DeepTag: Robust Image Tagging for DeepFake Provenance

Preprint · September 2020

CITATIONS
0

READS
599

7 authors, including:

Felix Juefei-Xu
Carnegie Mellon University
101 PUBLICATIONS **2,274** CITATIONS
SEE PROFILE

Qing Guo
Nanyang Technological University
73 PUBLICATIONS **1,449** CITATIONS
SEE PROFILE

Yang Liu
Nanyang Technological University
413 PUBLICATIONS **5,878** CITATIONS
SEE PROFILE

DeepTag: Robust Image Tagging for DeepFake Provenance

Run Wang[1*], **Felix Juefei-Xu**[2], **Qing Guo**[1], **Yihao Huang**[3], **Lei Ma**[4], **Yang Liu**[1], **Lina Wang**[5]

[1] Nanyang Technological University, Singapore
[2] Alibaba Group, USA
[3] East China Normal University, China
[4] Kyushu University, Japan
[5] Wuhan University, China

Abstract

In recent years, DeepFake is becoming a common threat to our society, due to the remarkable progress of generative adversarial networks (GAN) in image synthesis. Unfortunately, existing studies that propose various approaches, in fighting against DeepFake, to determine if the facial image is real or fake, is still at an early stage. Obviously, the current DeepFake detection method struggles to catch the rapid progress of GANs, especially in the adversarial scenarios where attackers can evade the detection intentionally, such as adding perturbations to fool DNN-based detectors. While passive detection simply tells whether the image is fake or real, DeepFake provenance, on the other hand, provides clues for tracking the sources in DeepFake forensics. Thus, the tracked fake images could be blocked immediately by administrators and avoid further spread in social networks.

In this paper, we investigated the potentials of image tagging in serving the DeepFake provenance. Specifically, we devise a deep learning-based approach, named *DeepTag*, with a simple yet effective encoder and decoder design to embed message to the facial image, which is to recover the embedded message after various **drastic** GAN-based DeepFake transformation with high confidence. The embedded message could be employed to represent the identity of facial images, which further contributed to DeepFake detection and provenance. Experimental results demonstrate that our proposed approach could recover the embedded message with an average accuracy of nearly 90%. Our research finding confirms effective privacy-preserving techniques for protecting personal photos from being DeepFaked. Thus, effective proactive defense mechanisms should be developed for fighting against DeepFakes, instead of simply devising DeepFake detection methods that can be mostly ineffective in practice.

1 Introduction

Capturing the exciting moments with camera and sharing them with friends over social networks (*e.g.*, Facebook, Twitter, Instagram) becomes a common activity in our daily life. However, with the recent development of GAN and its variants, our shared photos may suffer from being manipulated by various GANs to create DeepFakes (Mirsky and Lee 2020). Abusing the DeepFakes can bring potential threats and concerns to everyone, for example, releasing a realistic fake statement, creating fake pornography, *etc*. Additionally, many freely available tools (*e.g.*, FaceApp, ZAO) allow users to easily create DeepFakes on their own without any additional expertise. Thus, effective measures should be developed to fight against DeepFakes to protect our personal security and privacy.

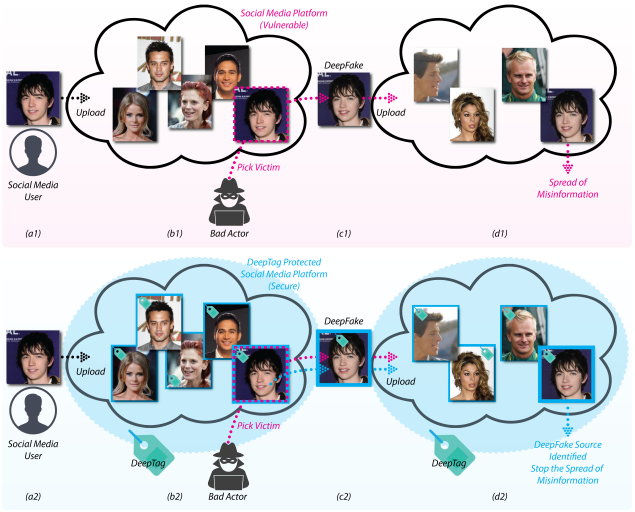


Figure 1: Comparison between a vulnerable social media platform (top panel) and a DeepTag protected social media platform (bottom panel) in handling malicious bad actors for spreading the misinformation by using DeepFake technology.

In fighting against DeepFakes, researchers are actively proposing various techniques to determine if a suspicious still image or video is real or fake passively. These studies mostly focus on the artifacts introduced in synthesizing the images with GANs. Identifying the synthesized images with observable artifacts (Li, Chang, and Lyu 2018; Yang, Li, and Lyu 2019) and detecting the synthesized images using deep neural networks (DNN) to spot the invisible artifacts (Dang et al. 2020; Wang et al. 2020d) are the two mainstream approaches in detecting DeepFakes. Unfortunately, our investigation into the artifact-based methods has revealed that they still suffer a lot from the following two issues.

- **Generalization**. Almost all the existing studies are focused on evaluating the effectiveness of their method on a limited number of known GANs. Since advanced GANs will be developed at an enormous speed and the artifacts

*Corresponding author, E-mail: wangrun@whu.edu.cn

which could be employed in previous GANs for distinguishing real and fake will likely be removed (Karras et al. 2020; Choi et al. 2020).

- **Robustness**. Simple image transformation (*e.g.*, resizing, compression, Gaussian noises) and adversarial attack with carefully crafted perturbations are two obstacles in developing robust detectors (Qian et al. 2020; Carlini and Farid 2020). Especially, the adversarial attacks by adding imperceptible noises can fool DNN-based detectors with high confidence in many cases (Carlini and Farid 2020; Huang et al. 2020b).

