Boosting the Adversarial Transferability of Surrogate Models with Dark Knowledge

Dingcheng Yang[1,2], Zihao Xiao[2], Wenjian Yu[1,*]
[1]Dept. Computer Science & Tech., BNRist, Tsinghua University, Beijing, China.
[2]RealAI.
ydc19@mails.tsinghua.edu.cn, zihao.xiao@realai.ai, yu-wj@tsinghua.edu.cn

***Abstract*—Deep neural networks (DNNs) are vulnerable to adversarial examples. And, the adversarial examples have transferability, which means that an adversarial example for a DNN model can fool another model with a non-trivial probability. This gave birth to the transfer-based attack where the adversarial examples generated by a ate model are used to conduct black-box attacks. There are some work on generating the adversarial examples from a given surrogate model with better transferability. However, training a special surrogate model to generate adversarial examples with better transferability is relatively under-explored. This paper proposes a method for training a surrogate model with dark knowledge to boost the transferability of the adversarial examples generated by the surrogate model. This trained surrogate model is named dark surrogate model (DSM). The proposed method for training a DSM consists of two key components: a teacher model extracting dark knowledge, and the mixing augmentation skill enhancing dark knowledge of training data. We conducted extensive experiments to show that the proposed method can substantially improve the adversarial transferability of surrogate models across different architectures of surrogate models and optimizers for generating adversarial examples, and it can be applied to other scenarios of transfer-based attack that contain dark knowledge, like face verification. Our code is publicly available at https://github.com/ydc123/Dark_Surrogate_Model.**

***Index Terms*—Deep learning, Image classification, Black-box adversarial attack, Transfer-based attack, Dark knowledge**

I. Introduction

Deep neural networks (DNNs) have achieved substantial success on many computer vision tasks. However, they are shown to be vulnerable to adversarial examples. Adversarial examples are carefully crafted data which could fool the DNNs by adding imperceptible noise on legitimate data. The generation of adversarial examples have been extensively researched in recent years. This is primarily due to its potential for preventing the malicious use of DNNs [1] and serving as a reliable evaluation criterion for DNN security [2], particularly in safety-critical scenarios like face verification.

The transferability of adversarial examples has attracted much attention. It means that, an adversarial example that fools one DNN model can fool another DNN model with a non-trivial probability. Thus, an adversary can train a surrogate model locally (training stage), and then generate adversarial examples to fool the surrogate model (generating stage). Finally, the generated adversarial examples can be directly used to attack an unknown black-box victim model (attacking stage). This process is called *transfer-based adversarial attack*.

The technique of adversarial example optimizer has been proposed for generating highly transferable adversarial examples [3], [4] (in generating stage). In contrast, we aim to train a better surrogate model (in training stage) so that it could yield adversarial examples with better success rates of transfer-based attacks. In analogy to the commonly used term "the transferability of adversarial example", we propose the concept "the adversarial transferability of surrogate models" to describe the ability of surrogate models on generating better adversarial examples for transfer-based attacks, using a fixed adversarial example optimizer. There are just a few works trying to train a surrogate model with better adversarial transferability [5], [6]. Specifically, adversarial training was used in [6] to improve the transferability, but at a significant computational cost. In [5], the knowledge distillation [7] was applied to improve the transferability without incurring excessive time overhead. However, it relies on a scheme of ensemble attack with multiple surrogate models and a combination of soft labels and one-hot labels.

Labels and data are two important components in training DNNs. Although the one-hot label is extensively used in normal DNN training, we notice that it does not well describe a data, because an image often contains the features of similar classes and even multiple objects in addition to the features of the true class. In contrast to one-hot labels, the soft labels which are the predicted distributions from a teacher model contain abundant information of image data, and have been used in knowledge distillation [7] for compressing neural networks. The soft label is also known as "dark knowledge" [7]. In this work, we propose that the dark knowledge is the key recipe to boost the adversarial transferability of surrogate models. Therefore, we propose to use the soft label to train the surrogate model, and enhance the dark knowledge by applying mixing augmentation skills to training data [8]–[10].

The surrogate model trained with dark knowledge is called "dark" surrogate model (DSM) in this work. The proposed method modifies the training stage, which enhances the dark knowledge by applying mixing augmentation to the training data and using soft labels extracted from a pretrained teacher model. We have conducted extensive experiments on attacking image classification models to show that the proposed method remarkably and consistently improves the adversarial

*corresponding author.

2375-0197/23/$31.00 ©2023 IEEE
DOI 10.1109/ICTAI59109.2023.00098

transferability of surrogate models. In addition, the proposed method can be applied to other transfer-based attack scenarios that contain dark knowledge, such as face verification, image retrieval, and text classification, to improve the success rate of the transfer-based attack. As an example, the experiments on applying DSM to attack face verification models are presented.

The major contributions and results are as follows.

- For improving the success rates of the transfer-based adversarial attack, we propose to use the dark knowledge during the training of the surrogate model, so as to obtain a "dark" surrogate model (DSM). The method for training the DSM is proposed, which modifies two key components of DNN training: labels and data, to make full usage of dark knowledge. Firstly, a pretrained DNN model, regarded as a teacher model, is employed to generate soft labels with dark knowledge. Secondly, the mixing augmentation skills are applied to enhance the dark knowledge of the training data explicitly.
- Extensive experiments on image classification are conducted to validate the proposed method. At first, the DSM is trained by using a pretrained model of the same architecture as the teacher model. Compared to directly using the pretrained model as the surrogate model, the proposed method improves the success rates of the untargeted attack by a TI-DI-MI-TG optimizer [11] on nine victim models by up to **21.6%**, **23.6%** and **11.0%** for the ResNet18, DenseNet121 and MobileNetv2 based surrogate models, respectively. Then, by using different teacher models, the maximum increments of attack success rate can be further improved to **25.7%**, **36.8%** and **26.3%**, respectively. Experimental results also validate that the proposed method performs better than the related work based on knowledge distillation [5].
- We have also applied the proposed method to the problem of attacking face verification models. On the widely-used ArcFace model [12], the proposed method improves the success rates of dodging attack by **12.9%** and impersonate attack by **16.2%**.

