electronics

MDPI

Article

Efficient Object Placement via FTOPNet

Guosheng Ye [1,2], Jianming Wang [1,2,*] and Zizhong Yang [1]

[1] Yunnan Provincial Key Laboratory of Entomological Biopharmaceutical R&D, Dali University, Dali 671000, China; ygs@stu.dali.edu.cn (G.Y.); yangzizhong@dali.edu.cn (Z.Y.)
[2] School of Mathematics and Computer Science, Dali University, Dali 671003, China
* Correspondence: wjm@dali.edu.cn

Abstract: Image composition involves the placement of foreground objects at an appropriate scale within a background image to create a visually realistic composite image. However, manual operations for this task are time-consuming and labor-intensive. In this study, we propose an efficient method for foreground object placement, comprising a background feature extraction module (BFEM) designed for background images and a foreground–background cross-attention feature fusion module (FBCAFFM). The BFEM is capable of extracting precise and comprehensive information from the background image. The fused features enable the network to learn additional information related to foreground–background matching, aiding in the prediction of foreground object placement and size. Our experiments are conducted using the publicly available object placement assessment (OPA) dataset. Both quantitative and visual results demonstrate that FTOPNet effectively performs the foreground object placement task and offers a practical solution for image composition tasks.

Keywords: image composition; deep learning; vision transformer; feature fusion; data enhancement

Citation: Ye, G.; Wang, J.; Yang, Z. Efficient Object Placement via FTOPNet. *Electronics* **2023**, *12*, 4106. https://doi.org/10.3390/electronics12194106

Academic Editor: Jenhui Chen

Received: 31 August 2023
Revised: 25 September 2023
Accepted: 29 September 2023
Published: 30 September 2023

1. Introduction

Image composition [1] is the process of combining foreground objects with a background image to create a new image that appears natural and realistic. Composite images have various applications, including poster design, user portrait editing, dataset augmentation, and data enhancement, among others. Traditionally, image composition necessitated manual adjustments, involving resizing the foreground, selecting spatial placement, and determining geometric rotation angles—a laborious process. However, the evolution of computer image processing technologies has witnessed the maturation of deep learning-based image composition methods, delivering superior performance. These methods not only simplify manual tasks but also enhance the realism of composite images.

The completion of image composition involves addressing several sub-problems, including (1) **geometric inconsistency** [1]: This pertains to issues such as the disproportionate sizing of foreground objects, improper positioning, and unrealistic occlusion between the foreground and background scenes; (2) **appearance inconsistency** [2]: this relates to challenges like visual disparities between foreground and background scenes, as well as blurred edge details in the foreground; (3) **semantic inconsistency** [1]: this encompasses scenarios where the foreground and background scenes fail to adhere to real-world logic. Among these sub-problems, addressing geometric inconsistency is often considered the initial step in image composition, particularly with regard to the precise placement of foreground objects within the background image.

Nevertheless, limited attention has been given to the placement of foreground objects in existing research. A few notable approaches, such as task-aware efficient realistic synthesis of example (TERSE) [3], hierarchy composition GAN (HIC-GAN) [4], and the object placement network (PlaceNet) [5], employ adversarial training to learn reasonable distribution information from real images. They subsequently use this information to determine the size and location parameters of foreground objects [6]. Spatial transformer

generative adversarial networks (ST-GANs) [7] leverage the spatial invariance properties of spatial transformer networks (STNs) [8] to enhance the realism of composite images by efficiently transforming foreground objects to match the background scenes. However, these approaches suffer from limitations. They often extract extraneous information when capturing background features, leading to a loss of detailed information and longer processing times due to redundant background feature extraction.

In this paper, we propose the **F**usion **T**ransformer **O**bject **P**lacement **Net**work (**FTOPNet**) to streamline the processing of background image feature information, facilitating image composition. Firstly, we introduce a novel background features extraction module (BFEM) designed to ensure the network captures richer and more accurate background information. This focused attention [9] allows the model to match the foreground with the most relevant areas of the background, ultimately determining the placement, size, and spatial locations of foreground objects. To achieve this, we designed a module that effectively combines global and local information from both foreground and background sources.

To summarize, this paper presents the following main contributions:

1. We introduce the two-stage fusion transformer object placement network (FTOPNet) for foreground object placement in image composition. This network comprises the innovative background feature extraction module (BFEM) and the foreground and background cross-attention feature fusion module (FBCAFFM), both contributing to enhanced plausibility and diversity in composite results.
2. Our BFEM departs from previous methods by not solely relying on global feature extraction. Instead, it captures features from both large and small regions, combining them to yield feature vectors that better match foreground objects, providing crucial information for the final placement.
3. The FBCAFFM is designed to predict the final placement area and foreground object sizes based on cross-attention mechanisms that deeply fuse foreground and background feature information from the BFEM. This results in more accurate hybrid feature representations of the placement results.
4. We retain a simple feature extraction and encoder stage in the model to blend prediction results with hybrid placement predictions. This balances global background image matching information with local foreground object details to yield final predictions, a strategy demonstrated to be highly effective in our experiments.

2. Related Works

In deep learning-based image composition tasks, most methods primarily emphasize the proper placement of foreground objects within a background image, the resizing of foreground objects, and the management of fine details along the object boundaries. These aspects are commonly referred to as geometric consistency and appearance consistency in the context of image composition. Successfully addressing these two tasks lays a solid foundation for various other image composition subtasks, such as image harmonization [10] and foreground object shadow generation [11]. Therefore, the deep learning methods related to geometric consistency and appearance consistency are mainly introduced as follows.

2.1. Geometric Consistency

The geometric consistency refers to whether the geometric information of foreground objects in the background image matches, mainly in terms of how to determine the sizes and positions of foreground objects, and whether foreground objects need to be geometrically rotated to match the perspective, etc. Tan et al. [12] use the network to predict reasonable bounding boxes for foreground objects, detecting both local and global scene appearance information to optimize the placements of these objects. They use alpha matting [13] to ensure that the foreground objects blend smoothly into the background. However, one limitation is that the model does not extract the background image information sufficiently, resulting in the foreground objects being more predominantly positioned in the middle of the background image. That is, the diversity of object placements is not satisfied.