Undoubtedly, advanced GANs will be developed to produce high-quality synthesized images with fewer artifacts. These advanced GANs will be applied for creating DeepFakes maliciously and pose real challenges for detection since existing detectors are not generalized to unknown GANs. Furthermore, recent studies have demonstrated that DNN-based detectors are susceptible to adversarial noise attacks by adding imperceptible perturbations into the facial images. To address these two issues in passively defending DeepFakes, we propose a novel approach, named *DeepTag*, by protecting the safety and privacy of faces with image tagging to embed messages into the victim images and recover them to determine whether they are DeepFaked and manipulated by GANs proactively. Specifically, our proposed approach can be employed in DeepFake forensics for both detection and provenance purposes.

Threat Model: In this paper, our threat model is described in Fig. 1. A user could upload his/ her personal photos to social networks like Facebook and share it with friends or anyone. Unfortunately, attackers can easily pick victim's photos and manipulate them with various GANs to create DeepFakes they wanted, like releasing a fake statement in a video. The created DeepFakes will cause panic and raise security and privacy concerns for victims when it spreads on social networks. Our proposed DeepTag embeds message into the images before uploading to the social networks, after which it tries to recover the embedded message from a suspicious photo in social network for DeepFake detection and DeepFake provenance by determining the sources based on the recovered message. The key idea here is that our image tagging method should be robust enough to survive the drastic image transformation and reconstruction by the DeepFake process. Finally, the confirmed DeepFakes could be blocked and avoid further spreading.

Here are more details regarding Fig. 1. In the top panel, after a user (Fig. 1-a1) uploads his/her personal photos to the public domain social media platform, the personal picture can be picked up by a malicious actor (Fig. 1-b1). The bad actor can apply off-the-shelf DeepFake technology to produce a DeepFaked version of the user's face image (Fig. 1-c1). In this case, the male face is transformed to exhibit female's attribute, which is one example of how DeepFake can alter any face image without noticeable artifacts. Then, the bad actor can upload the DeepFaked face image to the same social media platform again (Fig. 1-d1), impersonating the user, or aiming at other malicious activities such as spreading misinformation. As can be seen, the unprotected social media platform is quite vulnerable in this scenario in terms of identifying the DeepFake images and preventing the spread of misinformation since no mechanism is established to distinguish between a legitimate face image and a DeepFake one.

On the contrary, in the bottom panel where the social media platform is protected by the proposed DeepTag mechanism, the spread of misinformation can be effectively prohibited. When a user uploads his/her personal photo (Fig. 1-a2) to the social media platform, the DeepTag is invoked to check whether this picture has been tagged by a DeepTag message before (usually a UID that matches the user's identity). If this face image is new, DeepTag can embed a message in the image, which is sufficiently robust to survive drastic image transformation such as DeepFake reconstruction. When a malicious bad actor (Fig. 1-b2) picks out the victim's photo and applies the DeepFake technique (Fig. 1-c2), the DeepTag message will survive. Then, when the bad actor tries to upload the DeepFaked face image to the social media platform again (Fig. 1-d2), the embedded DeepTag message will trigger an alarm since the UID of the original picture does not match the one of the bad actors, indicating a perpetrating event has happened. In this way, proper measures can be taken to stop the spread of misinformation such as blocking the uploading of the DeepFake face image, and/ or raising a red flag for this bad actor. In the bottom panel, the DeepTag protected images are represented by a green tag as well as a blue picture frame. In both panels, the pink arrows depict the route that a bad actor can take from picking a victim to the spread of misinformation. The blue arrow route indicates where the DeepTag message remains active during the whole process.

Our DeepTag is motivated by the existing studies on the privacy-preserving of multimedia, for example, digital watermarking for digital multimedia copyright protection (Katzenbeisser and Petitcolas 2000). Digital watermarking allows users to embed visible and invisible watermarking into the target multimedia (*e.g.*, text, image, audio). Our proposed image tagging is similar to digital watermarking. However, the difference between our image tagging and digital watermarking lies in that image tagging should survive after various drastic image transformation with GAN, while digital watermarking need to robust against common image transformation. In tackling the GAN-based manipulation, the following challenges need to be addressed in our image tagging. 1) **Diverse GANs**. Existing GANs for face synthesis could be classified as entire synthesis and partial synthesis, but the manipulation intensity of them is obviously different. DeepTag needs to tackle diverse GANs with various manipulation intensities. 2) **Unclear manipulated region**. In creating DeepFakes using GAN, the manipulated regions are unknown, thus the embedded message should avoid the effects of position where manipulation is performed.

To address the aforementioned challenges in embedding a message into the images, in this paper, our proposed DeepTag is based on a simple yet effective encoder and decoder architecture that could recover messages effectively even after drastic GAN-based transformation. The encoder and decoder are both DNNs and jointly trained. In DeepTag, a

DeepFake simulator connects the encoder and decoder to simulate various manipulation with GANs on the encoded images to enforce that the decoder could recover the embedded messages effectively after GAN-based transformation. To comprehensively evaluate the effectiveness of our DeepTag, our experiments are conducted on three state-of-the-art (SOTA) GANs including STGAN (Liu et al. 2019), StarGAN (Choi et al. 2018), and StyleGAN (Karras, Laine, and Aila 2019). These three GANs involve all the two typical GAN-based transformations, entire synthesis and partial synthesis. Experimental results have demonstrated that DeepTag achieves an average accuracy of nearly 90% in recovering the embedded messages.

