Film-GAN: towards realistic analog film photo generation

Haoyan Gong[1] · Jionglong Su[1] · Kah Phooi Seng[1] · Anh Nguyen[2] · Ao Liu[1] · Hongbin Liu[1]

Abstract
In recent years, the art of film photography has reemerged as a topic of interest for both researchers and the community. Unlike digital photography, which relies on pixels to capture and store information, film photography employs silver halide to capture the scene. This process imbues film photos with a unique colour and textured graininess not present in digital photography. In this paper, we propose Film-GAN, the first Generative Adversarial Network (GAN)-based method for translating digital images to film. Film-GAN generates a corresponding film transformation of the input image based on the desired reference film style. To improve the realism of the generated images, we introduce the colour-noise-encoding (CNE) network, which extracts the colour and graininess of the reference image separately. Our experimental simulations demonstrate that Film-GAN outperforms other state-of-the-art approaches on multiple datasets. Based on evaluations from both professional photographers and amateur photography enthusiasts, the images generated by Film-GAN also received a higher number of votes, indicating its ability to produce better film-effect images.

Keywords GAN · Photo generation · Generative network · Image translation

1 Introduction

Film photography has a storied history spanning over a century. However, in the twenty-first century, digital photography has emerged as the dominant medium due to its substantial advancements in digital sensors and streamlined photo development processes. Despite these benefits, some of the most respected photographers maintain that film remains the superior form of photography. Film is an imaging equipment coated with silver halide and is typically loaded into a film camera. The film captures light as it passes through the lens, causing silver ions to be exposed and cured on the film base. Its photosensitive scale operates at the atomic level, resulting in more comprehensive information being recorded than digital photos of equivalent size. Furthermore, film produces a richer and more delicate visual perception. The resurgence of interest in film colour and photos has been observed among both the general community and researchers, particularly among the new generation on social media platforms. Despite the limited availability of film cameras and the complexity of film development, some individuals seek to learn the art of film photography, while others utilize mobile applications (APPs) to simulate a film-like effect on digital photos. These APPs that simulate film-like effects typically convert digital photos by linearly adjusting the colour filter. However, this approach often results in the loss of image details and fails to replicate the true characteristics of film photography. Recent advancements in style domain translation and image-to-image translation have made significant progress, particularly with the use of deep learning

✉ Hongbin Liu
hongbin.liu@xjtlu.edu.cn

Haoyan Gong
haoyan.gong21@student.xjtlu.edu.cn

Jionglong Su
jionglong.su@xjtlu.edu.cn

Kah Phooi Seng
jasmine.seng@xjtlu.edu.cn

Anh Nguyen
anh.nguyen@liverpool.ac.uk

Ao Liu
ao.liu19@student.xjtlu.edu.cn

[1] School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Jiangsu, People's Republic of China

[2] Department of Computer Science, University of Liverpool, Brownlow Hill, Liverpool L69 7ZX, UK

models that can extract different domain features for translation. For example, models based on Generative Adversarial Network (GAN) [1] can utilize both paired and unpaired images. In this paper, we propose a GAN-based model that can generate realistic film photos from digital images.

Prior image-to-image techniques have achieved tremendous success in various domains. Nevertheless, several obstacles remain to be overcome for the digital-to-film translation task. Firstly, one of the primary distinctions between film and digital photography lies in colour, encompassing factors such as hue and colour temperature. In regular circumstances, digital photographs strive to reproduce the colours of the actual scene, whereas developed film imparts a different colouration to the scene based on its style. For instance, photos taken using Kodak Gold 200 typically exhibit an overall yellowish hue. Therefore, accurately converting the colour feature of digital photos to the effect of film has certain difficulties. Secondly, digital photos consist of small units of pixels, and the noise is generated by the thermal motion of molecules. In contrast, film is composed of silver salt grains, and its graininess is a by-product of poor photographic quality. Additionally, the random noise produced by digital sensors blurs the entire image. Conversely, the silver halide grains produced by film are naturally arranged based on the light exposure, resulting in a more textured, sharper, and visually stylish image. Such film grain pattern is not randomly generated and it is hard to approximate without prior knowledge (see Fig. 1). Thirdly, there exist multiple types of films in the market, each possessing unique characteristics. The digital-to-film translation task involves a one-to-multiple domain conversion, transforming digital photos into various film effects. Consequently, addressing these challenges cannot be achieved solely by transferring prior approaches [2–5] on this task.

To address the aforementioned challenges, we propose the Film-GAN, which is the first approach to use Generative Adversarial Networks (GANs) to solve the digital-to-film translation task. Given that colour and graininess are the main differences between the two domains, we adapt our model to generate realistic images based on these aspects. Specifically, our approach introduces a colour-noise-encoding (CNE) network that comprises a noise-separation (NS) module, composed of a denoising module and an extracting noise module, as well as two encoders, namely, the colour encoder (CE) and the noise encoder (NE), to compress the image and noise into feature vectors, respectively. We then combine these with a source image as input for a single GAN, which generates the film image. Furthermore, our CE, NE, and discriminator units have multi-layer outputs to accommodate the transformation of the input to multiple film styles. During training, we only provide film datasets, such as Kodak Gold 200, Kodak Portra 160, and Ilford Hp5, and our model learns the transitions between film styles, while the trained model can be directly used for the digital-to-film conversion task.

