Zhu *et al.*
EURASIP Journal on Advances in Signal Processing (2023) 2023:66
https://doi.org/10.1186/s13634-023-01035-w

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

WDIG: a wavelet domain image generation framework based on frequency domain optimization

Qing Zhu[1,2], Xiumei Li[1,2], Junmei Sun[1,2] and Huang Bai[1,2*]

*Correspondence: bh667770@163.com

[1] School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China
[2] Key Laboratory of Cryptography of Zhejiang Province, Hangzhou Normal University, Hangzhou, China

Abstract

In the end-to-end image generation task, the spatial domain of pixel space cannot explicitly separate the low-frequency general information such as texture and color from the high-frequency detail information such as structure and identity. The loss function calculated in the spatial domain fails to effectively constrain the maintenance of detail information, and the generated image quality is insufficient. In this paper, a wavelet domain image generation (WDIG) framework is proposed to preserve the frequency information of images, in which the loss functions are constructed in the pixel space and wavelet space. In the pixel space, the low-frequency and high-frequency characteristic information of the signal are obtained by setting the appropriate Gaussian kernel and adopting the Gaussian fuzzy method. The loss function of ℓ_1 norm spatial domain is constructed for the low-frequency and high-frequency characteristic information. In the wavelet space, the corresponding channel sub-band coefficients are obtained by wavelet transform, and the image is explicitly separated into high-frequency information and low-frequency information. The ℓ_1 norm frequency domain loss function is constructed respectively for the sub-band coefficients. The WDIG can constrain model training more accurately and optimize model more precisely, so as to better maintain the details and quality of generated image. The WDIG framework is evaluated in the image generation applications including style transfer, image translation and Generative Adversarial Nets (GAN) Inversion. Experimental results show that the WDIG framework can effectively retain the details of images and generate more realistic images, and improve the image quality of the above applications in image generation.

Keywords: Image generation, Pixel space, Wavelet space, Frequency domain, Detail information

1 Introduction

As an important task in computer vision, image generation based on deep learning is widely used in art creation, industrial design, digital simulation and industry applications. General image generation usually includes image color generation, texture generation, and content generation. The related image generation tasks mainly include image style transfer, image translation, image repair, image attribute editing,

and generative adversarial nets (GAN) Inversion. Image generation is an end-to-end task based on autoencoder. The input and output of the network are images, which are encoded into latent codes during training and then decoded back to image space. According to Rate-Distortion theory [1], reversing a real-world image to a low-dimensional latent code would inevitably lead to information loss, and the lost information is primarily image details which result in poor image quality obtained by decoding latent codes. One of the strategies to solve this problem is to design more accurate loss function to constrain model training. The commonly used loss functions include per-pixel loss, adversarial loss and perceptual loss. However, these losses are all calculated in pixel space. Image generation task has a high requirement to maintain the detail information such as the structure and identity of the generated image. In pixel space, the detail information cannot be separated from the general information such as texture, color and contour of the image. Besides, the loss function calculated using pixel values cannot effectively improve the quality of the generated image.

Domain image generation has played an important role in many applications [2, 3]. Since the low-frequency of an image represents general information such as texture, color and contour, while the high-frequency represents detail information such as structure and identity, decomposing the image into low and high frequency components and calculating the loss separately will help to provide more accurate constraints for image generation. In the image preprocessing stage of computer vision tasks, even though the down-sampling operation could reduce computation complexity, it also obliviously removes both redundant and salient information, which results in accuracy degradation. In the task of image reconstruction and synthesis, a big gap exists between the real and generated images in the frequency domain. Narrowing gaps in the frequency domain can further ameliorate image reconstruction and synthesis quality. A learning-based frequency selection method is proposed to identify the trivial frequency components which can be removed without accuracy loss to reduce the size of the input image [4]. A novel focal frequency loss is proposed to directly optimize image reconstruction and synthesis in the frequency domain [5], which allows the model to adaptively focus on frequency components that are hard to synthesize by down-weighting the easy frequency components. To obtain better fidelity and visual quality for single image super resolution, Xie et al. [6] proposed a novel frequency-aware dynamic network for dividing the input into multiple parts according to its coefficients in the discrete cosine transform (DCT) domain. In order to solve the problem that synthesized images often over-adapt to the target domain, losing important structural characteristics and suffering from suboptimal visual quality in the image translation task, Cai et al. [7] constructed the loss function framework frequency domain image translation (FDIT) to preserve the frequency information in both pixel space and Fourier spectral space. FDIT constrains the image translation models training more accurately, and effectively preserves the identity of the source image while producing photo-realistic image hybrids. However, in the spectrum obtained by Fourier transform, it is compulsory to set the mask radius to separate the high-frequency and low-frequency of the image. The size of the mask radius is closely related to the size of the image, and plays an important role for the high and low frequency information separation. In the field of image generation, the size of training

images is not fixed for different tasks, therefore, it is difficult to select the mask radius and separate the high and low frequency under the FDIT framework.

Wavelet transform has good capability to extract details and perform multi-resolution analysis, and it can explicitly separate the high and low-frequency information of images. Using 2D discrete wavelet transform with Haar kernels, the input image can be decomposed into four sub-bands: A (average low-frequency information), V (vertical high-frequency information), H (horizontal high-frequency information) and D (diagonal high-frequency information). The low-frequency general information and high-frequency detail information can be represented separately on the sub-band, and its high-frequency information is low coupled and the separation is more thorough. The loss function calculated on the wavelet sub-bands can optimize the model accurately.

In this paper, a wavelet domain image generation framework WDIG is proposed based on wavelet transform. The loss function is constructed in pixel space and wavelet space respectively. The consistency of the generated image and the source image in low and high-frequency information is adjusted to enhance the quality of image generation. The WDIG framework is applied to image style transfer, image translation and GAN Inversion, which can effectively improve the quality of generated images. The main contributions of this paper are the following:

(1) A wavelet domain image generation framework WDIG is proposed, and its superiority is proved in image generation applications such as style transfer, image translation, and GAN Inversion.
(2) Gaussian blur is used to obtain the low and high frequency of the image in pixel space, and 2D discrete wavelet transform is used to obtain the low and high frequency of the image in wavelet space. The frequency information of the two spaces can be complementary. The WDIG framework is used to calculate the loss in the low and high frequency of the explicit separation of the image to achieve accurate constraints on the image generation model. The details of the generated image are better maintained and the quality of image generation is improved.
(3) The WDIG framework is applied to style transfer, image translation, GAN Inversion and other image generation tasks. Quantitative and qualitative evaluations on above tasks demonstrate the superiority of our framework. The experimental results show that when the WDIG framework is applied, the stylized image can retain the structure of the content image better in style transfer, the translated image is more photo-realistic and can preserve the identity of the source image in image translation, the generated reconstructed image is very similar to the input real image and improves the embedding accuracy of latent codes in GAN Inversion. In the above tasks, the WDIG framework significantly outperforms the baseline model, and in most cases outperforms the FDIT framework.

2 Related work

2.1 Image style transfer

Image style transfer is widely used in artistic creation, film and television entertainment and industrial design fields. It aligns style features in deep feature space, applies texture

and color styles of style images to content images, and generates stylized images with artistic characteristics. Image style transfer can be divided into optimization-based style transfer methods and feed-forward style transfer methods. The optimization-based style transfer methods optimize white noise images with high transfer quality but low efficiency, while the feed-forward style transfer methods optimize model parameters with high efficiency but low transfer quality. Currently, the most commonly used style transfer methods is feed-forward style transfer methods. The general process is as follows: Encode the content image I_c and style image I_s into the content feature $E(I_c)$ and style feature $E(I_s)$ in the feature space through encoder E; By aligning the mean and variance of the feature map [8] or whitening and coloring the feature map using the traditional algorithm [9], erase $E(I_c)$ from its own style, and then render the style of $E(I_s)$ to get the stylized feature t; Finally, the decoder D is used to decode t back to the stylized image $D(t)$ in the pixel space.

