# SPATIALLY TRANSFORMED ADVERSARIAL EXAMPLES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recent studies show that widely used Deep neural networks (DNNs) are vulnerable to the carefully crafted adversarial examples. Many advanced algorithms have been proposed to generate adversarial examples by leveraging the $\mathcal{L}_p$ distance for penalizing perturbations. Different defense methods have also been explored to defend against such adversarial attacks. While the effectiveness of $\mathcal{L}_p$ distance as a metric of perceptual quality remains an active research area, in this paper we will instead focus on a different type of perturbation, namely spatial transformation, as opposed to manipulating the pixel values directly as in prior works. Perturbations generated through spatial transformation could result in large $\mathcal{L}_p$ distance measures, but our extensive experiments show that such spatially transformed adversarial examples are more perceptually realistic and more difficult to defend against with existing defense systems. This potentially provides a new direction in adversarial example generation and the design of corresponding defenses. We visualize the spatial transformation based perturbation for different examples and show that our technique can produce realistic adversarial examples with smooth image deformation. Finally, we visualize the attention of deep networks with different types of adversarial examples to better understand behaviors of these examples.

## 1 INTRODUCTION

Deep neural networks (DNNs) have demonstrated their outstanding performance in different domains, ranging from image processing (Krizhevsky et al., 2012; He et al., 2016), text analysis (Collobert & Weston, 2008) to speech recognition (Hinton et al., 2012). Though deep networks have exhibited high performance for these tasks, recently they have been shown to be particularly vulnerable to adversarial perturbations added to the input images (Szegedy et al., 2013; Goodfellow et al., 2015). These perturbed instances are called *adversarial examples*, which can lead to undesirable consequences in many practical applications based on DNNs. For example, adversarial examples can be used to subvert malware detection, fraud detection, or even potentially mislead autonomous navigation systems (Papernot et al., 2016b; Evtimov et al., 2017; Grosse et al., 2016) and therefore pose security risks when applied to security related applications. A comprehensive study about adversarial examples is required to motivate effective defenses. Different methods have been proposed to generate such adversarial examples. Fast gradient sign methods (FGSM) (Goodfellow et al., 2015) have been proposed to produce adversarial instances rapidly. Optimization based methods (Opt) have been applied to search for adversarial examples with smaller magnitude of perturbation (Carlini & Wagner, 2017a).

One important criterion for adversarial examples is that the perturbed images should “look like" the original instances. The traditional attack strategies adopt $L_2$ (or other $\mathcal{L}_p$) norm distance as a perceptual similarity metric to evaluate the distortion (Gu & Rigazio, 2014). However, this is not an ideal metric (Johnson et al., 2016; Isola et al., 2017), as $L_2$ similarity is sensitive to lighting and viewpoint change of a pictured object. For instance, an image can be shifted by one pixel, which will lead to large $L_2$ distance, while the translated image can appear “the same" to human perception. Motivated by this example, in this paper we aim to look for other types of adversarial examples and propose to create more perceptually realistic examples by changing the positions of pixels instead of directly manipulating existing pixel values. This has been shown to better preserve the identity and structure of the benigh image (Zhou et al., 2016b). Thus, the proposed spatially transformed adversarial example optimization method (stAdv) can keep adversarial examples less distinguishable from real instances (such examples can be found in Figure 3).

Various defense methods have also been proposed to defend against adversarial examples. Adversarial training based methods have so far achieved the most promising results (Goodfellow et al., 2015; Tramèr et al., 2017; Mądry et al., 2017). They have demonstrated the robustness of improved deep networks under certain constraints. However, the spatially transformed adversarial examples are generated through a rather different principle, whereby what is being minimized is the local distortion rather than the $\mathcal{L}_p$ pixel error between the adversarial and original instances. Thus, the previous adversarial training based defense method may appear less effective against this new attack given the fact that these examples generated by stAdv have never been seen before. This opens a new challenge about how to defend against such attacks, as well as other attacks that are not based on direct pixel value manipulation.

We provide visualization of the spatial deformation generated by stAdv; it is seen to be very smooth and virtually imperceptible to the human eye. In addition, to better understand the properties of deep neural networks on different adversarial examples, we provide visualization of the attention of the DNN given adversarial examples generated by different attack algorithms. We find that the spatial transformation based attack is more resilient across different defense models, including adversarially trained robust models.

Our contributions are summarized as follows:

- We propose to generate adversarial examples based on spatial transformation instead of direct manipulation of the pixel values, and we show realistic and effective adversarial examples on MNIST, CIFAR10, and ImageNet datasets.
- We provide the visualization of optimized transformation and show that such geometric changes are small and locally smooth, leading to high perceptual quality.
- We empirically show that, compared to other attacks, adversarial examples generated by stAdv are more difficult to detect with current defense methods.
- Finally, we also provide the visualization for the attention of deep networks on different adversarial examples and demonstrate that adversarial examples based on stAdv can more consistently mislead the adversarial trained robust deep networks compared to other existing attack methods.

## 2 Related work

Here we first briefly summarize the existing adversarial algorithms as well as the current defense methods. We then discuss the spatial transformation used in our adversarial attacks.