The main contributions of this article are summarized as follows:

- We introduce Film-GAN, the first GAN-based approach for digital-to-film translation, which is capable of converting digital images into three different film styles as an example. To ensure the accuracy of feature extraction and maintain an appropriate distance between the GAN-generated image and the original

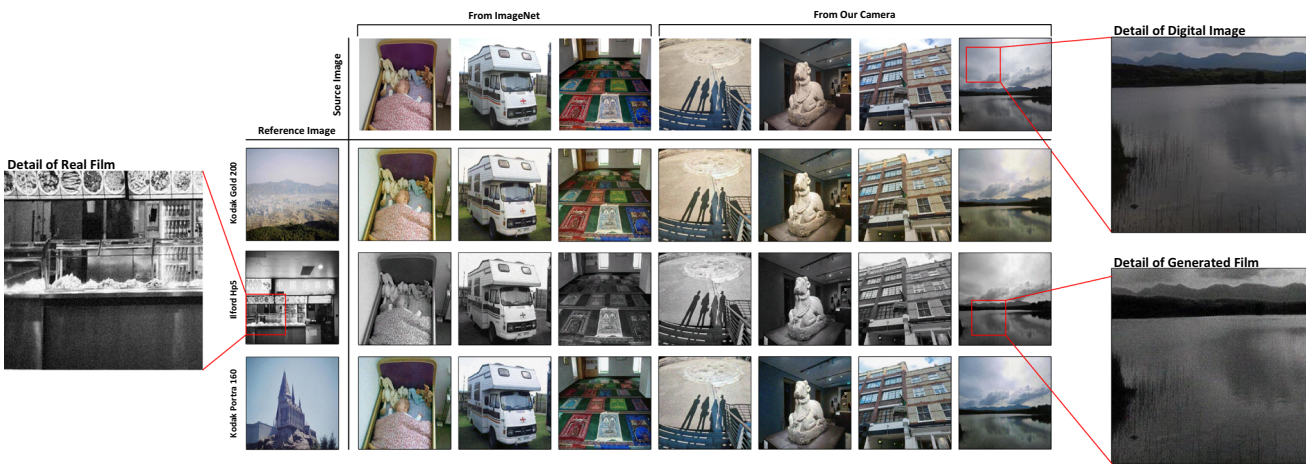

Fig. 1 The source images, reference images, generated images with film styles, including Kodak Gold 200, Kodak Portra 160 and Ilford Hp5, and three detail images. The source images contain the photos from the public dataset ImageNet and the photos we took with our smartphone. The difference in graininess and colour between film and digital images can be seen by comparing Detail of Real Film and Detail of Digital Image. Our model produces the film very close to the real film in colour and graininess, also guarantees the requirement of generating various styles. Please check the details in 200% zoom

image, we incorporate style loss and reconstruction loss. Our approach outperforms previous methods in terms of generating more visually appealing film-like images, as presented in Fig. 1.

- We propose the colour-noise-encoding (CNE) network, which consists of the noise-separation (NS) module, the colour encoder (CE), and the noise encoder (NE), to segregate and extract content colour and grains separately as multiple conditions for GAN. We pre-train the NS module for noise reduction and then perform noise separation on the film images to generate a noise map and a denoised map for subsequent feature extraction performed by CE and NE. Consequently, the CNE network enhances the quality of the images generated by GAN, and is better in impression.

2 Related work

Recently, a significant number of approaches are proposed for image translation. The approaches can be divided into deep learning-based approaches and GAN-based approaches. For deep learning-based methods, some are trained to learn specific styles from corresponding images. [2] introduce a neural algorithm of artistic style that can combine content and appearance of artworks to achieve image-to-image style transfer, and is the first to propose a Gram matrix for style representation. [3] propose a new loss calculation method for image style translation named perceptual loss. Instead of utilizing the per-pixel loss between the output and the reference image, it calculates the distance between the two in the feature domain. After the above methods are proposed, people realize the limitation that they can only generate fixed styles, so new approaches are developed to transfer arbitrary styles. [6] firstly present adaptive instance normalization (AdaIN) layer to combine content and arbitrary styles by transferring feature statistics. [7] take a pair of transforms, i.e., whitening and colouring, to the image reconstruction network for matching the content to target style. At the same time, a series of methods based on Generative Adversarial Networks [1] are proposed to generate arbitrary style images. [8] innovatively use conditional GAN [9] to complete the framework of image translation. [4] present CycleGAN using unpaired data to complete image style translation. [10] introduce a mask module to the adversarial network structure to spatially determine the stylization level for output. [11] propose a style-aware loss trained with a generator network, which better captures the style affects content. In addition, many other GAN-based approaches are proposed, such as UNIT [12], DualGAN [13] and DiscoGAN [14], that can generate arbitrary styles, i.e., one-to-one style translation. [15] use a multilayer and patch-based approach to conduct one-sided translation in the unpaired image translation. [16] introduce the attention mechanism to GAN, the generator can produce an attention mask to obtain high-quality translation results. [17] introduces FreezeSG and structure loss based on fine-tuning styleGAN v2, maintaining the structure between the source image and target image. [18] extend the latent space of StyleGAN with additional content features by developing a two-branch model. [19] enhance the handling of perform geometry changes and remove large objects scenarios with CAGAN. [20] use image translation to solve the problem of image culturalization. [21] propose a dual perceptual loss based on the complementarity between VGG and ResNet features to improve the effect of image reconstruction. [22] propose VecGAN for image translation by employing latent space decomposition to design learnable attribute editing. Other approaches focus on multi-modal image style translation. [23] propose a framework named MUNIT, which assumes that image features can be decomposed into content and style codes, to be reassembled after dismantling. [24] propose StarGAN, a novel and scalable approach that achieves multi-domain conversion by encoding the reference image and training multi-domain dataset at the same time, and further propose StarGAN v2 [5] that add encoder and mapping network for better handling of reference style features and latent code. [25] propose a new multi-domain image translation method by generating a new model in the target domain for the corresponding conversion. Then, many studies on the application of GAN-based image translation appeared. In order to handle geometric translation, [26] develop ObjectVariedGAN to learn the shapes mapping between source and target domains specifically. [27] propose SelectionGAN guide image-to-image translation by combining conditional semantics, multi-scale spatial pooling and multi-channel attention. [28] provide OutfitGAN, which uses the collocation classification module (CCM) to translate one extant fashion item to an entire outfit according to the mapping relationship identified by the semantic alignment module (SAM). [29] enhance the underwater images using a multi-stage generator network inspired by CycleGAN. [30] devise a cascade dense-channel attention (CDCA) module to adaptively distinguish noise feature and combine it with GAN structure, called UAGAN. Some other methods are introduced for multi-modal image style translation [31–35]. In the field of medical image processing, there are several GAN-based models proposed for image translation [35–37] (Figs. 2, 3).