The process of erasing styles in arbitrary style transfer models will destroy the image structure information. As shown in Fig. 1, the content feature $E(I_c)$ of the content image is subtracted by the mean and then divided by the variance to erase style, then the feature that erases style is decoded back to pixel space. It can be seen that the structure of the image is destroyed. As indicated by the red box part in the fourth column, the mountains in the content image are continuous and smooth, while the mountains in the erased style image are undulating. Therefore, the process of erasing the style of $E(I_c)$ will damage its structure and further lead to structure distortion of the resulting stylized image.

Another reason for the structure distortion of the stylized image is that the pre-trained VGG with deeper network layers is selected as the encoder in style transfer task. Although the extracted features are refined and can express rich semantics [10], lots of image structure details are lost. Image decoded through the rendering style feature in feature layer is structurally distorted and blurry. Reducing the number of encoder network layers can reduce the loss of structural details. However, the style rendering on shallow features can only transfer the color style rather than achieve rich stylization effect. To solve this problem, many scholars have designed and improved the model construction to ensure the stylized effect as well as maintain the structure. Yao et al. [11] proposed the attention-aware multi-stroke (AAMS) model, which employed self-attention to expand the participation of salient regions as well as kept the correlation between distant regions, then maintained the structure of main areas in the stylized image. Liu

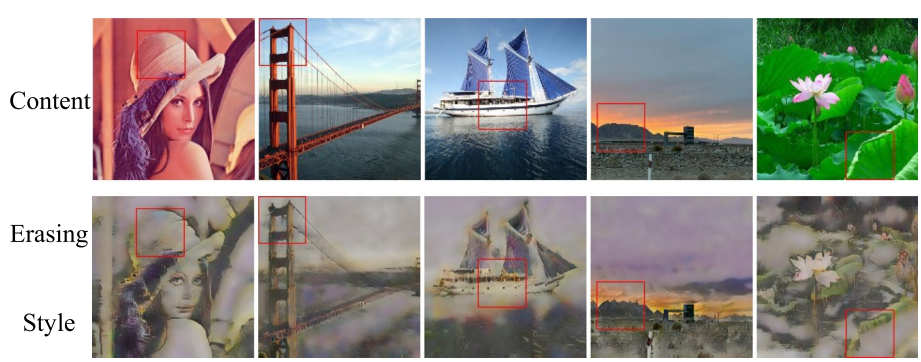

Fig. 1 Effects on image structure of erasing style manipulation

et al. [12] proposed adaptive attention normalization model (AdaAttn). The shallow and deep features were taken into account when calculating the attention weight. By aligning attention-weighted mean and variance of content feature maps and style feature maps in a per-point basis, the transfer quality of model was improved. Shen et al. [13] added an edge detection network to the neural style transfer model to extract the edge contour of the content image, then achieved a refined representation of the overall structure of the stylized image. Deng et al. [14] proposed a novel image style transfer transformer framework (StyTr^2), which can preserve the structure while enriching the stylization. However, transformer uses a large number of fully connected operations with amounts of parameters, and the model stylization speed is slow. Therefore, the model construction can achieve some performance improvement in the cost of additional computational consumption.

Figure 2 shows the stylized images generated by classical style transfer methods [8, 9, 11, 12, 15]. It can be seen that although the structure of stylized images generated by optimization-based method central moment discrepancy (CMD) [15] is well maintained, images are blurry and the transfer speed is far lower than feed-forward methods. The structure of stylized images by other methods is still commonly suffered from distortion. The proposed WDIG framework in this paper can accurately measure the lost structural information without additional computational consumption, and the structure of stylized images can be better maintained.

2.2 Image translation

Image translation can be applied to industrial design and digital simulation, such as the realization of satellite map to map conversion, real photo to sketch conversion and facial aging. Different from image style transfer which mainly transfers low-level styles such as textures and colors, image translation based on generative adversarial network can transform image from source domain to target domain through the adversarial training between generator and discriminator in pixel space, and the translated image is

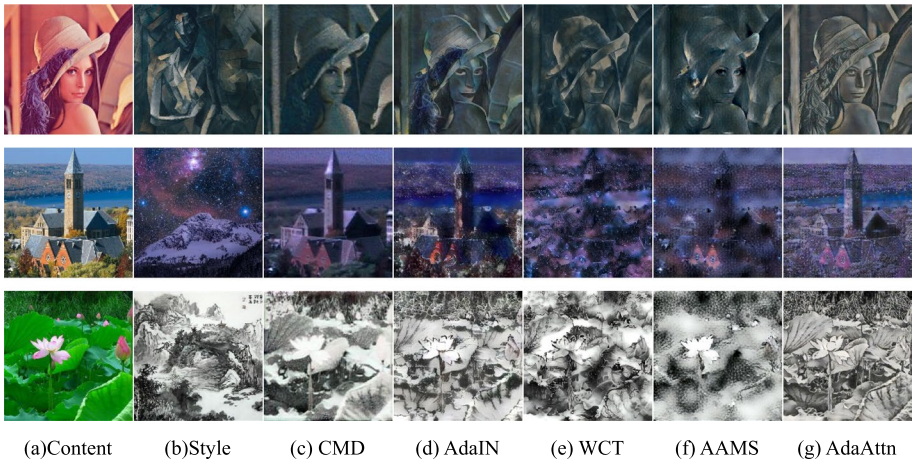

Fig. 2 Style transfer effects of different methods. **a** Content images. **b** Style images. **c** Style transfer effects of CMD [15]. **d** Style transfer effects of AdaIN [8]. **e** Style transfer effects of WCT [9]. **f** Style transfer effects of AAMS [11]. **g** Style transfer effects of AdaAttn [12]

more realistic. Image translation requires preserving the structural identity attribute of the source domain image and transforming it into the semantic attribute of the target domain image. The definitions of identity attributes and semantic attributes vary depending on the training dataset. If both the source domain image and the target domain image are faces, the identity attributes include hairstyle, gender and facial lines, and the semantic attributes include skin color, expression, age and illumination.

Image translation can be divided into single target domain image translation and multi-target domain image translation. Image translation methods in single target domain include supervised image translation Pix2Pix [16] and unsupervised image translation cycle-consistent adversarial networks (CycleGAN) [17]. The semantic attributes of a single target domain are encoded into the generator by training, and input source domain image can realize the transformation to target domain image at the semantic attribute level. Classical methods of image translation in multi-target domain mainly include multimodal unsupervised image-to-image translation (MUNIT) [18], swapping autoencoder (SWAE) [19], and StarGANv2 [20]. By unwrapping and recombining the structure and semantics of image in source domain and target domain, image translation from source domain to multiple target domain can be realized. The process is as follows. Given the source domain image x_1 and the target domain image x_2, x_1 and x_2 are encoded into features $E(x_1)$ and $E(x_2)$ by encoder E. $E(x_1)$ and $E(x_2)$ can be untangled for their respective content codes c_1, c_2 and style codes s_1, s_2. By combining the content code c_1 of the source domain image and the style code s_2 of the target domain image, the translated image $G(c_1, s_2)$ can be obtained by applying generator G.

However, the classical multi-target domain image translation methods have some shortcomings. Limited by the ability of feature resolution, translated images obtained by feature recombination are over-adapted to the target domain, which cannot maintain the identity of the image in the source domain, and the visual quality is poor. Figure 3 shows the translation effects of the classic image translation method StarGANv2 [20]. As shown in the red box, the images translated by StarGANv2 lose their identity information. For example, gender and hairstyle change, male is translated into female, and short hair becomes long hair. The proposed WDIG explicitly separates the image into low-frequency and high-frequency. The high frequency represents the structural identity property of the image, and the loss calculated at the high-frequency can accurately measure the lost identity information. Together with the generative adversarial loss of image translation itself, the model training is constrained to realize the transformation of semantic attributes without loss of identity attributes.