**Adversarial Examples** Given a benign sample $\mathbf{x}$, an attack instance $\mathbf{x}_{\text{adv}}$ is referred to as an adversarial example, if a small magnitude of perturbation $\epsilon$ is added to $\mathbf{x}$ (i.e. $\mathbf{x}_{\text{adv}} = \mathbf{x} + \epsilon$) so that $\mathbf{x}_{\text{adv}}$ is misclassified by the targeted classifier $g$. Based on the adversarial goal, attacks can be classified into two categories, targeted and untargeted attacks. In a targeted attack, the adversary's objective is to modify an input $\mathbf{x}$ such that the target model $g$ classifies the perturbed input $\mathbf{x}_{\text{adv}}$ in a *targeted* class chosen, which differs from its ground truth. In an untargeted attack, the adversary's objective is to cause the perturbed input $\mathbf{x}_{\text{adv}}$ to be misclassified in *any class* other than its ground truth. Based on the adversarial capabilities, these attacks can be categorized as white-box and black-box attacks, where an adversary has full knowledge of the classifier and training data in the white-box setting (Szegedy et al., 2014; Goodfellow et al., 2015; Carlini & Wagner, 2017a; Moosavi-Dezfooli et al., 2015; Papernot et al., 2016b; Biggio et al., 2013; Kurakin et al., 2016); while having zero knowledge about them in the black-box setting (Papernot et al., 2016a; Liu et al., 2017; Moosavi-Dezfooli et al., 2016; Mopuri et al., 2017). In this work, we will focus on the white-box setting to explore what a powerful adversary can do based on the Kerckhoffs's principle (Shannon, 1949) to better motivate defense methods.

**Spatial Transformation** In computer vision and graphics literature, two main aspects determine the appearance of a pictured object (Szeliski, 2010): (1) the *lighting and material*, which determine the brightness of a point as a function of illumination and object material properties, and (2) the *geometry*, which determines where the projection of a point will be located in the scene. Most previous adversarial attacks (Goodfellow et al., 2015) build on changing the *lighting and material*

aspect, while assuming the underlying geometry stays the same during the adversarial perturbation generation process.

Modeling geometric transformation with neural networks was first explored by "capsules," computational units that locally transform their input for modeling 2D and 3D geometric changes (Hinton et al., 2011). Later, Jaderberg et al. (2015) demonstrated that similar computational units, named spatial transformers, can benefit various recognition tasks. Zhou et al. (2016a) adopted the spatial transformers for synthesizing novel views of the same object and has shown that a geometric method can produce more realistic results compared to pure pixel-based method. Inspired by these works, we also use the spatial transformers to deform the input images, but with the different goal of generating realistic adversarial examples.

**Defensive Methods** Following the emergence adversarial examples, various of defense methods have been studied, including adversarial training (Goodfellow et al., 2015), distillation (Papernot et al., 2016c), gradient masking (Gu & Rigazio, 2014) and feature squeezing (Xu et al., 2017). However, it has been shown that these defenses can either be evaded by Opt attacks or only provide marginal improvements (Carlini & Wagner, 2017b; He et al., 2017). Among these defenses, adversarial training has achieved the state-of-the-art performance. Goodfellow et al. (2015) proposed to use the fast gradient sign attack as an adversary to perform adversarial training, which is much faster, followed by ensemble adversarial training (Tramèr et al., 2017) and projected gradient descent (PGD) adversarial training (Mądry et al., 2017). In this work, we explicitly analyze how effective the spatial transformation based adversarial examples are under these adversarial training based defense methods.

## 3 Generating Adversarial Examples

Here we first introduce several existing attack methods and then present our formulation for producing spatially transformed adversarial examples.

### 3.1 Problem Definition

Given a learned classifier $g : \mathcal{X} \to \mathcal{Y}$ from the feature space $\mathcal{X}$ to a set of classification outputs $\mathcal{Y}$ (e.g., $\mathcal{Y} = \{0, 1\}$ for binary classification), an adversary aims to generate adversarial example $\mathbf{x}_{\text{adv}}$ for an original instance $\mathbf{x} \in \mathcal{X}$ with its ground truth label $y \in \mathcal{Y}$, so that the classifier predicts $g(\mathbf{x}_{\text{adv}}) \neq y$ (untargeted attack) or $g(\mathbf{x}_{\text{adv}}) = t$ (targeted attack) where $t$ is the target class.

### 3.2 Background: Current pixel-value based attack methods

There have been a number of methods for generating adversarial examples, all built on directly modifying the pixel values of the original image.

The *fast gradient sign method* (FGSM) (Goodfellow et al., 2015) uses a first-order approximation of the loss function in order to construct adversarial samples for the adversary's target classifier $g$. The algorithm achieves untargeted attack by performing a single gradient ascent step: $\mathbf{x}_{\text{adv}} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \ell_g(\mathbf{x}, y))$, where $\ell_g(\mathbf{x}, y)$ is the loss function (e.g. cross-entropy loss) used to train the original model $g$, $y$ is the ground truth label and the hyper-parameter $\epsilon$ controls the magnitude of the perturbation. A targeted version of it can be done similarly.

*Optimization based attack* (Opt) produces an adversarial perturbation for a targeted attack based on certain constraints (Carlini & Wagner, 2017a; Liu et al., 2017) as formulated below:

$$\min ||\boldsymbol{\delta}||_p^2 \quad \text{s.t.} \qquad g(\mathbf{x} + \boldsymbol{\delta}) = t \quad \text{and} \quad \mathbf{x} + \boldsymbol{\delta} \in X,$$

where the $\mathcal{L}_p$ norm penalty ensures that the added perturbation $\epsilon$ is small. The same optimization procedure can achieve untargeted attacks with a modified constraint $g(\mathbf{x} + \boldsymbol{\delta}) \neq y$.

