EURASIP Journal on Advances
in Signal Processing

RESEARCH Open Access

# MT-GAN: toward realistic image composition based on spatial features

Xiang Li[1,2], Guowei Teng[1], Ping An[1*] and Hai-yan Yao[1]

*Correspondence:
anping@shu.edu.cn

[1] School of Communication and Information Engineering, Shanghai University, Shanghai, China
[2] Anyang Institute of Technology, Anyang, China

**Abstract**

The purpose of image composition is to combine the visual elements of different natural images to produce a natural image. The performance of most existing image composition methods drops significantly when they solve multiple issues, such as image harmonization, image blending, shadow generation, object placement, and spatial transformation. To address this problem, we propose a multitask GAN for image compositing based on spatial features, aiming to simultaneously address the geometric and appearance inconsistency. We use three related learning objective functions to train the network. Moreover, a new dataset including 7756 images with RoI region annotations is contributed to help evaluate the multitask image compositing results. Extensive experiments demonstrate that our proposed method is effective on our dataset.

**Keywords:** Image composition, Generative adversarial network, Multitask, Spatial features

## 1 Introduction

Image composition is the task of combining regions from different images to form a realistic composite image. It has a wide range of applications such as augmented reality, artistic creation, and e-commerce. However, due to the different resources of the source and target images, the geometric inconsistency and appearance inconsistency will cause the composite image to look disharmonious, which will lead to the degradation of its image quality.

There are many problems to be solved for geometric inconsistency and appearance inconsistency. As every problem could be very challenging, many efforts have been proposed to solve them. Image harmonization [1–3] aims to make the composite image look more harmonious. The usual strategy is to adjust the color and illumination statistics of the source object according to the composite target image to make it compatible with the composite target image. For appearance inconsistency, image blending [4, 5] aims to solve the issue of the discordant boundary between foreground and background. Shadow or reflection generation [6–8] targets generating plausible shadow or reflection for the source target based on the context. For geometric inconsistency, many methods have been proposed for object placement and spatial transformation. Object placement [9, 10] aims to translate and resize the source image, and spatial transformation [12, 13] aims to transform the foreground on a more complicated level such as perspective transformation.

Image composition has so many issues to be solved that some researchers focus on only one or two issues. We hope to solve multiple issues simultaneously and generate realistic and plausible images. At present, various GAN-based methods based on the original model [14] have been widely used in style transfer, semantic segmentation, image superresolution, image synthesis, etc. For image composition, many GAN-based approaches have been proposed in object placement [12, 15], image blending [4], source object generation [16], and image harmonization [17]. Inspired by the successful research, we propose a deep adversarial model, multitask GAN (MT-GAN) for image composition, which can generate realistic and plausible composite images, and address the geometric and appearance inconsistency simultaneously. Moreover, we introduce a geometric consistency network for geometric alignment and a spatial feature extraction network for performance enhancement.

As far as we know, there is no unified benchmark dataset for image compositing. The reason is that it is easy to obtain a large number of composite images by simply superimposing the source image on the target image, but these arbitrary composite images look unreal. Since manually adjusting the compositing results to make them look like the natural image is labor-intensive, time-consuming, and unreliable, many works use segmentation datasets [18–21] with annotated masks. Some efforts construct real-world dataset [22], synthetic real dataset [1], and rendered dataset [6, 23] for image harmonization or shadow generation. Therefore, we construct a multitask image compositing dataset where both the background image and the ground truth image are derived from natural images

In summary, the prime contributions of our work are as follows:

- We propose an effective and robust multitask GAN model trained for image compositing using spatial features. It can solve the geometric consistency and appearance consistency issues simultaneously. The proposed MT-GAN contains three different but related learning objectives that can alternatively assist each other to obtain better compositing results. A geometric consistency network and a spatial feature extraction network jointly assist the adversarial learning backbone for performance enhancement.
- We contribute a new dataset with RoI region annotations that can evaluate the multitask GAN. As we know, it is the largest dataset for evaluating multitask image synthesis algorithms. It can provide the ground truth of composite images from natural images for geometric consistency research, which is lacking in the existing datasets.

The rest of this paper is organized as follows. Section 2 introduces the related work. Our materials and methods are proposed in Sect. 3. In Sect. 4, we introduce the experimental results of the dataset. Finally, the conclusion is discussed in Sect. 5.

## 2 Related work

Methods for image composition mainly address two issues: geometric consistency and appearance consistency.

### 2.1 Geometric consistency

Geometric consistency aims to address the irrationality of the size, location, and shape of the source object in the task of image composition. It is an important issue for image compositing, and many works have been proposed on this task. STN [24] applies the perspective transformation to a deep learning network, so that the network can learn the geometric variations parameters of perspective transformation. Since it is difficult to obtain the corresponding ground truth of the composite image required in training, many studies [12, 25, 33] combine STN with a GAN network, and use the discriminator network in GAN to identify whether the composite image is real. Following the aforementioned methods, we propose a generative-based approach to address the geometric consistency.