2.3 GAN inversion

Image attribute manipulation is also an image generation task, which is different from image translation and image style transfer. Instead of training the generative model, the latent codes are edited in the latent space, and the trained generative model is used to decode the latent codes back to the image, so as to edit the image attributes such as the expression and skin color in the face image. Image attribute editing is based on generative adversarial network GAN. However, a key limitation is the lack of a coding mechanism for making inferences about real images. The latent codes z corresponding to a given real image x cannot be directly derived, which limits the application of

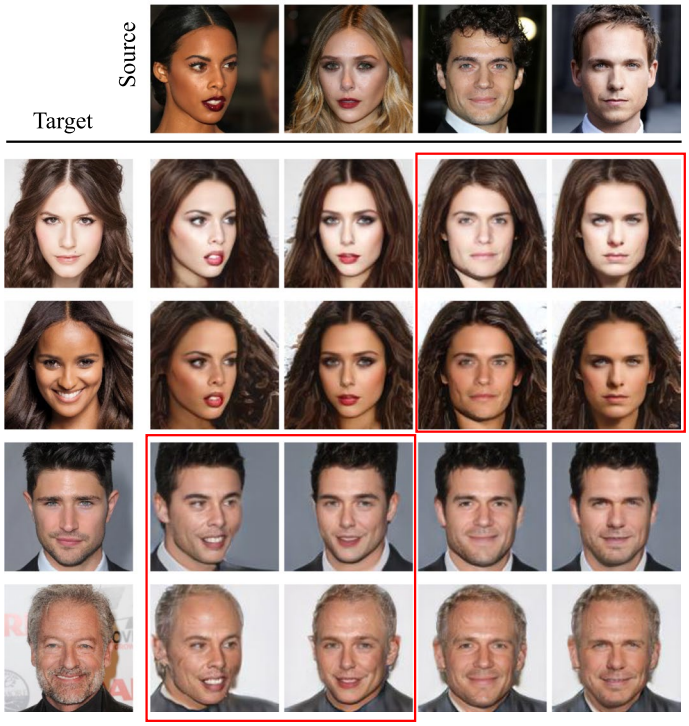

Fig. 3 Image translation effects of StarGANv2 [20] on CelebA-HQ [34]. The first row shows the source domain images, and the first column shows the target domain images

image attribute manipulation. The classical image generation method style-based generator architecture for generative adversarial networks (StyleGAN) [21] can generate high-resolution images through latent codes as well as generate style mixing images by editing and mixing the elements of two latent codes. However, due to the lack of mechanism to encode the real image into latent codes, style mixing of two real images is not possible. Therefore, many studies are devoted to GAN Inversion, a method for embedding real images into latent codes in latent space [22–24]. The more accurate the latent codes embedded in GAN Inversion, the more accurate the attribute of real images will be edited. Wang et al. [10] proposed a novel GAN inversion framework which consulted the observed distortion map as a high-rate reference to enhance the basic encoder model with high-quality reconstruction. Zhou et al. [25] presented an outfit generation framework COutfitGAN which used a silhouette and style fusion strategy for image synthesis and can overcome the spatial misalignment issue.

Currently, GAN Inversion methods fall into two categories, as shown in Fig. 4. The first type is to learn the encoder that can map a given real image to latent codes in latent space. A large number of latent codes are initialized and input into the pre-trained generator to generate corresponding images, and the training dataset is constructed. A large number of paired images and latent codes are used for training, and the loss is calculated by encoding latent codes and label latent codes. Then the encoder can learn to accurately encode the real image into latent codes. The second type of methods directly optimizes the latent codes and inputs the real image x as the label. In the training step, latent code z is initialized and input to the pre-trained

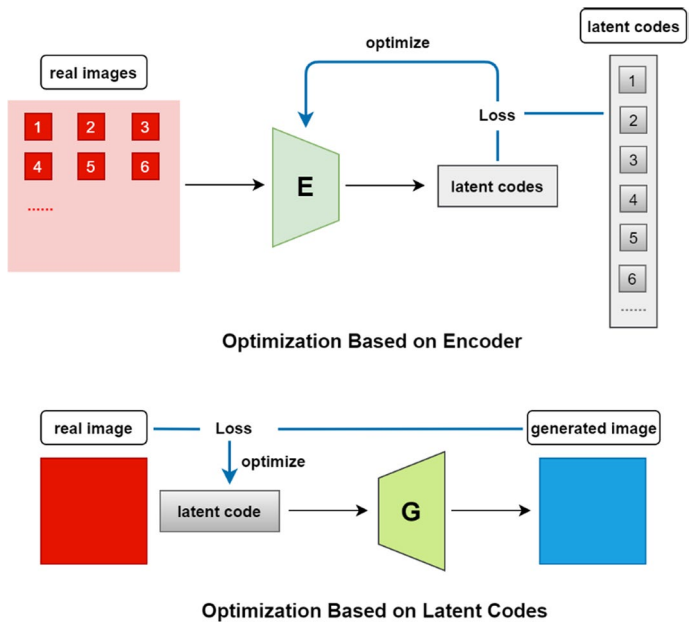

Fig. 4 Two types of GAN Inversion methods

generator G to generate the reconstructed image $G(z)$. The difference between $G(z)$ and x is calculated by the loss function to iteratively optimize z. When the loss is as small as possible and $G(z)$ is similar to x, the latent code z is equivalent to the latent code of x.

The second type of GAN Inversion method will generate images through latent codes and the labels are real images, and the WDIG framework can be applied. Image2StyleGAN [22] is a classic GAN Inversion method based on latent codes optimization which using StyleGAN pre-trained on the Flickr Faces HQ (FFHQ) dataset as a generator. However, it is highly sensitive to the initialization of latent codes. As the iteration progresses, the latent codes will continuously lose information and the generated reconstructed image will be quite different from the input real image. The embedded latent codes cannot accurately represent the input real image. The proposed WDIG calculates the loss in the frequency domain, which can reduce the difference between the reconstructed image generated by Image2StyleGAN and the input real image, and the embedded latent codes are more accurate.

2.4 Other important tasks

Besides image style transfer, image translation, and GAN inversion, there are many other important tasks in image generation such as image denoising and dehazing. In practical applications, denoising is an essential preprocessing step to recover improved image with high quality. Mahdaoui et al. [26] proposed an original image denoising method based on compressed sensing, which provided improved performance in terms of denoising efficiency and visual quality. Zhou et al. [27] proposed a novel effective dehazing method to restore a clear image, which can estimate transmission map and remove noise simultaneously.

2.5 Wavelet transform in deep learning

Some works have combined wavelet transform with deep learning for reducing the complexity of model calculation or improving the quality of image generation. In traditional U-Net, pooling layer and transposed convolution are often used as down-sampling and up-sampling layers. Liu et al. [28] embedded discrete wavelet transform (DWT) into CNN architecture, where DWT is introduced as a down-sampling operation without information loss and is helpful for preserving detail texture when using multi-frequency feature representation. Xue et al. [29] observed that most studies focused on designing deeper and wider architectures to improve the quality of image super-resolution at the cost of computational burden and speed. Therefore, a wavelet-based residual attention network was proposed. The input and label of network are four coefficients generated by the two-dimensional wavelet transform, which reduces the training difficulty of network by explicitly separating low-frequency and high-frequency details into four channels. Xin et al. [30] proposed an efficient and time-saving wavelet transform-based network architecture, where the image super-resolution processing is carried out in the wavelet domain. It can effectively guide the extraction of image feature maps by describing the high-frequency details in the wavelet domain. Ma et al. [31] used a scheme based on the frequency domain to reconstruct the high-resolution image at various frequency bands to achieve the most advanced performance in terms of super-resolution of remote sensing images. Guo et al. [32] designed a deep CNN to predict the missing details of wavelet coefficients of the low-resolution images to obtain the super-resolution results. To the best of our knowledge, no work has applied wavelet transform to the applications of image style transfer, image translation and GAN Inversion.

3 Proposed method

In this paper, a wavelet domain image generation framework WDIG is proposed based on the complementarity of Gaussian kernel and wavelet transform to preserve image frequency information. In the WDIG framework, loss functions are constructed in pixel space and wavelet space. In the pixel space, the low and high-frequency characteristic information of the signal are obtained by setting the appropriate Gaussian kernel and adopting the Gaussian fuzzy method. The loss function of ℓ_1 norm spatial domain is constructed for the low-frequency and high-frequency characteristic information. In the wavelet space, the corresponding channel sub-band coefficients are obtained by wavelet transform, and the image is explicitly separated into high-frequency information and low-frequency information. The ℓ_1 norm frequency domain loss functions are constructed respectively for the sub-band coefficients.