Tripathi et al. designed TERSE [3] with adversarial training to improve realism by judging the resultant images as true or false via a discriminator. However, as with Tan [12], TERSE [3] also does not achieve significant improvement in the diversity of foreground placement. Like them, Zhang et al., in their proposed PlaceNet [5], use generative adversarial networks to predict different location distribution information for foreground objects, further incorporating additional random parameter information to enhance the diversity of foreground object placement. However, the shortcoming is that the image compositing often relies on the foreground objects having similar domain information to the background image.

However, the self-consistent composition-by-decomposition (CoDe) proposed by Azadi et al. [14] overcomes the shortcomings of PlaceNet [5]. The network is no longer limited to domain-similar foreground objects and background images but is able to take images from two different distributions and calculate their joint distributions based on the texture, shape, and other information of the input content. At the same time, the model rotates, scales, and pans the foreground objects and takes part in the content masking in the composition process to obtain a more realistic result image.

2.2. Appearance Consistency

Appearance consistency primarily refers to how the appearance feature information of the foreground objects themselves fits into the background image. This includes blurring the edge details of the foreground objects when they are composited into the background image, and addressing instances where the foreground objects are obscured by other objects in the background image. To achieve more realistic results in composite images, some tasks often address the appearance consistency problem in conjunction with geometric consistency, as shown by CoDe [14].

Chen et al. designed geometrically and color-consistent GANs (GCC-GANs) [2] to address both geometric consistency and color harmony during adversarial learning, thereby managing occlusion and color harmony problems, respectively; the model is also able to automatically compose images from different sources. Tan et al. [15] conducted the image composition operation after discerning the occlusion relationship by estimating the depth information of the foreground objects. They detected the support area (e.g., ground plane) of the foreground objects based on this information and the objects' boundary information. However, its shortcoming is that the model extracts the region of interest (ROI) of the image by semantic segmentation, but inaccurate segmentation can cause the foreground objects to display artifact information, impacting the composite results. To avoid artifact information, Zhang et al. proposed the dense-connected multi-stream fusion image composition (MLF) network [16] for portrait composition tasks. This network can process feature information of portrait foreground and background images at different scales. It pays greater attention to the boundary artifact information induced by imperfect foreground objects' masks and color decontamination, working to enhance the geometric and appearance consistency of the resulting images.

Recently, Zhou et al. innovatively considered the object-placement task as a graph completion problem and proposed the graph completion network (GracoNet) [6], a model based on the graph completion module (GCM), which considers the background image as a graph with multiple nodes, treats foreground objects as special nodes, and inserts foreground nodes into the background graph with appropriate location and size information through model learning, resulting in significant improvement in foreground placement diversity; moreover, a dual-path framework is designed based on the GCM to address the mode collapse problem [17].

3. Proposed Method

To achieve image composition, the network must extract feature information from both foreground objects and the background image. It should also derive suitable placement information for foreground objects based on background feature information, including

spatial positioning. Simultaneously, it should utilize the foreground feature information to determine the sizes of the foreground objects and whether geometric angle adjustments are necessary. Ultimately, the goal is to produce the actual composite images.

Hence, our FTOPNet comprises two stages: one for completing the feature information extraction stage for the foreground objects and the background image, and the other for the adversarial training required to achieve image composition. In the initial stage, foreground objects are obtained based on the input source image and object mask image. Spatial transformer networks (STNs) [8] are employed for geometric and spatial transformations to adapt them to various background images. Subsequently, the foreground objects' encoder performs feature extraction and obtains the foreground feature vector (FFV). While extracting features from the background image, rather than utilizing direct global feature extraction, we adopt the feature extraction approach inspired by the Swin transformer (SwinViT) [18]. This method extracts features from the background image in layers and different local windows, allowing the network to concentrate on patch blocks suitable for foreground object placement, thus reducing the computational overhead to obtain the background feature vector (BFV).

Subsequently, upon completing the feature extraction from foreground objects and background images, they are employed as inputs to the FBCAFFM. Within this module, a multi-head cross-attention mechanism [19] is applied to emphasize the portion of the BFV that aligns well with the FFV. This signifies that foreground objects placed in the background image receive higher scores, and the resulting BFV is output alongside the FFV, forming the cross-attention foreground–background correlation feature vector.

In the second stage, a simple decoder generates the position prediction P_a by receiving only the FFV and BFV. Conversely, the hybrid decoder's input comprises not only the FFV and BFV but also the cross-attention foreground–background association feature vector. Simultaneously, introducing a random vector sampled from the $\mathcal{U}(0,1)$ uniform distribution space enhances the diversity of the foreground object placement [5]. It also provides information about the foreground object's position coordinates in the background image and its size, denoted as P_b. The final prediction result P is obtained through the addition operation. During the training process, the accuracies of the encoder-generated results are enhanced by iteratively updating the focus area of the FBCAFFM. The FTOPNet encompasses these two stages, as depicted in Figure 1.

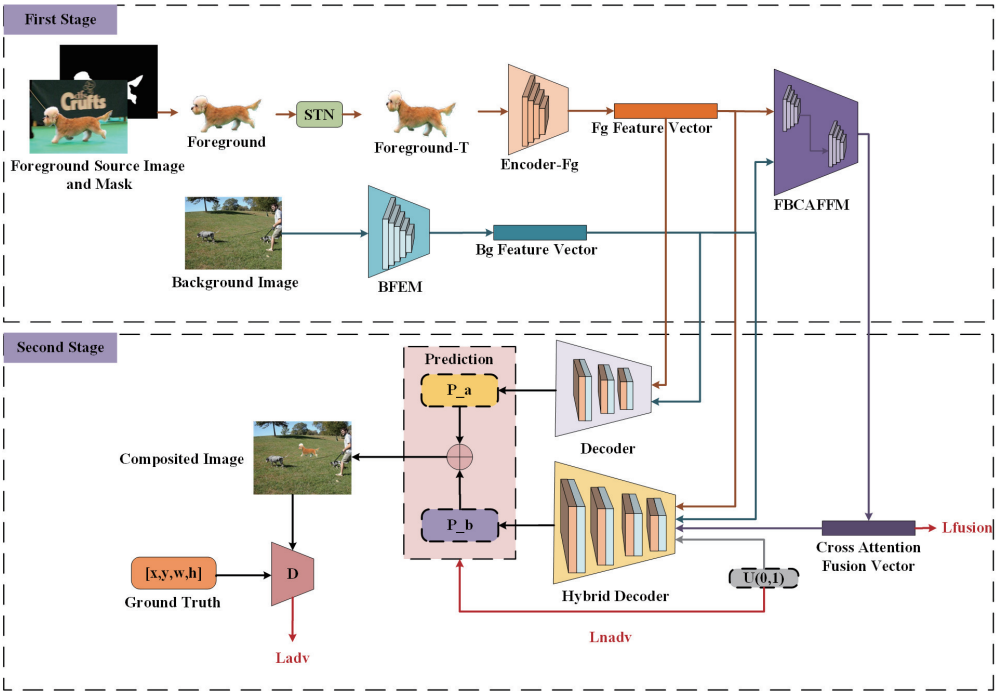

Figure 1. The overall framework of the fusion transformer object placement network (FTOPNet).

