SemanticAdv: Generating Adversarial Examples via Attribute-conditioned Image Editing

Haonan Qiu[*1] Chaowei Xiao[*2] Lei Yang[*3] Xinchen Yan[2,4]
Honglak Lee[2] Bo Li[5]

[1]The Chinese University of Hong Kong, Shenzhen
[2]University of Michigan, Ann Arbor
[3]The Chinese University of Hong Kong
[4] Uber ATG, [5] UIUC

Abstract. Recent studies have shown that DNNs are vulnerable to adversarial examples which are manipulated instances targeting to mislead DNNs to make incorrect predictions. Currently, most such adversarial examples try to guarantee "subtle perturbation" by limiting the L_p norm of the perturbation. In this paper, we propose *SemanticAdv* to generate a new type of *semantically realistic* adversarial examples via attribute-conditioned image editing. Compared to existing methods, our *SemanticAdv* enables fine-grained analysis and evaluation of DNNs with input variations in the attribute space. We conduct comprehensive experiments to show that our adversarial examples not only exhibit semantically meaningful appearances but also achieve high targeted attack success rates under both whitebox and blackbox settings. Moreover, we show that the existing pixel-based and attribute-based defense methods fail to defend against *SemanticAdv*. We demonstrate the applicability of *SemanticAdv* on both face recognition and general street-view images to show its generalization. We believe that our work can shed light on further understanding about vulnerabilities of DNNs as well as novel defense approaches. Our implementation is available at https://github.com/AI-secure/SemanticAdv.

1 Introduction

Deep neural networks (DNNs) have demonstrated great successes in various vision tasks [38,60,63,26,59,43,83,13,80]. Meanwhile, several studies have revealed the vulnerablity of DNNs against input variations [64,24,47,52,12,72,71,73,74,21,10,11,70,85,66]. For example, carefully crafted L_p bounded perturbations added to the pristine input images can introduce arbitrary prediction errors during testing time. While being visually imperceptible, L_p bounded adversarial attacks have certain limitations as they only capture the variations in the raw pixel space and cannot guarantee the semantic realism for the generated instances. Recent works [75,33,69] have shown the limitations of the L_p bounded perturbation (e.g., cannot handle variations in lighting conditions). Therefore, understanding the failure modes of deep

[*] The first three authors contributed equally.

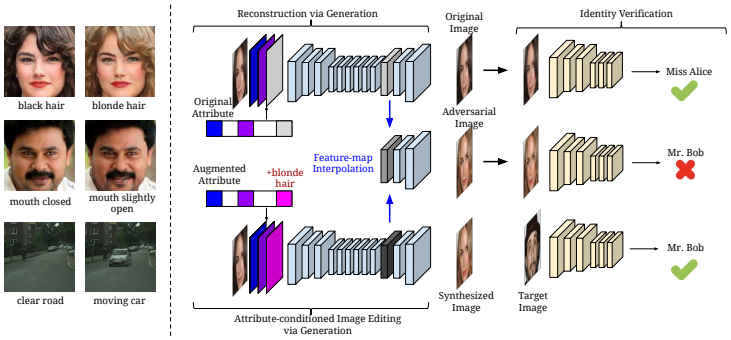

Fig. 1. Pipeline of *SemanticAdv* Left: Each row shows a pair of images differ in only one semantic aspect. One of them is sampled from the ground-truth dataset, while the other one is adversarial example created by our conditional image generator. Right: Overview of the proposed attribute-conditioned *SemanticAdv* against the face identity verification model

neural networks beyond raw pixel variations including semantic perturbations requires further understanding and exploration.

In this work, we focus on studying how DNNs respond towards semantically meaningful perturbations in the visual attribute space. In the visual recognition literature, visual attributes [22,39,53] are properties observable in images that have human-designated properties (e.g., *black hair* and *blonde hair*). As illustrated in Figure 1 (left), given an input image with known attributes, we would like to craft semantically meaningful (attribute-conditioned) adversarial examples via image editing along a single attribute or a subset of attributes while keeping the rest unchanged. Compared to traditional L_p bounded adversarial perturbations or semantic perturbations on global color and texture [5], such attribute-based image editing enables the users to conduct a fine-grained analysis and evaluation of the DNN models through removing one or a set of visual aspects or adding one object into the scene. We believe our attribute-conditioned image editing is a natural way of introducing semantic perturbations, and it preserves clear interpretability as: wearing a new pair of glasses or having the hair dyed with a different color.

To facilitate the generation of semantic adversarial perturbations along a single attribute dimension, we take advantage of the disentangled representation in deep image generative models [55,34,6,78,14,3,82,31]. Such disentangled representation allows us to explore the variations for a specific semantic factor while keeping the other factors unchanged. As illustrated in Figure 1 (right), we first leverage an attribute-conditioned image editing model [14] to construct a new instance which is very similar to the source except one semantic aspect (the source image is given as input). Given such pair of images, we synthesize the adversarial example by interpolating between the pair of images in the *feature-map space*. As the interpolation is constrained by the image pairs, the appearance of the resulting semantic adversarial example resembles both of them.

To validate the effectiveness of our proposed *SemanticAdv* by attribute-conditioned image editing, we consider two real-world tasks, including face verification and landmark detection. We conduct both qualitative and quantitative evaluations on CelebA dataset [42]. The results show that our *SemanticAdv* not only achieves high targeted attack success rate and also preserves the semantic meaning of the corresponding input images. To further demonstrate the applicability of our *SemanticAdv* beyond face domain, we extend the framework to generate adversarial street-view images. We treat semantic layouts as input attributes and use the layout-conditioned image editing model [27] pre-trained on Cityscape dataset [16]. Our results show that a well-trained semantic segmentation model can be successfully attacked to neglect the pedestrian if we insert another object by the side using our image editing model. In addition, we show that existing adversarial training-based defense method is less effective against our attack method, which motivates further defense strategies against such semantic adversarial examples.