II. Related works

A DNN model for classification can be considered as a function $f(x;\theta) : \mathbb{R}^d \to \mathbb{R}^K$, where K is the number of classes, θ denotes all the parameters, $x \in \mathbb{R}^d$ is an image, d denotes the dimensionality of x, and the predicted label is $\text{argmax}_{1\le i\le K} f(x;\theta)_i$.

Given an image x and its corresponding label y, an untargeted adversarial example (the example which is misclassified) can be generated to fool a DNN model parameterized by θ through maximizing a cross-entropy loss function:

$$x^* = \text{argmax}_{x'} \mathbf{CE}(e_y, \mathbb{S}(f(x';\theta))) \text{ , s.t. } \|x' - x\| \le \epsilon \ , \quad (1)$$

where e_y denotes a one-hot vector with true label y, and the cross-entropy loss function $\mathbf{CE}(p,q)$ is defined as $\mathbf{CE}(p,q) = -\sum_i p_i \log q_i$. The softmax function $\mathbb{S} : \mathbb{R}^K \to \mathbb{R}^K$ is used to normalize the outputs of a DNN to a probability distribution, which means $\mathbb{S}(z)_i = \exp(z_i)/\sum_{j=1}^K \exp(z_j)$. The $\|\cdot\|$ denotes a norm function, and we focus on $\mathcal{L}_\infty$ norm in this paper. The ϵ is the maximum allowed magnitude of perturbation. The generated adversarial example x^* looks similarly to x but can fool the DNN model parameterized by θ (also called victim model).

However, the victim model is often inaccessible in practice. To attack a black-box victim model, we should first train a white-box surrogate model θ, and use it to generate the adversarial example x^*. This adversarial example is then directly used to attack the victim model. Normally, a surrogate model θ is trained by solving the following optimization problem with a stochastic gradient descent optimizer:

$$\theta = \text{argmin}_{\theta'} \mathbf{CE}(e_y, \mathbb{S}(f(x;\theta'))) \ . \quad (2)$$

Many optimizers were proposed for generating untargeted adversarial examples by solving (1), using a surrogate model θ that was trained by solving (2), such as the one-step method FGSM [13], the MI-FGSM [3] that utilizes a momentum factor μ, the diverse inputs method (DIM) [4] that augments the inputs with a pre-defined probability p_t, and the translation-invariant method (TIM) [14] that convolves the gradient with a pre-defined kernel W. Recently, a more powerful optimizer has been proposed that utilizes transformed gradient (TG) to generate adversarial examples, and its combination with the above methods leads to a more effective TI-DI-MI-TG method [11]. These methods can be easily extended for the targeted attack, i.e. generating an adversarial example x^* which is misclassified as a pre-defined target label y_t. It has been recently shown that targeted attacks can be boosted by a logits-based loss and running more iterations [15].

Unlike the rapidly developing adversarial example optimizers, only a few of works were devoted to training a better surrogate model for the transfer-based attack. They include the knowledge-distillation based method [5] and the recent work [6] showing that a slightly robust model has better adversarial transferability. Notice that, the method in [6] costs large computational time for training the slightly robust model.

In the knowledge-distillation based method [5], a surrogate model is distilled using multiple teacher models. This method is inspired by previous works on ensemble attack [16], which show that attacking multiple surrogate models simultaneously is more effective than attacking a single surrogate model. By distilling from multiple teacher models, the resulting surrogate model shares similar characteristics with all the teachers, thereby mimicking the ensemble attack. Specifically, suppose there are M teacher models $F_1, \cdots, F_M$, a surrogate model θ is trained by solving the following optimization problem:

$$\begin{aligned} &\theta = \text{argmin}_{\theta'} \mathbf{CE}(\tilde{y}, \mathbb{S}(f(x;\theta'))) \ , \\ &\text{where} \quad \tilde{y} = \frac{\beta_{KD}}{M} \sum_{i=1}^{M} \mathbb{S}(F_i(x)) + (1-\beta_{KD}) e_y \ . \end{aligned} \quad (3)$$

The $F_i(x)$ is the output of the i-th teacher for image x, $\tilde{y}$ is a soft label generated for training the surrogate model, and β_{KD} is a hyper-parameter.

There are several studies on data augmentation by mixing multiple data for image classification, including Cutout [8], Mixup [9] and CutMix [10]. In Section IV.B, we will show that after enriching the dark knowledge of the images by these skills, the adversarial transferability can be further improved when using a teacher model to extract dark knowledge from the augmented images. To the best of our knowledge, we are the first to use conventional data augmentations in the *training* stage to boost transfer-based attacks, as opposed prior works that augment data during the *generating* stage [4], [11], [14].

III. Methodology

In this section, we propose the dark surrogate model (DSM) to yield adversarial examples with better transferability, which is illustrated in Fig. 1. We first introduce our idea of refining labels with dark knowledge. Then, we apply mixing augmentations to enhance the dark knowledge of training data. Finally, we describe the algorithm for training the proposed DSM.

