Generative Adversarial Examples for Sequential Text Recognition Models with Artistic Text Style

Yanhong Liu[1], Fengming Cao[2] and Yuqi Zhang[2]
[1]*Mashang Consumer Finance, China*
[2]*Pingan International Smart City, China*

Keywords: Sequential Text Recognition, Adversarial, Generative Adversarial Networks, Artistic Text Style.

Abstract: The deep neural networks (DNNs) based *sequential text recognition* (STR) has made great progress in recent years. Although highly related to security issues, STR has been paid rare attention on its weakness and robustness. Most existing studies have generated adversarial examples for DNN models conducting non-sequential prediction tasks such as classification, segmentation, object detection etc. Recently, research efforts have shifted beyond the L_p norm-bounded attack and generated realistic adversarial examples with semantic meanings. We follow this trend and propose a general framework of generating novel adversarial text images for STR models, based on the technique of artistic text style transfer. Experimental results show that our crafted adversarial examples are highly stealthy and the attack success rates for fooling state-of-the-art STR models can achieve up to 100%. Our framework is flexible to create natural adversarial artistic text images with controllable stylistic degree to evaluate the robustness of STR models.

1 INTRODUCTION

The success of deep neural networks (DNNs) has boosted the development of text recognition tasks such as Optical Character Recognition and scene text recognition in recent years. These tasks are typically applied in security-critical applications like human computer interaction, assistant reading and road sign recognition etc. To robustly processing text images with various visual appearance and light conditions, people have solved the text recognition tasks as a *sequence labeling* problem, thus we denote such sequential recognition of text images by *sequential text recognition* (STR).

Despite their wide applications, DNNs have been shown to be vulnerable to adversarial examples (attacks) with small crafted perturbations on normal images (Szegedy et al., 2014; Goodfellow et al., 2015; Papernot et al., 2016). Most existing works generate adversarial examples by limiting the L_p norm (Carlini and Wagner, 2017; Madry et al., 2018) of the perturbations, which are useful for evaluating the weakness of the learning models. However, L_p-norm bounded attacks have limited practical utility since the perturbations in the pixel space cannot handle the underlying real-word properties of image formation that lead to them, such as translation,rotation and illumination conditions etc. These perturbed adversarial examples are often unnatural, not semantically meaningful and can be easily detected since they are distinctly identified in the latent space.

Recently, researchers have moved forward to generate adversarial examples with semantic meanings. Various techniques have been explored such as spatial transformation (Xiao et al., 2018b), changes in 3D physical properties (Liu et al., 2019) that the images are rendered from, photo-realistic manipulation of the color and texture of the images etc. The semantic attributes of images are also manipulated by perturbing the latent or feature space via the *generative adversarial networks* (GANs) (Zhao et al., 2018; Song et al., 2018; Wang et al., 2020; Qiu et al., 2020). Very recently, the technique of *neural style transfer* is applied to generate realistic adversarial examples (Duan et al., 2020) for physical-world attacks.

Existing works on adversarial examples mainly focus on non-sequential vision tasks such as image classification, object detection, face recognition etc. People have rarely tried to attack STR models, which presents a more difficult sequence-labeling problem. As we know, the only few works for STR (Xu et al., 2020a; Xu et al., 2020b) are generating adversarial examples based on traditional L_p-norm bounded attack.

Liu, Y., Cao, F. and Zhang, Y.
Generative Adversarial Examples for Sequential Text Recognition Models with Artistic Text Style.
DOI: 10.5220/0010866800003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 71-79
ISBN: 978-989-758-549-4; ISSN: 2184-4313

In this paper, following the up-to-date trend of adversarial attacks, we explore the possibility of generating adversarial examples for STR models based on the technique of style transfer. Fortunately, there exists a line of work which transforms text images with artistic style transfer (Yang et al., 2019b). By utilizing these techniques, we propose a general framework of generating adversarial artistic text images for STR models. As shown in Figure 1 (c) and (d), we hide the adversarial perturbations in the style texture on the target text body and its near neighborhood only, while L_2-norm bounded attack generates noise-like perturbations spread over the image. Furthermore, our framework allows the parametric control of the stylistic degree in terms of the text shape deformation. Extensive experimental results show that our approach generates highly natural adversarial artistic text images and can successfully fool the state-of-the-art STR models at a rate of up to 100%. The proposed mechanism provides a new way of evaluating the weakness of STR models, which can also be used to protect the user privacy in STR scenarios from being recognized by automatic deep learning systems.

2 RELATED WORK

2.1 Adversarial Examples

Research efforts have been paid to generate adversarial examples to fool the DNNs. Classic methods like Projected Gradient Descent (PGD) (Madry et al., 2018) and Carlini & Wagner (C&W) (Carlini and Wagner, 2017) craft the perturbations along the direction of adversarial gradients, which are bounded by a small L_p norm ball $\|\cdot\|_p \leq \varepsilon$.