Our main contribution are summarized as follows:

- **New idea in defending DeepFake with image tagging.** To the best of our knowledge, this is the first work proposing image tagging to achieve DeepFake forensics for both DeepFake detection and DeepFake provenance. Our proactive defense techniques could overcome the generalization and robustness issues in the traditional artifact-based DeepFake detection.
- **Performing a comprehensive evaluation of the effectiveness on typical GANs.** Experiments are conducted on three SOTA GANs spanning entire synthesis and partial synthesis. Experimental results demonstrated the effectiveness in embedding messages and recovering them after drastic GAN-based transformation.
- **New insight for defending DeepFakes.** Detecting synthesized images based on artifacts passively for DeepFakes is not enough for defending DeepFakes since they are not generic to unknown GANs and robust against adversary attack. Our approach presents a new insight by employing image tagging to protect the safety of images proactively.

2 Related Work

2.1 DeepFake Creation and Detection

GANs (Goodfellow et al. 2014) have achieved remarkable progress in image synthesis (Zhu et al. 2017) and voice synthesis (Oord et al. 2016), which are widely employed in creating realistic DeepFakes. In this paper, we mainly focus on image synthesis which plays a key role in creating modern DeepFakes. Entire synthesis and partial synthesis are two typical manipulations in facial image synthesis with GANs (Tolosana et al. 2020). In the entire synthesis, the whole synthesized images are totally generated by GANs and it can be used for synthesizing a new face that does not exist in the world. PGGAN (Karras et al. 2017) and StyleGAN (Karras, Laine, and Aila 2019) can generate high-resolution facial images to improve the quality of a given face. Specifically, StyleGAN has the capability to synthesize a non-existent face by utilizing the idea of style transfer. In the partial synthesis, the face attributes like hair, expression, are manipulated by GANs automatically. StarGAN (Choi et al. 2018), STGAN (Liu et al. 2019), and AttGAN (He et al. 2019) can edit the attributes in a fine-grained manner, for example, changing the hair color, wearing eyeglasses, turning the smiling expression into scared, *etc*. Thus, determining whether a facial image is manipulated by GANs provides a straightforward idea for detecting DeepFake.

Due to the imperfection design of existing GANs, the manipulated images with GAN inevitably introduces various artifacts. Existing studies on identifying DeepFakes are mostly leveraging the artifacts as clues. The artifacts can be classified as observable-artifacts noticed by human eyes and invisible-artifacts learned by DNN-based classifiers (Wang et al. 2020c; Zhang, Karaman, and Chang 2019).

Lyu *et al*. proposed to spot DeepFake video by observing the lack of eye blinking in the synthesized face (Li, Chang, and Lyu 2018). The inconsistent head poses in the synthesized face is another observable-artifacts in DeepFake videos (Yang, Li, and Lyu 2019). Some researchers also investigated the invisible-artifacts which could be used for spotting DeepFakes. Wang *et al*. observed that CNN-generated images contain common artifacts that could be identified by careful pre- and post-processing and data augmentation (Wang et al. 2020d). Frank *et al*. addressed the GAN-generated image identification with a basic observation that the artifacts revealed in the frequency domain (Frank et al. 2020). AutoGAN (Zhang, Karaman, and Chang 2019) observed the upsampling design in GAN will introduce artifacts in the synthesized images, thus they developed a GAN simulator to produce fake images and train a classifier to detect GAN-generated images. These proposed methods all claimed the effectiveness on seen GANs, but their capabilities on unknown GANs are still unclear.

2.2 Digital Watermarking

In the past decades, digital watermarking plays a key role in digital multimedia copyright protection. Digital watermarking indicates that the embedded watermark could be visible and invisible by human eyes and the embedded watermark in the carrier should be recovered even after various image transformation. Thus, robustness is the main concerns in designing an effective embedding algorithm (Katzenbeisser and Petitcolas 2000; Podilchuk and Delp 2001; Siddaraju, Jayadevappa, and Ezhilarasan 2015).

The spatial and frequency domain are two lines in embedding watermark into the carrier. Spatial domain is more easily to perform than the frequency domain, but it can be easily corrupted or attacked by attackers with pixel perturbations (Singh et al. 2012). The spatial domain techniques embed watermark by modifying the pixels value, such as the least significant bit (LSB) (Bamatraf, Ibrahim, and Salleh 2010). In embedding on the frequency domain, the carrier will be first converted into a specific transformation, then the watermark will be embedded in the transformation coefficients. The common frequency domains adopted in embedding watermarks include discrete cosine transform (DCT), discrete wavelet transform (DWT), discrete Fourier transform (DFT), and singular value decomposition (SVD) (Jiansheng, Sukang, and Xiaomei 2009; Khan et al. 2013; Yavuz and Telatar 2007).

With the rapid development of deep learning, end-to-end watermark embedding techniques are proposed in recent years. HiDDeN (Zhu et al. 2018) proposed the first end-

to-end framework by jointly training encoder and decoder network which could robust to noises like Gaussian blurring, pixel-wise dropout, *etc*. StegaStamp (Tancik, Mildenhall, and Ng 2020) presented a steganographic algorithm for embedding arbitrary hyperlink into the photos, which comprises a deep neural network for encoding and decoding. In addressing the unknown image distortion, Luo *et al*. proposed a framework for distortion agnostic watermarking (Luo et al. 2020) which are generic to unseen distortions.

3 Methodology

We present the very first facial image tagging approach for DeepFake provenance. We give our motivation and summarize the qualifications of a desirable image tagging solution against DeepFake in Section 3.1. Then, we establish the image tagging pipeline in Section 3.2.

