electronics

MDPI

Article

High-Resolution Image Inpainting Based on Multi-Scale Neural Network

Tingzhu Sun [1,2], Weidong Fang [3], Wei Chen [1,2,4,*], Yanxin Yao [5], Fangming Bi [1,2] and Baolei Wu [1,2]

[1] School of Computer Science and Technology, China University of Mining Technology, Xuzhou 221000, Jiangsu, China; stz0526@163.com (T.S.); bfm@cumt.edu.cn (F.B.); blwu@cumt.edu.cn (B.W.)
[2] Mine Digitization Engineering Research Center of the Ministry of Education, China University of Mining and Technology, Xuzhou 221116, Jiangsu, China
[3] Key Laboratory of Wireless Sensor Network & Communication, Shanghai Institute of Micro-system and Information Technology, Chinese Academy of Sciences, Shanghai 201800, China; weidong.fang@mail.sim.ac.cn
[4] School of Earth and Space Sciences, Peking University, Beijing 100871, China
[5] School of Communication and Information Engineering, Beijing Information Science & Technology University, Beijing 100101, China; yanxin_buaa@126.com
[*] Correspondence: chenwdavior@163.com; Tel.: +86-1392-176-1978

Received: 10 October 2019; Accepted: 15 November 2019; Published: 19 November 2019

check for updates

Abstract: Although image inpainting based on the generated adversarial network (GAN) has made great breakthroughs in accuracy and speed in recent years, they can only process low-resolution images because of memory limitations and difficulty in training. For high-resolution images, the inpainted regions become blurred and the unpleasant boundaries become visible. Based on the current advanced image generation network, we proposed a novel high-resolution image inpainting method based on multi-scale neural network. This method is a two-stage network including content reconstruction and texture detail restoration. After holding the visually believable fuzzy texture, we further restore the finer details to produce a smoother, clearer, and more coherent inpainting result. Then we propose a special application scene of image inpainting, that is, to delete the redundant pedestrians in the image and ensure the reality of background restoration. It involves pedestrian detection, identifying redundant pedestrians and filling in them with the seemingly correct content. To improve the accuracy of image inpainting in the application scene, we proposed a new mask dataset, which collected the characters in COCO dataset as a mask. Finally, we evaluated our method on COCO and VOC dataset. the experimental results show that our method can produce clearer and more coherent inpainting results, especially for high-resolution images, and the proposed mask dataset can produce better inpainting results in the special application scene.

Keywords: image inpainting; content reconstruction; instance segmentation

1. Introduction

Every day, about 300 million pictures are captured and shared on social networks, and a large part of them are human-centered pictures (including selfies, street photos, travel photos, etc.,). Many human-related research directions have been produced in computer vision and machine learning in recent years. Among them, target tracking [1] (including pedestrian detection [2], pedestrian reidentification [3], human pose estimation [4], etc.,), face recognition [5], and face image inpainting (including pet eye fix [6], eye-closing to eye-opening [7], etc.,) are the research hotspots. Many researchers devote themselves in improving the performance of the existing network. However, integrating existing researches and enabling them to solve common problems in life is also of high practical significance.

In our social networks, we can often see street photos as shown on the left of Figure 1, but actually, the original images seem as shown on the right side of Figure 1. We can see redundant pedestrians in their background destroying beauty and artistic conception of the image. So the purpose of our study is to delete the redundant pedestrian in the image and ensure the reality of background inpainting. It involves pedestrian detection, identifying redundant pedestrians and filling in them with the seemingly correct content. This is a challenging problem because (1) the result largely depends on the accuracy of redundant pedestrian detection; (2) the diversity of background information under the redundant pedestrian area is difficult to recover; (3) the training data lacks real output samples to define the reconstruction loss. We want to deploy our work in the real world as a working application, so we took an interactive approach, removing unnecessary sections by manually selecting unnecessary pedestrian areas after pedestrian detection. After the user removed the unnecessary parts, our algorithm successfully filled the remaining holes with the surrounding background information.