In recent years, there is a movement beyond L_p norm-bounded attack, to generate perceptually realistic adversarial examples. Xiao et al. (Xiao et al., 2018a) proposed a general framework of adversarial GANs (AdvGAN) for this purpose. They also introduced the geometric image formation model and the perturbation of spatial transformation (Xiao et al., 2018b). Liu et al. (Liu et al., 2019) proposed a physically-based differentiable render that allows to propagate pixel gradients to the parametric 3D space of lighting and geometry. Bhattad et al. (Bhattad et al., 2020) manipulated the color and texture of the images to generate photo-realistic adversarial examples.

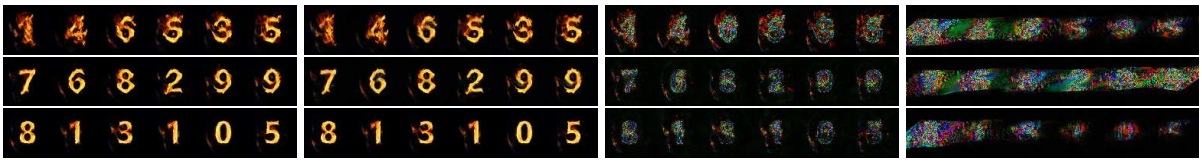
Semantically meaningful adversarial examples are also synthesized via GANs (Zhao et al., 2018; Song et al., 2018), by searching over the latent space. The semantic attributes of images are manipulated by perturbing disentangled latent codes (Wang et al., 2020) or using attribution-based image editing based on feature-space interpolation (Qiu et al., 2020). Semantic adversarial objects are synthesized (Shetty et al., 2020) by optimizing both appearance and positions of the objects for detectors.

Very recently, the technique of *neural style transfer* was explored for generating natural adversarial examples (Duan et al., 2020). We follow this line of utilizing style transfer for generating adversarial examples. However, Duan et al. (Duan et al., 2020) applied the traditional neural style transfer technique and transferred the texture of a style image to a user-specified region of the target image, where adversarial perturbations appear on the whole region in the target image. Instead of a global style transfer, we hide the adversarial perturbations in style texture which is limited on the text body and the very near neighborhood of the text. It is even more challenging for fooling the sequential recognition task models.

2.2 Sequential Text Recognition

The STR problem has been studied extensively in the area of scene text recognition. The state-of-the-art models treat the text recognition task as a sequence learning problem, which can be divided into four stages of consecutive operations (Baek et al., 2019): transformation (rectifying arbitrary text geometries), feature extraction (mapping the input image to a representation that focuses on the attributes relevant for character recognition, while suppressing irrelevant features such as font, color, size, and background), sequence modeling (capturing the contextual information within a sequence of characters) and prediction (estimating the output character sequence from the identified features of an image). The convolution neural network (CNN) and recurrent neural network (RNN) first encode the input image into a feature sequence. In the prediction phase, the connectionist temporal classification (CTC) or attention-mechanism (Attn) is used to predict the linguistic strings in the image, by constructing the alignment between the input images and their corresponding label sequence.

The only few work on generating adversarial examples for STR models (Song and Shmatikov, 2018; Xu et al., 2020a; Xu et al., 2020b) successfully attacked the CTC-based and attention-based STR models, using gradient-based optimization of the L_p norm ball of the perturbation. However, in this paper we explore a totally different attack mechanism, which aims to obtain natural and semantically meaningful adversarial examples for STR models.



(a) normal (b) adversarial (c) our perturbations (d) perturbs. by L_2 attack

Figure 1: Examples of (a) normal and (b) adversarial artistic *digit sequence* text images at three deformation levels, with the digits 0 to 9 in all the sequences recognized as 6,5,8,9,2,3,1,4,0,7 respectively, e.g. 146535 → 521393. Perturbations (amplified by 5x) of (c) ours are compared with (d) those by a L_2 PGD attack.

2.3 Artistic Text Style Transfer

Evolving from the problem of image style transfer, a series of work (Azadi et al., 2018; Yang et al., 2019a) has been conducted to transfer the source texture style to the target text glyph, forming a new text image with artistic style. Recently, the-state-of-the-art work (Yang et al., 2019b) along this line can stylize the text with arbitrary texture effects and control the degree of the glyph deformations with a parameterized fashion.

