.

Connection between Interpretability and Adversarial Attacks. While originally designed to explain model decisions by attributing importance to specific input features, IG naturally aligns with the core objective of transfer-based adversarial attacks. Interpretability methods highlight the features that most significantly influence a model's prediction. By treating these attribution maps as perturbation guides, attackers can precisely target and manipulate the most critical predictive features. Furthermore, standard single-point gradient attacks often overfit to the local, architecture-specific artifacts of the surrogate model. By accumulating gradients along a continuous path, IG provides a stable, global estimation of feature importance that mitigates local gradient noise. Consequently, adversarial methods leverage this integrated information to bypass surrogate-specific high-curvature regions and identify robust, architecture-agnostic features, thereby significantly enhancing cross-model transferability.



(a)MuMoDIG (b)BS-IG

Fig. 3: **Schematic comparison of integration paths on the surrogate loss landscape.** (a) **MuMoDIG**: Although multiple paths (dashed lines) are employed to mitigate instability, they inevitably traverse microscopic high-curvature peaks, resulting in residual gradient oscillations. (b) **BS-IG**: The proposed method applies scale-adaptive sampling and minimum region constraints to effectively smooth the rugged landscape, ensuring a stable and rectified integration trajectory.

B. Motivation and Analysis

Necessity of Adaptive Sampling. We first investigate the root cause of limited transferability by visualizing the gradient distributions of IG across different architectures.As shown in Figure 2, significant discrepancies exist between the gradients of the surrogate and target models. From a geometric perspective, this inconsistency is attributed to the high-curvature regions on the surrogate's loss landscape, which introduce noise and mislead the optimization.

Existing methods like MuMoDIG attempt to mitigate this by averaging gradients over multiple paths. However, as simulated in Figure 3(a), while the multi-path strategy alleviates instability through superposition, the individual integration paths inevitably still traverse these high-curvature areas, leading to residual gradient oscillations.

Inspired by AdaptGrad [23], we introduce an adaptive sampling strategy to fundamentally smooth the rugged surface. Crucially, this method injects noise dependent on the input intensity, ensuring that the perturbed samples strictly remain within the valid range while effectively flattening the local curvature.

Analysis of Boundary Collapse. Although adaptive sampling effectively smooths the landscape, directly applying it to adversarial generation introduces a critical failure mode: *Boundary Collapse*.

To satisfy the validity constraint, the sampling radius is mathematically forced to shrink as the pixel value approaches the boundaries (0 or 1).

As illustrated in Figure 3(b), considering that the integration often starts (baseline) or ends (input) near these high-curvature boundary regions, the calculated sampling range becomes infinitesimal.

Consequently, due to this restricted exploration area during the critical initialization phase, the algorithm fails to sense the smoothed global geometry and is instantly trapped in the local optima of the surrogate model. This necessitates imposing a minimum lower bound on the adaptive sampling region to maintain effective exploration.

C. Proposed Method

The core objective of BS-IG is to rectify the microscopic gradient oscillations that arise when the integration path traverses high-curvature regions of the surrogate manifold. We achieve this through a dual-component strategy: Scale-adaptive Sampling (Sa) to handle local ruggedness and Boundary-aware Rectification (Ba) to ensure continuous exploration at the data limits.

*1) **Scale-adaptive Sampling Strategy**:* To suppress curvature noise without violating the functional domain of the input, we employ a scale-adaptive mechanism. To ensure the perturbed samples remain within the valid input domain $[0, 1]$ with a high probability (confidence level c), we bound the standard normal noise $\eta \sim \mathcal{N}(0, 1)$. For a standard normal distribution, the probability that a sample falls within the interval $[-K, K]$ is given by $erf(K/\sqrt{2}) = c$. By solving for K, we derive the scaling factor using the inverse error function: $K = \sqrt{2} \cdot erf^{-1}(c)$. For each pixel x_i, the base sampling radius $\sigma_{adapt}(x_i)$ is dynamically scaled according to its proximity to the data boundaries:

$$\sigma_{adapt}(x_i) = \frac{\min(|x_i|, |1 - x_i|)}{K} \tag{4}$$

This allows for maximal smoothing in central intensity regions where the local geometry is most permissive, while automatically constricting the sampling range near 0 and 1 to maintain input validity.