As shown in Fig. 5, taking the image style transfer task as an example, the specific process of the proposed WDIG framework is as follows. For the content image X, reconstructed image X' and stylized image X_Y are obtained by style transfer model, and high-frequency and low-frequency of X, X' and X_Y are separated respectively. In pixel space, Gaussian kernel convolution is used to separate X, X' and X_Y into high-frequency and low-frequency images respectively. In wavelet space, wavelet transform is applied to decompose X, X' and X_Y into sub-bands A, V, H, D respectively, where A represents low-frequency information, V represents the vertical high-frequency

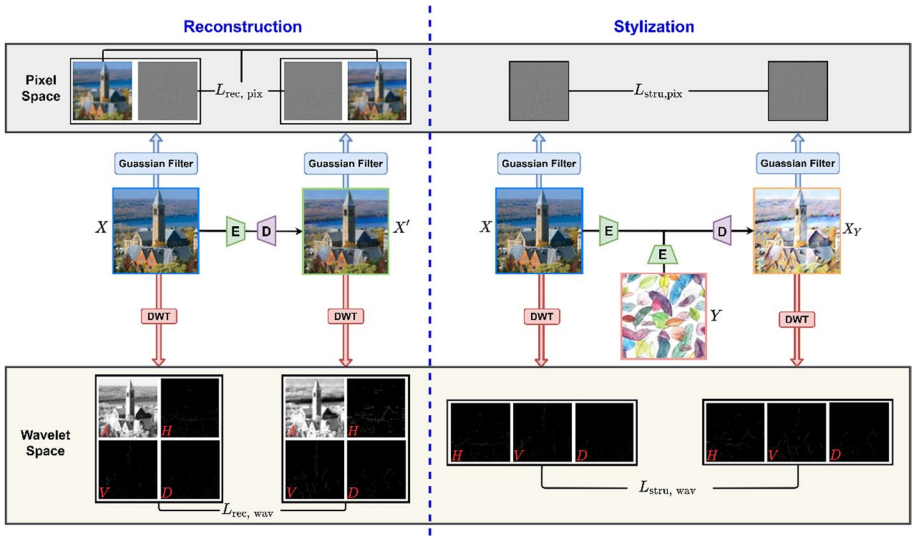

Fig. 5 Flowchart of the WDIG framework

information, H represents the horizontal high-frequency information, and D represents the diagonal high-frequency information. The pixel space loss includes reconstruction loss and structure loss. The reconstruction ℓ_1 loss is calculated at the high and low frequencies of X and X' respectively, and then summed. The structural ℓ_1 loss is calculated at the high frequencies of X and X_Y. The wavelet space loss includes reconstruction loss and structure loss. The reconstruction loss concatenates the four sub-bands which represent the high and low frequencies of X and X' into $[A, V, H, D]$, respectively, then the ℓ_1 loss is calculated. The structure loss concatenates the three sub-bands which represent high frequencies of X and X' into $[V, H, D]$, then the ℓ_1 loss is calculated.

When applying the WDIG framework to image translation and GAN Inversion, the operation in pixel space and wavelet space is consistent with the image style transfer task, but the object which is used to calculate the loss needs to be slightly adjusted. In image translation, the reconstruction loss is calculated by source domain image and generator reconstructed source domain image, and the structure loss is calculated by source domain image and translation image. In GAN Inversion, the reconstruction loss is calculated by the input image and the generator reconstruction image. Since there is no domain transformation, the structural loss is not used in GAN Inversion.

3.1 Pixel space loss

In the pixel space, Gaussian kernel function is applied to filter out high-frequency features and retain low-frequency information. As shown in Fig. 6, Gaussian blur is used to decompose the image into high-frequency and low-frequency parts with Gaussian kernels of different sizes. The first row is the low-frequency part, which mainly retains the texture, color and contour information of the image. The second row is the high-frequency part, which mainly retains the structure and identity information of the image. The larger the size of the Gaussian kernel is, the more blurred the

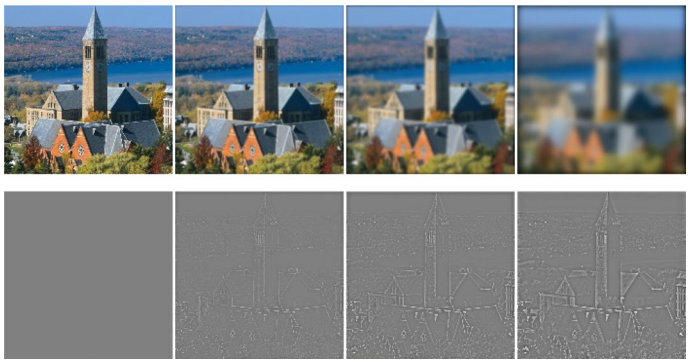

Fig. 6 Using Gaussian kernel to separate high and low frequencies

low-frequency image is and the clearer the high-frequency image is. In order to avoid distortion in the high-frequency and low-frequency regions, the size of Gaussian kernel k is set to 21 according to [7].

By using Gaussian blur, image x is converted into low-frequency image x_L and high-frequency image x_H, and the sizes of the three image are the same. Gaussian kernel is defined as Eq. (1).

$$k_\sigma[i,j] = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\left(\frac{i^2+j^2}{\sigma^2}\right)}, \tag{1}$$

where $[i,j]$ represents the spatial position of the image, and σ^2 represents the variance of the Gaussian which is proportional to the size of the Gaussian kernel. Applying the convolution of Gaussian kernel on the input image x, fuzzy low-frequency image x_L is obtained, as shown in Eq. (2).

$$x_L[i,j] = \sum_m \sum_n k[m,n] \otimes x[i+m, j+n], \tag{2}$$

where $[m, n]$ is the index of the 2D Gaussian kernel, $m, n \in \left[-\frac{k-1}{2}, \frac{k-1}{2}\right]$, $\otimes$ represents the convolution operation. After x is converted to the gray scale, the low-frequency information is subtracted to obtain the high-frequency x_H which represents the edge structure and identity of x, as shown in Eq. (3).

$$x_H = rgb2gray(x) - (rgb2gray(x))_L, \tag{3}$$

where *rgb2gray* represents the method of converting a color image to a gray scale image. Since the structural identity information is independent of color and illumination, the image is transferred to gray scale image to eliminate this information.

3.1.1 Reconstruction loss in the pixel space

The following ℓ_1 reconstruction loss term is employed for content image X and image X' reconstructed using autoencoder. The similarity between both low-frequency and high-frequency components of X and X' are enforced to enhance the ability of decoder to generate images. As shown in Eq. (4).

$$L_{\text{rec,pix}} = \left\|X_L - X'_L\right\|_1 + \left\|X_H - X'_H\right\|_1. \tag{4}$$

3.1.2 Structure loss in the pixel space

In order to make the stylized image X_Y retain the structure information of content image X, the ℓ_1 structure loss is employed to adjust the high-frequency component which represents the structure information, as shown in Eq. (5).

$$L_{\text{stru,pix}} = \|X_H - (X_Y)_H\|_1. \tag{5}$$

3.2 Wavelet space loss

In addition to the constraint of pixel space, the loss term of wavelet space is also introduced. Specifically, two-dimensional DWT is used to map image X from pixel space to wavelet space. As shown in Fig. 7, DWT has four convolutional filters, i.e. low-pass filter f_{LL} and high-pass filters f_{LH}, f_{HL}, f_{HH} to decompose x into four sub-band images, i.e. A, V, H and D. Taking Haar wavelet as an example, four filters are defined as

$$\begin{aligned} f_{LL} &= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad f_{LH} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \\ f_{HL} &= \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \quad f_{HH} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \end{aligned} \tag{6}$$

