Unrestricted Black-box Adversarial Attack Using GAN with Limited Queries

Dongbin Na, Sangwoo Ji, and Jong Kim

Pohang University of Science and Technology (POSTECH), Pohang, South Korea
{dongbinna,sangwooji,jkim}@postech.ac.kr

Abstract. Adversarial examples are inputs intentionally generated for fooling a deep neural network. Recent studies have proposed unrestricted adversarial attacks that are not norm-constrained. However, the previous unrestricted attack methods still have limitations to fool real-world applications in a black-box setting. In this paper, we present a novel method for generating unrestricted adversarial examples using GAN where an attacker can only access the top-1 final decision of a classification model. Our method, Latent-HSJA, efficiently leverages the advantages of a decision-based attack in the latent space and successfully manipulates the latent vectors for fooling the classification model.
With extensive experiments, we demonstrate that our proposed method is efficient in evaluating the robustness of classification models with limited queries in a black-box setting. First, we demonstrate that our targeted attack method is query-efficient to produce unrestricted adversarial examples for a facial identity recognition model that contains 307 identities. Then, we demonstrate that the proposed method can also successfully attack a real-world celebrity recognition service. The code is available at https://github.com/ndb796/LatentHSJA.

Keywords: Black-box adversarial attack, generative adversarial network, unrestricted adversarial attack, face recognition system

1 Introduction

Since state-of-the-art deep-learning models have been known to be vulnerable to adversarial attacks [20,40], a large number of defense methods to mitigate the attacks are proposed. These defense methods include adversarial training [32] and certified defenses [44,15]. Most previous studies have demonstrated their robustness against adversarial attacks that produce norm-constrained adversarial examples [32,44,15]. The common choices for the constraint are l_0, l_1, l_2, and l_∞ norms [10], because a short distance between two image vectors in a image space implies the visual similarity between them.

Recent studies show that adversarial examples can be legitimate even though the perturbation is not small norm-bounded. These studies have proposed unrestricted adversarial examples that are not norm-constrained but still shown as

Fig. 1. Showcases of our Latent-HSJA attack method against a facial identity recognition model. The first row shows target images and the second row shows source images. The adversarial examples in the last row are classified as target classes and require a feasible number of queries (only 20,000 queries).

natural images to humans [35,27,39]. For example, manipulating semantic information such as color schemes or rotation of objects in an image can cause a significant change in the image space while not affecting human perception. These unrestricted adversarial examples effectively defeat robust models to a norm-constrained perturbation [39,18]. However, only a few studies [43,27,39] have evaluated the effectiveness of unrestricted adversarial attacks for deep-learning models in a black-box setting.

In a black-box setting, an attacker can only access the output of a classification model. Real-world applications such as Clarifai and Google Cloud Vision provide only the top-k predictions of the highest confidence scores. Recent studies have proposed norm-constrained adversarial attack methods for the black-box threat models based on query-response to a classification model [9,26,14]. However, these black-box attacks have not yet successfully expanded to unrestricted adversarial examples. Although a few studies have demonstrated their unrestricted adversarial attacks in a black-box setting, their methods suffer from a large number of queries (more than hundreds of thousands of queries) compared with existing norm-based black-box attack methods [27] or only support an untargeted attack [43].

We propose a novel method, Latent-HSJA, for generating unrestricted adversarial examples using GAN in a black-box setting. To generate unrestricted adversarial examples, we utilize the disentangled style representations of StyleGAN2 [30]. Our method manipulates the latent vectors of GAN and efficiently leverages the decision-based attacks in a latent space. Especially, our targeted attack can be conducted with a target image classified as a target class and

a specific source image with limited queries (Figure 1). We mainly deal with the targeted attack because the targeted attack is more difficult to conduct and causes more severe consequences than an untargeted attack.

We show that the proposed method is query-efficient for the targeted attack in a hard-label black-box setting, where an attacker can only access the predicted top-1 label. It is noted that a black-box attack method should be query-efficient due to the high cost of a query (0.001$ in Clarifai). The proposed method for the targeted attack is able to generate an unrestricted adversarial example with a feasible number of queries (less than 20,000 queries) for a facial identity recognition model that contains 307 identities. This result is comparable to that of state-of-the-art norm-constrained black-box attacks [26,9,14]. Especially, our method generates more perceptually superior adversarial examples than previous methods in a super-limited query setting (less than 5,000 queries). Moreover, we demonstrate that our method can successfully attack a real-world celebrity recognition service.