3.1 Motivation

Existing techniques against DeepFake aim at observing the artifacts in the synthesized images with various methods. However, these studies suffer two issues, 1) they are not general to unknown GANs (Karras et al. 2020), 2) they are easily susceptible to adversarial attacks by adding perturbations intentionally or simple image transformation (*e.g*., compression, Gaussian noises) (Qian et al. 2020; Carlini and Farid 2020). Thus, the existing artifact-based techniques are not prepared in tackling the future emerging DeepFake threats.

Another straightforward idea for protecting facial images against DeepFake is that we can borrow the idea in privacy-preserving to defend DeepFakes proactively. Thus, we explore whether a robust image tagging can be served as a safeguard for protecting the safety of facial images in social networks against DeepFake. The image tagging allows us to easily conduct DeepFake detection and provenance with the embedded message. Our image tagging is similar to digital watermarking which is widely applied in protecting the copyright of digital multimedia (Ambadekar, Jain, and Khanapuri 2019), but ours has the following challenges which are vastly different to digital watermarking:

- Image tagging for DeepFake should be robust against GAN-based transformation, rather than simple image transformation like digital watermarking.
- The manipulated region in DeepFake is always unknown, however, the corrupted region in copyright protection can be figured out sometimes.

Inspired by the advances of deep learning in achieving end-to-end watermarking, we employ a DNN based encoder and decoder and jointly trained to enforce that the embedded message could survive various drastic GAN-based transformation. In the subsections, we introduce the pipeline of our proposed image tagging for DeepFake.

3.2 Image Tagging Pipeline

Overview Fig. 2 gives an overview of our proposed DeepTag overall architecture. Our method includes three key components, a DNN-based encoder F_{enc}, a GAN simulator G_{sim}, and a DNN-based decoder F_{dec}. The encoder and decoder are inspired by a previous work StegaStamp (Tancik, Mildenhall, and Ng 2020). Specifically, the functionalities of each component as follows.

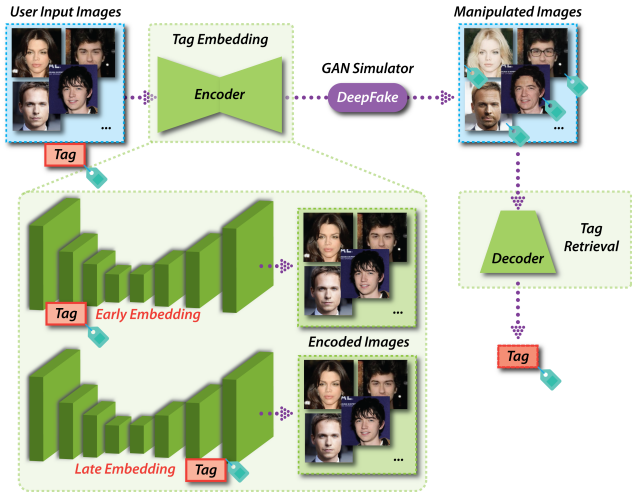


Figure 2: Training pipeline for DeepTag.

- The encoder F_{enc} embeds a message (usually a UID) into a facial image and ensures the embedded message invisible to human eyes. In other words, the encoded image needs to be perceptually similar to the input image.
- The GAN simulator G_{sim} is adopted for performing various GAN-based transformation, including entire synthesizing the encoded facial images, editing the attributes of encoded facial images.
- The decoder F_{dec} recovers the embedded message from the encoded facial images after drastic GAN-based transformation. The recovered UID is further used for the identity verification purpose.

Image tagging encoder-decoder training The DNN-based encoder and decoder are jointly trained to embed messages into the given input facial images. The encoder allows an arbitrary message to imperceptibly embed into the given arbitrary facial images. The decoder is trained to retrieve the embedded message even after drastic GAN-based manipulation. Here, the embedded message indicates n bits UID, but it can be easily extended to arbitrary binary bits.

Specifically, the encoder F_{enc} receives a facial image $\boldsymbol{i}$ and a message $\boldsymbol{w}$ as input, then the encoder output a tagged facial image $\widetilde{\boldsymbol{i}}$ with a mapping $F_{enc}(\boldsymbol{i}, \boldsymbol{w}) \mapsto \widetilde{\boldsymbol{i}}$. The input facial image $\boldsymbol{i}$ need to perceptually similar to the encoded facial image $\widetilde{\boldsymbol{i}}$, where $\boldsymbol{i} \approx \widetilde{\boldsymbol{i}}$. The encoded facial images may manipulated by GAN, where $G_{sim}(\widetilde{\boldsymbol{i}}) \mapsto \overline{\boldsymbol{i}}$. The decoder try to recover the embedded message $F_{dec}(\overline{\boldsymbol{i}}) \mapsto \widetilde{\boldsymbol{w}}$ or $F_{dec}(\widetilde{\boldsymbol{i}}) \mapsto \widetilde{\boldsymbol{w}}$, where $\widetilde{\boldsymbol{w}} \approx \boldsymbol{w}$.

To improve the capabilities of our decoder in recovering the embedded message from the encoded images, we need to explore where and when to embed message. In this paper, we embed the message in less manipulated regions and the late embedding level in the encoder. In manipulating faces

Type	Manipulation	GAN
Entire synthesis	full	StyleGAN
	identity swap	ZAO
Partial synthesis	facial attributes (*e.g.*, eyeglass, gender)	StarGAN
	facial expression (*e.g.*, smile, scared)	STGAN

Table 1: GANs adopted in creating DeepFake. The column *Type* indicates the fake type including entire synthesis and partial synthesis. The column *Manipulation* means the facial region that will be manipulated. The column *GAN* represents typical GANs. ZAO is an app for face manipulation, but the technical details are still unknown to us.

with GAN, the faces involve entire synthesis and partial synthesis, thus we employ masks to enforce more messages embedded in the region less manipulated by GANs. The level of embedding indicates when we can embed the message in the encoder. The late embedding in the encoder is less corrupted than early embedding since more layers are processed.

