electronics

MDPI

Article

Image Composition Method Based on a Spatial Position Analysis Network

Xiang Li [1,2], Guowei Teng [1,*], Ping An [1,*] and Haiyan Yao [2]

[1] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China; shdxlix@shu.edu.cn
[2] School of Electronic Information and Electrical Engineering, Anyang Institute of Technology, Anyang 455000, China; yaohywqh@shu.edu.cn
* Correspondence: tenggw@shu.edu.cn (G.T.); anping@shu.edu.cn (P.A.)

Abstract: Realistic image composition aims to composite new images by fusing a source object into a target image. It is a challenging problem due to the complex multi-task framework, including sensible object placement, appearance consistency, shadow generation, etc. Most existing researchers attempt to address one of the issues. Especially before compositing, there is no matching assignment between the source object and target image, which often leads to unreasonable results. To address the issues above, we consider image composition as an image generation problem and propose a deep adversarial learning network via spatial position analysis. We target the analysis network segment and classify the objects in target images. A spatial alignment network matches the segmented objects with the source objects, and predicts a sensible placement position, and an adversarial network generates a realistic composite image with the shadow and reflection of the source object. Furthermore, we use the classification information of target objects to filter out unreasonable image compositing. Moreover, we introduce a new test set to evaluate the network generalization for our multi-task image composition dataset. Extensive experimental results of the SHU (Shanghai University) dataset demonstrate that our deep spatial position analysis network remarkably enhances the compositing performance in realistic, shadow, and reflection generations.

Keywords: image composition; spatial position analysis; generator; deep learning

Citation: Li, X.; Teng, G.; An, P.; Yao, H. Image Composition Method Based on a Spatial Position Analysis Network. *Electronics* **2023**, *12*, 4413. https://doi.org/10.3390/electronics12214413

Academic Editors: Byung Cheol Song and George A. Papakostas

Received: 7 September 2023
Revised: 17 October 2023
Accepted: 24 October 2023
Published: 26 October 2023

1. Introduction

Image composition [1–3] aims to combine two different images or parts of images into an image that conforms to human visual common sense. It is a very complex task, involving a lot of details, such as image harmonization, object placement, object shadow generation, etc. We focus on the methods of placing (a foreground object) the source object onto the target image to make it look realistic. The main problems to be solved include placing the source object into the target image and adjusting the position and size to make it reasonable, coordinating the illumination and color between the source objects and the target images, and generating shadows and reflections that the source object does not have.

The main issues in image composition can be divided into two categories, i.e., appearance inconsistency and geometric inconsistency. The appearance inconsistency mainly includes (1) an unnatural boundary between the source object and the target image after compositing; (2) inconsistency in color, illumination, and contrast between the source object and the target image; (3) a lack of shadows and reflections in the target image as a backdrop. Image normalization aims to adjust the color and lighting statistics of the source object and target images to make them more compatible. In addition, shadow and reflection generations aim to train deep learning-based networks to generate shadows and reflections to better integrate the source object into the target image. Geometric inconsistency mainly includes (1) the size of the source object appearing too large or small in the target image; (2) the position of the source object not being associated with other objects in the

target image; (3) different perspectives between the source object and the target image; (4) unreasonable occlusion. To address geometric inconsistency, geometric corrections are performed on the source object, such as displacement, scaling, perspective transformation, etc. Object placement performs geometric transformations on the source object to adapt to the target image. The prediction through models aims to analyze the target image and predict the position and size of the source object through a deep learning-based network.

The image composition task is decomposed into several sub-tasks to solve separately. The object placement [4–6] task aims to find the reasonable location of the source object in the target image and give the appropriate size. The image harmonization task [7–9] is to solve the problem of color harmony between the source object and target image based on reasonable placement and size assumptions. The image blending task [10,11] mainly solves the problem of appearance inconsistency between the source object and the target image, especially in boundary regions. In addition, there are some studies that address the shadow and reflection [12–14] generations of the source object. MT-GAN [15] attempts to utilize image generation methods to simultaneously address object placement, appearance consistency, and shadow generation issues.

However, these studies were conducted on the premise that the two images are suitable for composition, and there are few studies on whether the two images are actually suitable for image composition. Before we perform the composition task, one important task is to estimate whether the target image contains reasonable spatial positions that are suitable for the placement of the source object. If not, then the two images are not suitable for image composition. Therefore, we consider analyzing the spatial position of target images before making synthetic images. One purpose of the analysis is to filter the images, and the other is to choose a suitable location for source object placement. Hence, we propose a generation method based on target image resolution to solve the above problems.

Our main contributions are listed as follows:

1. Deeply learned spatial position analysis for image composition—We propose a multi-task deep learning network for target object screening. We first analyze the spatial position of the target image, and then find an appropriate placement location for the source object. Furthermore, the source object is generated in the position box to obtain the image compositing result. Qualitative analysis shows that our method can generate clear composite images for source objects with different shapes and patterns, and quantitative analysis shows that the proposed framework achieves the highest score of authenticity.
2. The largest multi-task image composition dataset with spatial position annotations—We establish a large multi-task image composition dataset with 400 images with position masks for spatial position analysis. Moreover, a new generalization test set is collected as a supplement to the SHU dataset. It includes various target objects suitable for placing the source object, including stools, chests of drawers, tables, bedside cabinets, etc. We also add new annotations of source objects' locations for spatial position analysis.

2. Related Work

In this section, we will discuss the related works on image composition.