A. Refining Labels Using Dark Knowledge

Given an image x and its label y, the optimization problem (2) converges a minimum value only if the predicted distribution $\mathbb{S}(f(x;\theta))$ equals the one-hot label e_y, which means $f(x;\theta)_y - \max_{i\neq y} f(x;\theta)_i = \infty$, indicating that the trained surrogate model output an extremely high confidence score for the true class. However, the fitting target e_y does not describe an image well because an image often contains features of similar classes. For example, ImageNet [17], the most famous dataset for image classification, is a hierarchical dataset that contains many subcategories belonging to the category "dog", such as "papillon", "chihuahua", "Maltese dog". An image of "papillon" will have the features of other "dog" categories. Moreover, there may even be multiple objects appearing in an image. An example is illustrated in Fig. 1, where there are two images in the Raw Dataset labeled as "Persian cat" and "papillon", respectively, while they possess features of other cats, dogs, as well as pillows and cars. Even if the model achieves high accuracy on classification, the model trained with one-hot labels can not fully extract the features of an image for every class. This will be harmful for adversarial transferability of surrogate model, which directly depends on the working mechanism of the trained surrogate model, i.e. how it thinks "an image looks like a dog instead of a goldfish".

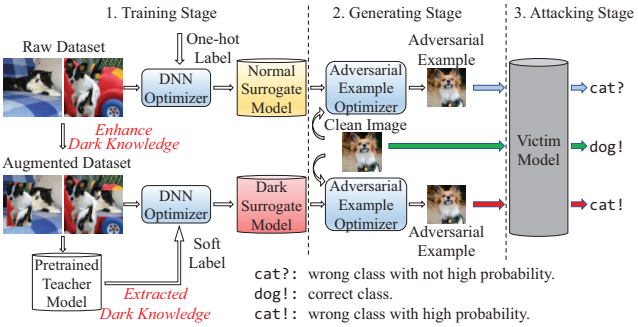


Fig. 1. An illustration of previous work and the proposed method for generating adversarial examples.

To overcome this weakness, we propose to leverage a pretrained teacher model to extract the dark knowledge from the training data, which is then utilized to train the surrogate model. Specifically, the predicted probability distribution of the teacher model serves as a soft label, which provides more information compared to the one-hot label, such as "which 2's look like 3's and which look like 7's" [7]. This information can help the surrogate model to learn image features better, and thus yield more transferable adversarial examples. Given a pretrained teacher model parameterized by θ_0, we can train a dark surrogate model parameterized by θ_d through solving the following optimization problem:

$$\theta_d = \text{argmin}_\theta \ \mathbf{CE}(\mathbb{S}(f(x;\theta_0)), \mathbb{S}(f(x;\theta))) \ . \tag{4}$$

The major difference to the normal training (2) is that the dark knowledge $\mathbb{S}(f(x;\theta_0))$ produced by the teacher model is used as the label.

Our work shares a similar process with the previous work [5], but we are motivated by a different goal and complement their work. Specifically, the objective in [5] is to make the surrogate model similar to multiple teacher models, and the success of [5] is entirely attributed to the ensemble attack [16], without any discussion on dark knowledge. This makes the method in [5] incomplete and lack theoretic support for the case of using a single teacher model. However, experimental results to be presented in Section IV show that using only a single teacher model can also improves the adversarial transferability. This improvement can not be explained by [5], but we provide a theoretical interpretation for this phenomenon based on the concept of dark knowledge. Additionally, while one-hot labels (e_y term in (3)) are still employed in [5], our perspective based on dark knowledge indicates that they are harmful for adversarial transferability and thus we do not use them. In Section IV.C, we will conduct a comparison experiment with [5] to highlight the differences between the two works.

B. Enhancing Dark Knowledge of Training Data

Although the soft label in (4) involves dark knowledge and thus is better than the one-hot label, it is still close to the one-hot label since the teacher model is obtained by training with the one-hot labels. To illustrate this point, we denote the confidence of a soft label $\tilde{y}$ as $\max_{i=1}^{K} \tilde{y}_i$. Thus, the confidence of a one-hot label achieves the maximum value of 1. Then, we train a ResNet18 teacher model on CIFAR-10 and provide a visualization of the empirical cumulative distribution function (CDF) of the confidences of soft labels it generates on the training images of CIFAR-10, as shown in Fig. 2. Fig. 2 shows that the empirical CDF of confidences on CIFAR-10 (red curve) is similar to the CDF of one-hot labels, namely a straight line with $x = 1$, which greatly weakens the effect of dark knowledge on boosting adversarial transferability.

To overcome this weakness, we propose to enhance the dark knowledge of training data by leveraging the data augmentation skills which explicitly mix a pair of images to synthesize image data containing features of different classes. Given an

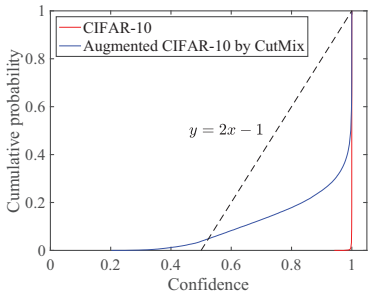
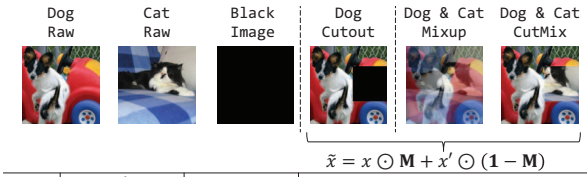


Fig. 2. The empirical cumulative distribution function (CDF) of confidence of soft labels generated by an ResNet18 teacher model.