3.1. Design of the Background Feature Extraction Module

Unlike PlaceNet [5], TERSE [3], HIC-GAN [4], and some other networks [20] in processing background images, global features are typically extracted directly from the background image. However, this approach is susceptible to the loss of local background feature information and may lead to the neglect of suitable foreground object placement. Therefore, in this work, we propose a reconfiguration of the background image feature extraction process and introduce the BFEM to improve the accuracy of background feature extraction while reducing computational complexity. Our focus is on identifying regions within the background image that are conducive to foreground object placement, as opposed to excessively emphasizing irrelevant or unsuitable local regions. The BFEM module is illustrated in Figure 2.

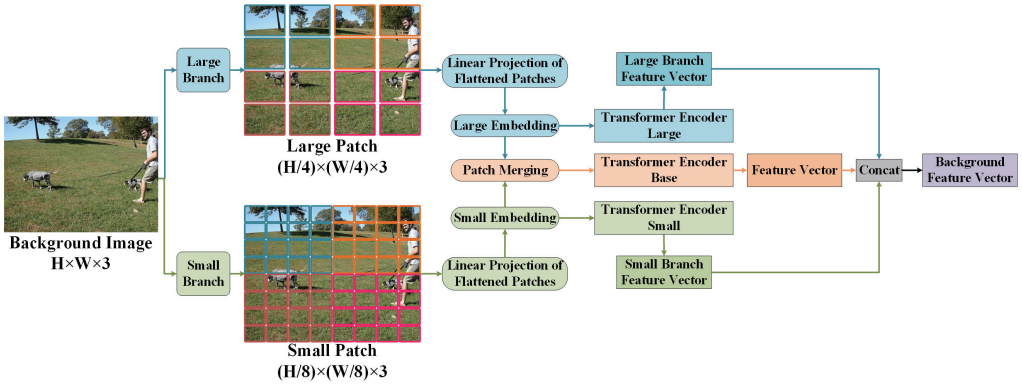

Figure 2. Background feature extraction module (BFEM).

In the BFEM, both large and small branches are utilized to extract features at different scales from the background image. In the large branch, the input image $X^{(H,\,W,\,C)}$ is divided into different patches $X_{large}^{(H/4,\,W/4,\,C)}$ by convolution; these patches are then flattened and sent to the linear projection layer, where the position information of each patch is added to the large embedding. The feature vector V_{large} is then obtained using the large transformer encoder. Similarly, in the small branch, a more detailed image-slicing operation is performed, reducing $X^{(H,\,W,\,C)}$ to $X_{small}^{(H/8,\,W/8,\,C)}$, and the feature vector V_{small} is derived through the small transformer encoder. To complete the fusion of features from the small and large branches at different scales, the two types of embedded patches are merged, and the fused feature vector V_{fusion} is generated using the base transformer. The aforementioned feature vectors V_{large}, V_{small}, and V_{fusion} are concatenated to produce the final feature vector from the background image. The purpose of this is to fully fuse large patch information, small patch information, and fusion feature information, which enables the network to better discover the background areas of suitable foreground objects during training.

The input of the transformer encoder is as follows:

$$Transformer\_Encoder\_Large/Small_{input} = \sum_{i=0}^{N-1} Patch_i\left(p^2 \cdot C\right) \tag{1}$$

where C represents the number of channels, p denotes the size $H \times W$ of each patch in the large branch and small branches, and N is calculated as $(H_{input} \times W_{input})/p^2$.

3.2. Design of Foreground–Background Cross-Attention Feature Fusion Module

After extracting the feature information of the foreground object and the background image, simply decoding the placement position and size based on the feature information of both often results in suboptimal outcomes. Therefore, we designed FBCAFFM to enhance the accuracy of foreground object placement and improve the visual realism of composites.

Inspired by the cross-attention module (CAM) employed in the cross-attention multi-scale vision transformer (CrossViT) [19], which splits images into different scales and employs cross-attention through two paths (the small branch and large branch), using a transformer encoder to derive the final classification results, we adopted a different approach in FBCAFFM. Here, we directly encode the foreground object into the FFV using an encoder. Simultaneously, the background image is encoded using three paths (large, small, and merge) within the BFEM to obtain the BFV. These FFVs and BFVs are then fed directly into the large and small branches in CAM. Additionally, *CLS* (class) tokens are introduced to learn abstract feature information within their respective branches. This setup allows the model to acquire information from different scale levels (background patches) from each other, facilitating enhanced extraction of background features more conducive to foreground object integration. The FBCAFFM is illustrated in Figure 3.

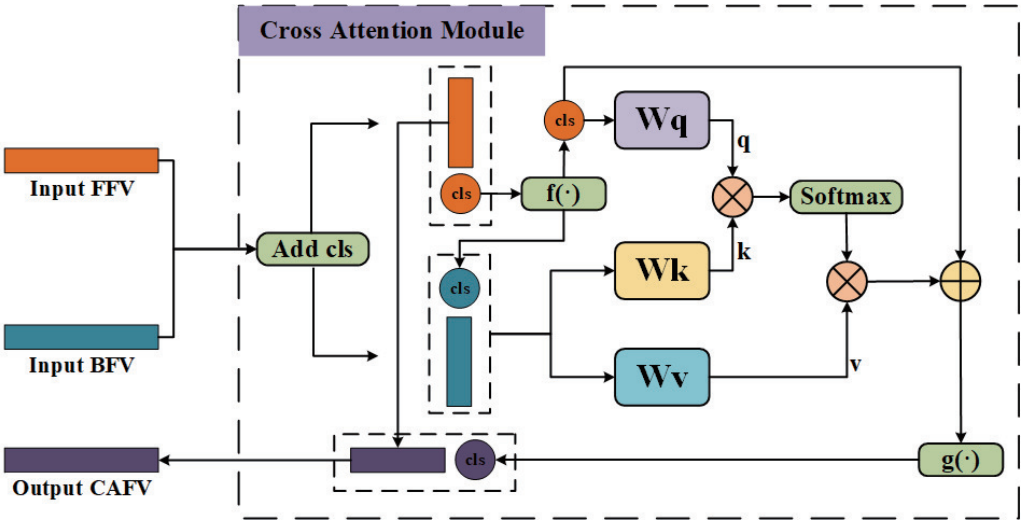

Figure 3. Foreground–background cross-attention feature fusion module (FBCAFFM).