Image composition has attracted wide attention in recent years. It aims to cut the source object from one image and process it or paste it seamlessly on another target image, resulting in a composite image. So far, the image composition task has been explored from a variety of perspectives. To complete a realistic-looking composite image, it is necessary to address issues such as geometric inconsistency, appearance inconsistency, and shadow and reflection generations. Deep learning-based methods have achieved widespread application, with some researchers [16,17] utilizing CNN frameworks in their computational methods for recognition tasks in the biological field. Similarly, many deep learning methods have emerged in image composition tasks. For example, Realism CNN [18] fine-tuned the VGG network to distinguish the authenticity of images, which assists in making

high-quality synthesis images and provides a criterion for evaluating the authenticity of composite images. Tsai et al. [9] proposed an effective method to collect large-scale and high-quality training data and designed an end-to-end deep convolutional neural network that can capture the contextual and semantic information of composite images during the coordination process. They proposed a solution to the image harmonization task. Then, Cong et al. [7] provided a dataset (iHarmony4) suitable for image harmony tasks and proposed an attention mechanism network to improve the effectiveness of image composition. Afterward, some researchers [19,20] conducted a series of studies on the basis of the iHarmony4 dataset, further improving the image composition effect.

During this period, the target placement task also made some progress with the emergence of ST-GAN [21]. ST-GAN integrates the STN and GAN frameworks to find realistic geometric corrections to the source object, so that it looks natural when composited into the target image. On this basis, Chen et al. [1] proposed a similar shape object replacing method. A GAN architecture with a pair of discriminators and a segmentation network was used to adjust the color of the source object for automatic image compositing. A transformation network and a refinement network were used to improve geometric consistency and polish the boundary of the composite image, respectively. Spatial fusion GAN [22] combines the spatial transformation network (STN) and style transformation network to achieve realism in both geometric and appearance spaces of the synthesized image. Tripathi et al. [5] presented a task-aware approach containing a trainable image synthesizer that can assess the strengths and weaknesses of a network to generate meaningful training samples. Subsequently, Zhang et al. [6] proposed PlaceNet, which can predict a diverse distribution of reasonable locations for source object placement. This had the benefit of a self-learning framework that could generate necessary training data without any manual labeling. Li et al. [23] presented a fast OPA model including foreground dynamic filters, background prior transfer, and composite feature simulation. Zhou et al. [24] proposed a graph completion module (GCM) with a dual-path framework to address the object placement problem. This GCM can fully exploit annotated composite images. In recent years, encouraging progress has been made in object placement and image harmonization, and novel solutions for shadow and reflection generations have emerged [12,13,25,26].

However, there are some limitations to the above methods. Image harmonization tends to adjust the color and illumination statistics of the source object to make the whole image harmonious, while shadow or reflection generation aims to generate plausible shadow or reflection for the source object to make the image more realistic. But they need to specify the location and size of the source object in advance. Object placement focuses on seeking information such as the appropriate size, shape, and position of the source object, but lacks the ability to adjust the appearance consistently.

Therefore, we propose a multi-task framework that can solve image harmonization, object placement, and shadow generation in a one-stop manner. In addition, existing image composition approaches do not judge whether the two images are suitable for compositing; our approach is an exception. The proposed method filters data by analyzing the target image.

3. Materials and Methods

In this section, we propose a spatial position analysis network (SPAN) for image compositing, matching positions before generating images. The overall process of the proposed SPAN is synthetically described in Figure 1. The network consists of five main components, including a target analysis network, spatial alignment network, composite generator, global discriminator, and local discriminator. Symbols and their representations in this section are shown in Table 1.

As the first step of the SPAN, the target analysis network is used to analyze the target image. It segments different objects in the target image and corresponds to different values in semantic segmentation image I_a. Then, we concatenate the analyzed data I_a with the source object I_s and feed it to the spatial alignment network. The spatial alignment network

predicts the locations of the source object based on the analysis results of the target image and provides the most reasonable placement position. The position is represented by the diagonal coordinates of the corresponding box, namely, the four numerical values of the upper left corner coordinate and the lower right corner coordinate, e.g., (x_0, y_0, x_1, y_1). These coordinate values are represented by a mask map I_m and fed to the subsequent composition generator. The values within the box are set to 1, and the values in other positions are set to 0. Then, the target image I_t, source object I_s, and position mask map I_m are concatenated and fed to the composite generator to predict the synthesized image I_c. The whole image is obtained through the adversarial generative learning process, and the generation process focuses on the region of the source object. Therefore, in addition to identifying the authenticity of the whole image, attention should also be paid to the details of the generated source object. We use two discriminators to train the generator: a global discriminator and a local discriminator. The global discriminator identifies the authenticity of the whole image between composite and real images. We extract the source objects from the composite image and the real image and then feed them to the local discriminator to verify the detailed authenticity of the generated source object. For all the experiments, we use the Adam solver with a batch size of 1. We set the learning rate as 0.00005.