GAN-based manipulation DeepFake involves facial images manipulation with various GANs. Specifically, existing GANs can be classified into entire synthesis and partial synthesis. The details refer to Tab. 1. Our encoded facial images will be corrupted by these GAN-based manipulations. Thus, a GAN simulator performs the two typical manipulations by connecting our encoder and decoder to enforce that the decoder could learn how to recover message after drastic GAN-based manipulations.

Losses To train the encoder and decoder jointly, we use a series of losses in training. Particularly, we adopt the losses defined in StegaStamp (Tancik, Mildenhall, and Ng 2020). Here, we use L_2 residual regularization $\mathcal{L}_R$, the LPIPS perceptual loss (Zhang et al. 2018) $\mathcal{L}_P$, and a critic loss $\mathcal{L}_C$ calculated between encoded image and input image. We use cross entropy loss $\mathcal{L}_M$ for the message. The training loss is calculated as follows.

$$\mathcal{L} = \lambda_R \mathcal{L}_R + \lambda_P \mathcal{L}_P + \lambda_C \mathcal{L}_C + \lambda_M \mathcal{L}_M \tag{1}$$

where $\lambda_{R,P,C}$ should be set to zero initially while the decoder trains to high accuracy, after which $\lambda_{R,P,C}$ are increased linearly.

4 Experiments

In this section, we conduct experiments to evaluate the effectiveness of our proposed DeepTag in recovering the embedded message after drastic GAN-based manipulation. Specifically, we evaluate the effectiveness of DeepTag against three typical GANs including entire and partial synthesis and its robustness against image perturbations. Furthermore, we explore the performance in tackling different length of embedded messages and the impact on the level of embedding.

4.1 Experiment Setup

GANs In our experiments, we employ three GANs, *i.e.*, StarGAN (Choi et al. 2020), STGAN (Liu et al. 2019), and StyleGAN (Karras, Laine, and Aila 2019), since they achieved the state-of-the-art performance in faces manipulation. StyleGAN can reconstruct a given face and generate a new face. Both StarGAN and STGAN involve partial synthesis including facial attributes editing (*e.g.*, wearing eyeglasses, changing hair color) and expression manipulation (*e.g.*, smile, scared).

Dataset We employ CelebA-HQ (Karras et al. 2017) that is a public face dataset consisting 30,000 facial images and contains several different size facial images, such as 128×128, 512×512, and $1,024 \times 1,024$, *etc*. In our experiments, we explore the effectiveness of DeepTag in tackling facial images with different input size.

Metrics To evaluate the performance of DeepTag quantitatively, we employ accuracy to measure the recovered message after GAN-based manipulations. The accuracy indicates the full message retrieval rate (FMRR). Furthermore, PSNR and SSIM are adopted for calculating the similarity between the input and encoded facial images with DeepTag.

4.2 Implementation

Encoder Our encoder is trained to embed messages into carrier images while preserving the perceptual similar to the input carrier. Here, we use a U-Net (Ronneberger, Fischer, and Brox 2015) style architecture for receiving the input carrier images and output an encoded three-channel image. In our experiments, we explore different size of input carrier images (*e.g.*, 128×128, 512×512) and different length of embedded message (*e.g.*, 20 bits, 30 bits, 50 bits). Furthermore, the embedded message could be embedded in different levels in our encoder for achieving better performance in recovering the message in the decoder.

Decoder Our decoder is trained to retrieve the embedded message from the encoded images that are the output of our encoder. The decoder consists of seven convolutional layers with kernel size 3×3 and strides ≥ 1, one dense layer, and finally output the decoded message with the sigmoid activation function. The size of the decoded message is the same as the embedded message.

GAN simulator We employ three GANs that have achieved state-of-the-art performance in their field, *i.e.*, StarGAN, STGAN, and StyleGAN. The three GANs involve all the two typical GAN manipulation (*e.g.*, entire synthesis, partial synthesis) in creating DeepFakes.

Encoder and decoder training The encoder and decoder are jointly trained with randomly generated messages. The input images are collected from the public dataset CelebA-HQ. In training, we use 4 different sizes input facial images to train the model to explore the performance of DeepTag in tackling input faces of different sizes.

4.3 Effectiveness

In this section, we mainly explore the effectiveness of our proposed DeepTag in recovering the embedded messages with different GANs manipulation. Three GANs are adopted in our experiments for evaluation, namely StyleGAN for entire synthesis, STGAN and StarGAN for partial synthesis. Here, the length of the message is set to 30 bits, and the late embedding is implemented.

Image Size	Facial Attributes			
	bald	mustache	eyeglasses	plain skin
128 × 128	0.912	0.923	0.891	0.872
256 × 256	0.920	0.931	0.913	0.893
512 × 512	0.928	0.941	0.918	0.902
Average	0.920	0.932	0.907	0.889

Table 2: Performance (FMRR) of DeepTag on STGAN. The facial attributes mean the encoded images will be manipulated on such attributes. Manipulating the color of skin is the most drastic one.

