ExS-GAN: Synthesizing Anti-Forensics Images via Extra Supervised GAN

Feng Ding, Zhangyi Shen, Guopu Zhu, *Senior Member, IEEE*, Sam Kwong, *Fellow, IEEE*,
Yicong Zhou, *Senior Member, IEEE*, and Siwei Lyu, *Fellow, IEEE*

***Abstract*—So far, researchers have proposed many forensics tools to protect the authenticity and integrity of digital information. However, with the explosive development of machine learning, existing forensics tools may compromise against new attacks anytime. Hence, it is always necessary to investigate anti-forensics to expose the vulnerabilities of forensics tools. It is beneficial for forensics researchers to develop new tools as countermeasures. To date, one of the potential threats is the generative adversarial networks (GANs), which could be employed for fabricating or forging falsified data to attack forensics detectors. In this article, we investigate the anti-forensics performance of GANs by proposing a novel model, the ExS-GAN, which features an extra supervision system. After training, the proposed model could launch anti-forensics attacks on various manipulated images. Evaluated by experiments, the proposed method could achieve high anti-forensics performance while preserving satisfying image quality. We also justify the proposed extra supervision via an ablation study.**

***Index Terms*—Anti-forensics, digital forensics, generative adversarial network (GAN), machine learning.**

I. Introduction

WHERE there is sunshine, there is also shadow. For images, there are many manipulations that can be used to attack images in different ways; nevertheless, there are also numerous forensics tools designed to defend images from all possible attacks [1]. Many digital forensics researchers are dedicated to creating algorithms [2], [3], [4] to secure images as reliable channels for communication. In the past, researchers have built mathematical models to trace the alteration of image statistics. In addition, many forensics tools have been developed based on designing handcrafted features to be classified by linear classifiers for a variety of forensics purposes, such as identifying source devices [5], [6], detecting manipulations [7], [8], [9], and exposing forgeries [10], [11].

In recent years, deep learning has made colossal progress. As the most well-known neural network, convolutional neural networks (CNNs) [12], [13], [14], [15] and recurrent neural networks (RNNs) [16], [17] are widely applied in various research fields to analyze data in different forms. In image forensics, it is quite common to adopt CNNs as classifiers to perform detection tasks [18], [19]. The feedforward structure and learning ability enabled by backward propagation make neural networks ideal forensic detectors. CNNs can learn high-dimensional features that cannot be comprehended by the human brain. These features are highly efficient for detection. It has been shown in many publications that well-trained CNN models achieve remarkable detection performance against various image editing manipulations and thoroughly outperform traditional methods [20], [21], [22], [23]. To the best of our knowledge, nearly all possible image editing manipulations can be precisely identified by deep neural networks with properly labeled training [24].

Whereas deep learning has been justified frequently as the perfect image forensics tool, with the more sophisticated architectures developed in recent years, new challenges have also appeared. Unlike the most commonly proposed models focusing on classification, generative adversarial networks (GANs) [25] are designed for creation. GANs are composed of multiple neural networks. A typical GAN model consists of two neural networks: one network functions as a discriminator, and the other network serves as a generator. Both the discriminator and generator simultaneously learn during training to enhance their designated ability to compete with each other in a game. In most cases, after training, GANs are capable of generating images that are similar to the input samples. Note

Manuscript received 17 April 2022; revised 20 July 2022; accepted 18 September 2022. Date of publication 20 October 2022; date of current version 17 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62262041, Grant 62172402, and Grant 61872350; in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant GK229909299001-019; in part by the Jiangxi Double Thousand Plan under Grant JXSQ201901075; in part by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA); in part by the Hong Kong RGC General Research Fund (GRF) under Grant 9042816 (CityU 11209819) and Grant 9042958 (CityU 11203820); in part by the Science and Technology Development Fund, Macau, under Grant 0049/2022/A1; in part by the University of Macau under Grant MYRG2022-00072-FST; in part by the Fundamental Research Funds for the Central Universities under Grant FRFCU5710011322; and in part by the National Science Foundation under Grant 2103450. This article was recommended by Associate Editor R. Yang. *(Corresponding author: Guopu Zhu.)*

Feng Ding is with the School of Software, Nanchang University, Nanchang 330031, Jiangxi, China (e-mail: fengding@ncu.edu.cn).

Zhangyi Shen is with the School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China (e-mail: shenzhangyi@hdu.edu.cn).

Guopu Zhu is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: guopu.zhu@hit.edu.cn).

Sam Kwong is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China, and also with the Department of Computer Science, City University of Hong Kong Shenzhen Research Institute, Shenzhen 51800, China (e-mail: cssamk@cityu.edu.hk).

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau, China (e-mail: yicongzhou@um.edu.mo).

Siwei Lyu is with the Department of Computer Science, State University of New York at Albany, Albany, NY 12222 USA (e-mail: slyu@albany.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2022.3210294.

Digital Object Identifier 10.1109/TCYB.2022.3210294

that "similar" here can be measured in many different ways. For instance, it could be objects of a homogeneous category, the same species, shapes with identical textures and colors, and analogous styles.

Because the images are generated by GANs without any natural information, they could be used by attackers to deliberately deliver false information. Usually, these images can easily deceive the human eyes. In addition, it is impossible for humans to process tremendous amounts of images. Thus, we rely on forensics algorithms to verify the authenticity of images. If these GAN-synthesized images are capable of defeating existing forensics tools, they may become huge threats to our community. In addition, most anti-forensics algorithms proposed in the past rely on specialists with the expertise to build corresponding anti-forensics models for different manipulations. However, unlike traditional approaches, this process is now significantly simplified in that ordinary people without any professional training can easily build their own attacks based on GANs by collecting the proper data. This makes GANs more dangerous than any anti-forensics methods previous.

Therefore, in this article, we would like to investigate the anti-forensics performance [26] of the GAN model to enlighten research on image forensics [27]. We propose our ExS-GAN model as a universal anti-forensics tool that features the extra supervision. The proposed GAN model can generate images that are capable of subtly hiding fingerprints of a variety of common image editing manipulations without altering the original image contents. Generally, these fingerprints are widely employed by forensics detectors to identify manipulations. By removing these fingerprints, the generated images are assumed to impede the forensics tools to make incorrect judgments.

To summarize the above, the main contributions of this article are as follows.

1) A GAN model is proposed as an anti-forensics tool for various common image editing manipulations to investigate the anti-forensics performance of GANs. The main novelty of the model is the enhanced supervision system. Justified by an ablation study, it can boost the anti-forensics performance for the proposed model.
2) Alternative GAN structures and generative networks are considered and studied to refine the proposed GAN model for achieving significant anti-forensics performance with ease of sacrifice for image quality. Evaluated by experiments, the proposed ExS-GAN could outperform the state-of-the-art method for counter forensics.

The remainder of this article is organized as follows. In Section II, we briefly introduce GANs. In Section III, we introduce several novel works on anti-forensics. The proposed model is described in Section IV. The assessment of the proposed model based on experiments is discussed in Section V. Finally, the conclusion is presented.

II. Generative Adversarial Networks

A GAN is a concept defined by Goodfellow *et al.* [25]. It is actually a class of machine learning systems consisting of generative networks and discriminative networks. In this system, given training samples, both the generative network and discriminative network are trained simultaneously for different purposes. The generative network generates new data to be evaluated by a discriminative network. The discriminative network is trained to discriminate the synthetic data from the training samples. Meanwhile, the generator learns from the discrimination procedure to generate new data with closer statistics to the training samples to fool the discriminator. Generally, this system can be regarded as a competition between two networks. Through backpropagation during training, both networks are optimized and become more intelligent. Typically, the new data synthesized by generators have attracted the most attention from researchers. They have been widely studied and applied in a variety of areas [28], [29], [30].

Many GAN models have been proposed, being driven by different motivations. Among all the GANs, the conditional GAN (cGAN) is a special category with fully supervised learning that concentrates on minimizing the process of setting the generating process conditions. Unlike many other GANs that only focus on generating vivid images, cGANs can generate vivid images with different characteristics. With proper supervision, cGANs are capable of delivering desired new data to satisfy different purposes. Because of this feature, cGAN is the preferred option to translate images from one style to another [31], [32].