Our contributions are listed as follows:

- We propose a query-efficient novel method, Latent-HSJA, for generating unrestricted adversarial examples in a black-box setting. To the best of our knowledge, our method is the first to leverage the targeted unrestricted adversarial attack in a query-limited black-box setting.
- We demonstrate that the proposed method successfully defeats state-of-the-art deep neural networks such as a gender classification model, a facial identity recognition model, and the real-world celebrity recognition model with a limited query budget.

2 Related Work

2.1 Adversarial Examples

Many deep-learning applications have been deployed in security-important areas such as face recognition [6], self-driving car [21], and malware detection [4]. However, the recent deep neural network (DNN) models have been known to be vulnerable to adversarial examples [12,20,34]. A lot of studies have presented adversarial attack methods in various domains such as image, text, and audio applications [3,13,17]. The adversarial examples can be also used to improve the robustness of automated authentication systems such as CAPTCHAs [38,33].

One key aspect in categorizing the threat model of adversarial attacks is the accessibility to components of DNN models, known as white-box or black-box (Figure 2). In the white-box threat model, the attacker can access the whole information of a DNN model, including its weights and hyper-parameters. Many adversarial attacks assume the white-box threat model. The fast gradient sign method (FGSM) is proposed to generate an adversarial example by calculating the gradient only once [20]. The projected gradient descent (PGD) is then proposed to efficiently generate a strong adversarial example with several gradient calculation steps [32]. The CW attack is commonly used to find an adversarial

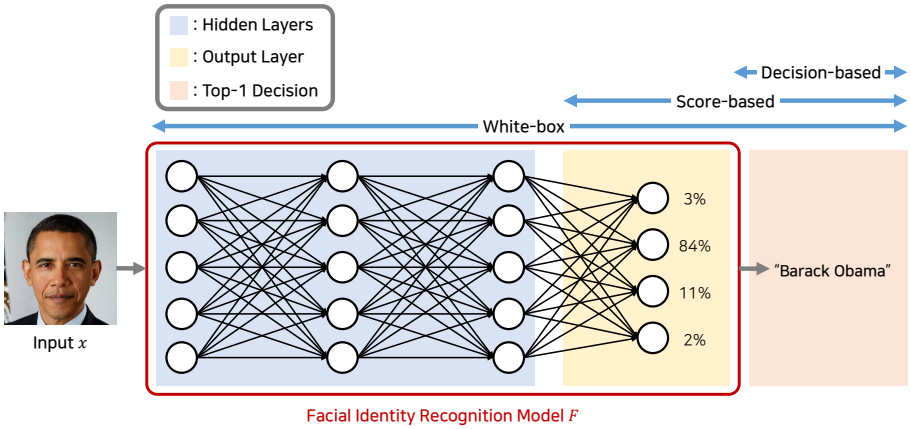

Fig. 2. An illustration of the common threat models. In a white-box threat model, the attacker can access the whole information of a DNN model including trained weights. In a score-based threat model, the attacker can access the output layer over all classes. In a decision-based threat model, the attacker can access the top-1 final decision.

example with a small perturbation by calculating a gradient vector typically more than thousands of times [12].

In the black-box threat model, the attacker can access only the output of a DNN model. The black-box threat model can be divided into two variants, the score-based threat model and the decision-based threat model. The score-based threat model assumes that an attacker is able to access the output of the softmax layer of a DNN model. Natural Evolution Strategy (NES) attack generates adversarial examples by estimating the gradient based on the top-k prediction scores of a DNN model [26]. On the other hand, the decision-based threat model assumes that an attacker is able to get the final decision of a DNN model, i.e., a predicted label alone. Boundary Attack (BA) uses random walks along the boundary of a DNN model to generate an adversarial example that looks similar to the source image [9]. HopSkipJump-Attack (HSJA) is then proposed to produce an adversarial example efficiently by combining binary search and gradient estimation [14]. As decision-based attacks (BA and HSJA) can access only the top-1 label, they start with an image already classified as the target class and maintain the classification result of the adversarial example during the whole attack procedure [9,14].

In this paper, we consider the decision-based threat model known as the most difficult black-box setting where an attacker can access only the top-1 label. This threat model is suitable for a real-world adversarial attack scenario because recent real-world applications such as the Clarifai service may provide only the top-1 label. Specifically, we present a variant of the HSJA [14], namely Latent-HSJA, by adding an encoding step to the original HSJA procedure. The

most significant difference between the original HSJA and our attack is that our attack leverages a latent space, not an input image space.