### 3.3 Our Approach: Spatially Transformed Adversarial Examples

All the existing approaches directly modify pixel values, which may produce noticeable artifacts. Instead, we aim to smoothly change the geometry of the scene while keeping the original appearance,

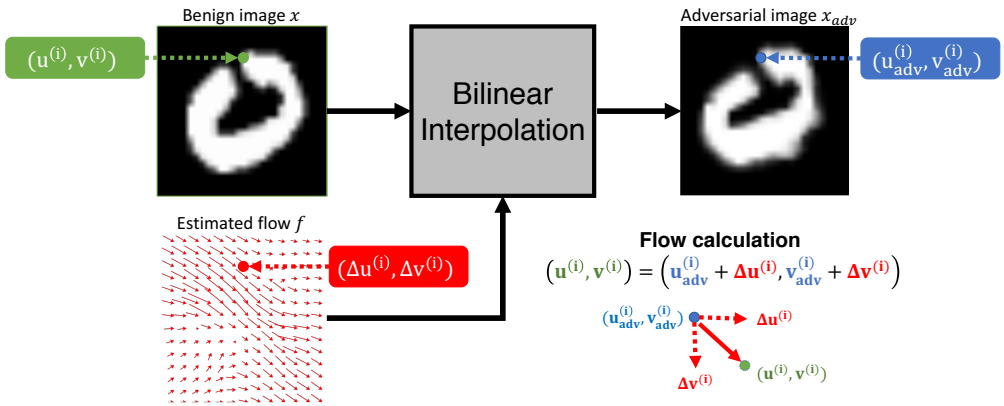



Figure 1: Generating adversarial examples with spatial transformation: the blue point denotes the coordinate of a pixel in the output adversarial image and the green point is its corresponding pixel in the input image. Red flow field represents the displacement from pixels in adversarial image to pixels in the input image.

producing more perceptually realistic adversarial examples. In this section, we introduce our geometric image formation model and then describe the objective for generating spatially transformed adversarial examples.

**Spatial transformation** We use $\mathbf{x}^{(i)}_{\text{adv}}$ to denote the value of the $i$-th pixel and 2D coordinates $(u^{(i)}_{\text{adv}}, v^{(i)}_{\text{adv}})$ to denote its location in the adversarial image $\mathbf{x}_{\text{adv}}$. We assume that $\mathbf{x}^{(i)}_{\text{adv}}$ is transformed from the pixel $\mathbf{x}^{(i)}$ from the original image. We use the per-pixel flow (displacement) field $f$ to synthesize the adversarial image $\mathbf{x}_{\text{adv}}$ using pixels from the input $\mathbf{x}$. For the $i$-th pixel within $\mathbf{x}_{\text{adv}}$ at the pixel location $(u^{(i)}_{\text{adv}}, v^{(i)}_{\text{adv}})$, we optimize the amount of displacement in each image dimension, with the pair denoted by the *flow vector* $f_i := (\Delta u^{(i)}, \Delta v^{(i)})$. Note that the flow vector $f_i$ goes from a pixel $\mathbf{x}^{(i)}_{\text{adv}}$ in the adversarial image to its corresponding pixel $\mathbf{x}^{(i)}$ in the input image. Thus, the location of its corresponding pixel $\mathbf{x}^{(i)}$ can be derived as $(u^{(i)}, v^{(i)}) = (u^{(i)}_{\text{adv}} + \Delta u^{(i)}, v^{(i)}_{\text{adv}} + \Delta v^{(i)})$. As the $(u^{(i)}, v^{(i)})$ can be fractional numbers and does not necessarily lie on the integer image grid, we use the differentiable bilinear interpolation (Jaderberg et al., 2015) to transform the input image with the flow field. We calculate $\mathbf{x}^{(i)}_{\text{adv}}$ as:

$$\mathbf{x}^{(i)}_{\text{adv}} = \sum_{q \in \mathcal{N}(u^{(i)}, v^{(i)})} \mathbf{x}^{(q)} (1 - |u^{(i)} - u^{(q)}|)(1 - |v^{(i)} - v^{(q)}|), \tag{1}$$

where $\mathcal{N}(u^{(i)}, v^{(i)})$ are the indices of the 4-pixel neighbors at the location $(u^{(i)}, v^{(i)})$ (top-left, top-right, bottom-left, bottom-right). We can obtain the adversarial image $\mathbf{x}_{\text{adv}}$ by calculating Equation 1 for every pixel $\mathbf{x}^{(i)}_{\text{adv}}$. Note that $\mathbf{x}_{\text{adv}}$ is differentiable with respect to the flow field $f$ (Jaderberg et al., 2015; Zhou et al., 2016b). The estimated flow field essentially captures the amount of spatial transformation required to fool the classifier.

**Objective function** Most of the previous methods constrain the added perturbation to be small regarding a $\mathcal{L}_p$ metric. Here instead of imposing the $\mathcal{L}_p$ norm, we introduce a new regularization loss $\mathcal{L}_{flow}$ on the local distortion $f$, producing higher perceptual quality for adversarial examples. Therefore, the goal of the attack is to generate adversarial examples which can mislead the classifier as well as minimizing the local distortion introduced by the flow field $f$.