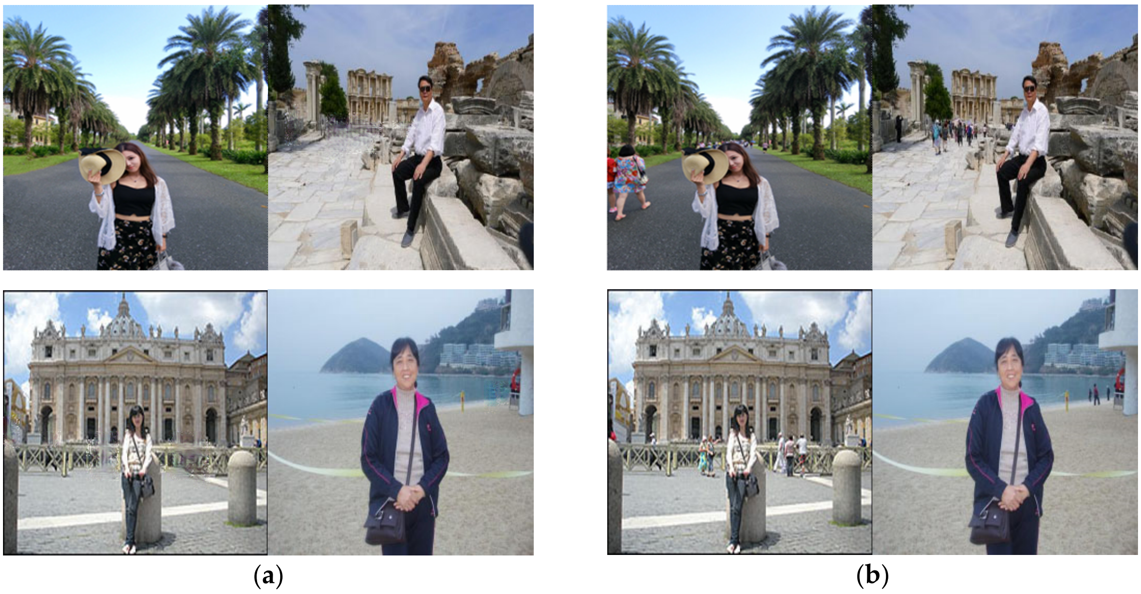

Figure 1. Street photo of the paper's special application scene. (**a**) The repaired figure, (**b**) the figure that is not repaired.

To counter the problems above, we combine the research of instance segmentation, image inpainting. Firstly, we need to complete the instance segmentation of human which detects the regions existing characters. Then it needs us to identify the target character and "protect" the region existing the target character. Finally, we use an image inpainting algorithm to repair other regions. To further improve the inpainting result of the task, we build a new mask dataset, which collects the characters in COCO dataset as a mask, representing various pose. The new mask dataset can produce a better inpainting result on character filtering tasks.

In recent years, deep network has achieved high-quality results in instance segmentation, image inpainting, and so on. Instance segmentation is the combination of object detection and semantic segmentation. First, it uses an object detection algorithm to locate each object in the image with positioning boxes. And then it adapts a semantic segmentation algorithm to mark the target objects in different positioning boxes to achieve the purpose of instance segmentation. The latest instance segmentation is Mask-R-CNN [8], which adds a mask branch of predictive segmentation for each region of interest based on Faster-R-CNN [9]. The mask branch only adds a small computational overhead but supports rapid systems and quick experiments.

Image inpainting can be defined as entering an incomplete image and filling in the incomplete area with semantically and visually believable content. Since Deepak [10] et al. adapts encoder-decoder to complete the inpainting of face images, image inpainting has two transformations from dealing with a fixed shape region to dealing with any non-central and irregular region [11,12], and from distortion to a smooth and clear inpainting result [13,14]. Image inpainting based on GAN network has made great progress in recent years, but for the high-resolution images, the inpainting results will still appear blurred texture and the unpleasant boundaries that are inconsistent with the surrounding area. We found SRGAN [15] has proved superiority on restoring finer texture details, therefore we put forward a new method based on deep generative model. The methods is a two-stage network consisting of content reconstruction and texture detail restoration. We further restore the finer texture details inspired by the architecture of SRGAN after holding the visually believable fuzzy texture. It can effectively solve the problem of structural distortion and texture blur to improve the quality of image inpainting.

In this paper, we propose a high-resolution image inpainting method based on the multi-scale neural network and build a new mask dataset for the special application scene. The main contributions of this paper are:

(1) Based on the current most advanced image inpainting network, we build a texture detail restoration network to restore the details of high-resolution images inspired by SRGAN. The experimental results show that our method can generate a smoother, clearer and more coherent inpainting result than other methods.

(2) To remove unnecessary pedestrians from the image, we proposed a new mask dataset, which collected various pose and could produce better inpainting results in the task of character filtering.

To train our network, we applied the new mask dataset to simulate the real pedestrian. Although we just built the mask dataset to represent the removed pedestrian area, we achieved good results in the real-world data.

The rest is organized as follows. The second part introduces the research status of instance segmentation and image inpainting at home and abroad. The third part describes the improved image inpainting network, and describes the construction of the mask dataset and the special application scene. The fourth part gives the experimental results. The fifth part gives the conclusion.