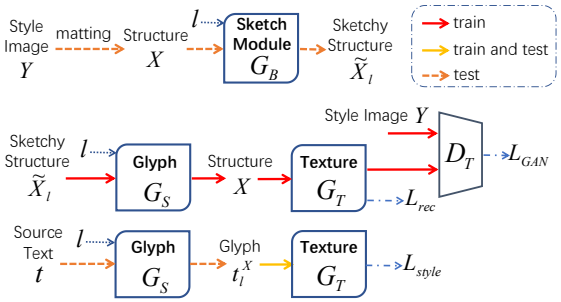


Figure 2: The process flow of Shape-Matching GAN.

3 SCALE-CONTROLLABLE ARTISTIC TEXT STYLE TRANSFER

In this paper, we propose a framework for generating adversarial text images, based on the technique of artistic text style transfer. We select the state-of-the-art work by (Yang et al., 2019b), called *Shape-Matching GAN* (SMG), since it can stylize the text with arbitrary textures and enable controllable glyph deformations, which shows promise for more application scenarios. Note that our proposed framework can also be extended for other artistic text style transfer techniques based on GANs (Azadi et al., 2018; Yang et al., 2019a).

In the following we briefly describe the concept of SMG. The reader can refer to (Yang et al., 2019b) for the full details. As shown in Figure 2, provided with the reference style image Y and the set of text images $\mathbb{T}$, the work designs a stylizing process to render each image $t \in \mathbb{T}$ with the texture of Y, where the deformation degree of the text glyphs can be controlled by a user-specified parameter $l \in [0,1]$. A larger value of l indicates a greater deformation degree. The render process is separated into two successive stages: structure transfer with the model G_S which generates text glyphs with controllable deformation degree, and texture transfer with the model G_T which renders the style texture on the text glyphs.

Glyph Network G_S. To obtain the structure transfer model G_S, a sketch module G_B is firstly trained on the set of source text images $\mathbb{T}$. A text image is smoothed at various level l by maintaining the contours of the text, after which the smoothed image is used for training to map it back to the text domain to learn the glyph characteristics.

The structure map X, which masks the shape of the style image Y, can be obtained by the existing image matting algorithms or Photoshop. A sketchy shape of X at the coarse level l, denoted by $\tilde{X}_l$, is then obtained with the shape characteristics of the text, by applying G_B (previously trained on the source text images) to X.

The glyph network G_S is trained to map $\tilde{X}_l$ to the original structure map X, so that it can characterize the shape features of X at the coarse level l. By applying G_S to the source text image t, the shape style of X is then transferred onto t and the structure transfer result t_l^X is obtained, showing text glyphs with deformation degree of l.

Texture Network G_T. As a generator component of SMG, the texture transfer network G_T is trained to render the texture of style image Y onto the text glyph image t_l^X to obtain the artistic text image t_l^Y, which is analogical to rendering the style texture of Y onto its structure map X to get Y.

The images of X and Y are randomly cropped to obtain adequate training pairs $\{x,y\}$. It is separated into two consecutive steps during each iteration of the training. Firstly with G_T fixed, a discriminator D_T is

trained to maximize the difference between the rendered $G_T(x)$ and the real image y. Then with D_T fixed, the generator G_T is trained to fool the discriminator. The adversarial loss function for the GAN can be restated with the formulation of Wasserstein GAN as follows:

$$L_{GAN} = \mathbb{E}_{x,y}[D_T(x,y)] - \mathbb{E}_{x,y}[D_T(x, G_T(x))] \quad (1)$$

A reconstruction loss L_{rec} is also used to minimize the difference between $G_T(x)$ and y in an L_1 sense:

$$L_{rec} = \mathbb{E}_{x,y}[\|G_T(x) - y\|_1] \quad (2)$$

A style loss L_{style} proposed in *neural style transfer* is also considered for the overall rendering performance on the source text image t. The final objective function for training the texture transfer network is defined as:

$$\begin{aligned} \mathcal{L} \quad = \quad & \min_{G_T} \max_{D_T} \{ \\ & \lambda_{GAN}\mathcal{L}_{GAN} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{style}\mathcal{L}_{style} \} \end{aligned} \quad (3)$$

4 ADVERSARIAL ARTISTIC TEXT GENERATION

As shown in Figure 3, in this paper we propose a framework for generating adversarial text images with artistic texture style which can mislead the STR models, by adapting the SMG technique presented in the last section and the adversarial GAN framework proposed in (Xiao et al., 2018a). We assume that the sketch module G_B and the glyph network G_S are already available, following the SMG process. We also pre-train a normal texture network G_T as a reference model, which renders the texture of style image Y on the text glyph images in a normal way. Our framework focuses on generating adversarial examples from the output of the pretrained glyph network G_S (i.e. t_l^X), and hiding the adversarial perturbations in the style texture rendered on/around the text glyphs. Note that it is not trivial to effectively generate style-based adversarial text examples without being perceived. We have to consider careful architecture adaptation and manipulation of loss functions.