The FFV and BFV obtained from the feature extraction stage have their dimensions changed from $[s, dim\_embedding]$ to match the input feature dimensions of FBCAFFM, resulting in $[s, 256, 96]$ and $[s, 196, 192]$, respectively. With the addition of *CLS* tokens, their dimensions become $[s, 257, 96]$ and $[s, 197, 192]$, which represent the input dimensions of the small branch and large branch, respectively. Then, position embeddings are added, and a multi-scale transformer is employed for feature fusion. Finally, to ensure that the output fusion feature vector can be directly accepted by the decoder in the second stage, its dimension is set to $[s, dim\_embedding]$, where s and $dim\_embedding$ denote the number of sampled random variables [5] and the dimension of embedding, respectively.

As a result, the cross-attention processing can be expressed as follows:

$$x = [f(FFV_{cls}) \,||\, FFV_{vector}],\ y = [f(BFV_{cls}) \,||\, BFV_{vector}] \tag{2}$$

x, y represent the FFV/BFV from small/large branches and concatenate their respective *CLS* tokens, as follows:

$$q = x_{cls}W_q,\ k = yW_k,\ v = yW_v \tag{3}$$

$$Attention = Softmax(\frac{qk^T}{\sqrt{C/H}}) \tag{4}$$

$$CrossAttention_{output}(FFV, BFV) = Attention \cdot v + f(FFV_{cls}) \tag{5}$$

$$CAFV = [g(CrossAttention_{output}) \,||\, FFV_{vector}] \tag{6}$$

This is to obtain the fusion feature vector. Following CrossViT [19], $f(\cdot)$ and $g(\cdot)$ are projections to align dimensions. $||$ denotes the operation of concatenation, C and H are the embedding dimensions and number of heads, W_q, W_k, $W_v \in \mathbf{R}^{(C \times C/H)}$ are linear learnable weights for the query, key, and value, respectively.

After obtaining the cross-attention feature vector (CAFV) by FBCAFFM, following conditional GAN loss [21], we define the fusion adversarial loss to enable the network to fuse the key feature information of the foreground and background and use it for placement location prediction, which is defined as follows:

$$L_{real\_fusion} = \alpha E_{x \in p_{data(x)}} \left[log(D(y|f,b)) + log\left(D\left(y|F_f, F_b, F\right)\right)\right] \tag{7}$$

$$L_{fake\_fusion} = \beta E_{z \in p_{(z)}} \left[log(1 - D(G(z|f,b)|f,b)) + log\left(1 - D\left(G\left(z|F_f, F_b, F\right)|F_f, F_b, F\right)\right)\right] \tag{8}$$

where D denotes the discriminator, G denotes the generator, f denotes the foreground, b denotes the background, y denotes the ground truth of the placement, z is the random variable in $\mathcal{U}(0,1)$ uniform distribution, F_f, F_b, and F denote the fusion of the foreground, background, and both, $G(z|f,b)$ and $G(z|F_f, F_b, F)$ denote the predicted placements, α , β, and λ denote the hyperparameters that we set at 0.9, 0.9.

In addition to that, reasonable prediction of foreground object placement is not our ultimate goal; we expect the network to learn other placement information to satisfy diverse requirements without loss of plausibility. Therefore, we first randomly sample the extracted feature vector BFV and FFV from BEFM in different dimensions to obtain F_f^i, F_f^j, F_b^i, and F_b^j, and dimensionally reconstruct F_f and F_b by fusing the corresponding feature vector. The further fusion of the two, together with the result of CAFV from FBCAFFM, constructs the reconstruction fusion loss, which we denote as $R_{i,j}^F$. Meanwhile, following the diversity loss of PlaceNet [5], we consider the placement prediction results y_a, y_b from the different generators. With the fusion results y, more placement schemes can be obtained by calculating the variation of pairwise distances with random variables [22]. And unlike the single diversity loss [5], which computes the distance between the predicted outcomes y and z, we merged the different predicted outcomes y_a, y_b, y with the distance between z, which we denote as $D_{i,j}^{z,y}$. Finally, the fusion diversity loss is as follows:

$$L_{gan\_fusion\_div}(y, z, F_f, F_b, F) = \lambda \left\| \frac{R_{i,j}^F}{S^2 + S} - \sum_{i=1}^{N} \sum_{i \neq j}^{N} D_{i,j}^{z,y} \right\| \tag{9}$$

$$R_{i,j}^F = \frac{1}{N} \sum_{i=j} \left(\sum_{i \neq j}^{S} F_f^{i,j} + \sum_{i \neq j}^{S} F_b^{i,j} \right) + \sum_{i}^{N} F^i \tag{10}$$

$$D_{i,j}^{z,y} = \frac{1}{N} \left[3D_{i,j}^z - \left(D_{i,j}^{y^a} + D_{i,j}^{y^b} + D_{i,j}^{y^a + y^b} \right) \right] \tag{11}$$

where y_a and y_b denote the predicted locations, z denotes the random variable from $\mathcal{U}(0,1)$ uniform distribution, $F_f^{i,j}$ and $F_b^{i,j}$ denote the feature vector of the foreground and background, F denotes the fusion feature vector, N and S denote the different numbers of sampled random variables in D and R; i,j indicate the sample indices, λ denotes the hyperparameter that we set at 0.3. Following PlaceNet [5], $D_{i,j}^z$, $D_{i,j}^{y^a}$, $D_{i,j}^{y^b}$ are the normalized pairwise distance matrices; they are defined as follows:

$$D_{i,j}^z = \frac{||z_i - z_j||}{\sum_j ||z_i - z_j||}, \; D_{i,j}^{y^a} = \frac{||y_i^a - y_j^a||}{\sum_j ||y_i^a - y_j^a||}, \; D_{i,j}^{y^b} = \frac{||y_i^b - y_j^b||}{\sum_j ||y_i^b - y_j^b||} \tag{12}$$

After designing the fusion adversarial loss and fusion diversity loss, we use θ_G and θ_D to represent the learnable weights in G and D, so the optimization objective can be expressed as follows:

$$\min_{\theta_G} \max_{\theta_D} L_{real\_fusion}(D) + L_{fake\_fusion}(G, D) + L_{gan\_fusion\_div}(y, z, F_f, F_b, F) \quad (13)$$

4. Experiments and Results

In this section, we describe the publicly available datasets used in our study and provide an overview of our experimental process and methodology. Furthermore, we conduct a comparative analysis with existing image composition techniques, followed by an ablation study of our model. Finally, we present visualizations of the experimental results.