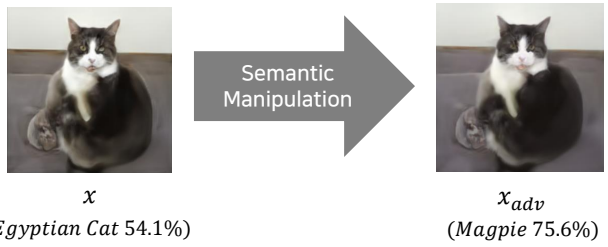

Fig. 3. An illustration of an unrestricted adversarial example based on an unrestricted attack method [35].

2.2 Generative Adversarial Networks

Generative Adversarial Network (GAN) has been proposed to generate plausible new data examples [19]. Especially, GANs with deep convolutional layers have shown remarkable achievements and are able to produce realistic examples in an image-specific domain [36]. Previous studies with GAN have shown that it is possible to generate high-resolution images up to 1024 $\times$ 1024 resolution in various domains such as the human face, vehicles, and animals [28,29]. Recently, StyleGAN architecture shows an outstanding quality of the synthesized image by combining progressive training and the idea of style transfer [29,30].

With the advent of GAN, some previous studies have shown that adversarial examples can exist in the distribution of GAN [35,5]. It implies that an attacker can generate various adversarial examples by manipulating latent vectors of GAN. From this intuition, we propose a novel method that efficiently utilizes GAN for generating unrestricted adversarial examples that look perceptually natural. With extensive experiments, we have found that the StyleGAN2 architecture is suitable for our Latent-HSJA to efficiently generate realistic unrestricted adversarial examples in a black-box setting [30].

2.3 Unrestricted Adversarial Attacks

Most previous studies consider attack methods that generate adversarial examples constrained to the specific p-norm bound [20,12,34]. Specifically, l_0, l_1, l_2, and l_∞ norms are commonly used [10]. Therefore, previous defense methods also focus on a norm-constrained adversarial perturbation. Especially, recently proposed defense methods [15,44] provide certified robustness against an adversarial perturbation with a specific size of p-norm bound. On the other hand, recent studies have proposed various unrestricted adversarial attack methods

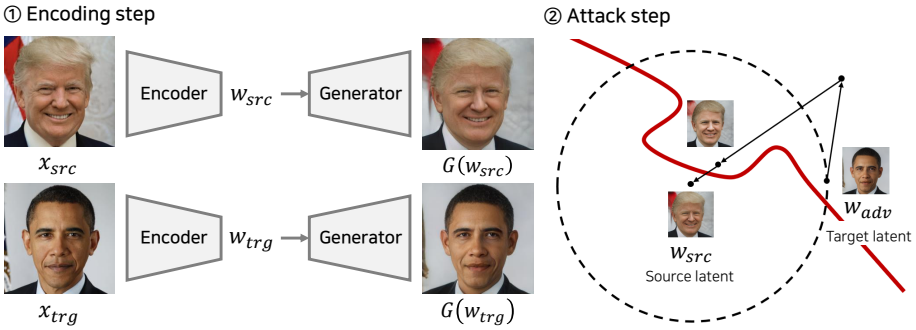

Fig. 4. An illustration of our attack method. Our proposed attack method consists of two steps. In the first encoding step, our method predicts two latent vectors (w_{src} and w_{trg}) according to input images. In the second attack step, our method drives the latent vector w_{trg} towards the latent vector w_{src} in a latent space. The red line denotes a decision boundary of a classification model in a latent space.

that are not norm-constrained [35,27,8,11] (Figure 3). Moreover, some studies have shown that these unrestricted attacks can bypass even the certified defense methods [39,18].

Nonetheless, the current unrestricted adversarial attacks have not yet successfully expanded to the black-box threat model. A related study has proposed an unrestricted black-box attack using an image-to-image translation network. However, it requires a large number of queries (more than hundreds of thousands of queries) in a black-box setting, which is not desirable for fooling real-world applications [27]. In this paper, we present a targeted attack method that generates unrestricted adversarial examples with a limited query budget (less than 20,000 queries). To the best of our knowledge, we are the first to propose a targeted black-box attack method that generates unrestricted adversarial examples with limited queries.

3 Proposed Methods

In a decision-based threat model, the common goal of an attacker is to generate an adversarial example x_{adv} that fools a DNN-based classification model $F(x)$ whose prediction output is the top-1 label for an input x [14,9]. In the targeted attack setting, the attacker finds an adversarial example x_{adv} that is similar to an x_{src} and is classified as a target class y_{trg} by the model F. The distance metric D is used to minimize the size of an adversarial perturbation. The common choice of the distance metric D is p-norm.