Image Size	Facial Attributes			
	blond hair	gender	angry	happy
128 × 128	0.901	0.761	0.831	0.842
256 × 256	0.909	0.779	0.846	0.849
512 × 512	0.913	0.803	0.852	0.853
Average	0.908	0.781	0.843	0.848

Table 3: Performance (FMRR) of DeepTag on StarGAN. The facial attributes mean the encoded images will be manipulated on such attributes. Changing the hair color to blond is the most moderate one than the three other facial attributes.

Tab. 2 summarizes the performance of DeepTag in tackling the attributes manipulation with STGAN. In the experiments, the manipulated attributes include removing hair into bald, adding mustache, wearing eyeglasses, and changing into pale skin. Experimental results have shown that our DeepTag can perform well in the three former attributes manipulation, but susceptible to the skin color changing. The main reason is that the manipulation region is larger and the intensity is drastic than others. We also observe that the size of the input image has a positive impact on performance. Large size image can provide more space for embedding message and can survive in GAN-based manipulation more easily.

Tab. 3 presents the performance of DeepTag in dealing with StarGAN. The manipulated attributes include turning hair color to blond, changing gender, and facial expressions (angry & happy). Among these four facial attributes manipulation, the hair color changing is the most moderate one involving less region manipulation. Experimental results show that our DeepTag achieve an average accuracy of 90.8% on the hair changing manipulation. However, DeepTag gives an accuracy of 78.1% on the gender manipulation that involves the whole skin modification. In tackling the facial expression manipulation, DeepTag reaches an accuracy nearly 85% on the two common facial expression manipulation. Compared with the performance on STGAN and StarGAN, we can notice that the performance of DeepTag on STGAN is better than StarGAN due to the less artifact existed in STGAN.

StyleGAN indicates the entire synthesis which receives an input image and reconstructs it with less observable artifacts. In our experiments, we evaluate the performance on 1024×1024 size image with a pretrained model provided by StyleGAN[1]. Experimental results show that DeepTag achieves an accuracy of more than 95.1% on the large resolution facial images. It is interesting to explore the performance of low resolution with StyleGAN, but training StyleGAN for entire synthesis is extremely time-consuming and computing resource-intensive.

According to the experimental results of DeepTag on the three GANs, we can easily find that the input image size has a positive impact on the recovering message, while the manipulated region has a negative impact on the embedded message retrieval. Furthermore, advanced GANs with less artifact in the synthesized images could also reduce the negative impact in message retrieval. Similarly, the advanced GANs could also be employed for creating realistic DeepFakes in the future.

4.4 Impacts on the message size

Capacity is an important factor for measuring the capability of our DeepTag in embedding message. A large capacity indicates that the carrier can contain more information which could represent a large number of UID in our work. Thus, we explore the impact of message size on the performance of DeepTag in recovering messages.

Fig. 3 shows the relation between the accuracy of DeepTag in recovering messages and the length of message on three GANs. For STGAN and StarGAN, the input image size is 256×256 which is the most common size in sharing images on the social networks. We select the bald attribute for STGAN and blond hair attribute for StarGAN. These two attributes involve less manipulated regions, which could be better for us to illustrate the problem.

Experimental results show that the length of message has a negative impact on the performance of DeepTag in recovering embedded message. DeepTag can achieve an accuracy of more than 95% on the three GNAs when the size of embedded message is 20 bits. However, the accuracy reduces to less than 70% when the size of embedded message is 50 bits. Actually, the 30 bits of message can represent more than 1 billion different UIDs and the 35 bits can represent more than 34 billion UIDs. We believe that message with the 30 bits or 35 bits is enough for a social media platform to assign each user a specified UID.

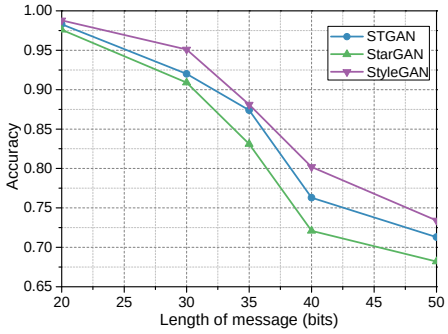


Figure 3: Performance of DeepTag on different size of messages.

4.5 Impact on the level of embedding

In DeepTag, the message could be embedded in the different levels of encoder, thus we explore the raw embedding, early embedding and late embedding. The raw embedding indicates the message embedded along with the carrier as

[1] https://github.com/Puzer/stylegan-encoder

Embedding	Facial Attributes			
	bald	mustache	eyeglasses	plain skin
Raw	0.880	0.898	0.903	0.861
Early	0.891	0.911	0.901	0.869
Late	0.920	0.931	0.913	0.893

Table 4: Performance (FMRR) of DeepTag on three different level embedding. STGAN is adopted for evaluation and the input image size is 256×256.

Metrics	GAN		
	STGAN	StarGAN	StyleGAN
PSNR ↑	27.48	26.32	29.89
SSIM ↑	0.901	0.862	0.927

Table 5: Image quality of the encoded images and input measured by PSNR and SSIM. For PSNR and SSIM, the higher the better.

input to the encoder. The early and late embedding means that the message embedded in the front and behind layer in the encoder, more details refer to Fig. 2.

Tab. 4 presents the performance of DeepTag on three different level embedding. The input image size is 256×256 and the adopted GAN is STGAN for its performance in partial synthesis. Experimental results show that the late embedding outperforms both raw and early embedding. Furthermore, the early embedding is better than raw embedding in most of the time, except wearing eyeglasses manipulated by STGAN. Experimental results in Tab. 4 indicates that the embedded message would be easily corrupted when more layers are processed in the encoder.