Although these methods are very successful, most of them are one-to-one domain translations, while multi-modal translations are mostly for portraits or colour prediction, which are not suitable for multi-modal-digital-film

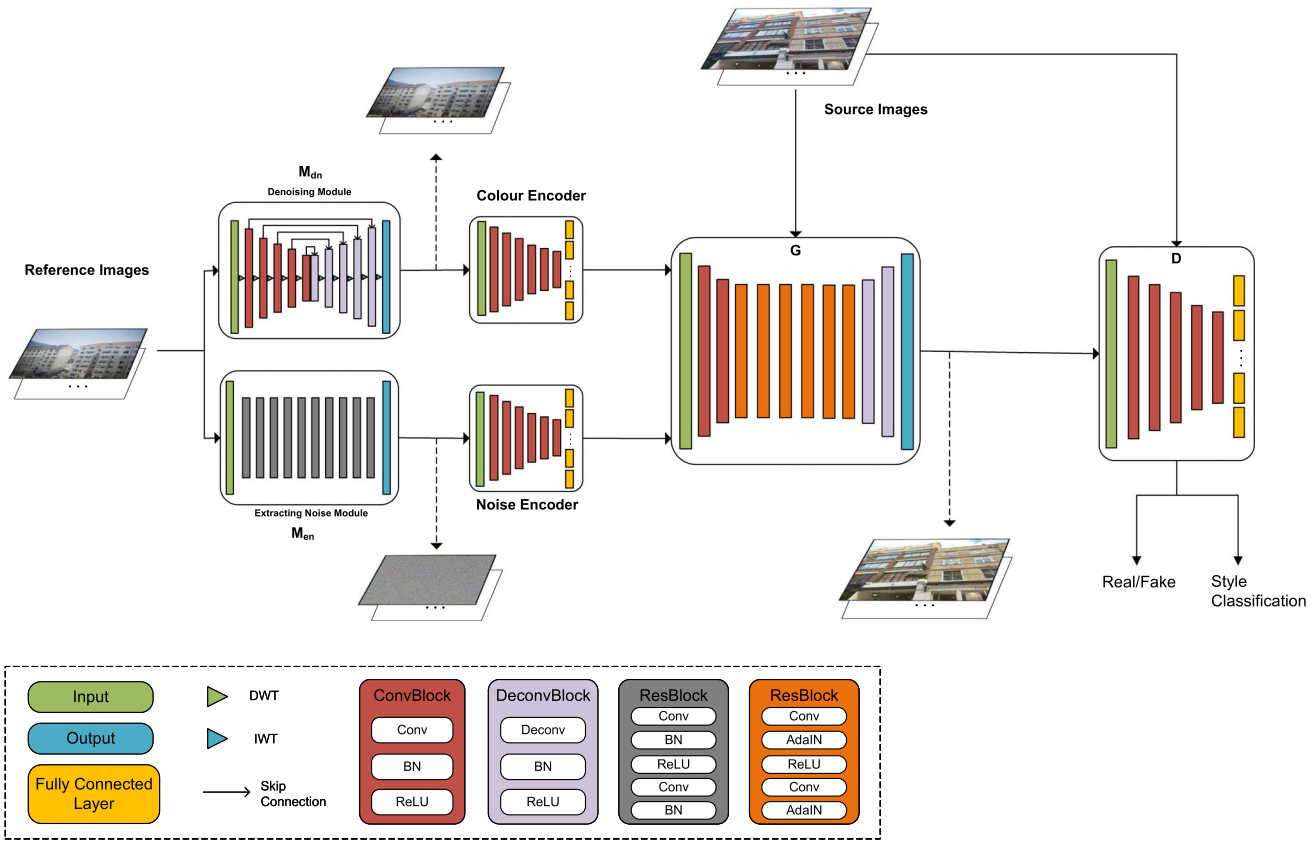

Fig. 2 The overall architecture of Film-GAN. It takes the reference image into the M_{dn} and the M_{en} to extract colour and noise features, respectively. The results are compressed by the E_i and the E_n into feature vectors, which are finally used as inputs of the generator G together with the source image to generate the film style we expected. During the training process, discriminator D accepts the result of G and the source image and then distinguishes their real-fake case and the style they belong to

style translation task that needs to focus on distinctive colour and textured graininess.

3 Method

3.1 Overview

We take P and F to represent the digital photo domain and film photo domain. We take $F = \{F_1, F_2, \cdots, F_n\}$, which contains n film styles. Film-GAN has two inputs, where the source image x_s is the object we want to transform and the reference image x_r is an example of the target film style. They have a style $f \in F$ and a style $\tilde{f} \in F$, respectively. After processing by NS module M_{ns}, reference image x_r is separated into denoised images and noise images, where the noise is actually the grains in film. The generator G receives the vectors compressed by colour encoder E_i and noise encoder E_n translates source image to target style. Inspired by StarGAN [5, 24], the discriminator D determines the image generated by G, and outputs n-dimensional results, each of which corresponds to a discriminant matrix for a style. In the training phase, source image x_s and reference image x_r are both divided into the n categories each representing a film style, and Film-GAN learns a mapping $\Phi_f : F_s \rightarrow F_r$. In the application stage, the source images are replaced by digital photos which every photo has a style $p \in P$ to achieve a transferring from film-film mapping Φ_f to digital-to-film mapping $\Phi_d : P \rightarrow F_r$.