Method	x'	$\mathbf{M}$	Pseudo Label
Cutout	Black Image	Binary mask	$\delta(x)$
Mixup	Another Image	$\lambda \cdot \mathbf{1}, \lambda \in [0,1]$	$\text{mean}(\mathbf{M}) \cdot \delta(x) + (1 - \text{mean}(\mathbf{M})) \cdot \delta(x')$
CutMix	Another Image	Binary mask	$\text{mean}(\mathbf{M}) \cdot \delta(x) + (1 - \text{mean}(\mathbf{M})) \cdot \delta(x')$

Fig. 3. An illustration of different data augmentation skills. The details of these skills are explained at bottom. $\delta(x)$ denotes the one-hot label of x.

original image x, we consider three popular skills of mixing augmentations in this work:

- Cutout [8], which randomly masks a fixed-size area of x to zero. The size of mask is set to 112×112 in this work.
- Mixup [9], which randomly samples a reference image x' and make a combination with x. This generates an image $\tilde{x} = \lambda x + (1 - \lambda)x'$, where $\lambda \sim U(0, 1)$ in this work. For this data $\tilde{x}$, a soft label $\tilde{y} = \lambda e_y + (1 - \lambda)e_{y'}$ should be used during the training, where y and y' are the true classes of x and x', respectively.
- CutMix [10], which randomly copies a rectangle area of x' to paste into x. If the area ratio of the rectangle to the whole image is $1 - \lambda$ where $\lambda \sim U(0, 1)$ in this work, a soft label $\tilde{y} = \lambda e_y + (1 - \lambda)e_{y'}$ is used for training, and y and y' are the true classes of x and x', respectively.

The data generated with the three mixing augmentation skills can be unified as $\tilde{x} = x \odot \mathbf{M} + x' \odot (\mathbf{1} - \mathbf{M})$, where $\mathbf{M}$ is a tensor of the same shape as x, $\odot$ is an element-wise product, and $\mathbf{1}$ is an all-one tensor. Fig. 3 illustrates these skills.

To demonstrate the effectiveness of mixing augmentation in mitigating the issue of soft labels being too similar to one-hot labels, we take CutMix as an example and apply it to the training images of CIFAR-10 to generate 50000 augmented images. We then plot the CDF of the confidences of generated soft labels on these augmented images in the blue curve of Fig. 2, which exhibit clear deviation from the red curve over about 40% data points, demonstrating that the mixing augmentations can construct data with more dark knowledge.

Moreover, Fig. 2 also reveals a huge discrepancy between the CDF of confidence of soft labels generated by the ResNet18 teacher model and those for the standard CutMix skill. Specifically, the soft labels employed in the standard CutMix is $\tilde{y} = \lambda e_y + (1 - \lambda)e_{y'}$ for an augmented image obtained by mixing by two images that belong to the category of y and y', respectively, where $\lambda \sim U(0, 1)$. As a result, the confidence of the soft label is $\max(\lambda, 1 - \lambda)$ and should obey a uniform distribution on $[0.5, 1]$, with a CDF of $y = 2x - 1$, which is completely different from the trend of the blue curve depicted in Fig. 2. This difference raises the question of whether the heuristic labeling strategy employed in CutMix is a reasonable strategy for improving adversarial transferability. Although CutMix has been shown to be effective on image classification, as discussed in Section III.A, higher classification accuracy does not necessarily correspond to better adversarial transferability. Notably, the labeling strategy of CutMix skill has a similar limitation to one-hot labels as it only considers the features of the categories to which the mixed images belong, while ignoring those of most other categories. This limitation is not aligned with the motivation of this work and can also be observed in other mixing augmentation skills. Therefore, CutMix and other mixing augmentation skills used in our method are only to enhance the dark knowledge of training data. The soft labels are obtained through the method introduced in Section III.A, specifically, by using the predicted probability distribution of the teacher model for the augmented images. This step is also illustrated in Fig. 1 and will be described as Algorithm 1 in Section III.C.

The experimental results in Section IV.B will show that the adversarial transferability of DSMs can be further improved by enhancing the dark knowledge of the training data with these mixing augmentations. Furthermore, Section IV.B also shows that the standard mixing augmentation will impair the adversarial transferability of surrogate models. It is noteworthy that the only difference between training with the standard mixing augmentation skills and the proposed DSM is the presence or absence of dark knowledge in the labeling strategy. This finding further emphasizes the crucial role of dark knowledge in boosting adversarial transferability.

C. The Proposed Algorithm for Training DSM

Combining the ideas in last two subsections, we propose the approach of training the DSM to boost the adversarial transferability of surrogate models, described as Algorithm 1. At each iteration during the training, we first enhance the dark knowledge by the mixing augmentation skills (Step 6), then train the surrogate model with dark knowledge extracted by the teacher model described with parameters θ_0 (Step 7).

Notice that any pretrained model for the same classification problem can be used as the teacher model. A simple choice of teacher model is the one with the same architecture as the DSM θ_d and trained by solving problem (2). Section IV.B will show that a teacher model with a different architecture from the DSM is also useful and sometimes makes the DSM exhibit better adversarial transferability of surrogate models. In addition, the proposed approach can be naturally combined with prior work on improving the adversarial transferability of surrogate models, through using their released model as a teacher model, as shown in Section IV.D.

Algorithm 1 Training the DSM for transfer-based attack

Input: Batch size m, learning rate η, training dataset $\mathcal{D}_{\mathcal{T}}$, a (pretrained) teacher model with parameters θ_0, dark surrogate DNN model parameterized by θ_d.
Output: the dark surrogate model with parameters θ_d.

Randomly initialize the parameters θ_d.
repeat ▷ solve the optimization problem (8)
  Read mini-batch $\{x_1, \cdots, x_m\}$ from $\mathcal{D}_{\mathcal{T}}$.
  $L \leftarrow 0$.
  for $i \leftarrow 1$ to m **do**
    Apply the mixing augmentation on x_i to obtain an augmented image x_i^{mix}.
    $L \leftarrow L + \mathbf{CE}(\mathbb{S}(f(x_i^{mix}; \theta_0)), \mathbb{S}(f(x_i^{mix}; \theta_d)))$
  end for
  $\theta_d \leftarrow \theta_d - \eta \nabla_{\theta_d} L$
until parameters θ_d are converged

Finally, the proposed approach can be applied to other scenarios of transfer-based attack that contain dark knowledge, like face verification. Training a face verification model consists of two steps, i.e., training a facial classifier and obtaining an embedding model based on that classifier. An adversary can train a facial classifier based on Algorithm 1 to obtain an embedding model. The obtained embedding model can be used as a surrogate to attack a black-box face verification model. We will show that the facial classifier trained by the proposed approach yields an embedding model with better adversarial transferability, with the experiments presented in Section IV.E.