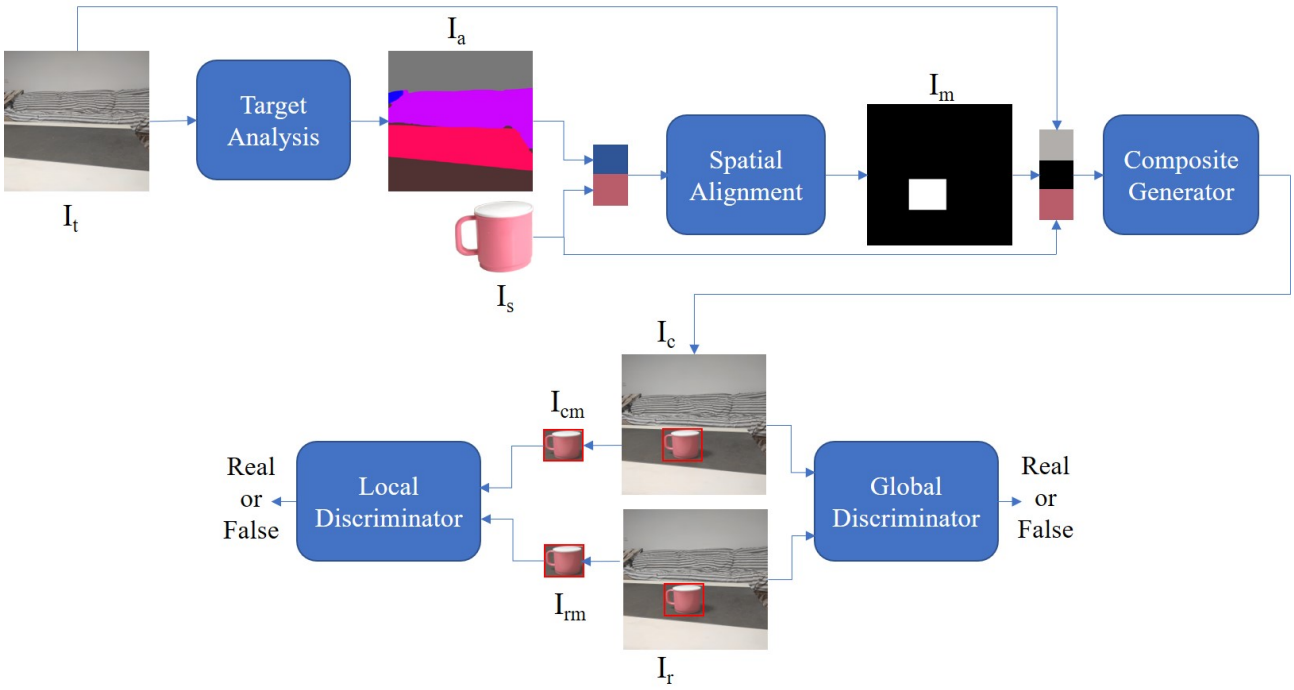

Figure 1. The architecture of the overall framework of the proposed spatial position analysis network (SPAN) for image composition. The target analysis network is used to analyze the target image. The spatial alignment network predicts the locations of the source object based on the analysis results of the target image and provides the most reasonable placement position. Then, the composite generator builds the synthesized image. The global discriminator identifies the authenticity of the whole image between composite and real images, and the local discriminator verifies the detailed authenticity of the generated source object.

3.1. Target Analysis Network

There are usually several different objects in an image. The relative size and position of each object in a real image are consistent with common sense. If an object in the image is placed in an inappropriate position, the whole image will look unreal. For example, normally, the train is running on the track, but if the train appears on the road, the picture will look fake.

Table 1. Related mathematical symbols and their representation.

Symbol	Representation
I_t	target image
I_s	source object
I_a	analyzed data; semantic segmentation data
I_m	mask map
I_{mr}	ground truth mask map
I_c	composite image
I_r	ground truth image
I_{cm}	source object cropped from the composite image
I_{rm}	source object cropped from the ground truth image
L_p	position loss function
L_c	compositing loss function
L_{Dg}, L_{Gg}	global discriminator loss function
L_{Dl}, L_{Gl}	local discriminator loss function
L_{final}	final loss function
P_p	predicted position
P_g	ground truth position
λ_p, λ_g	super parameters

The target analysis network first pre-processes the input data by analyzing the instances in the target image and labeling different target objects with different values. Then, the analyzed data are fed to the subsequent network to assist in training the image compositing. Through iterative training, the network learns to find the suitable location and scale of the target object that matches the source object. If the object attributes in the target image match the source object, image compositing can be performed; if there is no match, image compositing is not possible. The purpose is to filter the target images.

We adopt SegNeXt [27] as our target analysis network. We call the model parameters pre-trained on the ADE20K dataset and fine-tune them on our target images based on a semantic analysis task. The output of the target analysis network is the semantic information, in which different values represent different target objects using a three-channel RGB format.

3.2. Spatial Alignment Network

After the semantic analysis of the target image, the analyzed data (I_a) and the source object (I_s) are fed to the spatial alignment network to predict the location and size. Based on the input information, the spatial alignment network predicts the mask map of the source object, which corresponds to the position and scale distribution of the source object in the target image to be embedded. The mask map is set to a rectangle, and the network gives two coordinates of the diagonal of the rectangle: the coordinates in the upper left corner and the lower right corner.

The analyzed data I_a are three-channel and the source object I_s data are four-channel, in RGBA format. We resize both groups of data to 256 $\times$ 256 and feed the concatenation data to the spatial alignment network. The architecture is shown in Figure 2. Each blue block represents a sub-sampling convolution layer following a normalized layer and a ReLU activation function. The yellow block represents a ResNet module containing two convolutional layers. The input data first go through three sub-sampling convolution layers followed by three ResNet schemes for feature extraction. Next, the data go through another three sub-sampling convolution modules. Finally, the data are flattened, and four coordinate values are output through two fully connected layers. In order to facilitate the data concatenation in the subsequent image composite generator, we introduce the position mask map I_m according to the coordinate values by aligning the predicted position of the source object to 1 in a 256 $\times$ 256 all-zero map.

In the spatial alignment network, we define the position loss using the MSE between the predicted position P_p and the ground truth position P_g. The position loss function can be expressed via Equation (1).

$$L_p = MSE(P_p, P_g) \tag{1}$$

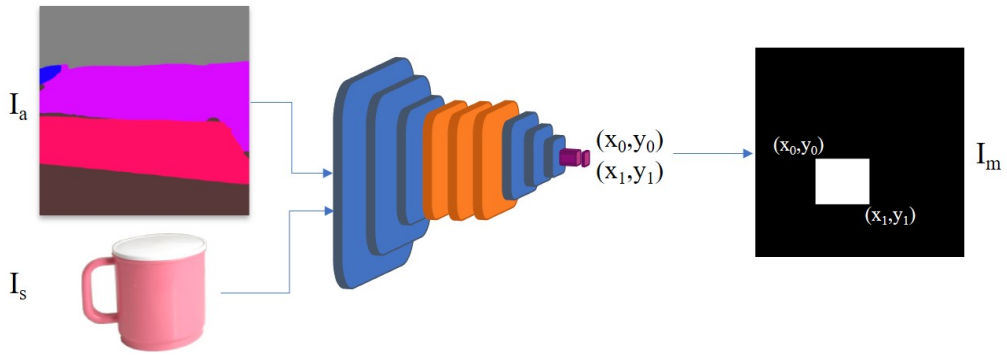

Figure 2. The architecture of the spatial alignment analysis network. (x_0, y_0) represent the diagonal coordinates of the corresponding box of the upper left corner coordinate and (x_1, y_1) represent the lower right corner coordinate. Set the value in the box to 1 and the other values to 0 to obtain the mask image I_m.