In image processing, many image editing manipulations leave unique traces in images producing particular visual effects. These visual effects can also be regarded as image styles. For example, sharpening can enhance the contrast of edges, which leads to sharp silhouettes as a visual feature. Taking a step further, theoretically, images without certain manipulation can also be considered as a style, that is, an untouched style. Therefore, given that cGANs can translate image styles, they are also assumed to be capable of transforming images from other styles into untouched styles. In other words, the fingerprints left by a variety of manipulations can be removed by cGANs such that the image may appear untouched by these manipulations. In image forensics, such operations could lead to the possibility of manipulations being applied to images becoming more difficult to detect. Thus, cGANs may serve as anti-forensics tools.

III. Anti-Forensics

Within digital forensics, anti-forensics, also known as counterforensics, is a branch that has raised much debate and discussion. Anti-forensics is a set of techniques that are used to combat digital forensics. Generally, anti-forensics tools are designed for malicious purposes. However, for scientific research, anti-forensics tools can also serve as countermeasures to forensics algorithms. By exposing the weaknesses of current forensics tools, anti-forensics helps researchers further develop powerful forensics tools for the future to guarantee that the collected data are authentic and dependable.

Anti-forensics falls into several subcategories, such as data hiding, artifact wiping, and trail obfuscation. In this article, as discussed in the previous section, we examine the anti-forensics performance of GANs from the perspective of artifact wiping to deceive forensics tools. In particular, we focus on wiping the traces of common signal processing manipulations that are widely studied in conventional image forensics works.

Although many anti-forensics works on erasing manipulated fingerprints have been reported, most of them focus on tackling anti-forensics problems for a single manipulation, that is, either JPEG compression or median filtering. The two manipulations are ideal counter-forensics targets because limiting data size and denoising are fundamental needs for image processing.

Among all the JPEG anti-forensics works conducted by different groups, it is recognized by most forensics researchers that Stamm *et al.* made the considerable contribution to this topic to date. They initiated related research for JPEG compression anti-forensics [33], [34]. Their works were followed by other groups [35], [36], where a variety of JPEG anti-forensics models were later proposed to enhance the JPEG anti-forensics performance under different circumstances. The majority of related work hides the compression trails by building anti-forensics models to tamper with the image statistics. A similar phenomenon also occurs in the history of median filter anti-forensics. After Fontani and Barni started counterforensics research on the median filter in 2012 [37], this topic has made considerable progress, with more median filter anti-forensics models being proposed [38], [39]. Most of them also achieve the anti-forensics effect by attacking the image statistics.

In addition to building anti-forensics models as described above, a few anti-forensics works based on adversarial networks have been proposed in recent years. By employing GANs for anti-forensics, researchers no longer need to analyze the image statistics or tamper with any specific fingerprints because GANs are capable of self-learning to achieve anti-forensics objectives automatically [40]. With supervised training, GANs can synthesize images that preserve exactly the same content as the attacked images. In addition, the manipulated fingerprints, once employed as clues for forensics detectors, are removed in synthesized images, that is, the GANs can serve as anti-forensics tools. Kim *et al.* [42] employed GANs to restore images [41] processed by median filters. The images reconstructed via their GAN can outperform images processed by other anti-forensics methods with higher undetectability, as reported in their paper. Luo *et al.* [43] proposed a GAN model that can reach acceptable undetectability with eased image quality degradation. Although there is no doubt that both works are brilliant efforts involving new methods of GANs, they are similar to other anti-forensics works that have contributed to improving anti-forensics performance for single manipulation.

For image manipulation anti-forensics, note that, other than the undetectability, the image quality is the other metric for evaluation. In most cases, the image quality must be sacrificed to enhance the undetectability. Thus, most anti-forensics works have made great efforts to achieve a tradeoff between undetectability and image quality. This is extremely important for our research, as a notional successful anti-forensics attack should be capable of deceiving forensic detectors and humans simultaneously. In summary, in this article, we concentrate on investigating the anti-forensics performance of GANs by proposing a GAN model that serves as an anti-forensics tool that targets multiple common image manipulations. The proposed method can remove traces of different manipulations while avoiding distortions or artifacts being introduced to the images. The entire procedure is depicted in Fig. 1.

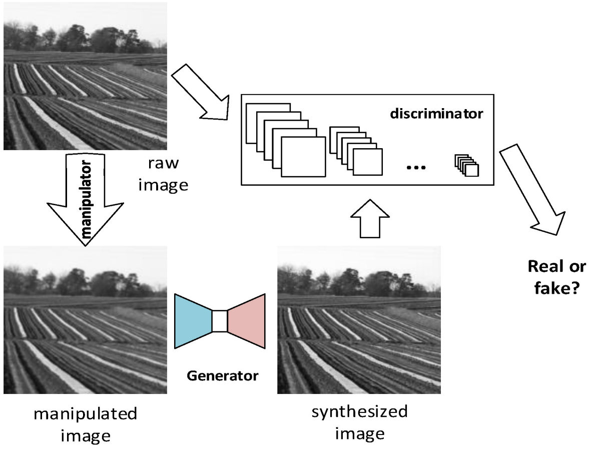

Fig. 1. Training GAN models to remove traces left by image editing manipulations.

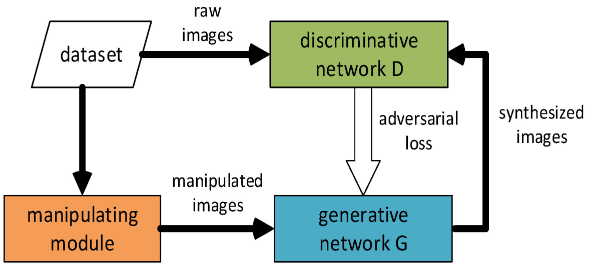

Fig. 2. Architecture of the prototype GAN model.

IV. Proposed Method

A. Prototype GAN Model

As mentioned above, both the discriminator and generator are important components in GANs. One of the most typical and fundamental GAN models, the deep convolutional GAN (DCGAN), consists of a single classification network as the discriminator and a single generative network as the generator. This GAN generates new images from random noise. However, the image content and texture generated in the DCGAN cannot be well supervised. Thus, in most application scenarios of GANs, the structures have to be refined and optimized. Thus, to function as an anti-forensics tool, we employ the prototype GAN model as illustrated in Fig. 2.

The only input to the GAN model is the untouched image dataset. Since our objective is to remove the manipulated fingerprints while keeping the image content untouched, we would like to have the image content be generated under strict supervision. This objective can be satisfied by inputting paired images to enhance the supervision of the image content. Hence, there must be two parallel input channels for feeding image pairs into the generator as source signals and target signals.

As observed from Fig. 1, the manipulating module in our model contains optional image manipulations that could be chosen to convert the untouched images to manipulated images. The manipulated images are source signals for the generator to synthesize new images. The output of the generator can be employed in association with untouched images to train the discriminator. In other words, the discriminator is trained to discriminate the untouched images and synthesized images. Note that for anti-forensics, our objective is to produce images that can deceive forensic detectors. Hence, we expect that the images synthesized from the generator will be close to their original manipulation-free version and that the discriminator fails to classify them. For this reason, the untouched images here can be regarded as the target signals. As a result of this arrangement, the source signals and target signals share the same image content, and the generation procedure for the image content is fully supervised. This is an advantage for the proposed GAN structure in that it ensures that the synthesized images from the generator preserve the same content. Thus, we only need to focus on transferring the image style for anti-forensics purposes.

G, the generator, is a vital part of the GAN model for generating images of the untouched style. Unlike most other GANs, the input to the generator is manipulated images. The anti-forensics effect can be achieved by removing the traces of manipulations in manipulated images. Behind the generator, the discriminator *D* is introduced to act as a supervisor. The weights learned in the discriminator are assumed to be backpropagated to the generator during training. Thus, the discriminator is arranged to concatenate to the generator. We would like to investigate the architecture of the generator and discriminator in detail later.

The architecture and networks introduced above are those of the proposed prototype GAN model. The loss of *G* and *D* would gradually stabilize after training with sufficient iterations. Subsequently, the *G* will be capable of synthesizing images with undetectability.