Our contributions are summarized as follows: (1) We propose a novel method *SemanticAdv* to generate semantically meaningful adversarial examples via attribute-conditioned image editing based on **feature-space** interpolation. Compared to existing adversarial attacks, our method enables fine-grained attribute analysis as well as further evaluation of vulnerabilities for DNN models. Such semantic adversarial examples also provide explainable analysis for different attributes in terms of their robustness and editing flexibility. (2) We conduct extensive experiments and show that the proposed feature-space interpolation strategy can generate high quality attribute-conditioned adversarial examples more effectively than the simple attribute-space interpolation. Additionally, our *SemanticAdv* exhibits high attack **transferability** as well as 67.7% query-free **black-box attack** success rate on a real-world face verification platform. (3) We empirically show that, compared to L_p attacks, the existing per-pixel based as well as attribute-based defense methods fail to defend against our *SemanticAdv*, which indicates that such semantic adversarial examples identify certain unexplored vulnerable landscape of DNNs. (4) To demonstrate the applicability and generalization of *SemanticAdv* beyond the face recognition domain, we extend the framework to generate adversarial street-view images that fool semantic segmentation models effectively.

2 Related Work

Semantic image editing. Semantic image synthesis and manipulation is a popular research topic in machine learning, graphics and vision. Thanks to recent advances in deep generative models [36,23,51] and the empirical analysis of deep classification networks [38,60,63], past few years have witnessed tremendous breakthroughs towards high-fidelity pure image generation [55,34,6], attribute-to-image generation [78,14], text-to-image generation [46,56,50,49,84,31], and image-to-image translation [29,87,41,68,27].

Adversarial examples. Generating L_p bounded adversarial perturbation has been extensively studied recently [64,24,47,52,12,72,71,73,21,70]. To further explore diverse adversarial attacks and potentially help inspire defense mechanisms, it is important to generate the so-called "unrestricted" adversarial examples which contain unrestricted magnitude of perturbation while still preserve perceptual realism [7]. Recently, [75,20] propose to spatially transform the image patches instead of adding pixel-wise perturbation, while such spatial transformation does not consider semantic information. Our proposed *semanticAdv* focuses on generating unrestricted perturbation with semantically meaningful patterns guided by visual attributes.

Relevant to our work, [61] proposed to synthesize adversarial examples with an unconditional generative model. [5] studied semantic transformation in only the color or texture space. Compared to these works, *semanticAdv* is able to generate adversarial examples in a controllable fashion using specific visual attributes by performing manipulation in the feature space. We further analyze the robustness of the recognition system by generating adversarial examples guided by different visual attributes. Concurrent to our work, [32] proposed to generate semantic-based attacks against a restricted binary classifier, while our attack is able to mislead the model towards arbitrary adversarial targets. They conduct the manipulation within the attribution space which is less flexible and effective than our proposed feature-space interpolation.

3 SemanticAdv

3.1 Problem Definition

Let $\mathcal{M}$ be a machine learning model trained on a dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}$ consisting of image-label pairs, where $\mathbf{x} \in \mathbb{R}^{H \times W \times D_I}$ and $\mathbf{y} \in \mathbb{R}^{D_L}$ denote the image and the ground-truth label, respectively. Here, H, W, D_I, and D_L denote the image height, image width, number of image channels, and label dimensions, respectively. For each image $\mathbf{x}$, our model $\mathcal{M}$ makes a prediction $\hat{\mathbf{y}} = \mathcal{M}(\mathbf{x}) \in \mathbb{R}^{D_L}$. Given a target image-label pair $(\mathbf{x}^{\text{tgt}}, \mathbf{y}^{\text{tgt}})$ and $\mathbf{y} \neq \mathbf{y}^{\text{tgt}}$, a *traditional attacker* aims to synthesize adversarial examples $\mathbf{x}^{\text{adv}}$ by adding pixel-wise perturbations to or spatially transforming the original image $\mathbf{x}$ such that $\mathcal{M}(\mathbf{x}^{\text{adv}}) = \mathbf{y}^{\text{tgt}}$. In this work, we consider a *semantic attacker* that generates semantically meaningful perturbation via attribute-conditioned image editing with a conditional generative model $\mathcal{G}$. Compared to the traditional attacker, the proposed attack method generates adversarial examples in a more controllable fashion by editing a single semantic aspect through attribute-conditioned image editing.

3.2 Attribute-conditioned Image Editing

In order to produce semantically meaningful perturbations, we first introduce how to synthesize attribute-conditioned images through interpolation.

Semantic image editing. For simplicity, we start with the formulation where the input attribute is represented as a compact vector. This formulation can be directly extended to other input attribute formats including semantic layouts. Let $\mathbf{c} \in \mathbb{R}^{D_C}$ be an attribute representation reflecting the semantic factors (e.g., expression or hair color of a portrait image) of image $\mathbf{x}$, where D_C indicates the attribute dimension and $c_i \in \{0, 1\}$ indicates the existence of i-th attribute. We are interested in performing semantic image editing using the attribute-conditioned image generator $\mathcal{G}$. For example, given a portrait image of a girl with `black hair` and the new attribute `blonde hair`, our generator is supposed to synthesize a new image that turns the girl's hair color from black to blonde while keeping the rest of appearance unchanged. The synthesized image is denoted as $\mathbf{x}^{\text{new}} = \mathcal{G}(\mathbf{x}, \mathbf{c}^{\text{new}})$ where $\mathbf{c}^{\text{new}} \in \mathbb{R}^{D_C}$ is the new attribute. In the special case when there is no attribute change ($\mathbf{c} = \mathbf{c}^{\text{new}}$), the generator simply reconstructs the input: $\mathbf{x}' = \mathcal{G}(\mathbf{x}, \mathbf{c})$ (ideally, we hope $\mathbf{x}'$ equals to $\mathbf{x}$). As our attribute representation is disentangled and the change of attribute value is sufficiently small (e.g., we only edit a single semantic attribute), our synthesized image $\mathbf{x}^{\text{new}}$ is expected to be close to the data manifold [4,57,55]. In addition, we can generate many similar images by linearly interpolating between the image pair $\mathbf{x}$ and $\mathbf{x}^{\text{new}}$ in the attribute-space or the feature-space of the image-conditioned generator $\mathcal{G}$, which is supported by the previous work [78,55,3]

Attribute-space interpolation. Given a pair of attributes $\mathbf{c}$ and $\mathbf{c}^{\text{new}}$, we introduce an interpolation parameter $\alpha \in (0, 1)$ to generate the augmented attribute vector $\mathbf{c}^* \in \mathbb{R}^{D_C}$ (see Eq. 1). Given augmented attribute $\mathbf{c}^*$ and original image $\mathbf{x}$, we produce the image $\mathbf{x}^*$ by the generator $\mathcal{G}$ through attribute-space interpolation.