3.2 Colour-noise-encoding network

3.2.1 Noise-separation module

The reference film image x_r is one of the important conditions for the generator G to produce an image. At first, we try to perform the feature extraction according to the original reference image by only an encoder; however, G generates a poor graininess effect. This shows that a single encoder cannot extract the features of the grains very well. Second, we add a pre-trained model for noise reduction, and in order to facilitate the simultaneous generation of noisy images, we assume that the grains in the film can be approximated by additive noise. Although grain is a by-

product of the photo reaction of film, which is generated in a different way from digital noise, it can be regarded as digital noise when viewed and processed with an electronic device. Based on the above assumption, we can use the properties of additive noise to calculate the noise image based on the reference film image and the clean image. However, since the denoising model cannot restore a perfectly clean image, its deviation will further lead to the inaccuracy of the noise image, so its performance is still not good. Therefore, we finally introduce two modules: denoising module M_{dn} and extracting noise module M_{en} to compose the noise-separation (NS) module M_{ns} together. The objective they are hoped to achieve is $x_r = M_{dn}(x_r) + M_{en}(x_r)$, We employ combinatorial training to jointly train two models and present a loss function to link them. Their loss functions are introduced in Sect. 3.3.

Denoising module. This part mainly focuses on restoring the input reference image to a clean image, that is, noise reduction. According to the practice of MWCNN [38], we develop M_{dn} by replacing the up and down sampling in U-Net [39] with discrete wavelet transform (DWT) and inverse wavelet transform (IWT) [40, 41]. It contains a five-layer encoder and a five-layer decoder, and finally obtains the output clean image through a fully connected layer. Each layer has a skip connection from the encoder to the decoder. In pre-training process, it accepts a noisy image I_n and outputs a denoised image I_c of the same size, that is $I_c = M_{dn}(I_n)$. In addition, we hope this module does not over-perform, because the side effect of excessive denoising capability will distort the values of some irrelevant pixels, the colour characteristics of the film itself will be changed and blurred, which makes it difficult for the encoder to extract features. And, this problem is well solved by the combinatorial training mentioned above.

Extracting noise module. This part mainly focuses on generating the noise image based on the input reference image. Inspired by DnCNN [42], we introduce M_{en} that use

Fig. 3 Performance comparison with various models. **a** Input source images. **b** Results of Film-GAN. **c**-**i** Results of Gatys et al., Fast NST, CycleGAN, MUNIT, StarGAN v2, SAVI2I and DiffuseIT, respectively

residual learning to extract the noise. There are twelve layers of the convolutional network, in which the first layer structure is a convolution operation and ReLU function, the last layer structure has only one convolution operation, and the middle part is the 10-layer structure of [Convolution, Batch Norm and ReLU function]. Finally, in pre-training process, it transforms noisy image I_n to its noise image $\tilde{I}_n$, that is, $\tilde{I}_n = M_{en}(I_n)$. Moreover, the model needs noisy and noise image pairs for training. Since there is no corresponding noise image in the SSID dataset, our combinatorial loss function avoids using the noise image directly.

3.2.2 Colour encoder and noise encoder

Since we process clean images and noisy images separately, we introduce two encoders to process these two features. They share the same standard encoder structure. It contains eight layers. The first layer is a convolution that increases the channel of input. The subsequent six-layer network is a combination of convolution operation and pooling operation, and they extract feature information from the image step by step. In the final part of the model, there are n groups of unshared fully connected layers. Each group corresponds to a 64-dimensional vector output, so the encoder has n outputs, where n is the number of styles included in the film dataset. In addition to the clean or noise image, the encoder also accepts another integer parameter $n_i \in \{\mathbf{0}, \mathbf{1}, \ldots, n\}$ that represents the style to which the current image belongs. The encoder selects the corresponding unshared layer as the final output according to n_i. As the final components of the CNE network, the outputs of the colour encoder and the noise encoder will be received by G as the generated reference in the next step.

3.3 Pre-training and training loss

In the pre-training stage, we train the NS module on the SSID noise dataset [43], including denoising module M_{dn} and extracting noise module M_{en}. For the training data $\{I_{\mathrm{n}}^i, I_{\mathrm{c}}^i\}_{i=1}^N$, the loss functions for M_{dn} is defined as follows:

$$L_{dn}(M_{dn}) = \mathbb{E}_{I_{\mathrm{n}}^i I_{\mathrm{c}}^i}\left[\left\|M_{dn}\left(I_{\mathrm{n}}^i\right) - I_{\mathrm{c}}^i\right\|_1\right], \tag{1}$$

where denoising module M_{dn} processes I_{n}^i and outputs a clean image $M_{dn}\left(I_{\mathrm{n}}^i\right)$. This loss function restricts the output $M_{dn}\left(I_{\mathrm{n}}^i\right)$ from deviating from the target clean image I_{c}^i

As mentioned above, since noise image is a missing component in the training set, we perform a training method as follows:

$$L_{\mathrm{com}}(M_{dn}, M_{\mathrm{en}}) = \mathbb{E}_{I_{\mathrm{n}}^i, I_{\mathrm{c}}^i}\left[\left\| \begin{matrix} M_{\mathrm{dn}}\left(I_{\mathrm{n}}^i\right) \\ +M_{\mathrm{en}}\left(I_{\mathrm{n}}^i\right) - I_{\mathrm{n}}^i \end{matrix} \right\|_{\mathbf{1}}\right], \tag{2}$$

where extracting noise module M_{en} outputs a noise image $M_{en}\left(I_{\mathrm{n}}^i\right)$ over the input I_{n}^i. Based on the property of additive noise, the summation of clean image $M_{dn}\left(I_{\mathrm{n}}^i\right)$ and noise image $M_{en}\left(I_{\mathrm{n}}^i\right)$ is specified to be equal to the input I_{n}^i. It not only avoids using noise image directly, but also strengthens the connection between outputs generated by M_{dn} and M_{en}. Combining the above two loss functions, the overall loss of the NS module is as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{NS}}(\mathrm{M_{dn}}, \mathrm{M_{en}}) = \lambda_{dn}\mathcal{L}_{\mathrm{dn}}(\mathrm{M_{dn}}) \\ + \lambda_{com}\mathcal{L}_{\mathrm{com}}(\mathrm{M_{dn}}, \mathrm{M_{en}}),\end{aligned} \tag{3}$$