### 2.2 Appearance consistency

The goal of appearance consistency is to pursue color, texture, and context consistency. There are many efforts for this task, for example, image harmonization, image blending, and shadow generation. Some image blending works [4] [148,170] utilize the deep learning network that can smooth the transition over the boundary and reduce the color and texture discrepancy between the source object and target image. For harmonization, Tsai et al. [3] proposed a CNN network with a context decoder and a semantic decoder to produce harmonized images. Cun et al. [2] design a spatial-separated attention module, aiming to learn regional appearance differences for image harmonization. Dovenet [1] proposes a domain verification discriminator in a basic GAN network. Shadow GAN [8] uses a local discriminator and a global discriminator to generate the shadow, and ARShdow-Gan [6] utilizes an attention-guided residual network for object shadow and reflection generation. SSN [7] proposes an interactive soft network to generate soft shadows that can be adjusted by users. So we propose a generative network to achieve a composite image with relatively consistent color, texture, and realistic shadow/ reflection simultaneously.

However, most of the geometric consistency works are more focused on variation, ignoring the appearance consistency issues such as color and lighting. Existing appearance consistency methods attempt to solve one issue or multiple issues separately. Each research direction diverges from image composition to solve different issues, without a unified network to address all the issues in image composition. This may increase the difficulty to composite realistic images. To sum up, there are many issues to be solved to obtain a realistic composite image. Thus, we propose an effective and robust multitask GAN model trained for image compositing that can solve the geometric and appearance inconsistency simultaneously. Moreover, we conduct a multitask dataset that can provide the ground truth of composite images from natural images for geometric and appearance consistency research, which is lacking in the existing datasets.

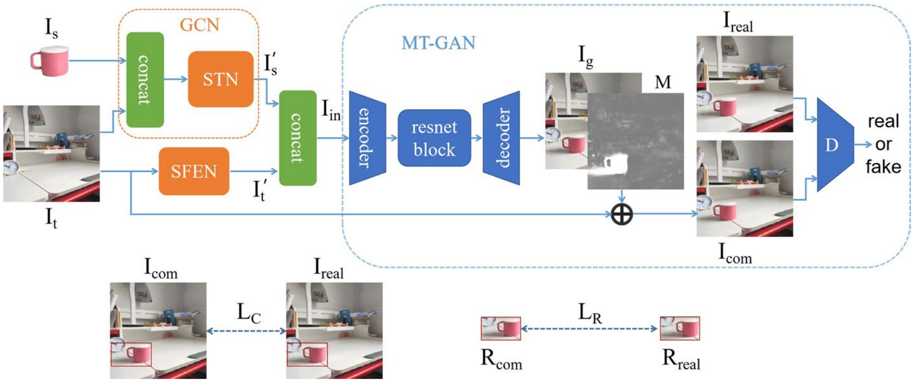

**Fig. 1** The architecture of our multitask GAN for image composition based on spatial features

**Table 1** Related mathematical symbols and their representation

| Symbol | Representation |
|---|---|
| $I_s$ | Source object |
| $I_t$ | Target image |
| $I_{con}$ | Concatenated $I_s$ and $I_t$ |
| $I_s'$ | $I_s$ geometrically aligned with $I_t$ |
| $I_t'$ | Output of spatial feature extraction network (SFEN) |
| $I_{in}$ | Input data concatenating $I_t'$ and $I_s'$ |
| $I_g$ | Output data of generator |
| $I_{com}$ | Compositing image |
| $I_{real}$ | Ground truth image |
| $I_g$ | Output data of generator |
| $L_{final}$ | Final objective |
| $L_A$ | Adversarial loss |
| $L_C$ | Compositing loss |
| $L_R$ | RoI consintency loss |

## 3 Materials and methods

The proposed multitask compositing network contains three components: the geometric consistency network (GCN) is designed to align source objects with target images for geometric consistency. The spatial feature extraction network (SFEN) changes target data from 3 channels to 1 channel, reducing the amount of data fed to the generated network and improving the training efficiency. On the other hand, in the input data of the network, the proportion of the source object and target image data increases, which improves the effectiveness of the source object. The multitask GAN for image composition (MT-GAN) takes the source object and target image together as input and generates realistic compositing results. Figure 1 shows our proposed framework. The details of our framework will be introduced in the following sections. In Table 1, we list the symbols used in this paper and their representations.

### 3.1 Geometric consistency network

Firstly, before adjusting the color of the source object, the size and position are adjusted by the geometric consistency network (GCN) to align with the target object. We consider appropriately deforming the foreground so that $I_s$ can learn the deformation of $I_t$. Therefore, the subsequent network can generate results with better geometric consistency with its assistance. Given an RGB target image $I_t \in \mathbb{R}^{H \times W \times 3}$ and a source image $I_s \in \mathbb{R}^{H \times W \times 3}$, we form the input $I_{con} \in \mathbb{R}^{H \times W \times 6}$ by concatenating them, where $H$ and $W$ are image dimensions. Then, we use the spatial transformer network (STN) [24] to achieve a source image $I_s^{'}$ geometrically aligned with the target image. The aligned $I_s^{'} = GCN(I_s)$ is prepared to be fed to the subsequent network for compositing.

