# Semantic Adversarial Attacks: Parametric Transformations That Fool Deep Classifiers Supplementary

## A. Related Work

**Adversarial Examples and Attacks.**
In 2014, Szegedy *et al*. [41] 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 [11] and Kurakin *et al*. [21] generate adversarial examples using Fast Gradient Sign method and its iterative variant under the $l_\infty$ constraint in less computation time. Other methods similar to FGSM have been mentioned in [42].

Papernot *et al*. [34] 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*. [30] 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*. [29] prove the existence of an image-agnostic adversarial perturbation. Fawzi *et al*. [10] 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 [31].

Carlini and Wagner [3] 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 $l_\infty$ 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*. [52] 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*. [1] demonstrate the creation of a real-world adversarial 3D model using optimization over affine transformations corresponding to real-world realizations. Eykhol *et al*. [9] also provide mechanisms for real-world realizable adversarial examples for stop signs using designed adversarial stickers.

Mopuri *et al*. [32] train a generative adversarial network to generate adversarial attacks for classifiers. Zhao *et al*. [54] 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 [53] 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*. [7] and Sharif *et al*. [38] which train generative networks for the specific purpose of creating ad-

versarial 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*. [40] uses an Auxiliary Class Generative Adversarial Network (AC-GAN) [33] 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 [14], 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*. [24] and Gardner*et al*. [45] 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*. [25] 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*. [39] 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) [12] are a popular approach for the generation of samples from a real-world data distribution. Recent advancements [36, 26, 46, 4] 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*. [55] recombine the information of the latent information of two images to swap a specific attribute between the given images. Liu *et al*. [26] 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*. [19], Goodfellow *et al*. [12],Larsen *et al*. [23], Mirza *et al*. [28], Radford *et al*. [36] have become quite popular recently. Perarnaue *et al*. [35] uses a Conditional Generative Adversarial Network [28] and encoder to learn the attribute invariant latent representation for attribute editing. Similar work has been seen in Fader Networks [22] 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 [14] 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 [6] 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*. [4] and Odena *et al*. [33] 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*. [16] 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*. [50] 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*. [17] 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*. [20] 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*. [49] and Xiao *et al*. [48] 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*. [13] 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 spe-

cial markers added to the inputs by the adversary. Turner *et al*. [44] 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*. [43] 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*. [5] 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, $f_{\boldsymbol{w}} : \mathbb{R}^d \to \{\pm 1\}$ defined by the unit vector, $\hat{\boldsymbol{w}} = sgn(\langle \hat{\boldsymbol{w}}, \mathbf{x} \rangle)$.

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

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*. [37] for the case of subspace attacks. The theorem statement is repeated here for convenience.

**Theorem 1.** *Let $\hat{\boldsymbol{w}}$ be such that $\langle \hat{\boldsymbol{w}}, \theta^\star \rangle \geq \sqrt{k}||U||_{\infty,1}||\hat{\boldsymbol{w}}^T\mathbf{U}||_2\epsilon$. Then, the linear classifier $f_{\hat{\boldsymbol{w}}}$ has a $\mathcal{S}_\epsilon$-robust classification error upper bounded as:*

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

*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,

$$\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{U}\mathbf{U}^T\delta \tag{2}$$

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, $\mathcal{B}^\epsilon_\infty$, which implies that

$$||\tilde{\mathbf{x}} - \mathbf{x}||_\infty \leq \epsilon \tag{3}$$

We attempt to bound the probability that a rank constrained adversarial example, $\tilde{\mathbf{x}}$, created using equation 2, exists under the constraint defined by equation 3.

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

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

$$\beta^\epsilon_\infty = \mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_{||\mathbf{U}\delta'|| \leq \epsilon} \langle y\mathbf{U}\delta', \hat{\boldsymbol{w}} \rangle \leq 0\right) \tag{4}$$

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 [2], 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,

$$\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_A f \leq \langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_B f \tag{5}$$

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,

$$\mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_B f \leq 0\right) \leq \mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_A f \leq 0\right) \tag{6}$$

We now upper bound the $RHS$ term using the same argument as that of Lemma 20 in [37].