The general form of the objective is as follows:

$$\underset{x_{adv}}{\text{minimize}} \quad D(x_{src}, x_{adv}) \quad \text{s. t.} \quad F(x_{adv}) = y_{trg}.$$

For the unrestricted attack, the attacker sets D as a metric to measure a distance of semantic information such as rotation, hue, saturation, brightness, or high-level styles between two images [8,18,23]. Our proposed method also minimizes the semantic distance between two images. We use D as the l_2 distance between the two latent vectors, i.e., w_{src} and w_{adv}. We postulate that if the distance between two latent vectors is short enough in the latent space of the GAN model G, the two synthesized images are similar in human perception. Especially when two latent vectors are exactly the same, the images generated by propagating two latent vectors into the GAN model should be the same.

Therefore, our attack uses the following objective:

$$\underset{w_{adv}}{\text{minimize}} \quad D(w_{src}, w_{adv}) \quad \text{s. t.} \quad F(G(w_{adv})) = y_{trg}.$$

3.1 Decision-based Attack in Latent Space

We propose a method to conduct an unrestricted black-box attack, namely, Latent-HSJA, consisting of two steps (Figure 4). First, we find latent vectors of source and target images (w_{src} and w_{trg}). We use the latent vector of the target image w_{trg} as an initial latent vector for an adversarial example w_{adv}. The corresponding adversarial example $G(w_{adv})$ should be adversarial (i.e., classified as the target class) at the start of the attack. Second, we conduct a decision-based update procedure in the latent space. Our method always maintains the predicted label of the adversarial example to be adversarial during the whole attack procedure ($F(G(w_{adv})) = y_{trg}$). We illustrate our attack algorithm in Figure 4.

We utilize the HSJA method [14] for the second step of the proposed method (the decision-based update). Our decision-based attack minimizes $D(w_{src}, w_{adv})$ while preserving that the model output $F(G(w_{adv}))$ is always classified as the target class y_{trg}. After we run the attack algorithm, we get an adversarial latent vector w_{adv} such that $G(w_{adv}) \approx x_{src}$ in human perception. Our method tends to change the semantic information of an adversarial example $G(w_{adv})$ because our method chooses to update the latent vector w_{adv} in the latent space rather than directly update the image x_{adv} in the image space.

Algorithm 1: Decision-based attack in a latent space

Require: *Encoder* denotes an image encoding network, G denotes a pre-trained GAN model, $LatentHSJA$ denotes our attack based on the HopSkipJump-Attack, and F denotes a classification model for the attack.
Input: Two input images x_{src}, x_{trg}.
Result: The adversarial example x_{adv}.
$w_{src} = Encoder(x_{src})$;
$w_{trg} = Encoder(x_{trg})$;
$w_{adv} = LatentHSJA(G, F, w_{src}, w_{trg})$;
$x_{adv} = G(w_{adv})$;

Table 1. The validation accuracies of our trained models. Our attack method is evaluated on these classification models.

Architectures	Identity Dataset	Gender Dataset
MNasNet1.0	78.35%	98.38%
DenseNet121	86.42%	98.15%
ResNet18	87.82%	98.55%
ResNet101	87.98%	98.05%

As a result, we can generate a perceptually natural adversarial example even in an early stage of the attack because our Latent-HSJA updates coarse-grained semantic features of the image. On the other hand, previous decision-based attack methods based on p-norm metrics suffer from a limitation that the generated adversarial example x_{adv} is not perceptually plausible in an early stage of the attack (less than 5,000 queries).

3.2 Encoding Algorithm

Our attack requires an accurate encoding method that maps an image x into a latent vector w such that $F(x) = F(G(Encoder(x)))$. Ideally, a perfect encoding method can satisfy this requirement. We first prepare w_{trg} by using the encoder so that $G(w_{trg})$ is classified as an adversarial class y_{trg}. Secondly, we also prepare w_{src} by using a given source image x_{src}. With this pair (w_{src} and w_{trg}), we could finally get the adversarial example $G(w_{adv})$ by conducting the Latent-HSJA (Algorithm 1).