IV. Experimental Results

In this section, we demonstrate the effectiveness of the proposed dark surrogate model (DSM) with experiments. We begin by evaluating the effectiveness of DSM on attacking image classification models. The ResNet18 (RN18) [18], DenseNet121 (DN121) [19] and MobileNetv2 (MNv2) [20] are chosen as the surrogate models. Unless explicitly stated, the CutMix skill is used for training DSM, and the teacher model employed is a normal pretrained model (trained with one-hot labels) with the same architecture as the DSM. The first three subsections are dedicated to the untargeted attack of image classifiaction models, including the comparison with the related work [5]. Then, the results of targeted attack are presented, which involves the combination of the proposed DSM and the method of training slightly robust model [6]. Lastly, the results on attacking face verification models are presented.

Adversarial examples are crafted with a maximum perturbation of $\epsilon = 16$ unless explicitly stated. We consider three adversarial example optimizers: FGSM [13], MI-FGSM [3], and TI-DI-MI-TG [11]. If the optimizer is not explicitly stated, the strongest TI-DI-MI-TG optimizer is employed. For the hyper-parameters of adversarial example optimizer, we set the step size β to 2, the momentum factor μ to 1.0, the probability of transformation p_t to 1.0, the size of kernel W to 7×7, the number of iterations N for untargeted attack to 10, following [11]. For targeted attack we set N to 200 and optimize the logits-based loss following the suggestion in [15].

Two image classification datasets are considered, including the small CIFAR-10 dataset and the large ImageNet dataset. In the experiments of CIFAR-10, the models are trained 200 epochs. We set the batch size to 256, the weight decay to 10^{-4}. The learning rate is set to 0.1 and updated by a cosine annealing scheduler. In the experiments of ImageNet, we follow the PyTorch official example[1] to train the models. We randomly sample 1000 images from CIFAR-10 dataset for generating adversarial examples, while for the experiments of ImageNet, the adversarial examples are generated on ImageNet-compatible dataset[2] since it was widely used in previous works [3]. This dataset comprises 1000 images and provides a true label and a target label of each image for untargeted and targeted attack, respectively.

A. Detailed Results of the Proposed Method

We first preliminarily show the performance of DSM on a simple dataset, CIFAR-10. Specifically, we train three normal surrogate models, ResNet18 (RN18), DenseNet121 (DN121) and MobileNetV2 (MNV2), on CIFAR-10, and then employ them as the teacher models to train DSMs with the same architecture. We denote a dark RN18 model trained without mixing augmentation as DSM(RN18, None), and a dark RN18 model trained with CutMix as DSM(RN18, CutMix). When there is no ambiguity, DSM(RN18, CutMix) is abbreviated as DSM(RN18), as CutMix is the default augmentation technique used in this paper. The naming convention of other models is done similarly. We employ the FGSM optimizer to efficiently generate adversarial examples by these normal/dark surrogate models, and use these normally trained models as victim models to evaluate the adversarial transferability. We list the experimental results in Table I, which first show that the dark surrogate models are only slightly better than or comparable to normal surrogate models when no mixing augmentation is used. This is because the dark knowledge extracted by the teacher models is too similar to the one-hot labels, as shown in Fig. 2. Then, Table I demonstrates that the attack success rates are remarkably improved by up to **27.0%** when CutMix is used to enhance the dark knowledge of the training data, emphasizing the crucial role of mixing augmentation.

Then, we consider a more challenging dataset, ImageNet, and used nine publicly available models as victim models. These models have been widely used in previous work [3]. The first three of them are normally trained models, including Inception-v3 (Inc-v3) [21], Inception-v4 (Inc-v4), and Inception-ResNet-v2 (IncRes-v2) [22]. The rest are robust models: Inc-v3$_{ens3}$, Inc-v3$_{ens4}$, and IncRes-v2$_{ens}$ [23], high-level representation guided denoiser (HGD) [24], input transformation through resizing and padding (R&P) [25], and the rank-3 submission in NIPS2017 adversarial competition

[1] https://github.com/pytorch/examples/tree/master/imagenet
[2] https://www.kaggle.com/google-brain/nips-2017-adversarial-learning-development-set

TABLE I
THE SUCCESS RATES (%) OF UNTARGETED ATTACKS ON CIFAR-10 DATASET. – INDICATES THE WHITE-BOX ATTACK.

Surrogate model	RN18	DN121	MNv2
Normal RN18	-	73.0	70.0
DSM(RN18,None)	-	74.1	68.2
DSM(RN18,CutMix)	-	**90.7**	**87.8**
Normal DN121	45.2	-	65.1
DSM(DN121,None)	46.2	-	65.5
DSM(DN121,CutMix)	**72.2**	-	**88.4**
Normal MNv2	38.2	60.6	-
DSM(MNv2,None)	41.3	59.7	-
DSM(MNv2,CutMix)	**64.2**	**85.7**	-

(NIPS-r3)[3]. To demonstrate the effectiveness of the proposed DSM across various adversarial example optimizers, we consider three optimizers, namely FGSM, MI-FGSM and TI-DI-MI-TG. The stronger TI-DI-MI-TG will be used as the default optimizer in the later section. We list the untargeted attack results in Table II, which shows that the proposed DSMs consistently outperform the normal surrogate models with same architecture. Notice that TI-DI-MI-TG represents a state-of-the-art method for generating adversarial examples without training a special surrogate model. Compared with it, using the three DSMs based on RN18, DN121 and MNV2 can improve the attack success rate by 10.7%-21.6%, 12.8%-23.6% and 8.5%-11.0%, respectively.

