mathematics

MDPI

Article

Enhance Domain-Invariant Transferability of Adversarial Examples via Distance Metric Attack

Jin Zhang [1,†], Wenyu Peng [2,3,*,†], Ruxin Wang [2,3], Yu Lin [1], Wei Zhou [2,3] and Ge Lan [1]

[1] Kunming Institute of Physics, Kunming 650223, China; zhangjin_211@163.com (J.Z.); lwlinyu@163.com (Y.L.); gelan_211@163.com (G.L.)
[2] School of Software, Yunnan University, Kunming 650500, China; ruxin.wang@ynu.edu.cn (R.W.); zwei@ynu.edu.cn (W.Z.)
[3] Engineering Research Center of Cyberspace, Yunnan University, Kunming 650500, China
* Correspondence: winniepeng@mail.ynu.edu.cn
† These authors contributed equally to this work.

Abstract: A general foundation of fooling a neural network without knowing the details (i.e., black-box attack) is the attack transferability of adversarial examples across different models. Many works have been devoted to enhancing the task-specific transferability of adversarial examples, whereas the cross-task transferability is nearly out of the research scope. In this paper, to enhance the above two types of transferability of adversarial examples, we are the first to regard the transferability issue as a heterogeneous domain generalisation problem, which can be addressed by a general pipeline based on the domain-invariant feature extractor pre-trained on ImageNet. Specifically, we propose a distance metric attack (DMA) method that aims to increase the latent layer distance between the adversarial example and the benign example along the opposite direction guided by the cross-entropy loss. With the help of a simple loss, DMA can effectively enhance the domain-invariant transferability (for both the task-specific case and the cross-task case) of the adversarial examples. Additionally, DMA can be used to measure the robustness of the latent layers in a deep model. We empirically find that the models with similar structures have consistent robustness at depth-similar layers, which reveals that model robustness is closely related to model structure. Extensive experiments on image classification, object detection, and semantic segmentation demonstrate that DMA can improve the success rate of black-box attack by more than 10% on the task-specific attack and by more than 5% on cross-task attack.

Keywords: deep learning; distance metric; adversarial attack; cross-task; transferability

MSC: 68T07

Citation: Zhang, J.; Peng, W.; Wang, R.; Lin, Y.; Zhou, W.; Lan, G. Enhance Domain-Invariant Transferability of Adversarial Examples via Distance Metric Attack. *Mathematics* **2022**, *10*, 1249. https://doi.org/10.3390/math10081249

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 8 March 2022
Accepted: 8 April 2022
Published: 11 April 2022

1. Introduction

The adversarial examples are crafted by adding the maliciously subtle perturbations to the benign images, which make the deep neural networks being vulnerable [1,2]. It is possible to employ such examples to interfere with real-world applications, thus raising concerns about the safety of deep learning [3–5]. While most of the adversarial attacks focus on a single task, we consider that the current vision-based systems usually consist of an ensemble of multiple pipelines with each addressing a certain task, such as object detection, tracking, or classification. Hence, for such a complex vision system, an adversarial example attacking multi-task or multi-model vulnerability is desired but challenging to be designed.

Generally, adversarial attacks can be divided into the white-box and the black-box cases [6]. The white-box attacks are known as attacking with the knowledge of the structure and the parameters of the given model, such as the fast gradient sign method [2], the basic iterative method [7], and the momentum-boosting iterative method [6]. On the contrary, the black-box attacks do not know the information of the model except the model outputs,

which describe a more common situation in real-world applications. The success of a black-box attack comes from either of two principles, i.e., the assumption of transferability or the feedback of queries. Hence, we could find two categories of black-box attacks, including transfer-based [8–11] and query-based [12,13]. While the latter has the problems such as poor attack effects and low query efficiency [14], in this paper, we focus on the transfer-based black-box attack, in which transferability is assumed to be an intriguing property of adversarial examples.

The assumption of transferability comes from the fact that different models are optimised based on similar distributions of training data, which means the adversarial examples generated by a given model can also fool the other unknown models. In details, transferability can be divided into the task-specific transferability and the cross-task transferability, according to the task of the victim model. Specifically, when the victim model and the given model are interested in the same task (e.g., classification), the assumed transferability is task-specific. On the other hand, when the victim model and the given model are interested in different tasks (e.g., classification vs. detection), the cross-task transferability is considered. To design the adversarial examples for multiple tasks, a natural question is: *are cross-task transferability and task-specific transferability incompatible*?

Regarding the task-specific transferability, it is known that the models are optimised from similar input distributions and similar label distributions, which could be viewed as in the same domain and hence, the characteristics revealed by the models are similar. Instead, the cross-task transferability can be regarded as a heterogeneous domain generalisation problem [15], where the label distributions are quite different although the input distributions are still similar. The heterogeneous domain generalisation problem is a typical problem in training neural networks. Learning the domain-invariant features has been proven as an effective way to solve the above issue [15], which could encourage good generalisation from the source domain to the unknown target domain. In this regard, when the feature extractor is aware of the underlying distribution of the source domain, the adversarial examples are the outliers of the distribution [16,17]. The question is then how to exploit the distribution of the outliers transferable across the domains. As shown in Figure 1, if the feature extractor has a good generalisation ability, the target domain and the source domain are well aligned, which helps to transfer both the benign examples and the adversarial examples. Hence, to enhance the domain-invariant transferability (i.e., both task-specific transferability and cross-task transferability) of the adversarial examples, a natural choice is to craft the adversarial examples based on a well-generalised feature extractor, e.g., pre-trained on ImageNet. As shown in Figure 2, the difference between the adversarial example and the benign image can be reflected in the feature-extraction stage and in the task-related stage of the model. While the task-specific transferability does not matter (since both stages have transferability), the cross-task transferability mostly relies on the transferability on the feature-extraction stage. However, most of the transfer-based attacks developed on image classification rely on the task-specific loss (e.g., the cross-entropy loss), which limits the cross-task transferability of the adversarial examples [18].

In this paper, we propose a novel cross-task attack method called distance metric attack (DMA) to enhance the domain-invariant transferability of the adversarial examples. Different from the normal gradient-based attacks that craft the benign input by maximising the cross-entropy loss, the goal of distance metric attack is to maximise the distance of the latent features between the adversarial example and the benign example. To ensure the basic transferability between different models, we consider the task-specific loss as the attack direction. To reasonably mitigate the effect of the task-specific loss, we use a weight factor to control the trade-off between the direction and the distance. We show that the adversarial examples crafted by distance metric attack can fool the models on image classification, object detection, and semantic segmentation. This indicates that distance metric attack can effectively improve the domain-invariant transferability of the adversarial examples.