B. Extra Supervision and Loss Function

As discussed above, our prototype GAN model is only supervised by a single discriminator to distinguish untouched images from synthesized images. Except for the synchronized image content, this strategy restrains the generated images from only one aspect, that is, the synthesized images should be close to the untouched images in terms of high-level patterns and statistics. Thus, the loss function can be designed with the following formula:

$$\mathcal{L}(G, D) = E\left[\log D(I, G(I_m, n))\right] \tag{1}$$

where I_m is the manipulated images that are employed as inputs to the generative network, I is the untouched images, which also represent target signals that supervise the generation procedure, and n is the noise that should also be fed to the generator. Note that here for our anti-forensics purpose, we define the n as the inverse residual of the manipulation fingerprints. If it is applied to the manipulated images, the synthesized images could be images without any manipulated fingerprints. n can be represented as

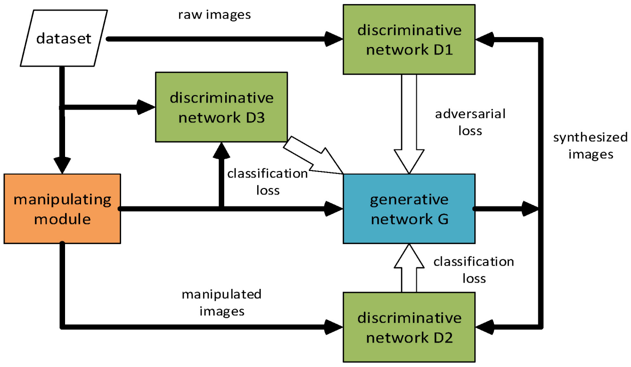

Fig. 3. Proposed GAN structure with Ex-S.

$$n = I - I_m. \tag{2}$$

Consequently, the synthesized image I_g is

$$I_g = G(I_m, n). \tag{3}$$

Since our objective is to introduce noise to the manipulated images to reconstruct the images I_g that are close to untouched images I, the loss of the generator must be minimized, while the loss of the discriminator should be maximized. Thus, the entire procedure of the GAN model can be described as

$$\begin{aligned} T &= \arg\min_{G}\max_{D}\{\mathcal{L}(G, D)\} \\ &= \arg\min_{G}\max_{D}\left\{E\left[\log D\left(I, I_g\right)\right]\right\}. \end{aligned} \tag{4}$$

Although the prototype model may satisfy the fundamental requirements to perform as a cGAN to remove manipulated fingerprints, we believe that the anti-forensics performance of the model can be improved if proper enhanced supervision can be conducted. Besides, it has been proved that the generating process could be more accurate if a more powerful supervision is applied in GANs. Therefore, we designed a refined supervision system to be associated with the prototype model to boost the performance. The proposed refined model is depicted in Fig. 3.

As seen in Figs. 2 and 3, the major difference is that two new discriminators, D2 and D3, are introduced in the refined structure in Fig. 3. These two discriminators serve as Extra-Supervisions (Ex-S) for the prototype. All three discriminators are equally weighted in our proposed model.

D2 is assumed to be trained to classify the output of the generative network from the input. Through backpropagation, the learned weights are transferred back to the generator. This strategy guarantees that the synthesized images should be far from the manipulated images in terms of high-level patterns and statistics while preserving the content.

Similar to many forensics detectors based on CNN, D3 is responsible for discriminating manipulated images from untouched images during training. The weights learned by D3 can also be used to update the parameters in the generator. This can enhance G and allow it to make wiser choices to avoid inputting features learned in D3 into synthesized images. As a result, the synthesized images may become more difficult to be distinguished from untouched images. Modules with similar functions to D3 can also be found in other works. It has been proven to have a positive impact on generating desired signals [44].

During training, all the discriminators are trained simultaneously, along with the generator. Nevertheless, we expect different convergence performances from each discriminator. As discussed in the prior discussion, the generative network deliberately deceives D1 to prevent it from converging. In contrast, we adopt D2 to monitor the generated images to be statistically further from the manipulated ones. D3 is required to converge with high classification performance to enhance the generation from different aspects. All three discriminators are equally weighted. Therefore, the loss function for this refined model can be defined as

$$\begin{aligned}\mathcal{L}_r(G, D_1, D_2, D_3) = {} & E\big[\log D_1\big(I, I_g\big)\big] \\ & + E\big[1 - \log D_2\big(I_m, I_g\big)\big] \\ & + E\big[\log D_3(I, I_m)\big]. \end{aligned} \tag{5}$$

In addition, as learned from recent reports [44], [45], [46], [47] on cGANs, we also deploy an $\mathcal{L}_1$ loss to enhance the performance of the generative network. This strategy has been proven to be capable of improving the quality of synthesized images. This loss can be described as

$$\mathcal{L}_1(G) = E_{I,I_m,I_g}\big[\big\|I_g - G(I_m, n)\big\|_1\big]. \tag{6}$$

Then, we have the complete form of the loss function for the refined model as

$$\mathcal{L}'(G, D_1, D_2, D_3) = \mathcal{L}_r(G, D_1, D_2, D_3) + \lambda\mathcal{L}_1(G). \tag{7}$$

This could lead to our final goal of the entire model which is defined as

$$T' = \arg \min_{(G,D_3)} \max_{(D_1,D_2)} \big\{\mathcal{L}'(G, D_1, D_2, D_3)\big\}. \tag{8}$$

C. Architectures of Discriminator and Generator

With the GAN architectures and loss functions studied, the remaining task is to discuss the architectures of the discriminative network and generative networks.

To the best of our knowledge, many proposed CNNs are serving as detectors in digital forensics. Although they have been employed to solve different problems successfully, most methods use simple, single lanes of feedforward structures, which can be considered homogeneous to AlexNet and LeNet. A similar arrangement can also be found for many discriminators in GANs. Thus, considering that the difficulty of discrimination tasks is not high, we also employ a simple structure of this type for all the discriminators in our proposed models. The architecture of our discriminators is depicted in Fig. 4.

The generator is the decisive component and can directly impact the anti-forensics performance. Therefore, we put greater effort into investigating generative networks with different architectures. Since the input signals to our generator are images and since the outputs are also images of uniform size, one of the most typical generative structures that suit this circumstance is the end-to-end model with a downsampling network and an upsampling network. In this model, the input images are first downsampled into feature vectors in the downsampling network, and then, the feature vectors are reconstructed as images by the upsampling network. This structure is illustrated in Fig. 4. Multiple convolutional layers are arranged in series to function as the downsampling network. Eventually, after processing by these convolutional layers, the images can be downsampled into feature vectors.

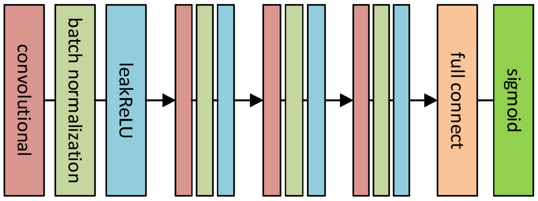

Fig. 4. Structure of discriminative network. The kernel size in all convolutional layers is 5 × 5, the stride is 2. The number of filters in the first convolutional layer is 64, it is always doubled in the next convolutional layer of the network. The slope for leakReLU is 0.2.

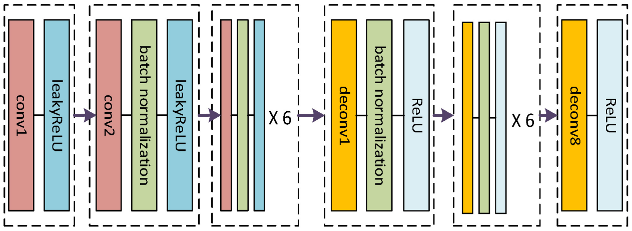

Fig. 5. Structure of basic generative network. For all convolutional and deconvolutional layers, the kernel size is fixed to 4, the stride is 2. The number of filters is n, $n = 64$ for conv1 and deconv8, $n = 128$ for conv2 and deconv7, $n = 256$ for conv3 and deconv6, and $n = 512$ for all the others.

After downsampling, we employ an upsampling network to reconstruct the images from the feature vectors. The upsampling network consists of multiple deconvolutional layers in series. The upsampling network is symmetric to the downsampling network to ensure a consistent image size. This is the simplest and most fundamental structure that can be considered the basic generator for our GANs.

