Semantic Adversarial Attacks: Parametric Transformations That Fool Deep Classifiers

Ameya Joshi **Amitangshu Mukherjee** **Soumik Sarkar** **Chinmay Hegde**[*]
Iowa State University
{ameya, amimukh, soumiks, chinmay}@iastate.edu

Abstract

Deep neural networks have been shown to exhibit an intriguing vulnerability to adversarial input images corrupted with imperceptible perturbations. However, the majority of adversarial attacks assume global, fine-grained control over the image pixel space. In this paper, we consider a different setting: what happens if the adversary could only alter specific attributes of the input image? These would generate inputs that might be perceptibly different, but still natural-looking and enough to fool a classifier. We propose a novel approach to generate such "semantic" adversarial examples by optimizing a particular adversarial loss over the range-space of a parametric conditional generative model. We demonstrate implementations of our attacks on binary classifiers trained on face images, and show that such natural-looking semantic adversarial examples exist. We evaluate the effectiveness of our attack on synthetic and real data, and present detailed comparisons with existing attack methods. We supplement our empirical results with theoretical bounds that demonstrate the existence of such parametric adversarial examples.

1 Introduction

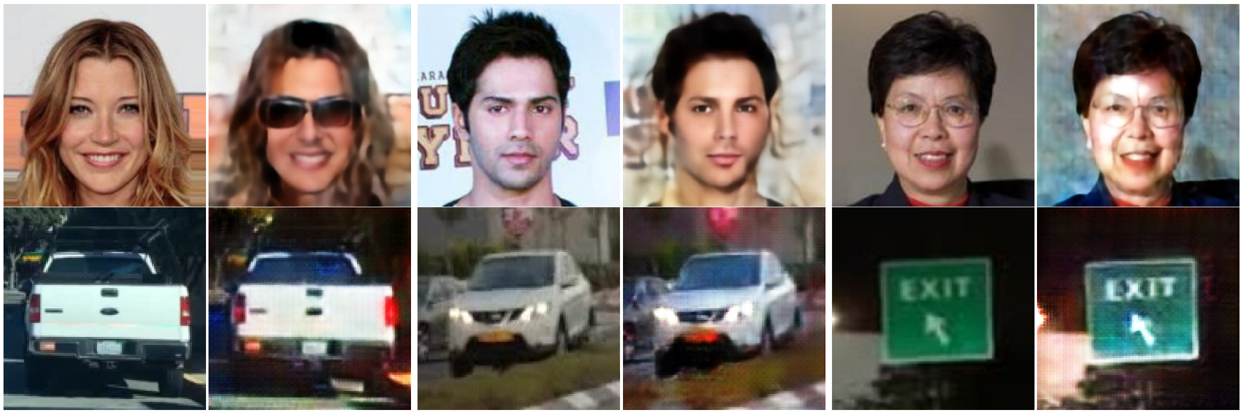

Figure 1: *Examples of semantic adversarial attacks with a single modifiable attribute. The first and third columns are original images. Semantic adversarial examples (Columns 2 and 4) are generated by optimizing over the manifold of a parametric generative model to fool a binary gender classifier.*

The existence of adversarial inputs for deep neural network-based classifiers has been well established by several recent works [5, 10, 16, 17, 58, 41]. The adversary typically confounds the classifier by adding an *imperceptible* perturbation

[*] This work was supported in part by NSF grants CCF-1750920, CNS-1845969, DARPA AIRA grant PA-18-02-02, AFOSR YIP Grant FA9550-17-1-0220, an ERP grant from Iowa State University, a GPU gift grant from NVIDIA corporation, and faculty fellowships from the Black and Veatch Foundation.

to a given input image, where the range of the perturbation is defined in terms of bounded pixel-space ℓ_p-norm balls. Such adversarial "attacks" appear to catastrophically affect the performance of state-of-the-art classifiers [1, 22, 23, 54].

Pixel-space norm-constrained attacks reveal interesting insights about generalization properties of deep neural networks. However, imperceptible attacks are certainly not the only means available to an adversary. Consider an input example that comprises salient, invariant features along with modifiable attributes. An example would be an image of a face, which consists of invariant features relevant to the identity of the person, and variable attributes such as hair color and presence/absence of eyeglasses. Such adversarial examples, though perceptually distinct from the original input, appear natural and acceptable to an oracle or a human observer but would still be able to subvert the classifier. Unfortunately, the large majority of adversarial attack methods do not port over to such natural settings.

A systematic study of such attacks is paramount in safety-critical applications that deploy neural classifiers, such as face-recognition systems or vision modules of autonomous vehicles. These systems are required to be immune to a limited amount of variability in input data, particularly when these variations are achieved through natural means. Therefore, a method to generate adversarial examples using natural perturbations, such as facial attributes in the case of face images, or different weather conditions for autonomous navigation systems, would shed further insights into the real-world robustness of such systems. We refer to such perceptible attacks as "semantic" attacks.

This setting fundamentally differs from existing attack approaches and has been (largely) unexplored thus far. Semantic attacks utilize nonlinear generative transformations of an input image instead of linear, additive techniques (such as image blending). Such complicated generative transformations display higher degrees of nonlinearity in corresponding attacks, the effects of which warrant further investigation. In addition, the role of the number of modifiable attributes (parameters in the generative models) in the given input is also an important point of consideration.

Contributions: We propose and rigorously analyze a framework for generating adversarial examples for a deep neural classifier by modifying *semantic* attributes.

We leverage generative models such as Fader Networks [30] that have semantically meaningful, tunable attributes corresponding to parameters into a continuous bounded space that implicitly define the space of "natural" input data. Our approach exploits this property by treating the range space of these attribute models as a manifold of semantic transformations of an image.

We pose the search for adversarial examples on this *semantic manifold* as an optimization problem over the parameters conditioning the generative model. Using face image classification as a running test case, we train a variety of parametric models (including Fader Networks and Attribute GANs), and demonstrate the ability to generate semantically meaningful adversarial examples using each of these models. In addition to our empirical evaluations, we also provide a theoretical analysis of a simplified semantic attack model to understand the capacity of parametric attacks that typically exploit a significantly lower dimensional attack space compared to the classical pixel-space attacks.

Our specific contributions are as follows:

1. We propose a novel optimization based framework to generate semantically valid adversarial examples using parametric generative transformations.
2. We explore realizations of our approach using variants of multi-attribute transformation models: Fader Networks [30] and Attribute GANs [20] to generate adversarial face images for a binary classifier trained on the CelebA dataset [37]. Some of our modified multi-attribute models are non-trivial and may be of independent interest.
3. We present an empirical analysis of our approach and show that increasing the *dimensionality* of the attack space results in more effective attacks. In addition, we investigate a sequence of increasingly nonlinear attacks, and demonstrate that a higher degree of *nonlinearity* (surprisingly) leads to weaker attacks.
4. Finally, we provide a preliminary theoretical analysis by providing upper bounds for the *classification error* for a simplified surrogate model under adversarial condition [52]. This analysis supports our empirical observations regarding the dimensionality of the attack space.

We demonstrate the effectiveness of our attacks on simple deep classifiers trained over complex image datasets; hence, our empirical comparisons are significantly more realistic than popular attack methods such as FGSM [16] and PGD [29, 39] that primarily have focused on simpler datasets such as MNIST [32] and CIFAR. Our approach also presents an interesting use-case for multi-attribute generative models which have been used solely as visualization tools thus far.

Outline: We begin with a review of relevant literature in Section 2. We describe our proposed framework, *Semantic Adversarial Generation*, in section 3. In Section 4 we describe two variants of our framework to show different methods of ensuring the semantic constraint. We provide empirical analysis of our work in Section 5. We further present

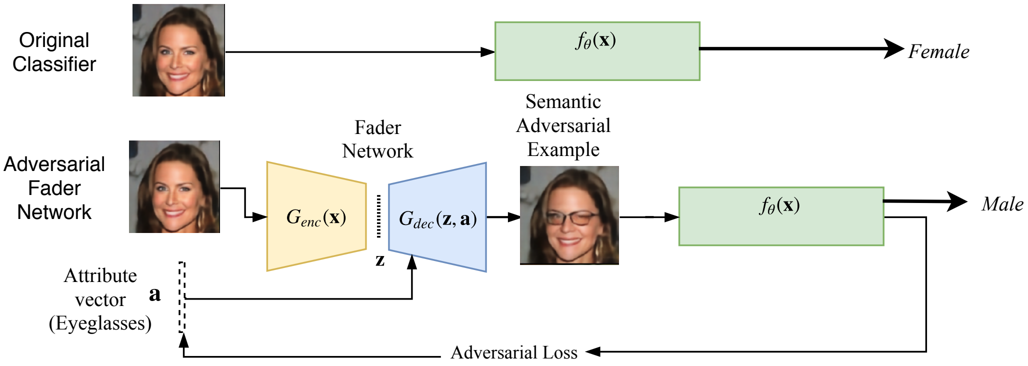

Figure 2: *An single-attribute* ***Adversarial*** *Fader Network. The semantic adversarial attack framework optimizes an adversarial loss to generate an adversarial direction. Backpropagating the adversarial direction through the Fader Network with respect to the attribute vector, $\mathbf{a}$, ensures that the adversarial example is only generated for that specific attribute. Here, the adversarial algorithm generates eyeglasses on a face of a Female by optimizing $\mathbf{a}$, thus forcing the gender classifier to misclassify the image as Male.*

empirical analysis and theoretical qualification on the dimensionality of the parametric attack space in Section 6, and conclude with possible extensions in Section 7.

2 Related Work

Due to space constraints coupled with the large amount of recent progress in the area of adversarial machine learning, our discussion of related work is necessarily incomplete. We defer a more detailed discussion to the appendix.

Our focus is on *white box, test-time* attacks on deep classification systems; other families of attacks (such as backdoor attacks, data poisoning schemes, and black-box attacks) are not directly relevant to our setting, and we do not discuss those methods here.

Adversarial Attacks: Evidence that deep classifiers are susceptible to imperceptible adversarial examples can be attributed to Szegedy *et al.* [58]. Goodfellow *et al.* [16] and Kurakin *et al.* [29] extend this line of work using the Fast Gradient Sign Method (FGSM) and its iterative variants. Carlini and Wagner [5] devise state-of-the-art attacks under various pixel-space l_p norm-ball constraints by proposing multiple adversarial loss functions. Athalye *et al.* [1] further analyze several defense approaches against pixel-space adversarial attacks, and demonstrate that most existing defenses can be surpassed by approximating gradients over defensively trained models.

Such attacks perturb the pixel-space under an imperceptibility constraint. On the contrary, we approach the problem of generating adversarial examples that have perceptible yet semantically valid modifications. Our method considers a smaller 'parametric' space of modifiable attributes that have physical significance.

Parametric Adversarial Attacks: Parametric attacks are a recently introduced class of attacks in which the attack space is defined by a set of parameters rather than the pixel space. Such approaches result in more "natural" adversarial examples as they target the image formation process instead of the pixel space. Recent works by Athalye *et al.* [2] and Liu *et al.* [35] use optimization over geometric surfaces in 3D space to create adversarial examples. Zhang *et al.* [71] demonstrate the existence of adversarially designed textures that can camouflage vehicles. Zhao *et al.* [72] generate adversarial examples by using the parametric input latent space of GANs[18]. Xiao *et al.* [65] employ spatial transforms to perturb image geometry for creating adversarial examples. Sharif *et al.* [55] propose a generative model to alter images of faces with eyeglasses in order to confound a face recognition classifier. Contrary to these methods, we consider the inverse approach of using a pre-trained multi-attribute generative model to transform inputs over multiple attributes for generating adversarial examples.

Song *et al.* [57] optimize over the latent space of a conditional GAN to generate unrestricted adversarial examples for a gender classifier. While our approach is thematically similar, we fundamentally differ in the context of being able to generate adversarial counterparts for given test samples while providing a finer degree of control using multi-attribute generative models. We discuss relevant literature regarding such multi-attribute generative models below.