3.3. Composite Generator

Although the reasonable placement of the source object has been determined, the resulting composite image may still be unrealistic. The reason for this is that the source object and target images are collected under different lighting conditions and shooting angles. Therefore, the source object should be further processed to make it consistent with the color of the target image. We use a generative network for image compositing. The generator takes the input data as a condition to generate a brand new image. The output of the generator tends to be plausible as long as the input of the discriminator is a real image. One original task of an adversarial generative network is generating real face images based on random noise [28], which is almost independent of the initial input data. So, the generated source objects are more easily integrated with the target image in terms of shape and color.

We use the merge data, source object I_s, mask map I_m, and target image I_t as the input of the composite generator. Where I_s is 256 × 256 × 4, mask map is 256 × 256 × 1, and I_t is 256 × 256 × 3. So, the input data are 256 × 256 × 8.

The structure schematic of the composite generator is shown in Figure 3a. We borrow the encoder–ResNet–decoder structure of the CycleNet [29] backbone network. We expect the encoder attention map to pay more attention to the target image of the encoder feature. The for this reason is that the source objects the of encoder feature may not be fully harmonized yet. Inspired by the attention module in DoveNet [7], we add a feature attention mechanism to the CycleNet. The encoder, ResNet, and decoder are represented as blue, yellow, and green blocks, respectively.

Each blue block corresponds to a convolution module, containing a convolutional layer, a regularization layer (InstanceNorm) [30], and a LeakyReLU activation function. Each yellow block corresponds to a ResNet module. The green block corresponds to a deconvolution module, containing a transposed convolution layer, a regularization layer (InstanceNorm), and a ReLU activation function. The attention module is shown in Figure 3b.

Firstly, we concatenate the corresponding encoder features (the output of the blue block) with the decoder features (the output of the green block). Then, the full attention maps containing the spatial attention and channel attention are learned. Specifically, we apply a 1 × 1 convolution layer following Sigmoid activation on the concatenation. Next,

we perform element-wise multiplication on the encoder (resp., decoder) attention map and the encoder (resp., decoder) feature.

We compare the compositing image I_c and the ground truth image I_r at pixel level. The $L1$ loss is used to promote the network to generate plausible images, and the compositing loss function can be expressed as follows:

$$L_c = ||I_c - I_r||_1 \tag{2}$$

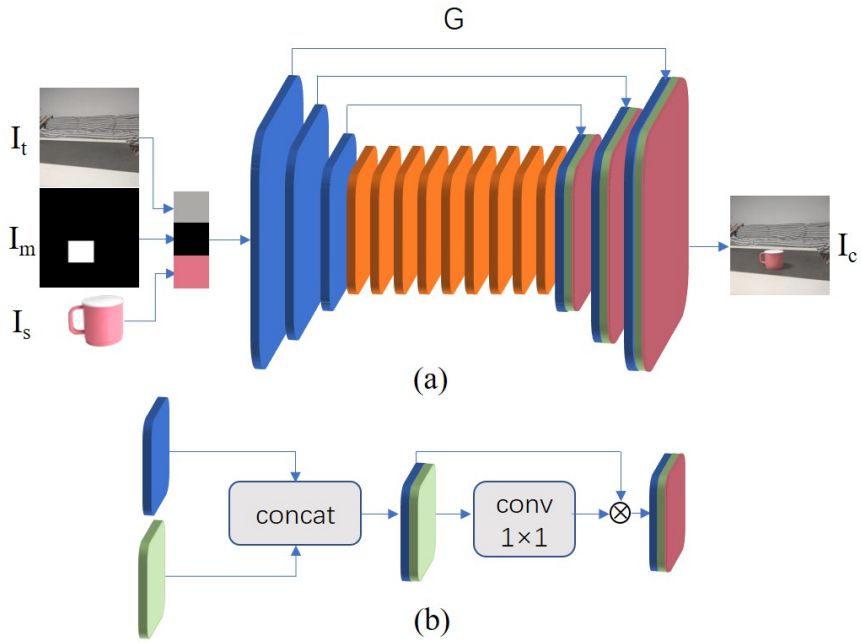

Figure 3. The architecture of the composite generator. The blue blocks correspond to the encoder module, the yellow blocks correspond to the ResNet module, the green blocks correspond to the deconvolution module, and the red blocks correspond to the attention module. (**a**) The architecture of the composite generator. (**b**) The attention module in the composite generator.

3.4. Discriminator

The purpose of the composite generator is to generate realistic source objects in the corresponding region in the mask map I_m. Two discriminators are used to discriminate the generated composite images. The global discriminator mainly focuses on the overall characteristics of the composite image, making the entire image close to the real image. Meanwhile, the local discriminator pays more attention to the details of the composite image in the corresponding region of the mask map. In addition, the global discriminator distinguishes the harmony of the whole image, and the local discriminator identifies the consistency of the generated source object with the ground truth. The two discriminators assist each other in the training process.

The global discriminator identifies the realism of the composite image. We feed the ground truth image and the composite image to the discriminator with different annotations. This can train the discriminator's ability to distinguish authenticity, and continuously improve the generator's ability to synthesize realistic images. The architecture of the global discriminator is shown in Figure 4, with fixed input data of 256 × 256. There are five convolution modules. The first one contains a convolution layer and a LeakyReLU activation function, and the subsequent three convolution modules contain a convolution layer, a regularization layer (InstanceNorm), and a LeakyReLU activation function. The last convolution module consists of only one convolution layer.

Figure 5 shows the architecture of the local discriminator, which is almost the same as the global discriminator. The difference is that the size of the input data is related to the mask map, which is not fixed. Specifically, due to the small size of some source objects, the

local discriminator may experience a size smaller than 1 during the down-sampling process. So, we increase the degree of padding in the convolution layer to avoid this situation in the local discriminator.

