석 사 학 위 논 문
Master's Thesis

전이 가능한 적대적 공격을 위한 이미지 어텐션 공간상의 섭동 생성 방법

Diverse Generative Perturbations on Attention Space for Transferable Adversarial Attacks

2023

김 우 재 (金 佑 載 Kim, Woo Jae)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

석 사 학 위 논 문

전이 가능한 적대적 공격을 위한 이미지 어텐션 공간상의 섭동 생성 방법

2023

김 우 재

한 국 과 학 기 술 원

전산학부

전이 가능한 적대적 공격을 위한 이미지 어텐션 공간상의 섭동 생성 방법

김 우 재

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2022년 11월 15일

심사위원장 윤 성 의 (인)

심 사 위 원 홍 승 훈 (인)

심 사 위 원 안 성 진 (인)

Diverse Generative Perturbations on Attention Space for Transferable Adversarial Attacks

Woo Jae Kim

Advisor: Sung-Eui Yoon

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in Computer Science

Daejeon, Korea
December 13, 2022

Approved by

Sung-Eui Yoon
Professor of School of Computing

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MCS 김우재. 전이 가능한 적대적 공격을 위한 이미지 어텐션 공간상의 섭동 생성 방법. 전산학부 . 2023년. 20+iii 쪽. 지도교수: 윤성의. (영문 논문)
Woo Jae Kim. Diverse Generative Perturbations on Attention Space for Transferable Adversarial Attacks. School of Computing . 2023. 20+iii pages. Advisor: Sung-Eui Yoon. (Text in English)

초 록

적대적 공격을 하고자 하는 타겟 모델의 정보가 주어지지 않은 상황에서도 정보가 알려져 있는 대리 모델에 적대적 이미지를 생성해 타겟 모델을 속이는 전이 가능한 적대적 공격은 그 실용성 덕분에 각광을 받아 왔다. 하지만 적대적 공격의 전이율을 높이는 기존의 기법은 결정론적방법으로 적대적 이미지를 생성한다는 단점을 가진다. 그렇게 생성된 적대적 이미지는 대리 모델의 좋지 않은 로컬 옵티멈에 빠져 과적합되고, 이는 전이율 손실을 일으킨다. 이 문제를 해결하기 위해 본 논문에선 이미지의 현저한 특징점을 다양하게 교란시키는 Attentive-Diversity Attack (ADA)을 제안한다. 다른 구조와 파라미터 값을 가진 모델이 공통적으로 학습하는 특징점을 교란시키기 위해 이미지 어텐션을 교란시킨다. 그리고 이러한 특징점을 다양한 방법으로 교란시킴으로써 더 다양한 전이 가능한 적대적 섭동을 학습하고, 좋지 않은 로컬 옵티멈에 빠지는 것을 방지한다. 이를 공격 생성기 모델을 사용하여 학습시키고, 잠재 코드를 이용하여 공격 생성기가 생성하는 적대적 섭동을 제어한다. 다양한 실험을 통해 기존 방법과 비교하여 본 방법의 높은 전이율을 확인하였다.

핵 심 낱 말 딥러닝, 컴퓨터 비전, 적대적 공격

Abstract

Improving the adversarial attack transferability, or the ability of an adversarial example crafted on a known model to also fool unknown models, has recently received much attention due to their practicality in real-world scenarios. However, existing methods that try to improve such attack transferability craft perturbations in a deterministic manner. Thus, adversarial examples crafted in this manner often fail to fully explore the loss surface and fall into a poor local optimum, suffering from low transferability. To solve this problem, we propose Attentive-Diversity Attack (ADA), which disrupts diverse salient features in a stochastic manner to improve transferability. We first disrupt the image attention to perturb universal features shared by different models. We also disturb these features in a stochastic manner to explore the search space of transferable perturbations more exhaustively and thus to avoid poor local optima. To this end, we use a generator to produce adversarial perturbations that each disturbs features in different ways depending on an input latent code. Extensive experimental evaluations demonstrate the effectiveness of our method, outperforming the transferability of state-of-the-art methods.

Keywords Deep Learning, Computer Vision, Adversarial Attack

Contents

List of Tables

List of Figures

Chapter 1. Introduction

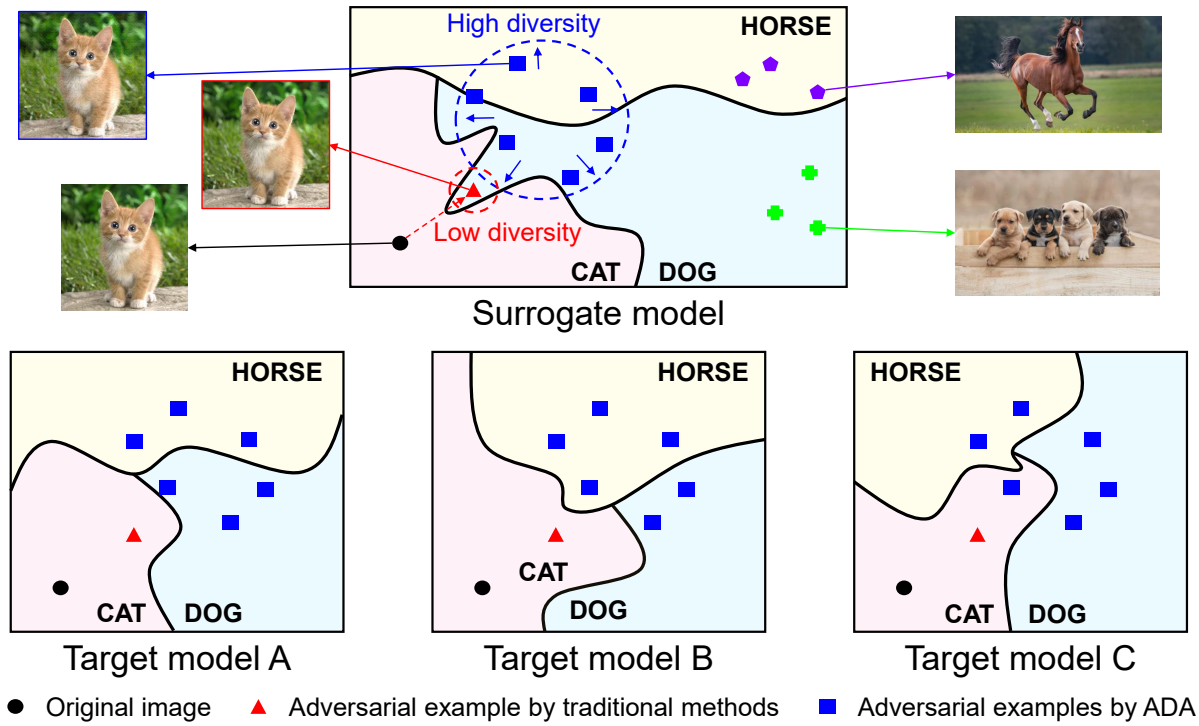


Figure 1.1: Conceptual illustration of class decision boundaries of a surrogate model and target models along with the adversarial examples crafted by traditional methods and our method (ADA). Adversarial examples by traditional methods are crafted in a deterministic manner and thus easily fall into a poor local optimum, overfitting to the surrogate model. In contrast, our method generates diverse perturbations and avoids such local optimum by exploring the search space of adversarial examples more exhaustively.