2. Related Works

In the past ten years, computer vision has made great progress in image processing tasks such as classification, target detection, segmentation, and so on. The performance of deep network has been greatly improved in these tasks, which lays a foundation for the new research difficult problems of image processing and provides support for image inpainting in this paper. We briefly review the relevant work in various sub-areas related to this article.

Instance segmentation integrates image classification, image segmentation, and target detection in computer vision. The earliest region-based CNN(R-CNN) [16] detecting object with a bounding box is to process a certain number of candidate object regions on each ROI independently. Faster-R-CNN [9] based on R-CNN improves by learning the attention mechanism of Region Proposal Network (RPN). Faster-R-CNN is flexible and robust for many later improvements [17–19], and leads the several current benchmarks. Li [20] et al. combines the two types of score map [21] and the target detection [22] to realize “full convolution instance segmentation” (FCIS). Different from the usual method, which predicts a set of position-sensitive channels with full convolution, this method abandons full connected layers for the shared subtasks of image segmentation and image classification, making the network more lightweight. In addition, no trainable parameters exist in either the integrated score map or the result, only the classifier exists. The Mask-R-CNN [8] used in the paper adds a mask branch of predictive segmentation to each region of interest based on Faster-R-CNN [9]. The mask branch only adds a small computational overhead and supports rapid systems and quick experiments.

Traditional inpainting approaches based on diffusion or patch typically use variational algorithms or patch similarity to spread information from background to holes, such as [23,24]. One of the most advanced methods for image inpainting at present is PatchMatch [25], without the use of deep learning, which fills in holes with statistical data of available images through iterated search for the most suitable patch. Although it produces a smoother result, it assumes the texture of the inpainting area can be found elsewhere in the image. This assumption does not always hold. Therefore, it is good at restoring patterned regions, such as background reconstruction, but has difficulty in reconstructing locally unique patterns.

Generative adversarial network makes the research of image inpainting to a peak. Vanilla GANs [26] shows good performance in generating clear images, but has difficulty extending to higher-resolution images due to the instability of training. Several techniques for stable training processes have been proposed, including DCGAN [27], energy-based GAN [28], Wasserstein GAN (WGAN) [29,30], WGAN-GP [31], BEGAN [32], and LSGAN [33]. A more relevant task of image inpainting is conditional image generation. For example, Pix2Pix [34], Pix2Pix HD [35], and CycleGAN [36] transform images in different domains using paired or unpaired data.

The commonly used loss function of image inpainting based on generative adversarial network is a combination of adversarial loss and L2 loss. L2 loss can excite the output of the generated network with variance computing, but cannot capture the high-frequency details and repair the clear texture structure. So the introduction of adversarial loss can effectively solve the problem.

The basic model of image inpainting based on generative adversarial networks is the encoder-decoder used by Deepak [10] et al. To improve the inpainting result of face images, it combines L2 loss with adversarial loss. The latest and effective image inpainting models based on deep learning are mostly developed on this basis. However, the shape of the repaired region is fixed so it has a strong limit in practical application. In response to this question, Liu [12] et al. introduces partial convolution, which can process any non-central, irregular region. However, the method still needs to establish a mask dataset based on deep neural network and conduct pre-training on the irregular masks of random lines. Iizuka [11] uses dilated convolution to increase the receptive field, which obtains the image information in a larger range as much as possible without missing extra information. This method is suitable for solving the inpainting problem of non-center and irregular region, but it has poor inpainting result on structural objects. In recent years, GAN has made a great breakthrough in the application of image inpainting. In the future, there will be more research progress on image inpainting based on deep learning.

Yu [14] et al. improves the generated network of image inpainting based on Iizuka's research [11], and proposes a unified feedforward generation network with a novel context attention layer. The proposed network consists of two phases. The first phase is to roughly extract the missing content after reconstruction loss training with dilated convolution. The second phase is to integrate the context attention. The core idea of context attention is to use the characteristics of known patch as convolution filters to generate patch. The two generating networks are similar to UNET.

Inspired by [14], the issue is divided into two subtasks: (1) The first subtask uses the context encoder (CE) [10] to fill in the large areas need repaired according to the environmental information. (2) As CE cannot recover high-frequency details, the second subtask uses a network similar to SRGAN to capture high-frequency details.

3. The Approach

3.1. Improved Image Inpainting Network

We first constructed our generated network of image inpainting by copying and improving the most advanced inpainting [14] model recently. The network shows a good inpainting result in natural images.