Attribute-Based Conditional Generative Models: Generative Adversarial Networks (GAN) [18] are a popular approach for the generation of samples from a real-world data distribution. Recent advancements [49, 36, 64, 6] in

GANs allow for creation of high quality realistic images. Chen *et al.* [6] introduce the concept of a attribute learning generative model where visual features are parametrized by an input vector.

Perarnaue *et al.* [48] use a Conditional Generative Adversarial Network [40] and an encoder to learn the attribute invariant latent representation for attribute editing. Fader Networks [30] improve upon this using an auto-encoder with a latent discriminator. He *et al.* [20] argue that such an attribute invariant constraint is too constrictive and replace it an attribute classification constraint and a reconstruction loss instead to alter only the desired attributes preserving attribute-excluding features. These models are primarily used for generation of a large variety of facial images. We provide a secondary (and perhaps practical) use case for such attribute models in the context of understanding generalization properties of neural networks.

3 Semantic Attacks

Conceptually, producing an adversarial semantic ("natural") perturbation of a given input depends on two algorithmic components: (i) the ability to navigate the manifold of parametric transformations of an input image, and (ii) the ability to perform optimization over this manifold that maximizes the classification loss with respect to a given target model. We describe each component in detail below.

Notation: We assume a white-box threat model, where the adversary has access to a target model $f(\mathbf{x}) : \mathbb{R}^d \to \{0, 1\}^c$ and the gradients associated with it. The model classifies an input image, $\mathbf{x}$ into one of c classes, represented by a one-hot output label, $\mathbf{y}$. In this paper, we focus on binary classification models ($c = 2$) while noting that our framework transparently extends to multi-class models. Let $G(\mathbf{x}, \boldsymbol{a}) : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$ denote parametric transformations, conditioned on a parameter vector, $\boldsymbol{a}$. Here, each element of $\boldsymbol{a}$ (say, a_i) is a real number that corresponds to a specific semantic attribute. For example, a_0 may correspond to facial hair, with a value of zero (or negative) denoting absence and a positive value denoting presence of hair on a given face example. We define a semantic adversarial attack as the deliberate process of transforming an input image, $\mathbf{x}$ via a parametric model to produce a new example $\tilde{\mathbf{x}} = G(\mathbf{x}, \mathbf{a})$ such that $f(\tilde{\mathbf{x}}) \neq f(\mathbf{x})$.

3.1 Parametric Transformation Models

First, let us consider the problem of generating semantic transformations of a given input example. In order to create semantically transformed examples, the defined parametric generative model $G(.)$ should satisfy two properties: $G(.)$ should reconstruct the invariant data in an image, and $G(.)$ should be able to independently perturb the semantic attributes while minimally changing the invariant data.

The parametric transformation model therefore, is trained to reconstruct the original example while disentangling the semantic attributes. This involves conditioning the generative model on a set of parameters corresponding to the modifiable attributes. The *semantic parameter vector* consists of these parameters and is input to the parametric model to control the expression of semantic attributes.

We argue that the range-space of such a model approximates the manifold of the semantic transformations of input images. Therefore, the transformation model can be used a projection operator to ensure that a solution to an optimization problem will lie in the set of semantic transformations of an input image. We also observe that the semantic parameter vectors will be much lower in dimension than the original image.

In this paper, we consider two variants of such conditional generative models: Fader Networks [30] and AttributeGANs (AttGAN) [20].

3.2 Adversarial Parameter Optimization

The problem of generating a semantic adversarial example essentially can be thought of as finding the right set of attributes that a classifier is adversarially susceptible to. In our approach, we model this as an optimization problem over the semantic parameters.

The generation of adversarial examples is generally modelled as an optimization problem that can be broken down into two sub problems: (1) Optimization of an adversarial loss over the target network to find the direction of an adversarial perturbation. (2) Projection of the adversarial vector on the viable solution-space.

In the first step, we optimize over an adversarial loss, L_{adv}. We model the second step as a projection of the adversarial vector onto the range space of a parametric transformation model. This is achieved by cascading the output of the transformation function to the input of our target network. The optimization problem can then be solved by back-propagating over both the network and the transform. We also modify the Carlini-Wagner untargeted adversarial

Algorithm 1 Adversarial Parameter Optimization

```
Require: $\mathbf{x}_0$:Input image, $\mathbf{a}_0$:Initial attribute vector, $E(.)$: Attribute encoder, $G(.,.)$:Pre-trained parametric transformation model, $f(.)$: Target classifier, $\mathbf{y}$ : Original label
$h_0 \leftarrow f(\mathbf{x})$, $l_{\text{adv}} \leftarrow \infty$, $i \leftarrow 0$
success = 0
while do $l_{\text{adv}} \neq 0$ and $i \leq$ MaxIter
    $\bar{\mathbf{a}} \leftarrow E(\mathbf{a})$
    $h_i \leftarrow f(G(\mathbf{x}_i, \bar{\mathbf{a}}_i))$
    $l_{\text{adv}} \leftarrow L_{\text{adv}}(\mathbf{y}, h_i)$
    $\mathbf{a}_{i+1} \leftarrow$ BackProp $\{\mathbf{a}_i, \nabla l_{\text{adv}}\left(f(G(\mathbf{x}_i, E(\mathbf{a}_i)))\right)\}$
    $\tilde{\mathbf{x}} \leftarrow G(\mathbf{x}, E(\mathbf{a}_{i+1}))$
    if $f(\mathbf{x}) \neq f(\tilde{\mathbf{x}})$ then
        return success, $\tilde{\mathbf{x}}$
    end if
    $i \leftarrow i + 1$
end while
```

loss [5] as shown in equation 1 to include our semantic constraint:

$$\max\left(0, \max_{t \neq i}\left(f(\tilde{\mathbf{x}})_t\right) - f(\tilde{\mathbf{x}})_i\right) \tag{1}$$
$$\text{s.t.} \quad \tilde{\mathbf{x}} = G(\mathbf{x}, \mathbf{a})$$

where i is the original label index and t are the class label indices for any of the other classes.

In comparison to the grid search method presented in Zhao *et al.* [72] and Engstrom *et al.* [12], our optimization algorithm scales better. In addition, we create semantic adversarial transformations with multiple attributes for a specific input allowing for a fine-grained analysis of the generalization capacities of the target model.

4 Semantic Transformations

While our semantic attack framework is applicable to any parametric transformation model that enables gradient computations, we instantiate it by constructing adversarial variants of two recently proposed generative models: Fader networks [30] and AttributeGANs (AttGAN) [20].

4.1 Adversarial Fader Network

A Fader Network [30] is an encoder-decoder architecture trained to modify images with continuously parameterized attributes. They achieve this by learning an invariance over the encoded latent representation while disentangling the semantic information of the images and attributes. The invariance of the attributes is learnt by an adversarial training step in the latent space with the help of a latent discriminator which is trained to identify correct attributes corresponding to each training sample.

Using our framework, we can adapt any pre-trained Fader Network to model the manifold of semantic perturbations of a given input. We note that minor adjustments are needed in our setting, since the parameter vector required by the approach of [30] requires each scalar attribute, a_i, to be represented by a tuple, $(1 - a_i, a_i)$. Since there is a one-to-one mapping between the two representations, we can project any real-valued parameter vector $\mathbf{a}$ into this tuple form via an additional, fixed affine transformation layer. Given this extra "attribute encoding" step, all gradient computations proceed as before. We quantitatively study the effect of allowing the attacker access to single or multiple semantic attributes. In particular, we construct three approaches for generating semantic adversarial examples: (i) A single attribute Fader Network; (ii) A multi-attribute Fader Network; and (iii) A cascaded sequence of single attribute Fader Networks.

Single Attribute Attack: For the single attribute attack, we use the range-space of a pre-trained, single attribute Fader Network to constrain our adversarial attack. The single attribute attack constrains an attacker to only modify a specified attribute for all the images. In the case of face images, such attributes might include presence/absence of eyeglasses, hair color, and nose shape.

In our experiments, we present examples of attacks on a gender classifier using three separate single attributes: eyeglasses, age, and skin complexion. Figure 2 describes the mechanism of a single-attribute adversarial Fader Network that generates an adversarial example by adding eyeglasses.

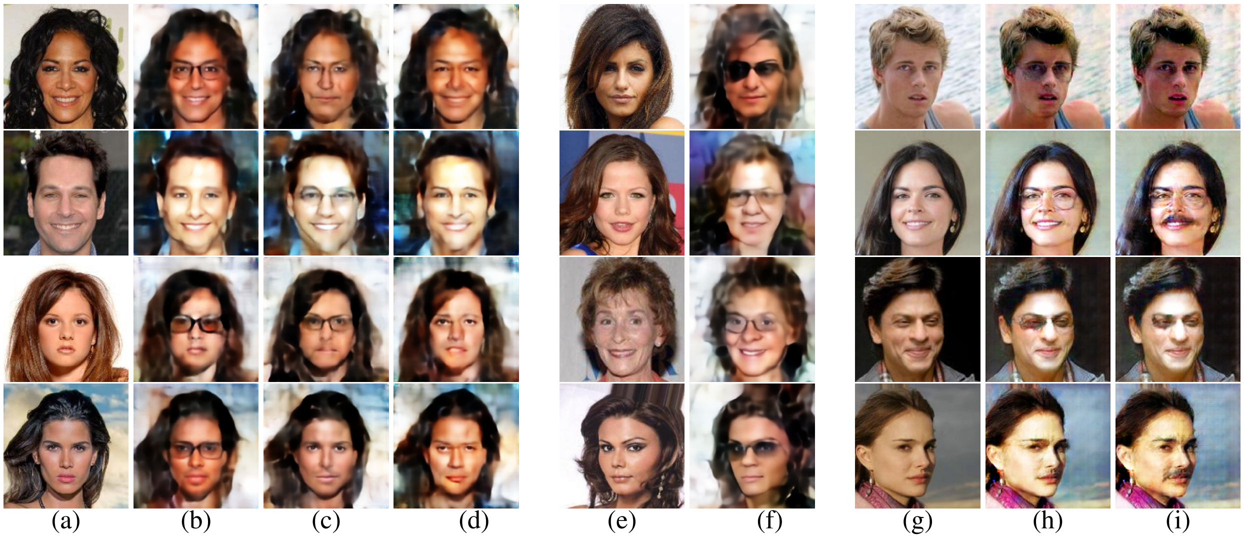

Figure 3: *Semantic adversarial examples generated with multiple attribute semantic models as in table 1. Columns (a), (e) and (g) are original images. Columns (b){Attribute category: A1,A5,A6} (c){Attribute category: A1,A2,A7} and (d){Attribute category: A2,A5,A6} show examples generated using multi-attribute Fader Networks as semantic transforms. Examples in (f){Attribute category: A1-A2-A3} were generated using cascaded single attribute Fader Network. Columns (h){Attribute category: A1,A2,A6,A8,A10} and (i){Attribute category: A1,A2,A6,A8,A9,A10} are images transformed using an AttGAN with 5 and 6 attributes respectively. Additional results of semantic attacks on multi-class classifiers for traffic scenes are provided in the appendix.*

Multi-Attribute Attack: Similar to the single-attribute case, we may also use pre-trained multi-attribute Fader Networks to model cases where the adversary has access to multiple modifiable traits.

A limitation of multi-attribute Fader Networks lies in the difficulty of their training. This is because a Fader Network is required to learn disentangled representations of the attributes while in practice, semantic attributes cannot be perfectly decoupled. We resolve this using a novel conditional generative model described as follows.

Cascaded Attribute Attack: We propose a novel method to simulate multi-attribute attacks by stage-wise concatenation pre-trained single attribute Fader networks. The benefit is that the computational burden of learning disentangled representations is now removed.

Each single-attribute model exposes a attribute latent vector. During execution of Alg. 1 we jointly optimize over all the attribute vectors. The optimal adversarial vector is then segmented into corresponding attributes for each Fader Network to generate an adversarial example.