$$\begin{aligned} \mathbf{x}^* &= \mathcal{G}(\mathbf{x}, \mathbf{c}^*) \\ \mathbf{c}^* &= \alpha \cdot \mathbf{c} + (1 - \alpha) \cdot \mathbf{c}^{\text{new}}, \text{ where } \alpha \in [0, 1] \end{aligned} \tag{1}$$

Feature-map interpolation. Alternatively, we propose to interpolate using the feature map produced by the generator $\mathcal{G} = \mathcal{G}_{\text{dec}} \circ \mathcal{G}_{\text{enc}}$. Here, $\mathcal{G}_{\text{enc}}$ is the encoder module that takes the image as input and outputs the feature map. Similarly, $\mathcal{G}_{\text{dec}}$ is the decoder module that takes the feature map as input and outputs the synthesized image. Let $\mathbf{f}^* = \mathcal{G}_{\text{enc}}(\mathbf{x}, \mathbf{c}) \in \mathbb{R}^{H_F \times W_F \times C_F}$ be the feature map of an intermediate layer in the generator, where H_F, W_F and C_F indicate the height, width, and number of channels in the feature map.

$$\begin{aligned} \mathbf{x}^* &= \mathcal{G}_{\text{dec}}(\mathbf{f}^*) \\ \mathbf{f}^* &= \boldsymbol{\beta} \odot \mathcal{G}_{\text{enc}}(\mathbf{x}, \mathbf{c}) + (\mathbf{1} - \boldsymbol{\beta}) \odot \mathcal{G}_{\text{enc}}(\mathbf{x}, \mathbf{c}^{\text{new}}) \end{aligned} \tag{2}$$

Compared to the attribute-space interpolation which is parameterized by a scalar α, we parameterize feature-map interpolation by a tensor $\boldsymbol{\beta} \in \mathbb{R}^{H_F \times W_F \times C_F}$ ($\beta_{h,w,k} \in [0, 1]$, where $1 \leq h \leq H_F$, $1 \leq w \leq W_F$, and $1 \leq k \leq C_F$) with the same shape as the feature map. Compared to linear interpolation over attribute-space, such design introduces more flexibility for adversarial attacks. Empirical results in Section 4.2 show such design is critical to maintain both attack success and good perceptual quality at the same time.

3.3 Generating Semantically Meaningful Adversarial Examples

Existing work obtains the adversarial image $\mathbf{x}^{\text{adv}}$ by adding perturbations or transforming the input image $\mathbf{x}$ directly. In contrast, our semantic attack method requires additional attribute-conditioned image generator $\mathcal{G}$ during the adversarial image generation through interpolation. As we see in Eq. 3, the first term of our objective function is the adversarial metric, the second term is a smoothness constraint to guarantee the perceptual quality, and λ is used to control the balance between the two terms. The adversarial metric is minimized once the model $\mathcal{M}$ has been successfully attacked towards the target image-label pair $(\mathbf{x}^{\text{tgt}}, \mathbf{y}^{\text{tgt}})$. For identify verification, $\mathbf{y}^{\text{tgt}}$ is the identity representation of the target image. For structured prediction tasks in our paper, $\mathbf{y}^{\text{tgt}}$ either represents certain coordinates (landmark detection) or semantic label maps (semantic segmentation).

$$\begin{aligned} \mathbf{x}^{\text{adv}} &= \text{argmin}_{\mathbf{x}^*} \mathcal{L}(\mathbf{x}^*) \\ \mathcal{L}(\mathbf{x}^*) &= \mathcal{L}_{\text{adv}}(\mathbf{x}^*; \mathcal{M}, \mathbf{y}^{\text{tgt}}) + \lambda \cdot \mathcal{L}_{\text{smooth}}(\mathbf{x}^*) \end{aligned} \tag{3}$$

Identity verification. In the identity verification task, two images are considered to be the same identity if the corresponding identity embeddings from the verification model $\mathcal{M}$ are reasonably close.

$$\mathcal{L}_{\text{adv}}(\mathbf{x}^*; \mathcal{M}, \mathbf{y}^{\text{tgt}}) = \max\{\kappa, \Phi^{\text{id}}_{\mathcal{M}}(\mathbf{x}^*, \mathbf{x}^{\text{tgt}})\} \tag{4}$$

As we see in Eq. 4, $\Phi^{\text{id}}_{\mathcal{M}}(\cdot, \cdot)$ measures the distance between two identity embeddings from the model $\mathcal{M}$, where the normalized L_2 distance is used in our setting. In addition, we introduce the parameter κ representing the constant related to the false positive rate (FPR) threshold computed from the development set.

Structured prediction. For structured prediction tasks such as landmark detection and semantic segmentation, we use Houdini objective proposed in [15] as our adversarial metric and select the target landmark (semantic segmentation) target as $\mathbf{y}^{\text{tgt}}$. As we see in the equation, $\Phi_{\mathcal{M}}(\cdot, \cdot)$ is a scoring function for each image-label pair and γ is the threshold. In addition, $l(\mathbf{y}^*, \mathbf{y}^{\text{tgt}})$ is task loss decided by the specific adversarial target, where $\mathbf{y}^* = \mathcal{M}(\mathbf{x}^*)$.

$$\mathcal{L}_{\text{adv}}(\mathbf{x}^*; \mathcal{M}, \mathbf{y}^{\text{tgt}}) = P_{\gamma \sim \mathcal{N}(0,1)}\Big[\Phi_{\mathcal{M}}(\mathbf{x}^*, \mathbf{y}^*) - \Phi_{\mathcal{M}}(\mathbf{x}^*, \mathbf{y}^{\text{tgt}}) < \gamma\Big] \cdot l(\mathbf{y}^*, \mathbf{y}^{\text{tgt}}) \tag{5}$$

Interpolation smoothness $\mathcal{L}_{smooth}$. As the tensor to be interpolated in the feature-map space has far more parameters compared to the attribute itself, we propose to enforce a smoothness constraint on the tensor α used in feature-map interpolation. As we see in Eq. 6, the smoothness loss encourages the interpolation tensors to

consist of piece-wise constant patches spatially, which has been widely used as a pixel-wise de-noising objective for natural image processing [45,30].