### 3.2 Spatial feature extraction network

We introduce a spatial feature extraction network (SFEN) to enhance performance. The source object and target image are both $3 \times 256 \times 256$. The target image provides spatial information for the source object. On this basis, it helps the network generate shadows and adjust the color of the generated foreground. The spatial position and size of the generated foreground are more important than color and shadow. Thus, we extract the spatial features of the target images before feeding the data to the backbone network. We use a lightweight network to convert the original three channels into one channel. Then, the target data becomes $1 \times 256 \times 256$. We express this process as $I_t^{'} = SFEN(I_t)$. Afterward, we concatenate the target data of channel 1 with the source data of channel 3 and feed them to the subsequent GAN network. In this way, on the one hand, the SFEN can extract the spatial information of the target images. On the other hand, the amount of background data is reduced by 2/3, which makes the backbone network pay more attention to foreground data and generate better composite images. By increasing the proportion of source data, the network can focus on the source object during training. During the training process, a clearer source object can be generated with fewer epochs after increasing the proportion of source data.

**Network configuration** Before feeding to the network, the input data size is adjusted to $256 \times 256$. The SFEN contains two convolution layers and two fractionally stridden convolution layers(fsconv). The details of the architecture are shown as follows: conv1(3, 64, 4, 2, 1)-conv2(64, 128, 4, 2, 1)-fsconv1(128, 64, 4, 2, 1)-fsconv2(64, 1, 4, 2, 1). The number in each parenthesis indicates the channel of input and output, kernel size, stride, and padding. Each convolution layer is filled with padding to ensure the correct output size. All convolutional layers are followed by a batch-normalization layer and a LeakyReLU layer. The final fractionally stridden convolution layer uses a Tanh layer as an activation function.

### 3.3 Muti-task GAN for image composition based on spatial features

We utilize a generative adversarial network (GAN) as our backbone to learn the function from the original source object and target image to a realistic compositing image.

We concatenate the $I_s^{'}$ and the $I_t^{'}$ to the generator G, and generate the composite image $I_{com}$, where the input of GAN is $4 \times 256 \times 256$. The first three channels are information related to the source image, and the last channel is to the target image. In order to highlight the importance of the source data during training, the output of GAN is 4 channels, including three channels of the generated image $I_g$, and another channel of mask data $M$. Mask $M$ is used to set the background to 0 and strengthen the source image with corresponding shadows. The compositing image $I_{com}$ can be achieved by the following equation. We do not set a specific ground truth for Mask $M$, the appropriate $M$ can be obtained by back-propagation. The final compositing image can be expressed as:

$$I_{com} = M \times I_g + (1 - M) \times I_t \tag{1}$$

Our architecture of GAN is adopted by Zhu et al. [28] and it is an encoder–transformer–decoder architecture. The encoder is used to extract the feature of the input and consists of three convolution layers. The transformer aims to transform the features into another distribution and contribute nine residual blocks [29]. Using the residual structure can preserve both the original input attributes and their size and shape. We also use instance normalization [31], following the existing work [28, 30]. In addition, the decoder restores the resolution to the same scale as the original input through three fractionally stridden convolutions.

The discriminator is an important part of the GAN network. In this task, the discriminator is mainly used to identify the authenticity of its input. Compared with the generator network, the network structure of the discriminator is generally much simpler. We used five convolution layers to build the discriminator network.

#### *3.3.1 Adversarial loss*

We apply an adversarial loss to the network which can be expressed as:

$$\begin{aligned} L_A = {} & \mathbb{E}_{x \sim Pdata(x)}[log(1 - D(G(x)))] \\ & + \mathbb{E}_{y \sim Pdata(y)}[logD(y)] \end{aligned} \tag{2}$$

where $x = I_{in}$ and $y = I_{real}$ are samples drawn from the fake data and the real data.

#### *3.3.2 Compositing loss*

The background of the whole image is often much larger than the foreground area so that the network will be greatly affected by the background during training. In order to make the network focus on the foreground region in the whole generation process, we add channel $M$ as the mask of the other three channels to the output of the generative network G. That is, the output of G has four channels, three of which are generated image $I_g$. Another channel is the mask channel $M$. In addition, this can reduce the error of generating results obviously, because the background generated by the G network will also bring deviation. Through the above weighting process, the original image background can be directly used, which can reduce the impact of the generated background on the composite image. At the same time, it ensures the harmony of the composite image.

$$L_C = MSE(I_{com}, I_{real}). \tag{3}$$

#### *3.3.3 RoI consistency loss*

To further strengthen the network's focus on the foreground area, we also add a loss function in the region of interest (RoI). We have labeled the source object and its shadow area in the ground truth with a bounding box. The labeled area is the area where the ground truth is different from the target image, which is also the core area in the training procedure. Although $L_A$ has certain constraints on the location of the source object, it is mainly for the whole image. So we add $L_R$ to enhance the constraint on the source object.

$$L_R = MSE(R_{com}, R_{real}). \tag{4}$$

#### *3.3.4 Final objective*

The final objective is called compositing loss, which is expressed as:

$$L_{final} = L_A + \lambda_1 L_C + \lambda_2 L_R \tag{5}$$

where $\lambda_1$ and $\lambda_2$ are super parameters keeping the balance of the three objectives. For training, we set $\lambda_1 = 10$ and $\lambda_2 = 1$ in our experiments. We use Adam solver [26] with the batch size of 1.