4.6 Quantitatively measuring encoded images

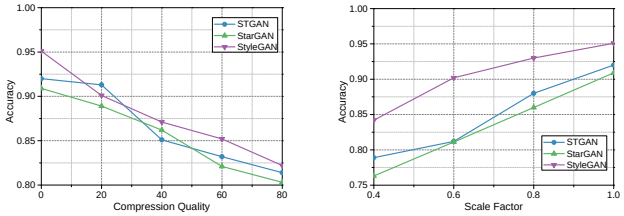
In DeepTag, the encoder outputs an encoded image with embedded message. Ideally, the encoded images should be perceptually similar to the input image. Here, we use two different metrics, PSNR and SSIM for measuring the distance between encoded image and input. Experimental result in Tab. 5 illustrates that our encoded image could maintain high visual quality. Furthermore, the StyleGAN achieved the best performance among the three GANs, due to that the entire synthesis exhibits less artifact.

4.7 Robustness against perturbations

In creating DeepFake videos, the manipulated images will be further processed by numerous image perturbations like compression, resizing, *etc*. In this section, we evaluated the robustness of DeepTag in tackling these image perturbations which are common appeared in producing videos.

Fig. 4 presents the robustness evaluation results of DeepTag on three GANs. In experiments, we employ two widely appeared perturbations in creating DeepFake videos, namely compression and resizing. For STGAN, we select the bald facial attributes for manipulation. For StarGAN, we select the blond hair attributes for manipulation. The StyleGAN involves the entire synthesis. The input image size for STGAN and StarGAN is 256×256, while the input image size for StyleGAN is 1024×1024.

Experimental results show that the accuracy in recovering message decreased when the compression quality increased,



(a) Compression (b) Resizing

Figure 4: Robustness evaluation with compression and resizing degradations.

but our DeepTag could achieve more than 80% in even when the compression quality is 80%. The higher compression quality indicates more messages will be discarded. In tackling the resizing perturbation, the performance of DeepTag is similar to compression where the scale factor has positive impacts on the accuracy. The small scale factor means a large resizing and it further results in a bad performance for DeepTag in recovering. In the real scenario, we can see that the compression quality for compression and the scale factor for resizing will not be too bad that touches the boundary. To some extent, our DeepTag can be well applied for real application in considering the robustness against perturbations.

Our pioneering work leverages image tagging for defending DeepFakes proactively. In the performance evaluation, we consider the most strict case where all the bits are fully recovered. DeepTag will have an even broader application, more robustness, and stronger resilience when partial errors could be tolerated in the retrieval or applying the redundancy code design techniques for embedding message.

5 Conclusions

In this paper, we proposed DeepTag that embeds messages into the images for DeepFake provenance. To the best of our knowledge, this is the first work that presents a new insight for fighting against DeepFake from the perspective of privacy-preserving, which aims to defend DeepFake proactively. Experiments on three typical GANs including the entire synthesis and partial synthesis demonstrate the effectiveness of our method in embedding watermarking into facial images and recovering them from facial images after drastic GAN-based transformation.

With the rapid development of AI-techniques, nobody can imagine future advances in producing DeepFakes. We can confirm that the DeepFake will become more and more realistic and everyone could fall victim. In this AI era, we are living in a world where we cannot believe our eyes anymore. However, detecting DeepFakes by observing the artifacts in the synthesized images is obviously insufficient for protecting us against this AI risk. Our work poses a new insight for fighting against DeepFakes proactively, instead of observing the artifacts by leveraging domain knowledge in synthesized images which could be easily invalid in unseen GANs. In future work, the community needs to develop more powerful defense strategies by considering how to protect images to avoid DeepFake threats.

Another orthogonal research direction is to investigate the interplay between DeepTag-based provenance technique

and the state-of-the-art DeepFake detection methods (Qi et al. 2020; Wang et al. 2020b; Huang et al. 2020d) as well as methods that can help DeepFakes more detection-evasive (Huang et al. 2020c,a). The effectiveness of the DeepTag under the presence of various adversarial perturbation, especially those that are not purely based on additive noise, such as (Gao et al. 2020; Cheng et al. 2020b; Tian et al. 2020; Zhai et al. 2020; Guo et al. 2020b; Wang et al. 2020a; Guo et al. 2020a; Cheng et al. 2020a) is also worth carefully studying.

References

Ambadekar, S. P.; Jain, J.; and Khanapuri, J. 2019. Digital image watermarking through encryption and DWT for copyright protection. In *Recent Trends in Signal and Image Processing*, 187–195. Springer.

Bamatraf, A.; Ibrahim, R.; and Salleh, M. N. B. M. 2010. Digital watermarking algorithm using LSB. In *2010 International Conference on Computer Applications and Industrial Electronics*, 155–159. IEEE.

Carlini, N.; and Farid, H. 2020. Evading Deepfake-Image Detectors with White-and Black-Box Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 658–659.

Cheng, Y.; Guo, Q.; Juefei-Xu, F.; Xie, X.; Lin, S.-W.; Lin, W.; Feng, W.; and Liu, Y. 2020a. Pasadena: Perceptually Aware and Stealthy Adversarial Denoise Attack. *arXiv preprint arXiv:2007.07097* .

Cheng, Y.; Juefei-Xu, F.; Guo, Q.; Fu, H.; Xie, X.; Lin, S.-W.; Lin, W.; and Liu, Y. 2020b. Adversarial Exposure Attack on Diabetic Retinopathy Imagery. *arXiv preprint arXiv* .

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8188–8197.

Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; and Jain, A. K. 2020. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5781–5790.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. *arXiv preprint arXiv:2003.08685* .