where λ_{dn} and λ_{com} are the parameters that control the importance of these two terms. In the main training stage, we take a source image x_s with a style $f \in F$, a reference image x_r with a style $\tilde{f} \in F$. After being processed by the trained NS module, the reference image's clean map $M_{dn}(x_r)$ and noise map $M_{en}(x_r)$ are generated. Then compressed by the colour encoder E_i and noise encoder E_n, we obtain colour vector $\tilde{c} = E_i^{\tilde{f}}(M_{dn}(x_r))$ and noise vector $\tilde{n} = E_n^{\tilde{f}}(M_{en}(x_r))$. $E_i^f(\cdot)$ and $E_n^f(\cdot)$ transforms an image to vector in corresponding style f. The generator G takes three inputs, the source input image x_s, colour vector $\tilde{c}$ and noise vector $\tilde{n}$ represents colour and noise style it will be transformed, tries to generate an image $G(x_s, \tilde{c}, \tilde{n})$. The real image x_s and generated image $G(x_s, \tilde{c}, \tilde{n})$ are passed to the discriminator D to distinguish. $D^f(\cdot)$ terms critical results in corresponding style f. The first part is adversarial loss [1] shown as follows,

$$\begin{aligned}\mathcal{L}_{adv}(D, G) = \mathbb{E}_{f, x_s}\left[\log D^f(x_s)\right] \\ + \mathbb{E}_{\tilde{f}, x_s, x_r}\left[\log\left(1 - D^{\tilde{f}}(G(x_s, \tilde{c}, \tilde{n}))\right)\right],\end{aligned} \tag{4}$$

where the discriminator D expects that it can accurately distinguish between real images and generated images. In the contrary, the generator G aims for its fake images to be judged as real images. The second part is colour restoration loss. In order to further strengthen the film effect of images generated by G and the learning ability of the colour encoder, we make the encoder extract the corresponding colour vector from the image generated by G. It is defined as follows:

$$\begin{aligned}\mathcal{L}_{col}(G, E_i) \\ = \mathbb{E}_{\tilde{f}, x_s, x_r}\left[\left\|\tilde{c} - E_i^{\tilde{f}}(M_{dn}(G(x_s, \tilde{c}, \tilde{n})))\right\|_1\right],\end{aligned} \tag{5}$$

where re-extracted colour vector is obtained from noise map of generated image $G(x_s, \tilde{c}, \tilde{n})$. It is expected that has

the minimum distance with the previous vectors $\tilde{c}$. In the case of high accuracy of G, this object can also maintain the stability of the encoder performance and prevent large deviations from expectations.

The third part is noise restoration loss. Like the colour restoration loss, it is similarly defined as follows,

$$\mathcal{L}_{noi}(G, E_n) = \mathbb{E}_{\tilde{f}, x_s, x_r}\left[\left\|\tilde{n} - E_n^{\tilde{f}}(M_{en}(G(x_s, \tilde{c}, \tilde{n}))\right\|_1\right], \tag{6}$$

where extracting noise module M_{en} produces noise map of generated image $G(x_s, \tilde{c}, \tilde{n})$, then noise encoder compresses noise map to noise vector approximates the previous $\tilde{n}$.

The fourth part is cycle consistency loss [4]. It guarantees the correlation and coherence of model output and input, and avoids training with paired dataset. Such loss function is defined as follows,

$$\mathcal{L}_{cyc}(G, E_i, E_n) = \mathbb{E}_{\substack{\sim \\ f, \tilde{f}, x_s, x_r}}\left[\|x_s - G(G(x_s, \tilde{c}, \tilde{n}), \hat{c}, \hat{n})\|_1\right], \tag{7}$$

where the colour vector and noise vector representing source style f is given by $\hat{c} = E_i^f(M_{dn}(x_s))$ and $\hat{n} = E_n^f(M_{en}(x_s))$, respectively. The generator G loads them and the generated fake image to the generator again, that is, request G to try to restore the image to the original source image style. Through comprehensive training generator, it ensures that images can be converted between different styles.

The final part is invariance loss. In a special case, the source image and reference image received by generator share the same style $f = \tilde{f}$. To handle this situation, invariance loss is defined as follows,

$$\mathcal{L}_{in}(G, E_i, E_n) = \mathbb{E}_{f, x_s}\left[\|x_s - G(x_s, \hat{c}, \hat{n})\|_1\right], \tag{8}$$

where the generator is expected to make zero change to the source image. We simulate this situation by utilizing the colour and noise vector generated from the source image itself, that is $\hat{c}$ and $\hat{n}$. Combining Equation 4-8, the total loss function for Film-GAN is defined as follows,

$$\mathcal{L}_{total}(D, G, E_i, E_n) = \mathcal{L}_{adv} + \lambda_{col}\mathcal{L}_{col} + \lambda_{noi}\mathcal{L}_{noi} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{in}\mathcal{L}_{in}, \tag{9}$$

where $\lambda_{\text{col}}, \lambda_{\text{noi}}, \lambda_{cyc}$ and λ_{in} are hyperparameters. During the training process, the primary objective of the generator G, colour encoder E_i and noise encoder E_n is to synthesize highly realistic images that are indiscernible from real images, thereby minimizing the loss function denoted as $\mathcal{L}_{total}(D, G, E_i, E_n)$. Simultaneously, due to the characteristics of adversarial learning, the primary goal of the discriminator D is to effectively distinguish between real and generated images, resulting in an increase in the loss function value. Consequently, the generator and two encoders try to minimize $\mathcal{L}_{total}(D, G, E_i, E_n)$, while the discriminator strives to maximize it, they gradually reach balance in the confrontation, and this process can be expressed as follows,

$$\min_{G, E_i, E_n} \max_{D} \mathcal{L}_{\text{total}}(\text{D}, \text{G}, \text{E}_\text{i}, \text{E}_\text{n}). \tag{10}$$