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

Let $\bar{\boldsymbol{w}} \triangleq \hat{\boldsymbol{w}}^T\mathbf{U}$

$$= \mathbb{P}\left(\langle y\mathbf{x}, \hat{\boldsymbol{w}} \rangle + \min_{||\delta'||_\infty \leq ||\mathbf{U}||_{\infty,1}\epsilon} y\bar{\boldsymbol{w}}\delta' \leq 0\right)$$

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

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

We now invoke Lemma 17 from [37] with $\mu = \theta^{\star}$ and $\rho = ||\mathbf{U}||_{\infty,1}\epsilon||\bar{\boldsymbol{w}}||_{\infty}$ to bound the $RHS$,

$$\beta \leq \exp\left(-\frac{\left(\langle \hat{\boldsymbol{w}}, \theta^{\star}\rangle - k||\mathbf{U}||_{\infty,1}||\hat{\boldsymbol{w}}^{T}\mathbf{U}||_{\infty}\epsilon\right)^{2}}{2\sigma^{2}}\right) \quad (7)$$

□

## C. Details of Experiments

**Dataset:** For our experiments, we use the CelebA dataset [27]. 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 $178 \times 178$ sub-image and resizing each crop to $256 \times 256$. 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 1. We train the classifier with 70% of the CelebA dataset [27] as training data and 20% as validation data using categorical cross-entropy. We use ADAM [18] 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 1. *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 1 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.

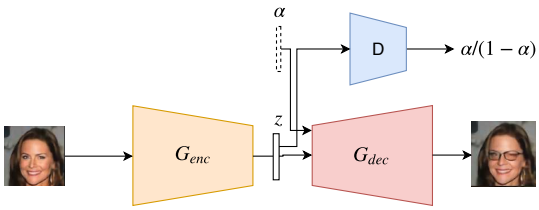



Figure 1. *Architecture of Fader Networks. The encoder converts the input image to a latent vector. The decoder takes as input the latent and the attribute vectors to generate the transformed image. Here, the discriminative classifier acts as an adversarial network to decouple the underlying invariant data from the semantic attributes.*

**Single and Multi Attribute Attacks.** We train three multi-attribute Fader networks with attributes presented in table 2. 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 [3] with respect to the attribute vectors using ADAM [18] with a learning rate of 0.01. We also experimented with various other optimizers including stochastic gradient descent, RMSProp [15], 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 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 (Refer Figure **??** for a block diagram). 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 [14] improve upon Fader Networks by using a discriminator-classifier pair to analyse the reconstructed images (Refer Figure 2 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.

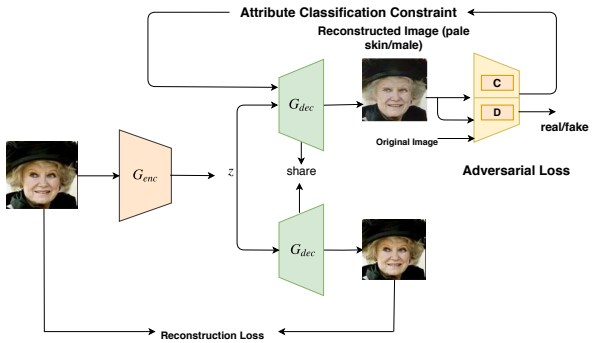



Figure 2. *Architecture of AttGAN Networks. As compared to Fader Networks, the discriminator/classifier pair is used to analyse the reconstruction of the image with the original semantic attributes. Enforcing the decoder to construct both the original and the semantically transformed image results in the decoupling of the semantic and the invariant data.*

## D. Results

| 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 2. *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 [8] 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.*

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 2). 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 [8].) This proves that our approach is successful at generating semantic adversarial examples.