Formally, given a benign instance $\mathbf{x}$, we obtain the flow field $f$ by minimize the following objective:

$$f^* = \underset{f}{\operatorname{argmin}} \quad \mathcal{L}_{adv}(x, f) + \tau \mathcal{L}_{flow}(f), \tag{2}$$

where $\mathcal{L}_{adv}$ encourages the generated adversarial examples to be misclassified by the target classifier. $L_{flow}$ ensures that the spatial transformation distance is minimized to preserve high perceptual

quality and $\tau$ balances the adversarial loss and the spatial transformation loss. The goal of $\mathcal{L}_{adv}$ is to guarantee the targeted attack $g(\mathbf{x}_{\text{adv}}) = t$ where $t$ is the targeted class, different from the ground truth label $y$. Recall that we transform the input image $\mathbf{x}$ to $\mathbf{x}_{\text{adv}}$ with the flow field $f$ (Equation 1).

In practice, directly enforcing $g(\mathbf{x}_{\text{adv}}) = t$ during optimization is highly non-linear, we adopt the objective function suggested in Carlini & Wagner (2017c).

$$\mathcal{L}_{adv}(x, f) = \max(\max_{i \neq t} g(\mathbf{x}_{\text{adv}})_i - g(\mathbf{x}_{\text{adv}})_t, 0), \tag{3}$$

where $g(x)$ represents the logit output of model $g$, and $g(x)_i$ denotes the $i$th element of the logit vector.

To compute $\mathcal{L}_{flow}$, we calculate the sum of spatial movement distance for any two adjacent pixels. Given an arbitrary pixel $p$ and its neighbors $q \in \mathcal{N}(p)$, we enforce the locally smooth spatial transformation perturbation $\mathcal{L}_{flow}$ based on the total variation (Rudin et al., 1992):

$$\mathcal{L}_{flow}(f) = \sum_{p}^{\textit{all pixels}} \sum_{q \in \mathcal{N}(p)} \sqrt{||\Delta u^{(p)} - \Delta u^{(q)}||_2^2 + ||\Delta v^{(p)} - \Delta v^{(q)}||_2^2}. \tag{4}$$

Intuitively, minimizing the spatial transformation can help ensure the high perceptual quality for stAdv, since adjacent pixels tend to move towards close direction and distance. We solve the above optimization with L-BFGS solver (Liu & Nocedal, 1989).

## 4 Experimental Results

In this section, we first show adversarial examples generated by the proposed spatial transformation method and analyze the properties of these examples from different perspectives. We then visualize the estimated flows for adversarial examples and show that with small and smooth transformation, the generated adversarial examples can already achieve a high attack success rate against deep networks. We also show that stAdv can preserve a high attack success rate against current defense methods, which motivates more sophisticated defense methods in the future. Finally, we analyze the attention regions of DNNs, to better understand the attack properties of stAdv.

**Experiment Setup** We set $\tau$ as 0.05 for all our experiments. We leverage L-BFGS (Liu & Nocedal, 1989) as our solver with backtracking linear search.

### 4.1 Adversarial Examples Based on Spatial Transformations

We show adversarial examples with high perceptual quality for both MNIST (LeCun & Cortes, 1998) and CIFAR-10 (Krizhevsky et al., 2014) datasets.

**stAdv on MNIST** In our experiments, we leverage three target models whose network architectures are shown in Table 4, to generate adversarial examples in the white-box setting on the MNIST dataset. Models A, B and C are derived from Tramèr et al. (2017), which represent different architectures. See Appendix A for more details. Table 1 presents the accuracy of pristine MNIST test data on each model as well as the attack success rate of adversarial examples generated by stAdv on these models. Figure 2 shows the adversarial examples against different models where the original instances appear in the diagonal. Each adversarial example achieves a targeted attack, with the target class shown on the top of the column. It is clear that the generated adversarial examples still appear to be in the same class as the original instance for humans. Another advantage for stAdv compared with traditional attacks is that examples based on stAdv seldom show noise pattern within the adversarial examples. Instead, stAdv smoothly deforms the digits and since such natural deformation also exists in the dataset digits, humans can barely notice such manipulation.

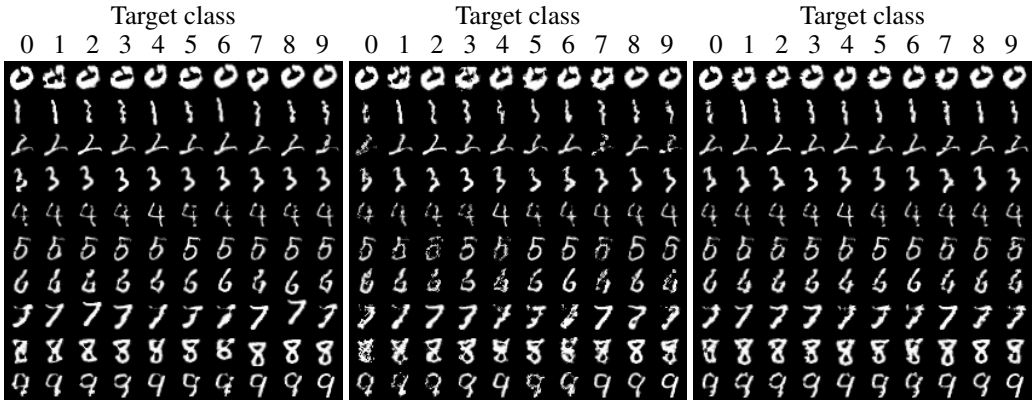
**stAdv on CIFAR-10** For CIFAR-10, we use models ResNet-32[1] and wide ResNet-34[2], as the target classifers (Zagoruyko & Komodakis, 2016; He et al., 2016; Mądry et al., 2017). We show