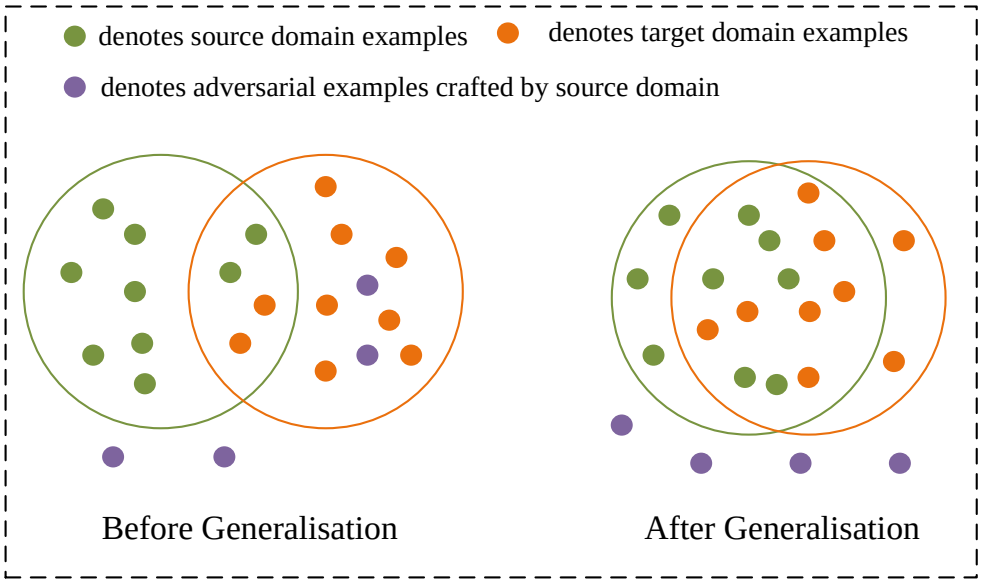

Figure 1. The relationship of the distribution of adversarial examples in different domains.

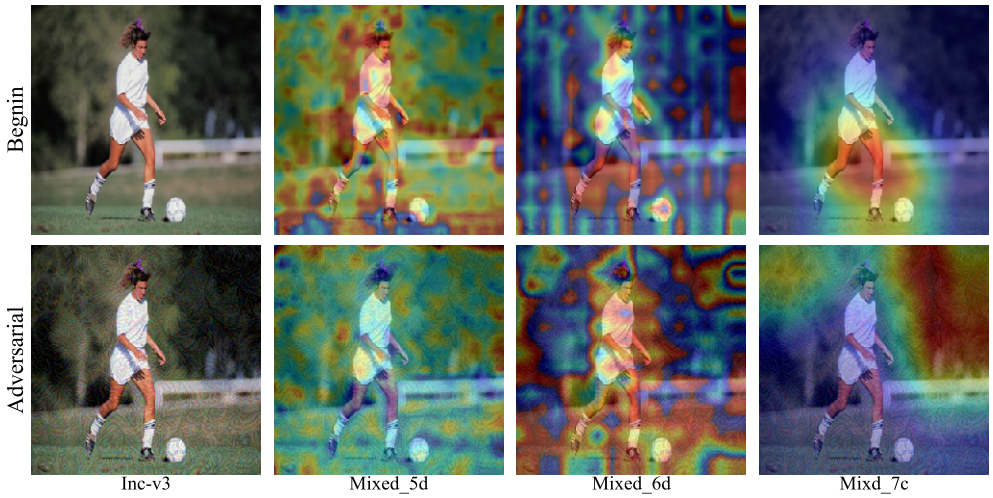

Figure 2. Visualisation of the features of adversarial and benign examples on different layers on the Inception-V3 model. The features are visualised by Grad-Cam [19].

Our main contributions are summarised as follows:

- We proposed a novel adversarial attack method, termed as distance metric attack (DMA), which enhances the domain-invariant transferability of the adversarial examples.
- We evaluate the robustness of the latent layers of different models by maximising the feature distance and find that the models with similar structures have consistent robustness at the same layer.
- Empirical results show that the attack success rate of the adversarial examples crafted by DMA is significantly improved on the multiple tasks, including image classification, object detection, and semantic segmentation.

The rest of the paper is arranged as follows: In Section 2, we review the related work about the adversarial attack on image classification, other vision tasks, and the cross-task case. In Section 3, we introduce the proposed distance metric attack (DMA) method. In Section 4, we present the attack results of DMA compared with multiple baselines on a variety of tasks. Finally, the conclusions are drawn in Section 5.

2. Related Work

In this section, we briefly review the adversarial attack methods on image classification, object detection, and semantic segmentation. Then, a brief explanation of the cross-task attack is given.

2.1. Adversarial Attacks on Image Classification

DNNs have shown vulnerability to the adversarial examples [1,2,20], which has attracted widespread attention. Many effective white-box attacks have been proposed, such as FGSM [2], BIM [7], C&W [21], DeepFool [22], and MIM [6], which rely on the details of the victim model. However, in real-world applications, the model details are often invisible. The transferability of the adversarial example motivates the black-box attacks. Inspired by data augmentation, many attacks enhance the transferability of adversarial examples through input transformations. For example, the diverse input method (DIM) [8] created diverse input patterns by applying random resizing and padding to the input at each iteration before feeding the image into the model for gradient calculation. The translation-invariant method (TIM) [9] optimised an adversarial example through an ensemble of multiple translated images and simplified the complex computation into a single convolutional operation according to the translation invariance of CNN. The scale-invariant method (SIM) [10] enhanced the transferability of adversarial examples by optimising the example with multi-scale copies which, however, yielded a huge cost of computation. In addition, the ILA method [23] aimed at attacking the latent layers, which also provided a new direction for improving the transferability of adversarial examples. Nevertheless, those methods only focused on the image classification task. The transferability of adversarial examples crafted by the above methods was based on the assumption that the victim model and the given model were trained on the same dataset. However, the real scenarios tell us that the real data are always changing and complex.

2.2. Adversarial Attacks on Other Vision Tasks

Compared with the image classification task, the adversarial examples for object detection and semantic segmentation are more challenging to be designed. Xie et al. [24] proposed DAG to generate adversarial examples for a wide range of segmentation and detection. Many adversarial patch attacks have been proposed to attack the object detection systems, such as Dpatch [25], person patch [26], and adversarial T-shirt [27]. However, those adversarial attacks require a huge cost of training time. Xiao et al. [28] characterised adversarial examples based on the spatial context information in semantic segmentation. However, the generated adversarial examples are barely transferred among models even in the same task.