4.2 Adversarial AttGAN

A second encoder-decoder architecture [20], known as AttGAN, achieves a similar goal as Fader Networks of editing attributes by manipulating the encoded latent representation; however, AttGAN disentangles the semantic attributes from the underlying invariances of the data by considering both the original and the flipped labels while training. This is achieved by training a latent discriminator and classifier pair to classify both the original and the transformed image to ensure invariance.

In order to generate semantic adversarial examples using AttGAN, we use a pretrained generator conditioned on 13 attributes. The attribute vector in this case, is encoded to be a perturbation of the original sequence of attributes for the image. We consider the two sets of attributes listed in Table 1 to generate adversarial examples. In our experience, the AttGAN architecture provides a more stable reconstruction, thus allowing for more modifiable parameters.

5 Experimental Results

We showcase our semantic adversarial attack framework using a binary (gender) classifier as the target model trained on the CelebA dataset [37]. While we restrict ourselves to results on binary classifers on faces in this paper, additional results with multi-class classifiers on the Berkeley Deep Drive dataset [69] can be found in the appendix (refer Fig. 8). All experiments were performed on a single workstation equipped with an NVidia Titan Xp GPU in PyTorch [47]

Attack Type	Attributes	Accuracy of target model (%)	Random Sampling (%)
Single Attribute Attack	A1	52.0	89.00
	A2	35.0	96.00
	A3	14.0	90.00
Multi Attribute Attack	A1,A5,A6	3.00	89.00
	A2,A5,A6	1.00	81.00
	A1,A2,A7	3.00	87.00
Cascaded Multi Attribute Attack	A1-A2-A3	18.00	55.6
	A2-A3-A4	20.00	93.00
Multi Attribute AttGAN Attack	A1,A2,A6,A8,A10	70.40	32.80
	A1,A2,A6,A8,A9,A10	39.40	40.40

Table 1: *Performance of the Semantic Adversarial Example under multiple implementations. Legend for attributes: A1-Eyeglasses, A2-Age, A3-Nose shape, A4-Eye shape, A5-Chubbiness, A6-Pale Skin, A7-Smiling, A8-Mustache, A9-Eyebrows, A10-Hair color. As the number of attributes increase, semantic attacks are more effective. Our optimization-based attack fares better as compared to worst-of-10 random sampling [12], showing the former's efficacy at finding semantic adversarial examples.*

v1.0.0.[2] We train the classifier using the ADAM optimizer [26] over the categorical cross-entropy loss.The training data is augmented with random horizontal flipping to ensure that the classifier does not overfit. The target model achieves a (standard) accuracy of 99.7% on the test set (10% of the dataset).

Our goal is to break this classifier model using semantic attacks. To do so, we use a subset of 500 randomly selected images from the test set. Each image is transformed by our algorithm using the various parametric transformation families described in Section 4. Our metric of comparison for all adversarial attacks is the target model accuracy on the generated adversarial test set.

Adversarial Fader Networks: We consider the three approaches documented in section 4.1. For every image in our original test set, we generate adversarial examples by optimizing the adversarial loss in equation 1 with respect to the corresponding attribute parameters.

In the cases of single-attribute and cascaded sequential attacks, we use the pre-trained single-attribute models provided by Lample *et al.* [30] to represent the manifold of semantic transformations. For the multi-attribute attack, we train 3 multi-attribute Fader Networks with the attributes presented in Table 1. We create an adversarial test set for each our approaches as described in Section 4.1 using our algorithm as defined in Algorithm 1.

Our experiments show that Adversarial Fader Networks successfully generate examples that confound the binary classifier in all cases; see Table 1. Visual adversarial examples are displayed in Figure 1 and Figure 3. We also observe that multi-attribute attacks outperform single-attribute attacks, which conforms with intuition; a more systematic analysis of the effect of the number of semantic attributes on attack performance is provided below in Section 6.

Adversarial AttGAN: We perform a similar set of experiments using the multi-attribute AttGAN implementation of He *et al.*[20]. We record the performance over two experiments: one using 5 attributes, and the second using 6 attributes, as seen in Table 1. We observe a significant improvement in performance as the number of semantic attributes increases (in particular, adding the eyebrows attribute results in nearly a 30% drop in model accuracy).

Comparison with parameter-space sampling: We compare our method with a previously-proposed approach that investigates parametric attacks *et al.* [12]. They propose picking s random samples from the parameter space and choose the adversarial example generated by the sample giving the worst cross entropy loss (we use $s = 10$).

We showcase the results in Table 1, and observe that in all cases (but one), our semantic adversarial attack algorithm outperforms random sampling. In addition, the table also reveals that random examples in the range of Fader Networks or AttGANs are mostly classified correctly. This suggests that the target model is generally invariant to the low reconstruction error incurred by the parametric transformation models[3].

Comparison with pixel-space attacks: In addition to our analyses described above, we also compare our attacks with the state-of-the-art Carlini-Wagner(CW) l_∞-attack [5] as well as several other attack techniques [16, 29, 12] in Table 2. To ensure fair comparison, we consider the maximum l_∞ distance over our multi-attribute attacks as the bound parameter ϵ for all pixel-norm based attacks. From the table, we observe that the CW attack is extremely effective; on the other hand, our semantic attacks are able to outperform other methods such as FGSM [17] and PGD [39].

[2] Code and models: https://github.com/ameya005/Semantic_Adversarial_Attacks

[3] We do not compare our work with other approaches such as the Differentiable Renderer [35] and 3D adversarial attacks [70], since these papers expect oracle access to a 3D rendering environment. We also do not compare with Song *et al.*[57] since they generate adversarial examples from scratch, whereas our attack targets specific inputs.

Attack ($\epsilon = 1.74$)	Accuracy(%)
Single Att. Semantic Attack	**14.01**
Multi Att. Semantic Attack	**1.00**
FGSM [16]	91.6
PGD [39, 29]	26.2
CW-l_∞ [5]	0.00
Spatial [12]	41.00

Table 2: *Comparison of adversarial attacks with other attack strategies. A lower target accuracy corresponds to a better attack. The pixel space attacks are allowed to generate adversarial examples under the l_∞ distance corresponding to our best performing multi-attribute attack model. Observe that semantic attacks are comparable to the state of the art pixel-space attack.*

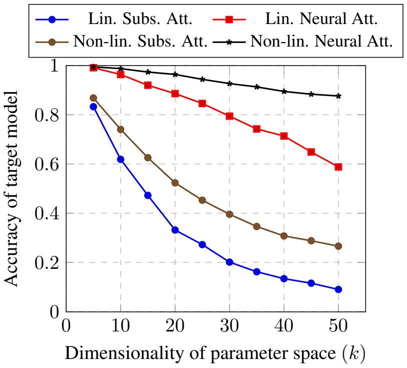

Figure 4: *Effect of dimensionality of the parametric attack space. Considering subspace and rank constrained transforms to generate adversarial examples, note that the target model accuracy decreases as the dimensionality of the attack space increases. The additive attack (surrogate to PGD) is more effective than multiplicative attack(similar to our approach) over all values of k.*

We also compare our approach to *Spatial Attacks* of [12], which uses a grid search over affine transformations of an input to generate adversarial examples; ℓ_∞ constraints do not apply here, and instead we use default parameters provided in [12]. Our proposed attack methods are considerably more successful. We provide additional detailed experiments on binary and multi-class classifiers for other attributes as well as other datasets in the appendix.

6 Analysis: Impact of Dimensionality

From our experiments, we observe that limiting the adversary to a low-dimensional, semantic parametric transformation of the input leads to less-effective attacks than pixel-space attacks (at least when the same loss is optimized). Moreover, single-attribute semantic attacks are more powerful than multi-attribute attacks. This observation makes intuitive sense: the dimension of the manifold of perturbed inputs effectively represents the *capacity* of the adversary, and hence a greater number of degrees of freedom in the perturbation should result in more effective attacks. In pixel-space attacks, the adversary is free to search over a high-dimensional ℓ_p-ball centered around the input example, which is perhaps why ℓ_p-norm attacks are so hard to defend against [1].

In this section, we provide experimental and theoretical analysis that precisely exposes the impact of the dimensionality of the attribute parameters. While our analysis is stylized and not directly applicable to deep neural classifiers, it constitutes a systematic first attempt towards upper bounds on what a semantically constrained adversary can possibly hope to achieve.

6.1 Synthetic Experiments

We propose and analyze the following synthetic setup which enables explicit control over the dimension of the semantic perturbations.

Data: We construct a dataset of $n = 500$ samples from a mixture of Gaussians (MoG) with 10 components (denoted by $\mathbb{P}_d$) defined over $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$. Each data sample is obtained by uniformly sampling one of the mixture component means, and then adding random Gaussian noise with standard deviation $\sigma \leq \sqrt{d}$. The component means are chosen as 10 randomly selected images (1 for each digit) from the MNIST dataset [32] rescaled to 10×10 (i.e., the ambient dimension is $d = 100$).

Target Model: We artificially define two classes: the first class containing images generated from digits 0-4 and the second class containing images from samples 5-9. We train a simple two-layer fully connected network, $f(\mathbf{x}) : \mathbb{R}^d \to \{\pm 1\}$ as the target model. The classifier is trained by optimizing cross-entropy using ADAM [26] for 50 epochs, resulting in training accuracy of 100%, validation accuracy of 99.8%, and test accuracy of 99.6%.

Parametric Transformations: We consider a stylized transformation function, $G(\mathbf{x}, \delta) : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$. We study the effect of varying k for two specific parametric transformation models.

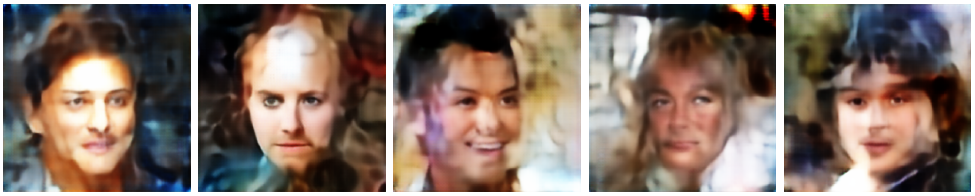

Figure 5: *Semantically transformed single-attribute examples which are classified correctly by the target model but show severe artifacts. This shows that neural networks are immune to significant changes in the semantic domain unlike the pixel domain.*

Subspace attacks: We first consider an additive (linear) attack model. Here, the manifold of semantic perturbations is constrained to lie a k-dimensional subspace spanned by an arbitrary matrix $\mathbf{U} \in \mathbb{R}^{d \times k}$, whose columns are assumed to be orthonormal, and $\delta \in \mathbb{R}^k$

$$G(\mathbf{x}, \delta) := \tilde{\mathbf{x}} = \mathbf{x} + \mathbf{U}\mathbf{U}^T \delta \tag{2}$$

Neural attacks: Next, we consider a *multiplicative* attack model. Here the manifold of perturbations corresponds to a rank-k transformation of the input.

$$G(\mathbf{x}, \delta) := \tilde{\mathbf{x}} = \mathbf{U}.diag(\delta).\mathbf{U}^T \mathbf{x} \tag{3}$$

Here, $\mathbf{U}$ and δ follow the definition presented earlier. This transformation can be interpreted as the action of a shallow (two-layer) auto-encoder network with k hidden neurons with scalar activations parameterized by δ.

Nonlinear ReLU variants: We also consider each of the above two attacks in the *rectified* setting where the transformation is passed through a rectified linear unit: $\tilde{\mathbf{x}} = \mathrm{ReLU}\left(G(\mathbf{x}, \delta)\right)$.

Results: We analyse the effect of the dimensionality of the attack space(k) by considering the performance of the subspace and neural attacks on the target binary classifier. Figure 4 shows the comparison of the constrained attacks for the linear and non-linear cases.