## 4 Results

### 4.1 Datasets and metrics

#### *4.1.1 Datasets*

Our goal is to generate a compositing image that is close to a real image using a training set of aligned image pairs. One is the target image with the source object, and the other is the image only with the same target scene. As far as we know, in the existing image compositing datasets, some datasets [1] only change color without position adjustment, and some [12] use a discriminator network to constrain the composite image without the compositing ground truth. These datasets can complete the single-task image compositing task. In order to implement the multitask image compositing, we propose a dataset with both position adjustment and color adjustment. In this work, we evaluate MT-GAN on our proposed SHU multitask image composition dataset. The information about the dataset is introduced in detail next.

We introduce a new SHU multitask image composition dataset. To our knowledge, this is the largest dataset for multitask image composition. SHU multitask image composition dataset contains 7756 images taken indoors. The training and testing sets consist of 3103 and 775 paired images (with and without source objects), respectively. The source objects are different kinds of cups, the target images are indoor scenes including tables, chairs, sofas, and other objects. We collect eight different cups as the source objects, with different colors, shapes, and appearances. For target images, we choose eight different scenes and collect paired images for each scene according to different heights, angles, and distances from the target. In addition, we collect images under two lighting conditions for the same scene: daylight and indoor lighting. In the same scene, the image without a source object is taken as the target image, and the image with a source object is taken as the ground truth. Then, we select an image from the ground truth and cut the foreground object to get the source image. Next, we label the region of the source object and its shadow in the ground truth images with a bounding box and store it in the label file. Finally, under the same lighting conditions of the same scene, 80% of the images are set as the training set, and the other 20% are for the test.

#### *4.1.2 Metrics*

For different tasks, there are many evaluation methods for image compositing. In the case of truth value, MSE and signal-to-noise ratio (SNR) are the two most commonly used evaluation methods. Therefore, we have adopted two evaluation methods.

Moreover, following [27], we train a classifier as an objective evaluation, imitating a person to make the subjective evaluation. The classifier aims to distinguish whether an image is a natural image or a computer-generated image composite. Then, it gives each image a score in the range of 0–1. The higher the score, the more realistic the image is, and vice versa.

### 4.2 Qualitative evaluation results

Figure 2 shows the qualitative compositing results of our approach. We can see that the source object and target image are basically consistent in terms of color, lighting, and context. Surprisingly, MT-GAN generates realistic shadows, and some of the compositing results also yield reflections. Although some composite images are not the same as the ground truth, they do look harmonious, but not incompatible with the context. The reason is that, under a single target scene, the reasonable placement and size of the source target are not unique, and the given real images are only a reference or a possibility.

There are slight differences in size, position, and shadow of the source target between the composite images and the ground truth, but the whole composite image

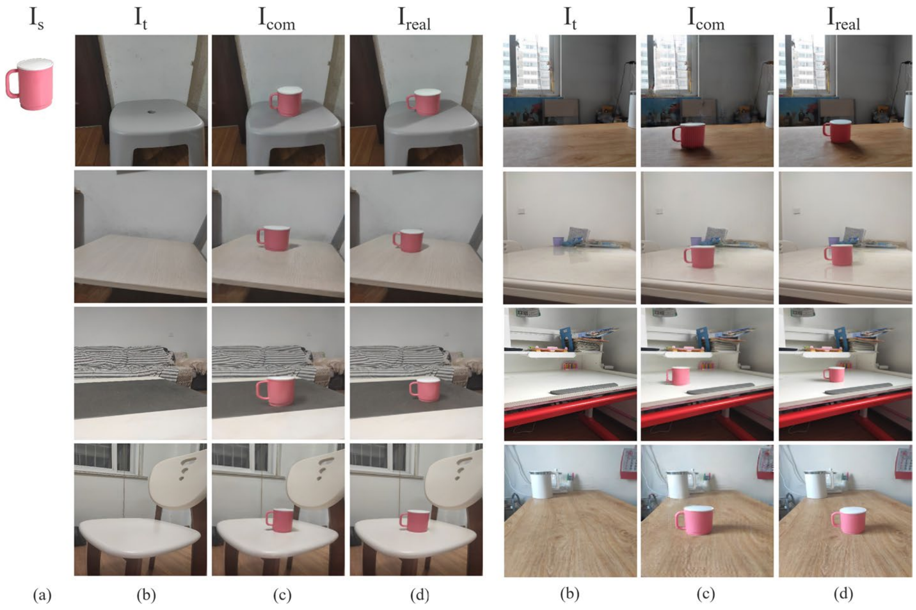

**Fig. 2** Image compositing results of one source sample and eight target samples on SHU dataset: **a** source object, **b** target image, **c** compositing image, and **d** ground truth: **a** source object, **b** target image, **c** compositing image, and **d** ground truth

has good harmony, which is the main goal of our network. For example, in the third line, the observed angle of the samples has a gradient. The position of the source object is closer to the shooting position, and the source object appears larger than the real one. In the second line, the plane in the target image has a reflective effect. Our network can also learn this character and generates reflections on the compositing images.

We also exploit experiments on other source objects, and the results are reported in Fig. 3. Our MT-GAN achieves realistic composite images as well.