2.3. Adversarial Defences

Corresponding to adversarial attack, adversarial defence has been developed vigorously in recent years. The methods that integrate the adversarial examples into the training dataset are called adversarial training [1,2], which is a promising adversarial defence scheme. Then, Tramer et al. [29] proposed the ensemble adversarial training, which generated adversarial examples by assembling multiple models. To improve adversarially robust generalisation and exploit robust local features, Song et al. [30] proposed a random block shuffle transformation, which cut up the adversarial example into blocks and then randomly combined those blocks to reassemble the example for adversarial training. However, the computational cost of adversarial training is too high, and adversarial training can only be designed for a single task.

In addition to adversarial training, mitigating the effects of adversarial perturbations is also an effective defence scheme. A set of image transformation methods were proposed by Guo et al. [31], which transformed the image before being input into the classifier. Xie et al. [32] randomly resized and padded the input image to mitigate the adversarial perturbations. However, all these defence schemes are developed for a specific single task.

2.4. Adversarial Attacks on Cross-Task

All the above adversarial attacks are designed for a single task, which limits the practicability of adversarial examples. A detection system based on computer vision (CV) techniques has been deeply applied in various security scenarios, which generally involves more than one model. Therefore, it is difficult for the above adversarial attacks for a specific task to attack the real-world CV systems successfully. Lu et al. [18] was the first to propose the cross-task attack (DR), which used the model of image classification to generate adversarial examples that could fool the models of object detection and semantic segmentation. Cross-task attack is a more challenging attack, where the source model is very different from the target models in the aspects of employed data and model structures. However, the DR attack has a low success rate in image classification. The main difference of performance between our proposed DMA and DR is that DMA can achieve a high attack success rate on image classification, object detection, and semantic segmentation. Namely, DMA can effectively enhance the domain-invariant transferability of the adversarial examples.

3. Methodology

3.1. Notation

Let x and y be the clean image and the corresponding label, respectively. $\ell_f(x, y)$ is the cross-entropy loss of the image classifier $f(x)$. The adversarial example x^{adv} is indistinguishable from the clean image x but fools the classifier, i.e., $f(x^{adv}) \neq y$. Following the previous work, we use the L_∞ norm to constrain the adversarial perturbation level as $||x^{adv} - x||_\infty \leq \epsilon$. The goal of adversarial attack is to find an adversarial example x^{adv} that maximises the loss $\ell_f(x^{adv}, y)$. Regarding the feature space, let the latent feature $f_l(x)$ be the l-th layer of the classifier when the input is x. The distance function $D(f_l(x), f_l(x^{adv}))$ is used to measure the distance (e.g., L2 distance) between the latent layers of those examples. Thus, the optimisation problem in the normal gradient-based attacks can be written as:

$$\arg\max_{x^{adv}} \ell_f(x^{adv}, y), s.t. ||x^{adv} - x||_\infty \leq \epsilon. \tag{1}$$

3.2. Motivation

The domain-invariant transferability of adversarial examples includes the task-specific transferability and the cross-task transferability. Recent advances of adversarial attacks focus on enhancing the task-specific transferability, where the adversarial examples crafted by the given model can also fool unknown models on the same task. The task-specific transferability of the adversarial examples is due to the given model and the unknown models being trained on the same domain. On the other hand, the cross-task transferability can be regarded as the heterogeneous domain generalisation problem, where the domains have different label spaces [15]. To address the heterogeneous domain generalisation problem, many methods [15,33,34] aim to generate a domain-invariant feature representation. In this case, the whole network is split into the feature extractor and the classifier. To match various classifiers, the feature extractor is trained to be as general as much. Fortunately, the feature extractor pre-trained on ImageNet is a general model.

As can be seen from Figure 2, the difference between the adversarial example and the original image is reflected from the difference in features, which is eventually evolved into the difference in the identification regions. Therefore, the feature extractor based on ImageNet can solve the problem of heterogeneous domain generalisation and can improve the domain-invariant transferability by expanding the distance of the latent features.

3.3. Distance Metric Attack

Based on the above analyses, by attacking the feature space of the model, the domain-invariant transferability of the adversarial examples can be enhanced. ILA [23] points out that the adversarial perturbation is constrained by the norm, but the perturbation in the latent features of the model is not constrained. So the perturbation on the latent features

can be maximised to perform attack. Motivated by this, we propose distance metric attack (DMA), which aims to maximise the distance between the latent features of the benign image and the adversarial image.

The gradient-based attack algorithms imply that the direction of attack is as important as the magnitude of the perturbation. ILA crafts the adversarial examples by introducing external adversarial examples that are used as the direction of attack, so that the latent layer of the adversarial examples generated by the current algorithm is close to the corresponding latent layer of the adversarial examples by other algorithms. Different from ILA, DMA does not need to introduce external adversarial examples. The whole framework is illustrated in Figure 3. We assume that the pre-trained model can be split into the feature extractor part and the classifier part. DMA can be directly combined with the other adversarial attacks, where the resultant loss involves both the cross-entropy loss for image classification and the distance metric loss on the latent layers. The optimisation makes the latent feature distance between the benign examples and the adversarial examples farther and farther in the process of generating adversarial examples iteratively. In this way, DMA can maximise the distance between the latent features of the benign images and the adversarial images. In addition, the cross-entropy loss dependent on the gradient-based attack serves as the attack direction of DMA. Therefore, we define the problem of finding an adversarial example as an optimisation problem:

$$\underbrace{\max_{x^{adv}} D(f_l(x), f_l(x^{adv}))}_{\text{maximise distance}} + \underbrace{\arg\max_{x^{adv}} \ell_f(x^{adv}, y)}_{\text{attack direction}}, s.t. ||x^{adv} - x||_\infty \leq \epsilon. \tag{2}$$

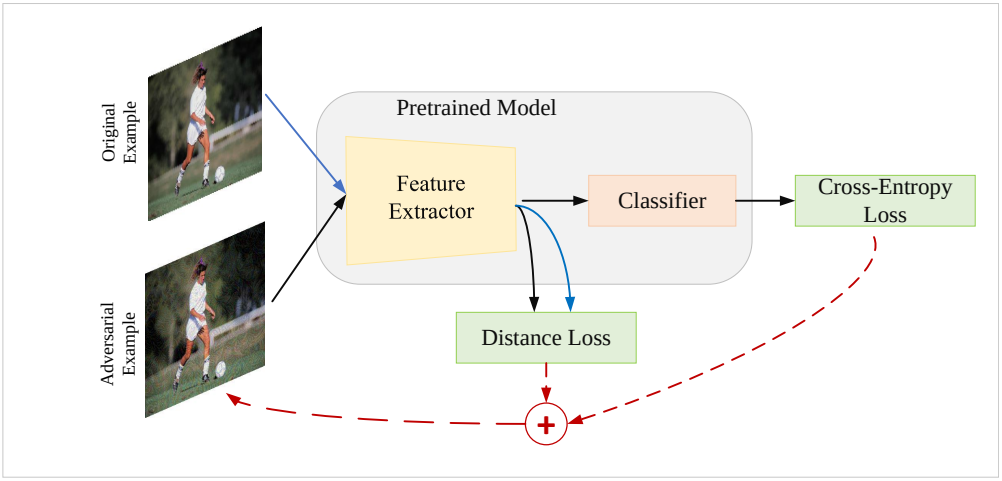

Figure 3. Illustration of the distance metric attack framework.