$$\mathcal{L}_{\text{smooth}}(\boldsymbol{\beta}) = \sum_{h=1}^{H_F-1}\sum_{w=1}^{W_F} \|\boldsymbol{\beta}_{h+1,w} - \boldsymbol{\beta}_{h,w}\|_2^2 + \sum_{h=1}^{H_F}\sum_{w=1}^{W_F-1} \|\boldsymbol{\beta}_{h,w+1} - \boldsymbol{\beta}_{h,w}\|_2^2 \quad (6)$$

4 Experiments

In the experimental section, we mainly focus on analyzing the proposed *SemanticAdv* in attacking state-of-the-art face recognition systems [62,59,86,67,81,79,28] due to its wide applicability (e.g., identification for mobile payment) in the real world. We attack both face verification and face landmark detection by generating attribute-conditioned adversarial examples using annotations from CelebA dataset [42]. In addition, we extend our attack to urban street scenes with semantic label maps as the condition. We attack the semantic segmentation model DRN-D-22 [83] previously trained on Cityscape [16] by generating adversarial examples with dynamic objects manipulated (e.g., insert a car into the scene).

4.1 Experimental Setup

Face identity verification. We select `ResNet-50` and `ResNet-101` [26] trained on MS-Celeb-1M [25,17] as our face verification models. The models are trained using two different objectives, namely, `softmax` loss [62,86] and `cosine` loss [67]. For simplicity, we use the notation "R-N-S" to indicate the model with N-layer residual blocks as backbone trained using `softmax` loss, while "R-N-C" indicates the same backbone trained using `cosine` loss. The distance between face features is measured by normalized L2 distance. For R-101-S model, we decide the parameter κ based on the commonly used false positive rate (FPR) for the identity verification task [37,35]. Four different FPRs have been used: 10^{-3} (with $\kappa = 1.24$), 3×10^{-4} (with $\kappa = 1.05$), 10^{-4} (with $\kappa = 0.60$), and $< 10^{-4}$ (with $\kappa = 0.30$). Supplementary provides more details on the performance of face recognition models and their corresponding κ. To distinguish between the FPR we used in generating adversarial examples and the other FPR used in evaluation, we introduce two notations "Generation FPR (G-FPR)" and "Test FPR (T-FPR)". For the experiment with query-free black-box API attacks, we use two online face verification services provided by Face++ [2] and AliYun [1].

Semantic attacks on face images. In our experiments, we randomly sample $1,280$ distinct identities form CelebA [42] and use the StarGAN [14] for attribute-conditional image editing. In particular, we re-train our model on CelebA by aligning the face landmarks and then resizing images to resolution 112×112. We select 17 identity-preserving attributes as our analysis, as such attributes mainly reflect variations in facial expression and hair color.

In feature-map interpolation, to reduce the reconstruction error brought by the generator (e.g., $\mathbf{x} \neq \mathcal{G}(\mathbf{x}, \mathbf{c})$) in practice, we take one more step to obtain the updated feature map $\mathbf{f}' = \mathcal{G}_{\text{enc}}(\mathbf{x}', \mathbf{c})$, where $\mathbf{x}' = \text{argmin}_{\mathbf{x}'} \|\mathcal{G}(\mathbf{x}', \mathbf{c}) - \mathbf{x}\|$.

For each distinct identity pair $(\mathbf{x}, \mathbf{x}^{\text{tgt}})$, we perform *semanticAdv* guided by each of the 17 attributes (e.g., we intentionally add or remove one specific attribute while keeping the rest unchanged). In total, for each image $\mathbf{x}$, we generate 17 adversarial images with different augmented attributes. In the experiments, we select a commonly-used pixel-wise adversarial attack method [12] (referred as CW) as our baseline. Compared to our proposed method, CW does not require visual attributes as part of the system, as it only generates one adversarial example for each instance. We refer the corresponding attack success rate as the instance-wise success rate in which the attack success rate is calculated for each instance. For each instance with 17 adversarial images using different augmented attributes, if one of the 17 produced images can attack successfully, we count the attack of this instance as one success, vice verse.

Face landmark detection. We select Face Alignment Network (FAN) [9] trained on 300W-LP [88] and fine-tuned on 300-W [58] for 2D landmark detection. The network is constructed by stacking Hour-Glass networks [48] with hierarchical block [8]. Given a face image as input, FAN outputs 2D heatmaps which can be subsequently leveraged to yield 68 2D landmarks.

Semantic attacks on street-view images. We select DRN-D-22 [83] as our semantic segmentation model and fine-tune the model on image regions with resolution 256×256. To synthesize semantic adversarial perturbations, we consider semantic label maps as the input attribute and leverage a generative image manipulation model [27] pre-trained on CityScape [16] dataset. The details are shown in supplementary materials.

4.2 *SemanticAdv* on Face Identity Verification

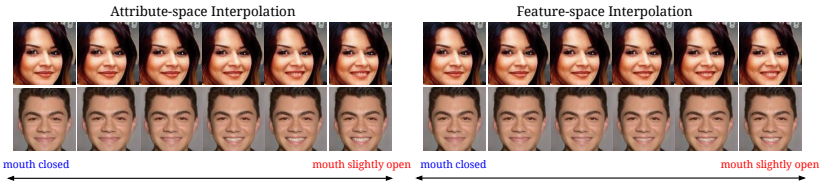

Fig. 2. Qualitative comparisons between attribute-space and feature-space interpolation In our visualization, we set the interpolation parameter to be $0.0, 0.2, 0.4, 0.6, 0.8, 1.0$

Attribute-space vs. feature-space interpolation. First, we qualitatively compare the two interpolation methods and found that both attribute-space and feature-space interpolation can generate reasonably looking samples (see Figure 2) through interpolation (these are not adversarial examples). However, we found the two

Table 1. Attack success rate by selecting attribute or different layer's feature-map for interpolation on R-101-S(%) using G-FPR = T-FPR = 10^{-3}. Here, $\mathbf{f}_i$ indicates the feature-map after i-th `up-sampling` operation. $\mathbf{f}_{-2}$ and $\mathbf{f}_{-1}$ are the first and the second feature-maps after the last `down-sampling` operation, respectively. Due to the effectiveness of **feature-space** interpolation, we only use **feature-space** interpolation in the following experiments.