We infer the following: (i) as expected, increasing dimensionality of the semantic attack space leads to less accurate target models; (ii) adding a non-linearity to the transformation function *reduces* the viability of both subspace- and rank-constrained attacks; (iii) subspace-constrained attacks are more powerful than neural attacks. In general, the degree of "nonlinearity" in the transformation model appears to be inversely proportional to the power of the corresponding semantic attack. We believe this phenomenon is somewhat surprising, and defer further analysis to future work.

6.2 Theory

In the case of subspace attacks, we explicitly derive upper bounds on the generalization behavior of target models. Our derivation follows the recent approach of Schmidt *et al.* [52], who consider a simplified version of the data model defined in Section 6.1 and bound the performance of a linear classifier in terms of its *robust classification error*.

Def. 6.1 (Robust Classification Error)**.** Let $\mathbb{P}_d : \mathbb{R}^d \times \{\pm 1\} \rightarrow \mathbb{R}$ be a distribution and let $\mathcal{S}$ be any set containing $\mathbf{x}$. Then the $\mathcal{S}$-robust classification error of any classifier $f : \mathbb{R}^d \rightarrow \{\pm 1\}$ is defined as $\beta = P_{(\mathbf{x},y)\sim\mathbb{P}_d}\left[\exists \tilde{\mathbf{x}} \in \mathcal{S} : f(\tilde{\mathbf{x}}) \neq y\right]$.

Using this definition, we analyze the efficacy of subspace attacks on a simplified linear classifier trained using a mixture of two spherical Gaussians. Consider a dataset with samples $(\mathbf{x}, y) \in \mathbb{R}^d \times \{\pm 1\}$ sampled from a mixture of two Gaussians with component means $\pm\theta^\star$ and standard deviation $\sigma \leq \sqrt{d}$. We assume a linear classifier $f_{\hat{\boldsymbol{w}}}$, defined by the unit vector $\hat{w}$, as $f_{\hat{\boldsymbol{w}}}(\mathbf{x}) = \mathrm{sign}(\langle \hat{\boldsymbol{w}}, \mathbf{x} \rangle)$.

Let $\mathcal{S}_\epsilon = \{\tilde{\mathbf{x}} \mid \tilde{\mathbf{x}} = \mathbf{x} + \mathbf{U}\mathbf{U}^T\delta,\ ||\mathbf{x} - \tilde{\mathbf{x}}||_\infty \leq \epsilon\}$. Assuming that the target classifier is well-trained (i.e., $\hat{\boldsymbol{w}}$ is sufficiently well-correlated with the true component mean $\theta^\star$), we can upper bound the probability of error incurred by the classifier when subjected to any subspace attack.

Theorem 1 (Robust classification error for subspace attacks)**.** *Let $\hat{\boldsymbol{w}}$ be such that $\langle \hat{\boldsymbol{w}}, \theta^\star \rangle \geq k||\mathbf{U}||_{\infty,1}||\hat{\boldsymbol{w}}^T\mathbf{U}||_\infty\epsilon$. Then, the linear classifier $f_{\hat{\boldsymbol{w}}}$ has a $\mathcal{S}_\epsilon$-robust classification error upper bounded as:*

$$\beta \leq \exp\left(-\frac{\left(\langle \hat{\boldsymbol{w}}, \theta^\star \rangle - k||\mathbf{U}||_{\infty,1}||\hat{\boldsymbol{w}}^T\mathbf{U}||_\infty\epsilon\right)^2}{2\sigma^2}\right) \tag{4}$$

The proof is deferred to the appendix, but we provide some intuition. Lemma 20 of [52] recovers a similar result, albeit with the $\sqrt{k}$ term in the exponent being replaced by $\sqrt{d}$. This is because they only consider bounded ℓ_∞-perturbations in pixel-space, and hence their bound on the robust classification error scales exponentially according to the *ambient dimension* d, while our bound is expressed in terms of the *number of semantic attributes* $k \ll d$. A natural next step would be to derive *sample complexity* bounds analogous to [52] but we do not pursue that direction here.

7 Discussion and Conclusions

We conclude with possible obstacles facing our approach and directions for future work. We have provided evidence that there exist adversarial examples for a deep neural classifier that may be perceptible, yet are semantically meaningful and hence difficult to detect. A key obstacle is that parameters associated with semantic attributes are often difficult to decouple. This poses a practical challenge, as it is difficult to train a conditional generative model with independent latent semantic dimensions. However, the success of recent efforts in this direction, including Fader Networks [30], AttGans [20], and StarGANs [8] demonstrate promise of our approach: any newly developed conditional generative models can be used to mount a semantic attack using our framework.

Despite the existence of semantic adversarial examples, we have found that enforcing semantic validity confounds the adversary's task, and that target models are generally able to classify a significant subset of the examples generated under our semantic constraint. Figure 5 are examples of images generated with severe artifacts, yet that are successfully classified. This presents the question: is "naturalness" a strong defense?

This intuition is the premise of a recent defense strategy called DefenseGAN [51]. Indeed, our approach can be viewed as converse of this strategy: DefenseGAN uses the range-space of a generative model (specifically, a GAN) to *defend* against pixel-space attacks, while conversely, we use the same principle to *attack* trained target models. A closer look into the interplay between the two approaches is worthy of future study.

Acknowledgements

We thank Gauri Jagatap, Mohammedreza Soltani, and Anuj Sharma for helpful discussions.

References

[1] A. Athalye, N. Carlini, and D. A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. 2, 3, 8

[2] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *ICML*, 2018. 3, 13

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 15

[4] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[5] N. Carlini and D. A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 2017. 1, 3, 5, 7, 8, 12, 16

[6] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2016. 3, 4, 13, 14

[7] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *arxiv preprint*, abs/1712.05526, 2017. 14

[8] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 10, 14

[9] A. Dabouei, S. Soleymani, J. M. Dawson, and N. M. Nasrabadi. Fast geometrically-perturbed adversarial faces. *WACV*, 2019. 13

[10] S. Dathathri, S. Zheng, S. Gao, and R. Murray. Measuring the Robustness of Neural Networks via Minimal Adversarial Examples. In *NeurIPS-W*, volume 35, 2017. 1

[11] R. R. de Castro and H. A. Rabitz. Targeted nonlinear adversarial perturbations in images and videos. *arxiv preprint*, abs/1809.00958, 2018.

[12] L. Engstrom, D. Tsipras, L. Schmidt, and A. Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arxiv preprint*, abs/1712.02779, 2017. 5, 7, 8, 17, 18

[13] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. X. Song. Robust physical-world attacks on deep learning visual classification. *CVPR*, 2018. 13

[14] A. Fawzi, H. Fawzi, and O. Fawzi. Adversarial vulnerability for any classifier. In *NeurIPS*, 2018.

[15] A. Fawzi, O. Fawzi, and P. Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107, 2018. 12

[16] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1, 2, 3, 7, 8, 12

[17] I. J. Goodfellow. Defense against the dark arts: An overview of adversarial example security research and future research directions. *arxiv preprint*, abs/1806.04169, 2018. 1, 7

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3, 13

[19] T. Gu, B. Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arxiv preprint*, abs/1708.06733, 2017. 14

[20] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. *arxiv preprint*, 2017. 2, 4, 5, 6, 7, 10, 13, 14, 17

[21] G. Hinton, N. Srivastava, and K. Swersky. Lecture 6a, overview of mini-batch gradient descent. 16

[22] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. In *PMLR*, volume 80, 2018. 2

[23] A. Ilyas, L. Engstrom, and A. Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arxiv preprint*, abs/1807.07978, 2018. 2

[24] T. Kaneko, K. Hiramatsu, and K. Kashino. Generative attribute controller with conditional filtered generative adversarial networks. *CVPR*, 2017. 14

[25] T. Kim, B. Kim, M. Cha, and J. Kim. Unsupervised visual attribute transfer with reconfigurable generative adversarial networks. *arxiv preprint*, abs/1707.09798, 2017. 14

[26] D. Kingma and J. Ba. Adam: a method for stochastic optimization (2014). In *ICLR*, 2015. 7, 8, 16

[27] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arxiv preprint*, abs/1312.6114, 2014. 13

[28] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *JMLR*, volume 70, 2017. 14

[29] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arxiv preprint*, abs/1607.02533, 2017. 2, 3, 7, 8, 12

[30] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, et al. Fader networks: Manipulating images by sliding attributes. In *NeurIPS*, 2017. 2, 4, 5, 7, 10, 14

[31] A. B. L. Larsen, S. K. Sønderby, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016. 13

[32] Y. LeCun and C. Cortes. MNIST handwritten digit database, 2010. 2, 8

[33] M. Li, W. Zuo, and D. Zhang. Convolutional network for attribute-driven and identity-preserving human face generation. *arxiv preprint*, abs/1608.06434, 2016. 13

[34] M. Li, W. Zuo, and D. Zhang. Deep identity-aware transfer of facial attributes. *arxiv preprint*, abs/1610.05586, 2016. 13

[35] H.-T. D. Liu, M. Tao, C.-L. Li, D. Nowrouzezahrai, and A. Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. In *ICLR*, 2019. 3, 7

[36] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 3, 13

[37] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 2, 6, 16

[38] Y. Lu, Y.-W. Tai, and C.-K. Tang. Attribute-guided face generation using conditional cyclegan. In *ECCV*, 2018.

[39] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2, 7, 8

[40] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arxiv preprint*, abs/1411.1784, 2014. 4, 13

[41] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. *CVPR*, 2017. 1, 12

[42] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *CVPR*, 2016. 12

[43] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard. Robustness via curvature regularization, and vice versa. In *CVPR*, 2019. 12

[44] K. R. Mopuri, U. Ojha, U. Garg, and R. V. Babu. Nag: Network for adversary generation. *CVPR*, 2018. 13

[45] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 13, 14

[46] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. *EuroS&P*, 2016. 12

[47] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NeurIPS-W*, 2017. 6

[48] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. *arxiv preprint*, abs/1611.06355, 2016. 4, 13

[49] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arxiv preprint*, abs/1511.06434, 2016. 3, 13

[50] J. Rauber, W. Brendel, and M. Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

[51] P. Samangouei, M. Kabkab, and R. Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *ICLR*, 2018. 10

[52] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. In *NeurIPS*, 2018. 2, 9, 10, 14, 15, 16

[53] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NeurIPS*, 2018.

[54] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein. Are adversarial examples inevitable? In *ICLR*, 2019. 2

[55] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. *arxiv preprint*, abs/1801.00349, 2018. 3, 13

[56] W. Shen and R. Liu. Learning residual images for face attribute manipulation. *CVPR*, 2017. 13

[57] Y. Song, R. Shu, N. Kushman, and S. Ermon. Constructing unrestricted adversarial examples with generative models. In *NeurIPS*, 2018. 3, 7, 13

[58] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 3, 12

[59] O. Tange. Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47, Feb. 2011.

[60] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. D. McDaniel. Ensemble adversarial training: Attacks and defenses. *arxiv preprint*, abs/1705.07204, 2017. 12

[61] B. Tran, J. Li, and A. Madry. Spectral signatures in backdoor attacks. In *NeurIPS*, 2018. 14

[62] A. Turner, D. Tsipras, and A. Madry. Clean-label backdoor attacks, 2019. 14

[63] P. Upchurch, J. R. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala, and K. Q. Weinberger. Deep feature interpolation for image content changes. *CVPR*, 2017. 13

[64] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016. 3, 13

[65] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. X. Song. Spatially transformed adversarial examples. *arxiv preprint*, abs/1801.02612, 2018. 3, 23

[66] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. M. Eckert, and F. Roli. Support vector machines under adversarial label contamination. *Neurocomputing*, 160, 2015. 14

[67] H. Xiao, H. Xiao, and C. M. Eckert. Adversarial label flips attack on support vector machines. In *ECAI*, 2012. 14

[68] T. Xiao, J. Hong, and J. Ma. Dna-gan: Learning disentangled representations from multi-attribute images. *arxiv preprint*, abs/1711.05415, 2018. 14

[69] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018. 6, 23

[70] X. Zeng, C. Liu, Y.-S. Wang, W. Qiu, L. Xie, Y.-W. Tai, C.-K. Tang, and A. L. Yuille. Adversarial attacks beyond the image space. *arxiv preprint*, abs/1711.07183, 2017. 7, 13

[71] Y. Zhang, H. Foroosh, P. David, and B. Gong. Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *ICLR*, 2019. 3, 13

[72] Z. Zhao, D. Dua, and S. Singh. Generating natural adversarial examples. In *ICLR*, 2018. 3, 5, 13

[73] S. Zhou, T. Xiao, Y. Yang, D. Feng, Q. He, and W. He. Genegan: Learning object transfiguration and attribute subspace from unpaired data. *arxiv preprint*, abs/1705.04932, 2017. 13

[74] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017.