4.1. Datasets

In previous studies on image composition, various datasets were employed by researchers to address specific tasks [2–5,7,12,14–16]. Prior to the introduction of GracoNet [6], there was no standardized public dataset available for this purpose. We adopted the OPA dataset for our experiments [23], which is a public dataset that was designed for image composition. To ensure a comprehensive assessment of our model, we exclusively utilized the OPA dataset for both training and testing. The OPA dataset comprises 62,074 composite images for training, of which 21,376 are positive samples and 40,698 are negative samples. For testing, it includes 11,396 composite images, with 3588 positive samples and 7808 negative samples. The dataset contains annotations for 1389 distinct background scenes and 4137 unique foreground objects, spanning across 47 categories. Consequently, our training process involved using the OPA training set to train FTOPNet, and subsequently, we evaluated its performance on the OPA test set, specifically utilizing the 3588 positive samples. During the validation phase, FTOPNet processed the foreground/background pairs of each positive test sample as input, generating 10 composite images through random sampling for quality assessment [6,23].

4.2. Implementation Details

All the images were resized to 256 $\times$ 256 and normalized before being input to the model. The foreground object and background image were converted into feature vectors (FFV and BFV) by their respective encoders. It is important to note that BFV is obtained using BFEM, and the dimensions of FFV and BFV are unified as $[256, 96]$ and $[196, 192]$, respectively. The decoder consists of two groups (linear, BatchNorm, ReLU) that generate $locations_a$ from the input FFV and BFV. Subsequently, FBCAFFM is used to obtain the fusion feature vector CAFV. At this stage, we input the FFV, BFV, CAFV, and a vector randomly sampled from $\mathcal{U}(0, 1)$ into the hybrid decoder. Finally, the output $Locations_b$ of the hybrid decoder underwent an 'addition' operation with $Locations_a$ to obtain the predicted location.

Our model was trained with batch sizes of 16 for 12 epochs on a single RTX A5000 GPU (NVIDIA Corporation, Santa Clara, California, United States). The initial learning rate was set to 1×10^{-4}, and the weight decay value was configured as 5×10^{-4}. We selected the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For the embedding dimension C, we set it to 512, and the fully connected layer (FC) dimension of G and D was also 512. The hyperparameters were configured as follows: the random noise sampling count N was set to 4, the number of sampling points in the fusion feature S was set to 96, and the values of α, β, and λ for the fusion loss were set to 0.9, 0.9, and 0.3, respectively.

4.3. Evaluation Metrics

In order to evaluate the quality of the composite image, we consider the following aspects: (1) the plausibility of the composite images, (2) the disparity with the labeled image, (3) the diversity in foreground placement, and so on. Therefore, following the methodology introduced in GracoNet [6], we employ accuracy [6] and *Fŕechet Inception Distance*

(FID) [24] to assess the plausibility of the resulting images, while *Learned Perceptual Image Patch Similarity* (LPIPS) [25] is utilized to gauge the diversity of foreground object placement. The specific details of each metric are described as follows:

- **Accuracy**. It was first introduced in GracoNet [6], which enables the quantitative assessment of composite image quality. This approach extends the simple yet effective baseline in the OPA (SimOPA) [23] model to evaluate the accuracy of the object placement generation results. The extended model operates as a binary classifier that distinguishes between reasonable and unreasonable object placements. Accuracy is then defined as the proportion of generated composite images classified as positive by the binary classifier during inference [6]. Using the accuracy metric not only streamlines the process of determining the realism of composite images, but also mitigates potential subjective discrepancies among different human evaluators, resulting in more consistent ratings for the same composite images.
- **FID**. To enable the comparison with ground truth, the realism of the placement can be quantified by calculating the FID [24], which measures the dissimilarity between the two distributions. A lower FID suggests that the distribution of composite images is closer to that of the ground truth, indicating a higher likelihood of the image being considered realistic.
- **LPIPS**. Following GracoNet [6], we utilized LPIPS [25] to quantify the diversity of the model's generated results. By sampling the random vector 10 times, they obtained 10 different composite results. LPIPS was calculated for all 10 different results for each test sample, and the average LPIPS was computed across all test samples. Higher LPIPS scores indicate greater diversity among the images, signifying enhanced generative diversity.

4.4. Comparison between Existing Methods

To further validate the effectiveness of the proposed method, three baselines: TERSE [3], PlaceNet [5], and GracoNet [6] were selected for comparison on the OPA dataset. To verify their performance in the image composition task, the redundant modules in TERSE [3] and PlaceNet [5] were removed [6], and the quantitative comparison results are shown in Table 1.

Table 1. Comparison between different methods on the OPA dataset.

Method	Accuracy ↑	FID ↓	LPIPS ↑
TERSE [3]	68.2	47.45	0
PlaceNet [5]	71.5	34.91	13.7
GracoNet [6]	84.1	29.62	20.5
FTOPNet (Ours)	84.4	24.46	11.3

From the results, it is easy to see that our model outperformed the latest model, GracoNet [6], in terms of accuracy and FID values, but the results are not as effective in terms of diversity (LPIPS), which we explain in detail here in the discussion section.

4.5. Ablation Studies

4.5.1. Different Types of Background Encoders

To assess the effectiveness of our proposed BFEM, we conducted a comparison with state-of-the-art feature extraction methods, including ResNet-18 [26], pyramidal convolution (PyConv) ResNet [27], pure ConvNet (ConvNeXt) [28], and vision transformer (ViT) [29]. The experimental results are presented in Table 2.

Table 2. Comparison between different types of encoders in background feature extraction.