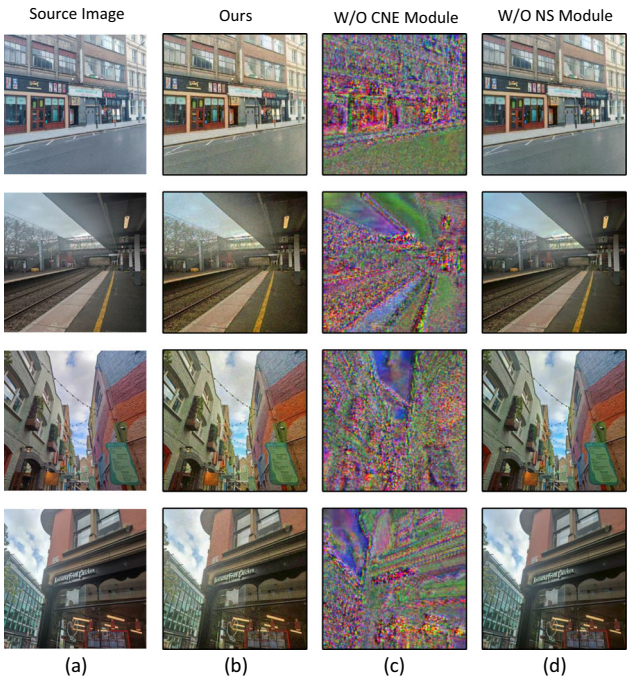

Fig. 4 Ablation experiment on Film-GAN. **a** Input source images. **b** Results of Film-GAN with complete CNE network. **c** Results of Film-GAN without CNE network. **d** Results of Film-GAN without NS module

4 Experiment

In this section, we first introduce the dataset and training details in Sects. 4.1 and 4.2. Then, we conduct an ablation study to test the effect of the NS module in Sect. 4.3, and several comparisons of other models and ours. All experiments are implemented in the PyTorch environment running on Linux with NVIDIA GeForce RTX 2060 GPU. Code for the models is available at https://github.com/haoyGONG/FilmGAN.

4.1 Dataset

There are three datasets in this paper, in which D_{pre} is for pre-training and D_{train} and D_{test} are for training process. In the pre-training process, dataset D_{pre} contains 2000 noisy-clean image pairs sampled from the public dataset

Table 1 Quantitative comparison in three film effects translation

Method	Kodak Gold 200		Kodak Portra 160		Ilford Hp5	
	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓
Gatys [2]	5.69	157.84	5.53	161.42	6.67	141.45
Fast NST [3]	3.02	166.38	3.95	157.12	3.82	138.49
CycleGAN [4]	5.99	157.16	6.13	163.49	7.58	145.13
MUNIT [23]	5.84	150.56	5.69	154.93	7.35	160.06
StarGAN v2 [5]	6.48	139.94	**12.17**	135.25	8.01	141.44
SAVI2I [46]	5.62	157.28	3.41	169.47	6.76	148.15
DiffuseIT [47]	2.21	252.15	2.12	263.89	2.52	190.9
FilmGAN	**7.92**	**132.83**	**12.23**	**131.49**	**14.45**	**123.4**

Smartphone Image Denoising Dataset (SIDD) [43]. Each sample of D_{pre} is a size of 256×256, normalized and optimized by cropping and horizontal flipping. In the training process, dataset D_{train} has two sub-datasets, which are source image set and reference image set. There are 786 images for Kodak Gold 200, 807 images for Kodak Portra 160 and 775 images for Ilford HP5 in each of the two sets. They have the same film styles and quantity, without sharing the image content, i.e., the images from two sub-datasets are different. Each sample of D_{train} is normalized and optimized by horizontal flipping and angle-correction with final size of 256×256. In our experiment, we utilize all of the samples as training set. For testing set D_{test}, in order to test the performance better, we sample 100 photos for source images, which are 50 digital photos from the public dataset ImageNet [44], 50 digital photos from a smartphone with following camera parameters: 48 MP 1/1.32 inch Quad-Bayer sensor, 25 mm equivalent f/1.85 lens, Phase Detection Autofocus (PDAF) and Optical Image Stabilization (OIS). Additionally, we take 100 reference-image sets in the structure of [Kodak Gold 200, Kodak Portra 160, Ilford HP5]. For each digital image, the model generates its three film-style transformations according to the reference. D_{test} has the same normalization, optimization and image size as D_{train}.

4.2 Training details

For pre-training denoising module M_{dn} and extracting noise module M_{en}, each sample of D_{pre} are fixed to 3 channels in RGB domain. We set parameters $\lambda_{dn} = 1$ and $\lambda_{com} = 1$ in Equation 3, and employ Adam solver [45] with batch size of 4. In the first 100 epochs, the learning rate is set to 0.0001, and the learning rate decreases linearly to 0 in the next 100 epochs.

For training part, each sample of D_{train} and D_{test} are fixed in RGB domain. The parameters in $\mathcal{L}_{total}$ are set to $\lambda_{col} = 1, \lambda_{noi} = 1, \lambda_{cyc} = 1, \lambda_{in} = 1$. We employ Adam solver [45] with a batch size of 5. In the first 100 epochs, the learning rate is set to 0.00001, and the learning rate decreases linearly to 0 in the next 300 epochs.