### 4.3 Comparison with other methods

We compare with other deep learning-based image compositing methods. AGCP [33] is an adversarial geometric consistency pursuit model, using the Poisson solution algorithm to perform seamless appearance harmonization. Ding et al. [34] propose a conditional generative adversarial network to generate the object in the bounding box. DoveNet [1] aims to boost the quality of the composite image by harmonizing the foreground and background. It consists of a generator and two discriminators. The generator uses U-Net added the attention mechanism as the backbone network. One discriminator is the common GAN discriminator, and another is the domain verification discriminator.

Table 2 shows the comparison with other methods on the SHU dataset. The first row (arbitrary composite) is the result of the composite image with an arbitrary deformed source object and target image. It can be seen that the geometric consistency methods, e.g., ST-GAN and AGCP, achieve better results than arbitrary composite, but worse than the generative-based methods in the last three rows. Our proposed MT-GAN achieves the best results and outperforms all the baselines, which indicates the advantage of our multitask compositing strategy.

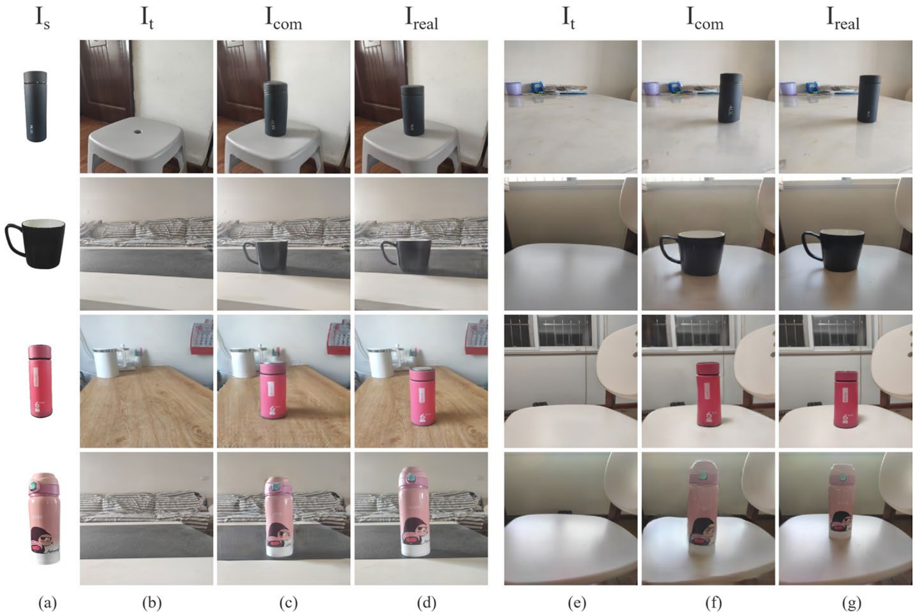

**Fig. 3** Image compositing results of other source samples on SHU dataset: **a** source object, **b** target image, **c** compositing image, **d** ground truth, **e** target image, **f** compositing image, and **g** ground truth

**Table 2** Result comparison with other deep learning-based methods for multitask image compositing on SHU dataset

| Methods | MSE↓ | PSNR↑ | Score↑ |
|---|---|---|---|
| Arbitrary composite | 423.98 | 20.86 | 0.15 |
| ST-GAN [12] | 200.87 | 23.52 | 0.64 |
| AGCP [33] | 198.40 | 24.56 | 0.689 |
| Ding et al. [34] | 196.61 | 24.66 | 0.691 |
| DoveNet [1] | 92.80 | 27.83 | 0.902 |
| MT-GAN | **82.30** | **29.51** | **0.954** |

Bold values indicate the best results

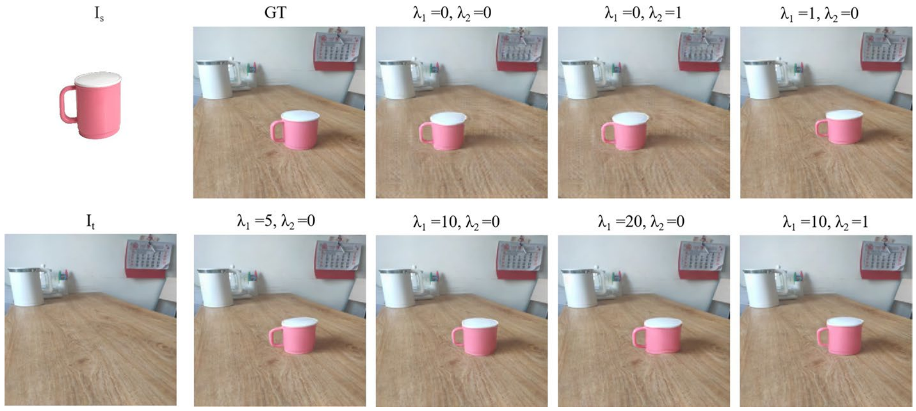
$I_s$
GT
$\lambda_1=0, \lambda_2=0$
$\lambda_1=0, \lambda_2=1$
$\lambda_1=1, \lambda_2=0$
$I_t$
$\lambda_1=5, \lambda_2=0$
$\lambda_1=10, \lambda_2=0$
$\lambda_1=20, \lambda_2=0$
$\lambda_1=10, \lambda_2=1$