Method	Accuracy ↑	FID ↓	LPIPS ↑
ResNet-18 [26]	81.3	29.72	6.5
PyConv-ResNet [27]	76.6	24.92	16.7
ConvNeXt [28]	82.8	29.42	5.8
ViT [29]	80.7	27.55	8.1
BFEM (Ours)	84.4	24.46	11.3

When it comes to feature extraction from the background image, different modules exhibit notable differences in performance. For instance, upon employing PyConv-ResNet [27], the LPIPS score for the composite result reaches the highest value of 16.7, surpassing the 13.7 achieved by PlaceNet [5]. However, the results for accuracy and FID are less satisfactory. ResNet-18 [26], ConvNeXt [28], and ViT [29] all achieve high accuracy scores of 80, but their FID results are not particularly outstanding, and they exhibit poor diversity. On the other hand, when BFEM is utilized as the feature extraction module for the background image, it leads to ideal accuracy and FID values, with an acceptable LPIPS score.

4.5.2. Different Values of Hyperparameters

In the FBCAFFM, the random vector z and the number of hybrid feature samples s play pivotal roles in determining the diversity and plausibility of the composite results. The purpose of z is to aid in achieving foreground placement by generating a random sampling vector from the $\mathcal{U}(0,1)$ space. On the other hand, the hybrid feature-sampling parameter s is responsible for defining the feature dimensionalities of FFV and BFV to ensure they can capture sufficient and precise feature information. We conducted a comparison using different values of z and s, and the results are presented in Tables 3 and 4. Ultimately, we selected $z = 4$ and $s = 96$ for all our experiments.

Table 3. Comparison with hyperparameter $\mathcal{Z}$.

$\mathcal{Z}$	Accuracy ↑	FID ↓	LPIPS ↑
2	70.3	24.73	9.6
4 (Ours)	84.4	24.46	11.3
6	77.2	36.37	4.4
8	77.5	26.52	8.2
12	77.7	26.67	8.8

Table 4. Comparison with hyperparameter $\mathcal{S}$.

$\mathcal{S}$	Accuracy ↑	FID ↓	LPIPS ↑
16	67.1	19.84	15.3
32	77.0	28.31	8.3
64	71.7	22.82	10.8
80	81.0	27.83	7.8
96 (Ours)	84.4	24.46	11.3

The foreground and background features undergo distinct feature extraction modules, resulting in feature dimensions of $[256, 96]$ and $[196, 192]$, respectively. To maintain uniformity in random sampling across the same feature dimension, we selected a value of dim = 2 in FFV as the maximum value for s. Subsequently, for different comparisons, we sequentially decrease s by 16.

4.5.3. Practicality of Different Loss Functions

In Section 3.2, we introduced the fusion adversarial loss, which modifies the conditional GAN Loss [21] to create L_{real_fusion} and L_{fake_fusion} by incorporating fusion foreground features, fusion background features, or both. To validate the effectiveness of the

fusion adversarial loss function, we conducted experiments where we removed the parts with fusion features while keeping the original components for comparison. These experiments are presented in the first and second rows of Table 5. Additionally, we designed the fusion diversity loss $L_{gan\_fusion\_div}$ to enhance the diversity of foreground object placement. To assess whether it improves the LPIPS score, we conducted an experiment in the third row of Table 5 where we removed it.

Furthermore, in rows 4 and 5, we retained the fusion adversarial loss and fusion diversity Loss, respectively, to investigate their combined impact on composite results. The experiment results in rows 6 demonstrate that both fusion adversarial loss and fusion diversity loss with fusion feature significantly contribute to enhancing the plausibility and diversity of the composite results.

Table 5. Comparisons between the loss functions of different combinations.

Method	Accuracy ↑	FID ↓	LPIPS ↑
w/o L_{real_fusion}	77.9	27.17	8.3
w/o L_{fake_fusion}	62.8	26.33	8.1
w/o $L_{gan\_fusion\_div}$	70.0	26.27	6.6
w/ $L_{fusion\_adversarial}$	74.2	27.07	1.9
w/ $L_{gan\_fusion\_div}$	82.3	27.50	7.6
All	84.4	24.46	11.3

In addition, we incorporated the fusion diversity loss into the objective function to act as a balancing factor during the optimization process. This allows the network to effectively utilize the diversity of fusion learning. We introduced the parameter λ in Equation (9) to control the impact of the fusion diversity loss. We explored a range of values, from $1/\lambda$ to $[1, 5]$, and the results in Table 6 demonstrate the influence of different lambda values on the final results. Consequently, we selected $\lambda = 1/3$ for all our experiments, as it yielded the most plausible and diverse composite images.

Table 6. Coefficient λ in fusion diversity loss.

1/λ	Accuracy ↑	FID ↓	LPIPS ↑
1	71.3	38.63	7.0
2	82.3	26.17	8.1
3 (Ours)	84.4	24.46	11.3
4	82.6	22.76	10.7
5	71.9	20.24	13.9

4.5.4. Finalization of Placement Parameters

In FTOPNet, the FFV and BFV are retained, and predicted parameters *P_a* are obtained by the decoder, as well as predicted parameters *P_b* obtained by the hybrid decoder. To obtain optimal parameter information for foreground placements, we applied different treatments to *P_a* and *P_b*, including *P_a*-only , *P_b*-only, concatenation operation, and addition operation. The results for the different treatments of the placement parameters are presented in Table 7.

Table 7. Comparisons between different operations of placement results.

Method	Accuracy ↑	FID ↓	LPIPS ↑
only *P_a*	64.7	24.96	15.5
only *P_b*	58.9	46.63	1.5
concat (*P_a*,*P_b*)	61.4	26.67	15.0
add (*P_a*,*P_b*) (Ours)	84.4	24.46	11.3

4.6. Visualization of Foreground Object Placement Results

To visually demonstrate the compositional effects of different models, we first randomly selected six composite images from the pool of results. These composite images feature various combinations of foreground objects and background images. The comparison results of different models are presented in Figure 4. Subsequently, we showcase the results of placing different foreground objects within the same background images in Figure 5. Additionally, in Figure 6, we display the results of placing the same foreground objects against different background images.