A Related Work

Adversarial Examples and Attacks.
In 2014, Szegedy *et al.* [58] shows that deep neural networks had mainly two counter intuitive properties, stating that the space described by higher layers of neural networks captures semantic information and there exists adversarial examples which questioned the generalization ability of a neural network. They generate such adversarial examples under the L_2 distance constraint which look similar to the original images but are classified with a different label by the classifier using a box constrained L-BFGS attack.

Goodfellow et. al [16] and Kurakin *et al.* [29] generate adversarial examples using Fast Gradient Sign method and its iterative variant under the l_∞ constraint in less computation time. Other methods similar to FGSM have been mentioned in [60].

Papernot *et al.* [46] implements an attack under the l_0 constraint where they modify the pixel having the most significant contribution in changing the classification of the model to the target class. Moosavi-Dezfooli *et al.* [42] describe an untargeted attack algorithm under the L_2 constraint with the assumption that neural networks are linear in nature which they further extend to non-linear neural networks. Another family of attacks relates to a single universal adversarial direction for a dataset. Moosavi-Dezfooli *et al.* [41] prove the existence of an image-agnostic adversarial perturbation. Fawzi *et al.* [15] extend this to theoretically show that every classifier is vulnerable to adversarial attacks. Moosavi-Dezfooli *et al.*further consider the effect of the curvature of the decision boundaries on the existence of adversarial examples in [43].

Carlini and Wagner [5] propose three attacks for adversarial image generation and shows that defensive distillation is not an effective defence mechanism. They devise attacks under the three norms in literature l_1, l_2 and l_∞ to measure the deviation of adversarial perturbation from the original sample over seven different surrogate loss functions and finally selecting one of them which we use in our attack algorithm as well. The attack that they implement in this work is proven to be the most effective attack in literature and is a benchmark for comparison.

The primary difference between the aforementioned attacks and our attack is that these attacks perturb the image and make imperceptible changes in the pixel space and thereby not modifying the image in a semantic way. On the other hand, our attack focuses making naturalistic perceptible changes to the image which are semantic in nature and realistic.

Parametric adversarial attacks.

The use of parametric transformations to generate adversarial examples has been tackled by several previous works. Most of these parametric attacks target the image formation process to create adversarial example. A recent work by Liu *et al.*perturbs geometrical surfaces or lighting by optimizing over the relevant parameters for a 3D environment. They show convincing results with realistic looking adversarial examples. Zeng *et al.* [70] use FGSM to perturb 3D models of objects to create adversarial examples. The primary caveat to such approaches is that they require precise 3D models of the objects that they create adversarial examples. Athalye *et al.* [2] demonstrate the creation of a real-world adversarial 3D model using optimization over affine transformations corresponding to real-world realizations. Eykhol *et al.* [13] also provide mechanisms for real-world realizable adversarial examples for stop signs using designed adversarial stickers.

Mopuri *et al.* [44] train a generative adversarial network to generate adversarial attacks for classifiers. Zhao *et al.* [72] show an interesting use of a GAN and an inverter network where they search over the input space of the GAN to generate semantically valid adversarial examples. These approaches are morally similar to our approach though we focus on specific physically perturbed attributes of images rather than imperceptible perturbations. CAMOU [71] is a more recent work that learns a neural approximator for physical camouflage and then optimizes over the same to generate an adversarial version to fool object detectors.

The space of generating adversarial examples using GANs for face recognition systems has also been touched upon by Dabouei *et al.* [9] and Sharif *et al.* [55] which train generative networks for the specific purpose of creating adversarial examples. Sharif *et al.*especially show a realizable attack by adding glasses using a generative network to fool a face recognition classifier. We, in comparison, provide a more diverse attack space allowing for various semantic attributes. In addition, since our attack involves physically realizable perceptible attributes, it can be used to characterize a classifier's performance against physical adversarial attacks as well.

Song *et al.* [57] uses an Auxiliary Class Generative Adversarial Network (AC-GAN) [45] to generate unrestricted adversarial examples from noise and then optimizes over the latent space of the conditional GAN to find such adversarial examples which get missclassified by a gender classifier. The paper describes the use of *Mechanical Turk* as a checker for naturalness and validation for the generated images belonging to the desired class. We approach the more complex problem of finding an adversarial transformation for an input image instead of generating a random semantic adversarial example.

Attribute based generative models.

Our approach relies on the use of attribute based generative models for enforcing the semantic constraint and representing attributes as a real-valued semantic variable. we discuss a few relevant approaches published recently.

As mentioned in [20], the literature related to facial attribute editing can be broadly divided into two sections, optimization based approaches and learning based approaches. Optimization approaches include Li *et al.* [33] and Gardner*et al.* [63] where the former optimizes the CNN feature difference between the input face image and the face images with the desired attributes with respect to the input face while the latter optimizes the input face in order to match the deep feature along the direction vector between the faces with and without the attributes.

Li *et al.* [34] describe a method to optimize over an adversarial attribute loss and a deep identity feature loss in order to train a deep identity aware transfer model to add or remove facial attributes to/from a face. Shen *et al.* [56] learn the difference between images before and after manipulation to simultaneously train two networks for respectively adding and removing a specific attribute.

Generative Adversarial Networks(GAN) [18] are a popular approach for the generation of samples from a real-world data distribution. Recent advancements [49, 36, 64, 6] in GANs allow for creation of high dimensional, high quality realistic images. These have been incorporated into the several attribute swapping generative models. Zhou *et al.* [73] recombine the information of the latent information of two images to swap a specific attribute between the given images. Liu *et al.* [36] generate high quality images by coupling GANs in order to learn a shared latent representation in order to tackle several unsupervised image translation tasks including domain adaptation and face image translation.

For multiple attribute swapping, models based on Kingma *et al.* [27], Goodfellow *et al.* [18],Larsen *et al.* [31], Mirza *et al.* [40], Radford *et al.* [49] have become quite popular recently. Perarnaue *et al.* [48] uses a Conditional Generative Adversarial Network [40] and encoder to learn the attribute invariant latent representation for attribute editing. Similar work has been seen in Fader Networks [30] where the model learns the attribute invariant latent space in order to identify a face as one and the same with or without a specific attribute. On the other hand, AttGAN [20] argues that such attribute invariant constraint is a bit too excessive and imposes an attribute classification constraint and a reconstruction loss instead to alter only the desired attributes preserving attribute-excluding features. StarGAN [8] uses a cyclic consistency loss to preserve information and instead of learning a latent representation, it trains a conditional attribute transfer network to modify attributes. Chen *et al.* [6] and Odena *et al.* [45] map the generated images back to the conditional signals with the help of an auxiliary classifier to learn this conditional generation of the images. Kaneko *et al.* [24] uses a conditional filtered generative adversarial network to present a generative attribute controller to edit attributes of an image while preserving the variations of an attribute.

Xiao *et al.* [68] swaps blocks of the latent distribution containing relevant attributes between a given pair of images. A similar approach has been seen in Kim*et al.* [25] where the latent representation is divided in blocks corresponding different attributes and these latent blocks are swapped in order to achieve multiple attribute swapping.

Data poisoning.
Much of the prior work mentioned discuss about adversarial attacks during inference. Data poisoning is a technique where the adversary injects false data to hinder the generalization capability of a deep neural network. Koh *et al.* [28] present the seminal work on data poisoning for deep neural networks where they construct approximate upper bounds to provide certificates to a large class of attacks. Xiao *et al.* [67] and Xiao *et al.* [66] also present a similar approach but on shallow learning models. Another class of data poisoning attack is referred to as a *backdoor attack*, where an adversary corrupts the model to misclassify either a specific input or a group of inputs to a target label thus engineering a *backdoor* that can be used to corrupt the learned model. Gu *et al.* [19] demonstrate a method to train a network maliciously with good performance on training and validation datasets but persistent poor performance on inputs associated with backdoor triggers.

These attacks can be realistic in nature, for e.g., a stop sign can be identified by the classifier as a speed limit sign in the presence of backdoor triggers which are mainly special markers added to the inputs by the adversary. Turner *et al.* [62] show that an adversary is able to gain whole control over the target model during inference, by training with samples generated with a GAN. More recently Tran *et al.* [61] identify a property related to all backdoor attacks known as spectral signatures with which poisoned examples from real image datasets can be detected and removed effectively. Chen *et al.* [7] demonstrate an application of such backdoor attacks on a visual recognition system where they were able to break a weak threat model with a limited number of poisoned data examples with semantic attribute changes. This is perhaps the first attempt at considering the effect of semantic changes.

B Theoretical Results

Robust classification error for subspace attacks

We present a proof for the upper bound of the robust classification error in the case of subspace attacks. Recall the data model we use; a Mixture of Gaussians data model, $\mathbb{P}_d(\theta^\star, \sigma) \sim \mathbb{R}^d \times \pm 1$ with two components and $\sigma \leq \sqrt{d}$. Each of the components are regarded as classes. We additionally assume a linear classifier, $f_{\boldsymbol{w}} : \mathbb{R}^d \rightarrow \{\pm 1\}$ defined by the unit vector, $\hat{\boldsymbol{w}} = sgn(\langle \hat{\boldsymbol{w}}, \mathbf{x} \rangle)$.

Let $\mathcal{S}_\epsilon = \{\tilde{\mathbf{x}} \mid \tilde{\mathbf{x}} = \mathbf{x} + \mathbf{U}\mathbf{U}^T\delta,\ ||\tilde{\mathbf{x}} - \mathbf{x}||_\infty \leq \epsilon\}$ and $rank\ \mathbf{U} = k$.

Under the assumption that the linear classifier is well trained, *i.e.*, $\hat{\boldsymbol{w}}$ is sufficiently correlated with the true component mean, $\theta^\star$, we upper bound the robust classification error. This involves considering the sample generalization error of a linear classifier on Gaussian data. We adapt arguments from Schmidt *et al.* [52] for the case of subspace attacks. The theorem statement is repeated here for convenience.

Theorem 1. *Let $\hat{\boldsymbol{w}}$ be such that $\langle \hat{\boldsymbol{w}}, \theta^\star \rangle \geq k||U||_{\infty,1}||\hat{\boldsymbol{w}}^T\mathbf{U}||_\infty\epsilon$. Then, the linear classifier $f_{\hat{\boldsymbol{w}}}$ has a $\mathcal{S}_\epsilon$-robust classification error upper bounded as:*

$$\beta \leq \exp\left(-\frac{\left(\langle \hat{\boldsymbol{w}}, \theta^\star \rangle - k||U||_{\infty,1}||\hat{\boldsymbol{w}}^T\mathbf{U}||_\infty\epsilon\right)^2}{2\sigma^2}\right) \tag{5}$$

Proof. For proving the above statement, we consider the probability of adversarial misclassification under a rank constrained attack.