**Qualitative Results for Attacks on Binary Gender Classifier**


Figure 3. *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.*

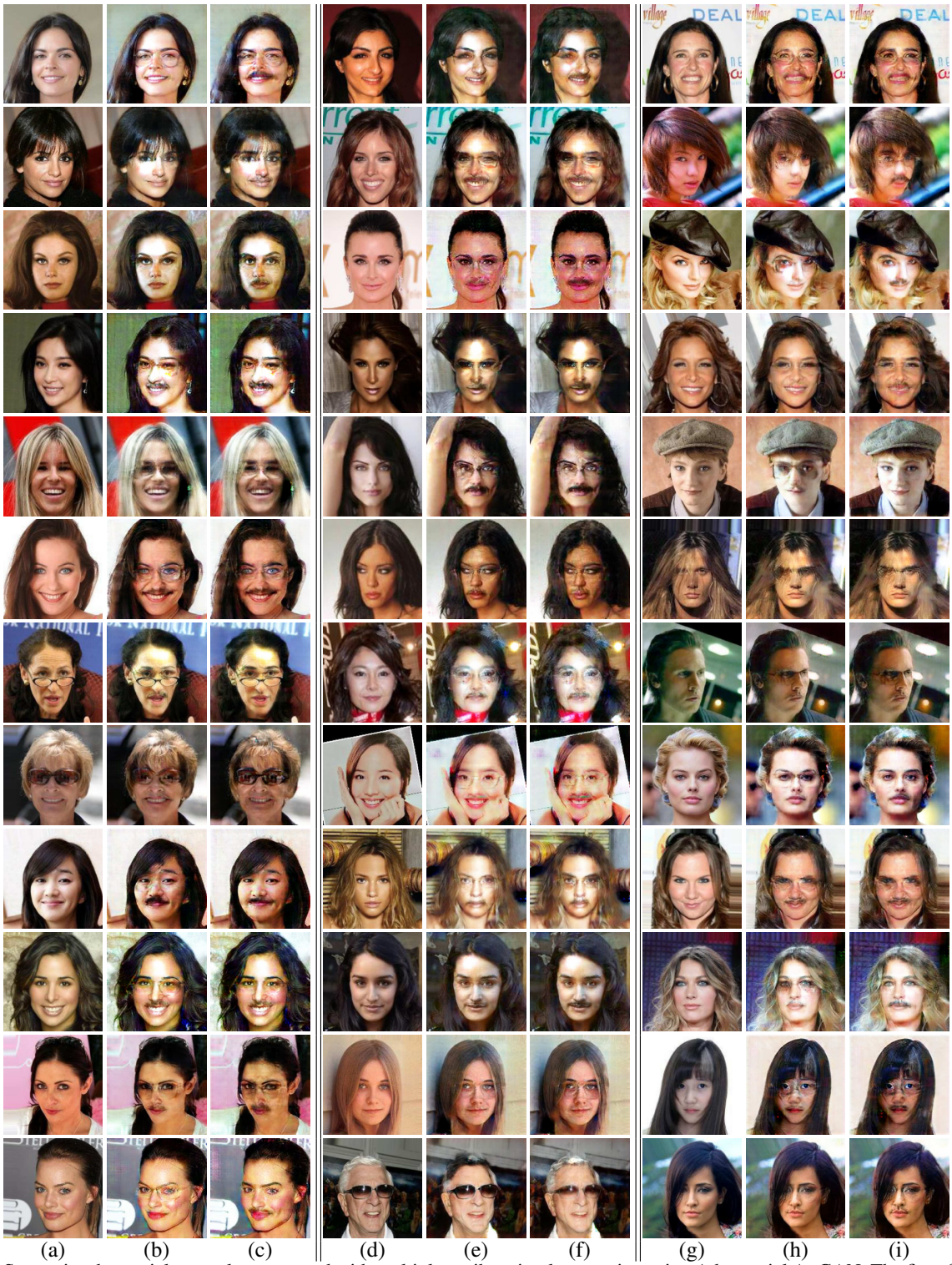



Figure 4. *Semantic adversarial examples generated with multiple attribute implementation using Adversarial AttGAN. The first, fourth and seventh columns contain the original images. We show adversarial examples generated under the attributes: (b),(e) and (h) Eyeglasses-Mustache-Age-Pale Skin-Young-Black 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.*