The operations of DWT are defined as

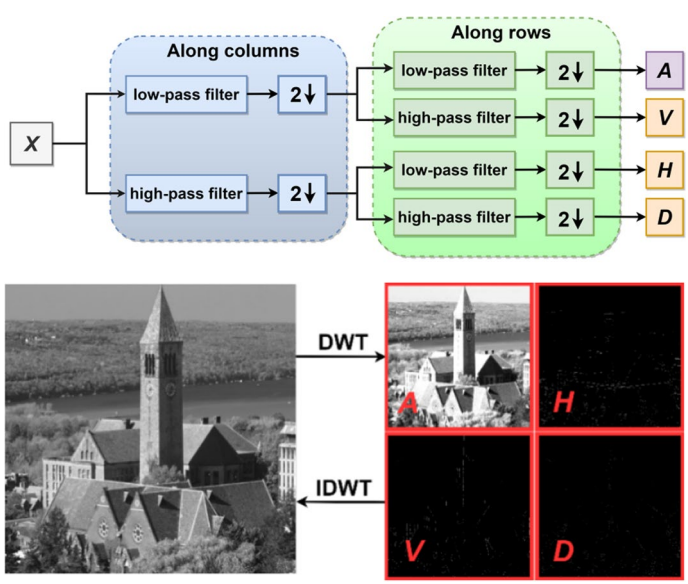

Fig. 7 Schematic diagram of two-dimensional wavelet transform for a image

$$
\begin{aligned}
[A, V, H, D] &= \mathrm{DWT}(x), \\
A &= (f_{LL} \otimes x) \downarrow_2 \\
V &= (f_{LH} \otimes x) \downarrow_2 \\
H &= (f_{HL} \otimes x) \downarrow_2 \\
D &= (f_{HH} \otimes x) \downarrow_2
\end{aligned} \tag{7}
$$

where $\otimes$ denotes convolution operator, $\downarrow_2$ represents the standard down-sampling operation with factor 2. In other words, DWT mathematically involves four fixed convolution filters with stride 2 to implement down-sampling operation.

3.2.1 Reconstruction loss in the wavelet space

The following ℓ_1 reconstruction loss of the wavelet space is employed. The four sub-band images of the content image X and the reconstructed image X' are spliced into matrix [A, V, H, D] respectively to calculate the loss, as shown in Eq. (8).

$$
L_{\mathrm{rec,wav}} = \left\| \mathrm{DWT}(X)_{[A,V,H,D]} - \mathrm{DWT}(X')_{[A,V,H,D]} \right\|_1. \tag{8}
$$

3.2.2 Structure loss in the wavelet space

Similar to Eq. (5), the ℓ_1 structure loss in the wavelet space is also employed. Only the sub-band images V, H, D of content image X and the stylized image X_Y represent high-frequency structure information are used, and V, H, D are spliced into matrix [V, H, D] to calculate the loss, as shown in Eq. (9).

$$
L_{\mathrm{stru,wav}} = \left\| DWT(X)_{[V,H,D]} - DWT(X_Y)_{[V,H,D]} \right\|_1. \tag{9}
$$

3.3 Overall loss

Considering all the above loss, the overall loss function is defined as follows.

$$
L_{\mathrm{WDIT}} = L_{\mathrm{org}} + \lambda_1 L_{\mathrm{rec,pix}} + \lambda_2 L_{\mathrm{stru,pix}} + \lambda_3 L_{\mathrm{rec,wav}} + \lambda_4 L_{\mathrm{stru,wav}}, \tag{10}
$$

where L_{org} is the loss function of the image generation model itself, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the weighting coefficients of the losses, which are set as $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$ in the experiment for simplicity. Gaussian kernel and wavelet transform are complementary for preserving frequency information. Gaussian kernel extracts local frequency features through convolution, while wavelet utilizes the information from all pixels to obtain the wavelet value for each spatial frequency, characterizing the frequency distribution globally. We will show in ablation study in Sects. 4.2 and 4.3 that the ability to preserve frequency information of both Gaussian kernel and wavelet transform is effective in improving the quality of generated images in image generation tasks.

4 Experiments

In this section, the performance of the proposed WDIG framework is evaluated on several classical image generation methods such as Image Style Transfer AdaIN, StyTr^2, Image Translation StarGANv2, and GAN Inversion Image2StyleGAN. The experimental environment is PyCharm and Pytorch, the graphics card model is Nvidia GeForce RTX

2080Ti with 12G video memory, and the processor model is Core i7-9700 K with 32 GB running memory.

Qualitative and quantitative experimental results show that the WDIG framework can preserve the structure in image style transfer task, retain the identity information of source domain in image translation task, and improve the embedding accuracy in GAN Inversion task.

4.1 Applying WDIG to image style transfer

Most image style transfer models use the framework of CNN-based autoencoder. Recently, Deng et al. [14] proposed StyTr^2, a Transformer-based style transfer model, which can better maintain the structure of stylized image compared with CNN-based models. In this subsection, FDIT framework and WDIG framework are applied to the classical CNN-based style transfer model AdaIN and Transformer-based style transfer model StyTr^2. In the experiments, models are trained using dataset MS-COCO [33] as content images and a dataset of paintings mostly collected from WikiArt [34] as style images. On the AdaIN, during training, the learning rate size is 1×10^{-4}, the aspect ratio of image is preserved and the size of the image is rescaled to 512×512 pixels, and then randomly crop to 256×256 pixels. A batch size of 8 content-style image pairs for 160 k iterations. Since StyTr^2 is based on Transformer structure, the number of parameters is large and amounts of the video memory is required. To reduce video memory usage, during training, the learning rate size starts at 5×10^{-4} and then decays slowly. The aspect ratio of the image is preserved and the size of the image is rescaled to 512×512 pixels, and then randomly crop to 128×128 pixels. A batch size of 8 content-style image pairs for 160 k iterations.

Style transfer effects are shown in Fig. 8. As shown in Fig. 8c, the structure distortion is serious in the stylized images of AdaIN. In Fig. 8d, applying FDIT to AdaIN, the structure of stylized images is maintained, but the style rendering is not sufficient. In Fig. 8e, applying WDIG to AdaIN, the structure of stylized images is maintained and the style rendering is more sufficient, especially the part of the sky framed in the second, third and fifth rows. In Fig. 8f, StyTr^2 is based on the structure of relational modeling Transformer and has strong ability of feature representation. It can avoid the details loss in the process of feature extraction and can well preserve the structure of stylized images. However, there are artifacts in some areas of stylized images, such as the sea surface in the third row and the sky in the fifth row. In Fig. 8g, applying FDIT to StyTr^2, the artifacts are alleviated, but the style rendering is not sufficient. In Fig. 8h, applying WDIG to StyTr^2, the artifacts are alleviated and the style rendering is more sufficient, such as the face area in the first row and the sky in the fourth row.

In this part, quantitative evaluations are adopted. The goal of the style transfer is that the stylized image should be consistent with the content image in terms of semantic structure. Structural similarity (SSIM) is a metric to measure the similarity of two images. The closer the SSIM value is to 1, the more similar the structure is. Peak signal-to-noise ratio (PSNR) is often used as a matric for signal reconstruction quality in image compression and other fields. The higher the value, the better the reconstructed image quality. However, due to the subjectivity of human vision, sometimes images with low PSNR values may have better visual effects than images with high PSNR values.