Given $(\mathbf{x}, y) \in \mathbb{R}^d \times \pm 1$ where $\mathbf{x} \sim MoG(\theta^\star, \sigma; \sigma \leq \sqrt{d}$, we consider a linear additive attack under a rank constraint,

$$\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{U}\mathbf{U}^T\delta \tag{6}$$

Here, $\mathbf{U} \in \mathbb{M}_{d,k}$ is a random matrix with the columns forming an orthonormal basis of dimensionality k. In addition, we consider that the adversarial example thus created is constrained to be in the norm ball, $\mathcal{B}_{\infty}^{\epsilon}$, which implies that

$$||\tilde{\mathbf{x}} - \mathbf{x}||_{\infty} \leq \epsilon \tag{7}$$

We attempt to bound the probability that a rank constrained adversarial example, $\tilde{\mathbf{x}}$, created using equation 6, exists under the constraint defined by equation 7.

$$\begin{aligned}
\beta_{\infty}^{\epsilon} =& \mathbb{P}\left(\exists \tilde{\mathbf{x}}, \tilde{\mathbf{x}} \in \mathcal{B}_{\infty}^{\epsilon} s.t. \langle y\tilde{\mathbf{x}}, \hat{\boldsymbol{w}} \rangle \leq 0\right) \\
=& \mathbb{P}(\exists \delta \ : \ ||\mathbf{U}\mathbf{U}^T\delta||_{\infty} \leq \epsilon, \left\langle y(\mathbf{x} + \mathbf{U}\mathbf{U}^T\delta), \hat{\boldsymbol{w}} \right\rangle \leq 0) \\
=& \mathbb{P}(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_{||\mathbf{U}\mathbf{U}^T\delta||_{\infty} \leq \epsilon} \left\langle y\mathbf{U}\mathbf{U}^T\delta, \hat{\boldsymbol{w}} \right\rangle \leq 0)
\end{aligned}$$

Let $\delta' \triangleq \mathbf{U}^T\delta$.

Now,

$$\beta_{\infty}^{\epsilon} = \mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_{||\mathbf{U}\delta'|| \leq \epsilon} \langle y\mathbf{U}\delta', \hat{\boldsymbol{w}} \rangle \leq 0\right) \tag{8}$$

Consider the domain of the minimization,

$$||\mathbf{U}\delta'||_{\infty} \leq \epsilon$$

Now using the definition of the $(1, \infty)$ operator norm for rectangular matrices (See [3], Sec A.1.5) and the fact that $\mathbf{U}$ is orthonormal,

$$||\delta'||_{\infty} \leq ||\delta'||_1 = ||\mathbf{U}^T\mathbf{U}\delta'||_1 \leq ||\mathbf{U}^T||_{1,\infty}||\mathbf{U}\delta'||_{\infty} \leq ||\mathbf{U}||_{\infty,1}\epsilon$$

Let set $A \triangleq \{\delta' : ||\delta'||_{\infty} \leq ||\mathbf{U}||_{\infty,1}\epsilon\}$ and set $B \triangleq \{\delta' : ||\mathbf{U}\delta'||_{\infty} \leq \epsilon\}$. We can clearly see that $B \subseteq A$. Now considering the $f = \langle y\mathbf{U}\delta', \hat{w} \rangle$, as

$$\min_A f \leq \min_B f$$

Thus we show that,

$$\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_A f \leq \langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_B f \tag{9}$$

From the above inequality, $RHS \leq 0 \implies LHS \leq 0$ but not vice versa.

By using the inclusion argument of probability measure, we can therefore show that,

$$\mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_B f \leq 0\right) \leq \mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_A f \leq 0\right) \tag{10}$$

We now upper bound the RHS term using the same argument as that of Lemma 20 in [52].

$$\begin{aligned}
&\mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_{||\delta'||_{\infty} \leq ||\mathbf{U}||_{\infty,1}\epsilon} \langle y\mathbf{U}\delta', \hat{\boldsymbol{w}} \rangle \leq 0\right) \\
=&\mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_{||\delta'||_{\infty} \leq ||\mathbf{U}||_{\infty,1}\epsilon} y\hat{\boldsymbol{w}}^T\mathbf{U}\delta' \leq 0\right) \\
&\text{Let } \bar{\boldsymbol{w}} \triangleq \hat{\boldsymbol{w}}^T\mathbf{U} \\
=&\mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_{||\delta'||_{\infty} \leq ||\mathbf{U}||_{\infty,1}\epsilon} y\bar{\boldsymbol{w}}\delta' \leq 0\right)
\end{aligned}$$

We now drop y as the constraint is symmetric and use definition of dual norm,

$$\begin{aligned}
&\mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle - ||\mathbf{U}||_{\infty,1}||\bar{\boldsymbol{w}}||_{\infty}^{\star}\epsilon \leq 0\right) \\
=&\mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle \leq ||\mathbf{U}||_{\infty,1}||\bar{\boldsymbol{w}}||_{\infty}^{\star}\epsilon\right) \\
=&\mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle \leq k||\mathbf{U}||_{\infty,1}||\bar{\boldsymbol{w}}||_{\infty}\epsilon\right)
\end{aligned}$$

We now invoke Lemma 17 from [52] with $\mu = \theta^\star$ and $\rho = ||\mathbf{U}||_{\infty,1}\epsilon||\bar{\boldsymbol{w}}||_\infty$ to bound the RHS,

$$\beta \leq \exp\left(-\frac{\left(\langle\hat{\boldsymbol{w}}, \theta^\star\rangle - k||\mathbf{U}||_{\infty,1}||\hat{\boldsymbol{w}}^T\mathbf{U}||_\infty\epsilon\right)^2}{2\sigma^2}\right) \tag{11}$$

□

C Details of Experiments

Dataset: For our experiments, we use the CelebA dataset [37]. The dataset has approximately 200k images of faces. Each image is annotated with 65 binary attributes. Examples of these attributes are gender, age and skin complexion. We preprocess the images by cropping the central 178 × 178 sub-image and resizing each crop to 256 × 256. The resized images are then normalized to be between -1 and 1.

Target Binary Classifier: We attack a pre-trained gender binary classifier using our approach. The architecture used for the classifier is shown in Table 3. We train the classifier with 70% of the CelebA dataset [37] as training data and 20% as validation data using categorical cross-entropy. We use ADAM [26] as our optimizer. Our model is 95.6% accurate on the test set (10% of the dataset). We additionally train a binary *age* classifier with the same architecture.

Layers	Size
Convolutional Layer with Relu	32x3x3
Maxpooling Layer	2x2
Convolutional Layer with Relu	64x3x3
Maxpooling Layer	2x2
Convolutional Layer with Relu	128x3x3
Maxpooling Layer	2x2
Fully Connected Layer	1024
Fully Connected Layer	2

Table 3: *Architecture for the binary classifier.*

Adversarial Fader Networks

Architecture of Fader Networks. Fader Networks are an encoder-decoder architecture that disentangles semantic attributes during the reconstruction process. This is achieved by training a discriminator on the encoded latent vector while simultaneously reconstructing the original image from the concatenated latent vector and the semantic attribute vector. Figure 6 shows the architecture of the Fader Networks. An intriguing effect of the training process is that the attribute vector space can be treated as a continuous and bounded space. We further can optimize over this space to generate adversarial examples.

Single and Multi Attribute Attacks. We train three multi-attribute Fader networks with attributes presented in table 4. The pre-trained Fader networks are used as semantic constraints with the attribute vectors as the optimization variables. We then process examples from the CelebA test set with the semantic attack algorithm to generate adversarial examples. In order to make our optimization algorithm compatible with Fader Networks, we create a non-parametric forward model to convert the attribute vector to a compatible form. We call this forward model "Attribute Encoding".

We generate semantic adversarial images by optimizing over a modified Carlini-Wagner loss [5] with respect to the attribute vectors using ADAM [26] with a learning rate of 0.01. We also experimented with various other optimizers including stochastic gradient descent, RMSProp [21], but find that ADAM generates sharper images as well is the most successful.

Our experiments show that successful multi-attribute models tend to be deeper and wider. In addition, these networks are extremely susceptible to mode collapse unless the hyperparameters are carefully tuned. We hypothesize that this is an effect of the strong coupling of facial attributes, thus making the generator-discriminator optimization difficult. An unconditioned generative neural network generally learns to associate these entangled representations to a latent vector space where dimension represents some combination of attributes. In order to get past this, we model the

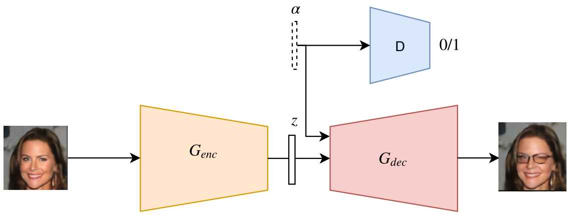

Figure 6: *Architecture of Fader Networks. The encoder converts the input image to a latent vector. The decoder takes as input the latent and the attribute vectors to generate the transformed image. Here, the discriminative classifier acts as an adversarial network to decouple the underlying invariant data from the semantic attributes.*

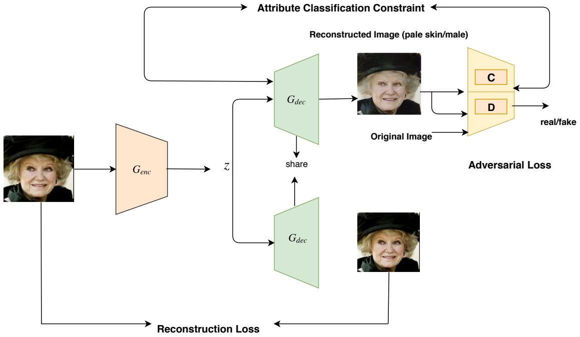

Figure 7: *Architecture of AttGAN Networks. As compared to Fader Networks, the discriminator/classifier pair is used to analyse the reconstruction of the image with the original semantic attributes. Enforcing the decoder to construct both the original and the semantically transformed image results in the decoupling of the semantic and the invariant data.*

Attack Type	Attributes	Accuracy of target model (%)	Random Sampling (%)
Single Attribute Attack	A1	70.0	87.0
	A2	61.0	93.0
	A3	48.0	88.0
Multi Attribute Attack	A1,A5,A6	12.0	86.0
	A2,A5,A6	7.00	85.0
	A1,A2,A7	28.0	84.0
Cascaded Multi Attribute Attack	A1-A2-A3	30.0	68.0
	A1-A3-A4	31.0	80.0
	A2-A3-A4	42.0	68.0

Table 4: *Performance of the Semantic Adversarial Example under various Adversarial Fader Networks implementations for the binary age classifier. Legend for attributes: A1-Eyeglasses, A2-Gender, A3-Nose shape, A4-Eye shape, A5-Chubbiness, A6-Pale Skin, A7-Smiling. Observe that as the number of perturbed attributes increase, the semantic attacks become more effective. In comparison to worst-of-10 random sampling [12] of the attribute space, our optimization framework is more effective at finding semantic adversarial examples. Note that the performance of our semantic attacks fare very well to decrease the accuracy of the age classifier as well like the gender classifier.*

multi-attribute perturbation problem as a sequential perturbation of single attributes.

Cascaded Attribute Attack. For the cascaded attribute attack, we cascade several smaller single attribute models one after the other to sequentially transform the input image. In this case, the problem of decoupling facial features from the underlying invariant data is divided among multiple models. The transformed image is then input to the target model. We generate adversarial examples as in the previous two cases for the CelebA test set by optimizing the Carlini-Wagner loss. In this case, we also modify the attribute encoding module to treat each attribute tuple separately.

We find that the semantically transformed images tend to be less sharp as compared to the ones generated single or multi-attribute attacks. This can be attributed to the concatenation of several reconstruction steps. Sequential reconstruction leads to loss of information and the reconstruction error compounding.

Attribute Encoding. Each attribute is represented by a tuple of real numbers that sum up to one. These tuples are concatenated into an attribute vector. To ensure that this structure is preserved over the optimization framework, we use a non-parametric forward model to algebraically manipulate our optimization variables to this specific representation. The encoding module also implements the box constraint for the optimized attribute values to lie between -3.0 and 3.0 in order to ensure that the generated images are valid.