We retrain a new adversarial texture transfer network $\bar{G}_T$, with the output of G_T as the reference artistic text image. The output of $\bar{G}_T$ is fed into the STR model f such that f is fooled. The work flow of our framework is detailed as follows.

4.1 Preprocessing

There could be several potential scenarios to apply our approach. For example, we may want to attack a given set of text images that may be stylized. Or we would just like to produce a graphic verification code, a poster or advertising board containing titles, brands, phone/address numbers etc., which we want to protect from being recognized by automatic deep learning systems.

Before we train the adversarial texture transfer network, we need to preprocess the source text images to obtain the set of text glyph images. First of all, we prepare the source text images as follows. Given a set of target text images to be attacked, we may apply the technique of destylization (Yang et al., 2019a) to remove the text effects, if any, from the existing images and acquire the set of source images $\mathbb{T}$ with only content features. Or else, we may just know the text labels to create adversarial examples for. In this case we prepare the images of individual characters (digits) for a word (digit sequence number), and then concatenate them to obtain the source image t.

Then, given the style image Y and the source text images $\mathbb{T}$, the process presented in Section 3 are then followed to train the sketch module G_B and the glyph network G_S. The text glyph image t_l^X under different deformation degrees, can be obtained by applying G_S to the source image $t \in \mathbb{T}$, with various pre-specified values of l.

4.2 Adversarial Texture Style Transfer

Based on the preprocessed text glyph images with deformation degree of l, we would retrain a new texture style transfer network $\bar{G}_T$ for generating adversarial text images with the texture of the style image Y. To enhance the stealthiness of the adversarial text images, it is not enough to just distinguish between the rendered style image $\bar{G}_T(x)$ and its real one y like that of SMG during the training.

Following the framework as shown in Figure 3, each text glyph image t_l^X is input into the normal texture network G_T to render it with style texture of Y without adversarial effects. The output of G_T, i.e. t_l^Y, is used as the reference artistic text image. At the same time, the adversarial texture transfer network $\bar{G}_T$ renders the input glyph image t_l^X with adversarial style texture, the output of which is denoted by $\bar{t}_l^Y$. The new discriminator $\bar{D}_T$ is also trained to distinguish between $\bar{t}_l^Y$ and the normally rendered t_l^Y.

The generated adversarial text image $\bar{t}_l^Y$ is used as the input of the target STR model f for recognition. We train the adversarial GAN including $\bar{G}_T$ and $\bar{D}_T$ such that the model f mis-recognizes the real text content in $\bar{t}_l^Y$. We achieve this goal by manipulating the loss functions based on the proposed framework.

Firstly, we keep the adversarial loss of the GAN

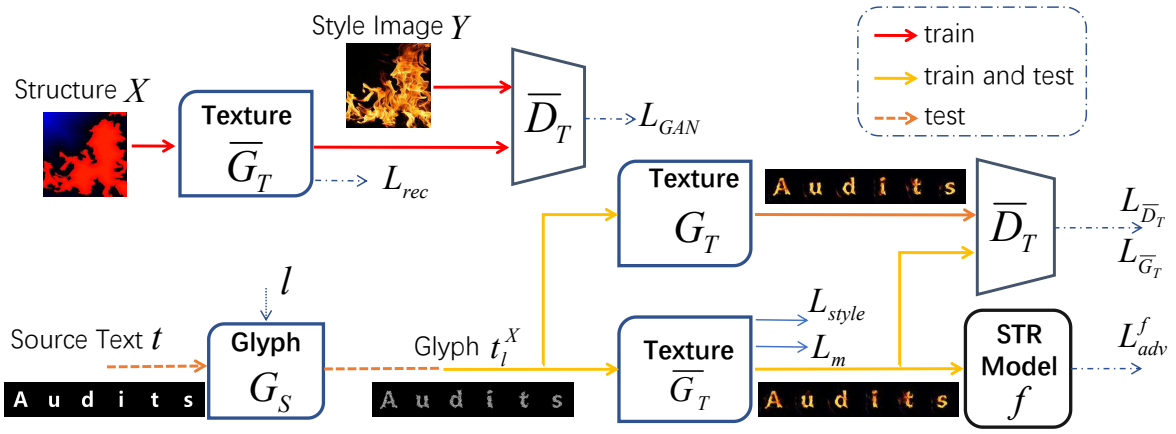


Figure 3: Overview of our framework for generating adversarial text images with artistic texture style transfer.