The upsampling here is used to restore the image to be of equal size to the input image. In most cases, deconvolution is the preferred upsampling method over linear interpolation in GANs. From the literature, it can be assumed that the input images should be downsampled by a series of convolutional layers to produce feature vectors as output. The output of the downsampling network should be the input for the upsampling network to have the image reconstructed. Thus, the upsampling network, consisting of multiple deconvolutional layers in series, should be connected behind the downsampling network to rebuild the images. This structure is shown in Fig. 5. It is the simplest and most fundamental structure for building our desired generative network.

In addition to the structure of the basic generator, there are also other advanced end-to-end architectures. These architectures can be considered refined versions of the basic model. They can serve as optional structures for our generator. Here,

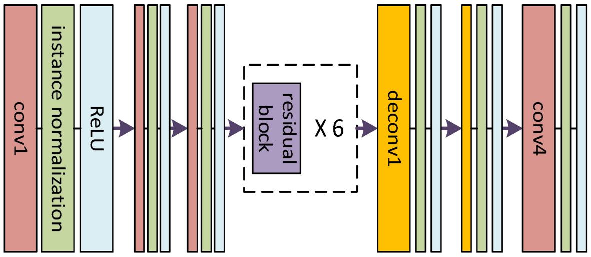

Fig. 6. Generative network of T-Net. The kernel size for conv1 and conv4 is 7, and for other layers and residual blocks is 3. The stride for conv2, conv3, deconv1, and deconv2 is 2, and for other layers and residual blocks is 1. The number of filters for conv1 and deconv2 is 64, for conv2 and deconv1 is 128, for conv4 is 3, and for conv3 and residual blocks is 256.

we introduce two types of refined versions and evaluate them later with experiments. The first optional adjustment is inserting a transformation network between the upsampling and downsampling networks [48]. The transformation network is composed of multiple residual blocks which can be regarded as convolutional layers. It has been proved to be efficient for transforming images that can be employed as a generator in GANs. We name this structure T-Net in this article. The second optional adjustment is inspired by U-Net, which was proposed in Ronneberger *et al.*'s work [49]. It establishes channels for the symmetrically located layers in upsampling and downsampling networks to enable one-way communication in corresponding layers from the downsampling network to the upsampling network. As a result of their strategy, the deconvolutional layers in the upsampling network can reconstruct the images with the assistance of corresponding convolutional layers to improve the accuracy of the details in the synthesized images. Consequently, the synthesized images generated via this model are expected to be of higher quality than the other methods. The architectures of these advanced generative models are illustrated in Figs. 6 and 7.

V. Experimental Results and Discussion

In order to serve as a general anti-forensics tool, the evaluation for the proposed ExS-GAN model is based on four datasets to increase the diversity of data. The BOSS image dataset contains 10 000 grayscale images of size 512 × 512. RAISE is a relatively new image dataset released in 2015. It consists of 8156 high-resolution images of size 4288 × 2848 or 4928 × 3264 and is intended for research on digital forensics. UCID includes 1338 uncompressed color images of size 384 × 512. The NRCS dataset consists of 1000 grayscale images of size 768 × 512. The BOSS and RAISE datasets are our training datasets, while the UCID and NRCS datasets are the validation set. All images are randomly cropped to a uniform size of 256 × 256. In addition, all color images are converted to grayscale images for our experiments. All experiments are simulated with TensorFlow 1.1.0 and CUDA 8.0. The generator along with all discriminators in the GAN models are trained simultaneously with the Adam optimizer of learning rate 0.0002 for 50 epochs. We manually terminate training if the losses of generators and discriminators tended to be stable.

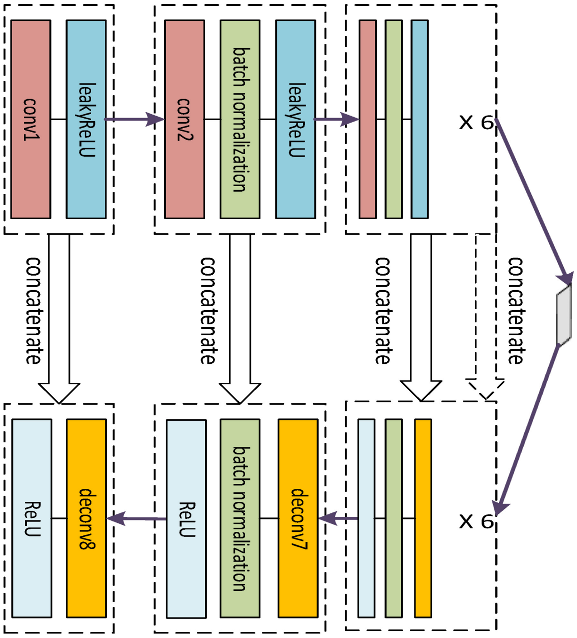

Fig. 7. Generative network of U-Net. For all convolutional and deconvolutional layers, the kernel size is fixed to 4, the stride is 2. The number of filters is n, $n = 64$ for conv1 and deconv8, $n = 128$ for conv2 and deconv7, $n = 256$ for conv3 and deconv6, and $n = 512$ for all the others.

TABLE I
Employed Manipulations and Applied Parameters

Editing manipulations (EM)	Parameters
Gaussian filtering (GF)	3×3 window size, $\sigma = 0.8$
Median filtering (MF)	3×3 window size
Average filtering (AF)	3×3 window size
USM sharpening (US)	$\sigma = 1, \lambda = 1$
Gaussian noising (AGN)	$\sigma = 0.01$
JPEG compression (JC)	$Q = 50$

In our experiment, as introduced in Section III, we would like to assess the anti-forensics performance of the GAN models with some common signal processing manipulations. Hence, the following manipulations were selected in the manipulation module for the GAN models: Gaussian filtering, median filtering, average filtering, USM sharpening, adding Gaussian noise, and JPEG compression. The parameters applied for each manipulation in our experiments can be found in Table I.

A. Ablation Study of ExS and Generator Structure

First, we study the structures of the generator and the proposed Ex-S enhanced supervision system. To achieve this goal, we employ the prototype GAN model α with different generator structures introduced in Section IV: the encoder–decoder E, the U-Net U, and the Transformation-Net T as options for assessment. In addition, we conducted several experiments in ablation studies of Ex-S. In this case, individual D2, D3, and Ex-S are tested along with the prototype model α. The generator in α is fixed with E to ensure the ablation study is professional.

TABLE II
CLASSIFICATION PERFORMANCE OF TRAINED CCNN FOR ABLATION STUDY

EM	Detection accuracy
GF	99.12%
MF	99.67%
AF	99.52%
US	99.25%
AGN	99.93%
JC	99.34%

TABLE III
PRECISION RESULTS OF ABLATION STUDY FOR MODELS WITH DIFFERENT GENERATORS AND SUPERVISION MODULES

EM	α + E	α + U	α + T	α + D2	α + D3	α+ExS
GF	13.64%	13.59%	14.38%	**9.13%**	10.21%	9.16%
MF	1.93%	1.66%	3.47%	1.02%	1.58%	**0.29%**
AF	0.25%	0.57%	0.96%	**0.01%**	0.06%	**0.01%**
US	22.33%	20.51%	26.21%	15.46%	17.01%	**12.89%**
AGN	0.07%	0.01%	0.19%	0.03%	0.01%	**0.00%**
JC	19.51%	19.63%	23.70%	16.32%	17.30%	**13.17%**

To evaluate the undetectability, it is necessary to employ a forensics tool as a benchmark. In this experiment, we choose the constrained CNN (CCNN) [24] to play this role. Although many famous classifiers can be employed as forensics detectors, the CCNN [24] proposed in 2018 is generally considered the state-of-the-art detector in digital forensics. The reported results outperform almost all digital forensics tools proposed in past years. In addition, the CCNN covers a wide range of image editing manipulations and can also serve as a universal tool. Given all the advantages, our ideal evaluation tool should be the CCNN. Therefore, several CCNNs are trained against the manipulations listed in Table I. The observed classification performance reported in Table II demonstrates that it is an effective and reliable tool as a benchmark.

The anti-forensics images synthesized by GANs are then predicted by corresponding detectors. For each manipulation, the ratio of synthesized images detected as manipulated images is listed in Table III.