*2) **Boundary-aware Sampling Rectification**:* While the Sa strategy ensures validity, it introduces a critical failure mode known as Boundary Collapse. As $x_i \to 0$ or 1, the sampling radius σ_{adapt} mathematically vanishes. Since IG typically initiates from a black baseline ($x = 0$), the algorithm begins in an exploration blind spot. In this state, the sampling range becomes infinitesimal, stripping the estimator of its ability to sense the smoothed global geometry and trapping the optimization in surrogate-specific local optima.

To resolve this, we introduce a stochastic exploration floor σ_b. The final rectified sampling range $\sigma_{BS}(x_i)$ is defined as:

$$\sigma_{BS}(x_i) = \max\left(\sigma_{adapt}(x_i), \sigma_b\right), \tag{5}$$

where $\sigma_b > 0$ serves as a predefined minimum exploration threshold. This constraint strictly enforces a **minimum sampling range** determined by σ_b, ensuring consistent exploration capabilities even at the integration endpoints. Consequently, the gradient estimator retains sufficient stochasticity to escape high-curvature traps throughout the entire path, effectively eliminating exploration blind spots at the data boundaries.

*3) **BS-IG Transferable Attack Method**:* Based on the sampling range σ_{BS} constructed in the previous section, we formally present the BS-IG transferable attack method.

We model the gradient estimation as an expectation problem. For any interpolation point x_k along the integration path, we utilize $\sigma_{BS}(x_k)$ to construct a local Gaussian distribution and approximate the expected gradient via Monte Carlo sampling, thereby replacing the partial derivative term in Equation 3.

Specifically, we generate N noise samples $\eta \sim \mathcal{N}(0,1)$ from a standard normal distribution. The n-th perturbed sample $\tilde{x}_{k,n}$ is calculated as:

$$\tilde{x}_{k,n} = \text{Clip}_{[0,1]}\left(x_k + \sigma_{BS}(x_k) \odot \eta\right), \tag{6}$$

where $\odot$ denotes element-wise multiplication. The final BS-IG gradient estimator is derived by averaging the gradients of these perturbed samples. By substituting the partial derivative term in Equation 3 with the Monte Carlo estimation, we derive the final formulation of BS-IG:

$$BS\text{-}IG(x)_i \approx \frac{x_i - b_i}{m} \sum_{k=0}^{m-1} \left(\frac{1}{N} \sum_{n=1}^{N} \frac{\partial f(\tilde{x}_{k,n})}{\partial x_i} \right), \tag{7}$$

where $\frac{\partial f(\tilde{x}_{k,n})}{\partial x_i}$ denotes the gradient with respect to the n-th perturbed sample.

Compared to the original Equation 3, the proposed Equation 7 effectively suppresses gradient oscillations caused by non-robust features by incorporating boundary-aware local smoothing at each integration step.

D. Algorithm Framework

We integrate the proposed BS-IG estimator into an iterative attack framework, with the complete procedure summarized in **Algorithm 1**.

Algorithm 1: Framework of BS-IG Transferable Attack

Input : Surrogate model f, clean image x, true label y, baseline b.
Parameter: Budget ϵ, iterations T, step size α; BS params: steps m, samples N, confidence c, bound σ_b.
Output : Adversarial image x_{adv}.

Initialize $x_0 = x$ and scaling factor $K = \sqrt{2} \cdot \text{erf}^{-1}(c)$;
for $t = 0$ **to** $T - 1$ **do**
  `// BS-IG gradient estimation`
  $G_{acc} = 0$;
  for $k = 0$ **to** $m - 1$ **do**
    $x_k = b + \frac{k}{m}(x_t - b)$;
    `// Sampling range`
    $\sigma_{adapt} = \min(|x_k|, |1 - x_k|)/K$;
    $\sigma_{BS} = \max(\sigma_{adapt}, \sigma_b)$;
    `// Monte Carlo sampling`
    $g_{sum} = 0$;
    for $n = 1$ **to** N **do**
      Sample $\eta \sim \mathcal{N}(0,1)$;
      $\tilde{x} = \text{Clip}_{[0,1]}(x_k + \sigma_{BS} \odot \eta)$;
      $g_{sum} = g_{sum} + \nabla_x f(\tilde{x})$;
    $G_{acc} = G_{acc} + \frac{1}{N} g_{sum}$;
  $\bar{g}_t = \frac{x_t - b}{m} \odot G_{acc}$;
  `// Update adversarial image`
  $x_{t+1} = \text{Clip}_{x,\epsilon}\{x_t + \alpha \cdot \text{sign}(\bar{g}_t)\}$;
$x_{adv} = x_T$;
return x_{adv}