(denoted by $\mathcal{L}'_S$) and reconstruction loss (denoted by $\mathcal{L}'_{rec}$) for the style reference images $\{x,y\}$, as expressed in Eqns. (1) and (2), by replacing G_T and D_T with $\bar{G}_T$ and $\bar{D}_T$ respectively. Additionally, as we also apply the discriminator $\bar{D}_T$ on text images, we calculate the adversarial loss of the GAN for the text images as follows:

$$\mathcal{L}'_T = \mathbb{E}_{t_l^X}[\bar{D}_T(G_T(t_l^X))] - \lambda_{\bar{G}_T}\mathbb{E}_{t_l^X}[\bar{D}_T(\bar{G}_T(t_l^X))] \quad (4)$$

The hyper-parameter $\lambda_{\bar{G}_T}$ is used to control how the generated adversarial text image $\bar{G}_T(t_l^X)$ resembles the reference one $G_T(t_l^X)$.

Another adversarial loss is added to fool the target STR model f:

$$\mathcal{L}_{adv}^f = \mathbb{E}_{\bar{t}_l^Y}\mathcal{F}(\bar{t}_l^Y, \vec{W}) \quad (5)$$

$\mathcal{F}$ is the original loss function (CTC loss or cross entropy loss) for the target STR model. $\mathcal{L}_{adv}^f$ aims to fool the STR model f to incorrectly recognize the rendered adversarial image $\bar{t}_l^Y$ as the target sequence label $\vec{W}$.

Finally, we add a smoothness loss to reduce the variance between adjacent pixels in the adversarial text images:

$$\begin{aligned}\mathcal{L}_m &= \textstyle\sum_{i,j} \| \bar{t}_l^Y(i,j) - \bar{t}_l^Y(i+1,j) \|_2^2 \\ &\quad + \textstyle\sum_{i,j} \| \bar{t}_l^Y(i,j) - \bar{t}_l^Y(i,j+1) \|_2^2\end{aligned} \quad (6)$$

where $\bar{t}_l^Y(i,j)$ is the pixel value at coordinate (i,j) of image $\bar{t}_l^Y$. The smoothness loss helps to enhance the stealthiness and robustness of the adversarial images.

The total objective function for training the adversarial texture style transfer network can then be summarized as:

$$\begin{aligned}\mathcal{L}_{adv} &= \min_{\bar{G}_T}\max_{\bar{D}_T}\{ \\ &\quad \mathcal{L}'_T + \lambda_{adv}^f\mathcal{L}_{adv}^f + \lambda_m\mathcal{L}_m + \\ &\quad \lambda_{style}\mathcal{L}_{style} + \lambda_S\mathcal{L}'_S + \lambda_{rec}\mathcal{L}'_{rec}\}\end{aligned} \quad (7)$$

5 EXPERIMENTAL RESULTS

5.1 Setup

Datasets. During the experimentation, we prepared the datasets of the source text images as follows. Based on the images of the 10 Arabic digits, 26 capital and 26 small English letters that are available at the open source website of SMG (*https://github.com/VITA-Group/ShapeMatchingGAN, with MIT License*), we generated two types of datasets: one containing digit numbers and the other containing English words only, which represent the typical STR scenarios of a board containing address/phone numbers and brands/titles respectively. Note that our framework applies to other text shapes as well, only if the technique of artistic text style transfer works for. We may also apply a differentiable augmentation (Zhao et al., 2020) module after the texture transfer networks $\bar{G}_T$ and G_T shown in Figure 3, to get even more diverse examples for the STR model f. However, we concentrated on the style transfer based attack mechanism and put such augmentation out of the scope of this study.

For the digit dataset, we randomly generated 1000 six-digit numbers. The text image of a number was obtained by concatenating the corresponding image of each digit. The results of 1000 digit text images were then split into 800 and 200 ones respectively for the training and testing of the adversarial network.

For the word dataset, we sampled around 1800 English words of length 6 from the widely used synthetic dataset MJSynth (Jaderberg et al., 2014) designed for scene text recognition. The text image of each word was then generated by concatenating the corresponding image of each character. We split the dataset into around 1600 and 200 ones respectively

for the training and testing of the adversarial network.

We also used the style images provided from the website of SMG. All the digit/letter images downloaded from its website were resized to 256x256 pixels. Hence the created source text images of digit numbers and English words are of size 1536x256 pixels. Note that we used the fixed length of 6 digits/characters just for speeding up the training.

Target STR Models. We experimented with the five state-of-the-art models as implemented by (Baek et al., 2019), i.e. three CTC-based models: CRNN (None-VGG-BiLSTM-CTC), Rosetta (None-ResNet-None-CTC), STAR-Net (TPS-ResNet-BiLSTM-CTC) and two attention-based ones: RARE (TPS-VGG-BiLSTM-Attn), TRBA (TPS-ResNet-BiLSTM-Attn). These models cover the different combinations of the four-stage operations of STR. Different DNN network architectures of VGG and Resnet are applied for visual feature extraction. The Bidirectional LSTM (Bi-LSTM) is used as the (de-)selection in sequence modeling. CTC and attention schemes are adopted for sequence prediction. Although these models were originally proposed for scene text recognition, we believe that they are also good choices for general STR problems.