From the table, we can see that each extra discriminator boosts the undetectability of the GAN model. Above that, it can also be observed that the model with the joint supervision from both extra discriminators achieves the best anti-forensics performance of all the models. Hence, ExS is the ideal strategy to boost the anti-forensics performance. Although the structure of the generator has a certain impact on the undetectability of GAN models, this effect is not prominent that can be ignored.

Then, we examine the qualities of the synthesized images via three criteria: PSNR, SSIM, and VIF. PSNR and SSIM are two popular criteria widely applied for image quality assessments. VIF is an image quality assessment method proposed in 2006 [50]. Unlike the other methods, VIF evaluates the image quality in a perceptually consistent manner that matches the human vision system. The quality assessment can be found in Tables IV–VI.

TABLE IV
AVERAGE PSNR FOR IMAGES SYNTHESIZED BY DIFFERENT MODELS

EM	α + E	α + U	α + T	α + D2	α + D3	α + ExS	I_m
GF	31.03	**32.72**	25.31	31.07	30.88	30.97	29.89
MF	27.86	**30.62**	25.27	27.99	27.86	27.85	28.35
AF	27.69	**29.31**	23.57	27.72	27.75	27.76	27.68
US	35.13	**36.39**	26.77	35.08	35.10	35.12	35.38
AGN	24.55	**26.90**	16.77	24.60	24.53	24.64	19.87
JC	31.80	**33.77**	23.98	31.65	31.71	31.74	33.09

TABLE V
AVERAGE SSIM FOR IMAGES SYNTHESIZED BY DIFFERENT MODELS

EM	α + E	α + U	α + T	α + D2	α + D3	α + ExS	I_m
GF	0.913	**0.941**	0.850	0.916	0.918	0.916	0.883
MF	0.827	**0.903**	0.726	0.828	0.825	0.826	0.836
AF	0.887	**0.922**	0.715	0.900	0.902	0.892	0.817
US	0.978	**0.988**	0.852	0.979	0.976	0.981	0.975
AGN	0.617	**0.731**	0.387	0.638	0.620	0.652	0.389
JC	0.919	**0.933**	0.735	0.922	0.920	0.920	0.923

TABLE VI
AVERAGE VIF FOR IMAGES SYNTHESIZED BY DIFFERENT MODELS

EM	α + E	α + U	α + T	α + D2	α + D3	α + ExS	I_m
GF	0.809	**0.861**	0.576	0.806	0.801	0.800	0.612
MF	0.493	**0.547**	0.408	0.495	0.492	0.492	0.526
AF	0.680	**0.759**	0.406	0.667	0.666	0.671	0.552
US	0.898	0.945	0.788	0.895	0.908	0.902	**0.947**
AGN	0.182	**0.278**	0.075	0.191	0.180	0.193	0.247
JC	0.670	**0.714**	0.437	0.675	0.681	0.668	0.712

The quality assessments demonstrate that the structure of the generator has a strong impact on the quality of the synthesized images. In the meanwhile, the supervisions have a quite limited effect on image quality. Therefore, after all these evaluations, we pick ExS as the supervision system to pursue higher undetectability. On the other hand, the U-Net is chosen as the generator structure to improve the quality of synthesized images. The proposed model is determined to be the GAN model illustrated in Fig. 3 with U-Nets as generators.

B. *Evaluation of the Proposed ExS-GAN Model*

Since the structure of the proposed ExS-GAN model is determined via the justifications above, we thoroughly evaluate the model by conducting more experiments.

With the trained proposed models, image sets can be synthesized. Since we also want to investigate the effect of different patch sizes, the images are generated of size 256 × 256, 128 × 128, and 64 × 64. For the quality assessments of these images, the average value of PSNR, SSIM, and VIF is reported in Table VII.

The results of the quality assessment demonstrate that the quality of the synthesized image is quite high. It is well known in traditional anti-forensics that attacking images produces distortions. Consequently, the image quality is always sacrificed to enhance the undetectability. However, this rule does not apply to our proposed method. Surprisingly, in contrast, summarized from the observed results, the synthesized images tend to have higher quality than the attacked images I_m. This can be considered as a tremendous advantage for anti-forensics based

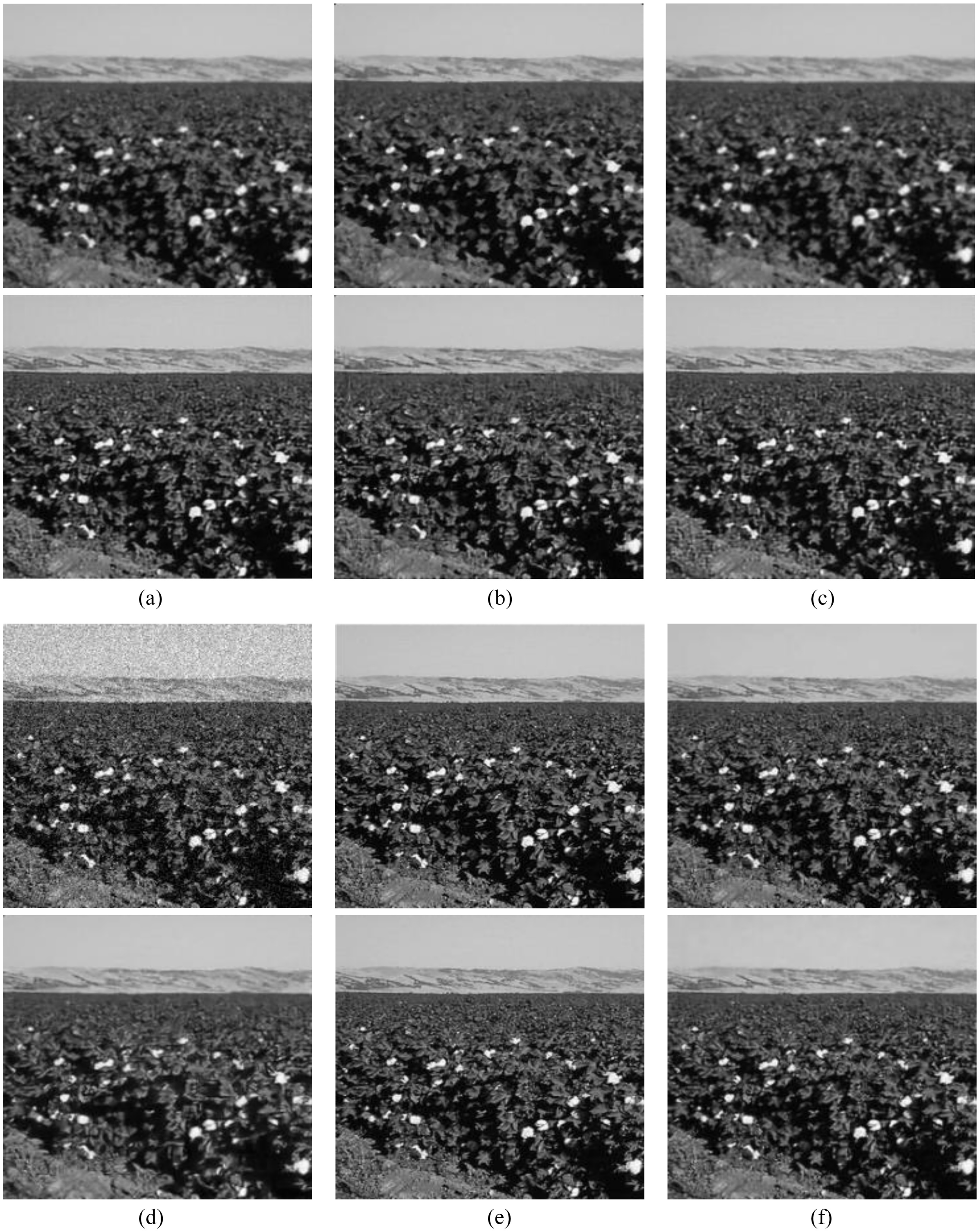

Fig. 8. Sample images. The images on the top are the manipulated images, the ones on the bottom are the synthesized images. (a) Gaussian filtering, (b) median filtering, (c) average filtering, (d) Gaussian noising, (e) USM sharpening, and (f) JPEG compression.

on GANs. In addition, it can also be observed that the patch size has little impact on the quality of the synthesized images.