**Fig. 4** Results testing on one sample of our approach for multitask compositing network: **a** source object, **b** target image, **c** compositing image, and **d** ground truth

## 4.4 Ablation study

### *4.4.1 Effect on different key parametre*

As can be seen in Fig. 4, compared with GT (ground truth), the position deviation of the source object is obviously large when $\lambda_1$=0 and $\lambda_1 = 0$. In other cases, the source object position is consistent with GT. Especially when $\lambda_1$=10 and $\lambda_1 = 1$, the size, location, and shadow of the source object are closer to the GT, which means the better appearance and geometric consistency with the real image. For quantitative evaluation, we study the detailed impact of each objective in our method based on investigating their effectiveness.

Training a model which can predict whether a given image will be judged to be realistic by a person are difficult, because of the prohibitive amount of manually labeled data. Therefore, we train a classifier as an objective evaluation. The effect of the composite image can be judged by its similarity with the natural image. The higher the similarity, the higher the score should be. Based on this idea, we use a pre-trained 16-layer VGG [32] network to score the similarity between composite images and real images. We initialize the weights on the ImageNet classification challenge and then fine-tune our binary classification task. The score range is 0–1, and a higher score means a more realistic image. More specifically, the training images are divided into three categories. The first category is the background images without foreground, the second category is the images composited with randomly deformation foreground and background, and the third category is the ground truth images. Then, we fine-tune the VGG network. Then, the scores of the third category images are set as the scores of a realistic image. Moreover, we calculate the MSE and PSNR as well. The smaller the MSE, the smaller the error with ground truth, and the larger the PSNR, the better the effect of the composite image.

Table 3 shows the results of ablating each learning objective. By comparing with the performance of the $L_A$, $L_C$ can significantly boost the results, and $L_R$ also has a certain improvement effect. The reason is that $L_A$ has certain constraints on the location of the source object, it is mainly for the whole image. $L_R$ can enhance the constraint on the source object. We suppose that the scope of $L_C$ is larger than that of $L_R$. Finally, our full method, i.e., "$L_A + L_C + L_R$," achieves the best performance.

Then, we did some experiments to seek out the best values of $\lambda_1$ and $\lambda_2$. First, we set $\lambda_2 = 0$, and test the performance on different $\lambda_1$, i.e., 0, 1, 5, 10, 20. Table 4 reports that MSE, PSNR, and Score get the best results when $\lambda_1 = 10$. Hence, we add $\lambda_2$ on the bases of $\lambda_1 = 10$. Table 5 shows that when $\lambda_2 = 1$ we achieve the optimal results. Therefore, we set $\lambda_1 = 10$ and $\lambda_2 = 1$ as the final key parameters.

Although the MSE and PSNR are not optimal when $\lambda_1 = 0, \lambda_2 = 0$, the score is still high. In other words, the position of the source object in the composite image deviates greatly from the position in the ground truth, resulting in poor MSE and PSNR, but the score is equivalent to others. MSE is the difference between the composite image and the ground truth, but the size and position of the source object in the whole image are not unique, while many positions seem reasonable. Not all positions

**Table 3** Result comparison on SHU dataset with different learning objectives

| Methods | MSE↓ | PSNR↑ | Score↑ |
|---|---|---|---|
| $L_A$ | 161.27 | 26.54 | 0.940 |
| $L_A + L_C$ | 84.66 | 29.34 | 0.951 |
| $L_A + L_R$ | 104.87 | 28.14 | 0.941 |
| $L_A + L_C + L_R$ | **82.30** | **29.51** | **0.954** |

Bold values indicate the best results

**Table 4** Result comparison on SHU dataset with different values of the key parameter $\lambda_1$

| $\lambda_1$ | MSE↓ | PSNR↑ | Score↑ |
|---|---|---|---|
| 0 | 161.27 | 26.54 | 0.94 |
| 1 | 86.26 | 29.15 | 0.943 |
| 5 | 85.61 | 29.24 | 0.945 |
| 10 | **84.66** | **29.34** | **0.951** |
| 20 | 84.93 | 29.31 | 0.945 |

Bold values indicate the best results

**Table 5** Result comparison on SHU dataset with different values of the key parameter $\lambda_2$

| $\lambda_2$ | MSE↓ | PSNR↑ | Score↑ |
|---|---|---|---|
| 0.5 | 84.47 | 29.35 | 0.952 |
| 1 | **82.30** | **29.51** | **0.954** |
| 2 | 83.12 | 29.34 | 0.952 |

Bold values indicate the best results

in the target image are reasonable. With the constraint of ground truth, the deviation and the unreasonable possibility are reduced. We achieve obviously better MSE and PSNR, and competitive scores.