We pretrained the five STR models with datasets of normal artistic text images. Firstly, we generated 1000/5000 text images for the source digit/word dataset, following the way as described above for dataset preparation. Then, following the process as shown in Figure 2, we created the normal artistic text datasets with different style images and glyph deformation degrees, which contain around 9000 and 45000 samples for the digit and word set respectively.

The STR models were then trained on the digit and word datasets, so that they can recognize the normal artistic text images (resized to 384x64 pixels). The recognition accuracy of the five STR models achieved 100% on the digit dataset, and 99.95% (CRNN), 100.0% (Rosetta), 99.80% (STAR-Net), 99.93% (RARE), 100% (TRBA) respectively on the word dataset.

Implementation Details. Our generation of the adversarial artistic text examples was mainly based on the implementations of SMG. Given a specific style image, we adopted the pretrained glyph transfer network G_S and texture transfer network G_T that is used as the reference model for generating normal artistic text images. We generated adversarial examples at three coarse levels of $l = 0.0, 0.6, 1.0$ respectively, representing the *slight*, *moderate* and *heavy* deformation degrees in the text glyphs.

For all experiments, we set $\lambda_S = 1.0, \lambda_{rec} = 100$ and $\lambda_{style} = 0.01$, same as SMG. The number of epochs for training adversarial models was set to 300.

Threat Model. Our framework allows the generation of adversarial text images with artistic style. However, it is based on the training of a GAN structure and each digit or English character can be learned to be targeted to a pre-specified one. The untargeted attack for a text image can be naturally achieved by just assigning a target sequence label $\vec{W}$ (as specified in Eqn. (5)) which is different to the whole or part of the original digits/letters, so that the STR models incorrectly predict the text labels of the adversarial image. For the targeted attack case, the STR models are expected to recognize the labels of an adversarial image as the pre-specified ones. In practical use, we can apply our framework for attacking a small source text set where the target label for each digit/letter should be uniform for all samples. An extended dataset of moderate size (e.g. around 1000 and 1800 on digit and word datasets respectively in our case), which contains those digits/letters to be attacked, can be easily crafted for training the adversarial texture transfer network.

During the experiments, we firstly assumed a white-box scenario, where the network architecture and weight parameters of the STR models are known. Then we conducted a *cross-model* transfer attack, where the examples generated for a STR model are used to fool a different one.

5.2 Overall Results

Digit Dataset. During our experiments, we reshuffled the 10 digits randomly and assigned each reshuffled digit as the target label for its original one in all the digit text images. We then trained the adversarial network G_T for each STR model on the digit dataset as described previously. We set the parameters $\lambda_{G_T} = 0.1$, $\lambda^f_{adv} = 1.0$ and $\lambda_m = 0.001$. The *attack success rates* (ASRs), defined as the ratio of successful generation of adversarial examples, achieved 100% for all the five models. Figure 1 compares a few examples of the normal and adversarial artistic digit text images from the first row to the bottom, at the *heavy*, *moderate* and *slight* deformation levels respectively. It also shows that the perturbations generated by our framework possess the semantic meanings with style texture, compared to the noise-like perturbations by a L_2 norm based PGD attack.

Word Dataset. We randomly reshuffled 52 English letters and assigned the small capital version of each

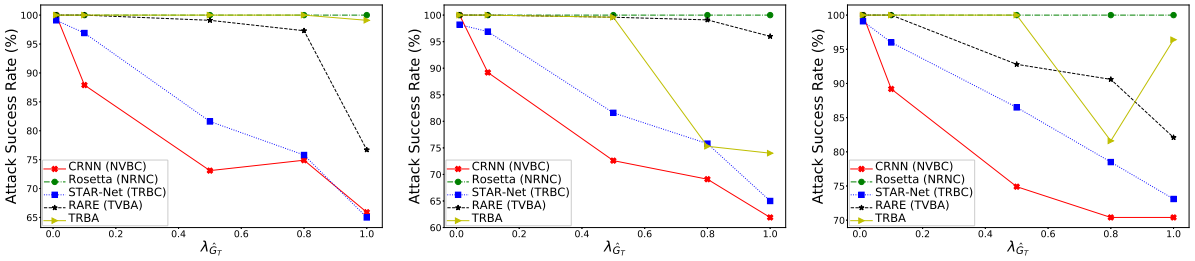
(a) normal (b) adversarial (c) our perturbations (d) perturbs. by L_2 attack