While deep neural networks (DNNs) have achieved impressive performance on numerous computer vision tasks [1, 2, 3], they have been known to be vulnerable against adversarial examples [4, 5]. Performed by adding a maliciously designed perturbation to the image, such adversarial attack is categorized as white-box or black-box depending on the knowledge of the model accessible to the attacker, such as its weights or structures. Recent works have focused on more challenging black-box attacks due to their practicality in real-world scenarios. Query-based attacks [6, 7, 8], which is a branch of black-box attacks, exploit the query outputs of an unknown model to estimate the gradients of an unknown model, but the excessive number of required queries limits their practicality. Instead, more attention has been given to transfer-based attacks that rely on the attack transferability, which is the ability of an adversarial example crafted on a white-box surrogate model to fool black-box target models.

However, traditional gradient-based white-box attacks (*e.g.*, BIM [9], PGD [10], *etc.*) often suffer from poor transferability because they easily overfit to the surrogate model. To this end, some have proposed more advanced optimization algorithms for these gradient-based methods to reduce overfitting. Dong *et al.* [11] applied a momentum strategy, Xie *et al.* [12] applied random transformations to the

image, Dong *et al.* [13] proposed a translation-invariant attack, and Wang *et al.* [14] applied variance tuning for more stable momentum. Based on findings that different models learn similar feature representations for same images, another branch of works has focused on disrupting the intermediate feature representations of image; Zhou *et al.* [15] maximized the distance between the features of the original image and the adversarial image. However, classifiers tend to also learn model-specific features [16], and naïvely disrupting these features can overfit the attack to the surrogate model. More recent works perturbed salient features; Wu *et al.* [17] disrupted the attention heatmaps, and Wang *et al.* [18] proposed aggregated gradients to perturb object-aware features.

Nevertheless, these methods still rely on a gradient-based method [9] that generates perturbations in a deterministic manner. They iteratively update a perturbation in a single, specific direction that maximizes the given objective function, and without sufficient stochasticity in this process, they often fail to fully explore the entire loss surface of the model. Thus, as shown in Fig. 1.1, with low diversity, adversarial examples crafted by these attacks can easily fall into a poor local optimum and *overfit* to the surrogate model, suffering from low transferability.

To solve this problem, we propose Attentive-Diversity Attack (ADA), which improves the transferability of adversarial examples by disrupting salient features in a diverse manner. Primarily, we step away from the gradient-based method and use a generator to craft adversarial perturbations for a given image. Then, based on recent findings [17, 18] that specifically disrupting the salient features of the image boosts transferability, we perturb the image attention representation, which highlights features that are responsible for model decision and thus are likely to be shared across different models. Then, to prevent the generator from corrupting the attention in a deterministic manner and thus falling into a poor local optimum, we guide the generator to disrupt these features in a diverse and stochastic manner. More specifically, we pass a latent code as an input to the generator and guide it to craft diverse perturbations for different latent codes. In that way, as shown in Fig. 1.1, the generator explores the search space of transferable adversarial examples more exhaustively and can learn to craft diverse perturbations that are located outside the poor local optimum. These adversarial examples effectively fool the target models, while those crafted by existing deterministic methods become overfitted to the surrogate model.

In summary, our contributions are as follow:

- For the first time, we introduce stochasticity to the generation of adversarial examples in the feature level to improve their transferability.
- We propose Attentive-Diversity Attack (ADA), an effective generator-based adversarial attack framework that perturbs image attention in a diverse, non-deterministic manner.
- Extensive experiments exhibit the superior transferability of our method as compared to existing state-of-the-art methods.

Chapter 2. Related Works

This section discusses the recent trend in both white-box and black-box adversarial attacks.

2.1 White-box Attacks

White-box attacks exploit the information of a model, such as its weights and structures, to generate adversarial images. One popular branch of white-box attacks rely on the gradient-based method [4, 9, 10], which directly exploits the loss gradient to generate adversarial perturbations. These attacks, however, are known to generate deterministic adversarial examples [19]. Another branch of works has focused on generator-based attacks [20, 21, 22, 23, 24], which employ a generator to produce adversarial perturbations. While these white-box attacks have shown to be transferable to some extent [25, 26], such ability remains very limited, leading to boosted interests on more practical black-box attacks.

2.2 Black-box Attacks

Unlike white-box attacks, black-box attacks fool models that are kept hidden from the attacker [27, 28, 11, 13, 29, 30, 31, 32, 14, 17, 33, 12, 15].

Query-based method is a branch of black-box attacks that relies on the query results available from black-box models to estimate the gradients. Papernot *et al.*used query information and knowledge distillation to build a white-box approximation of the black-box model [30]. Brendel *et al.* [27] and Chen *et al.* [28] used query access to estimate the gradients of the black-box model. AutoZOOM increased the query search efficiency of ZOO [28] by employing an autoencoder [31]. Ilyas *et al.* [29] proposed a method of fooling black-box models with limited number of queries. Nevertheless, query-based methods still require an excessive number of queries and trial-and-errors to craft a successful attack, limiting their practicability in real-world usages [13].

In contrast, transfer-based method aims to boost the transferability of white-box attacks to fool unknown target models. Since this method does not require any query access, the same attack can be used to fool multiple target models and does not require extensive computational costs. Thus, they have shown to be much more practical than query-based attacks [13]. Numerous works on transfer-based black-box attacks have focused on improving the optimization strategy of existing gradient-based white-box attacks. In order to prevent the attacks from falling in to poor local maxima, MI-FGSM added a momentum term to the loss gradient [11], DI^2-FGSM applied random transformations to inputs [12], TI-FGSM applied a predefined kernel to the loss gradient [13], and VMI-FGSM applied variance tuning on gradients [14]. Noting that transferable attacks are also often robust against image transformations, ATTA employed adversarial transformation networks [33]. There also have been attempts to modify the objective functions of the attack. TAP perturbed image features in the intermediate layers of the white-box model [15], ATA disrupted images on their attention representation [17], and FIA used aggregated gradients to perturb object-aware features [18]. Nevertheless, all these attempts still rely on the gradient-based method that leads to simple, deterministic attacks that easily overfit to surrogate models.