IV. Experiments

In this section, we empirically validate the effectiveness of the proposed BS-IG. First, we detail the experimental settings. Next, we demonstrate the superior cross-architecture transferability of BS-IG against state-of-the-art attacks and evaluate its compatibility with existing IG variants. Finally, ablation studies and sensitivity analyses verify the contributions of the scale-adaptive strategy and boundary-aware rectification, alongside the robustness of hyper-parameter selection.

A. Experimental Settings

Dataset and Models. Following previous works [1], [6], [13], [28], we conduct experiments on 1,000 images randomly selected from the ILSVRC 2012 validation set [29], ensuring all images are correctly classified by the surrogate models.

Surrogate models include three CNNs, *i.e.*, RN-50 [24], VGG-19 [30], and DN-121 [31], and three ViTs, *i.e.*, ViT-B [25], TNT-S [32], and Swin-T [33]. Target models include four CNNs, *i.e.*, RN-50, VGG-19, DN-121, and Inc-v3 [34], and four ViTs, *i.e.*, ViT-B, TNT-S, Swin-T, and DeiT-B [35], and two defense methods, *i.e.*, Bit Depth Reduction (Bit) [36] and JPEG Compression (JPEG) [37].

Baselines and Parameters. We compare BS-IG with four state-of-the-art transferable attacks, *i.e.*, MIG [5], MuMoDIG [6], SIA [13], and BSR [14]. The baselines adopt the default parameter settings reported in their corresponding papers. For all attacks, we consistently set the maximum perturbation $\epsilon = 16/255$, iteration number $T = 10$, and step size $\alpha = 1.6/255$. For our proposed BS-IG, we specifically set the sampling number $N = 20$, confidence level $c = 0.95$, and minimum boundary threshold $\sigma_b = 0.05$. In terms of computational cost, the average generation time per image for BS-IG is 2.13s, while the baselines require 0.21s (MIG), 1.95s (MuMoDIG), 0.26s (SIA), and 0.28s (BSR), respectively.

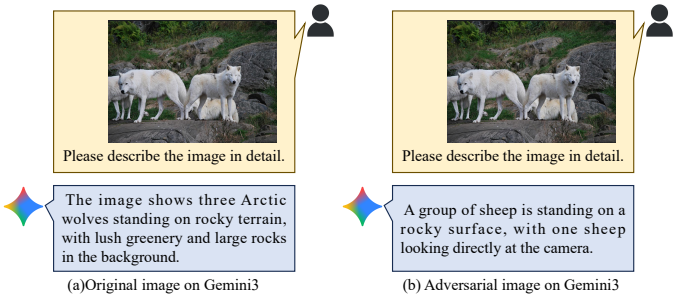
B. Comparison with Other Attacks

Performance against IG-based Baselines. We first evaluate BS-IG against representative IG variants, *i.e.*, MIG and MuMoDIG. The results in Table I demonstrate the distinct advantage of our method. Specifically, the MIG yields the lowest transferability due to overfitting. Although MuMoDIG improves performance by employing multi-monotonic paths to diversify the integration, it still falls short of BS-IG. For instance, with ResNet-50 as the surrogate, BS-IG achieves a MASR of **71.0%** on ViT targets, outperforming MuMoDIG (66.7%) by a clear margin. Overall, BS-IG outperforms these state-of-the-art IG-based methods by up to **5.9%**, indicating that smoothing microscopic noise induced by high curvature (BS-IG) is more effective than merely adding monotonic integration paths to bypass high curvature.

Effectiveness against State-of-the-art Transformation-based Attacks. We further compare BS-IG with advanced

TABLE I. Attack Success Rate (ASR %) and Mean Attack Success Rate (MASR %) of BS-IG and other attacks. The evaluation is conducted across six surrogate models. **MASR (CNNs)**, **MASR (ViTs)**, and **MASR (Defensed)** represent the mean ASR on CNN targets, ViT targets, and defense methods (Bit & JPEG), respectively, while **MASR** denotes the average ASR across all target models. The best results for each surrogate are highlighted in **bold**.