Figure 4: Examples of (a) normal and (b) adversarial artistic *word sequence* text images at three deformation degrees (rows 1/4, 2/5, 3/6 for heavy, moderate and slight one respectively), with $a \rightarrow w, e \rightarrow h, i \rightarrow f, d \rightarrow a, t \rightarrow c$ etc. Perturbations (amplified by 5x) of (c) ours are compared with (d) those by a L_2 PGD attack.

reshuffled character as the target label for its original one in all the word text images. The adversarial model $\bar{G}_T$ was trained on the word dataset so that the generated adversarial examples can mislead the STR models to predict each letter in a word as the targeted label. It is a rather tough task since each character in the word was attacked, while in the previous work for STR attack (Xu et al., 2020a) only small *edit distances* were applied. Figure 4 shows a few adversarial examples on the word dataset for the five STR models at three deformation degrees. It can be similarly observed that our generated adversarial examples have better perception, compared to the L_2 norm based PGD attack with noise-like perturbations. Note that the state-of-the-art L_2 STR attack by (Xu et al., 2020a) should exhibit similar noise-like perturbations. Our focus is the naturalness of the adversarial examples, and hence we did not bother with experiments with the work of (Xu et al., 2020a) since its source code is unavailable.

There should be a trade-off between the similarity of a generated example with its reference normal artistic text image, and its adversarial strength. We conducted extensive experiments on the word dataset by setting $\lambda_{\bar{G}_T}$ at various values to control how the adversarial text images are similar to their reference normal ones, with the adversarial parameters λ^f_{adv} and λ_m fixed to be 10 and 0.001 respectively. Figure 5 shows how the ASRs vary for the different settings of $\lambda_{\bar{G}_T}$ at three deformation degrees. It can be observed that our approach has the potential to successfully attack the STR models at a rate of up to 100%. The ASRs generally decease as the similarity with the reference images increases. The Rosetta model is the most vulnerable since it has no sequence modeling stage. The CRNN model is the most difficult to attack due to the CTC prediction scheme and the RNN sequence modeling. The CTC prediction scheme shows more robustness than the Attn scheme, when coupled with the RNN. It is different from the intuition that our generated examples show similar adversarial strengths at different deformation levels. However, it indeed has some effects on the two models using Attn scheme, where TRBA is more robust than RARE to the examples at the moderate deformation degree.

Cross-model Transfer Attack. We also conducted a *cross-model* transfer attack where the examples generated for one STR model is used to fool another one. We selected three adversarial texture style transfer models $\bar{G}_T$ corresponding to the three deformation degrees of $l = 1.0, 0.6, 0.0$ for each of the five STR models, all of which were trained with the parameter $\lambda_{\bar{G}_T} = 0.5$. Table 1 shows the ASRs of the examples generated from each of the adversarial models trained for one STR model, while attacking the other STR models. It can be observed that the CRNN model obtains the highest average ASR scores, while



(a) heavy deformation (b) moderate deformation (c) slight deformation

Figure 5: Attack success rates at different settings of $\lambda_{\bar{G_T}}$ and deformation degrees on the word dataset.

Table 1: Results of *cross-model* transfer attack on the word dataset.

Models		ASRs(%)					
	l	CRNN	Rosetta	STAR-Net	RARE	TRBA	Avg.
CRNN	1.0	*	73.5	60.5	73.1	75.8	70.7
	0.6	*	100	69.1	85.7	98.7	88.4
	0.0	*	100	55.2	76.7	76.7	77.2
Rosetta	1.0	7.2	*	27.8	30.0	57.4	30.6
	0.6	3.6	*	64.1	38.1	42.2	37.0
	0.0	4.9	*	58.3	32.7	44.8	35.2
STAR-Net	1.0	1.8	71.7	*	56.5	79.4	52.4
	0.6	3.6	58.7	*	67.3	77.1	51.7
	0.0	1.3	62.8	*	40.4	65.0	42.4
RARE	1.0	1.8	4.5	6.3	*	21.5	8.5
	0.6	1.8	56.1	23.8	*	49.8	32.9
	0.0	2.7	0.9	18.8	*	36.3	14.7
TRBA	1.0	16.6	7.6	53.4	83.6	*	40.3
	0.6	10.8	70.4	46.2	55.6	*	45.8
	0.0	23.3	73.5	62.3	78.9	*	59.5

the RARE model has the lowest scores. It indicates the mixed effects of different visual feature extraction (i.e. VGG and ResNet) and prediction schemes (i.e. CTC and Attn) on the results of cross-model transfer attack.

Human Perception Study. To quantify the perceptual realism of our generated adversarial artistic text images, we conducted a user study (Zhao et al., 2018; Song et al., 2018) to ask human participants to choose the more *visually realistic* image from a pair of an adversarial text image and its reference benign one generated with the normal texture transfer network. We selected 100 adversarial text images at various coarse levels from the results generated for the five state-of-the-art STR models. During each trial, an adversarial example is shown side-by-side with its reference one for 2 seconds. The user was then asked to make a decision.