[1] https://github.com/tensorflow/models/blob/master/research/ResNet/ResNet_model.py

[2] https://github.com/MadryLab/cifar10_challenge/blob/master/model.py

Table 1: Accuracy of pristine data (p) on different models, and attack success rate of adversarial examples generated by stAdv on MNIST dataset.

| Model | A | B | C |
|---|---|---|---|
| Accuracy (p) | 98.58% | 98.94% | 99.11% |
| Attack Success Rate | 99.95% | 99.98% | 100.00% |




(a) Model A (b) Model B (c) Model C

Figure 2: Adversarial examples generated by stAdv against different models on MNIST. The ground truth images are shown in the diagonal and the rest are adversarial examples that are misclassified to the targeted class shown on the top.

the classification accuracy of pristine CIFAR-10 test data (p) and attack success rate of adversarial examples generated by stAdv on different models in Table 2. Figure 3 shows the generated examples on CIFAR-10 against different models. The original images are shown in the diagonal. The other images are targeted adversarial examples, with the index of the target class shown at the top of the column. Here we use "0-9" to denote the ground truth labels of images lying in the diagonal for each corresponding column. These adversarial examples based on stAdv are randomly selected from the instances that can successfully attack the corresponding classifier. Humans can hardly distinguish these adversarial examples from the original instances.

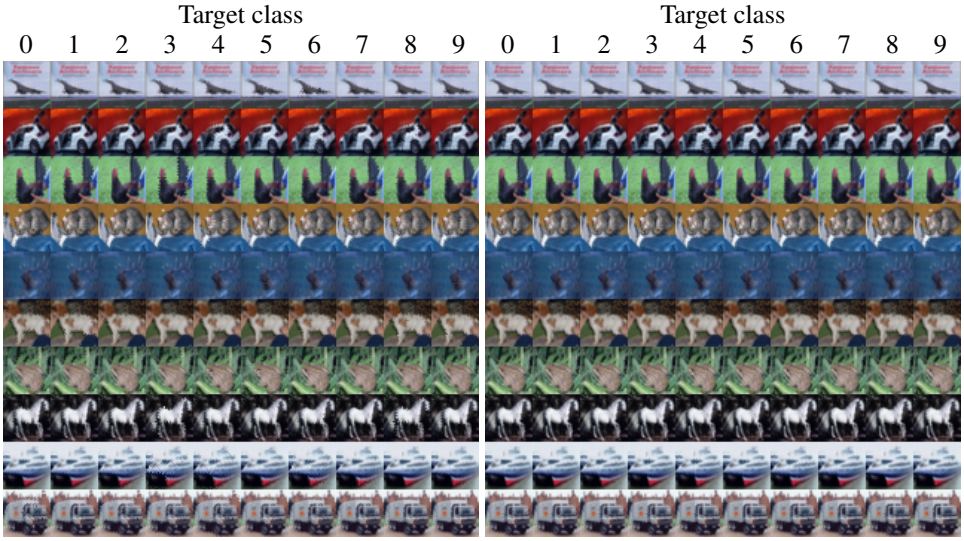
Table 2: Performance of different models on pristine (p) and stAdv adversarial examples on CIFAR-10 dataset. The number in parentheses is the number of parameters.

| Model | ResNet32 (0.47M) | Wide ResNet34 (46.16M) |
|---|---|---|
| Accuracy (p) | 93.16% | 95.82% |
| Attack Success Rate | 99.56% | 98.84% |

**Comparison of different adversarial examples** In Figure 4, we show adversarial examples that are targeted attacked to the same class ("0" for MNIST and "airplane" for CIFAR-10), which is different from their ground truth. We compare adversarial examples generated from different methods and show that those based on stAdv clearly preserve higher perceptual quality compared with FGSM and Opt methods.

### 4.2 Visualizing Spatial Transformation

In order to better understand the spatial transformation applied to the original images, we visualize the optimized transformation flow for different datasets, respectively. Figure 5 visualizes a transformation on an MNIST instance, where the digit "0" is misclassified as "2." We can see that the adjacent flows move in a similar direction in order to generate smooth figure. The flows are more focused on the edge of the digit and sometimes these flows move in different directions along the




(a) wide ResNet34 (b) ResNet32

Figure 3: Adversarial examples generated by stAdv against different models on CIFAR-10. The ground truth images are shown in the diagonal while the adversarial examples on each column are classified into the same class as the ground truth image within that column.