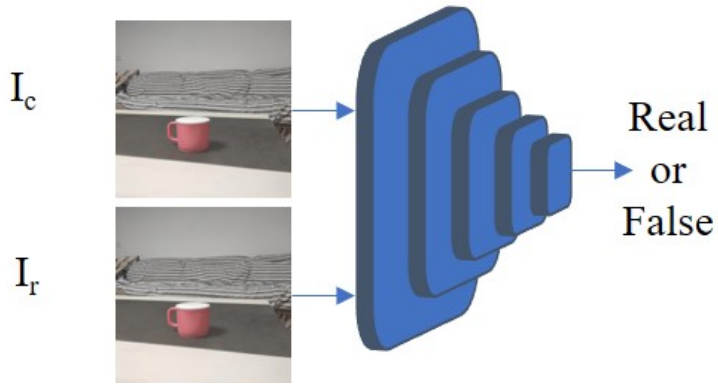

Figure 4. The architecture of the global discriminator with fixed input data of 256 × 256.

The loss function of the global discriminator is shown as follows:

$$L_{D_g} = E[max(0, 1 - D_g(I_r))] + E[max(0, 1 + D_g(I_c))]$$
$$L_{G_g} = -E[D_g(G(I_t, I_m, I_s))] \quad (3)$$

We minimize L_{D_g} to train global discriminator D_g, which is encouraged to produce large scores for real images and small scores for generated images. In the same way, we minimize L_{G_g} to train G, and the generated samples are expected to fool global discriminator D_g into taking them for large scores.

The loss of the local discriminator is given by the following:

$$L_{D_l} = E[max(0, 1 - D_l(I_r, I_m))] + E[max(0, 1 + D_l(I_r, I_m))]$$
$$L_{G_l} = -E[D_l(G(I_t, I_m, I_s), I_m)] \quad (4)$$

We train D_l by minimizing L_{D_l}, and train G by minimizing L_{G_l}.

Overall, the final loss function for training the whole network is shown in the following:

$$L_{final} = L_c + \lambda_p L_p + \lambda_g (L_{G_g} + L_{G_l}) \quad (5)$$

where λ_p and λ_g are super parameters. We set λ_p as 1, and λ_g as 0.1 in our experiment.

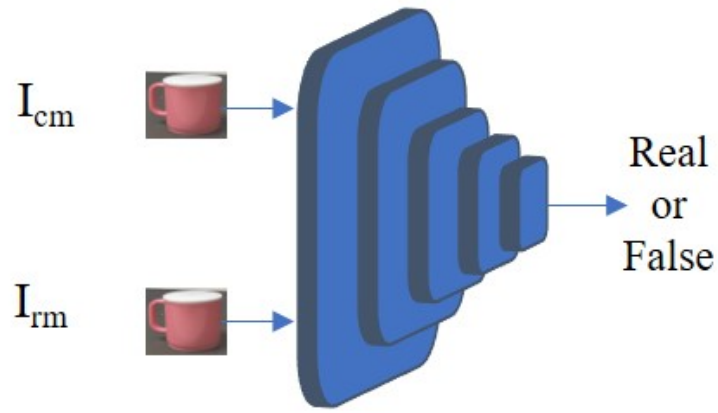

Figure 5. The architecture of the local discriminator. The input data are not fixed and are related to the mask map.

4. Results

We evaluate the SPAN model on the SHU dataset, and achieve optimal performance in the literature. All training and testing of the proposed model was completed based on the PyTorch framework.

4.1. Experiment Setup

Our SPAN method contains three sub-networks: the target analysis network, the spatial alignment network, and the composite generator. The target analysis network uses the parameters pre-trained on the ADE20K dataset, while the spatial alignment network and the composite generator are trained and evaluated on the proposed multi-task image composite dataset, named the SHU dataset.

ADE20K dataset is a publicly available dataset published by the Computer Vision team at the Massachusetts Institute of Technology (MIT) in 2016. This dataset fully marks the targets in the image, with more than 3000 object categories, which can be used for semantic segmentation, instance segmentation, scene analysis, and other tasks. Currently, the data publicly provide 25,574 training images and 2000 validation images. We utilize the SegNeXt network parameters, which are obtained by training on this dataset.

The SHU dataset is proposed by MT-GAN and contains a total of 7756 images across eight source object categories, including 6206 images in the training set and 1550 for the test. The source objects are different in color, texture, and pattern. The data are taken in different scenes with diverse angles, distances, and lighting conditions. The paired samples are a target image without the source object and a ground truth image with the source object under the same photographic conditions. Each ground truth sample also marks the RoI containing the source object and corresponding shadows. In this paper, the spatial alignment network inputs the source object region, excluding the shadow part. Thus, we add a bounding box annotation containing only the source objects. The spatial alignment network and the composite generator are trained and tested on the SHU dataset, of which 20% is used for the test set and 80% for training. Further, we add new test data to test the generalization of the network, called the generalization test set (GTS), to distinguish them from the previous test set (PTS). The images in the PTS exhibit different angles and different lighting conditions compared to the training data, but the same scene. To test the compositing effect of our method in multiple brand-new scenarios, we collect a total of 400 images from 15 completely different scenes in the GTS, using different angles and distances to those used with the PTS. Considering that the GTS needs to be adapted for image composition with source objects, we adopt a similar shooting method with the SHU dataset. The difference between the SHU dataset and other datasets is that there are multiple paired images corresponding to a source object. Each pair of data contains a target image (without the source object) and a ground truth image (with the source object) under the same shooting condition. To our knowledge, there are currently no other datasets with large quantities of paired sample data in different scenes. In addition, we name the eight categories of source objects, which are shown in Figure 6.

Figure 6. The eight categories of source objects and corresponding names (S1, S2, S3, S4, S5, S6, S7, S8).

4.2. Results Comparison with Other Methods

We evaluate our method and report the results of the comparison with other methods, including quantitative results, qualitative results, ablation study results, and the objective estimation of other objects of the SHU dataset. We used mean squared error (MSE), peak signal-to-noise ratio (PSNR), intersection over union (IoU), the user study, and the objective estimation score [18] as the evaluation metrics.