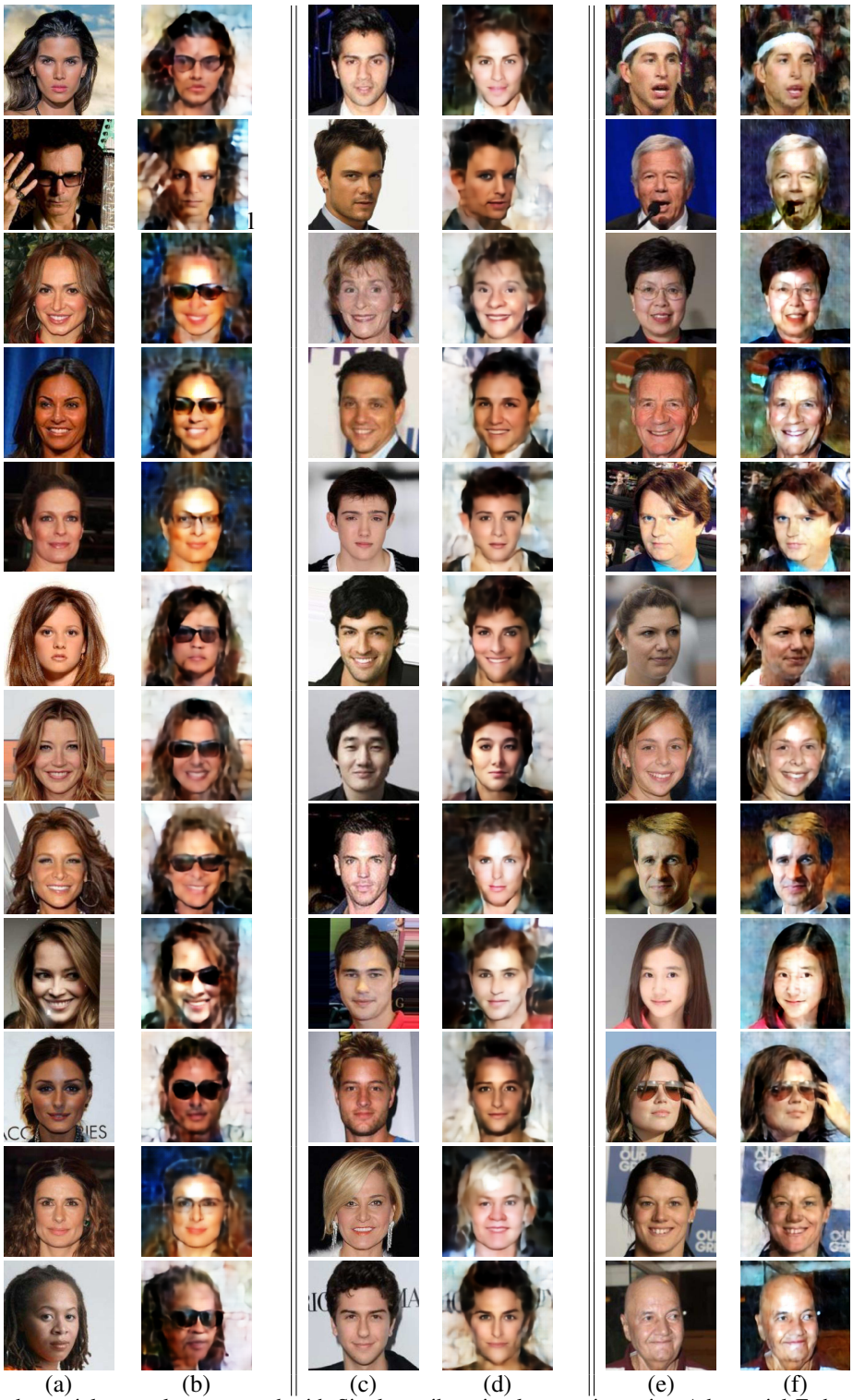



Figure 5. *Semantic adversarial examples generated with Single attribute implementation using Adversarial Fader Networks. Columns (a),(c) and (e) contain the original images. We show adversarial examples generated under the attributes: (b),(d) and (f) Eyeglasses, Nose shape and Age respectively.*