In this paper, we focus on boosting the transferability of transfer-based black-box attacks. Existing gradient-based attacks often overfit to vulnerable features that are unique to the surrogate model.

Also, since they greedily exploit the loss gradient, they generate deterministic attacks which may easily make the adversarial examples fall into poor local optima of the surrogate models. We solve these problems by first adopting a generator-based attack and then guiding the generator to produce diverse, non-deterministic adversarial examples. Additionally, we perturb the image attention to boost the transferability even further.

Chapter 3. Preliminaries

Let f_ψ be a target classifier. The objective of an untargeted adversarial attack is to create an adversary x^{adv} of an image x in class t such that it leads to a misclassification on the target classifier (*i.e.*, $f_\psi(x^{adv}) \neq t$). In this paper, we consider a black-box attack where we do not have access to the target classifier. Instead, we employ an accessible surrogate model h_θ that shares the same output space with f_ψ but has different architectures and/or parameters. We then generate a transferable adversarial example on the surrogate model as follows:

$$\arg\max_{x^{adv}} L_\theta(x^{adv}, t), \quad \text{s.t.} \quad \|x - x^{adv}\|_\infty \leq \epsilon, \tag{3.1}$$

where $L_\theta(\cdot,\cdot)$ is the classification loss on the surrogate model h_θ, and ϵ is a constraint set on the magnitude of perturbation.

The success of the black-box attack highly depends on the transferability assumption; an effective adversary on one network can transfer to another. However, it has been observed that many existing black-box attacks suffer from a limited transferability because they tend to *overfit* to the surrogate classifier. Also, even though an ideal untargeted adversary should be able to populate a diverse set of misclassification results (*i.e.*, $f_\psi(x^{adv}) \in \mathcal{T} \setminus \{t\}$), existing methods tend to produce a deterministic one. This is because they are mostly based on gradient-based optimization, which prefers a solution that maximizes the classification loss (Eq. (3.1)). We argue and empirically demonstrate that such deterministic property can also lower the chance of black-box attack being successful.

To address these challenges, we propose Attentive-Diversity Attack (ADA), an untargeted black-box attack method with high transferability. We adopt *attention perturbation* to disrupt images on highly transferable attention space. We also apply the *feature diversification* to encourage our method to produce stochastic perturbations, each of which leads to different misclassification label by exploring various transferable features and increrases the chance of improving transferability. We optimize these objectives using an *attack generator* that has higher expressive power than previous gradient-based methods.

Chapter 4. Attentive-Diversity Attack

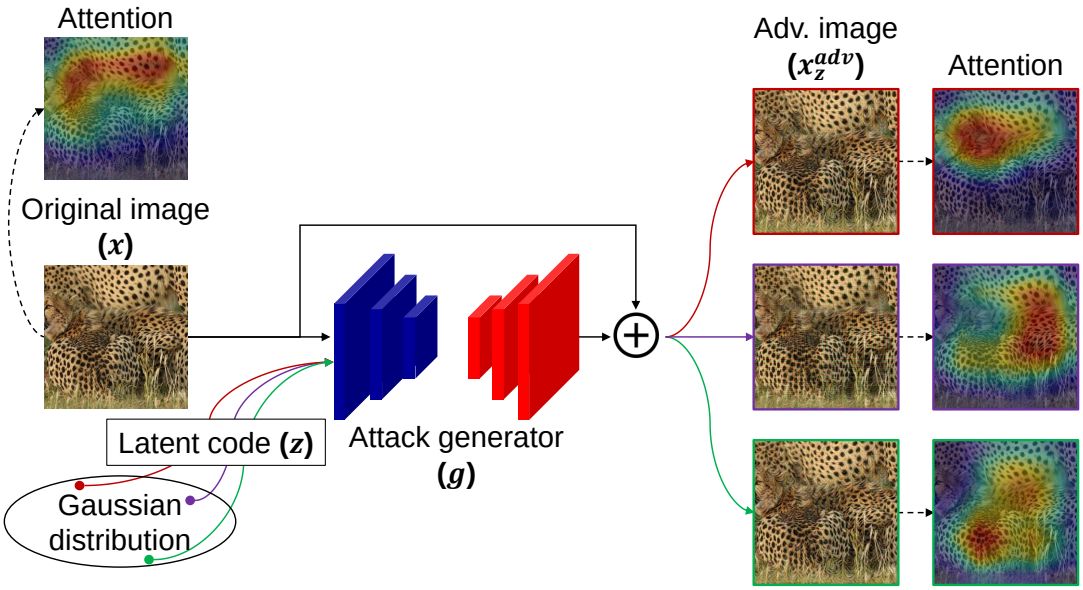


Figure 4.1: Overview of Attentive-Diversity Attack (ADA). Given an image and a latent code, the attack generator produces a perturbation that disrupts the image attention in a diverse manner.

In the following subsections, we elaborate on the main components of our Attentive-Diversity Attack (ADA) (Fig. 4.1). First, we explain the rationale behind using attack generator (Section 4.1). Then, we analyze how attention guidance can boost attack transferability and our approach to perturbing attention (Section 4.2). Lastly, we elaborate on the role of diversity regularization, especially on boosting transferability (Section 4.3).

4.1 Attack Generator

In this work, we step away from a widely-used gradient-based method that crafts perturbations in an iterative and deterministic manner and instead use a generator to parameterize the adversary with a DNN. As shown in Fig. 4.1, given an image x and a latent code z sampled from a Gaussian distribution, the generator g learns to output an adversarial perturbation that is dependent on the latent code. We then form an adversarial image x_z^{adv} as follows:

$$x_z^{adv} = Clip_{x,\epsilon}\{x + \epsilon \cdot g(x, z)\}, \tag{4.1}$$

where $Clip_{x,\epsilon}$ [9] clips the perturbation in a per-pixel manner so that it is bounded to ϵ-ball of L_∞ norm. How we exploit the latent code z to generate diverse adversarial perturbations will be discussed in Sec. 4.3.