### 4.4.2 Effect on spatial feature extraction network (SFEN)

We introduce a spatial feature extraction network (SFEN) to enhance performance. The source object and target image are both $3 \times 256 \times 256$. The target image provides spatial information for the source object. On this basis, it helps the network generate shadows and adjust the color of the generated foreground. The spatial position and size of the generated foreground are more important than color and shadow. Thus, we extract the spatial features of the target images before feeding the data to the backbone network. We use a lightweight network to convert the original three channels into one channel. Then, the target data becomes $1 \times 256 \times 256$. We express this process as $I_t^{'} = SFEN(I_t)$. Afterward, we concatenate the target data of channel 1 with the source data of channel 3 and feed them to the subsequent GAN network. In this way, on the one hand, the SFEN can extract the spatial information of the target images. On the other hand, the amount of background data is reduced by 2/3, which makes the backbone network pay more attention to foreground data and generate better composite images. By increasing the proportion of source data, the network can focus on the source object during training.

More specifically, when SFEN trained 500 epoch, the test results show that MSE $=$ 91.42 / PSNR $=$ 28.09. When the source object image and target image are concatenated and fed to the network directly, the test results show that MSE $=$ 264.62 / PSNR $=$ 23.71 at 500 epoch, and MSE $=$ 137.68 / PSNR $=$ 16.87 at 3000 epoch. Thus, the training speed is faster and the test results are better.

## 5 Conclusions

An efficient multitask GAN model for image compositing is proposed. In this method, a geometric consistency network and a spatial feature extraction network jointly assist the adversarial learning backbone. Benefiting from the design of the learning objectives, e.g., adversarial loss, compositing loss, and RoI consistency loss, the proposed method can generate realistic and plausible images. As the experiment results show, our proposed method achieves higher quality composite images with better performance than the baseline methods for both geometric and appearance consistency.

**Abbreviations**

| | |
|---|---|
| MT-GAN | Multitask GAN |
| GCN | Geometric consistency network |
| SFEN | Spatial feature extraction network |
| STN | Spatial transformer network |
| RoI | Region of interest |
| GAN | Generative adversarial network |
| MSE | Mean square error |
| PSNR | Peak signal-to-noise ratio |

**Authors' information**

**Xiang Li** He received the B.S. degree from the Tianjin University of Technology and Education, Tianjin, China, in 2001, and the M.S. degree from Jiangsu University, Zhenjiang, China, in 2008. He is currently pursuing the Ph.D. degree in signal and information processing with Shanghai University, China. His current research interests include artificial intelligence and deep learning.

**Guowei Teng** He received his Ph.D. degree in Communication and Information Systems from Shanghai University in 2005 and his M.S. degree in electronics from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. He is currently a senior engineer at the School of Communication and Information Engineering, Shanghai University, China. His research interests include video compression, image and video processing, and computer vision.

**Guowei Teng** He received his Ph.D. degree in Communication and Information Systems from Shanghai University in 2005 and his M.S. degree in electronics from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. He is currently a senior engineer at the School of Communication and Information Engineering, Shanghai University, China. His research interests include video compression, image and video processing, and computer vision.

**Ping An** She received the B.E. and M.S. degrees from the Hefei University of Technology, Hefei, China, in 1990 and 1993, respectively, and the Ph.D. degree from Shanghai University, Shanghai, China, in 2002. She is currently a professor at the School of Communication and Information Engineering, Shanghai University, China. Her research interests include image and video processing, with a focus on immersive video processing

**Hai-yan Yao** She received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2004, the M.S. degree from Nanjing University of Science and Technology, Nanjing, China, in 2011, and the Ph.D. degree from Shanghai University, Shanghai, China, in 2022. She was with the Department of Electronic Information and Electrical Engineering, Anyang Institute of Technology, China. Her current research interests include computer vision and machine learning.

### Acknowledgements

This research was funded by National Natural Science Foundation of China [Grant 62071287], Science and Technology Commission of Shanghai Municipality [Grant 20DZ2290100], and Project of Science and Technology of Henan [Grant 212102310383].

### Author contributions

All the authors contributed equally in data collection, processing, experiments, and article writing. The authors read and approved the final manuscript.

### Funding

This research was funded by National Natural Science Foundation of China [Grant 62071287], Science and Technology Commission of Shanghai Municipality [Grant 20DZ2290100] and Project of Science and Technology of Henan [Grant 212102310383].

### Availability of data and materials

All the data are available upon request from the corresponding author.

## Declarations

### Ethics approval and consent to participate

Not applicable. All the databases were obtained from the literature that are publicly available.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 14 November 2022 Accepted: 4 April 2023
Published online: 27 April 2023

### References

1. W. Cong, J. Zhang, L. Niu, L. Liu, Z. Ling, W. Li, L. Zhang, DoveNet: Deep image harmonization via domain verification.*In Proceedings of the Computer Vision and Pattern Recognition (CVPR)* (2020)
2. X. Cun, C. Pun, Improving the harmony of the composite image by spatial-separated attention module. IEEE Trans. Image Process. **29**, 759–4771 (2020)
3. Y. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, M. Yang, Deep image harmonization.*In Proceedings of the Computer Vision and Pattern Recognition (CVPR)* (2017)
4. H. Wu, S. Zheng, J. Zhang, K. Huang, GP-GAN: Towards realistic high-resolution image blending.*the ACM International Conference on Multimedia* (2019)
5. L. Zhang, Y. Gao, R. Zimmermann, Q. Tian, X. Li, Fusion of multichannel local; global structural cues for photo aesthetics evaluation. IEEE Trans. Image Process. **23**(3), 419–1429 (2014)
6. D. Liu, C. Long, H. Zhang, H. Yu, X. Dong, C. Xiao, ARshadow-Gan: Shadow generative adversarial network for augmented reality in single light scenes.*In Proceedings of the Computer Vision and Pattern Recognition (CVPR)* (2020)
7. Y. Sheng, J. Zhang, B. Benes, SSN: Soft shadow network for image compositing.*In Proceedings of the Computer Vision and Pattern Recognition (CVPR)* (2021)
8. S. Zhang, R. Liang, M. Wang, ShadowGAN: shadow synthesis for virtual objects with conditional adversarial networks. Comput. Visual Media **5**(1), 05–115 (2019)