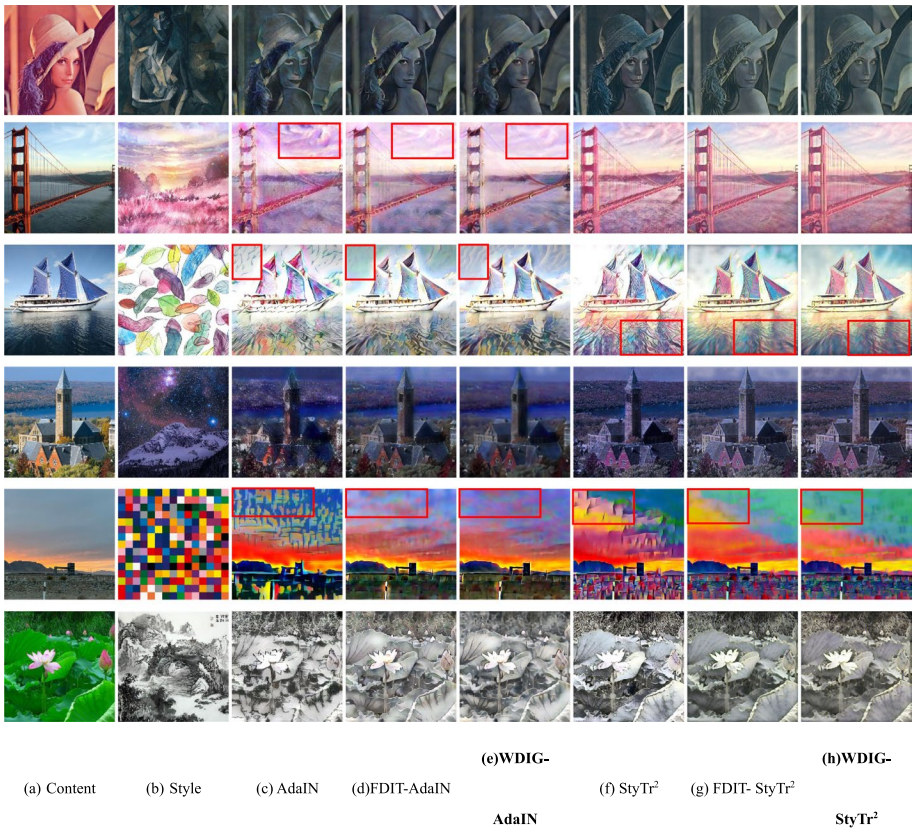

Fig. 8 Comparison of Style transfer effects. **a** Content images. **b** Style images. **c** Style transfer effects of AdaIN. **d** Style transfer effects of FDIT-AdaIN. **e** Style transfer effects of WDIG-AdaIN. **f** Style transfer effects of StyTr^2. **g** Style transfer effects of FDIT-StyTr^2. **h** Style transfer effects of WDIG-StyTr^2

Table 1 Structural similarity SSIM, PSNR, style transfer time comparison between stylized images obtained by AdaIN and content images

Methods	SSIM		PSNR		Time (s)
AdaIN	0.281		10.8		0.056
FDIT-AdaIN	0.457	62.6%↑	**13.1**	**21.3%↑**	0.056
WDIG-AdaIN	**0.463**	**64.8%↑**	12.9	19.4%↑	0.057

The best results are highlighted in bold

Table 2 Structural similarity SSIM, PSNR, style transfer time comparison between stylized images obtained by StyTr^2 and content images

Methods	SSIM		PSNR		Time (s)
StyTr^2	0.415		11.4		0.552
FDIT-StyTr^2	**0.624**	**50.4%↑**	**13.8**	**21.1%↑**	0.556
WDIG-StyTr^2	0.600	44.6%↑	13.7	20.2%↑	0.547

The best results are highlighted in bold

Therefore, higher PSNR value as well as that the stylized image with artistic beauty is preferred in Style Transfer. In this paper, 10 content images and 10 style images are selected, and 100 stylized images are obtained by using the models trained with different frameworks. The average SSIM and PSNR are calculated on the corresponding images, and the stylized execution time is also calculated. As shown in Tables 1 and 2, for the style transfer method AdaIN, WDIG increases SSIM by 64.8% and PSNR by 19.4%. For the style transfer method StyTr^2, WDIG increases SSIM by 44.6% and PSNR by 20.2%. The improvement of WDIG in SSIM is higher than that of FDIT, which indicates that the structures of stylized images produced by WDIG are more similar to the structure of content images. WDIG renders the style more fully, which causes stylized images to have more artistic beauty, however, also with more noise. Therefore, the improvement in PSNR of WDIG is slightly lower than that of FDIT. It can also be seen from Tables 1 and 2 that FDIT and WDIG could provide improved performance without further increasing the style transfer time.

4.2 Applying WDIG to image translation

StarGANv2 is an advanced image translation model, which generates translation images guided by target domain images or latent codes. The proposed WDIG can be used to optimize the performance of StarGANv2. In this subsection, dataset CeleBA-HQ [35] is employed to train StarGANv2 with the FDIT framework and the WDIG framework. During training, the learning rate size is 1×10^{-4}, the aspect ratio of the image is preserved and the size of the image is rescaled to 512×512 pixels, and then randomly crop to 128×128 pixels.

A batch size of 6 content-style image pairs for 30 k iterations. The goal of image translation is to keep the identity attributes of the image in the source domain such as gender and hair style, and transform the semantic attributes of the image in the target domain

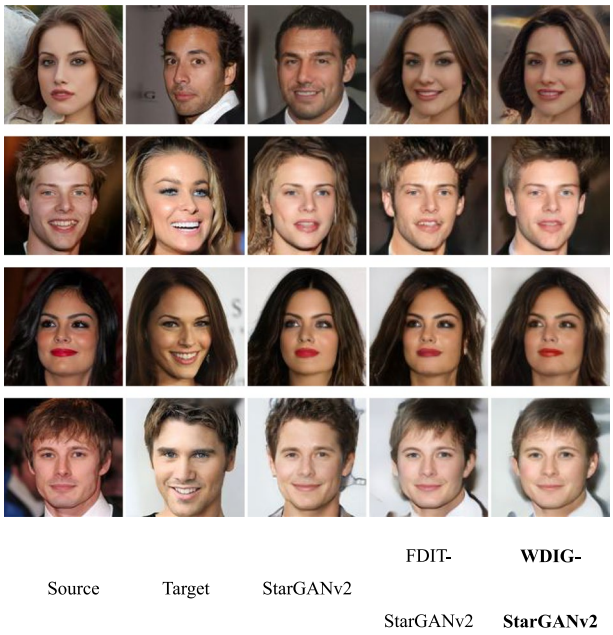

Fig. 9 Comparison of image translation effects on CelebA-HQ [28]

such as expression, skin color and illumination. Image translation effects are shown in Fig. 9. The identity attribute of source domain cannot be maintained when StarGANv2 is applied. FDIT and WDIG can strictly preserve identity features while modifying semantic attributes of faces. As shown in the first and second rows, in the translated images of StarGANv2, the gender changes, while both FDIT and WDIG can ensure that the gender remains unchanged after translation. As shown in the fourth row, StarGANv2 changes the hairstyle, while both FDIT and WDIG can ensure that the hairstyle remains the same. Compared with FDIT, the image translated by WDIG is more realistic. In the second row, the structure of the mouth corner in the face image translated by FDIT is distorted, while WDIG can avoid this problem.

In this part, quantitative evaluations are calculated to compare the translation image quality produced by different methods. Frechet Inception Distance (FID) metric is widely used in image translation. It is calculated using the feature vector output from the last average pooling layer of inception-V3 pre-trained on ImageNet to measure the difference between two groups of images. The lower the FID is, the more similar the two groups of images are, that is, the closer the identity attributes are. The perceptual distance LPIPS is calculated using the feature vectors extracted from the AlexNet pre-trained on ImageNet to measure the perceptual similarity of the two groups of images, which is consistent with human perception. The lower the LPIPS is, the more perceptually similar the generated image is to the input image, representing the closer the identity attribute is. In this paper, the loss of pixel space, wavelet space and Fourier space in FDIT are individually added to the StarGANv2 loss function for training. WDIG framework and FDIT framework are also used to train StarGANv2, and different image translation models are obtained. 1000 source domain images and 10 target domain images are selected, and 10,000 translated images are obtained by using different models. Then the average FID value and LPIPS value are calculated on the corresponding images. As shown in Table 3, the loss of pixel space, Fourier space or wavelet space can improve StarGANv2 on these metrics, no matter the image is translated by style coding guided by latent codes or target domain images. The FID is improved by 16.3% and the LPIPS is improved by 33.3% by using WDIG, which indicates that the image identity details generated by the WDIG are well preserved. In terms of FID metrics, WDIG provides the best performance. In terms of LPIPS, FDIT provides better performance than WDIG. The possible reason is that WDIG destroys some identity attributes while converting more semantic attributes. However, WDIG does not need to manually set the mask

Table 3 The FID and LPIPS comparisons between source domain images and the translated images obtained by different methods

Methods	Latent codes guided		Reference guided	
	FID	LPIPS	FID	LPIPS
StarGANv2	16.1	0.320	18.8	0.275
StarGANv2 + pixel space	14.2	0.211	16.1	0.183
StarGANv2 + Fourier space	14.9	0.242	16.0	0.196
StarGANv2 + wavelet space	14.6	0.233	15.7	0.194
FDIT	14.2 (11.8%↑)	**0.210 (34.4%↑)**	16.3 (13.3%↑)	**0.179 (34.9%↑)**
WDIG	**13.5 (16.1%↑)**	0.212 (33.8%↑)	**15.7 (16.5%↑)**	0.185 (32.7%↑)

The best results are highlighted in bold

radius to separate the high and low frequencies, which is more simple and stable in the applications.