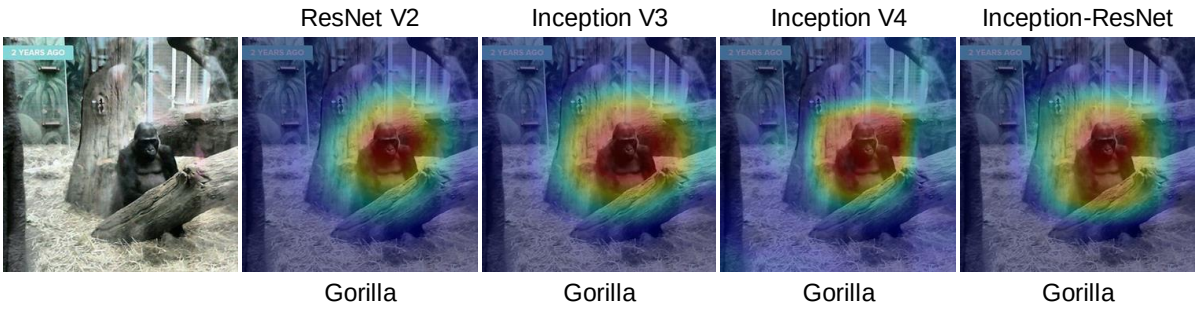


Figure 4.2: Attention representations of an image across different models. While these models have different weights and structures, they all focus on similar regions of the image for correct inference.

4.2 Attention Perturbation.

In order to boost transferability, we disrupt the image attention, which highlights features that are responsible for model decision and are likely to be relevant to the main objects of the image [17, 18]. Fig. 4.2 visualizes the attention representations of four different models given an image. As shown in the figure, such salient features are also likely to be shared across different models. As different classifiers similarly rely on these object-related features to make decisions, perturbations on these feature on the surrogate model will effectively transfer to other models.

Based on Grad-CAM [34], we define attention A as a weighted representation of features F, which we set as the output from the last convolutional layer (*e.g.*, *Mixed_7c* for Inc-v3), as follows:

$$A'(x;t) = \alpha_t F = \texttt{GAP}\left(\frac{\partial y_t}{\partial F}\right) F. \tag{4.2}$$

The weight α_t denotes the importance of the feature F given the ground truth class t. It is obtained by taking the gradient of y_t – the prediction for class t – with respect to F and applying global average pooling ($\texttt{GAP}(\cdot)$) over the spatial dimension. To prevent the generator from perturbing only the few channels with the highest magnitudes, we further apply channel-wise normalization on $A'(x;t)$ as follows:

$$A = \frac{\alpha_t A'}{\|\alpha_t A'\|_2}. \tag{4.3}$$

The generator learns to maximize the distance between the attention representations of the original image and the adversarial image by maximizing the following *attention loss* L_{attn}:

$$L_{attn} = \|A(x_z^{adv};t) - A(x;t)\|_2. \tag{4.4}$$

ATA [17] has also similarly disrupted the attention heatmaps extracted using the techniques of Grad-CAM [34]. Our method differs from their approach on that we additionally apply channel normalization. Without channel normalization, the generator perturbs only the few feature channels with highest magnitudes and reduces the diversity of perturbations it can generate. To prevent this, unlike ATA, we normalize each feature channel and enable the generator to disrupt more diverse features.

4.3 Feature Diversification.

Without any guidance, the generator may still greedily maximize the attention loss in a deterministic manner just like gradient-based methods by learning to generate a same perturbation for different input

latent codes. Thus, we explicitly guide the generator to explore and corrupt diverse features in a stochastic manner. We train it to disturb the attention representations differently for two distinct input latent codes z_1 and z_2, each sampled from a Gaussian distribution, by applying a diversity regularization [35] and maximizing the following *diversity loss* L_{div}:

$$L_{div} = \frac{\|A(x^{adv}_{z_1}; t) - A(x^{adv}_{z_2}; t)\|}{\|z_1 - z_2\|}. \tag{4.5}$$

We craft two adversarial examples $x^{adv}_{z_1}$ and $x^{adv}_{z_2}$ each by passing z_1 and z_2, respectively, into the generator (Eq. 4.1) and obtain their respective attention representations $A(x^{adv}_{z_1}; t)$ and $A(x^{adv}_{z_2}; t)$ (Eq. 4.2, 4.3). Then, by maximizing the distance between the two representations, we force the generator to craft semantically diverse perturbations.

While Yang *et al.* [35] originally proposed the diversity regularization, their applications have been limited to pixel or feature levels. Diversity on these levels, however, may not necessarily translate to diversity on the attention space and may fail to guide our generator to disrupt the *salient* features in a diverse manner. To explicitly guide it to perturb the meaningful features in a diverse manner, unlike existing approaches, we apply the diversity regularization on the attention level.

Overall, we learn the attack generator g to *maximize*:

$$L = L_{cls} + \lambda_{attn} \cdot L_{attn} + \lambda_{div} \cdot L_{div}, \tag{4.6}$$

where L_{cls} is the cross-entropy loss between the adversarial image and the ground truth label, and λ_{attn} and λ_{div} control the weights of the attention loss (L_{attn}) and the diversity loss (L_{div}), respectively.

There have been several attempts to craft diverse adversarial examples. Jang *et al.* [19] and Dong *et al.* [36] modeled diverse perturbations from a single image, but their approaches are limited to pixel-level diversity and improving adversarial robustness. Xie *et al.* [12] boosted transferability by crafting perturbations on randomly transformed images, but their approach can only implicitly perturb features in a diverse manner as a result of pixel-level transformations. In contrast, for the first time, we craft semantically diverse perturbations by explicitly disrupting diverse features. As a result, we effectively avoid poor local optimum, improving transferability as also shown by the experiment results (Table 5.1).

Chapter 5. Experiments

In this section, we report the experiment results on our proposed Attentive-Diversity Attack (ADA). We first report the experimental setups (Sec. 5.1). Then, we make comparisons on the transferability of our method with that of existing state-of-the-art transfer-based black-box attacks (Sec. 5.2) and verify that our method indeed generates diverse adversarial perturbations (Sec. 5.3). Lastly, we perform ablation studies and hyperparameter analysis to verify the individual components of our method (Sec. 5.4).

5.1 Experimental Setups

We implement the attack generator by using a U-Net [23, 37] based convolutional encoder-decoder consisting of three encoding blocks and three decoding blocks. Each encoding and decoding block consists of a convolutional layer and a transposed convolutional layer, respectively, followed by a batch normalization layer and a ReLU layer. At each encoding block, the latent code z is spatially expanded and concatenated to the input of the block. The generator is trained for 100 epochs with learning rate of 1e-4 and the batch size of 8 using an Adam optimizer [38] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and weight decay 1e-5.