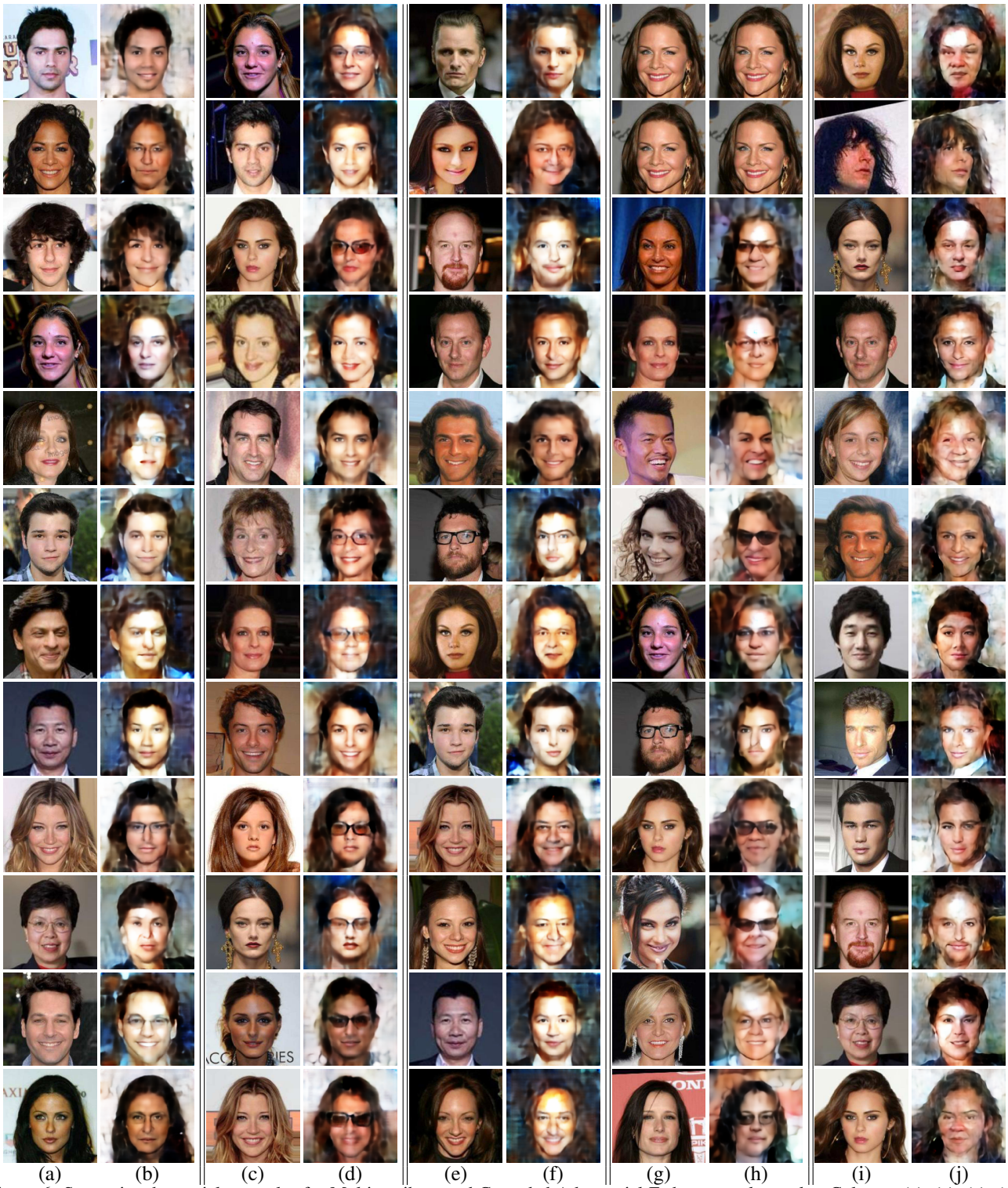
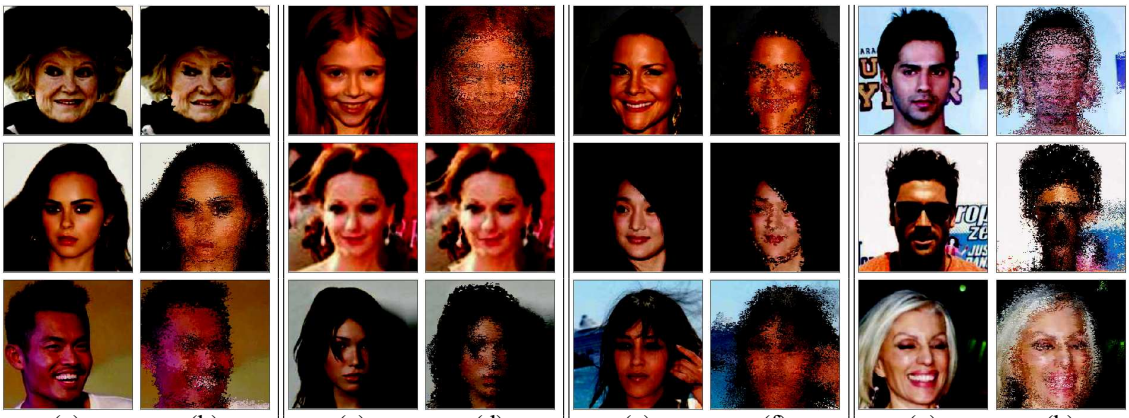