To solve the problem in Equation (2), we need to calculate both the gradient of the cross-entropy loss with respect to the input x and the gradient of the distance metric loss to the input x. However, the cross-entropy loss limits the cross-task transferability of the adversarial examples. To mitigate the influence of the cross-entropy loss, we set a hyper-parameter $\beta >= 1$ to flexibly increase the weight of the distance loss. Hence, Equation (2) can be written in detail as:

$$\begin{aligned} &\text{maximise } L(x^{adv}, x, y), \\ &\text{where } L(x^{adv}, x, y) = \ell_f(x^{adv}, y) + \beta \cdot D(f_l(x) - f_l(x^{adv})), s.t. ||x^{adv} - x||_\infty \leq \epsilon. \end{aligned} \tag{3}$$

For fair comparisons, we use the MI-FGSM as the optimisation method to craft the adversarial example, which is an efficient iterative gradient-based attack. Therefore, when $\beta = 0$, DMA degenerates to the vanilla gradient-based attack (MI-FGSM).

Specifically, the manufacture of adversarial examples in MI-FGSM be formulated as:

$$\begin{aligned} x_0^{adv} &= x, \\ g_{t+1} &= \mu \cdot g_t + \frac{\nabla_x L(x_t^{adv}, x, y)}{||\nabla_x L(x_t^{adv}, x, y)||_1} \\ x_{t+1}^{adv} &= x_t^{adv} + \alpha \cdot sign(g_{t+1}) \end{aligned} \tag{4}$$

where g_t is the accumulative gradient in the t-th iteration in the attack process, and μ is a decay factor. The DMA algorithm for crafting adversarial examples iteratively is summarised in Algorithm 1, where DMA is combined with MI-FGSM.

Algorithm 1 Distance Metric Attack

Input: A deep model f and the loss function ℓ_f; the latent layer f_l of the model f; a benign example x and its ground-truth label y.
Input: The maximum perturbation ϵ, the number of iteration T, the decay factor μ, and the distance weight β.
Output: An adversarial example x^{adv}.

$\alpha = \epsilon/T$, $g_0 = 0$, $x_0^{adv} = x$;
for $t = 0 \to T - 1$ **do**
  Get the latent feature $f_l(x)$ of the model by inputting x; Obtain the latent feature $f_l(x_t^{adv})$ of the model by inputting x_t^{adv};
  Calculate the distance between the latent features $D(f_l(x) - f_l(x_t^{adv})) = ||f_l(x) - f_l(x_t^{adv})||_2$
  Get the softmax cross-entropy loss $\ell_f(x_t^{adv}, y)$.
  Calculate the loss $L(x_t^{adv}, x, y) = \ell_f(x_t^{adv}, y) + \beta \cdot D(f_l(x) - f_l(x_t^{adv}))$.
  Calculate the gradient $\nabla_x L(x_t^{adv}, x, y)$.
  Update g_{t+1} by $g_{t+1} = \mu \cdot g_t + \frac{\nabla_x L(x_t^{adv}, x, y)}{||\nabla_x L(x_t^{adv}, x, y)||_1}$.
  Update x_{t+1}^{adv} by $x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(g_{t+1})$
end for
return x_T^{adv};

Note that DMA generates adversarial examples based on a highly generalised image classification model, expecting that the adversarial example can fool models that are not only image classification models but also object detection and semantic segmentation models. However, for the image on object detection and semantic segmentation, there are no ground-truth labels for the source model. Before the craft adversarial example, DMA would give the image an alternative label in the source model labels by $y = f(x)$. Then, feeding the original image to Algorithm 1, we get the adversarial examples. At the end, we input the adversarial examples into the target models to get the attack results.

4. Experiments

4.1. Setup

4.1.1. Datasets

In experiments, we evaluate the performance of the proposed method on cross-domain tasks, including image classification, detection, and segmentation. For the image classification task, we randomly choose 1000 images from the ILSVRC 2012 validation set, which are almost correctly classified by all the image classification victim models. For object detection and semantic segmentation, we randomly select 1000 images from the COCO2017 and PASCAL VOC2012 datasets, respectively. All images are resized to the size of $3 \times 299 \times 299$.

4.1.2. Models

We use four normally trained image classification models as the target models to craft the adversarial examples, including Inception-v3 (Inc-v3) [35], Inception-v4 (Inc-v4) [36], Inception-Resnet-v2 (IncRes-v2) [36], and Resnet-v2-101 (Res-101) [37]. For the image classification task, we also employ three adversarially trained models as the victims, including Inc-v3$_{ens3}$, Inc-v3$_{ens4}$, and IncRes-v2$_{ens}$ [29]. In addition, we evaluate the attack performance of object detection on Yolov3-DarkNet53 [38], Faster R-CNN-ResNet101 [39], RetinaNet-ResNet101 [40], YoloF-ResNet50 [41], and Sparse R-CNN-ResNet101 [42], which are available on mmdetection [43]. The performance on the semantic segmentation task is tested on FCN-ResNet50 [44], DeepLabv3-ResNet50 [45], ANN-ResNet50 [46], OCRNet-HRNetV2p [47], and GCNet-ResNet101 [48], which are available on mmsegmentation [49].

4.1.3. Hyper-Parameters

We consider MI-FGSM [6], DIM [8], TIM, DI-TIM [9], ILA [23], and DR attack [18] as the baselines. For the settings of hyper-parameters, we set the maximum perturbation to be $\epsilon = 16$ in the pixel range of [0, 255]. All the baselines are iterative attack, where we set the iteration as $T = 10$ and the step size as $\alpha = 1.6$. For MI-FGSM, DIM, TIM, and DI-TIM, we set the decay factor as $\mu = 1.0$. For DIM and DI-TIM, we set the transformation probability to 0.7. For TIM and DI-TIM, the kernel size is set to 7×7.