Adversarial Attribute GANs

Architecture of Attribute GANs. Attribute GANs [20] improve upon Fader Networks by using a discriminator-classifier pair to analyse the reconstructed images (Refer Figure 7 for the architecture). They optimize over a combination of a reconstruction loss, an adversarial loss and an attribute constraint loss to ensure the editing of the exact desired attribute while preserving the attribute excluding details at the same time. The encoded latent vector is conditioned on the attribute vector during the decoding process. This results in the decoupling of semantic attributes from the underlying identity data. AttGAN takes as input an image and an attribute vector where each element represents an attribute. We select k attributes to perturb for our semantic attack.

We use a pretrained AttGAN model with 13 semantic attributes. For our experiments we consider 5 and 6 attributes respectively for transforming input images.

Attacks. We adapt our adversarial Fader Network approach to the AttGANs by modifying the "Attribute Encoding" module to mask attributes that we do not perturb. The encoding module also constrains the elements to lie between -1.0 and 1.0 as required by our algorithm to generate valid images.

D Results

Our additional experiments on the binary age classifier show that our approach is able to generate adversarial examples for other classifiers trained on the CelebA dataset (See Table 4). Note that our observations regarding the increasing effectiveness of our attack approach as the number of attributes we perturb increase, holds even for a new classifier. We also compare the performance of our attack with worst-of-10 random sampling (similar to the approach in [12].) This proves that our approach is successful at generating semantic adversarial examples.

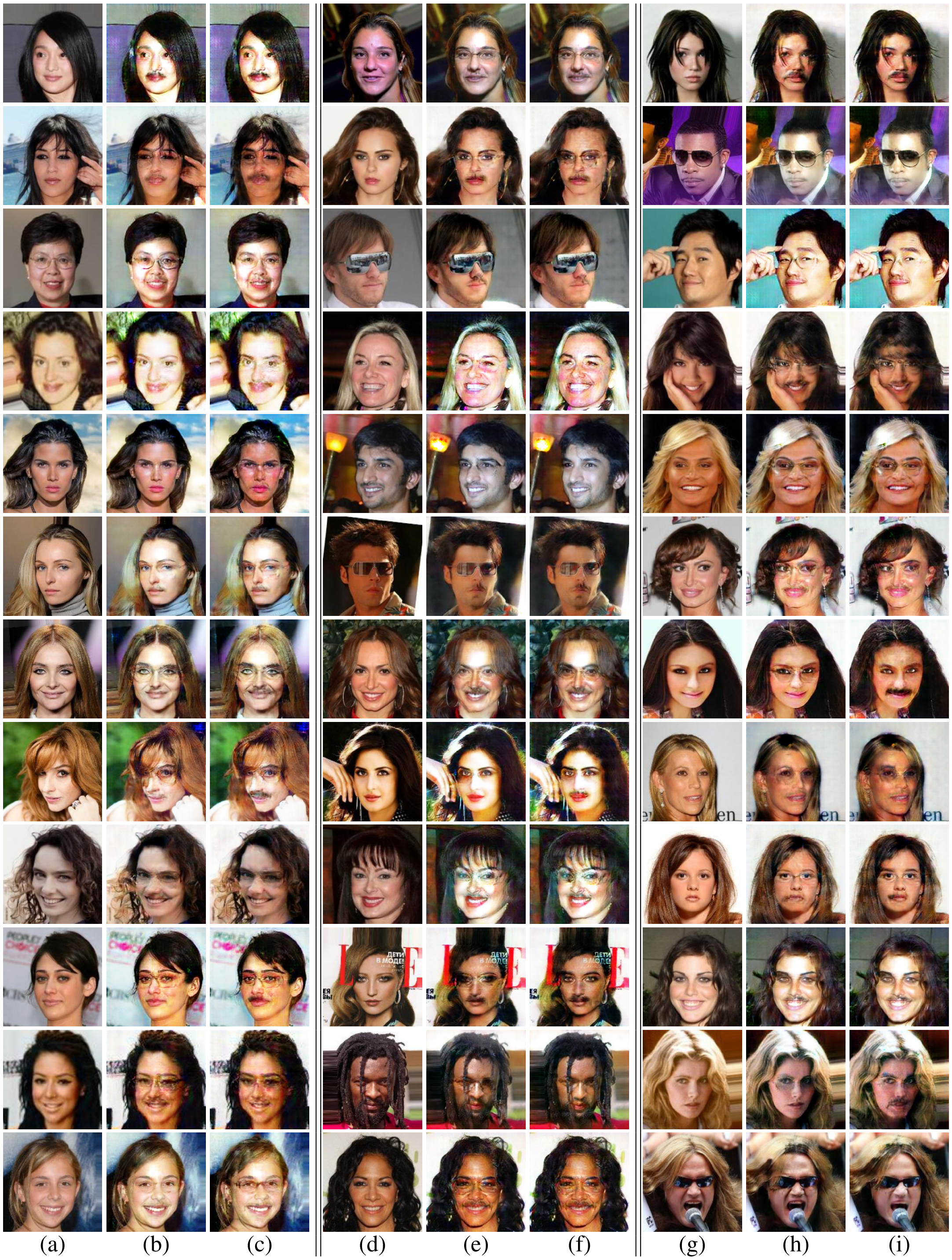

Figure 8: *Semantic adversarial examples generated with multiple attribute implementation using Adversarial AttGAN. The first, fourth and seventh columns contain the original images. We show adversarial examples generated under the attributes: (b),(e) and (h) Eyeglasses-Mustache-Age-Pale Skin-Young-Black Hair and (c),(f) and (i)Eyeglasses-Mustache-Pale skin-Age-Bushy eyebrows-Black hair. The quality of the images produced by the Adversarial AttGAN are sharper than those produced by the Adversarial Fader Networks.*

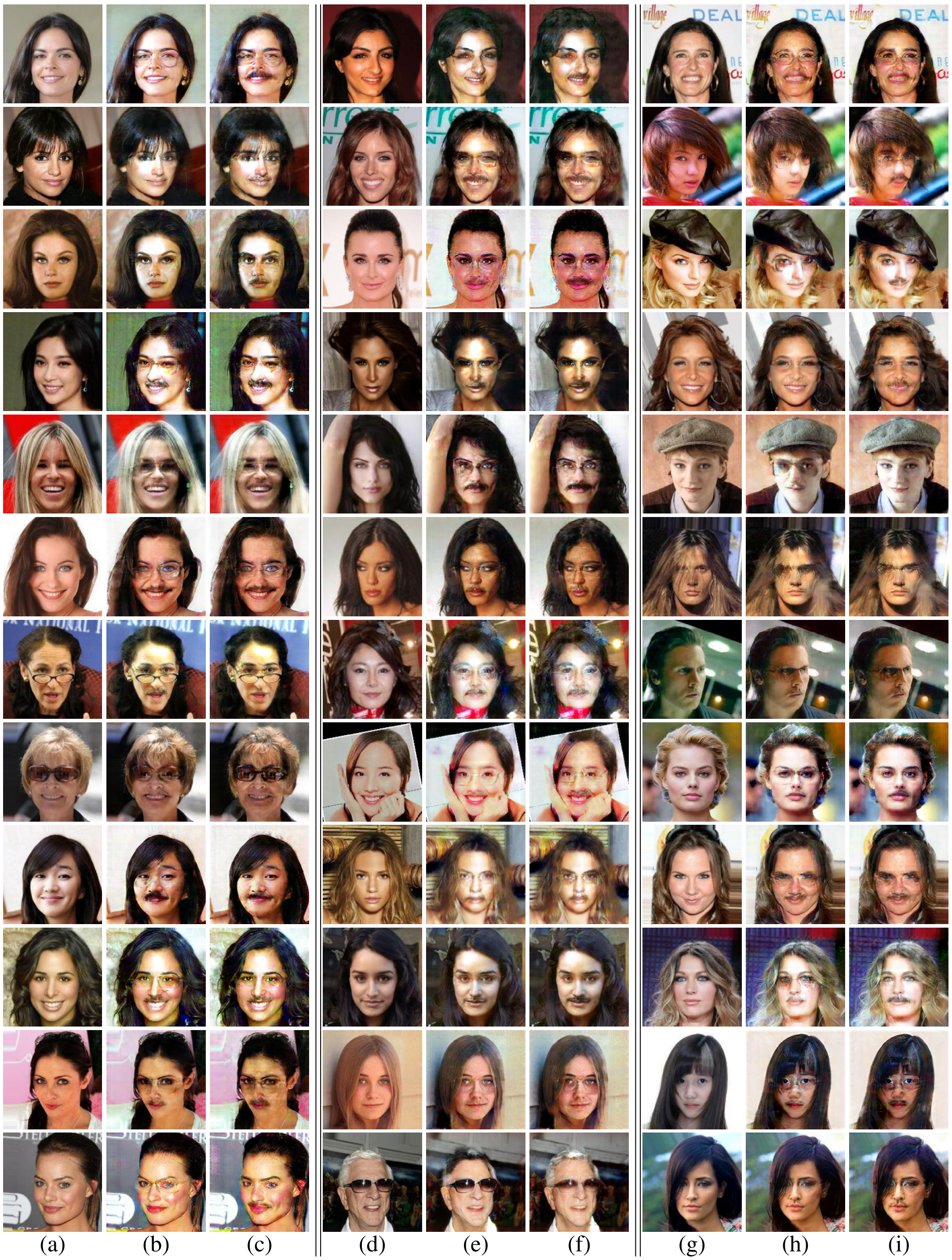

Figure 9: *Semantic adversarial examples generated with multiple attribute implementation using Adversarial AttGAN. The first, fourth and seventh columns contain the original images. We show adversarial examples generated under the attributes: (b),(e) and (h) Eyeglasses-Mustache-Age-Pale Skin-Young-Black Hair and (c),(f) and (i)Eyeglasses-Mustache-Pale skin-Age-Bushy eyebrows-Black hair. The quality of the images produced by the Adversarial AttGAN are sharper than those produced by the Adversarial Fader Networks.*