4.3 Applying WDIG to GAN inversion

Image2StyleGAN is a classical GAN Inversion method based on optimizing latent codes, which can embed the user-specified image into latent codes of GAN latent space. The accuracy of the embedding is verified by the similarity between the reconstructed image generated by latent codes and the input real image. In this subsection, WDIG framework and FDIT framework are applied to Image2StyleGAN. During training, the batch size is set to 1, the learning rate size is 1×10^{-2}, image size is set to 1024×1024, and the number of iterations is set to 5000. The GAN Inversion effects are shown in Fig. 10. It can be seen that WDIG can better preserve the details of the overall structure and color distribution of the input image, and the reconstructed image is more similar to the input image. While for images in the fourth row, the animation face reconstructed by Image2StyleGAN is inconsistent with the input image in color, and the area in the left eye is distorted and unrecognizable. After applying the FDIT, the color of the animation face is roughly the same as the input image, but the area in the left eye and hair tip still have some distortion and blurring. After applying the WDIG, the structural details of the animation face are well reconstructed and highly consistent with the input image.

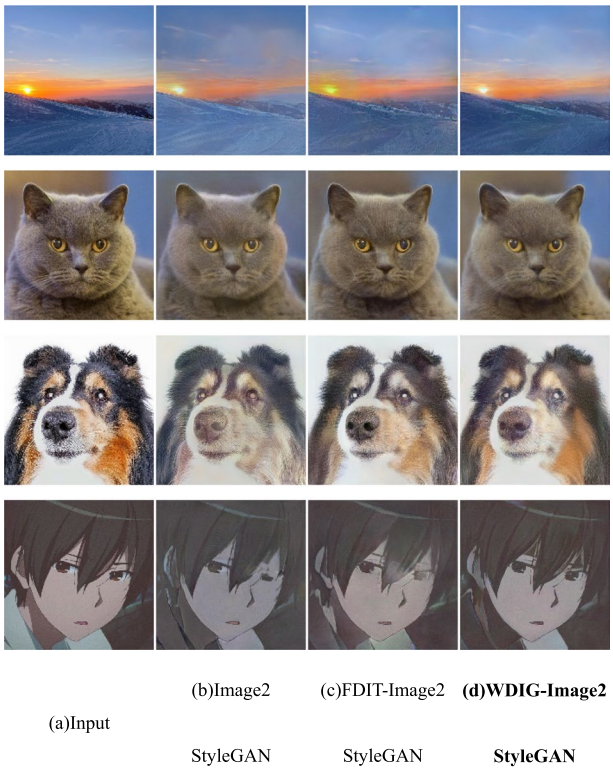

Fig. 10 Comparison of GAN Inversion effects. **a** High-resolution real image. **b** Reconstruct images of Image2StyleGAN. **c** reconstruct images of FDIT-Image2StyleGAN. **d** reconstruct images of WDIG-Image2StyleGAN

Table 4 Structural similarity, peak signal-to-noise ratio, MSE and MAE comparison between the reconstructed images of different models and the input images

Methods	SSIM	PSNR	MSE	MAE
Image2StyleGAN	0.686	23.2	0.0058	0.0539
FDIT-Image2StyleGAN	0.691 (0.7%↑)	24.1 (3.9%↑)	0.0044 (24.1%↑)	0.0445 (17.4%↑)
WDIG-Image2StyleGAN	**0.696 (1.5%↑)**	**24.2 (4.3%↑)**	**0.0044 (24.1%↑)**	**0.0443 (17.8%↑)**

The best results are highlighted in bold

Table 5 Ablation experiments in pixel space, Fourier space, wavelet space

Loss Terms added in Image2StyleGAN			FID
Pixel space	Fourier space	Wavelet space	
×	×	×	28.23
✓	×	×	27.80 (1.5%↑)
×	✓	×	28.03 (0.7%↑)
×	×	✓	27.66 (2.0%↑)
✓	✓	×	27.22 (3.6%↑)
✓	×	✓	**25.49 (9.7%↑)**

The best results are highlighted in bold

In this subsection, quantitative metrics are calculated to compare the similarity between the reconstructed image and the input image produced by different frameworks applied to Image2StyleGAN. 50 pairs of 1024×1024 high-resolution input images and reconstructed images are selected, and the average SSIM, PSNR, MSE and MAE are calculated on the corresponding images. The MSE and MAE are commonly used in the field of image reconstruction to measure the difference between the reconstructed image and the original image. As shown in Table 4, in terms of SSIM, PSNR, MSE and MAE, WDIG outperforms Image2StyleGAN and FDIT. Compared with the baseline model Image2StyleGAN, WDIG improved the SSIM by 1.5%, PSNR by 4.3%, MSE by 24.1%, and MAE by 17.8%. The improvements of these metrics indicate that the reconstructed images of WDIG are more similar to the input images and WDIG improves embedding accuracy. The losses of pixel space, wavelet space and Fourier space are separately added to the Image2StyleGAN model for training, and the average FID is calculated on the reconstructed image and the input image. As shown in Table 5, the FID can be improved by the addition of the loss of pixel space, wavelet space and Fourier space, but only when the pixel space loss and the wavelet space loss are used jointly, the FID increases the most, which is 9.7% higher than the baseline model Image2StyleGAN, showing the effectiveness of WDIG.

5 Conclusion

In order to solve the problem of image quality deficiency caused by the details loss in image generation tasks, a wavelet domain image generation framework WDIG is proposed based on the complementarity of Gaussian kernel and wavelet transform to preserve image frequency information. Loss functions are constructed in pixel space and

wavelet space. The separated high frequencies and low frequencies are used to calculate the loss to measure missing details of the generated image. The training of the model is more accurately constrained, so as to accurately optimize the model and improve the quality of image generation. Qualitative and quantitative experimental results on image generation tasks such as image style transfer, image translation, GAN Inversion are provided to show that the WDIG framework can effectively preserve the details of the generated image. In the image style transfer task, the structure of the content image can be retained in stylized images; In the image translation task, the identity information of the source domain image can be preserved in the translated image, and the generated image is more realistic; In the GAN Inversion task, the difference between the generated reconstructed image and the input image is smaller, and the accuracy of latent codes embedding is improved.

In the future work, the proposed WDIG framework needs to be trained with large data sets to provide more effective performance. Moreover, the applications of the proposed WDIG in other image generations are also worth exploring.

Abbreviations

WDIG	Wavelet domain image generation
DCT	Discrete cosine transform
FDIT	Frequency domain image translation
AAMS	Attention-aware multi-stroke
AdaAttn	Adaptive attention normalization
$StyTr^2$	Style transfer transformer framework
CMD	Central moment discrepancy
AdaIN	Adaptive instance normalization
WCT	Whitening and coloring transforms
CycleGAN	Cycle-consistent adversarial networks
MUNIT	Multimodal unsupervised image-to-image translation
SWAE	Swapping autoencoder
GAN	Generative adversarial networks
StyleGAN	Style-based generator architecture for generative adversarial networks
FFHQ	Flickr faces HQ
DWT	Discrete wavelet transform
MS-COCO	Microsoft common objects in context
SSIM	Structural similarity
PSNR	Peak signal-to-noise ratio
FID	Frechet inception distance
LPIPS	Learned perceptual image patch similarity
MSE	Mean squared error
MAE	Mean absolute error

Acknowledgements
This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61801159 and 61571174.