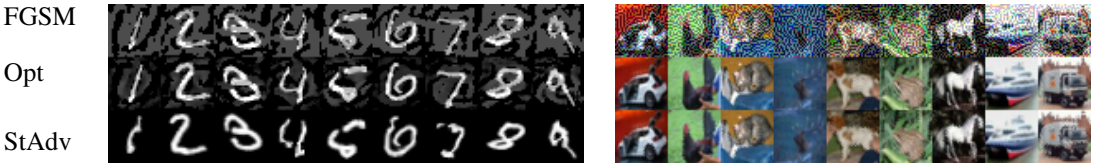



Figure 4: Comparison for adversarial examples generated by FGSM, Opt and stAdv. (Left: MNIST, right: CIFAR-10) The target class for MNIST is “0" and “air plane" for cifar.

edge, which implies that the object boundary plays an important role in our stAdv optimization. Figure 6 illustrates a similar visualization on CIFAR-10. It shows that the optimized flows often focus on the area of the main object, such as the air plane. We also observe that the magnitude of flows near the edge are usually larger, which similarly indicates the importance of edges for misleading the classifiers. This observation confirms the observation that when DNNs extract edge information in the earlier layers for visual recognition tasks (Viterbi, 1998). In addition, we visualize the similar flow for ImageNet (Krizhevsky et al., 2012) in Figure 7. The top-1 label of the original image in Figure 7 (a) is “mountain bike". Figure 7 (b)-(d) show targeted adversarial examples generated by stAdv, which have target classes “goldfish,” “maltese dog,” and “tabby cat,” respectively, and which

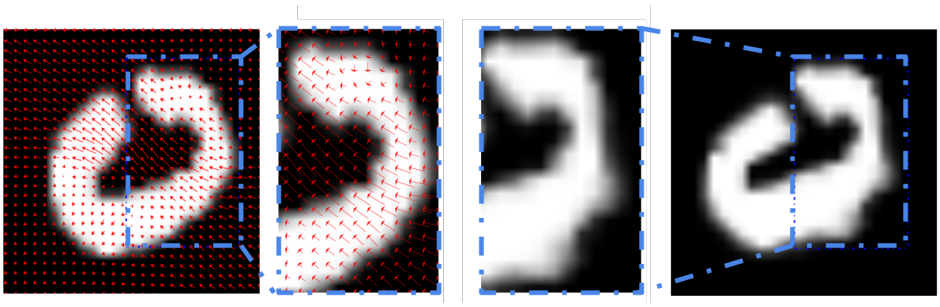


Figure 5: Flow visualization on MNIST. The digit “0" is misclassified as “2".

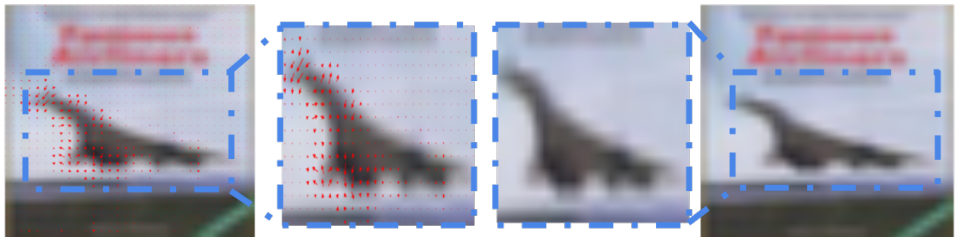


Figure 6: Flow visualization on CIFAR-10. The example is misclassified as bird.

are predicted as such as the top-1 class. An interesting observation is that, although there are other objects within the image, nearly 90% of the spatial transformation flows tend to focus on the target object bike. Different target class correspond to different directions for these flows, which still fall into the similar area.

(a) mountain bike (b) goldfish (c) Maltese dog (d) tabby cat

Figure 7: Flow visualization on ImageNet. (a): the original image, (b)-(c): images are misclassified into goldfish, dog and cat, respectively. Note that to display the flows more clearly, we fade out the color of the original image.

### 4.3 Attack Efficiency Under Defense Methods

Here we generate adversarial examples in the white-box setting and test defense methods against these samples to evaluate the strength of these attacks under defense. To evaluate the efficiency of defenses, we directly apply different defense methods on these generated adversarial examples to make predictions. We mainly focus on the adversarial training defenses, since they have demonstrated the state-of-the-art performance. We apply three defense strategies in our evaluation: the FGSM adversarial training (Adv.) (Goodfellow et al., 2015), ensemble adversarial training (Ens.) (Tramèr et al., 2017), and projectile gradient descent (PGD) adversarial training (Mądry et al., 2017) methods. For adversarial training purpose, we generate adversarial examples with 0.3 as the $L_\infty$ bound (Carlini & Wagner, 2017a). We test adversarial examples generated against model A, B, and C on MNIST as shown in table 4, and similarly adversarial examples generated against ResNet32 and wide ResNet34 on CIFAR-10.

The results on MNIST and CIFAR-10 dataset are shown in Table 3. We observe that the three defense strategies can achieve high performance (less than 10% attack success rate) against FGSM and Opt attacks. Note that adversarial examples generated by FGSM and Opt apply $0.3$ as the $L_\infty$ bound. For simplicity, we use confidence $\kappa = 0$ for both Opt and stAdv for a fair comparison. These defense methods only achieve low defense performance on stAdv, which improve the attack success rate to more than $30\%$ among all defense strategies. These results indicate that new type of adversarial strategy, such as our spatial transformation-based attack, may open new directions for developing better defence systems. However, for stAdv, we cannot use $\mathcal{L}_p$ norm to bound the distance as translating a image by one pixel may introduce large $\mathcal{L}_p$ penalty. We instead constrain the spatial transformation flow and show that our adversarial examples have high perceptual quality in Figures 2, 3, and 4.