Afterward, we conducted several experiments to investigate the undetectability of the proposed model. In most conventional cases, for anti-forensics, an attack is successful if an attacked image can be falsely determined as untouched image by a binary classifier. Therefore, since there were only two categories for classification, and only the synthesized images are tested, the anti-forensics performance can be regarded as higher if less attacked images are classified as manipulated images.

Here, we employ more forensics detectors to fulfill the task. Along with the CCNN introduced in the above section, the VGG16 and the rich model are also chosen for validation. VGG16 is a famous CNN model for classification and detection [51]. The rich model is a non-CNN forensics tool proposed in 2012. Although the original purpose of the rich model was steganalysis, it has been proven by many researchers to be a successful algorithm in revealing manipulations in images [52]. We employ it along with ensemble learning for multiclass classification. Both the VGG16 and rich

TABLE VII
QUALITY ASSESSMENT FOR IMAGE OF DIFFERENT SIZES SYNTHESIZED VIA PROPOSED ExS-GAN

EM	256 × 256			128 × 128			64 × 64		
	PSNR	SSIM	VIF	PSNR	SSIM	VIF	PSNR	SSIM	VIF
GF	32.77	0.938	0.864	32.67	0.931	0.865	32.66	0.933	0.865
MF	30.71	0.898	0.559	30.70	0.900	0.562	30.67	0.903	0.560
AF	29.33	0.919	0.763	29.36	0.917	0.760	29.37	0.915	0.760
US	36.41	0.989	0.948	36.38	0.991	0.948	36.40	0.990	0.947
AGN	26.92	0.742	0.290	26.73	0.726	0.288	26.34	0.721	0.281
JC	33.72	0.937	0.715	33.79	0.935	0.717	33.80	0.935	0.714

TABLE VIII
ANTIFORENSICS ASSESSMENT FOR IMAGE OF DIFFERENT SIZES SYNTHESIZED VIA PROPOSED ExS-GAN

EM	256 × 256			128 × 128			64 × 64		
	Rich model	VGG16	CCNN	Rich model	VGG16	CCNN	Rich model	VGG16	CCNN
GF	3.50%	7.72%	5.34%	3.03%	7.17%	5.45%	1.59%	6.85%	6.09%
MF	0.01%	0.97%	0.35%	0.00%	0.23%	0.52%	0.00%	0.68%	0.39%
AF	0.55%	2.15%	0.82%	0.20%	2.23%	1.67%	0.00%	1.86%	1.34%
US	22.25%	23.68%	14.03%	17.52%	25.16%	18.59%	12.77%	24.39%	17.21%
AGN	0.00%	0.01%	0.01%	0.00%	0.03%	0.01%	0.00%	0.05%	0.07%
JC	7.21%	20.89%	9.27%	4.19%	23.26%	12.79%	3.45%	22.63%	11.52%

model could easily reach excellent detection accuracy over 99% toward all manipulations in Table II after training. All the GAN-generated images are tested by the three detectors. The detection rate representing the ratios of images that are detected as "manipulated" on images of different patch sizes are reported in Table VIII.

As observed from the table, most synthesized images were falsely judged as untouched images. We can also notice that the impact of patch sizes is also not prominent. Hence, the proposed enhanced supervision system along with the GAN structure has been justified to be a successful general anti-forensics tool.

After the evaluation based on the conventional binary classification detector, we conduct further experiments to investigate the anti-forensics performance of these GAN-generated images toward universal forensics tools. For universal detectors, the most important feature is that they can be employed against a wide range of attacks. Thus, the three forensics detectors were trained with untouched images I_o and all kinds of manipulated images I_m to be powerful multiclass classifiers that can achieve overall classification accuracies over 90%. Then, the GAN-generated images were classified by these general forensics detectors. Along with the overall detection accuracy reported in Table IX, the confusion matrices based on the classification of image patch size 256 × 256 for each detector are shown in Tables X–XII.

It can be observed that most synthesized images are falsely detected as original images. However, unlike the binary classifications, for multiclass classification, the synthesized images may also be falsely predicted as images attacked by other manipulations. For example, the sharpened images could be incorrectly labeled as compressed images after being processed by the GAN model. Nevertheless, such results can also be regarded as successful anti-forensics attacks because detectors are disrupted.

TABLE IX
PREDICTION PRECISION OF DIFFERENT MULTICLASS CLASSIFIERS ON IMAGES OF DIFFERENT SIZES

Classifiers	256 × 256	128 × 128	64 × 64
Rich model	6.20%	3.85%	2.52%
VGG16	8.07%	8.76%	8.68%
CCNN	4.18%	4.39%	4.15%

Summarizing the above experiments, the proposed model is found to be a reliable anti-forensics tool that can produce images with high quality while maintaining satisfying undetectability.

C. Comparisons With Prior Arts

Recall that there are existing anti-forensics approaches for median filtering and JPEG compression. In the following experiments, we compare the performance of the proposed model with these prior arts.

For median filtering, Kim *et al.*‘s method [42] is reported as a state-of-the-art median anti-forensics model. Their work is also based on supervised training of the GAN model. Therefore, their model is an ideal approach to be compared with. Validation images are generated with their model from the identical dataset as ours to guarantee that the comparison is fair. Here, the CCNN is trained with median filtered images and untouched images as binary classifiers for validation. For comparison, we also considered the effect of different parameters for median filtering. Hence, the comparison results for

TABLE X
CONFUSION MATRIX OF RICH MODEL; PREDICTION (COLUMNS) VERSUS GROUND TRUTH (ROWS)

Manipulations	GF	MF	AF	US	AGN	JC	I_o
GF	1.39%	0.00%	6.28%	31.55%	0.00%	12.64%	48.14%
MF	0.01%	0.01%	0.36%	29.44%	0.00%	4.15%	66.03%
AF	0.00%	0.00%	9.72%	27.66%	0.00%	1.72%	60.90%
US	0.00%	0.00%	0.10%	19.18%	0.01%	6.39%	74.42%
AGN	0.00%	0.00%	0.00%	39.37%	0.00%	5.11%	55.52%
JC	0.83%	0.12%	0.03%	26.32%	0.00%	6.91%	65.79%

TABLE XI
CONFUSION MATRIX OF VGG16; PREDICTION (COLUMNS) VERSUS GROUND TRUTH (ROWS)

Manipulations	GF	MF	AF	US	AGN	JC	I_o
GF	7.66%	0.52%	0.03%	4.15%	0.00%	5.47%	82.17%
MF	1.82%	0.55%	0.27%	4.98%	0.00%	6.25%	86.13%
AF	0.67%	0.58%	1.43%	3.73%	0.00%	4.08%	89.51%
US	0.03%	0.00%	0.01%	20.21%	0.00%	16.32%	63.43%
AGN	0.03%	0.12%	0.01%	29.16%	0.00%	10.75%	59.94%
JC	1.73%	0.03%	0.00%	12.45%	0.00%	18.57%	67.22%

TABLE XII
CONFUSION MATRIX OF CCNN; PREDICTION (COLUMNS) VERSUS GROUND TRUTH (ROWS)

Manipulations	GF	MF	AF	US	AGN	JC	I_o
GF	5.69%	0.07%	0.12%	6.61%	0.00%	6.12%	81.39%
MF	0.03%	0.22%	0.38%	5.96%	0.03%	5.40%	88.00%
AF	1.15%	0.79%	0.03%	6.22%	0.00%	7.78%	84.03%
US	0.12%	0.03%	0.03%	9.69%	0.00%	19.73%	70.40%
AGN	0.01%	0.03%	0.00%	26.63%	0.03%	9.82%	63.48%
JC	1.12%	0.66%	0.01%	20.57%	0.00%	8.33%	69.31%

TABLE XIII
COMPARISON RESULTS FOR ANTIFORENSICS OF MEDIAN FILTERING

Window size	Methods	PSNR	SSIM	VIF	Precision
3 × 3	Kim *et al.*'s	28.45	0.870	0.553	4.07%
	Proposed	**30.71**	**0.898**	**0.559**	**0.35%**
5 × 5	Kim *et al.*'s	24.52	0.741	0.312	4.52%
	Proposed	**26.19**	**0.763**	**0.347**	**0.50%**

TABLE XIV
COMPARISON RESULTS FOR ANTIFORENSICS OF JPEG COMPRESSION

Quality factor	Methods	PSNR	SSIM	VIF	Prescision
30	Stamm *et al.*'s	26.74	0.811	0.598	**0.13%**
	Luo *et al.*'s	30.94	0.901	**0.676**	17.43%
	Proposed	**31.15**	**0.904**	0.670	13.28%
50	Stamm *et al.*'s	27.72	0.840	0.616	**0.27%**
	Luo *et al.*'s	32.90	0.920	0.696	13.21%
	Proposed	**33.72**	**0.937**	**0.715**	9.27%
70	Stamm *et al.*'s	28.55	0.859	0.641	**0.00%**
	Luo *et al.*'s	34.94	**0.958**	0.728	6.55%
	Proposed	**35.47**	0.956	**0.730**	2.26%

window sizes 3 and 5 are displayed in Table XIII with the detection accuracy and image quality. The detection accuracy here is the ratio of synthesized images that are classified as median filtered images.