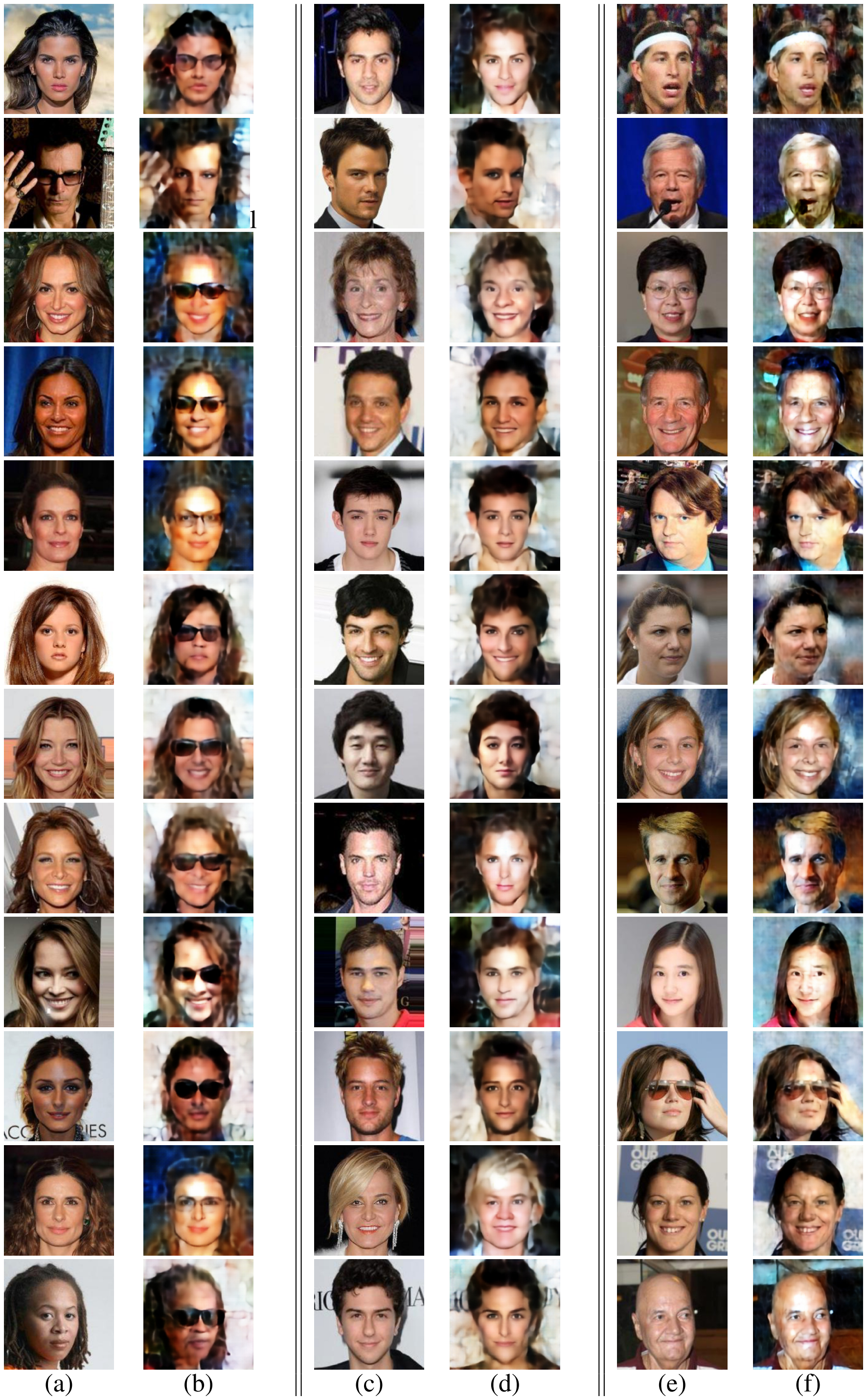

Figure 10: *Semantic adversarial examples generated with Single attribute implementation using Adversarial Fader Networks. Columns (a),(c) and (e) contain the original images. We show adversarial examples generated under the attributes: (b),(d) and (f) Eyeglasses, Nose shape and Age respectively.*

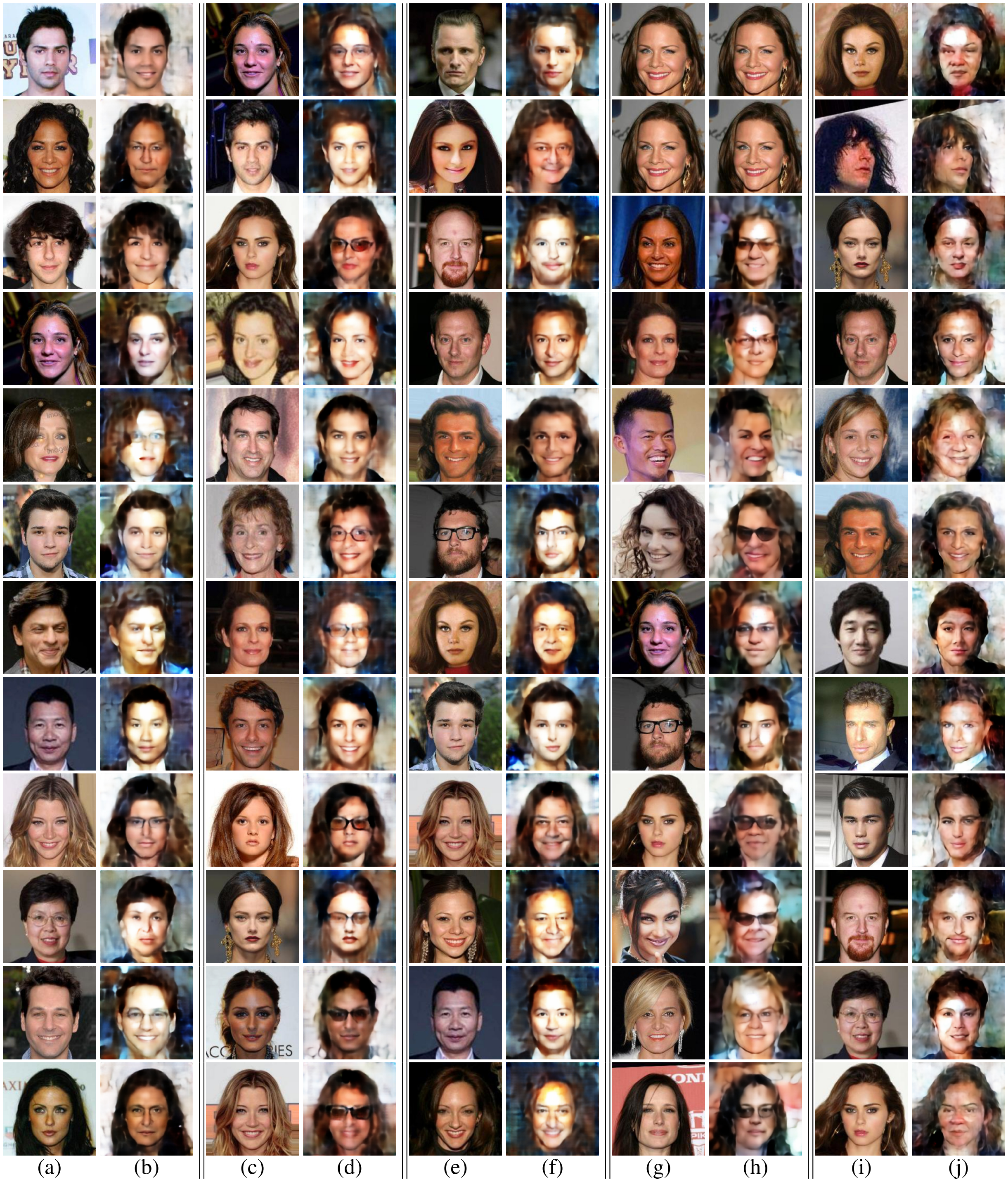

Figure 11: *Semantic adversarial examples for Multi-attribute and Cascaded Adversarial Fader network attacks. Columns (a), (c), (e), (g), (i) are original images. Columns (b): Multi-attribute Eyeglasses,Age,Smile, (d): Multi-attribute Pale Skin,Eyeglasses,Chubbiness, (f): Multi-attribute Age,Chubbiness,Pale Skin, (h): Cascaded Eyeglasses-Age-Nose shape, (j): Cascaded Nose shape-Narrow Eyes-Age*

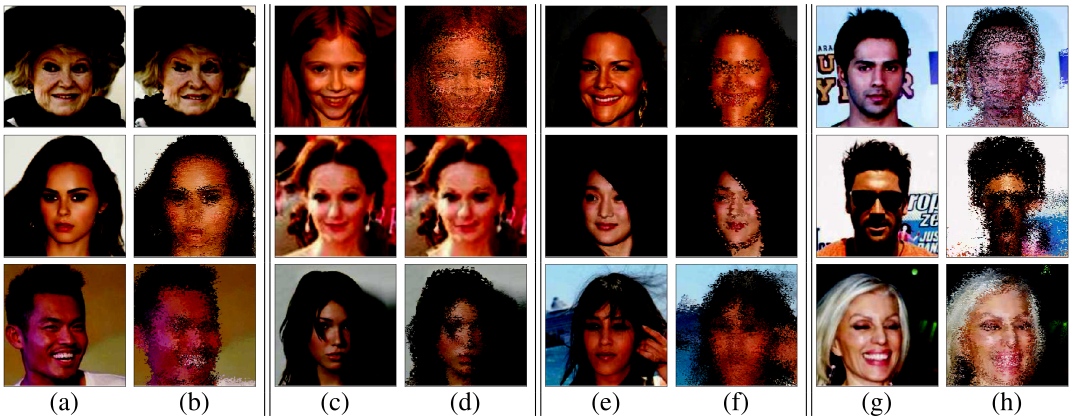

Figure 12: *Columns (a), (c), (e), (g) show original images and columns (b), (d), (f) and (h) show the corresponding adversarial images produced by implementing the algorithm from Xiao* et al.*[65]. As can be clearly seen although the adversarial examples are missclassified by the deep gender classifier, the produced adversarial images are not semantically valid.*

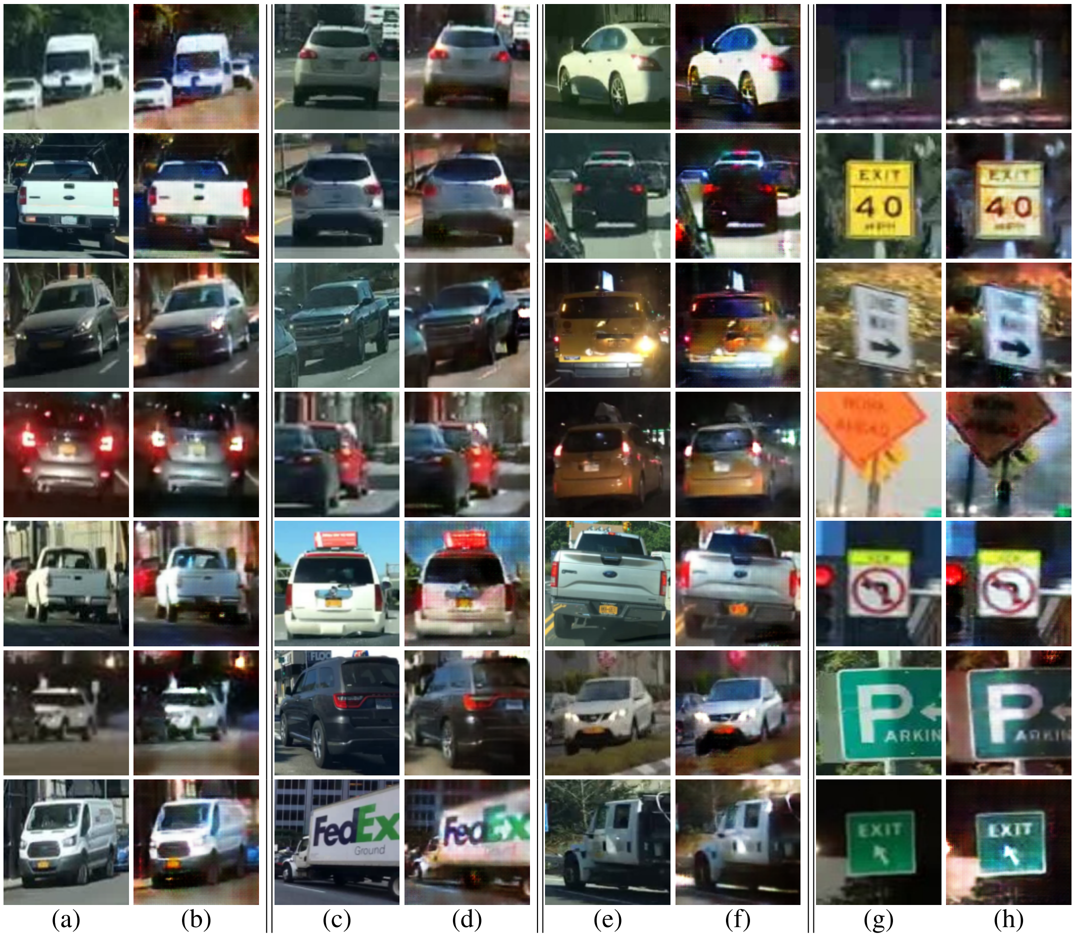

Figure 13: *Semantic adversarial examples produced by Attribute GAN trained on Time of Day labels from BDD dataset[69]. Columns (a), (c), (e), (g) are original images. Rows (1) through (6) and columns (a) to (f) show adversarial examples of cars getting miss-classified as traffic signs or trucks. Column (h) shows adversarial examples of traffic signs being miss-classified as cars. Row 7, columns (b) and (f) shows examples as trucks getting miss-classified as cars. Row 7, column d shows an adversarial example of a truck getting miss-classified as a traffic sign.*