For developing an encoding method that finds a latent vector according to an image for the StyleGAN-based generators, two approaches are commonly used. The first is an optimization-based approach that updates a latent vector using gradient descent steps [1,2]. The second approach trains an additional encoder network that embeds an image to a latent vector [16,37,42]. Recent studies have also shown that hybrid approaches combining the two methods could return better-encoded results [46,7]. We have found that the optimization-based encoding method is unsuitable for our attack because it could cause severe overfitting. Previous studies have also shown that such overfitting is not desirable for semantic image manipulation [46,42]. When the initial latent vector is highly overfitted, our attack could fail. We have observed that the recent pSp encoder provides successful encoding results for our attack method [37].

4 Experiments

4.1 Experiment Settings

Dataset For experiments, we use the CelebA-HQ dataset [31], a common baseline dataset for face attribute classification tasks. The CelebA-HQ dataset contains 30,000 face images that are all 1024×1024 resolution images. There are

Fig. 5. Showcases of our Latent-HSJA attack method against a face gender classification model. The first row shows target images and the second row shows source images. The adversarial examples in the last row are classified as target classes and require a feasible number of queries (only 20,000 queries).

6,217 unique identities and 40 binary attributes in the CelebA-HQ dataset. First, we filter the CelebA-HQ dataset so that each identity contains more than 15 images for training the facial identity recognition models. As a result, our filtered facial identity dataset contains 307 identities, and there are 4,263 face images for training and 1,215 face images for validation. Secondly, we utilize the original CelebA-HQ dataset to train the facial gender classification models. The CelebA-HQ dataset contains 11,057 male images and 18,943 female images. We split these two datasets into 4:1 as training and validation.

Classification Models For validating our Latent-HSJA, we have trained classification models on the two aforementioned datasets. We have fine-tuned MNasNet1.0, DenseNet121, ResNet18 and ResNet101 [41,24,22] that are pre-trained on the ILSVRC2012 dataset. All classification models resize the resolution of inputs to 256×256 in the input pre-processing step. The validation accuracies of all trained models are reported in Table 1. We have found the ResNet18 models show good generalization performance for both tasks, thus we report the main experimental results using the fine-tuned ResNet18 models. We also evaluate our attack method on a real-world application to verify the effectiveness of our method. The celebrity recognition service of Clarifai contains a large number of facial identities of over 10,000 recognized celebrities. For each face object of an image, this service returns the top-1 class and its probability.

Attack Details In our attack method, we utilize the StyleGAN2 architecture [30]. For updating a encoded latent vector, a previous study utilizes w^+ space

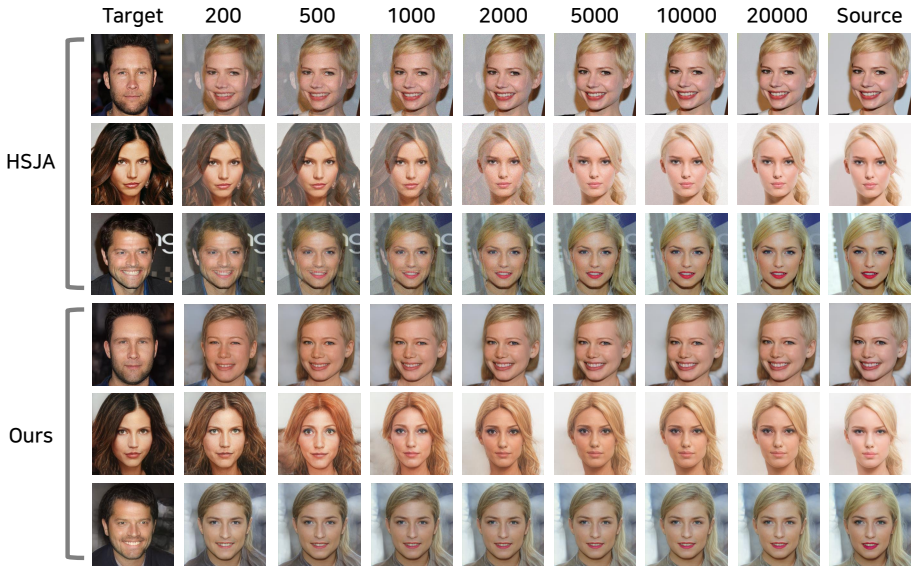

Fig. 6. An illustration of our targeted black-box attack results for a facial identity classification model. Our method (Latent-HSJA) is comparable with state-of-the-art norm-constrained adversarial attacks (HSJA) in a black-box setting.

whose dimension is 18×512 to get better results [1]. Following the previous work, we use w^+ space and normalize all the latent vectors so that the values of latent vectors are between [0, 1] and utilize the HSJA in the normalized latent space. In our experiments, we randomly select 100 (image x, encoded latent w) pairs such that $F(x)$ is equal to $F(G(w))$ in the validation datasets for facial identity recognition and gender recognition. As mentioned in the previous section, the goal of our unrestricted attack is to minimize the l_2 distance between w_{adv} and w_{src}. We note that the attack success rate is always 100% because Latent-HSJA maintains the adversarial example to be always adversarial in the whole attack procedure, and the objective of attacks is to minimize the similarity distance D. For whole experiments, we report the targeted adversarial attack results.