We use 10,000 images randomly selected from the ImageNet validation set [39] for train data and 1,000 images from the NeurIPS 2017 adversarial competition [40] for test data. We test our method on Inception-v3 (Inc-v3) [2], Inception-v4 (Inc-v4) [41], Inception-ResNet-V2 (IncRes-v2) [41], ResNet-V2 (Res-v2) [42], and VGG16 (Vgg-16) [43][1]. We compare our method with various state-of-the-art attacks – MI-FGSM [11], DIM [12], VMI-FGSM [14], TAP [15], and FIA [18] – for which we set the number of iterations $T = 10$, the step size $\alpha = 1.6$, and the rest of the hyperparameters as specified in their respective references. The maximum perturbation constraint ϵ is set to 16 under L_∞ norm. For our method, we use $\lambda_{attn} = 10$, $\lambda_{div} = 1000$, and 16 for the length of latent code z.

5.2 Comparison of Transferability

We first compare the attack transferability of our method and five existing transfer-based black-box attacks, which are all based on a gradient-based method. We craft adversarial examples on four surrogate models – Inc-v3, Inc-v4, IncRes-v2, and Res-v2 – shown on the second column and measure their attack success rates (ASR), or the misclassification rate of a model [14], on five target models shown on the second row (Table 5.1). Numbers marked in parantheses () represent the white-box setting; the surrogate model is the same as the target model. We also report the ASR based on the ensemble logit of all black-box target models, which are all different from the surrogate model, on the column **Ensemble**. The results indicate that our method fools black-box target models with higher ASR than existing methods in most cases, outperforming FIA by average of 6.4% on the ensemble model. Also, while our method generally shows lower ASR on the white-box target models, it shows much higher ASR on black-box target models, showing that it overfits less to the surrogate model and generalizes well to unknown models.

[1] These models are publicly available at: https://github.com/Cadene/pretrained-models.pytorch

			Target models						
		Attack	Inception V3	Inception V4	Inception-ResNet V2	ResNet V2	VGG16	Ensemble	Rank
Surrogate models	Inception V3	MI-FGSM	(97.9)	42.9	39.9	41.2	53.1	35.7	6
		DIM	(98.0)	68.3	61.9	53.1	68.6	58.2	5
		VMI-FGSM	(97.9)	69.6	66.7	57.6	70.0	61.8	4
		TAP	(**100.0**)	77.9	75.3	53.1	70.6	69.1	3
		FIA	(98.5)	84.2	80.1	69.3	85.6	77.6	2
		Ours	(96.1)	**88.9**	**82.9**	**82.4**	**95.2**	**85.3**	**1**
	Inception V4	MI-FGSM	59.4	(98.9)	44.9	47.8	63.6	44.9	6
		DIM	75.5	(97.9)	66.5	60.5	74.9	66.5	5
		VMI-FGSM	76.6	(98.5)	70.0	65.1	76.8	68.9	4
		TAP	75.6	(**100.0**)	70.2	59.7	82.6	72.2	3
		FIA	83.3	(95.0)	**78.5**	74.6	85.2	78.3	2
		Ours	**85.2**	(97.7)	67.6	**79.0**	**89.5**	**79.5**	**1**
	Inception-ResNet V2	MI-FGSM	58.0	52.6	(**99.4**)	46.9	63.5	47.0	6
		DIM	73.0	70.6	(94.8)	57.7	70.4	65.6	4
		VMI-FGSM	78.4	77.4	(99.3)	64.5	74.2	71.6	3
		TAP	74.1	66.8	(95.2)	46.7	63.0	51.8	5
		FIA	81.6	77.1	(88.6)	66.9	81.9	74.8	2
		Ours	**82.5**	**89.4**	(93.0)	**80.7**	**89.4**	**85.9**	**1**
	ResNet V2	MI-FGSM	54.7	47.9	44.1	(99.6)	61.4	44.4	6
		DIM	75.3	70.6	68.8	(99.1)	73.9	70.0	3
		VMI-FGSM	72.7	67.4	64.7	(97.6)	72.3	65.3	4
		TAP	51.8	44.3	44.5	(92.4)	68.2	52.6	5
		FIA	**81.2**	76.7	**74.5**	(**99.9**)	82.9	74.1	2
		Ours	79.7	**90.9**	71.9	(94.0)	**93.1**	**80.6**	**1**

Table 5.1: Attack success rates (%) of different attacks against various target models. The leftmost column and the uppermost row show surrogate models and target models, respectively. Parentheses () indicates white-box attack where the target model is the surrogate model. Best results are highlighted in bold, and "Rank" denotes the order of highest average ASR on black-box models.

Interestingly, our method outperforms existing methods by a significantly larger margin when the architectures of the target models are more different from that of the surrogate model. For example, while our attack crafted on Inc-v3 outperforms FIA by 4.7% on a more similarly structured Inc-v4 [44], it outperforms FIA by a larger margin of 9.6% on a more differently structured Vgg-16 [44]. While existing methods overfit to the surrogate model and show low transferability on models with more distinct structures, our method shows high transferability regardless of the model structure.

5.3 Verification of Attack Diversity

Application to Adversarial Training. We now show that our method indeed generates diverse adversarial perturbations to escape poor local optimum of the surrogate model. To do so, we first evaluate the robustness of adversarial training models trained with different adversarial attacks including our method. Adversarial training [4, 10], which learns a model to be robust against adversarial attacks by training it with adversarially generated data, has been widely considered as one of the most effective defense strategies. It solves the following minimax optimization problem:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{x^{adv}} L(h_\theta(x^{adv}), t) \right], \tag{5.1}$$

where h is a model with parameter θ, x^{adv} is an adversarial example crafted from a natural image with label t from dataset $\mathcal{D}$, and $L(\cdot,\cdot)$ represents the classification loss. Recent findings [19, 36] have shown that training an adversarial training model with a diverse set of adversarial examples improves its

	No Attack	FGSM	BIM	PGD	Ours
No defense†	91.85	35.51	2.89	2.59	4.45
FGSM	79.24	82.25	81.02	83.99	83.77
BIM	83.17	82.80	83.17	83.10	82.28
PGD	85.84	84.73	85.62	85.54	85.10
Ours (w/o L_{div})	89.18	80.80	85.62	86.66	86.95
Ours (w/ L_{div})	**90.88**	**85.92**	**88.21**	**89.25**	**89.62**

Table 5.2: Classification accuracy (%) of adversarial training models under various attacks. Leftmost column and uppermost row represent the attacks used for training and the attacks used for evaluation, respectively. Adversarial examples used for evaluation are crafted on a classifier trained with original images (marked †). Best results are highlighted in bold.