Interpolation / Attack Success (%)	Feature					Attribute
	$\mathbf{f}_{-2}$	$\mathbf{f}_{-1}$	$\mathbf{f}_0$	$\mathbf{f}_1$	$\mathbf{f}_2$	
$\mathbf{x}^{\text{adv}}$, G-FPR = 10^{-3}	99.38	100.00	**100.00**	100.00	99.69	0.08
$\mathbf{x}^{\text{adv}}$, G-FPR = 10^{-4}	59.53	98.44	**99.45**	97.58	73.52	0.00

interpolation methods perform differently when we optimize using the adversarial objective (Eq. 3). We measure the attack success rate of attribute-space interpolation (with G-FPR = T-FPR = 10^{-3}): 0.08% on R-101-S, 0.31% on R-101-C, and 0.16% on both R-50-S and R-50-C, which consistently fails to attack the face verification model. Compared to attribute-space interpolation, generating adversarial examples with feature-space interpolation produces much better quantitative results (see Table 1). We conjecture that this is because the high dimensional feature space can provide more manipulation freedom. This also explains one potential reason of poor samples (e.g., blurry with many noticeable artifacts) generated by the method proposed in [32]. We select $\mathbf{f}_0$, the last `conv` layer before `up-sampling` layer in the generator for feature-space interpolation due to its good performance. Note that due to the effectiveness of **feature-space** interpolation, we only use **feature-space** interpolation for *semanticAdv* in the following experiments.

Qualitative analysis. Figure 3 (top) shows the generated adversarial images and corresponding perturbations against R-101-S of *SemanticAdv* and CW respectively. The text below each figure is the name of an augmented attribute, the sign before the name represents "adding" (in red) or "removing" (in blue) the corresponding attribute from the original image. Figure 3 (bottom) shows the adversarial examples with 17 augmented semantic attributes, respectively. The attribute names are shown in the bottom. The first row contains images generated by $\mathcal{G}(\mathbf{x}, \mathbf{c}^{\text{new}})$ with an augmented attribute $\mathbf{c}^{\text{new}}$ and the second row includes the corresponding adversarial images under feature-space interpolation. It shows that our *SemanticAdv* can generate examples with reasonably-looking appearance guided by the corresponding attribute. In particular, *SemanticAdv* is able to generate perturbations on the corresponding regions correlated with the augmented attribute, while the perturbations of CW have no specific pattern and are evenly distributed across the image.

To further measure the perceptual quality of the adversarial images generated by *SemanticAdv* in the most strict settings (G-FPR $< 10^{-4}$), we conduct a user study using Amazon Mechanical Turk (AMT). In total, we collect $2,620$ annotations from 77 participants. In $39.14 \pm 1.96\%$ (close to random guess 50%) of trials, the adversarial images generated by our *SemanticAdv* are selected as

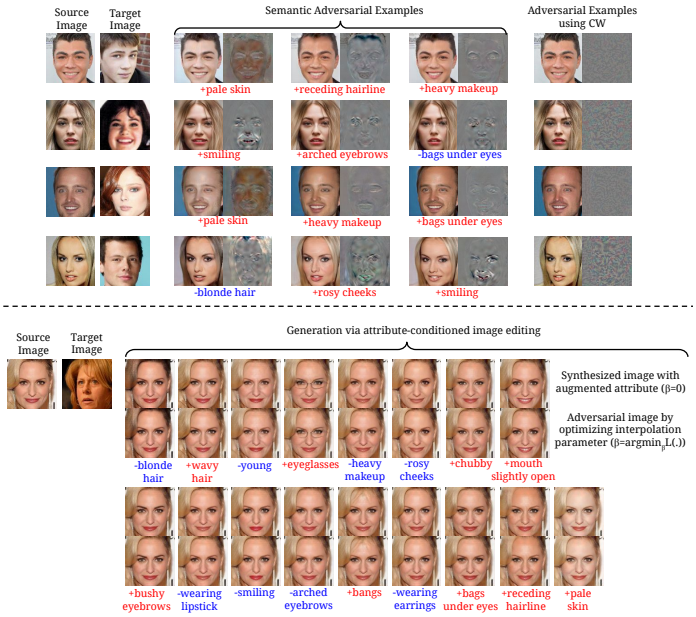

Fig. 3. Top: Qualitative comparisons between our proposed *SemanticAdv* and pixel-wise adversarial examples generated by CW [12]. Along with the adversarial examples, we also provide the corresponding perturbations (residual) on the right. Perturbations generated by our *SemanticAdv* (G-FPR = 10^{-3}) are unrestricted with semantically meaningful patterns. Bottom: Qualitative analysis on single-attribute adversarial attack (G-FPR = 10^{-3}). More results are shown in the supplementary

reasonably-looking images, while $30.27 \pm 1.96\%$ of trials by CW are selected as reasonably-looking. It indicates that *SemanticAdv* can generate more perceptually plausible adversarial examples compared with CW under the most strict setting (G-FPR $< 10^{-4}$). The corresponding images are shown in supplementary materials.

Single attribute analysis. One of the key advantages of our *SemanticAdv* is that we can generate adversarial perturbations in a more controllable fashion guided by the selected semantic attribute. This allows analyzing the robustness of a recognition system against different types of semantic attacks. We group the adversarial examples by augmented attributes in various settings. In Figure 4, we present the attack success rate against two face verification models, namely, R-101-S and R-101-C, using different attributes. We highlight the bar with light blue for G-FPR = 10^{-3} and blue for G-FPR = 10^{-4}, respectively. As shown in Figure 4, with a larger T-FPR = 10^{-3}, our *SemanticAdv* can achieve almost 100% attack success rate across different attributes. With a smaller T-FPR = 10^{-4}, we observe that *SemanticAdv* guided by some attributes such as `Mouth Slightly Open` and `Arched Eyebrows` achieve less than 50% attack success rate,

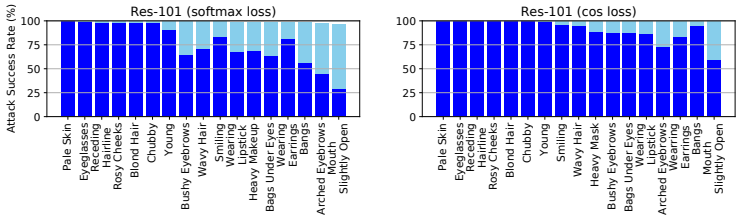

Fig. 4. Quantitative analysis on the attack success rate with different single-attribute attacks. In each figure, we show the results correspond to a larger FPR (G-FPR = T-FPR = 10^{-3}) in `skyblue` and the results correspond to a smaller FPR (G-FPR = T-FPR = 10^{-4}) in `blue`, respectively

while other attributes such as `Pale Skin` and `Eyeglasses` are relatively less affected. In summary, the above experiments indicate that *SemanticAdv* guided by attributes describing the local shape (e.g., mouth, earrings) achieve a relatively lower attack success rate compared to attributes relevant to the color (e.g., hair color) or entire face region (e.g., skin). This suggests that the face verification models used in our experiments are more robustly trained in terms of detecting local shapes compared to colors. In practice, we have the flexibility to select attributes for attacking an image based on the perceptual quality and attack success rate.