4.3 Ablation experiment

We introduce a CNE network to assist GAN in better generating target styles, as shown in Sect. 3.2, including a pretrained NS module, colour encoder and noise encoder. NS module is pretrained on dataset D_{pre} to learn to divide film image into clean image and noise image. The colour encoder and noise encoder are trained to find the feature representations of clean images and noise images from the NS module. These modules can enhance the model's ability to extract detailed features and encourage it to restore colour and noise as equally important features. Without this part, the model will automatically focus on the colour feature, while ignoring the noise feature, and since the grains in the film image also have a certain influence on the colour distribution, the model cannot correctly extract the colour feature. As shown in 3rd column of Fig. 4, if the entire CNE network is removed, the model receives the entire reference image without any target features, and the generated images are indistinguishable. As shown in 4th column of Fig. 4, if we eliminate the NS module, the performance of the resulting images is unsatisfactory in terms of colour and simulated graininess.

4.4 Comparison with the state-of-art method

We performed a comparison between Film-GAN and other seven state-of-art approaches that can translate digital image to film image, included GAN-based methods CycleGAN [4], MUNIT [23] and StarGAN v2 [5], other methods [2], Fast NST [3], SAVI2I [46] and DiffuseIT [47]. Since other datasets do not have paired film-digital images, we adopt the commonly used Inception Score (IS) [48] and Frechet Inception Distance (FID) [49] to measure the quality of the generated images and the similarity to the target domain distribution. IS directly evaluates the quality

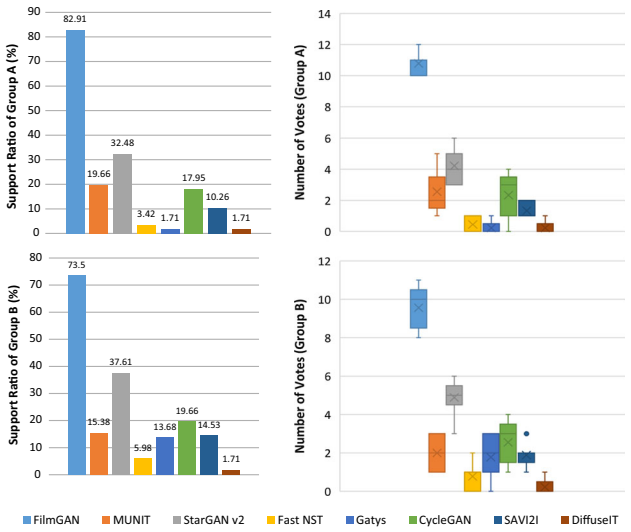

Fig. 5 A user survey on the performance of various models. The histograms (left) show the support ratio for each model, calculated as the ratio of votes to participation. The boxplots (right) show the number of votes each model received

and diversity of generated images, while FID calculates the divergence between the generated images and the real images in the feature space. We evaluated the dataset of Kodak Gold 200, Kodak Portra 160 and Ilford Hp5, respectively, as shown in Table 1. Our model achieves the best IS and FID on both Kodak Gold 200 and Ilford Hp5, and has similar IS values to StarGAN on the Kodak Portra 160 dataset. We provide the generated samples for visual comparison, which is conducted in processing film effect of Kodak Gold 200, taking seven digital images as input and using the same Kodak Gold 200 image for reference. In Fig. 3, the results of all referenced methods have a common defect, which is the reduction of graininess. This issue exists in both deep learning-based methods and GAN-based methods, and they probably ignore this feature to a certain extent without learning the features of graininess. On the other hand, in terms of colour, because they process reference image without separating noise, and the contents of the reference image and the source image are irrelevant, the models lose some information about the original image during the restoration process, resulting in overlapping or confusing colours, and overall visual is unsatisfactory.

4.5 User survey

We invited 26 volunteers to participate in the user survey. They are divided into two groups: Group A contains 13 people with film photography experience, and Group B contains other people without relevant experience. We randomly select nine photos in the test dataset D_{test} and convert them into different film styles using various models, including [2], Fast NST [3], CycleGAN [4], MUNIT [23], StarGAN v2 [5], SAVI2I [46] and DiffuseIT [47]. For each image, all volunteers need to vote for one or two models that they think have the best film. Since it is multiple choice, we use the ratio of votes to the number of participants to calculate the support rate for each model. The results are shown in Fig. 5, Film-GAN obtained a support ratio of 82.91% and 73.50% in Group A and Group B, respectively, which is more popular than the results of other models.

5 Conclusions

Generating analogue-film-like photos from digital images is a practical and challenging task. In this paper, we propose Film-GAN, the first GAN-based approach for unpaired digital-to-film translation. We implement one-to-three film style translations as an example, converting digital images into Kodak Gold 200, Kodak Portra 160, and Ilford Hp5 film styles, respectively. By separately processing the two important film features, colour and graininess, using the innovated CNE network, Film-GAN significantly improves the digital-to-film translation over existing solutions. Experiment results demonstrate the outstanding generative photo quality of our model, and statistical measures confirm that Film-GAN outperforms other state-of-the-art methods.

Although this research makes some progress in generating realistic film images, there are still limitations that need to be addressed. This research solely focuses on the authenticity of the generated images, while other aspects, such as control over grain density and diversity, have not been thoroughly explored. Therefore, future research could consider further optimizing the generator model to enhance grain diversity, for instance, by introducing parameters that control the level of grain. Additionally, further research can explore the application and generalization of the model on different film datasets to validate its universality and robustness.