Surrogate	Attack	Target Model										MASR (CNNs)	MASR (ViTs)	MASR (Defensed)	MASR
		RN-50	VGG-19	DN-121	Inc-v3	ViT-B	TNT-S	Swin-T	DeiT-B	Bit	JPEG				
RN-50	SIA	99.9*	93.3	95.5	80.3	49.2	71.0	82.4	61.9	90.8	**94.5**	92.3	66.1	92.7	81.9
	BSR	99.9*	**94.6**	95.2	87.6	51.9	75.5	84.5	66.2	92.2	95.3	94.3	69.5	93.7	84.3
	MIG	**100.0***	66.9	70.3	57.1	27.0	41.4	52.3	33.5	62.9	69.1	73.6	38.6	66.0	58.1
	MuMoDIG	99.8*	91.2	93.4	86.3	50.6	70.6	80.9	64.5	88.5	92.3	92.7	66.7	90.4	81.8
	BS-IG	99.7*	94.5	**97.5**	**93.4**	**53.8**	**76.0**	**86.6**	**67.4**	**93.8**	93.9	**96.3**	**71.0**	**93.9**	**85.7**
VGG-19	SIA	**83.9**	**100.0***	90.1	80.0	35.6	59.7	71.2	46.5	66.5	73.3	88.5	53.3	79.7	72.6
	BSR	82.4	**100.0***	90.3	83.9	33.8	56.9	70.2	45.6	61.2	69.7	89.2	51.6	79.8	72.3
	MIG	61.3	**100.0***	78.5	74.5	25.7	43.6	56.2	32.8	44.7	50.5	78.6	39.6	68.9	61.0
	MuMoDIG	83.3	**100.0***	91.1	86.0	34.6	57.4	67.8	43.1	64.8	71.5	90.1	50.7	80.3	72.4
	BS-IG	83.6	**100.0***	**93.4**	**89.1**	**38.0**	**61.3**	**71.6**	**47.8**	**68.1**	**75.5**	**91.5**	**54.7**	**82.3**	**74.9**
DN-121	SIA	**94.9**	96.2	**100.0***	95.1	60.4	82.3	85.2	**70.9**	89.8	90.2	96.6	74.7	90.0	86.5
	BSR	94.2	95.2	**100.0***	94.6	55.0	75.3	84.3	68.9	88.5	93.0	96.0	70.9	90.8	84.9
	MIG	79.9	88.4	**100.0***	83.3	36.8	55.5	68.1	47.2	67.0	74.2	87.9	51.9	70.6	70.0
	MuMoDIG	92.9	95.8	**100.0***	93.8	52.2	72.0	82.5	64.9	85.9	89.8	95.6	67.9	87.9	83.0
	BS-IG	94.8	**97.0**	**100.0***	**96.0**	**61.1**	**83.2**	**86.9**	70.2	**93.0**	**96.5**	**97.0**	**75.4**	**94.8**	**87.9**
ViT-B	SIA	83.7	88.9	89.2	87.7	**100.0***	95.4	94.8	96.1	77.9	82.0	87.4	96.6	80.0	89.6
	BSR	82.4	89.0	89.7	84.7	**100.0***	94.7	91.7	93.6	81.2	83.2	86.4	95.0	82.2	89.0
	MIG	65.5	79.5	72.8	69.0	**100.0***	90.8	83.0	87.8	58.2	60.6	71.7	90.4	59.4	76.7
	MuMoDIG	78.6	87.8	85.8	82.0	**100.0***	92.3	89.5	92.3	77.5	79.2	83.6	93.5	78.4	86.5
	BS-IG	**85.7**	**90.2**	**91.7**	**88.3**	**100.0***	**98.8**	**96.8**	**99.2**	**85.5**	**87.8**	**89.0**	**98.7**	**86.7**	**92.4**
TNT-S	SIA	91.8	**90.2**	89.1	86.5	**90.3**	**100.0***	96.7	95.9	78.9	75.8	89.4	95.7	77.4	89.5
	BSR	**94.5**	88.9	90.2	84.3	86.9	**100.0***	**97.6**	93.9	77.6	78.4	89.5	94.6	78.0	**89.2**
	MIG	53.5	71.9	59.0	53.8	59.7	**100.0***	84.6	71.3	42.2	46.2	59.6	78.9	44.2	64.2
	MuMoDIG	86.8	88.4	84.4	87.4	86.5	**100.0***	93.3	94.6	72.0	73.3	86.8	93.6	72.7	86.7
	BS-IG	92.9	89.6	**91.2**	**91.9**	90.0	**100.0***	97.5	**96.5**	**82.7**	**80.1**	**91.4**	**96.0**	**81.4**	91.3
Swin-T	SIA	87.6	91.8	89.3	78.3	69.8	96.3	**100.0***	82.4	76.2	78.2	86.8	87.1	77.2	85.0
	BSR	92.8	93.5	92.2	87.4	76.8	94.1	**100.0***	90.8	84.1	82.8	91.5	90.4	83.5	90.5
	MIG	54.1	66.1	60.1	52.5	42.7	75.6	**100.0***	51.5	42.3	45.6	58.2	67.5	44.0	59.1
	MuMoDIG	89.9	90.7	88.9	88.2	76.9	95.4	**100.0***	88.4	81.0	82.3	89.4	90.2	81.7	88.1
	BS-IG	**94.5**	**94.0**	**93.7**	**89.8**	**83.3**	**98.2**	**100.0***	**93.3**	**84.7**	**83.3**	**93.0**	**93.7**	**84.0**	**92.5**