Evaluation Metrics The adversarial example x_{adv} should be close to the source image x_{src} such that $F(x_{adv}) = y_{trg}$ in the targeted adversarial attack setting. Therefore, we calculate how different the adversarial example x_{adv} is from the source image x_{src} using several metrics. We consider the similarity score SIM [25] and perceptual loss LPIPS [45] for evaluating our attack compared with the previously proposed p-norm based attack. Previous studies have demonstrated that the similarity score and LPIPS can be used for measuring perceptual similarity between two human face images [37].

Table 2. Experimental results of our Latent-HSJA compared with previous p-norm based HSJA against the gender recognition model and identity classification model. All adversarial examples are generated for fooling the ResNet18 models. The SIM and LPIPS scores are calculated using x_{src} and adversarial example x_{adv} ($G(w_{src})$ and $G(w_{adv})$ for Latent-HSJA).

		Gender Recognition					
Method	Metric	Model Queries					
		500	1000	3000	5000	10000	20000
HSJA [14]	SIM↑	0.546	0.621	0.724	0.780	**0.840**	**0.886**
	LPIPS↓	0.484	0.340	0.181	0.122	0.068	0.037
Ours	SIM↑	**0.769**	**0.794**	**0.817**	**0.821**	0.824	0.828
	LPIPS↓	**0.066**	**0.052**	**0.041**	**0.039**	**0.037**	**0.036**
		Identity Classification					
Method	Metric	Model Queries					
		500	1000	3000	5000	10000	20000
HSJA [14]	SIM↑	0.538	0.569	0.663	0.728	**0.821**	**0.895**
	LPIPS↓	0.313	0.287	0.187	0.132	0.067	**0.029**
Ours	SIM↑	**0.665**	**0.719**	**0.779**	**0.797**	0.811	0.819
	LPIPS↓	**0.164**	**0.119**	**0.072**	**0.061**	**0.053**	0.049

4.2 Gender Classification

Gender classification is a common binary classification task for classifying a face image. For the gender classification model, the targeted attack setting is the same as the untargeted attack setting. We demonstrate that adversarial examples made from our Latent-HSJA are sufficiently similar to the source images x_{src} (Figure 5). Especially, our method is more efficient in the initial steps than the norm-based adversarial attack (Table 2). Our adversarial examples show better results in the evaluation metrics of SIM and LPIPS compared to the norm-based adversarial attack below 5,000 queries.

4.3 Identity Recognition

For the evaluation of our method against the facial identity recognition task, we have experimented with the targeted attack. We demonstrate that our method is query-efficient for the targeted attack. Our targeted attack is based on Algorithm 1. As illustrated in Figure 6, our method mainly changes the coarse-grained semantic features in the initial steps (Table 2). This property is desirable for the black-box attack since our adversarial example will quickly be semantically far away from the target image. Moreover, the previous decision-based attacks often generate disrupted images that contain perceptible artificial noises with a limited query budget (under 5,000 queries). In contrast, our method maintains the adversarial examples to be always perceptually feasible.

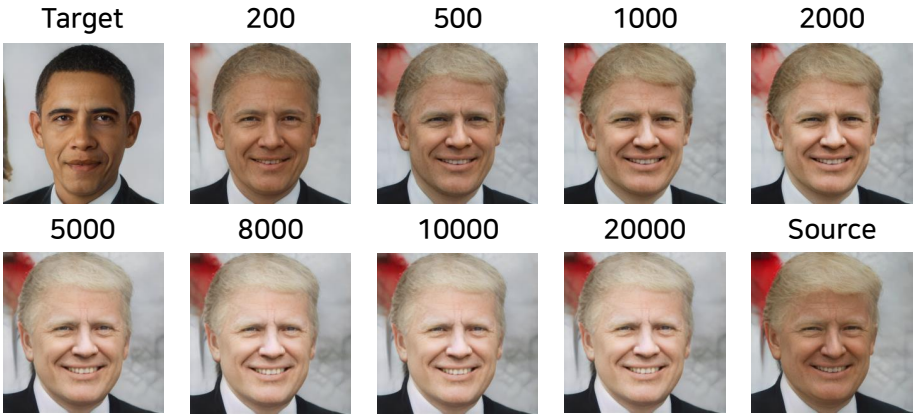

Fig. 7. An illustration of our targeted unrestricted black-box attack results for Clarifai service.