4.2.1. Quantitative Experimental Results

We compare our approach with other deep learning-based approaches. Table 2 shows the comparison of the composite results of source object S1. Following MT-GAN [15], we use MSE, PSNR, and the objective estimation score of the compositing images. The objective estimation scores are tested on the PTS and GTS, respectively. In addition, in order to obtain the objective estimation of the spatial alignment network, we introduce IoU to evaluate the accuracy of mask map generation and obtain a Mask-IoU score.

With the structure of the network backbone anchored, we first try a two-stage approach, in which the spatial alignment network and the composite generator are trained separately. Then, we attempt an end-to-end approach, in which the spatial alignment network and the composite generator are trained simultaneously. Importantly, the spatial alignment network is trained with only one constraint of the bounding box in a two-stage method. Meanwhile, in the end-to-end method, there are two more constraints for training, e.g., the global image ground truth constraint and adversarial constraint. We can see that the end-to-end method is superior to other methods. Specifically, the end-to-end method is 46.99% higher than the two-stage method in terms of the Mask-IoU score.

Our method achieves the best performance in both PTS score and GTS score, but is slightly worse than MT-GAN and DoveNet in terms of MSE and PSNR. The reason for this is the insufficient accuracy of the generated target position resulting in the deviation between the generated and the ground truth position of the source target.

Table 2. Comparison results of objective estimation with other methods. The best results are denoted in boldface.

Method	MSE ↓	PSNR ↑	Mask-IoU Score ↑	PTS Score ↑	GTS Score ↑
Arbitrary Composite	423.98	20.86	-	0.1510	0.1391
ST-GAN [21]	200.87	23.52	-	0.6429	0.4639
AGCP [31]	198.40	24.56	-	0.6887	0.4954
DoveNet [7]	92.80	27.83	-	0.9021	0.6861
MT-GAN [15]	**82.30**	**29.51**	-	0.9453	0.7561
Ours (two-stage)	115.45	28.00	0.6103	0.9751	0.8688
Ours (proposed)	103.51	28.69	**0.8971**	**0.9958**	**0.8933**

Samples of the mask map are shown in Figure 7. We can see that the differences between the predicted mask maps and the generated source objects and their corresponding ground truths are mainly in size, while the position, color, texture, pattern, and shadow are very realistic. This resulted in our method not achieving optimal results for MSE and PSNR. However, the position and size of the source object are reasonable. Moreover, the objects far from the camera are smaller, while those closer are larger. Thus, the composite images still look authentic.

4.2.2. Qualitative Experimental Results

Figure 8 shows samples of our composite results of the eight categories of the source objects, which demonstrate that the composite images are very similar to the real images in terms of location, size, shape, texture, color, shadow, and detail. This indicates that our method can generate realistic compositing images for different source objects in different scenes. Particularly, both the geometric pattern details and the characters are consistent with the ground truth source objects. For glossy color source objects, our method can

generate appropriate luster results through integration with the target images. In addition, we can achieve satisfactory shadows of the source objects.

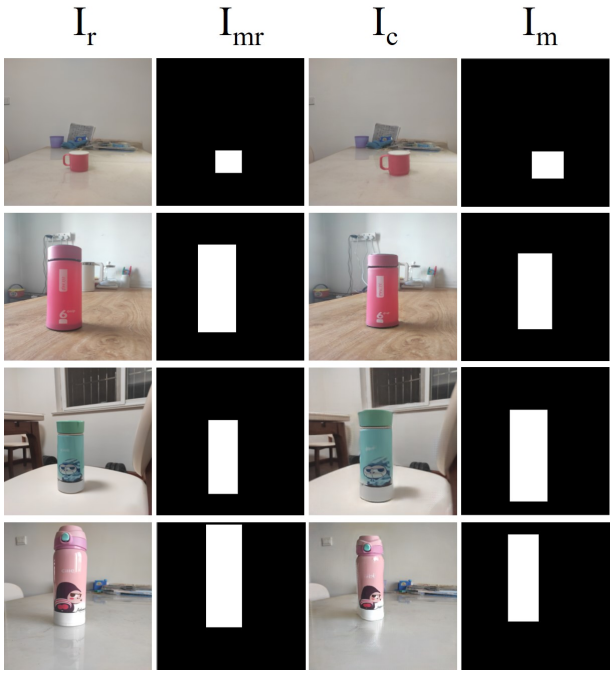

Figure 7. Samples of the mask map. I_r represents the ground truth image, I_{mr} represents the ground truth mask map, I_c represents the compositing image, and I_m represents the predicted mask map.

Figure 8. Samples of composite results of the previous test set (PTS). We test the compositing effect of 8 source objects and different target images. I_s represents the source object, I_t represents the target image, I_c represents the compositing result, and GT represents the ground truth image.

Figure 9 shows that the source objects have been resized and placed on the matching target objects, making the composite images appear reasonable. In addition, the network adjusts the angle of the source objects. Taking the composite image of source object s1 as an example, the white lid of the cup has an inclination angle. This indicates that the source object in the composite image has an angle adjustment in the Z-axis direction relative to the original source object.

Figure 9. Samples of composite results of the GTS. We test the compositing effect of 8 source objects and different target images. I_s represents the source object, I_t represents the target image, and I_c represents the compositing result.

We use data augmentation to avoid overfitting. We annotate the bounding boxes of the source objects in each real image, and randomly crop the samples while ensuring the inclusion of the source objects.

4.3. Ablation Study

We investigate the effectiveness of each learning objective in our method, and Table 3 shows the results. Because of the inherent deviation between the predicted position of the source object and the ground truth, the non-uniqueness of the reasonable location of the source object in the composite image, and no ground truth on the generalization dataset, we utilize score evaluation when measuring the compositing effect.

In Table 3, we can see that the objective evaluation score is lower without the spatial alignment network in the PTS and the GTS. The scores rise noticeably after adding a global discriminator. The addition of a spatial alignment network significantly surges the score of the GTS, indicating an enhanced authenticity and visibility of the compositing images. The local discriminator mainly improves the detail and clarity of the generated source object. Our proposed method reaches the highest score of objective authenticity and also reaches a peak in the GTS score.