Data availability The datasets generated and analyzed during the current study are not publicly available, as they originate from multiple private film collections and cannot be disclosed without authorization from the original authors. However, they can be obtained from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Goodfellow I, Pouget-Abadie J, Mirza M, et al. (2014) Generative adversarial nets. Advances in neural information processing systems 27
2. Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
3. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision, Springer, pp 694–711
4. Zhu JY, Park T, Isola P, et al. (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232
5. Choi Y, Uh Y, Yoo J, et al. (2020) Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8188–8197
6. Huang X, Belongie S (2017) Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision, pp 1501–1510
7. Li Y, Fang C, Yang J, et al. (2017) Universal style transfer via feature transforms. Advances in neural information processing systems 30
8. Isola P, Zhu JY, Zhou T, et al. (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
9. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784
10. Xu Z, Wilber M, Fang C, et al. (2018) Learning from multi-domain artistic images for arbitrary style transfer. arXiv preprint arXiv:1805.09987
11. Sanakoyeu A, Kotovenko D, Lang S, et al. (2018) A style-aware content loss for real-time hd style transfer. In: proceedings of the European conference on computer vision (ECCV), pp 698–714
12. Liu MY, Breuel T, Kautz J (2017) Unsupervised image-to-image translation networks. Advances in neural information processing systems 30
13. Yi Z, Zhang H, Tan P, et al. (2017) Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE international conference on computer vision, pp 2849–2857
14. Kim T, Cha M, Kim H, et al. (2017) Learning to discover cross-domain relations with generative adversarial networks. In: International conference on machine learning, PMLR, pp 1857–1865
15. Park T, Efros AA, Zhang R, et al. (2020) Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, Springer, pp 319–345
16. Tang H, Liu H, Xu D, et al. (2021) Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE Transactions on Neural Networks and Learning Systems
17. Back J (2021) Fine-tuning stylegan2 for cartoon face generation. arXiv preprint arXiv:2106.12445
18. Yu Y, Kamran G, HsiangTao W, et al. (2022) Expanding the latent space of stylegan for real face editing. arXiv preprint arXiv:2204.12530
19. Hou X, Song J, Liu H (2022) Unpaired image-to-image translation using generative adversarial networks with coordinate attention loss. In: 2022 4th International Conference on Intelligent Information Processing (IIP), IEEE, pp 68–76
20. Zaino G, Recchiuto CT, Sgorbissa A (2022) Culture-to-culture image translation with generative adversarial networks. arXiv preprint arXiv:2201.01565
21. Song J, Yi H, Xu W, et al. (2022) Dual perceptual loss for single image super-resolution using esrgan. arXiv preprint arXiv:2201.06383
22. Dalva Y, Altındiş SF, Dundar A (2022) Vecgan: Image-to-image translation with interpretable latent directions. In: European Conference on Computer Vision, Springer, pp 153–169
23. Huang X, Liu MY, Belongie S, et al. (2018) Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV), pp 172–189
24. Choi Y, Choi M, Kim M, et al. (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, pp 8789–8797
25. Huang J, Liao J, Kwong S (2021) Unsupervised image-to-image translation via pre-trained stylegan2 network. IEEE Trans. Multimed 24:1435–1448
26. Qin Z, Chen Q, Ding Y et al (2022) Segmentation mask and feature similarity loss guided gan for object-oriented image-to-image translation. Inform Process Manag 59(3):102,926
27. Tang H, Torr PH, Sebe N (2022) Multi-channel attention selection gans for guided image-to-image translation. IEEE Trans Pattern Anal Mach Intell 45(5):6055–6071
28. Zhou D, Zhang H, Yang K, et al. (2022) Learning to synthesize compatible fashion items using semantic alignment and collocation classification: An outfit generation framework. IEEE Transactions on Neural Networks and Learning Systems
29. Hu K, Weng C, Shen C et al (2023) A multi-stage underwater image aesthetic enhancement algorithm based on a generative adversarial network. Eng Appl Artificial Intell 123(106):196
30. Wang N, Chen T, Kong X, et al. (2023) Underwater attentional generative adversarial networks for image enhancement. IEEE Transactions on Human-Machine Systems
31. Zhu JY, Zhang R, Pathak D, et al. (2017) Toward multimodal image-to-image translation. Advances in neural information processing systems 30
32. Lee HY, Tseng HY, Huang JB, et al. (2018) Diverse image-to-image translation via disentangled representations. In: Proceedings of the European conference on computer vision (ECCV), pp 35–51
33. Lee HY, Tseng HY, Mao Q et al (2020) Drit++: Diverse image-to-image translation via disentangled representations. Int J Comput Vis 128(10):2402–2417
34. Lian Y, Shi X, Shen S, et al. (2023) Multitask learning for image translation and salient object detection from multimodal remote sensing images. The Visual Computer pp 1–20
35. Cao B, Bi Z, Hu Q, et al. (2023) Autoencoder-driven multimodal collaborative learning for medical image synthesis. Int J Comput Vis pp 1–20
36. Tan C, Yang M, You Z et al (2022) A selective kernel-based cycle-consistent generative adversarial network for unpaired low-dose ct denoising. Precision Clin Med 5(2):pbac011
37. Wang Y, Chen Y, Wang W, et al. (2022) Msgan: Multi-stage generative adversarial networks for cross-modality domain adaptation. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, pp 520–524
38. Liu P, Zhang H, Zhang K, et al. (2018) Multi-level wavelet-cnn for image restoration. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 773–782
39. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer, pp 234–241

40. Edwards T (1991) Discrete wavelet transforms: Theory and implementation. Universidad de pp 28–35
41. Shensa MJ et al (1992) The discrete wavelet transform: wedding the a trous and mallat algorithms. IEEE Trans Signal Process 40(10):2464–2482
42. Zhang K, Zuo W, Chen Y et al (2017) Beyond a gaussian denoiser: residual learning of deep cnn for image denoising. IEEE Trans Image Process 26(7):3142–3155
43. Abdelhamed A, Lin S, Brown MS (2018) A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1692–1700
44. Deng J, Dong W, Socher R, et al. (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp 248–255
45. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
46. Mao Q, Tseng HY, Lee HY et al (2022) Continuous and diverse image-to-image translation via signed attribute vectors. Int J Comput Vis 130(2):517–549
47. Kwon G, Ye JC (2022) Diffusion-based image translation using disentangled style and content representation. arXiv preprint arXiv:2209.15264
48. Salimans T, Goodfellow I, Zaremba W, et al. (2016) Improved techniques for training gans. Adv Neural Inform Process Syst 29
49. Heusel M, Ramsauer H, Unterthiner T, et al. (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv Neural Inform Process Syst 30