B. Ablation Studies

In this subsection, we first conducted experiments using teacher models with different architectures to investigate the effect of the teacher model and report the results in Table III. Notice that the results for the DSM sharing same architecture as the teacher model are the same as those in Table II for TI-DI-MI-TG. From Table III we see that using different teacher model may further improve the attack success rates. Comparing the results in Table II, we find out that DSMs can improve the attack success rates by **25.7%**, **36.8%** and **26.3%** at most for the situations with RN18, DN121 and MNV2 based surrogate models, respectively. Although it is still an open problem that what teacher model is best for the adversarial transferability of DSM, just using the teacher model with the same architecture as DSM is a simple yet effective choice.

We proceeded to investigate the impact of different mixing augmentations on adversarial transferability. Note that some preliminary results in this aspect were previously shown in Table I. Here, we additionally consider two commonly used mixing augmentation skills, namely Cutout and Mixup. The RN18 is considered as the architecture of the surrogate model. The results are shown in Fig. 4, which shows that all mixing augmentation skills can improve the attack success rates. It is noteworthy that the improvement is smaller for ImageNet-compatible dataset than for CIFAR-10 (Table I) due to the greater complexity of ImageNet. Consequently, the output of the teacher model does not degenerate to one-hot labels as depicted in Fig. 2. Nevertheless, the use of mixing augmentation consistently improves adversarial transferability in all cases, with negligible additionally computational overhead.

[3]https://github.com/anlthms/nips-2017/tree/master/mmd

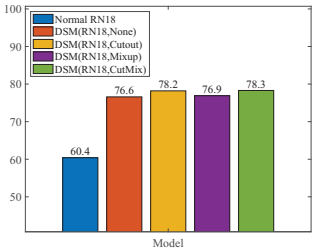


Fig. 4. The average success rates (%) of untargeted attack on ImageNet-compatible dataset against nine victim models when using different mixing augmentations to train a (dark) RN18.

Since the labeling strategy in DSM is different from the standard mixing augmentation skills, we also conduct experiments on a RN18 surrogate model trained with standard Cutout/Mixup/CutMix skills. Note that the only difference between a normal RN18 trained with standard mixing augmentation skill and a dark RN18 trained with the same mixing augmentation skill is the labeling strategy. The results of the RN18 models trained with these skills are presented in Table IV. Surprisingly, we observed that the normal RN18 achieves better results than the models trained with mixing augmentation skills, indicating that such skills actually impairs the adversarial transferability of surrogate model when there is no teacher model to extract dark knowledge. This finding highlights the crucial role of extracted dark knowledge in boosting adversarial transferability.

C. Comparison with the Knowledge Distillation Based Method [5]

In this subsection, we compare the proposed DSM with the knowledge-distillation based method [5], since it also uses knowledge distillation to train the surrogate model. We refer to the model trained in this way as KDSM (knowledge-distillation based surrogate model). If the number of teacher models $M = 1$, the difference between our DSM and KDSM is that the latter uses one-hot labels to interpolate with dark knowledge (see (3)), while the former additionally uses the mixing augmentations to enhance dark knowledge. Using normal RN18 as the teacher model to train a surrogate model in same architecture, the results of KDSM and our DSM are shown in Fig. 5. When $\beta_{KD} = 0$, the KDSM degenerates to a normally trained model, and when $\beta_{KD} = 1$ the KDSM is equivalent to our DSM without mixing augmentation. Fig. 5 shows that the attack success rate increases as β_{KD} increases, and it reaches the maximum at $\beta_{KD} = 1$. This means the DSM is better than KDSM, and it is unnecessary and inadvisable to interpolate with the one-hot labels. Furthermore, as Fig. 4 once demonstrated, the dash line in Fig. 5 shows better results for DSMs with mixing augmentation. The effectiveness of mixing augmentation to enhance the dark knowledge is also observed in other dataset. For example, the results presented in Table I demonstrate that mixing augmentation improves the attack success rates by up to **27.0%** on CIFAR-10, while there is almost no improvement without using mixing augmentation. Additionally, Section IV.E will verify the effectiveness of mixing augmentation in attacking face verification models.

TABLE II
THE SUCCESS RATES (%) OF UNTARGETED ATTACKS ON IMAGENET-COMPATIBLE DATASET WITH DIFFERENT OPTIMIZERS.