Table 3. Comparison results of the SHU dataset with different learning objectives. The best results are denoted in boldface.

Method	PTS Score ↑	GTS Score ↑
L_c	0.3563	0.0601
$L_c + L_{G_g}$	0.9279	0.5960
$L_c + L_p + L_{G_g}$	0.9821	0.7417
$L_c + L_p + L_{G_g} + L_{G_l}$	**0.9958**	**0.8933**

To pursue the optimal performance of our method, we conduct the key parameter study. When adjusting the weight of one loss function, the weights of others are fixed. Firstly, when $\lambda_p = 1$, we set λ_g as four different options: 0.001, 0.01, 0.1, and 1. As shown in Table 4, the optimal result is under $\lambda_g = 0.1$. The PTS scores are similar in other cases, but the GTS scores fluctuate wildly. Then, when $\lambda_g = 0.1$, we set λ_p as five different options: 0.1, 0.5, 1, 5, and 10. Table 5 shows that the optimal performance occurs under $\lambda_p = 1$.

Table 4. Comparison results of the SHU datasets with different values of the key parameter λ_g, which is used to keep the balance of the loss functions. λ_g is set to 0.1 in training. The best results are denoted in boldface.

λ_g	PTS Score ↑	GTS Score ↑
0.001	0.9787	0.6086
0.01	0.9923	0.8231
0.1	**0.9958**	**0.8933**
1	0.9943	0.8555

Table 5. Comparison results of the SHU datasets with different values of the key parameter λ_p, which is used to keep the balance of the loss functions. λ_p is set to 1 in training. The best results are denoted in boldface.

λ_p	PTS Score ↑	GTS Score ↑
0.1	0.9826	0.8564
0.5	0.9936	0.8754
1	**0.9958**	**0.8933**
5	0.9947	0.8807
10	0.9921	0.8641

4.4. The Objective Estimation of Other Source Objects of the SHU Dataset

We conduct the objective estimation of the PTS score and GTS score on other source objects of the SHU dataset. In addition, we perform a user study to further evaluate the perceptual quality of our method, the results of which are shown in Table 6. For different source objects, the PTS score of objective evaluation is higher than 0.9 points, and the GTS score of objective evaluation is higher than 0.8 points. The user study scores have strong subjectivity, since different observers and different environments affect the score results. We can see that the ground truth samples receive scores between 0.7821 and 0.7124. The mean relative error between the PTS score of different source objects and the ground truth is 16.24%. The mean relative error between the GTS score of different source objects and the ground truth is 22.40%.

Table 6. Comparison results of the evaluation of different source objects. The error of the same evaluation metric for different source objects does not exceed 0.1, indicating that the proposed method achieves similar results in image compositing tasks with different source objects.

I_s	Objective Evaluation		User Study		
	PTS Score ↑	GTS Score ↑	Ground Truth Score ↑	PTS Score ↑	GTS Score ↑
S1	0.9958	0.8933	0.7205	0.6128	0.5630
S2	0.9313	0.8281	0.7304	0.6032	0.5652
S3	0.9219	0.8140	0.7135	0.5937	0.5562
S4	0.9653	0.8981	0.7469	0.6265	0.5648
S5	0.9385	0.8672	0.7173	0.6586	0.5913
S6	0.9537	0.8749	0.7821	0.6071	0.5714
S7	0.9150	0.8075	0.7124	0.5965	0.5632
S8	0.9886	0.8927	0.7549	0.6202	0.5826

User study: This user study involves 30 participants. During this study, each participant was shown 240 samples, consisting of the ground truth, the composite image in the PTS, and the GTS in a 1:1:1 ratio. They watch for 3 s and rate the authenticity of each sample. Participants were asked to rate the results based on two levels: 1 for real and 0 for fake. Firstly, we obtain the scores of each source object from participants. Then, we sum them and divide them by the number of participants. The average scores (higher the better) are calculated via Equation (6).

$$S_{av} = \frac{1}{N_p} \sum_{j=1}^{N_p} (\frac{1}{N_s} \sum_{i=1}^{N_s} x) \tag{6}$$

where x is the score of the images of each source object from participants, N_s is the number of samples of each source object, N_p is the number of participants, and S_{av} is the average score of images with the corresponding source object.

5. Conclusions

In this work, we supplemented the multi-task image composition dataset, including position masks for spatial position analysis and a generalization test set for generalization ability testing. We also propose a novel image composition method using spatial position analysis. The target analysis network can effectively segment and classify target objects. The purpose is to filter the target image and choose suitable samples for image composition. According to the target analysis results, the spatial alignment network effectively matched the source object with the target image, and predicted regions with appropriate position and reasonable size. As the experimental results of the SHU dataset and our newly proposed test set show, the trained network achieves good performance in terms of the image composition task, and generates realistic composite images. However, there are some limitations. When compositing a new source object, it is necessary to collect a lot of relevant data and train the model again. That is to say, our method relies on training data, and obtaining a more realistic source object compositing result requires targeted data support. In the future, we will aim to optimize the model. Furthermore, we would like to apply spatial position analysis to improve video compositing performance.

Author Contributions: Conceptualization, X.L. and P.A.; methodology, X.L.; figure/table preparation, X.L.; data curation, X.L. and H.Y.; writing—original draft preparation, X.L. and H.Y.; writing—review and editing, X.L.; visualization, X.L.; supervision, G.T. and P.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant 62071287), the Science and Technology Commission of Shanghai Municipality (Grant 20DZ2290100), and the Anyang Science and Technology Program (Grant 2021C01SF052).

Data Availability Statement: Not applicable.