Table 3: Attack success rate of adversarial examples generated by stAdv against models A, B, and C under standard defenses on MNIST, and against ResNet and wide ResNet on CIFAR-10.

| Model | Def. | FGSM | Opt. | stAdv |
|---|---|---|---|---|
| A | Adv. | 4.3% | 4.6% | 32.62% |
| | Ens. | 1.6% | 4.2% | 48.07% |
| | PGD | 4.4% | 2.96% | 48.38% |
| B | Adv. | 6.0% | 4.5% | 50.17% |
| | Ens. | 2.7% | 3.18% | 46.14% |
| | PGD | 9.0% | 3.0% | 49.82% |
| C | Adv. | 3.22% | 0.86% | 30.44% |
| | Ens. | 1.45% | 0.98% | 28.82% |
| | PGD | 2.1% | 0.98% | 28.13% |

| Model | Def. | FGSM | Opt. | stAdv |
|---|---|---|---|---|
| ResNet32 | Adv. | 13.10% | 11.9% | 43.36% |
| | Ens. | 10.00% | 10.3% | 36.89% |
| | PGD | 22.8% | 21.4% | 49.19% |
| wide ResNet34 | Adv. | 5.04% | 7.61% | 31.66% |
| | Ens. | 4.65% | 8.43% | 29.56% |
| | PGD | 14.9% | 13.90% | 31.6% |

### 4.4 Visualizing Attention of Networks on adversarial examples

In addition to the analyzing adversarial examples themselves, in this section, we further characterize these spatially transformed adversarial examples from the perspective of deep neural networks.

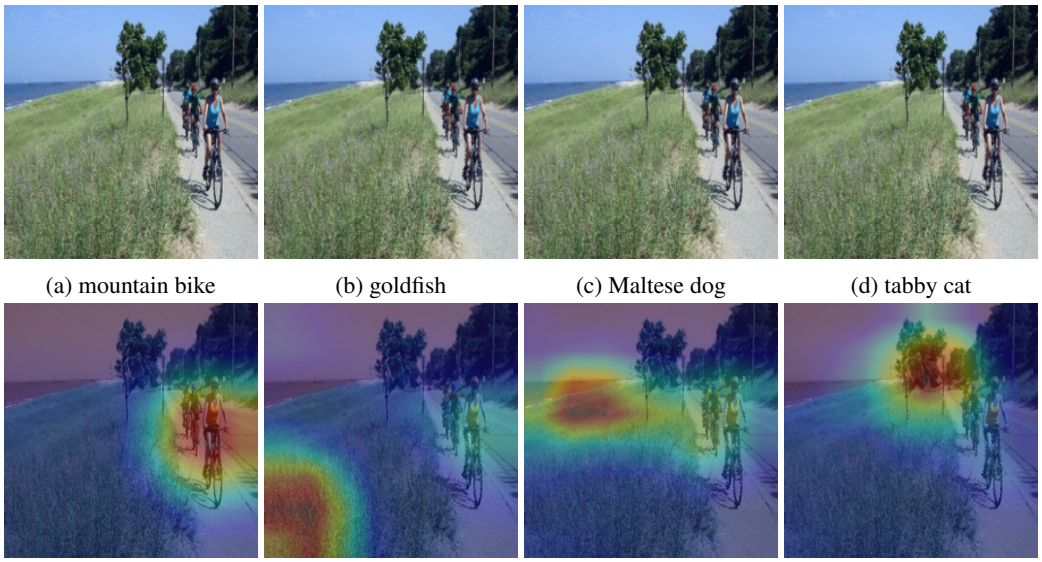
Here we apply Class Activation Mapping (CAM) (Zhou et al., 2016a), an implicit attention visualization technique for localizing the discriminative regions detected by a DNN. We use it to show the attention of the target ImageNet inception_v3 model (Szegedy et al., 2016)) for both the original image and generated adversarial examples. Figure 8(a) shows an input bike image and Figure 8(b)-(d) show the targeted adversarial examples based on stAdv targeting three different classes (goldfish, dog, and cat). Figure 8(e) illustrates that the target model draws attention to the bicycle region. Interestingly, attention regions on examples generated by stAdv varies for different target classes as shown in Figure 8(f)-(h). Though humans can barely distinguish between the original image and the ones generated by stAdv, CAM map focus on completely different regions, implying that our attack is able to mislead the network's attention.

In addition, we also compare and visualize the attention regions of both naturally trained and the adversarial trained inception_v3 model[3] on adversarial images generated by different attack algorithms (Figure 9). The ground truth top-1 label is "cinema," so the attention region for the original image (Figure 9 (a)) includes both tower and building regions. However, when the adversarial examples are targeted attacked into the adversarial label "missile," the attention region focuses on only the tower for all the attack algorithms as shown in Figure 9 (b)-(d) with slight different attention region sizes. More interestingly, we also test these adversarial examples on the public adversarial trained robust inception_v3 model. The result shows in Figure 9 (f)-(h). This time, the attention region is drawn to the building again for both FGSM and Opt methods, which is close to the attention region of the original image. The top-1 label for Figure 9 (f) and (g) are again the ground truth "cinema", which means they fail to attack the robust model. However, Figure 9 (h) is still misclassified as "missile" under the robust model and the CAM visualization shows that the attention region still focuses on the tower. This example again implies that adversarial examples generated by stAdv are challenging to defend for the current "robust" ImageNet models.