9. S. Tripathi, S. Chandra, A. Agrawal, A. Tyagi, J. Rehg, V. Chari, Learning to generate synthetic data via compositing.*In Proceedings of the Computer Vision and Pattern Recognition (CVPR)* (2019)
10. L. Zhang, T. Wen, J. Min, J. Wang, D. Han, J. Shi, Learning object placement by inpainting for compositional data augmentation. *The European Conference on Computer Vision(ECCV)* (2020)
11. L. Lin, ST-GAN: spatial transformer generative adversarial networks for image compositing.*In Proceedings of the Computer Vision and Pattern Recognition (CVPR)* (2018)
12. C. Lin, E. Yumer, O. Wang, E. Shechtman, S. Lucey, ST-GAN: spatial transformer generative adversarial networks for image compositing. *In Proceedings of the Computer Vision and Pattern Recognition (CVPR)* (2018)
13. K. Kikuchi, K. Yamaguchi, E. Simo-Serra, T. Kobayashi, Regularized adversarial training for single-shot virtual try-on. *The IEEE International Conference on Computer Vision (ICCV) Workshop* (2019)
14. I. Goodfellow, A. Abadie J, M. Mirza, Generative adversarial nets.*International Conference on Neural Information Processing Systems* 2672–2680 (2014)
15. S. Azadi, D. Pathak, S. Ebrahimi, T. Darrell, Compositional GAN: learning image conditional binary composition. Int. J. Comput. Vis. **128**(10), 570–2585 (2020)
16. K. Zheng, M. Wei, G. Sun, B. Anas, Using vehicle synthesis generative adversarial networks to improve vehicle detection in remote sensing images. ISPRS Int. J. Geo-Inf **8**(9), 390 (2019)
17. B. Chen, A. Kae, Toward realistic image compositing with adversarial learning. *In Proceedings of the Computer Vision and Pattern Recognition (CVPR)* (2019)
18. T. Lin, M. Maire, S. Belongie, J. Hays, P. Pietro, R. Deva, D. Piotr, C. L. Zitnick, Microsoft COCO: common objects in context. *The European Conference on Computer Vision (ECCV)* 8693, 740–775 (2014)
19. M. Everingham, G. Van, C. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 03–338 (2010)
20. M. Cordts, M. Omran, S. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding. *In Proceedings of the Computer Vision and Pattern Recognition (CVPR)* (2016)
21. G. Georgakis, M. Md, Alimoor Reza, Arsalan Mousavian, Phi-Hung Le, and Jana Kosecka. Multiview RGB-D dataset for object instance detection. In *Proceedings of the International Conference on 3D Vision* (2016)
22. P. Laffont, Z. Ren, X. Tao, C. Qian, J. Hays, Transient attributes for high level understanding and editing of outdoor scenes. ACM Trans. Graph. **33**(4), 1–11 (2014)
23. W. Cong, J. Cao, L. Niu, J. Zhang, X. Gao, Z. Tang, L. Zhang, Deep image harmonization by bridging the reality gap. arXiv:2103.17104 (2021)
24. M. Jaderberg, K. Simonyan, K. Zisserman, Spatial transformer networks. *Advances in Neural Information Processing Systems* 2017–2025 (2015)
25. F. Zhang, H. Zhu, S. Lu. Spatial Fusion GAN for Image Synthesis. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)* (2019)
26. D. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)
27. J. Zhu, P. Krahenbuhl, E. Shechtman, A. Efros, Learning a discriminative model for the perception of realism in composite images. *The IEEE International Conference on Computer Vision* 3943–3951 (2015)
28. J. Zhu, T. Park, P. Isola, Unpaired image-to-image translation using cycle-consistent adversarial networks. *The IEEE International Conference on Computer Vision* 2242–2251 (2017)
29. K. He, X. Zhang, S. Ren, Deep residual learning for image recognition. *In Proceedings of the Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016)
30. J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution. *The European Conference on Computer Vision (ECCV)* 694–711 (2016)
31. D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: the missing ingredient for fast stylization. arXiv: 1607.08022 (2016)
32. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
33. X. Li, G. Teng, P. An, H. Yao, Image synthesis via adversarial geometric consistency pursuit. Signal Process.-Image Commun. **99**, 116489 (2021)
34. Y. Ding, G. Teng, Y. Yao, Context-Aware Natural Integration of Advertisement Object. *In Proceedings of the 2019 International Conference on Image Processing* 4689–4693 (2019)

## Publisher's Note