(a)Original image on Gemini3 (b) Adversarial image on Gemini3

Fig. 4: **Adversarial transferability to VLMs.** While the VLM correctly identifies the original image as *Arctic wolves* (a), the adversarial example generated by BS-IG successfully misleads it into hallucinating *a group of sheep* (b), verifying the robust transferability against commercial black-box systems.

transformation-based methods (SIA and BSR), which typically enhance transferability by augmenting input diversity. As shown in Table I, BS-IG consistently achieves superior performance. Taking ResNet-50 as the surrogate, our method obtains a total MASR of **85.7%**, surpassing BSR (84.3%) and SIA (81.9%). Notably, BS-IG outperforms these baselines without relying on complex input transformations, achieving an average improvement of **2.42%** over the latest transformation-based competitor (BSR). This suggests that refining the intrinsic gradient quality via scale-adaptive sampling serves as a highly effective approach for enhancing transferability, offering a distinct and powerful perspective alongside input augmentation.

TABLE II. **Compatibility Analysis.** Comparison of Mean Attack Success Rates (MASR %) when integrating BS-IG into state-of-the-art IG-based attacks (MIG and MuMoDIG). RN-50 and ViT-B serve as representative surrogate models. **Gain** indicates the absolute improvement in MASR achieved by our method. The best results are highlighted in **bold**.

Surrogate	Attack	CNNs		ViTs	
		MASR	Gain	MASR	Gain
RN-50	MIG	73.6	-	38.6	-
	MIG + BS-IG	**96.5**	**+22.9**	**71.8**	**+33.2**
	MuMoDIG	92.7	-	66.7	-
	MuMoDIG + BS-IG	**96.2**	**+3.5**	**72.3**	**+5.6**
ViT-B	MIG	71.7	-	90.4	-
	MIG + BS-IG	**89.5**	**+17.8**	**99.0**	**+8.6**
	MuMoDIG	83.6	-	93.5	-
	MuMoDIG + BS-IG	**90.2**	**+6.6**	**98.8**	**+5.3**

Transferability of Adversarial Examples on VLMs. To verify the practical robustness of BS-IG, we further evaluate it on a commercial Large Vision-Language Model (e.g., Gemini). As illustrated in Figure 4, the generated adversarial perturbations successfully mislead the target system into hallucinating semantically unrelated content, despite the model's accurate understanding of the original image.

C. Compatibility with Advanced IG Variants

Existing IG variants (MIG, MuMoDIG) prioritize macroscopic path planning, whereas BS-IG targets microscopic gradient rectification. These complementary granularities allow BS-IG to consistently boost macroscopic methods, as shown in Table II. Specifically, integrating BS-IG resolves MIG's high-frequency oscillations, increasing MASR (RN-50

surrogate) by **33.2%** on ViTs and **22.9%** on CNNs. Even for MuMoDIG, BS-IG yields consistent gains. Overall, the combination boosts performance by an average of **12.94%**, validating the robustness of uniting macroscopic planning with microscopic smoothing.