## 5 Conclusions

Different from the previous works that generate adversarial examples by directly manipulating pixel values, in this work we propose a new type of perturbation based on spatial transformation, which aims to preserve high perceptual quality for adversarial examples. We have shown that adversarial examples generated by stAdv are much more difficult for humans to distinguish from original instances. We also analyze the attack success rate of these examples under existing defense methods and demonstrate they are harder to defend against, which opens new directions for developing more robust defense algorithms. Finally, we visualize the attention regions of DNNs on our adversarial examples to better understand this new attack.

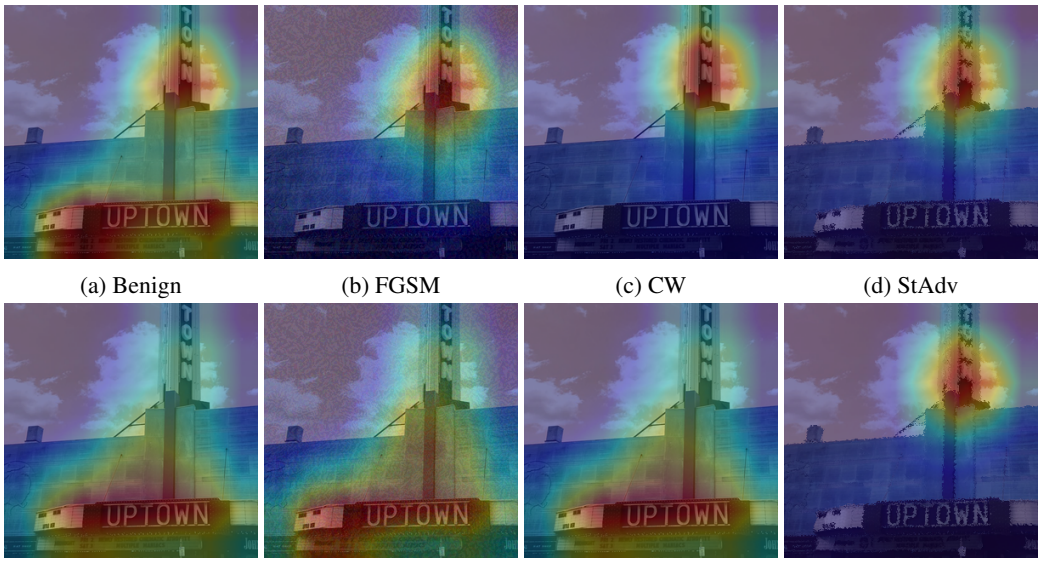
[3]https://github.com/tensorflow/cleverhans/tree/master/examples/nips17_adversarial_competition/



(a) mountain bike (b) goldfish (c) Maltese dog (d) tabby cat

(e) (f) (g) (h)

Figure 8: CAM attention visualization for ImageNet inception_v3 model. (a) the original image and (b)-(d) are stAdv adversarial examples targeting different classes. Row 2 show the attention visualization for the corresponding images above.




(a) Benign (b) FGSM (c) CW (d) StAdv

(e) Benign (f) FGSM (g) CW (h) StAdv

Figure 9: CAM attention visualization for ImageNet inception_v3 model. Column 1 shows the CAM map corresponding to the original image. Column 2-4 show the adversarial examples generated by different methods. The visualization is drawn for Row 1: inception_v3 model; Row 2: (robust) adversarial trained inception_v3 model. (a) and (e)-(g) are labeled as the ground truth "cinema", while (b)-(d) and (h) are labeled as the adversarial target "missile".

## REFERENCES

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 387–402. Springer, 2013.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy, 2017*, 2017a.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *arXiv preprint arXiv:1705.07263*, 2017b.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 39–57. IEEE, 2017c.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167. ACM, 2008.

Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2017.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. Adversarial perturbations against deep neural networks for malware classification. *arXiv preprint arXiv:1606.04435*, 2016.

Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defenses: Ensembles of weak defenses are not strong. *arXiv preprint arXiv:1706.04701*, 2017.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pp. 44–51. Springer, 2011.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, pp. 2017–2025, 2015.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 2014.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Yann LeCun and Corrina Cortes. The MNIST database of handwritten digits. 1998.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *arXiv preprint arXiv:1511.04599*, 2015.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. *arXiv preprint arXiv:1610.08401*, 2016.

Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. *arXiv preprint arXiv:1707.05572*, 2017.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083 [cs, stat]*, June 2017.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016a.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387. IEEE, 2016b.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pp. 582–597. IEEE, 2016c.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

Claude E Shannon. Communication theory of secrecy systems. *Bell Labs Technical Journal*, 28(4): 656–715, 1949.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Andrew J. Viterbi. An intuitive justification and a simplified implementation of the map decoder for convolutional codes. *IEEE Journal on Selected Areas in Communications*, 16(2):260–264, 1998.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016a.

Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, pp. 286–301. Springer, 2016b.

# A MODEL ARCHITECTURES

Table 4: Architecture of models applied on MNIST

| A | B | C |
|---|---|---|
| Conv(64,5,5) + Relu | Conv(64,8,8) + Relu | Conv(128,3,3) + Relu |
| Conv(64,5,5) + Relu | Dropout(0.2) | Conv(64,3,3) + Relu |
| Dropout(0.25) | Conv(128, 6, 6) + Relu | Dropout(0.25) |
| FC(128) + Relu | Conv(128, 5, 5) + Relu | FC(128) + Relu |
| Dropout(0.5) | Dropout(0.5) | Dropout(0.5) |
| FC(10) + Softmax | FC(10) +Softmax | FC(10)+Softmax |