Transferability analysis. To generate adversarial examples under black-box setting, we analyze the transferability of *SemanticAdv* in various settings. For each model with different FPRs, we select the successfully attacked adversarial examples from Section 4.1 to construct our evaluation dataset and evaluate these adversarial samples across different models. Table 2(a) illustrates the transferability of *SemanticAdv* among different models by using the same FPRs (G-FPR = T-FPR = 10^{-3}). Table 2(b) illustrates the result with different FPRs for generation and evaluation (G-FPR = 10^{-4} and T-FPR = 10^{-3}). As shown in Table 2(a), adversarial examples generated against models trained with `softmax` loss exhibit certain transferability compared to models trained with `cosine` loss. We conduct the same experiment by generating adversarial examples with CW and found it has weaker transferability compared to our *SemanticAdv* (results in brackets of Table 2).

As Table 2(b) illustrates, the adversarial examples generated against the model with smaller G-FPR = 10^{-4} exhibit strong attack success rate when evaluating the model with larger T-FPR = 10^{-3}. Especially, we found the adversarial examples generated against R-101-S have the best attack performance on other models. These findings motivate the analysis of the query-free black-box API attack detailed in the following paragraph.

Query-free black-box API attack. In this experiment, we generate adversarial examples against R-101-S with G-FPR = $10^{-3}(\kappa = 1.24)$, G-FPR = $10^{-4}(\kappa =$

Table 2. Transferability of *SemanticAdv*: cell (i, j) shows attack success rate of adversarial examples generated against j-th model and evaluate on i-th model. Results of CW are listed in brackets. Left: Results generated with G-FPR = 10^{-3} and T-FPR = 10^{-3}; Right: Results generated with G-FPR = 10^{-4} and T-FPR = 10^{-3}

$\mathcal{M}_{\text{test}}$ / $\mathcal{M}_{\text{opt}}$	R-50-S	R-101-S	R-50-C	R-101-C
R-50-S	1.000 (1.000)	**0.108** (0.032)	**0.023** (0.007)	**0.018** (0.005)
R-101-S	**0.169** (0.029)	1.000 (1.000)	**0.030** (0.009)	**0.032** (0.011)
R-50-C	**0.166** (0.054)	**0.202** (0.079)	1.000 (1.000)	**0.048** (0.020)
R-101-C	**0.120** (0.034)	**0.236** (0.080)	**0.040** (0.017)	1.000 (1.000)

$\mathcal{M}_{\text{test}}$ / $\mathcal{M}_{\text{opt}}$	R-50-S	R-101-S
R-50-S	1.000 (1.000)	**0.862** (0.530)
R-101-S	**0.874** (0.422)	1.000 (1.000)
R-50-C	**0.693** (0.347)	**0.837** (0.579)
R-101-C	**0.617** (0.218)	**0.888** (0.617)

0.60), and G-FPR $< 10^{-4}(\kappa = 0.30)$, respectively. We evaluate our algorithm on two industry level face verification APIs, namely, Face++ and AliYun. Since attack transferability has never been explored in concurrent work that generates semantic adversarial examples, we use $\mathcal{L}_p$ bounded pixel-wise methods (CW [12], MI-FGSM[18], M-DI^2-FGSM[76]) as our baselines. We also introduce a much strong baseline by first performing attribute-conditioned image editing and running CW attack on the editted images, which we refer as and StarGAN+CW. Compared to CW, the latter two devise certain techniques to improve their transferability. We adopt the ensemble version of MI-FGSM[18] following the original paper. As shown in Table 3, our proposed *SemanticAdv* achieves a much higher attack success rate than the baselines in both APIs under all FPR thresholds (e.g., our adversarial examples generated with G-FPR $< 10^{-4}$ achieves 67.69% attack success rate on Face++ platform with T-FPR = 10^{-3}). In addition, we found that lower G-FPR can achieve higher attack success rate on both APIs within the same T-FPR (see our supplementary material for more details).

Table 3. Quantitative analysis on query-free black-box attack. We use ResNet-101 optimized with `softmax` loss for evaluation and report the attack success rate(%) on two online face verification platforms. Note that for PGD-based attacks, we adopt MI-FGSM ($\epsilon = 8$) in [18] and M-DI^2-FGSM ($\epsilon = 8$) in [76], respectively. For CW, StarGAN+CW and *SemanticAdv*, we generate adversarial samples with G-FPR $< 10^{-4}$

API name	Face++		AliYun	
Attacker / Metric	T-FPR = 10^{-3}	T-FPR = 10^{-4}	T-FPR = 10^{-3}	T-FPR = 10^{-4}
CW [12]	37.24	20.41	18.00	9.50
StarGAN+CW	47.45	26.02	20.00	8.50
MI-FGSM [18]	53.89	30.57	29.50	17.50
M-DI^2-FGSM [76]	56.12	33.67	30.00	18.00
SemanticAdv	**67.69**	**48.21**	**36.50**	**19.50**

SemanticAdv against defense methods. We evaluate the strength of the proposed attack by testing against five existing defense methods, namely, `Feature squeezing` [77], `Blurring` [40], `JPEG` [19], `AMI` [65] and adversarial training [44].