Gao, R.; ; Guo, Q.; Juefei-Xu, F.; Yu, H.; Ren, X.; Feng, W.; and Wang, S. 2020. Making Images Undiscoverable from Co-Saliency Detection. *arXiv preprint arXiv* .

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Guo, Q.; Juefei-Xu, F.; Xie, X.; Ma, L.; Wang, J.; Feng, W.; and Liu, Y. 2020a. ABBA: Saliency-Regularized Motion-Based Adversarial Blur Attack. *arXiv preprint arXiv:2002.03500* .

Guo, Q.; Xie, X.; Juefei-Xu, F.; Ma, L.; Li, Z.; Xue, W.; Feng, W.; and Liu, Y. 2020b. SPARK: Spatial-aware online incremental attack against visual tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

He, Z.; Zuo, W.; Kan, M.; Shan, S.; and Chen, X. 2019. AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing* 28(11): 5464–5478.

Huang, Y.; Juefei-Xu, F.; Guo, Q.; Xie, X.; Ma, L.; Miao, W.; Liu, Y.; and Pu, G. 2020a. FakeRetouch: Evading DeepFakes Detection via the Guidance of Deliberate Noise. *arXiv preprint arXiv* .

Huang, Y.; Juefei-Xu, F.; Wang, R.; Guo, Q.; Ma, L.; Xie, X.; Li, J.; Miao, W.; Liu, Y.; and Pu, G. 2020b. FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction. *arXiv preprint arXiv:2006.07533* .

Huang, Y.; Juefei-Xu, F.; Wang, R.; Guo, Q.; Ma, L.; Xie, X.; Li, J.; Miao, W.; Liu, Y.; and Pu, G. 2020c. FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

Huang, Y.; Juefei-Xu, F.; Wang, R.; Guo, Q.; Xie, X.; Ma, L.; Li, J.; Miao, W.; Liu, Y.; and Pu, G. 2020d. FakeLocator: Robust Localization of GAN-Based Face Manipulations. *arXiv preprint arXiv:2001.09598* .

Jiansheng, M.; Sukang, L.; and Xiaomei, T. 2009. A digital watermarking algorithm based on DCT and DWT. In *Proceedings. The 2009 International Symposium on Web Information Systems and Applications (WISA 2009)*, 104. Citeseer.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* .

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

Katzenbeisser, S.; and Petitcolas, F. 2000. Digital watermarking. *Artech House, London* 2.

Khan, M. I.; Rahman, M.; Sarker, M.; Hasan, I.; et al. 2013. Digital watermarking for image authenticationbased on combined dct, dwt and svd transformation. *arXiv preprint arXiv:1307.6328* .

Li, Y.; Chang, M.-C.; and Lyu, S. 2018. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7. IEEE.

Liu, M.; Ding, Y.; Xia, M.; Liu, X.; Ding, E.; Zuo, W.; and Wen, S. 2019. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3673–3682.

Luo, X.; Zhan, R.; Chang, H.; Yang, F.; and Milanfar, P. 2020. Distortion Agnostic Deep Watermarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13548–13557.

Mirsky, Y.; and Lee, W. 2020. The Creation and Detection of Deepfakes: A Survey. *arXiv preprint arXiv:2004.11138* .

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* .

Podilchuk, C. I.; and Delp, E. J. 2001. Digital watermarking: algorithms and applications. *IEEE signal processing Magazine* 18(4): 33–46.

Qi, H.; Guo, Q.; Juefei-Xu, F.; Xie, X.; Ma, L.; Feng, W.; Liu, Y.; and Zhao, J. 2020. DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues. *arXiv preprint arXiv:2007.09355* .

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Siddaraju, P. M.; Jayadevappa, D.; and Ezhilarasan, K. 2015. Digital image watermarking techniques: a review. *Int. J. Comput. Sci. Secur* 9(3): 140–156.

Singh, A. K.; Sharma, N.; Dave, M.; and Mohan, A. 2012. A novel technique for digital image watermarking in spatial domain. In *2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing*, 497–501. IEEE.

Tancik, M.; Mildenhall, B.; and Ng, R. 2020. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2117–2126.

Tian, B.; Guo, Q.; Juefei-Xu, F.; Chan, W.; Cheng, Y.; Li, X.; Xie, X.; and Qin, S. 2020. Bias Field Poses a Threat to DNN-based X-Ray Recognition. *arXiv preprint arXiv* .

Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; and Ortega-Garcia, J. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv preprint arXiv:2001.00179* .

Wang, R.; Juefei-Xu, F.; Guo, Q.; Huang, Y.; Xie, X.; Ma, L.; and Liu, Y. 2020a. Amora: Black-box Adversarial Morphing Attack. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

Wang, R.; Juefei-Xu, F.; Huang, Y.; Guo, Q.; Xie, X.; Ma, L.; and Liu, Y. 2020b. DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*.

Wang, R.; Juefei-Xu, F.; Ma, L.; Xie, X.; Huang, Y.; Wang, J.; and Liu, Y. 2020c. FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020d. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 7.

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8261–8265. IEEE.

Yavuz, E.; and Telatar, Z. 2007. Improved SVD-DWT based digital image watermarking against watermark ambiguity. In *Proceedings of the 2007 ACM symposium on Applied computing*, 1051–1055.

Zhai, L.; Juefei-Xu, F.; Guo, Q.; Xie, X.; Ma, L.; Feng, W.; Qin, S.; and Liu, Y. 2020. It's Raining Cats or Dogs? Adversarial Rain Attack on DNN Perception. *arXiv preprint arXiv* .

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, X.; Karaman, S.; and Chang, S.-F. 2019. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6. IEEE.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, 657–672.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.