4.2. Ablation Studies

4.2.1. The Effect of the Distance Loss and the Factor β

To further gain insight into the performance of DMA, we conduct the ablation studies to examine the effect of various factors. We attack the Inc-v3 model by DMA with four distance losses and different factor values β, which range from 0 to 200. Note that when $\beta = 0$, DMA degenerates to MI-FGSM. As shown in Figure 4, we observe that the MSE loss, the L1 loss, and the cosine distance loss can improve the transferability of the adversarial examples compared with MI-FGSM.

It can be found from Figure 4 that in the cases of the MSE loss and the L1 loss, the transferability of the adversarial examples increases with the factor β, which indicates that the task-specific loss not only limited the transferability on cross-tasks but also on cross-models. When the distance loss is the MSE loss and the factor β is 200, DMA exhibits the best transferability on all models. Specifically, on the first picture in Figure 4, it can be seen that as β increases, the success rate gradually increases. When $\beta = 200$, the increase in the attack success rate gradually becomes flat. Therefore, we adopt the MSE loss and the factor $\beta = 200$ in the following experiments.

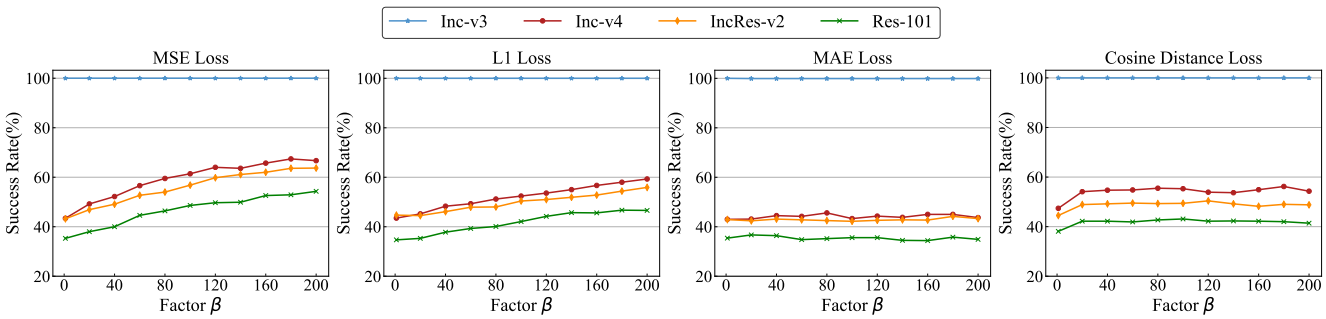

Figure 4. The success rate of different distance losses with the factor β from 0 to 200. The adversarial examples are crafted by Inception-v3 where the selected latent layer is the 6th layer.

4.2.2. The Performance on Attacking Different Layers

To evaluate the robustness of the latent layers in various networks, we compare the transferability of the adversarial examples crafted by DMA on different layers in four normally trained models. As Figure 5 reports, for the models with similar structures, the

robustness on the same layer is consistent. This indicates that the robustness of the model is related to the model structure. Interestingly, Inception-Resnet-v2 performs similarly to ResNet-v2 on the black-box attack, with the first few layers of attacks working well. For the white-box attack, layer 6 and layer 7 work better, which is similar to the Inception series.

In the following experiments, for Inc-v3 and Inc-v4, we select the sixth layer as the latent layer. For Res-101, we adopt the third layer as the latent layer. Since the success rate of the previous layers of Inc-Res is insufficient in the white-box setting, we use the sixth layer as the latent layer.

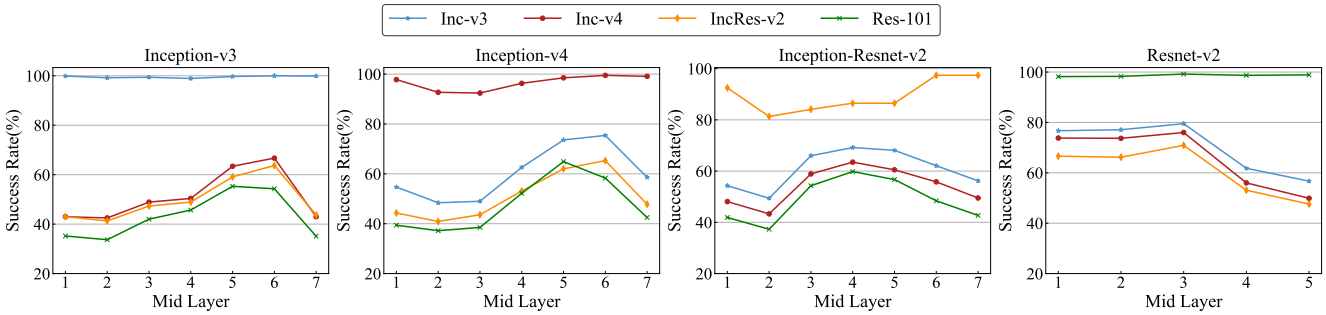

Figure 5. Evaluation of the robustness of the latent layers in four models.

4.3. Adversarial Attack on Image Classification

In this section, we present the attack results on the image classification task. To verify the effectiveness of DMA, we use MI-FGSM, TI-DIM, and ILA (where the proxy is crafted by MI-FGSM and TI-DIM, respectively) as the competitors. For cross-task attack, we first evaluate the task-specific transferability of DMA and DR. We report the success rates of MI-FGSM, ILA (where the proxy is crafted by MI-FGSM), and DMA in Table 1, the success rates of TI-DIM, ILA (proxy crafted by TI-DIM), and DMA in Table 2, and the success rates of DR and DMA in Table 3.

Table 1. The success rates of MI-FGSM, ILA, and DMA. The proxy of ILA is crafted by MI-FGSM. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2, and Res-101, respectively. * indicates the white-box attacks.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	MI-FGSM	99.9 *	43.8	43.6	33.7	12.2	9.9	5.6
	ILA	99.9 *	48.4	40.3	33.5	6.3	5.6	3.5
	DMA	**100.0** *	**66.7**	**63.7**	**54.3**	**15.9**	**13.6**	**7.2**
Inc-v4	MI-FGSM	56.3	**99.9** *	46.1	40.5	13.7	11.1	7.1
	ILA	58.5	99.6 *	43.2	36.2	8.0	5.6	5.2
	DMA	**75.4**	99.5 *	**65.3**	**58.3**	**17.6**	**15.5**	**7.7**
IncRes-v2	MI-FGSM	57.0	50.1	97.3 *	43.0	18.1	15.6	10.5
	ILA	**71.5**	**64.3**	**97.8** *	**56.5**	**21.2**	15.1	**12.3**
	DMA	62.1	55.8	97.4 *	48.4	19.8	**16.6**	11.7
Res-101	MI-FGSM	55.9	50.2	48.5	**99.4** *	22.7	19.5	11.5
	ILA	66.9	63.3	55.1	**99.4** *	18.5	13.3	9.1
	DMA	**79.5**	**76.0**	**70.9**	99.2 *	**31.9**	**27.4**	**15.8**

As shown in Table 1, we observe that the proposed DMA outperforms the baseline attacks in most cases. Compared with the baselines, DMA can significantly improve the task-specific transferability of the adversarial examples by 2–25%. For IncRes-v2, compared with MI-FGSM, the transferability of the adversarial examples generated by DMA is also enhanced.