Table 3. The results of brute-forcing for finding feasible target image instances that can be used for starting points of our Latent-HSJA using StyleGAN [29] (higher $|Class|$ is better). The facial identity classification model contains 307 identities.

Dataset	Model Queries	$\|Class\|$	$\|Class\|_{>50}$	$\|Class\|_{>90}$
FFHQ [29]	1000	183	117	36
	5000	252	203	90
	10000	267	241	111
	40000	300	275	168
	80000	305	288	200
CelebA-HQ [31]	1000	242	170	46
	5000	301	268	131
	10000	304	289	171
	40000	307	306	245
	80000	307	306	274

4.4 Real-world Application

Ethical Considerations We are aware that it is important not to cause any disturbance to commercial services when evaluating real-world applications. Prior to experiments with the Clarifai service, we received permission from Clarifai to use the public API with a certain budget for the research purpose.

Attack Results As illustrated in Figure 7, our method is suitable for evaluating real-world black-box applications' robustness. We demonstrate that our method requires a feasible number of queries (about 20,000) with a specific source image. We note that the targeted attack for this real-world application is challenging because this service contains more than 10,000 identities. When running our attack method 5 times with random celebrity image pairs (x_{src}, x_{trg}), we have

gotten the average value of SIM is 0.694 and the average value of LPIPS is 0.125. As a result, the Clarifai celebrity recognition service shows better robustness than our trained facial identity recognition model.

5 Discussion

We have found a generative model can efficiently create various images that are classified as a specific target class if the distributions of both a generative model and a classification model are similar to each other. We present a brute-forcing method for finding numerous target images. For validating the brute-forcing method, we simply use a pre-trained StyleGAN model [29]. We have demonstrated that sampling random image $G(w)$ with a large number of queries can find various images that could be classified as a target class (Table 3). For example, the brute-forcing with 80,000 queries can find the majority classes of the facial identity recognition model including high confidence ($> 90\%$) images. We note that the instance generated by brute-forcing is not appropriate as an adversarial example itself because this instance may have similar semantic features to the target class images. However, our result shows that if the GAN learns a similar distribution of the classification model and has enough capability for sampling various images, it is easy to find various images that are classified as a specific target class. For example, we can use a found image by brute-forcing as an initial starting point in our attack procedure.

6 Conclusion

In this work, we present Latent-HSJA, a novel method for generating unrestricted adversarial examples in a black-box setting. We demonstrate that our method can successfully attack state-of-the-art classification models, including a real-world application. Our method can explore the latent space and generate various realistic adversarial examples in terms of that the latent space of GAN contains a large number of adversarial examples. We especially utilize the valuable features of the StyleGAN2 architecture that have highly disentangled latent representations for a black-box attack. The experimental results show that our attack method has the potential as a new adversarial attack method. However, we have observed that the adversarial example is sometimes not feasible even though $MSE(w_{adv}, w_{src})$ is small enough. Therefore, future work may seek a better evaluation metric useful for our unrestricted adversarial attack. In addition, for generating strong a adversarial example, an attacker can use a hybrid approach that combines our Latent-HSJA with other norm-based decision-based attacks in terms of that our attack method shows better results in an early stage of the attack. Although we focus on the unrestricted attack on the face-related applications in this work, we believe our method is scalable to other classification tasks because GAN networks can be trained on various datasets. We hope our work has demonstrated new possibilities for generating semantic adversarial examples in a real-world black-box scenario.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4432–4441 (2019)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8296–8305 (2020)
3. Alzantot, M., Sharma, Y., Elgohary, A., Ho, B., Srivastava, M.B., Chang, K.: Generating natural language adversarial examples. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018)
4. Alzaylaee, M.K., Yerima, S.Y., Sezer, S.: Dl-droid: Deep learning based android malware detection using real devices. Computers & Security **89**, 101663 (2020)
5. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International Conference on Machine Learning. pp. 274–283. PMLR (2018)
6. Balaban, S.: Deep learning and face recognition: the state of the art. In: Biometric and Surveillance Technology for Human and Activity Identification XII. vol. 9457, p. 94570B. International Society for Optics and Photonics (2015)
7. Bau, D., Zhu, J.Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., Torralba, A.: Seeing what a gan cannot generate. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4502–4511 (2019)
8. Bhattad, A., Chong, M.J., Liang, K., Li, B., Forsyth, D.A.: Unrestricted adversarial examples via semantic manipulation. In: 8th International Conference on Learning Representations, ICLR 2020 (2020)
9. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In: 6th International Conference on Learning Representations, ICLR 2018 (2018)
10. Brendel, W., Rauber, J., Kümmerer, M., Ustyuzhaninov, I., Bethge, M.: Accurate, reliable and fast robustness evaluation. In: Annual Conference on Neural Information Processing Systems 2019 (2019)
11. Brown, T.B., Carlini, N., Zhang, C., Olsson, C., Christiano, P., Goodfellow, I.: Unrestricted adversarial examples. arXiv preprint arXiv:1809.08352 (2018)
12. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57 (2017)
13. Carlini, N., Wagner, D.: Audio adversarial examples: Targeted attacks on speech-to-text. In: 2018 IEEE Security and Privacy Workshops (SPW). pp. 1–7 (2018)
14. Chen, J., Jordan, M.I., Wainwright, M.J.: Hopskipjumpattack: A query-efficient decision-based attack. In: 2020 ieee symposium on security and privacy (sp). pp. 1277–1294 (2020)
15. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: International Conference on Machine Learning. pp. 1310–1320 (2019)
16. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. CoRR **abs/1605.09782** (2016)
17. Ebrahimi, J., Rao, A., Lowd, D., Dou, D.: HotFlip: White-box adversarial examples for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics (2018)