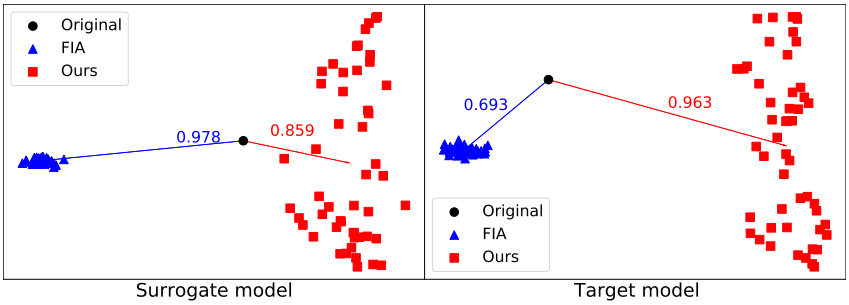


Figure 5.1: PCA visualization on surrogate model (Inc-v3) and target model (Res-v2) for features of adversarial examples crafted by FIA (blue) and our method (red).

robustness. Based on this idea, we test the robustness of a model trained against adversarial examples crafted by ADA to show that our method indeed generates diverse perturbations. We train Inc-v3 for 30 epochs using batch size of 8, an SGD optimizer with learning rate of 0.001, momentum of 0.9, and weight decay of 5e-4 on Caltech101 dataset [45] with 8,681 images and 101 classes randomly split into 7,332/1,349 for training/test set.

We report the classification accuracy of each adversarial training model trained by a different attack in Table 5.2. The model trained by our method with L_{div} records the highest robustness in all cases, outperforming that trained without L_{div} by 3.24% in average. Our method without L_{div} and gradient-based methods (FGSM [4], BIM [9], and PGD [10]) train the classifiers only against a limited set of deterministic adversarial examples. In contrast, our method with L_{div} trains the model against diverse adversarial examples and makes it more robust.

Effects of Diversity on Transferability. To show that disrupting diverse features indeed improves transferability, we visualize in Fig. 5.1 the features of adversarial examples crafted by FIA and our method by projecting them on the 2D space spanned by eigenvectors obtained from PCA. As shown in the figure, feature representations of FIA adversarial examples form a dense cluster, while those of our method are widely spread out. Also, FIA generates stronger adversarial examples compared to our method on the surrogate model (FIA); the average ℓ_2 distance from the feature of the original image to the features of adversarial examples crafted by FIA is 0.978, which is higher than that of our method,

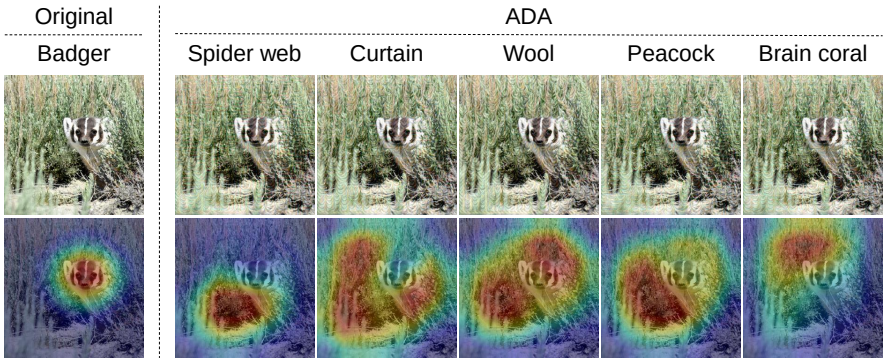
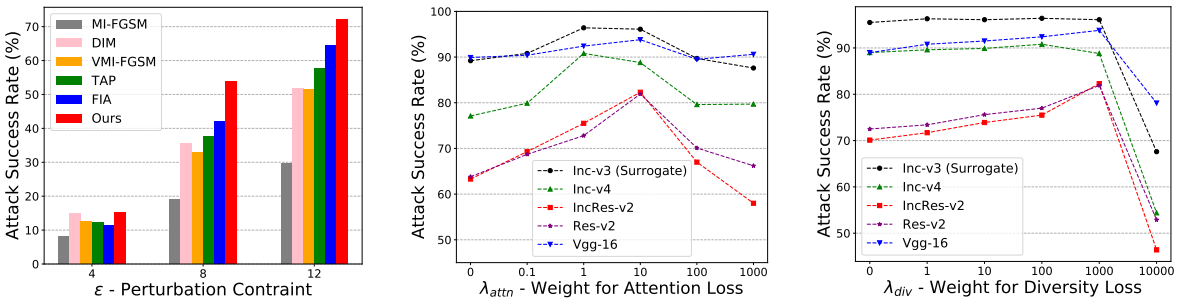


Figure 5.2: Original/adversarial images (top row) crafted by our attack and their attention heatmaps (bottom row). Our attack crafts adversarial examples that disrupt the attention and final predictions in diverse ways.



(a) (b) (c)

Figure 5.3: Analysis of our method with all adversarial examples crafted on Inc-v3 as the surrogate model. (a) comparison of attack success rates (ASR) on the ensemble of black-box target models with varying perturbation constraint ϵ, (b) ASR on varying weights for attention loss λ_{attn} when $\lambda_{div} = 1000$, and (c) ASR on varying weights for diversity loss λ_{div} when $\lambda_{attn} = 10$.

which is 0.859. However, FIA suffers from poor transferability on the target model, showing a far lower average distance of 0.693 as compared to 0.963 of our method. Our method effectively avoids poor local optimum and improves transferability by learning to craft semantically diverse perturbations.

Visualization of Diverse Adversarial Examples. We additionally visualize semantically diverse adversarial examples crafted by our method in Fig. 5.2. From a single image, our method generates various adversarial examples that disrupt the image attention and model predictions in a diverse manner.

5.4 Ablation Studies

Various Perturbation Constraint ϵ. In Fig. 5.3(a), we compare the transferability of our attack with the existing transfer-based black-box attacks under lower perturbation constraints ϵ (*i.e.*, 4, 8, and 12) by reporting the ASR on the ensemble of black-box target models. Our method exhibits superior transferability over all of the existing methods under all constraints, outperforming FIA by average ASR of 7.7%. Even when perturbations are less visible, our method exhibits high transferability.