TI-DIM is one of the best gradient-based adversarial attack methods, which can also be combined with DMA to generate adversarial examples. From Table 2, it can be found that ILA cannot consistently improve the transferability of adversarial examples, while the proxy examples are generated by TI-DIM. However, TI-DI-DMA outperforms TI-DIM by 2% to 15% in most cases. In particular, in the black-box manner, the adversarial examples crafted by TI-DI-DMA achieve the success rate of more than 60% on the normally trained models, with some cases even reaching 80%.

Table 2. The success rates of TI-DIM, ILA, and TI-DI-DMA. The proxy of ILA is crafted by TI-DIM. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2, and Res-101, respectively. * indicates the white-box attacks.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	TI-DIM	98.8 *	65.8	62.1	54.9	35.2	32.2	20.9
	ILA	**99.7** *	51.0	46.1	36.0	9.2	7.9	4.8
	TI-DI-DMA	99.6 *	**84.0**	**79.8**	**68.0**	**42.5**	**38.6**	**23.9**
Inc-v4	TI-DIM	72.9	**97.8** *	64.3	55.4	34.9	31.5	23.5
	ILA	61.7	99.0 *	49.1	38.3	10.7	9.6	5.3
	TI-DI-DMA	**83.2**	97.7 *	**71.0**	**64.5**	**41.6**	**36.5**	**25.1**
IncRes-v2	TI-DIM	68.1	65.6	91.9 *	59.2	43.0	37.3	35.1
	ILA	**74.5**	67.5	**95.2** *	59.6	26.8	19.3	18.2
	TI-DI-DMA	70.7	**70.1**	92.0 *	**62.5**	**45.1**	**39.9**	**36.8**
Res-101	TI-DIM	75.0	70.6	69.3	99.2 *	**54.3**	**50.2**	40.0
	ILA	72.7	68.4	64.4	**99.3** *	22.9	19.2	12.4
	TI-DI-DMA	**80.6**	**78.6**	**76.8**	98.5 *	54.1	49.0	**41.0**

The cross-task attacks should not only focus on other tasks but also the current task. Therefore, we also present the DR attack result on image classification. As shown in Table 3, we observe that the adversarial examples crafted by DR yields a low success rate on both white-box and black-box attacks. However, the attack success rate of DMA is higher than DR by a large margin on all models. Note that the empirical result of DR is deeply inferior compared with other adversarial attacks which focus on image classification. DR only focuses on enhancing the cross-task transferability, which implies that the practicality of adversarial examples is questionable.

Table 3. The success rates of DR and DMA. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2, and Res-101, respectively. * indicates the white-box attacks.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	DR	96.3 *	14.1	14.4	16.2	4.9	3.9	3.1
	DMA	**100.0** *	**66.7**	**63.7**	**54.3**	**15.9**	**13.6**	**7.2**
Inc-v4	DR	21.7	76.0 *	14.8	14.3	5.7	4.9	3.4
	DMA	**75.4**	**99.5** *	**65.3**	**58.3**	**17.6**	**15.5**	**7.7**
IncRes-v2	DR	32.5	24.3	63.3 *	25.8	12.3	10.4	7.0
	DMA	**62.1**	**55.8**	**97.4** *	**48.4**	**19.8**	**16.6**	**11.7**
Res-101	DR	35.6	30.0	27.8	98.1 *	8.8	7.6	5.8
	DMA	**79.5**	**76.0**	**70.9**	**99.2** *	**31.9**	**27.4**	**15.8**

4.4. Cross-Task Attack on Object Detection

We next evaluate the cross-task transferability of the adversarial examples generated by DR, MI-FGSM, TI-DIM, and the proposed DMA in the object detection task. For cross-task attack, all the adversarial examples come from the COCO dataset and are crafted by the models trained on the ImageNet dataset, including Inc-v3, Inc-v4, IncRes-v2, and Res-101. The label y required by MI-FGSM, TI-DIM, and DMA is obtained by inferring the image classification model on the original COCO image, which corresponds to one of the ImagNet labels.

The results of cross-task attack on object detection are presented in Table 4, which shows that compared with MI-FGSM, the attack success rate of the adversarial examples crafted by DMA with different models can lead to an improvement of 0.4–9% in all the object detection models. Except for IncRes-v2, DMA outperforms DR by 2–5%. Although TI-DIM can effectively enhance the specific-task transferability of adversarial examples, it cannot greatly improve the cross-task transferability of adversarial examples. Therefore, DMA is the first attack method that focuses on the domain-invariant transferability of adversarial examples.

Table 4. The detect results (mAP) of the adversarial examples crafted by MI-FGSM, TI-DIM, DR, and DMA on Inc-v3, Inc-v4, IncRes-v2, and Res-101, respectively.

Model	Attack	Faster RCNN ResNet101	Retinanet ResNet101	Yolov3 DarkNet53	YoloF ResNet50	Sparse RCNN ResNet101
	clean	50.5	48.1	46.8	47.1	55.2
Inc-v3	MI-FGSM	33.1	32.3	31.4	30.6	39.2
	TI-DIM	31.7	30.9	29.2	26.9	37.3
	DR	30.9	30.8	28.1	27.7	37.2
	DMA	**28.6**	**28.2**	**26.0**	**25.3**	**34.4**
Inc-v4	MI-FGSM	30.6	29.9	28.2	27.0	36.3
	TI-DIM	29.1	28.3	26.5	24.8	35.1
	DR	30.0	30.2	27.7	27.4	36.0
	DMA	**25.0**	**24.6**	**22.3**	**22.0**	**29.4**
IncRes-v2	MI-FGSM	29.5	29.8	28.1	28.0	35.3
	TI-DIM	28.3	28.5	26.9	**24.6**	34.4
	DR	**26.2**	**26.1**	**24.6**	**24.6**	**30.3**
	DMA	29.1	28.6	27.9	27.3	34.6
Res-101	MI-FGSM	30.8	30.8	29.2	29.0	35.7
	TI-DIM	30.1	30.0	28.2	27.5	35.4
	DR	25.7	25.9	23.2	23.1	31.2
	DMA	**21.8**	**22.6**	**19.3**	**20.2**	**25.8**

4.5. Cross-Task Attack on Semantic Segmentation

In this section, we further investigate the cross-task transferability of the adversarial examples generated by DR, MI-FGSM, TI-DIM, and the proposed DMA in the semantic segmentation task. All the adversarial examples are selected from the PASCAL VOC2012 dataset and are crafted by the models trained on the ImageNet dataset, including Inc-v3, Inc-v4, IncRes-v2, and Res-101. Similar to the object detection task, the label y required by MI-FGSM TI-DIM, and DMA is obtained by inferring the image classification model on the original image. The evaluation metric for the semantic segmentation is mIoU, where a lower value indicates a better attack effect.