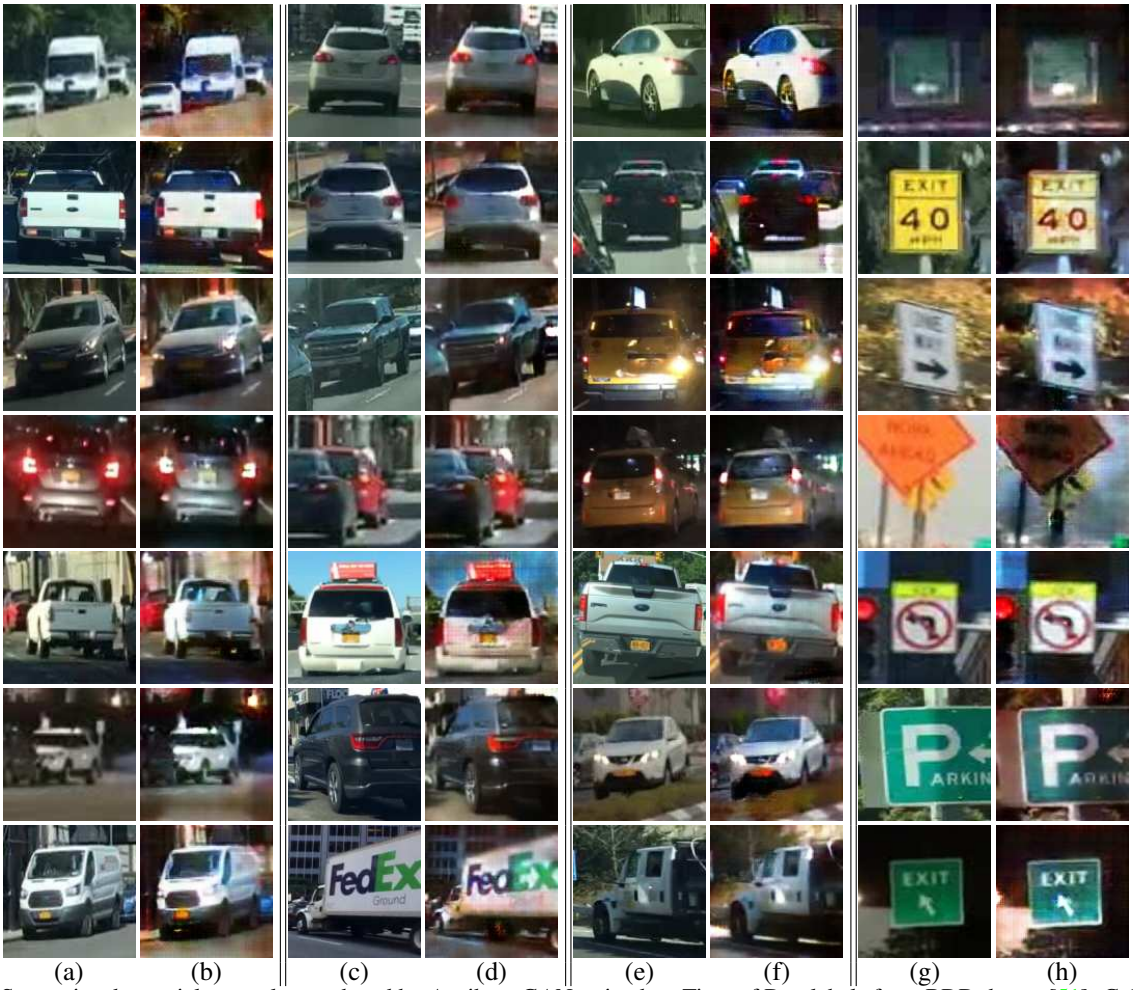