Acknowledgments: This research was funded by the National Natural Science Foundation of China (Grant 62071287), the Science and Technology Commission of Shanghai Municipality (Grant 20DZ2290100), and the Anyang science and technology program (Grant 2021C01SF052).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SPAN	Spatial position analysis network
STN	Spatial transformation network
GCM	Graph completion module
MIT	Massachusetts Institute of Technology
GTS	Generalization test set
PTS	Previous test set
MSE	Mean squared error
PSNR	Peak signal-to-noise ratio
IoU	Intersection over union

References

1. Chen, B.C.; Kae, A. Toward realistic image compositing with adversarial learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8407–8416.
2. Weng, S.; Li, W.; Li, D.; Jin, H.; Shi, B. Misc: Multi-condition injection and spatially-adaptive compositing for conditional person image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7738–7746.
3. Niu, L.; Cong, W.; Liu, L.; Hong, Y.; Zhang, B.; Liang, J.; Zhang, L. Making images real again: A comprehensive survey on deep image composition. *arXiv* **2021**, arXiv:2106.14490.
4. Lee, D.; Liu, S.; Gu, J.; Liu, M.Y.; Yang, M.H.; Kautz, J. Context-aware synthesis and placement of object instances. In Proceedings of the Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; pp. 10414–10424.
5. Tripathi, S.; Chandra, S.; Agrawal, A.; Tyagi, A.; Rehg, J.M.; Chari, V. Learning to generate synthetic data via compositing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 461–470.
6. Zhang, L.; Wen, T.; Min, J.; Wang, J.; Han, D.; Shi, J. Learning object placement by inpainting for compositional data augmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 566–581.
7. Cong, W.; Zhang, J.; Niu, L.; Liu, L.; Ling, Z.; Li, W.; Zhang, L. DoveNet: Deep image harmonization via domain verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8391–8400.
8. Cun, X.; Pun, C. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Trans. Image Process.* **2020**, *29*, 759–4771. [CrossRef] [PubMed]
9. Tsai, Y.; Shen, X.; Lin, Z.; Sunkavalli, K.; Lu, X.; Yang, M. Deep image harmonization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2799–2807.
10. Wu, H.; Zheng, S.; Zhang, J.; Huang, K. GP-GAN: Towards realistic high-resolution image blending. In Proceedings of the the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2487–2495.
11. Zhang, L.; Gao, Y.; Zimmermann, R.; Tian, Q.; Li, X. Fusion of multichannel local and global structural cues for photo aesthetics evaluation. *IEEE Trans. Image Process.* **2014**, *23*, 1419–1429. [CrossRef] [PubMed]
12. Liu, D.; Long, C.; Zhang, H.; Yu, H.; Dong, X.; Xiao, C. ARshadow-Gan: Shadow generative adversarial network for augmented reality in single light scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8136–8145.
13. Sheng, Y.; Zhang, J.; Benes, B. SSN: Soft shadow network for image compositing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 4380–4390.
14. Zhang, S.; Liang, R.; Wang, M. ShadowGAN: Shadow synthesis for virtual objects with conditional adversarial networks. *Comput. Vis. Media* **2019**, *5*, 105–115. [CrossRef]
15. Li, X.; Teng, G.; Yao, P.A. MT-GAN: Toward realistic image composition based on spatial features. *EURASIP J. Adv. Signal Process.* **2023**, *2023*, 46. [CrossRef]
16. Yuan, Q.; Chen, K.; Yu, Y.; Le, N.Q.K.; Chua, M.C.H. Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding. *Brief. Bioinform.* **2023**, *24*, bbac630. [CrossRef] [PubMed]
17. Kha, Q.; Ho, Q.; Yu, Y.; Le N.Q.K. Identifying SNARE Proteins Using an Alignment-Free Method Based on Multiscan Convolutional Neural Network and PSSM Profiles. *J. Chem. Inf. Model.* **2022**, *62*, 4820–4826. [CrossRef] [PubMed]
18. Zhu, J.Y.; Krahenbuhl, P.; Shechtman, E.; Efros, A.A. Learning a discriminative model for the perception of realism in composite images. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3943–3951.
19. Cong, W.; Niu, L.; Zhang, J.; Liang, J.; Zhang, L. Bargainnet: Background-guided domain translation for image harmonization. In Proceedings of the IEEE International Conference on Multimedia and Expo, Shenzhen, China, 5–19 July 2021; pp. 1–6.
20. Cong, W.; Tao, X.; Niu, L.; Liang, J.; Gao, X.; Sun, Q.; Zhang, L. High-resolution image harmonization via collaborative dual transformations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 5–19 July 2022; pp. 18449–18458.
21. Lin, C.H.; Yumer, E.; Wang, O.; Shechtman, E.; Lucey, S. ST-GAN: Spatial transformer generative adversarial networks for image compositing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9455–9464.
22. Zhan, F.; Zhu, H.; Lu, S. Spatial Fusion GAN for Image Synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3653–3662.
23. Niu, L.; Liu, Q.; Liu, Z.; Li, J. Fast Object Placement Assessment. *arXiv* **2022**, arXiv:2205.14280.
24. Zhou, S.; Liu, L.; Niu, L.; Zhang, L. Learning Object Placement via Dual-path Graph Completion. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 373–389.
25. MA, S.; Shen, Q.; Hou, Q.; Ren, Z.; Zhou, K. Neural compositing for real-time augmented reality rendering in low-frequency lighting environments. *Sci. China-Inf. Sci.* **2021**, *64*, 135–149. [CrossRef]
26. Hong, Y.; Niu, L.; Zhang, J. Shadow generation for composite image in real-world scenes. In Proceedings of the Association for the Advance of Artificial Intelligence, Online, 22 February–1 March 2022; pp. 914–922.

27. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. SegNeXt: Rethinking Convolutional Attention Design for Semantic Segmentation. In Proceedings of the Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 1140–1156.
28. Goodfellow, I.J.; Abadie, J.P.; Mirza, M.; Xu, B.; Farley, D.W.; Ozairy, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
29. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2242–2251.
30. Ulyanov, D.; Vedaldi, A.; Lempitsky, V.S. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv* **2016**, arXiv:1607.08022.
31. Li, X.; Teng, G.; An, P.; Yao, H.Y. Image synthesis via adversarial geometric consistency pursuit. *Signal Process. Image Commun.* **2021**, *99*, 116489. [CrossRef]