From Table 5, we observe that DMA effectively reduces mIoU compared to MI-FGSM by 4–19% on the five semantic segmentation networks. With the exception of IncRes-v2, DMA reduces mIoU by 5–13% compared to DR and 2–20% compared to TI-DIM. In addition, among the four source models, the adversarial examples crafted by Resnet-V2 can highly reduce the mIOU of the semantic segmentation model. Meanwhile, four semantic segmentation models are based on ResNet, which indicates that the more similar the source model is to the target structure, the higher the attack success rate is.

Table 5. The segmentation results (mIoU) of the adversarial examples crafted by DR and DMA on Inc-v3, Inc-v4, IncRes-v2, and Res-101, respectively.

Model	Attack	deeplabv3 ResNet50	ANN ResNet50	FCN ResNet50	OCRNet HRNetV2p	GCNet ResNet101
	clean	66.8	66.3	58.8	64.6	67.1
Inc-v3	MI-FGSM	53.0	52.0	42.8	52.1	55.2
	TI-DIM	51.6	50.5	41.7	50.1	55.8
	DR	50.7	50.0	40.9	51.4	53.7
	DMA	**44.3**	**43.1**	**35.2**	**40.7**	**47.3**
Inc-v4	MI-FGSM	48.8	48.6	39.9	49.9	53.8
	TI-DIM	50.4	48.9	39.4	49.8	54.1
	DR	44.0	42.4	34.3	45.3	48.0
	DMA	**39.2**	**38.8**	**31.5**	**38.4**	**42.8**
IncRes-v2	MI-FGSM	49.2	48.6	38.5	48.4	51.3
	TI-DIM	47.8	47.1	38.2	47.9	51.5
	DR	**44.4**	**43.6**	**33.3**	**42.0**	**47.4**
	DMA	47.4	47.3	37.9	47.5	50.0
Res-101	MI-FGSM	48.8	48.7	39.1	48.7	52.2
	TI-DIM	50.5	48.9	39.5	48.8	53.7
	DR	42.2	41.2	32.1	41.6	46.3
	DMA	**32.0**	**31.5**	**26.0**	**28.0**	**33.0**

4.6. Discussions

The image classification model predicts a classification score on the whole image, while the object detection and semantic segmentation models focus on the localisation and classification of the objects in the image. Hence, it is undoubtedly difficult to attack the object detection and semantic segmentation models using the adversarial examples generated by the image classification model. However, as we describe in Section 3.2, the domain-invariant features facilitate the attack by our model. Figure 6 shows the sample results of object detection and semantic segmentation models. We are surprising to find that compared to other methods, DMA can interfere with the results of the model by adding semantics and objects. The addition of semantics and objects, in a real CV system, is enough to create a barrier to recognition. However, the victim models still have a strong ability to detect the original semantics and objects, which is a limitation of DMA. Future work can revolve around how to reduce the detection of benign semantics and object by models.

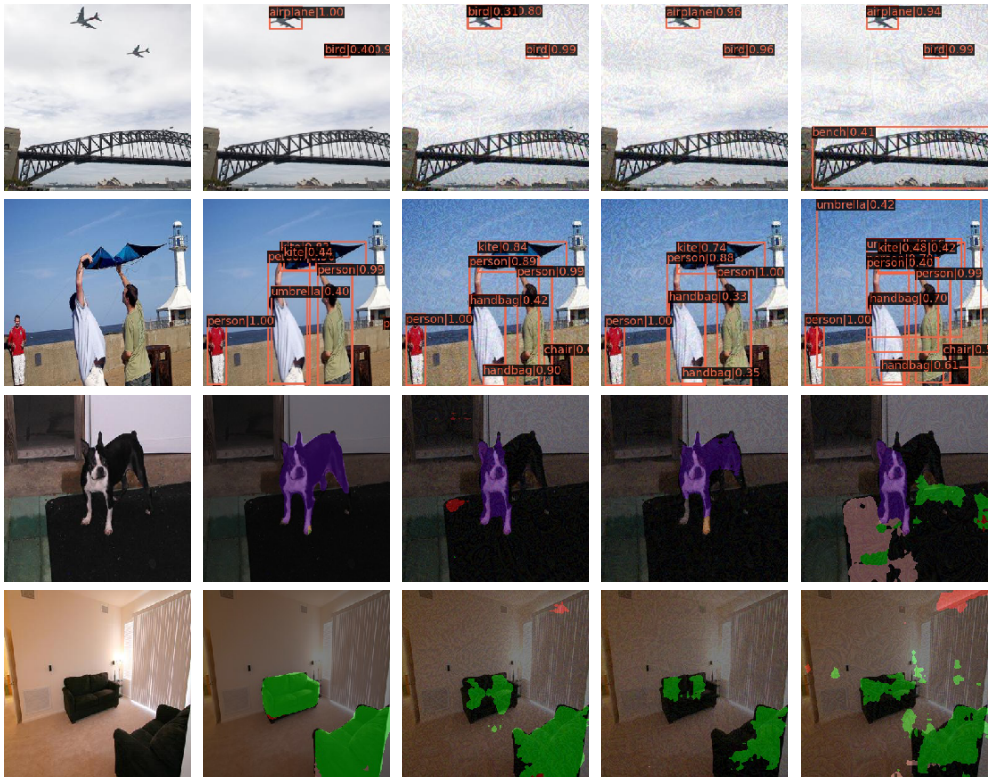

Figure 6. The samples of object detection and semantic segmentation.

5. Conclusions

In this paper, we extend the transferability of adversarial examples to the domain-invariant transferability (both the task-specific transferability and the cross-task transferability) of adversarial examples. Relying on the well-generalised features pre-trained on ImageNet, we propose the distance metric attack (DMA) method, which maximises the distance of the latent features between the adversarial example and the benign example. The adversarial examples crafted by DMA are highly transferable to various models on different tasks. Extensive experiments on image classification, object detection, and semantic segmentation indicate that the model robustness is highly related to the model structure. In addition, it is demonstrated that DMA can improve the success rate of black-box attack by more than 10% on specific-tasks and by more than 5% on cross-tasks compared with the state-of-the-art competitors.