D. Ablation Study

TABLE III. **Ablation Study.** Investigation of the contribution of individual components: Scale-adaptive (Sa) strategy and Boundary-aware (Ba) rectification. RN-50 and ViT-B serve as representative surrogate models. **Gain** indicates the absolute improvement in MASR compared to the baseline (IG). The best results are highlighted in **bold**.

Surrogate	Method	Components		CNNs		ViTs	
		Sa	Ba	MASR	Gain	MASR	Gain
RN-50	Baseline (IG)	-	-	69.5	-	35.8	-
	+ Sa	✓	-	93.9	+24.4	62.8	+27.0
	+ Sa + Ba (Ours)	✓	✓	**96.3**	**+26.8**	**71.0**	**+35.2**
ViT-B	Baseline (IG)	-	-	67.4	-	83.6	-
	+ Sa	✓	-	81.2	+13.8	93.1	+9.5
	+ Sa + Ba (Ours)	✓	✓	**89.0**	**+21.6**	**98.7**	**+15.1**

To verify the effectiveness of the proposed components in BS-IG, *i.e.*, the Scale-adaptive (Sa) strategy and Boundary-aware (Ba) rectification, we conduct an ablation study on both ResNet-50 and ViT-B surrogates. The results are reported in Table III.

Effectiveness of Scale-adaptive (Sa) Strategy. The vanilla IG (Baseline) generates deterministic gradients that are highly susceptible to high-curvature noise. The proposed Sa strategy introduces dynamic noise injection defined by Equation 4, which adapts to input intensity while strictly maintaining data validity. As shown in Table III, this strategy yields the primary performance boost. For instance, with ResNet-50 as the surrogate, Sa alone improves the MASR on ViT targets by **27.0%** (from 35.8% to 62.8%) and on CNN targets by **24.4%**. These improvements demonstrate that adaptively smoothing the manifold based on local geometry is far more effective than the deterministic baseline.

Importance of Boundary-aware (Ba) Rectification. However, as analyzed in Section III-C, the strict validity constraint in Sa leads to *Boundary Collapse*, where the sampling range vanishes at the integration baseline ($x = 0$).The Ba component resolves this critical issue by enforcing a minimum sampling floor σ_b (Equation 5).Integrating Ba ensures sufficient exploration during the crucial initialization phase, providing consistent additional gains.Notably, for the ResNet-50 surrogate, adding Ba further boosts the MASR on ViT targets to **71.0%** (a total gain of **35.2%**).Likewise, for the ViT-B surrogate, the full BS-IG achieves a remarkable MASR of **89.0%** on CNN targets, surpassing the Sa-only version by **7.8%**.This confirms that preventing boundary collapse is essential for eliminating exploration blind spots and achieving optimal transferability.

E. Hyper-parameter Sensitivity

We evaluate the impact of the sampling number N and boundary threshold σ_b on ResNet-50 and ViT-B (Figure 5), fixing the confidence level $c = 0.95$ as derived in AdaptGrad [23].

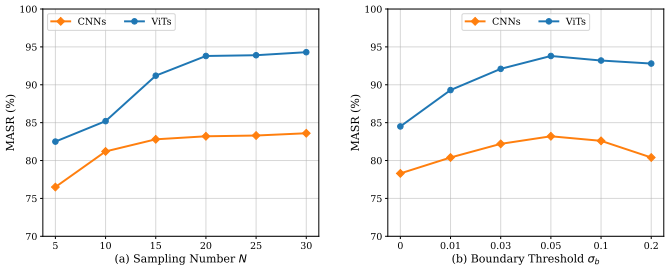


Fig. 5: **Hyper-parameter Sensitivity.** Ablation studies on the impact of (a) sampling number N and (b) boundary threshold σ_b. The Mean Attack Success Rate (MASR) is evaluated using CNNs (ResNet-50) and ViTs (ViT-B) as surrogate models to demonstrate cross-architecture robustness.

Impact of Sampling Number N. N determines gradient estimation granularity. As Figure 5(a) shows, MASR improves rapidly but saturates around $N = 20$. Further increasing N to 30 yields marginal gains ($\leq 0.5\%$) but increases computational cost by 50%. Thus, we set $N = 20$ to optimally balance effectiveness and efficiency.