Figure 5 illustrates *SemanticAdv* is more robust against the pixel-wise defense methods comparing with CW. The same G-FPR and T-FPR are used for evalua-

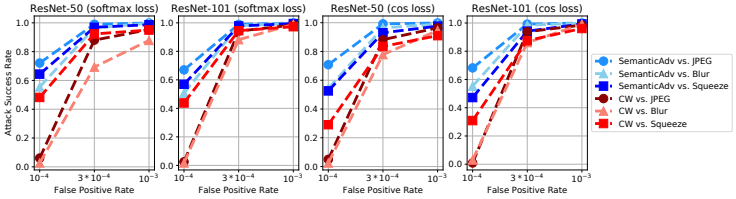

Fig. 5. Quantitative analysis on attacking several defense methods including `JPEG` [19], `Blurring` [40], and `Feature Squeezing` [77]

tion. Both *SemanticAdv* and CW achieve a high attack success rate when T-FPR $= 10^{-3}$, while *SemanticAdv* marginally outperforms CW when T-FPR goes down to 10^{-4}. While defense methods have proven to be effective against CW attacks on classifiers trained with ImageNet [38], our results indicate that these methods are still vulnerable in the face verification system with small G-FPR.

We further evaluate *SemanticAdv* on attribute-based defense method `AMI` [65] by constructing adversarial examples for the pretrained VGG-Face [54] in a black-box manner. From adversarial examples generated by R-101-S, we use `fc7` as the embedding and select the images with normalized L2 distance (to the corresponding benign images) beyond the threshold defined previously. With the benign and adversarial examples, we first extract attribute witnesses with our aligned face images and then leverage them to build a attribute-steered model. When misclassifying 10% benign inputs into adversarial images, it only correctly identifies 8% adversarial images from *SemanticAdv* and 12% from CW.

Moreover, we evaluate *SemanticAdv* on existing adversarial training based defense (the detailed setting is presented in supplementary materials). We find that accuracy of adversarial training based defense method is 10% against the adversarial examples generated by *SemanticAdv*, while is 46.7% against the adversarial examples generated by PGD [44]. It indicates that existing adversarial training based defense method is less effective against *SemanticAdv*, which further demonstrates that our *SemanticAdv* identifies an unexplored research area beyond previous L_p-based ones.

4.3 *SemanticAdv* on Face Landmark Detection

We evaluate the effectiveness of *SemanticAdv* on face landmark detection under two tasks, "Rotating Eyes" and "Out of Region". For the "Rotating Eyes" task, we rotate the coordinates of the eyes with 90° counter-clockwise. For the "Out of Region" task, we set a target bounding box and push all points out of the box. Figure 6 indicates that *semanticAdv* could attack landmark detection models.

4.4 *SemanticAdv* on Street-view Semantic Segmentation

We further generate adversarial perturbations on street-view images to show the generalization of *semanticAdv*. Figure 7 illustrates the adversarial examples on

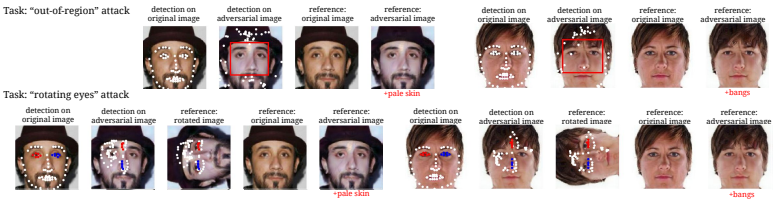

Fig. 6. Qualitative results on attacking face landmark detection model

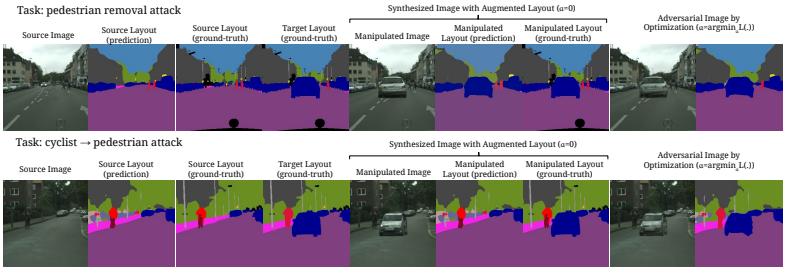

Fig. 7. Qualitative results on attacking street-view semantic segmentation model

semantic segmentation. In the first example, we select the leftmost pedestrian as the target object instance and insert another car into the scene to attack it. The segmentation model has been successfully attacked to neglect the pedestrian (see last column), while it does exist in the scene (see second-to-last column). In the second example, we insert an adversarial car in the scene by *SemanticAdv* and the cyclist has been recognized as a pedestrian by the segmentation model. Details can be found in the supplementary material.

5 Conclusions

Overall we presented a novel attack method *SemanticAdv* to generate semantically meaningful adversarial perturbations guided by single semantic attribute. Experimental evaluations demonstrate several unique properties of *semanticAdv* including attack transferability. We believe *semanticAdv* would open up new research opportunities and challenges in adversarial learning domain. For instance, how to generate semantic adversarial example in physical world and leverage semantic information to defend against such attacks.

Acknowledgments. This work was supported in part by AWS Machine Learning Research Awards, National Science Foundation under grants CNS-1422211, CNS-1616575, CNS-1739517, and NSF CAREER Award IIS-1453651.

References

1. Alibaba Cloud Computing Co. Ltd. https://help.aliyun.com/knowledge_detail/53535.html
2. Megvii Technology Co. Ltd. https://console.faceplusplus.com/documents/5679308
3. Bau, D., Zhu, J.Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A.: Gan dissection: Visualizing and understanding generative adversarial networks. arXiv preprint arXiv:1811.10597 (2018)
4. Bengio, Y., Mesnil, G., Dauphin, Y., Rifai, S.: Better mixing via deep representations. In: ICML (2013)
5. Bhattad, A., Chong, M.J., Liang, K., Li, B., Forsyth, D.: Unrestricted adversarial examples via semantic manipulation. In: International Conference on Learning Representations (2020)
6. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: ICLR (2019)
7. Brown, T.B., Carlini, N., Zhang, C., Olsson, C., Christiano, P., Goodfellow, I.: Unrestricted adversarial examples. arXiv preprint arXiv:1809.08352 (2018)
8. Bulat, A., Tzimiropoulos, G.: Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3706–3714 (2017)
9. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: ICCV (2017)
10. Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q.A., Fu, K., Mao, Z.M.: Adversarial sensor attack on lidar-based perception in autonomous driving. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. pp. 2267–2281 (2019)
11. Cao, Y., Xiao, C., Yang, D., Fang, J., Yang, R., Liu, M., Li, B.: Adversarial objects against lidar-based autonomous driving systems. arXiv preprint arXiv:1907.05418 (2019)
12. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (S&P). IEEE (2017)
13. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
14. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR (2018)
15. Cisse, M., Adi, Y., Neverova, N., Keshet, J.: Houdini: Fooling deep structured prediction models. In: NIPS (2017)
16. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
17. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
18. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)