Optimizer	Surrogate model	Inc-v3	Inc-v4	IncRes-v2	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&D	NIPS-r3
FGSM [13]	normal RN18	47.3	40.7	33.8	33.0	34.6	23.1	26.2	24.9	26.2
	DSM(RN18)	**56.3**	**48.4**	**43.6**	**40.0**	**42.1**	**29.2**	**34.2**	**32.2**	**34.7**
	normal DN121	44.6	39.3	34.4	31.0	32.2	22.1	24.0	23.8	25.2
	DSM(DN121)	**54.1**	**47.9**	**43.0**	**40.0**	**39.9**	**29.3**	**33.5**	**31.1**	**32.9**
	normal MNv2	42.4	34.2	28.2	26.8	28.3	17.9	18.4	19.2	22.3
	DSM(MNv2)	**46.0**	**40.0**	**32.7**	**30.2**	**30.8**	**20.9**	**21.5**	**22.1**	**24.7**
MI-FGSM [3]	normal RN18	62.0	54.2	43.7	39.2	39.0	26.2	35.2	28.7	31.7
	DSM(RN18)	**80.9**	**71.5**	**66.3**	**56.8**	**56.0**	**40.2**	**56.1**	**43.0**	**48.3**
	normal DN121	58.3	52.8	49.2	38.1	38.6	27.1	38.1	29.6	31.2
	DSM(DN121)	**79.7**	**77.4**	**70.4**	**55.1**	**51.9**	**40.7**	**56.9**	**41.8**	**46.3**
	normal MNv2	49.2	42.5	34.6	30.5	30.3	20.2	25.0	21.4	25.9
	DSM(MNv2)	**59.2**	**50.2**	**42.4**	**36.4**	**37.0**	**24.3**	**30.6**	**27.2**	**30.3**
TI-DI-MI-TG [11]	normal RN18	81.7	75.3	66.4	58.9	58.8	42.8	55.8	49.5	54.4
	DSM(RN18)	**92.4**	**89.9**	**84.0**	**78.4**	**76.1**	**63.0**	**77.4**	**69.1**	**74.1**
	normal DN121	79.8	75.5	70.2	56.8	54.4	44.1	58.1	48.7	52.1
	DSM(DN121)	**92.6**	**92.1**	**89.6**	**77.9**	**75.5**	**63.0**	**81.7**	**69.2**	**74.4**
	normal MNv2	72.2	65.3	59.5	49.4	49.6	34.4	43.9	38.7	44.8
	DSM(MNv2)	**82.4**	**75.3**	**70.0**	**59.7**	**59.0**	**43.4**	**54.9**	**47.2**	**54.9**

TABLE III
THE SUCCESS RATES (%) OF UNTARGETED ATTACKS USING THE DSMS WITH VARIOUS TEACHER MODELS ON IMAGENET-COMPATIBLE DATASET.

Surrogate model	Teacher model	Inc-v3	Inc-v4	IncRes-v2	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&D	NIPS-r3
DSM(RN18)	RN18	92.4	89.9	84.0	78.4	76.1	63.0	77.4	69.1	74.1
	DN121	90.5	87.4	80.5	70.3	67.9	55.4	70.3	60.1	64.8
	MNv2	**95.0**	**91.3**	**87.6**	**83.4**	**80.2**	**66.9**	**81.5**	**71.3**	**77.6**
DSM(DN121)	RN18	**97.8**	**95.9**	**94.5**	**89.7**	**88.9**	**80.9**	**91.0**	**85.4**	**87.6**
	DN121	92.6	92.1	89.6	77.9	75.5	63.0	81.7	69.2	74.4
	MNv2	96.4	95.6	92.5	89.1	86.7	75.0	88.0	80.3	85.2
DSM(MNv2)	RN18	**89.5**	**86.7**	**80.7**	**71.3**	**67.9**	**55.6**	**70.2**	**60.6**	**67.0**
	DN121	82.3	78.6	68.9	57.9	56.1	41.9	53.1	44.9	52.5
	MNv2	82.4	75.3	70.0	59.7	59.0	43.4	54.9	47.2	54.9

TABLE IV
THE SUCCESS RATES (%) OF UNTARGETED ATTACK ON IMAGENET-COMPATIBLE DATASET WHEN USING DIFFERENT STRATEGIES.

Surrogate model	Inc-v3	Inc-v4	IncRes-v2	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&D	NIPS-r3
Normal RN18	**81.7**	**75.3**	**66.4**	**58.9**	**58.8**	**42.8**	**55.8**	**49.5**	**54.4**
RN18+Cutout	79.5	73.5	65.8	58.4	55.7	41.9	55.0	45.9	50.3
RN18+Mixup	77.8	72.0	62.6	53.7	50.5	36.6	47.5	39.1	42.0
RN18+CutMix	74.0	67.0	57.4	51.7	50.6	37.1	46.5	40.0	43.7

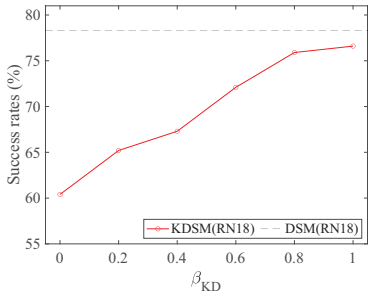


Fig. 5. The average success rates (%) of untargeted attack on ImageNet-compatible dataset against nine victim models with the KDSM [5] as the surrogate model, versus the value of β_{KD} in (3). The dash line indicates the result of DSM.

D. The Results of Targeted Attack

We have shown that the proposed DSM performs very well on untargeted attack. While for the much more difficult task of targeted attack, a state-of-the-art work is the method with slightly robust surrogate model [6] despite that it costs large computational time for generating adversarial examples online. Below, we will show that it can be further improved with the proposed DSM while almost not inducing extra time cost. The combination of DSM and the method with slightly robust surrogate model [6] can make a favorable success rate of targeted attack. In [6], it is demonstrated that by using a slightly robust model trained with small-magnitude adversarial examples as the surrogate model, the state-of-the-art success rates on targeted attack are achieved. Specifically, using N rounds of iteration to generate adversarial examples would make the training time to be $N+1$ times that of standard training, where N is about 10 as suggested by [26]. We use the slightly robust ResNet18 model (SR-RN18), which was trained with maximum perturbation of 0.1 (the recommend valued in [6]), as the teacher model to train a DSM denoted by DSM(SR-RN18), taking about the same time as standard training. We conduct experiments of targeted attacks and report the results in Table V, which shows that the proposed DSM can be naturally combined with [6] and it again remarkably improves the success rates of black-box attack by 6.9%-18.1%.

TABLE V
THE SUCCESS RATES (%) OF TARGETED ATTACKS ON IMAGENET-COMPATIBLE DATASET.