Impact of Minimum Boundary Threshold σ_b. σ_b prevents gradient vanishing at data limits. Figure 5(b) reveals a climb-and-drop pattern. Without rectification ($\sigma_b = 0$), MASR drops significantly, empirically validating that boundary collapse hinders optimization. Performance peaks at $\sigma_b = 0.05$ (reaching 93.8% for ViT-B) and declines beyond 0.1 due to excessive deviation from the original manifold. We default to $\sigma_b = 0.05$.

V. Conclusion

To address the challenge of cross-architecture transferability, we identify the high-curvature bottleneck and the *Boundary Collapse* issue in adaptive sampling, proposing Boundary-aware Scale-adaptive Integrated Gradients (BS-IG) to overcome these limitations. By combining a scale-adaptive mechanism with a minimum sampling region constraint, BS-IG effectively rectifies microscopic gradient oscillations and ensures robust smoothing along the integration path. Extensive experiments show that our approach significantly outperforms state-of-the-art methods, achieving superior transferability across CNNs and ViTs, while exhibiting high compatibility with existing IG-based methods as a reliable robustness benchmark. Future work will explore BS-IG's generalization on lower-resolution datasets (e.g., CIFAR), where compressed feature spaces may require further calibration of the scale-adaptive parameters.

Acknowledgment

This work was sponsored by the National Natural Science Foundation of ChinaGeneral Program with grant number(No. 62272352). This research is supported in part by the Humanities and Social Sciences of Ministry of Education Planning Fund (No. 21YJAZH073).

REFERENCES

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[2] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*. PMLR, 2019, pp. 7472–7482.

[3] G. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein, "Adversarial examples that fool both computer vision and time-limited humans," *Advances in neural information processing systems*, vol. 31, 2018.

[4] Y. Huang and A. W.-K. Kong, "Transferable adversarial attack based on integrated gradients," in *International Conference on Learning Representations*.

[5] W. Ma, Y. Li, X. Jia, and W. Xu, "Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4630–4639.

[6] Y. Ren, Z. Zhao, C. Lin, B. Yang, L. Zhou, Z. Liu, and C. Shen, "Improving integrated gradient-based transferable adversarial examples by refining the integration path," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 7, 2025, pp. 6731–6739.

[7] M. Cheng, T. Le, P.-Y. Chen, H. Zhang, J. Yi, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," in *International Conference on Learning Representations*.

[8] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

[9] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[10] J. Wang, Z. Chen, K. Jiang, D. Yang, L. Hong, P. Guo, H. Guo, and W. Zhang, "Boosting the transferability of adversarial attacks with global momentum initialization," *Expert Systems with Applications*, vol. 255, p. 124757, 2024.

[11] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1924–1933.

[12] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2730–2739.

[13] X. Wang, Z. Zhang, and J. Zhang, "Structure invariant transformation for better adversarial transferability," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4607–4619.

[14] K. Wang, X. He, W. Wang, and X. Wang, "Boosting adversarial transferability by block shuffle and rotation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 24 336–24 346.

[15] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[16] K. Mahmood, R. Mahmood, and M. Van Dijk, "On the robustness of vision transformers to adversarial examples," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7838–7847.

[17] Z. Wei, J. Chen, M. Goldblum, Z. Wu, T. Goldstein, and Y.-G. Jiang, "Towards transferable adversarial attacks on vision transformers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2668–2676.

[18] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," in *International conference on learning representations*, 2018.

[19] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard, "Robustness via curvature regularization, and vice versa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9078–9086.

[20] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel, "Explanations can be manipulated and geometry is to blame," *Advances in neural information processing systems*, vol. 32, 2019.

[21] A. Kapishnikov, S. Venugopalan, B. Avci, B. Wedin, M. Terry, and T. Bolukbasi, "Guided integrated gradients: An adaptive path method for removing noise," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5050–5058.

[22] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Advances in neural information processing systems*, vol. 32, 2019.

[23] L. Zhou, C. Ma, Z. Wang, L. Wu, and X. Shi, "Adaptgrad: Adaptive sampling to reduce noise," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[26] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[27] K. Bykov, A. Hedström, S. Nakajima, and M. M.-C. Höhne, "Noisegrad—enhancing explanations by introducing stochasticity to model weights," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6132–6140.

[28] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the transferability of adversarial attacks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 158–16 167.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[32] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in neural information processing systems*, vol. 34, pp. 15 908–15 919, 2021.

[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[35] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[36] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[37] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *International Conference on Learning Representations (ICLR)*, 2018.