The experimental results demonstrate that both models can reach quite high anti-forensibility, as the detection performance for both is almost perfect. Under such circumstances, our proposed model can still outperform Kim *et al.*'s method by at least 3%. For qualities of reconstructed images, the images synthesized with our proposed model lead the competition with higher quality. It can be also observed that a larger window size may lead to lower quality of synthesized images. Despite that, there is not any impact on window size that can be observed for the detection accuracy.

We follow the same pipeline to implement comparisons with Stamm *et al.*'s method [34] and Luo *et al.*'s method [43] as JPEG compression anti-forensics models. For JPEG compression, the quality factors of 30 and 70 are also considered for comparison. We still employ the trained CCNN as our validation tool. The detection accuracy is still the ratio of synthesized images that are classified as compressed images. The performance, including detection accuracy and image quality, can be found in Table XIV.

As observed from the experimental results, Stamm *et al.*'s method can achieve the highest undetectability, that almost all compression fingerprints left in compressed images can be removed. However, this merit comes at the price of sacrificing image quality. Albeit the undetectability is relatively low in contrast to [34], the image quality can be satisfactory for the ones synthesized by GANs. For the performance of two GAN models, our proposed model outperforms Luo *et al.*'s method, with slight improvements in both undetectability and image quality for most cases.

VI. Conclusion

In this article, we investigate the capability of GANs to perform as image anti-forensics tools. Discussions are made on this topic after proposing a GAN model as an anti-forensics tool. With extra supervision, the proposed ExS-GAN can synthesize images with high anti-forensics performance. As proved by our experiments, most synthesized images are undetectable by forensic detectors regardless of whether they are based on CNNs. In addition, the images synthesized by GANs are also of higher quality when compared with the traditional anti-forensics approach. Anti-forensics via GANs could be a potentially huge threat to information security. The development of forensic tools with higher robustness should be encouraged against this situation.

Acknowledgment

The authors gratefully acknowledge the contributions of Yinming Luo and Dr. Xiangui Kang for kindly sharing their codes with them for the performance comparison. The views, opinions, and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the National Science Foundation, or the U.S. Government.

References

[1] S. Lyu and H. Farid, "How realistic is photorealistic?" *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 845–850, Feb. 2005.

[2] A. Tuama, F. Comby, and M. Chaumont, "Camera model identification with the use of deep convolutional neural networks," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2016, pp. 1–6.

[3] F. Ding, G. Zhu, J. Yang, J. Xie, and Y.-Q. Shi, "Edge perpendicular binary coding for USM sharpening detection," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 327–331, Mar. 2015.

[4] Y. Hong, S. Kwong, Y. Chang, and Q. Ren, "Consensus unsupervised feature ranking from multiple views," *Pattern Recognit. Lett.*, vol. 29, no. 5, pp. 595–602, 2008.

[5] E. Flor, R. Aygun, S. Mercan, and K. Akkaya, "PRNU-based source camera identification for multimedia forensics," in *Proc. IEEE 22nd Int. Conf. Inf. Reuse Integr. Data Sci. (IRI)*, 2021, pp. 168–175.

[6] H. Zeng, J. Liu, J. Yu, X. Kang, Y. Q. Shi, and Z. J. Wang, "A framework of camera source identification Bayesian game," *IEEE Trans. Cybern.*, vol. 47, no. 7, pp. 1757–1768, Jul. 2017.

[7] F. Ding, G. Zhu, W. Dong, and Y.-Q. Shi, "An efficient weak sharpening detection method for image forensics," *J. Vis. Commun. Image Represent.*, vol. 50, pp. 93–99, Jan. 2018.

[8] D. Wang, T. Gao, and Y. Zhang, "Image sharpening detection based on difference sets," *IEEE Access*, vol. 8, pp. 51431–51445, 2020.

[9] Y. Zhang, F. Ding, S. Kwong, and G. Zhu, "Feature pyramid network for diffusion-based image inpainting detection," *Inf. Sci.*, vol. 572, pp. 29–42, Sep. 2021.

[10] A. J. Fridrich, B. D. Soukal, and A. J. Lukáš, "Detection of copy-move forgery in digital images," in *Proc. Digit. Forensic Res. Workshop*, 2003, pp. 1–10.

[11] K.-T. Huynh, T.-N. Ly, and T. Le-Tien, "A deep learning-based method for image tampering detection," in *Proc. Int. Conf. Future Data Security Eng.*, 2021, pp. 170–184.

[12] Y. Rao and J. Ni, "A deep learning approach to detection of splicing and copy-move forgeries in images," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2016, pp. 1–6.

[13] G. Singh and P. Goyal, "SDCN2: A shallow densely connected CNN for multi-purpose image manipulation detection," *ACM Trans. Multimidia Comput. Commun. Appl.*, to be published.

[14] Z. Guo, K. Yu, A. Jolfaei, F. Ding, and N. Zhang, "Fuz-spam: Label smoothing-based fuzzy detection of spammers in Internet of Things," *IEEE Trans. Fuzzy Syst.*, early access, Nov. 24, 2021, doi: 10.1109/TFUZZ.2021.3130311.

[15] F. Ding, Y. Shi, G. Zhu, and Y.-Q. Shi, "Real-time estimation for the parameters of Gaussian filtering via deep learning," *J. Real-Time Image Process.*, vol. 17, no. 1, pp. 17–27, 2020.

[16] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveillance (AVSS)*, 2018, pp. 1–6.

[17] Y. Al-Dhabi and S. Zhang, "Deepfake video detection by combining convolutional neural network (CNN) and recurrent neural network (RNN)," in *Proc. IEEE Int. Conf. Comput. Sci. Artif. Intell. Electron. Eng. (CSAIEE)*, 2021, pp. 236–241.

[18] J. Chen, X. Kang, Y. Liu, and Z. J. Wang, "Median filtering forensics based on convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 1849–1853, Nov. 2015.

[19] Y. Zhang, G. Zhu, L. Wu, S. Kwong, H. Zhang, and Y. Zhou, "Multi-task SE-network for image splicing localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4828–4840, Jul. 2022.

[20] F. Ding, H. Wu, G. Zhu, and Y.-Q. Shi, "METEOR: Measurable energy map toward the estimation of resampling rate via a convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4715–4727, Dec. 2020.

[21] W. Lu, Q. Zhang, S. Luo, Y. Zhou, J. Huang, and Y.-Q. Shi, "Robust estimation of upscaling factor on double JPEG compressed images," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10814–10826, Oct. 2022.

[22] I. C. Camacho and K. Wang, "A comprehensive review of deep-learning-based methods for image forensics," *J. Imag.*, vol. 7, no. 4, p. 69, 2021.

[23] J. Ye, Z. Shen, P. Behrani, F. Ding, and Y.-Q. Shi, "Detecting USM image sharpening by using CNN," *Signal Process. Image Commun.*, vol. 68, pp. 258–264, Oct. 2018.

[24] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Trans. Inf. Forensics Security*, vol. 13, pp. 2691–2706, 2018.

[25] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[26] S. Garfinkel, "Anti-forensics: Techniques, detection and countermeasures," in *Proc. 2nd Int. Conf. i-Warfare Security*, 2007, pp. 77–84.

[27] R. Böhme and M. Kirchner, "Counter-forensics: Attacking image forensics," in *Digital Image Forensics*. New York, NY, USA: Springer, 2013, pp. 327–366.