Surrogate model	Inc-v3	Inc-v4	IncRes-v2	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&D	NIPS-r3
SR-RN18 [6]	46.2	41.6	49.0	38.3	40.4	34.2	49.5	39.0	41.0
DSM(SR-RN18)	**63.1**	**59.7**	**65.5**	**48.9**	**49.0**	**41.1**	**62.2**	**48.2**	**51.7**

TABLE VI
THE SUCCESS RATES (%) OF THE DODGING/IMPERSONATE ATTACKS TO FACE VERIFICATION MODELS ON LFW DATASET.

Surrogate model	Dodging attack				Impersonate attack			
	FaceNet	SphereFace	CosFace	ArcFace	FaceNet	SphereFace	CosFace	ArcFace
IR50	79.2	95.6	93.2	77.4	45.4	84.5	76.3	60.6
DSM(IR50,None)	86.2	97.7	96.2	84.2	53.6	88.6	82.0	69.9
DSM(IR50,CutMix)	**92.5**	**99.4**	**98.8**	**90.3**	**63.0**	**93.8**	**87.2**	**76.8**

E. Application to Attacking Face Verification Model

DNN models for face verification have been widely deployed in many safety-critical scenarios like mobile unlocking. To show the versatility of the proposed method, we present the experimental results on attacking face verification models in this subsection. A face verification model is used to judge whether a pair of facial images belong to the same identity. It is built based on a classifier trained on a dataset of facial images to separate images of different identities. Given a pair of facial images, a pair of embedding features are extracted by the classifier, i.e. the outputs of the penultimate layer of the model. Then, the cosine similarity between them is calculated for judging whether they belong to the same identity.

The dodging attack and impersonate attack are two kinds of attack to face verification model. Given a pair of facial images x and x_r belonging to the same identity, dodging attack aims to generate an adversarial example x^{adv} which is similar to x but be recognized as a different identity from x_r. On the contrary, impersonate attack aims to generate an adversarial example x^{adv} which is similar to x but be recognized as the same as x_r if x and x_r do not belong to the same identity. We conduct experiments on the standard LFW [27] protocol, which means we select both 3000 pairs of images for dodging attack and impersonate attack. The IResNet50 (IR50) [28] is chosen as the surrogate model and four publicly available face verification models as the victim models are considered, including FaceNet, SphereFace, CosFace and ArcFace since they have different architectures and are considered in prior works on attacking face verification models [29].

We train an IR50 classifier on CASIA-WebFace [30] following previous work [29], and use it as a teacher model to train the dark surrogate model. We conduct dodging/impersonate attack experiments on them with $\epsilon = 8$ and list the results in Table VI, which shows that adversarial transferability can be remarkably improved through using dark knowledge, and can be further improved by introducing CutMix. Specifically, the proposed DSM can improve the success rates of dodging attack and impersonate attack by **12.9%** and **16.2%** respectively, when the ArcFace [12] is used as the victim model.

V. CONCLUSIONS

In this paper, we propose a method to train the surrogate model for transfer-based adversarial attack on image classification, which boosts the adversarial transferability of surrogate models. The trained surrogate model is named dark surrogate model (DSM). The proposed method includes two key components: using a teacher model to generate dark knowledge (soft label) for training the surrogate model, and using the mixing augmentation skills to enhance the dark knowledge of training data. The effectiveness of the proposed method is validated by extensive experiments and the comparisons with counterparts. Besides, we show that the proposed method can be extended to other transfer-based attack scenarios that contain dark knowledge, like face verification.

VI. ACKNOWLEDGMENT

This work was supported by the National Key Research and Development Plan of China (2020AAA0103502), and National Key Research and Development Project of China (No. 2021ZD0110502).

REFERENCES

[1] C. Shi, X. Xu, S. Ji, K. Bu, J. Chen, R. Beyah, and T. Wang, "Adversarial captchas," *IEEE transactions on cybernetics*, vol. 52, no. 7, pp. 6095–6108, 2021.
[2] Z. Zhao, H. Zhang, R. Li, R. Sicre, L. Amsaleg, and M. Backes, "Towards good practices in evaluating transfer adversarial attacks," *arXiv preprint arXiv:2211.09565*, 2022.
[3] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *CVPR*, 2018, pp. 9185–9193.
[4] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *CVPR*, 2019, pp. 2730–2739.
[5] W. Cui, X. Li, J. Huang, W. Wang, S. Wang, and J. Chen, "Substitute model generation for black-box adversarial attack based on knowledge distillation," in *ICIP*, 2020, pp. 648–652.
[6] J. Springer, M. Mitchell, and G. Kenyon, "A little robustness goes a long way: Leveraging robust features for targeted transfer attacks," *NeurIPS*, vol. 34, 2021.
[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
[8] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.
[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.
[10] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV*, 2019, pp. 6023–6032.
[11] Z. He, Y. Duan, W. Zhang, J. Zou, Z. He, Y. Wang, and Z. Pan, "Boosting adversarial attacks with transformed gradient," *Computers & Security*, vol. 118, p. 102720, 2022.
[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.
[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.
[14] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *CVPR*, 2019, pp. 4312–4321.
[15] Z. Zhao, Z. Liu, and M. Larson, "On success and simplicity: A second look at transferable targeted attacks," *NeurIPS*, vol. 34, 2021.
[16] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *ICLR*, 2017.
[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.
[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.
[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.
[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017.

[23] F. Tramèr, D. Boneh, A. Kurakin, I. Goodfellow, N. Papernot, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *ICLR*, 2018.
[24] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *CVPR*, 2018, pp. 1778–1787.
[25] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in *ICLR*, 2018.
[26] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, "Bag of tricks for adversarial training," in *ICLR*, 2020.
[27] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV*, 2016, pp. 630–645.
[29] X. Yang, D. Yang, Y. Dong, W. Yu, H. Su, and J. Zhu, "Delving into the adversarial robustness on face recognition," *arXiv preprint arXiv:2007.04118*, 2020.
[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.