Effects of L_{attn}. Fig. 5.3(b) shows ASR of our attack with varying λ_{attn}, or the weights for L_{attn}, when λ_{div} is fixed to 1000. In general, L_{attn} plays an important role at boosting the transferability, achieving

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-v2	Vgg16	Ensemble
Ours (w/o channel-norm)	79.3	87.2	92.9	72.0	88.9	82.9
Ours	**85.2**	**89.4**	**93.0**	**80.7**	**89.4**	**85.9**

Table 5.3: Comparison of attack success rates (%) as we remove channel-wise normalization from the attention loss L_{attn}. All adversarial examples are crafted using IncRes-v2 as the surrogate model. Best results are highlighted in **bold**.

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-v2	Vgg16	Ensemble
Pixel-level	76.3	79.0	91.7	73.8	88.2	74.8
Feature-level	76.5	85.2	92.4	76.7	88.4	82.5
Ours	**85.2**	**89.4**	**93.0**	**80.7**	**89.4**	**85.9**

Table 5.4: Comparison of attack success rates (%) as we apply the diversity loss L_{div} on the pixel/feature-level. All adversarial examples are crafted using IncRes-v2 as the surrogate model. Best results are highlighted in **bold**.

the highest performance when $\lambda_{attn} = 10$. Interestingly, using a slightly lower weight of $\lambda_{attn} = 1$ leads to higher ASR on Inc-v4 but significantly lower ASR on the other target models. This is because a lower weight on attention loss causes the attack to overfit to features specific to the surrogate model Inc-v3, which has a similar architecture as Inc-v4. Thus, the adversarial examples fool Inc-v4 with a high success rate but suffers from lower transferability to more differently structured models, such as Res-v2 and Vgg-16.

Effects of L_{div}. Fig. 5.3(c) shows ASR as we alter λ_{div}, or the weights for L_{div}, when $\lambda_{attn} = 10$. Similarly, L_{div} generally boosts transferability, achieving the highest performance when $\lambda_{div} = 1000$, showing that exploring diverse features is vital for the attack to escape from poor local optimum. Also in this scenario, our attack shows highest ASR on Inc-v4 when we apply a smaller weight on L_{div}, (*i.e.*, $\lambda_{div} = 100$). Less emphasis on the feature diversification overfits the attack to features specific to Inc-v3, which may be effective at fooling a similarly structured Inc-v4, but not the other black-box models. Too much weight on diversity loss prevents the generator from disrupting features in a destructive manner and lowers ASR.

Effects of Channel Normalization on L_{attn}. In Sec. 4.2, we claim that our work differs from the work of Wu *et al.* [17] on that we apply channel-wise normalization (Eq. 4.3) on the attention loss L_{attn} to prevent perturbation on only the few feature channels with the highest magnitudes. To verify this claim, we test the attack transferability upon removing the channel-wise normalization process from our framework, whose results are shown in Table 5.3. Our method using channel-wise normalization shows higher ASR on all target models than our method without channel-wise normalization, showing that it plays an important role perturbing the image attention in a diverse manner.

Applying Diversity Regularization on Different Levels. DSGAN [35], which originally proposes the diversity regularization, applies the diversity loss on a pixel- and feature-level. We propose in Sec. 4.3 that our approach of applying diversity on the attention level guides our attack generator to disrupt the *salient* features in a diverse manner and leads to higher transferability. To verify this claim, we modify our diversity loss L_{div} (Eq. 4.5) such that it maximizes the distance between the two adversarial examples on the pixel- or the feature-level. As shown in Table 5.4, applying diversity on the feature-level leads

to higher transferability than applying diversity on the pixel-level. This is because the diversity on the feature level guides the generator to disrupt image features in a diverse manner. However, naïvely disrupting diverse features may not necessarily lead to diversity on the attention space, and our scheme of applying diversity on the attention space leads to higher transferability.

Chapter 6. Conclusion

In this paper, we have proposed Attentive-Diversity Attack (ADA) that generates highly transferable adversarial examples. ADA relies on a generator to generate perturbations that disrupt image-salient features in a non-deterministic manner. Consequently, it avoids overfitting to model-specific features, to which existing attacks easily overfit and thus suffer from poor transferability. Exhaustive experiments validate the superior performance of ADA against state-of-the-art methods and the effectiveness of its individual components. In the future, we hope that our method can serve as a benchmark for evaluating the robustness of various models.

This work has been published at the International Conference on Image Processing (ICIP) 2022.

Bibliography

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition", in *CVPR*, 2016.

[2] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision", in *CVPR*, 2016.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation", in *CVPR*, 2015.

[4] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples", in *ICLR*, 2015.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks", in *ICLR*, 2014.

[6] Wieland Brendel, Jonas Rauber, and Matthias Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models", in *ICLR*, 2018.

[7] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin, "Black-box adversarial attacks with limited queries and information", in *ICML*, 2018.

[8] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord, "Adversarial risk and the dangers of evaluating against weak attacks", in *ICML*, 2018.

[9] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, "Adversarial examples in the physical world", in *ICLR Workshop*, 2017.

[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks", in *ICLR*, 2018.

[11] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, "Boosting adversarial attacks with momentum", in *CVPR*, 2018.

[12] 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 *CVPR*, 2019.

[13] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks", in *CVPR*, 2019.

[14] Xiaosen Wang and Kun He, "Enhancing the transferability of adversarial attacks through variance tuning", in *CVPR*, 2021.

[15] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang, "Transferable adversarial perturbations", in *ECCV*, 2018.

[16] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry, "Adversarial examples are not bugs, they are features", in *NeurIPS*, 2019.

[17] 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 *CVPR*, 2020.

[18] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren, "Feature importance-aware transferable adversarial attacks", in *ICCV*, 2021.

[19] Yunseok Jang, Tianchen Zhao, Seunghoon Hong, and Honglak Lee, "Adversarial defense via learning to generate diverse attacks", in *ICCV*, 2019.

[20] Shumeet Baluja and Ian Fischer, "Learning to attack: Adversarial transformation networks", in *AAAI*, 2018.

[21] Xiaofeng Mao, Yuefeng Chen, Yuhong Li, Yuan He, and Hui Xue, "Gap++: Learning to generate target-conditioned adversarial examples", 2020.

[22] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli, "Cross-domain transferability of adversarial perturbations", 2019.

[23] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie, "Generative adversarial perturbations", in *CVPR*, 2018.

[24] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song, "Generating adversarial examples with adversarial networks", in *ICJAI*, 2018.

[25] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song, "Delving into transferable adversarial examples and black-box attacks", in *ICLR*, 2017.

[26] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami, "Practical black-box attacks against machine learning", in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017.

[27] Wieland Brendel, Jonas Rauber, and Matthias Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models", in *ICLR*, 2018.

[28] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui 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 (AISEC)*, 2017.