[28] S. Xun *et al.*, "Generative adversarial networks in medical image segmentation: A review," *Comput. Biol. Med.*, vol. 140, Jan. 2022, Art. no. 105063.

[29] F. Ding, K. Yu, Z. Gu, X. Li, and Y. Shi, "Perceptual enhancement for autonomous vehicles: Restoring visually degraded images for context prediction via adversarial training," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9430–9441, Jul. 2022.

[30] B. Sun, "Methodology and application of GAN algorithms," in *Proc. Int. Conf. Comput. Graph. Artif. Intell. Data Process. (ICCAID)*, 2022, pp. 663–668.

[31] J. Yu *et al.*, "Toward realistic face photo–sketch synthesis via composition-aided GANs," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4350–4362, Sep. 2021.

[32] R. Li, "Image style transfer with generative adversarial networks," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 2950–2954.

[33] M. C. Stamm, S. K. Tjoa, W. S. Lin, and K. R. Liu, "Undetectable image tampering through JPEG compression anti-forensics," in *Proc. IEEE Int. Conf. Image Process.*, 2010, pp. 2109–2112.

[34] M. C. Stamm and K. J. R. Liu, "Anti-forensics of digital image compression," *IEEE Trans. Inf. Forensics Security*, vol. 6, pp. 1050–1065, 2011.

[35] C. Pasquini and G. Boato, "JPEG compression anti-forensics based on first significant digit distribution," in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process. (MMSP)*, 2013, pp. 500–505.

[36] Z. Qian and X. Zhang, "Improved anti-forensics of JPEG compression," *J. Syst. Softw.*, vol. 91, pp. 100–108, May 2014.

[37] M. Fontani and M. Barni, "Hiding traces of median filtering in digital images," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, 2012, pp. 1239–1243.

[38] Z.-H. Wu, M. C. Stamm, and K. R. Liu, "Anti-forensics of median filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 3043–3047.

[39] K. Singh, A. Kansal, and G. Singh, "An improved median filtering anti-forensics with better image quality and forensic undetectability," *Multidimensional Syst. Signal Process.*, vol. 30, no. 4, pp. 1951–1974, 2019.

[40] F. Ding, G. Zhu, Y. Li, X. Zhang, P. K. Atrey, and S. Lyu, "Anti-forensics for face swapping videos via adversarial training," *IEEE Trans. Multimedia*, vol. 24, pp. 3429–3441, 2021.

[41] Y. Chen, W. He, N. Yokoya, and T.-Z. Huang, "Hyperspectral image restoration using weighted group sparsity-regularized low-rank tensor decomposition," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3556–3570, Aug. 2020.

[42] D. Kim, H.-U. Jang, S.-M. Mun, S. Choi, and H.-K. Lee, "Median filtered image restoration and anti-forensics using adversarial networks," *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 278–282, Feb. 2018.

[43] Y. Luo, H. Zi, Q. Zhang, and X. Kang, "Anti-forensics of JPEG compression using generative adversarial networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, 2018, pp. 952–956.

[44] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[45] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[46] K. Regmi and A. Borji, "Cross-view image synthesis using conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3501–3510.

[47] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by Inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2536–2544.

[48] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[49] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[50] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, pp. 430–444, 2006.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[52] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, pp. 868–882, 2012.

Feng Ding received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the M.S. and Ph.D. degrees in electrical and computer engineering from the New Jersey Institute of Technology, Newark, NJ, USA, in 2011 and 2017, respectively.

He was a Specially Appointed Researcher with the Osaka University of Japan, Suita, Japan, in 2018. He was a Postdoctoral Researcher with the State University of New York at Albany, Albany, NY, USA, from 2019 to 2020. He is currently a Professor with Nanchang University, Nanchang, China. His current research interests mainly include digital forensics, machine learning, medical imaging, and digital image processing.

Zhangyi Shen received the B.S. degree in information security from Hangzhou Dianzi University, Hangzhou, China, in 2015, and the M.S. and Ph.D. degrees in electrical and computer engineering from the New Jersey Institute of Technology, Newark, NJ, USA, in 2016 and 2021, respectively.

He is currently an Instructor with the School of Cyberspace, Hangzhou Dianzi University. His research interests include multimedia security, information security, image forensics, image anti-forensics, medical imaging, and digital image processing.

Guopu Zhu (Senior Member, IEEE) received the B.S. degree in transportation from Jilin University, Changchun, China, in 2002, and the M.S. and Ph.D. degrees in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2004 and 2007, respectively.

He is currently a Professor with the Harbin Institute of Technology. He has authored or coauthored more than 50 papers in peer-reviewed international journals. His main research areas are multimedia security, image processing, and control theory.

Prof. Zhu serves as an Associate Editor for several journals, including IEEE Transactions on Cybernetics, IEEE Systems Journal, *Journal of Information Security and Applications*, and *Electronics Letters*.

Sam Kwong (Fellow, IEEE) received the B.Sc. degree from the State University of New York at Buffalo, Buffalo, NY, USA, in 1983, the M.A.Sc. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1986, and the Ph.D. degree from FernUniversität Hagen, Hagen, Germany, in 1996.

He is currently the Chair Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong, where he previously served as the Department Head and a Professor from 2012 to 2018.

Prof. Kwong led the IEEE SMC Hong Kong Chapter to win the Best Chapter Award in 2011 and was awarded the Outstanding Contribution Award for his contributions to SMC 2015. He is currently the Associate Editor of leading IEEE transaction journals, including IEEE Transactions on Evolutionary Computation, IEEE Transactions on Industrial Informatics, and IEEE Transactions on Cybernetics. His involvement in the multiple facets of IEEE has been extensive and committed throughout the years. For IEEE Systems, Man and Cybernetics Society (SMCS), he serves as the Hong Kong SMCS Chapter Chairman, a Board Member, a Conference Coordinator, a Membership Coordinator, and a member of the Long Range Planning and Finance Committee, the Vice President of Conferences and Meetings, and the Vice President of Cybernetics. He was the President-Elect of the IEEE SMC Society in 2021. He currently serves as the President of the IEEE SMC Society.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree in electrical engineering from Hunan University, Changsha, China, and the M.S. degree in electrical engineering and the Ph.D. degree in 2010 in electrical engineering from Tufts University, Medford, MA, USA.

He is a Professor with the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include image processing, computer vision, machine learning, and multimedia security.

Dr. Zhou received the Third Price of Macao Natural Science Award as a sole winner in 2020 and a co-recipient in 2014. He serves as an Associate Editor for IEEE Transactions on Cybernetics, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Geoscience and Remote Sensing. He is a Fellow of the Society of Photo-Optical Instrumentation Engineers and was recognized as one of the "Highly Cited Researchers" in 2020 and 2021.

Siwei Lyu (Fellow, IEEE) received the B.S. degree in information science and the M.S. degree in computer science from Peking University, Beijing, China, in 1997 and 2000, respectively, and the Ph.D. degree in computer science from Dartmouth College, Hanover, NH, USA, in 2005.

He is an Empire Innovation Professor with the Department of Computer Science and Engineering and the Founding Director of UB Media Forensic Lab (UB MDFL), University at Buffalo (UB), State University of New York, Buffalo, NY, USA. Before joining UB, he was an Assistant Professor from 2008 to 2014, a Tenured Associate Professor from 2014 to 2019, and a Full Professor from 2019 to 2020, with the Department of Computer Science, University at Albany, State University of New York, Albany, NY, USA. From 2005 to 2008, he was a Postdoctoral Research Associate with the Howard Hughes Medical Institute and the Center for Neural Science, New York University, New York, NY, USA. He was an Assistant Researcher with Microsoft Research Asia (then Microsoft Research China), Beijing, in 2001. He has published over 150 refereed journal and conference papers. His research interests include digital media forensics, computer vision, and machine learning.

Dr. Lyu is the recipient of the IEEE Signal Processing Society Best Paper Award in 2011, the National Science Foundation CAREER Award in 2010, the SUNY Albany's Presidential Award for Excellence in Research and Creative Activities in 2017, the SUNY Chancellor's Award for Excellence in Research and Creative Activities in 2018, and the Google Faculty Research Award in 2019.