Author Contributions: Conceptualisation, J.Z. and W.P.; methodology, W.P.; software, J.Z.; validation,J.Z., W.P. and Y.L.; formal analysis, G.L.; investigation, G.L.; resources, Y.L.; data curation, Y.L.; writing—original draft preparation, W.P.; writing—review and editing, R.W.; visualisation, J.Z.; supervision, W.Z.; project administration, W.Z.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China under Grant 62162067, 62101480, 61762089, 61763048, in part by the Yunnan Province Science Foundation for Youths under Grant No.202005AC160007, and in part by the fundamental research plan of "Release Management Service" in Yunnan Province: Research on Multi-source Data Platform and Situation Awareness Application for Cross-border Cyberspace Security (No. 202001BB050076).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
2. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
3. Liu, Y.; Chen, X.; Liu, C.; Song, D. Delving into Transferable Adversarial Examples and Black-box Attacks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
4. Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. Synthesizing Robust Adversarial Examples. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 284–293.
5. Yang, P.; Gao, F.; Zhang, H. Multi-Player Evolutionary Game of Network Attack and Defense Based on System Dynamics. *Mathematics* **2021**, *9*, 3014. [CrossRef]
6. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting Adversarial Attacks with Momentum. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 9185–9193.
7. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
8. Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; Yuille, A.L. Improving Transferability of Adversarial Examples With Input Diversity. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2730–2739.
9. Dong, Y.; Pang, T.; Su, H.; Zhu, J. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4312–4321.
10. Lin, J.; Song, C.; He, K.; Wang, L.; Hopcroft, J.E. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
11. Zhou, W.; Hou, X.; Chen, Y.; Tang, M.; Huang, X.; Gan, X.; Yang, Y. Transferable Adversarial Perturbations. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11218, pp. 471–486.
12. Zhang, Y.; Li, Y.; Liu, T.; Tian, X. Dual-Path Distillation: A Unified Framework to Improve Black-Box Attacks. In Proceedings of the International Conference on Machine Learning (ICML), Virtual Event, 12–18 July 2020; pp. 11163–11172.
13. Ilyas, A.; Engstrom, L.; Athalye, A.; Lin, J. Black-box Adversarial Attacks with Limited Queries and Information. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 2142–2151.
14. Bhagoji, A.N.; He, W.; Li, B.; Song, D. Practical Black-Box Attacks on Deep Neural Networks Using Efficient Query Mechanisms. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11216, pp. 158–174.
15. Li, Y.; Yang, Y.; Zhou, W.; Hospedales, T.M. Feature-Critic Networks for Heterogeneous Domain Generalization. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 10–15 June 2019; Volume 97, pp. 3915–3924.
16. Peng, W.; Liu, R.; Wang, R.; Cheng, T.; Wu, Z.; Cai, L.; Zhou, W. EnsembleFool: A method to generate adversarial examples based on model fusion strategy. *Comput. Secur.* **2021**, *107*, 102317. [CrossRef]
17. Shang, Y.; Jiang, S.; Ye, D.; Huang, J. Enhancing the Security of Deep Learning Steganography via Adversarial Examples. *Mathematics* **2020**, *8*, 1446. [CrossRef]
18. Lu, Y.; Jia, Y.; Wang, J.; Li, B.; Chai, W.; Carin, L.; Velipasalar, S. Enhancing Cross-Task Black-Box Transferability of Adversarial Examples with Dispersion Reduction. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 937–946.
19. Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
20. Paluzo-Hidalgo, E.; Gonzalez-Diaz, R.; Gutiérrez-Naranjo, M.A.; Heras, J. Simplicial-Map Neural Networks Robust to Adversarial Examples. *Mathematics* **2021**, *9*, 169. [CrossRef]
21. Carlini, N.; Wagner, D.A. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
22. Moosavi-Dezfooli, S.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
23. Huang, Q.; Katsman, I.; Gu, Z.; He, H.; Belongie, S.J.; Lim, S. Enhancing Adversarial Example Transferability with an Intermediate Level Attack. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 4732–4741.
24. Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; Yuille, A.L. Adversarial Examples for Semantic Segmentation and Object Detection. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1378–1387.
25. Liu, X.; Yang, H.; Liu, Z.; Song, L.; Chen, Y.; Li, H. DPATCH: An Adversarial Patch Attack on Object Detectors. In Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, HI, USA, 27 January 2019; Volume 2301.

26. Thys, S.; Ranst, W.V.; Goedemé, T. Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 49–55.
27. Xu, K.; Zhang, G.; Liu, S.; Fan, Q.; Sun, M.; Chen, H.; Chen, P.; Wang, Y.; Lin, X. Adversarial T-Shirt! Evading Person Detectors in a Physical World. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Volume 12350, pp. 665–681.
28. Xiao, C.; Deng, R.; Li, B.; Yu, F.; Liu, M.; Song, D. Characterizing Adversarial Examples Based on Spatial Consistency Information for Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11214, pp. 220–237.
29. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.J.; Boneh, D.; McDaniel, P.D. Ensemble Adversarial Training: Attacks and Defenses. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
30. Song, C.; He, K.; Lin, J.; Wang, L.; Hopcroft, J.E. Robust Local Features for Improving the Generalization of Adversarial Training. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
31. Guo, C.; Rana, M.; Cissé, M.; van der Maaten, L. Countering Adversarial Images using Input Transformations. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
32. Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; Yuille, A.L. Mitigating Adversarial Effects Through Randomization. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
33. Li, H.; Pan, S.J.; Wang, S.; Kot, A.C. Domain Generalization With Adversarial Feature Learning. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 5400–5409.
34. Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; Tao, D. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11219, pp. 647–663.
35. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
36. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; Volume 9908, pp. 630–645.
38. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
39. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
40. Lin, T.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.
41. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You Only Look One-level Feature. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 13039–13048.
42. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse R-CNN: End-to-End Object Detection With Learnable Proposals. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 14454–14463.
43. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
44. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
45. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
46. Zhu, Z.; Xu, M.; Bai, S.; Huang, T.; Bai, X. Asymmetric Non-Local Neural Networks for Semantic Segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 593–602.
47. Yuan, Y.; Chen, X.; Wang, J. Object-Contextual Representations for Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Volume 12351, pp. 173–190.
48. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 1971–1980.
49. MMSegmentation Contributors. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. 2020. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 20 April 2021).