Figure 6. *Semantic adversarial examples for Multi-attribute and Cascaded Adversarial Fader network attacks. Columns (a), (c), (e), (g), (i) are original images. Columns (b): Multi-attribute Eyeglasses,Age,Smile, (d): Multi-attribute Pale Skin,Eyeglasses,Chubbiness, (f): Multi-attribute Age,Chubbiness,Pale Skin, (h): Cascaded Eyeglasses-Age-Nose shape, (j): Cascaded Nose shape-Narrow Eyes-Age*



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 7. *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. *[47]. As can be clearly seen although the adversarial examples are missclassified by the deep gender classifier, the produced adversarial images are not semantically valid.*




(a) (b) (c) (d) (e) (f) (g) (h)

Figure 8. *Semantic adversarial examples produced by Attribute GAN trained on Time of Day labels from BDD dataset[51]. 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.*

## References


[1] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *ICML*, 2018. 1

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

[3] N. Carlini and D. A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 2017. 1, 4

[4] 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. 2

[5] 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. 3

[6] 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. 2

[7] A. Dabouei, S. Soleymani, J. M. Dawson, and N. M. Nasrabadi. Fast geometrically-perturbed adversarial faces. *WACV*, 2019. 1

[8] 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

[9] 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. 1

[10] A. Fawzi, O. Fawzi, and P. Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107, 2018. 1

[11] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1

[12] 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. 2

[13] T. Gu, B. Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arxiv preprint*, abs/1708.06733, 2017. 2

[14] 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, 5

[15] G. Hinton, N. Srivastava, and K. Swersky. Lecture 6a, overview of mini-batch gradient descent. 4

[16] T. Kaneko, K. Hiramatsu, and K. Kashino. Generative attribute controller with conditional filtered generative adversarial networks. *CVPR*, 2017. 2

[17] T. Kim, B. Kim, M. Cha, and J. Kim. Unsupervised visual attribute transfer with reconfigurable generative adversarial networks. *arxiv preprint*, abs/1707.09798, 2017. 2

[18] D. Kingma and J. Ba. Adam: a method for stochastic optimization (2014). In *ICLR*, 2015. 4

[19] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arxiv preprint*, abs/1312.6114, 2014. 2

[20] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *JMLR*, volume 70, 2017. 2

[21] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arxiv preprint*, abs/1607.02533, 2017. 1

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

[23] A. B. L. Larsen, S. K. Sønderby, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016. 2

[24] M. Li, W. Zuo, and D. Zhang. Convolutional network for attribute-driven and identity-preserving human face generation. *arxiv preprint*, abs/1608.06434, 2016. 2

[25] M. Li, W. Zuo, and D. Zhang. Deep identity-aware transfer of facial attributes. *arxiv preprint*, abs/1610.05586, 2016. 2

[26] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 2

[27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 4

[28] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arxiv preprint*, abs/1411.1784, 2014. 2

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

[30] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *CVPR*, 2016. 1

[31] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard. Robustness via curvature regularization, and vice versa. In *CVPR*, 2019. 1

[32] K. R. Mopuri, U. Ojha, U. Garg, and R. V. Babu. Nag: Network for adversary generation. *CVPR*, 2018. 1

[33] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017. 2

[34] 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. 1

[35] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. *arxiv preprint*, abs/1611.06355, 2016. 2

[36] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arxiv preprint*, abs/1511.06434, 2016. 2

[37] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. In *NeurIPS*, 2018. 3, 4

[38] 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. 1

[39] W. Shen and R. Liu. Learning residual images for face attribute manipulation. *CVPR*, 2017. 2

[40] Y. Song, R. Shu, N. Kushman, and S. Ermon. Constructing unrestricted adversarial examples with generative models. In *NeurIPS*, 2018. 2

[41] 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

[42] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. D. McDaniel. Ensemble adversarial training: Attacks and defenses. *arxiv preprint*, abs/1705.07204, 2017. 1

[43] B. Tran, J. Li, and A. Madry. Spectral signatures in backdoor attacks. In *NeurIPS*, 2018. 3

[44] A. Turner, D. Tsipras, and A. Madry. Clean-label backdoor attacks, 2019. 3

[45] 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. 2

[46] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016. 2

[47] 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. 10

[48] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. M. Eckert, and F. Roli. Support vector machines under adversarial label contamination. *Neurocomputing*, 160, 2015. 2

[49] H. Xiao, H. Xiao, and C. M. Eckert. Adversarial label flips attack on support vector machines. In *ECAI*, 2012. 2

[50] T. Xiao, J. Hong, and J. Ma. Dna-gan: Learning disentangled representations from multi-attribute images. *arxiv preprint*, abs/1711.05415, 2018. 2

[51] 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. 10

[52] 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. 1

[53] Y. Zhang, H. Foroosh, P. David, and B. Gong. Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *ICLR*, 2019. 1

[54] Z. Zhao, D. Dua, and S. Singh. Generating natural adversarial examples. In *ICLR*, 2018. 1

[55] 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. 2