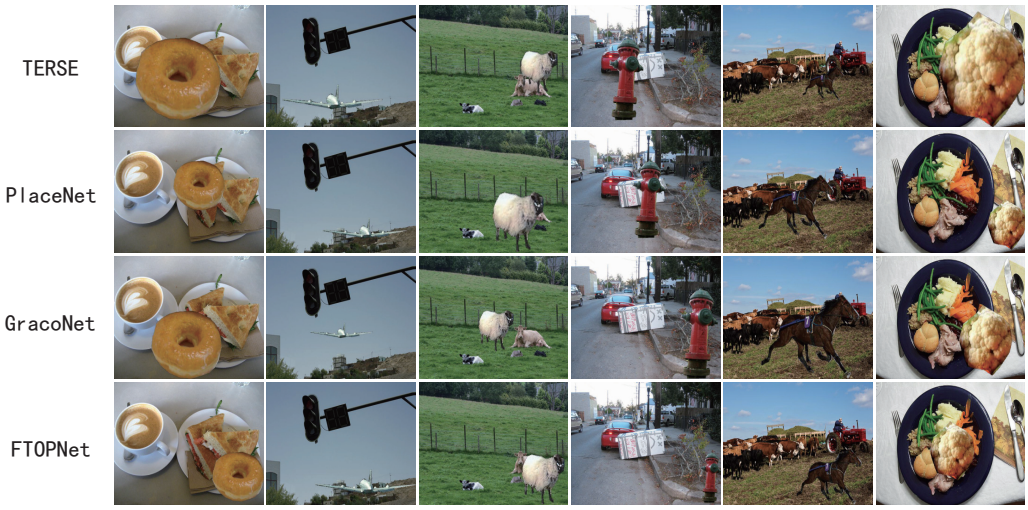

Figure 4. Visualization of object placement results for different methods on the OPA test set with different foreground objects and different background images.

Figure 5. Visualization of object placement results for different methods on the OPA test set with different foreground objects and the same background images.

Figure 6. Visualization of object placement results for different methods on the OPA test set with same foreground objects and different background images.

To assess the diversity in the placement of foreground objects, we conducted 10 placements using the same background and foreground objects. In Figure 7, we present the comparison results of different models, showcasing 6 of these placements.

Figure 7. Visualization of object placement diversity results for different methods on the OPA test set (fire hydrant) by sampling different random vectors.

From the results presented in Figures 7 and 8, it is evident that TERSE [3] consistently produces identical placement results in multiple instances, indicating a lack of diversity in its placement ability. This is reflected in its LPIPS result, which is 0. In contrast, the placement results of other models vary, and the greater the variation, the higher the LPIPS score, suggesting a stronger diversity in their placement abilities.

Although the FTOPNet LPIPS scores in the quantitative results may not be as impressive as those of GracoNet [6] or even PlaceNet [5], one possible explanation is that in some composite images, the available positions for placing foreground objects are very limited. The model may have recognized that offering too many position choices would compromise the overall plausibility of the composite images, resulting in relatively stable position choices.

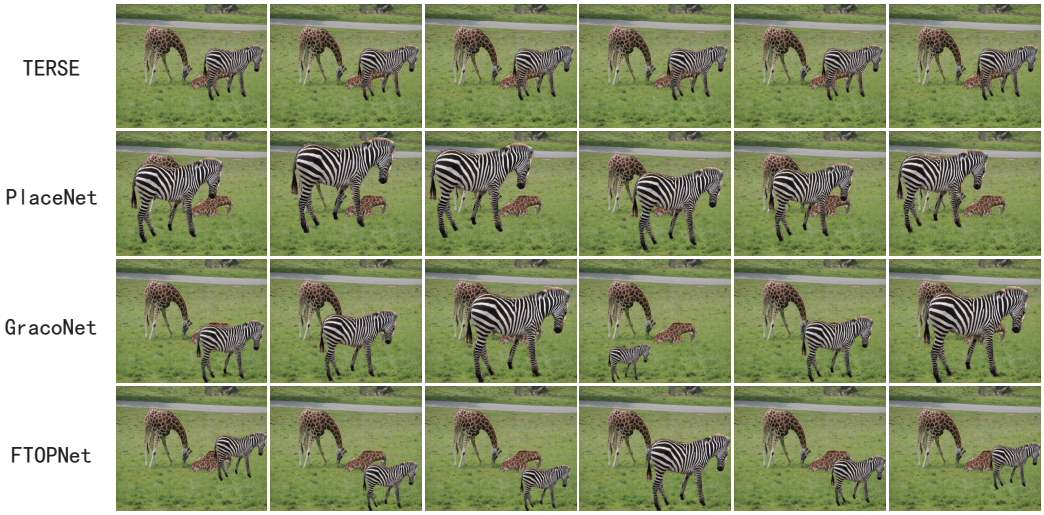

Figure 8. Visualization of object placement diversity results for different methods on the OPA test set (zebra) by sampling different random vectors.

5. Discussion

From the experimental results, it becomes evident that in the domain of image composition, exemplified by TERSE [3], HIC-GAN [4], and PlaceNet [5], extracting global feature information solely from the background image often leads to suboptimal performance in the final composition. This limitation is rooted in the network's tendency to overlook critical interaction information between foreground objects and background features. In contrast, our FTOPNet benefits from an enhanced background information extraction module and the fusion of foreground and background features. As a result, the network effectively captures the association information between foreground and background elements, facilitating more accurate and intuitive placement predictions. Although our results demonstrate modest enhancements over GracoNet [6] concerning accuracy and FID, they still exhibit a deficiency in diversity. This implies that our FTOPNet is designed with a strong emphasis on goal-oriented training, potentially hindering its flexibility in generating a wider range of diverse composite results. The model tends to favor compositions that closely resemble real images, avoiding overly diverse outcomes that might compromise realism (as illustrated in Figure 7, where the lower-right corner demonstrates a realistic foreground placement).

Moreover, the current state of image composition research fails to address a fundamental challenge: the automatic and precise matching of foreground and background images. While our approach instructs the network to compose paired images through tagging, the ideal image composition model should possess the capability to autonomously select suitable pairs from the pool of foreground objects and background images, subsequently completing the composition task. This necessitates that the model acquire additional logical reasoning abilities to comprehend high-level semantic and contextual information pertaining to foreground objects and background scenes.

Related studies, such as CAIS [30], UFO [31], Li et al. [32] , Wu et al. [33], and GALA [34], have explored foreground object search (FoS) within the context of matching the foreground object to the target background image. These approaches typically involve marking the location information using bounding boxes within the background image. Subsequently, they employ techniques like knowledge distillation [32,33] or adversarial learning [31,34] to search for suitable foreground objects from predefined categories. However, this strategy introduces a conflict with the image composition task, as it requires manual annotation of location information, which should ideally be learned automatically by the composition network. Therefore, future research in image composition should explore avenues to address the FoS challenge, with the goal of establishing a more holistic and perfected composition process.