In total, we collected around 1000 annotations from 40 users. Our generated adversarial text images were chosen as the more realistic in $49.60\% \pm 4.26\%$ of the trials (50% represents that users are unable to distinguish if an image is adversarial or not). This indicates that our framework can generate adversarial examples perceptually indistinguishable from their reference ones. Note that it is especially challenging to generate adversarial examples with high stealthiness in our setup that the images have clean background.

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed a framework of generating novel adversarial examples for state-of-the-art STR models, based on the technique of *artistic text style transfer*. Our framework is flexible in that it allows users to control the stylistic degree and can achieve the trade-off between the stealthiness and adversarial strength of the examples. Extensive experiments validated the effectiveness of our approach in fooling the STR models with visually realistic adversarial artistic text images.

Currently, our approach is dependent on the capability of the technique of artistic text style transfer. In the future, we may incorporate differential post-processing schemes (Zhan et al., 2019) into our framework, to generate rich and varied adversarial examples with real-world scenes. We may also explore to combine the techniques of manipulating latent codes with style transfer, to further enhance the generation process and the smoothness of the adversarial style texture.

REFERENCES

Azadi, S., Fisher, M., Kim, V. G., Wang, Z., Shechtman, E., and Darrell, T. (2018). Multi-content GAN for few-shot font style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S. J., and Lee, H. (2019). What is wrong with scene text recognition model comparisons? dataset and model analysis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Bhattad, A., Chong, M. J., Liang, K., Li, B., and Forsyth, D. A. (2020). Unrestricted adversarial examples via semantic manipulation. In *International Conference on Learning Representations (ICLR)*.

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*.

Duan, R., Ma, X., Wang, Y., Bailey, J., Qin, A. K., and Yang, Y. (2020). Adversarial camouflage: Hiding physical-world attacks with natural styles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Synthetic data and artificial neural networks for natural scene text recognition. *CoRR*, abs/1406.2227.

Liu, H. D., Tao, M., Li, C., Nowrouzezahrai, D., and Jacobson, A. (2019). Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. In *International Conference on Learning Representations (ICLR)*.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Papernot, N., Mcdaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. (2016). The limitations of deep learning in adversarial settings. In *IEEE Symposium on Security and Privacy*.

Qiu, H., Xiao, C., Yang, L., Yan, X., Lee, H., and Li, B. (2020). Semanticadv: Generating adversarial examples via attribute-conditional image editing. In *European Conference on Computer Vision (ECCV)*.

Shetty, R., Fritz, M., and Schiele, B. (2020). Towards automated testing and robustification by semantic adversarial data generation. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J., editors, *European Conference on Computer Vision (ECCV)*.

Song, C. and Shmatikov, V. (2018). Fooling OCR systems with adversarial text images. *CoRR*, abs/1802.05385.

Song, Y., Shu, R., Kushman, N., and Ermon, S. (2018). Constructing unrestricted adversarial examples with generative models. In *Annual Conference on Neural Information Processing Systems 2018 (NeurIPS)*.

Szegedy, H., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Wang, S., Chen, S., Chen, T., Nepal, S., Rudolph, C., and Grobler, M. (2020). Generating semantic adversarial examples via feature manipulation. *ArXiv*, abs/2001.02297.

Xiao, C., Li, B., Zhu, J., He, W., Liu, M., and Song, D. (2018a). Generating adversarial examples with adversarial networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., and Song, D. (2018b). Spatially transformed adversarial examples. In *International Conference on Learning Representations*.

Xu, X., Chen, J., Xiao, J., Gao, L., Shen, F., and Shen, H. T. (2020a). What machines see is not what they get: Fooling scene text recognition models with adversarial text images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, X., Chen, J., Xiao, J., Wang, Z., Yang, Y., and Shen, H. T. (2020b). Learning optimization-based adversarial perturbations for attacking sequential recognition models. In *International Conference on Multimedia*.

Yang, S., Liu, J., Wang, W., and Guo, Z. (2019a). TET-GAN: text effects transfer via stylization and destylization. In *AAAI Conference on Artificial Intelligence*, pages 1238–1245.

Yang, S., Wang, Z., Wang, Z., Xu, N., Liu, J., and Guo, Z. (2019b). Controllable artistic text style transfer via shape-matching GAN. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhan, F., Zhu, H., and Lu, S. (2019). Spatial fusion GAN for image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, S., Liu, Z., Lin, J., Zhu, J., and Han, S. (2020). Differentiable augmentation for data-efficient GAN training. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Zhao, Z., Dua, D., and Singh, S. (2018). Generating natural adversarial examples. In *International Conference on Learning Representations (ICLR)*.