[29] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin, "Black-box adversarial attacks with limited queries and information", in *ICML*, 2018.

[30] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami, "Practical black-box attacks against machine learning", in *The Asia Conference on Computer and Communications Security (ASIA CCS)*, 2017.

[31] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng, "Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks", in *AAAI*, 2019.

[32] Jonathan Uesato, Brendan O'Donoghue, Pushmeet Kohli, and Aaron van den Oord, "Adversarial risk and the dangers of evaluating against weak attacks", in *ICML*, 2018.

[33] Weibin Wu, Yuxin Su, Michael R Lyu, and Irwin King, "Improving the transferability of adversarial samples with adversarial transformations", in *CVPR*, 2021.

[34] 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 *ICCV*, 2017.

[35] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee, "Diversity-sensitive conditional generative adversarial networks", in *ICLR*, 2019.

[36] Yinpeng Dong, Zhijie Deng, Tianyu Pang, Hang Su, and Jun Zhu, "Adversarial distributional training for robust deep learning", in *NeurIPS*, 2020.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation", in *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[38] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization", in *ICLR*, 2015.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge", in *International Journal of Computer Vision*, 2015.

[40] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al., "Adversarial attacks and defences competition", in *The NIPS'17 Competition: Building Intelligent Systems*, 2018.

[41] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning", in *AAAI*, 2017.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks", in *ECCV*, 2016.

[43] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition", in *ICLR*, 2015.

[44] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao, "Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models", in *ECCV*, 2018.

[45] Li Fei-Fei, Rob Fergus, and Pietro Perona, "One-shot learning of object categories", in *TPAMI*, 2006.

Acknowledgments in Korean

연구자의 길을 걷기 위해 처음 연구실에 개별연구생으로 발을 들였던 2020년 봄부터 석사 졸업을 앞둔 지금까지 도와주셨던 모든 분께 감사의 인사를 드립니다.

우선 올바른 연구 방법을 항상 일깨워 주시고, 잘못된 방향으로 어긋나지 않도록 격려 해주시고 바로 잡아주신 윤성의 교수님께 가장 큰 감사를 올립니다. 교수님의 지도 덕분에 포기하지 않고 꾸준히 노력할 수 있는 원동력을 얻을 수 있었습니다. 제 부족했던 연구의 문제점을 짚어주시고 큰 가르침을 주신 홍승훈 교수님께도 진심으로 감사드립니다. 두 교수님께 지도받을 수 있어 행운이었고, 주신 가르침을 마음에 새기며 박사과정 동안에도 열심히 노력하도록 하겠습니다.

대학원 생활을 즐겁게 할 수 있도록 도와주신 SGVR 연구실 구성원분께도 감사드립니다. 항상 제 연구의 부족한 점을 짚어주고 조언을 아끼지 않았던 비전/렌더링팀의 태영이형, 우빈이형, 재윤이형(이미지), 준식이형, Xu, Guoyuan, 윤기형, 재윤이형(렌더링), 규범이, 주민이, 주형군, 우정군, Yaxin에게 감사드립니다. 비록 같은 팀은 아니지만 연구의 이모저모에 대해 조언해주시고 연구실 생활에 적응할 수 있도록 도와주신 (강)민철이형, 인규형, 희찬이형, 충수형님, (김)민철형님, 민성이형, 세빈이형, 진원이형께도 감사드립니다. 이제는 졸업하셨지만, 제가 연구실에 처음 왔을 때 적응할 수 있도록 도와주신 Yuchi 박사님, 김수민 박사님, 권용선 박사님, 훈민님, 창호군, 진혁님, 인영님, 형열님, 규연님, Harry, Pei에게도 감사의 말씀을 드립니다. 연구에 몰두할 수 있도록 복잡한 행정일을 처리해주신 김슬기나 선생님께도 감사의 말씀을 올립니다. 함께 시간을 보낼 수 있어서 행복했고, 박사과정 때도 함께 할 생각에 기대가 됩니다.

마지막으로 지금까지의 시간 동안 웃음을 잃지 않게 해준, 그리고 앞에 놓여진 박사과정도 함께 할 시목이에게도 고마운 마음을 전합니다. 항상 바쁘다는 핑계로 자주 찾아 뵙지 못했지만, 자신감을 잃지 않게 사랑으로 응원해주시고, 학계의 길을 걸을 수 있도록 격려해주신 부모님과 형에게도 감사의 마음을 전합니다. 항상 초심을 잃지 않고 열심히, 그리고 꾸준히 나아가겠습니다.

Curriculum Vitae in Korean

이　　름: 김 우 재

학　　력

2013. 1. – 2016. 5.　Northview High School, GA, USA

2016. 9. – 2021. 2.　한국과학기술원 전산학부 및 전기및전자공학부 (학사, 부전공)

2021. 3. – 2023. 2.　한국과학기술원 전산학부 (석사)

경　　력

2018. 12. – 2019. 2.　SK 하이닉스 인턴

2021. 3. – 2021. 6.　한국과학기술원 전산학부 조교 (CS101, 프로그래밍기초)

2021. 9. – 2022. 12.　한국과학기술원 전산학부 조교 (CS206, 데이타구조)

2022. 3. – 2022. 6.　한국과학기술원 전산학부 조교 (CS101, 프로그래밍기초)

연 구 업 적

1. **Woo Jae Kim**, Yoonki Cho, Junsik Jung, and Sung-Eui Yoon, "Feature Separation and Recalibration for Adversarial Robustness", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023 (under review).
2. Guoyuan An, **Woo Jae Kim**, Saelyne Yang, Yuchi Huo, Rong Li, and Sung-Eui Yoon, "Towards Content-based Pixel Retrieval in Revisited Oxford and Paris", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023 (under review).
3. **Woo Jae Kim**, Seunghoon Hong, and Sung-Eui Yoon, "Diverse Generative Perturbations on Attention Space for Transferable Adversarial Attacks", IEEE International Conference on Image Processing (ICIP), 2022.
4. Yoonki Cho, **Woo Jae Kim**, Seunghoon Hong, and Sung-Eui Yoon, "Part-based Pseudo Label Refinement for Unsupervised Person Re-identification", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
5. Kyuyeon Kim, Junsik Jung, **Woo Jae Kim**, and Sung-Eui Yoon, "Deep Video Inpainting Guided by Audio-Visual Self-Supervision", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2022.
6. **Woo Jae Kim**, and Sung-Eui Yoon, "적대적 공격에 견고한 피처 신뢰도 기반 다운샘플링", Workshop on Image Processing and Image Understanding (IPIU), 2022.