19. Dziugaite, G.K., Ghahramani, Z., Roy, D.M.: A study of the effect of jpg compression on adversarial images. arXiv preprint arXiv:1608.00853 (2016)
20. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., Madry, A.: A rotation and a translation suffice: Fooling cnns with simple transformations. arXiv preprint arXiv:1712.02779 (2017)
21. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1625–1634 (2018)
22. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. IEEE (2009)
23. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
24. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2014)
25. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV. Springer (2016)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
27. Hong, S., Yan, X., Huang, T.S., Lee, H.: Learning hierarchical semantic image manipulation through structured representations. In: NeurIPS (2018)
28. Huang, Q., Yang, L., Huang, H., Wu, T., Lin, D.: Caption-supervised face recognition: Training a state-of-the-art face model without manual annotation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
29. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017)
30. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. Springer (2016)
31. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: CVPR. pp. 1219–1228 (2018)
32. Joshi, A., Mukherjee, A., Sarkar, S., Hegde, C.: Semantic adversarial attacks: Parametric transformations that fool deep classifiers. arXiv preprint arXiv:1904.08489 (2019)
33. Kang, D., Sun, Y., Hendrycks, D., Brown, T., Steinhardt, J.: Testing robustness against unforeseen adversaries. arXiv preprint arXiv:1908.08016 (2019)
34. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018)
35. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: CVPR. pp. 4873–4882 (2016)
36. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
37. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: CVPR (2015)
38. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
39. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV. IEEE (2009)
40. Li, X., Li, F.: Adversarial examples detection in deep networks with convolutional filter statistics. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5764–5772 (2017)

41. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NIPS (2017)
42. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
43. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
44. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
45. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: CVPR (2015)
46. Mansimov, E., Parisotto, E., Ba, J.L., Salakhutdinov, R.: Generating images from captions with attention. In: ICLR (2015)
47. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2574–2582 (2016)
48. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. pp. 483–499. Springer (2016)
49. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: ICML. JMLR (2017)
50. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. In: NIPS (2016)
51. Oord, A.v.d., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: ICML (2016)
52. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: Security and Privacy (EuroS&P), 2016 IEEE European Symposium on (2016)
53. Parikh, D., Grauman, K.: Relative attributes. In: ICCV. IEEE (2011)
54. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: bmvc. vol. 1, p. 6 (2015)
55. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2015)
56. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
57. Reed, S., Sohn, K., Zhang, Y., Lee, H.: Learning to disentangle factors of variation with manifold interaction. In: ICML (2014)
58. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: ICCV Workshop (2013)
59. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
60. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
61. Song, Y., Shu, R., Kushman, N., Ermon, S.: Constructing unrestricted adversarial examples with generative models. In: Advances in Neural Information Processing Systems. pp. 8312–8323 (2018)
62. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: CVPR (2014)

63. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al.: Going deeper with convolutions. In: CVPR (2015)
64. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
65. Tao, G., Ma, S., Liu, Y., Zhang, X.: Attacks meet interpretability: Attribute-steered detection of adversarial samples. In: NeurIPS (2018)
66. Tong, L., Li, B., Hajaj, C., Xiao, C., Zhang, N., Vorobeychik, Y.: Improving robustness of {ML} classifiers against realizable evasion attacks using conserved features. In: 28th {USENIX} Security Symposium ({USENIX} Security 19). pp. 285–302 (2019)
67. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: CVPR (2018)
68. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)
69. Wong, E., Schmidt, F.R., Kolter, J.Z.: Wasserstein adversarial examples via projected sinkhorn iterations. ICML (2019)
70. Xiao, C., Deng, R., Li, B., Lee, T., Edwards, B., Yi, J., Song, D., Liu, M., Molloy, I.: Advit: Adversarial frames identifier based on temporal consistency in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3968–3977 (2019)
71. Xiao, C., Deng, R., Li, B., Yu, F., Liu, M., Song, D.: Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In: ECCV (2018)
72. Xiao, C., Li, B., Zhu, J.Y., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. In: IJCAI (2018)
73. Xiao, C., Pan, X., He, W., Peng, J., Sun, M., Yi, J., Liu, M., Li, B., Song, D.: Characterizing attacks on deep reinforcement learning. arXiv preprint arXiv:1907.09470 (2019)
74. Xiao, C., Yang, D., Li, B., Deng, J., Liu, M.: Meshadv: Adversarial meshes for visual recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6898–6907 (2019)
75. Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M., Song, D.: Spatially transformed adversarial examples. In: ICLR (2018)
76. 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 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2730–2739 (2019)
77. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017)
78. Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: ECCV. Springer (2016)
79. Yang, L., Chen, D., Zhan, X., Zhao, R., Loy, C.C., Lin, D.: Learning to cluster faces via confidence and connectivity estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
80. Yang, L., Huang, Q., Huang, H., Xu, L., Lin, D.: Learn to propagate reliably on noisy affinity graphs. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)

81. Yang, L., Zhan, X., Chen, D., Yan, J., Loy, C.C., Lin, D.: Learning to cluster faces on an affinity graph. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
82. Yao, S., Hsu, T.M., Zhu, J.Y., Wu, J., Torralba, A., Freeman, B., Tenenbaum, J.: 3d-aware scene manipulation via inverse graphics. In: Advances in neural information processing systems. pp. 1887–1898 (2018)
83. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Computer Vision and Pattern Recognition (CVPR) (2017)
84. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017)
85. Zhang, H., Chen, H., Xiao, C., Gowal, S., Stanforth, R., Li, B., Boning, D., Hsieh, C.J.: Towards stable and efficient training of verifiably robust neural networks. ICLR 2020 (2019)
86. Zhang, X., Yang, L., Yan, J., Lin, D.: Accelerated training for massive classification via dynamic class selection. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
87. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
88. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: CVPR (2016)