Author contributions
QZ: conceptualization and original draft preparation, XL: simulation and revision, JS: verification and revision, HB: methodology and revision.

Funding
National Natural Science Foundation of China under Grant Nos. 61801159 and 61571174.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare no conflict of interest.

Received: 3 November 2022 Accepted: 12 June 2023
Published online: 19 June 2023

References

1. C.E. Shannon, Coding theorems for a discrete source with a fidelity criterion. IRE Natl. Convent. Rec. **4**, 142–163 (1959)
2. Zhou K, Yang Y X, Hospedales T, et al. Deep Domain-Adversarial Image Generation for Domain Generalisation, in *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), pp. 13025–13032. https://doi.org/10.1609/aaai.v34i07.7003.
3. X. Jiang, F.R. Yu, T. Song et al., Blockchain-enabled cross-domain object detection for autonomous driving: a model sharing approach. IEEE Internet Things J. **7**(5), 3681–3692 (2020)
4. K. Xu, M. Qin, F. Sun, et al. Learning in the frequency domain, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Seattle, WA, 2020), pp. 1740–1749. https://doi.org/10.1109/CVPR42600.2020.00181.
5. L. Jiang, B. Dai, W. Wu, et al. Focal frequency loss for image reconstruction and synthesis, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, Montreal Canada, 2021), pp. 13919–13929. https://doi.org/10.1109/ICCV48922.2021.01366.
6. W. Xie, D. Song, C. Xu, et al. Learning frequency-aware dynamic network for efficient super-resolution, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, Montreal, Canada, 2021), pp. 4308–4317. https://doi.org/10.1109/ICCV48922.2021.00427.
7. M. Cai, H. Zhang, H. Huang, et al. Frequency domain image translation: more photo-realistic, better identity-preserving, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, Montreal, Canada, 2021), pp. 13930–13940. https://doi.org/10.1109/ICCV48922.2021.01367.
8. X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, Venice, Italy, 2017), pp. 1501–1510. https://doi.org/10.1109/ICCV.2017.167
9. Y. Li, C. Fang, J. Yang, Z. Wang et al., *Universal Style Transfer Via Feature Transforms, in Advances in Neural Information Processing Systems* (MIT Press, Long Beach, CA, USA, 2017), pp.386–396
10. T. Wang, Y. Zhang, Y. Fan, et al. High-fidelity Gan inversion for image attribute editing, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New Orleans, LA, USA, 2022), pp. 11379–11388. https://doi.org/10.1109/CVPR52688.2022.01109
11. Y. Yao, J.Q. Ren, X.S. Xie, et al. Attention-aware multi-stroke style transfer, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Long Beach, CA, USA, 2019), pp. 1467–1475. https://doi.org/10.1109/CVPR.2019.00156
12. S.H. Liu, T.W. Lin, D.L. He, et al. Adaattn: revisit attention mechanism in arbitrary neural style transfer, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, Montreal Canada, 2021), pp. 6649–6658. https://doi.org/10.1109/ICCV48922.2021.00658
13. Yu. Shen, Y. Qian, C. Xiaopeng et al., Structure refinement neural style transfer. J. Electron. Inf. Technol. **43**(08), 2361–2369 (2021). https://doi.org/10.11999/JEIT200211
14. Y. Deng, F. Tang, X. Pan, et al. StyTr^2: Image Style Transfer with Transformers [OL]. https://arxiv.org/abs/2105.14576.
15. N. Kalischek, J.D. Wegner, K. Schindler, In the light of feature distributions: moment matching for neural style transfer, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Nashville, TN, USA, 2021), pp. 9382–9391. https://doi.org/10.1109/CVPR46437.2021.00926
16. P. Isola, J.Y. Zhu, T. Zhou, et al. Image-to-image translation with conditional adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Honolulu, HI, USA, 2017), pp. 1125–1134. https://doi.org/10.1109/CVPR.2017.632
17. J. Y. Zhu, T. Park, P. Isola, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, Venice, Italy, 2017). https://doi.org/10.1109/ICCV.2017.244
18. X. Huang, M.Y. Liu, S. Belongie, et al. Multimodal unsupervised image-to-image translation, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 172–189
19. T. Park, J.Y. Zhu, O. Wang et al., Swapping autoencoder for deep image manipulation. Adv. Neural. Inf. Process. Syst. **33**, 7198–7211 (2020)
20. Y. Choi, Y. Uh, J. Yoo, et al. StarGAN v2: Diverse image synthesis for multiple domains, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Seattle, WA, USA, 2020), pp. 8188–8197. https://doi.org/10.1109/CVPR42600.2020.00821.
21. T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Long Beach, CA, USA, 2019), pp. 4401–4410. https://doi.org/10.1109/TPAMI.2020.2970919
22. R. Abdal, Y. Qin, P. Wonka, Image2StyleGAN: how to embed images into the StyleGAN latent space? in*Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (IEEE, Seoul, Korea (South), 2019), pp. 4432–4441. https://doi.org/10.1109/ICCV.2019.00453

23. E. Richardson, Y. Alaluf, O. Patashnik, et al. Encoding in style: a stylegan encoder for image-to-image translation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 2287–2296
24. S. Guan, Y. Tai, B. Ni, et al. Collaborative Learning for Faster Stylegan Embedding [EB/OL]. https://arxiv.org/abs/2007.01758 (2020)
25. D. Zhou, H. Zhang, Q. Li et al., COutfitGAN: learning to synthesize compatible outfits supervised by silhouette masks and fashion styles. IEEE Trans. Multimed. (2022). https://doi.org/10.1109/TMM.2022.3185894
26. A.E. Mahdaoui, A. Ouahabi, M.S. Moulay, Image denoising using a compressive sensing approach based on regularization constraints. Sensors **22**(6), 2199 (2022)
27. H. Zhou, H. Xiong, C. Li et al., Single image dehazing based on weighted variational regularized model. IEICE Trans. Inf. Syst. **104**(7), 961–969 (2021)
28. P. Liu, H. Zhang, W. Lian et al., Multi-level wavelet convolutional neural networks. IEEE Access **7**, 74973–74985 (2019). https://doi.org/10.1109/ACCESS.2019.2921451
29. S. Xue, W. Qiu, F. Liu et al., Wavelet-based residual attention network for image super-resolution. Neurocomputing **382**, 116–126 (2020). https://doi.org/10.1016/j.neucom.2019.11.044
30. J. Xin, J. Li, X. Jiang et al., Wavelet-based dual recursive network for image super-resolution. IEEE Transactions on Neural Networks and Learning Systems (2020). https://doi.org/10.1109/TNNLS.2020.3028688
31. W. Ma, Z. Pan, J. Guo et al., Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive res-net. IEEE Trans. Geosci. Remote Sens. **57**, 3512–3527 (2019). https://doi.org/10.1109/TGRS.2018.2885506
32. T. Guo, H.S. Mousavi, T.H. Vu, et al. Deep wavelet prediction for image super-resolution, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (IEEE, Honolulu, HI, USA, 2017), pp. 1100–1109. https://doi.org/10.1109/CVPRW.2017.148
33. T.Y. Lin, M. Maire, S. Belongie, et al. Microsoft COCO: common objects in context, in *Proceedings of the European Conference on Computer Vision (ECCV)* (IEEE, Zurich, Switzerland, 2014), pp. 740–755
34. K. Nichol, Painter by Numbers, Wikiart. https://www.kaggle.com/c/painter-by-numbers/. Assessed 20 Nov 2021.
35. T. Karras, T. Aila, S. Laine, et al. Progressive growing of GANs for improved quality, stability, and variation [EB/OL]. arXiv:1710.10196 (2018)

Publisher's Note