18. Ghiasi, A., Shafahi, A., Goldstein, T.: Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates. In: 8th International Conference on Learning Representations, ICLR 2020 (2020)
19. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Annual Conference on Neural Information Processing Systems 2014 (2014)
20. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations, ICLR 2015 (2015)
21. Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G.: A survey of deep learning techniques for autonomous driving. Journal of Field Robotics **37**(3), 362–386 (2020)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Hosseini, H., Poovendran, R.: Semantic adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1614–1619 (2018)
24. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
25. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: adaptive curriculum learning loss for deep face recognition. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5901–5910 (2020)
26. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. In: International Conference on Machine Learning. pp. 2137–2146. PMLR (2018)
27. Kakizaki, K., Yoshida, K.: Adversarial image translation: Unrestricted adversarial examples in face recognition systems. In: Proceedings of the Workshop on Artificial Intelligence Safety, co-located with 34th AAAI 2020 (2020)
28. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: Proceedings of International Conference on Learning Representations (ICLR) 2018 (2018)
29. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
30. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
31. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
32. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, ICLR 2018 (2018)
33. Na, D., Park, N., Ji, S., Kim, J.: Captchas are still in danger: An efficient scheme to bypass adversarial captchas. In: International Conference on Information Security Applications. pp. 31–44. Springer (2020)

34. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P). pp. 372–387 (2016)
35. Poursaeed, O., Jiang, T., Goshu, Y., Yang, H., Belongie, S., Lim, S.N.: Fine-grained synthesis of unrestricted adversarial examples. arXiv preprint arXiv:1911.09058 (2019)
36. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: 4th International Conference on Learning Representations, ICLR 2016 (2016)
37. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2287–2296 (2021)
38. Shi, C., Xu, X., Ji, S., Bu, K., Chen, J., Beyah, R., Wang, T.: Adversarial captchas. IEEE transactions on cybernetics (2021)
39. Song, Y., Shu, R., Kushman, N., Ermon, S.: Constructing unrestricted adversarial examples with generative models. In: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018 (2018)
40. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (2014)
41. Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., Le, Q.V.: Mnasnet: Platform-aware neural architecture search for mobile. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2820–2828 (2019)
42. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) **40**(4), 1–14 (2021)
43. Wang, R., Juefei-Xu, F., Guo, Q., Huang, Y., Xie, X., Ma, L., Liu, Y.: Amora: Black-box adversarial morphing attack. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1376–1385 (2020)
44. Wong, E., Kolter, J.Z.: Provable defenses against adversarial examples via the convex outer adversarial polytope. In: Proceedings of the 35th International Conference on Machine Learning, ICML 2018 (2018)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
46. Zhu, J., Shen, Y., Zhao, D., Zhou, B.: In-domain gan inversion for real image editing. In: European Conference on Computer Vision. pp. 592–608. Springer (2020)