6. Conclusions

In this paper, we introduced the fusion transformer object placement network (FTOPNet) as a novel solution for image composition tasks. FTOPNet comprises two key components: a background feature extraction module (BFEM) and a foreground–background cross-attention feature fusion module (FBCAFFM). These modules are designed to enhance the plausibility and diversity of foreground object placement in composite images. To assess the model's performance in image composition, we conducted experiments using the publicly available OPA dataset. Our results demonstrate that FTOPNet excels in the task of foreground object placement during image composition. Furthermore, there is significant potential for future improvements to FTOPNet. Specifically, we aim to further enhance the diversity of foreground object placement while maintaining the plausibility of the results.

In conclusion, our research is dedicated to the task of image composition, specifically foreground object placement, and it efficiently accomplishes data augmentation with a small dataset. This not only reduces time and labor costs but also provides a substantial data foundation for our subsequent computer vision research tasks, such as image classification, image segmentation, and object detection. Currently, numerous research efforts have focused on utilizing composite image data, yielding excellent results. Our work can greatly facilitate these efforts in dataset creation. Furthermore, image composition has wide-ranging applications in various aspects of our daily lives, including user portrait editing, poster and advertising production, game promotion, and autonomous driving, among others.

Author Contributions: Conceptualization, G.Y. and J.W.; methodology, G.Y.; software, G.Y.; validation, G.Y. and J.W.; formal analysis, G.Y.; investigation, J.W.; resources, G.Y.; data curation, G.Y.; writing—original draft preparation, G.Y.; writing—review and editing, G.Y.; visualization, G.Y. and J.W.; supervision, G.Y. and J.W.; project administration, G.Y. and Z.Y.; funding acquisition, Z.Y. and G.Y. 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 no. 32001313), the Yunnan Fundamental Research Project (grant no. 202201AT070006), the Open Project of the Yunnan Provincial Key Laboratory of Entomological Biopharmaceutical R&D (grant no. AP2022008), the Yunnan Postdoctoral Research Fund Project (grant no. ynbh20057), and the Fundamental Research Joint Special Youth Project of Local Undergraduate Universities in Yunnan Province (grant no. 2018FH001-106).

Data Availability Statement: Not applicable.

Acknowledgments: We thank the Center for Brain-Like Computing and Machine Intelligence (BCMI), Shanghai Jiao Tong University (SJTU), for their great contributions to the image composition work based on deep learning.

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

References

1. 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* **2022**, arXiv:2106.14490.
2. Chen, B.C.; Kae, A. Toward realistic image compositing with adversarial learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8415–8424.
3. 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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 461–470.
4. Zhan, F.; Huang, J.; Lu, S. Hierarchy composition gan for high-fidelity image synthesis. *arXiv* **2019**, arXiv:1905.04693.
5. 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 Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 566–581.
6. Zhou, S.; Liu, L.; Niu, L.; Zhang, L. Learning Object Placement via Dual-Path Graph Completion. In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XVII; Springer: Berlin/Heidelberg, Germany, 2022; pp. 373–389.

7. 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 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9455–9464.
8. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9.
9. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
10. Liu, S.; Huynh, C.P.; Chen, C.; Arap, M.; Hamid, R. LEMaRT: Label-efficient masked region transform for image harmonization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, QC, Canada, 18–22 June 2023; pp. 18290–18299.
11. Hong, Y.; Niu, L.; Zhang, J. Shadow generation for composite image in real-world scenes. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 914–922.
12. Tan, F.; Bernier, C.; Cohen, B.; Ordonez, V.; Barnes, C. Where and who? automatic semantic-aware person composition. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1519–1528.
13. Chen, Q.; Li, D.; Tang, C.K. KNN matting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2175–2188. [CrossRef] [PubMed]
14. Azadi, S.; Pathak, D.; Ebrahimi, S.; Darrell, T. Compositional gan: Learning image-conditional binary composition. *Int. J. Comput. Vis.* **2020**, *128*, 2570–2585. [CrossRef]
15. Tan, X.; Xu, P.; Guo, S.; Wang, W. Image composition of partially occluded objects. *Comput. Graph. Forum* **2019**, *38*, 641–650. [CrossRef]
16. Zhang, H.; Zhang, J.; Perazzi, F.; Lin, Z.; Patel, V.M. Deep image compositing. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2021; pp. 365–374.
17. Zhu, J.Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A.A.; Wang, O.; Shechtman, E. Toward Multimodal Image-to-Image Translation. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
18. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
19. Chen, C.F.R.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 357–366.
20. Zhan, F.; Zhu, H.; Lu, S. Spatial fusion gan for image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3653–3662.
21. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
22. Liu, S.; Zhang, X.; Wangni, J.; Shi, J. Normalized diversification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10306–10315.
23. Liu, L.; Liu, Z.; Zhang, B.; Li, J.; Niu, L.; Liu, Q.; Zhang, L. OPA: Object placement assessment dataset. *arXiv* **2021**, arXiv:2107.01889.
24. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–12.
25. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Duta, I.C.; Liu, L.; Zhu, F.; Shao, L. Pyramidal convolution: Rethinking convolutional neural networks for visual recognition. *arXiv* **2020**, arXiv:2006.11538.
28. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
30. Zhao, H.; Shen, X.; Lin, Z.; Sunkavalli, K.; Price, B.; Jia, J. Compositing-aware image search. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 502–516.
31. Zhao, Y.; Price, B.; Cohen, S.; Gurari, D. Unconstrained foreground object search. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2030–2039.
32. Li, B.; Zhuang, P.Y.; Gu, J.; Li, M.; Tan, P. Interpretable foreground object search as knowledge distillation. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXVIII; Springer: Berlin/Heidelberg, Germany, 2020; pp. 189–204.

33. Wu, Z.; Lischinski, D.; Shechtman, E. Fine-grained foreground retrieval via teacher-student learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 3646–3654.
34. Zhu, S.; Lin, Z.; Cohen, S.; Kuen, J.; Zhang, Z.; Chen, C. GALA: Toward Geometry-and-Lighting-Aware Object Search for Compositing. In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XXVII; Springer: Berlin/Heidelberg, Germany, 2022; pp. 676–692.