Our improved image inpainting network is shown in Figure 2. We follow the same input and output configurations as in [14] for training and inference. The improved generated network takes an image with white pixels filled in the holes A and a binary mask indicating the hole regions as the input pair, and it outputs the final repaired image. The size of the input image is 256 × 256 and the size of the output image is also 256 × 256. We trained our improved image inpainting network on two datasets including COCO and VOC datasets, as described in the Section 4.

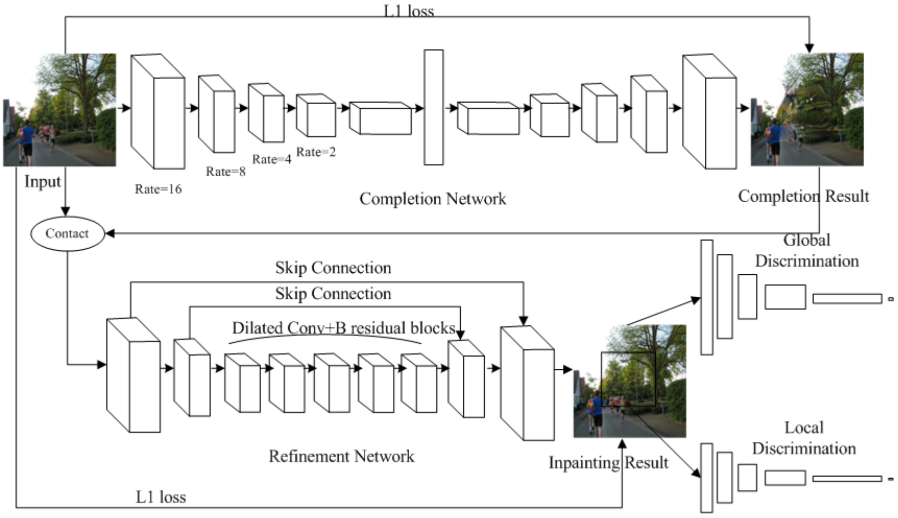

Figure 2. Improved image inpainting network.

To further improve the visual effect of high-resolution image inpainting and reduce the blurred texture and the unpleasant boundaries that inconsistent with the surrounding area, the network we introduced consists of two stages: 1) content reconstruction network, 2) texture details restoration network. The first network is a completion network used to complete the image and obtain the rough prediction results, and it adopts reconstruction loss when training. The second network is a refinement network. It takes coarse prediction results as input to further restore finer texture details of high-resolution images without changing semantic information of coarse prediction results, and it adopts reconstruction loss and adversarial loss when training. The goal of the texture details restoration network is to ensure that the image texture of the holes is "similar" to the surrounding area.

Content Reconstruction Network (Completion Network): Different from [14], we use the VGG network as the encoder, which can better obtain the detailed features of the images. We use continuous 3 × 3 convolution kernels (using small convolution kernels is superior to the use of bigger convolution kernels) for a given receptive field. Also, we alternately use four layers of dilated convolution (rate 16, 8, 4, 2, corresponding feature map size 128, 64, 32, 8) in the intermediate convolution layer. The purpose of dilated convolution is to capture a larger field of view with fewer parameters so that the part under the remaining holes is consistent with its surrounding environment. Then we take the output information of the encoder through the decoder. In our implementation, the content reconstruction network adapts to the context encoder network.

As shown in Figure 2, the five-layer encoder gradually samples down, and each layer of the encoder is composed of Convolution, Relu, BN, and Dilated Convolution. The rate of dilated convolution decreases with the decrease of the size of the feature map. The decoder gradually samples features up to the input image scale. We use transposed convolution instead of convolution in the decoder.

Texture Details Restoration Network (Refinement Network): Inspired by SRGAN, we add multiple residual blocks and skip connections between input and output in the middle layers of the texture detail restoration network. Each residual block uses two 3 x 3 convolution layers, 64 characteristic figures, and the batch normalized layer (BN) after every convolution layer, and uses ReLU as the activation function. The texture detail restoration network uses two sub-pixel convolution layers instead of deconvolution to enlarge the feature size. Reducing invalid information through the sub-pixel convolution layer can make the high-resolution image smoother, reduce the blurred texture and the unpleasant boundaries that inconsistent with the surrounding area, and obtain a better visual result.

3.2. Loss Function

Inspired by Iizuka [11], this paper uses L1 loss while attaching the loss of WGAN-GP [31] to the global and local output of the second-stage network, so as to enhance the consistency between the global and local. The original WGAN used the Wasserstein distance $\mathrm{W}(P_\mathrm{r}, P_g)$ to compare the distribution differences between the generated data and the actual data. Wasserstein is defined as follows:

$$W(P_r, P_g) = \inf_{Y \in \Pi(P_r, P_g)} E_{(x,y)\sim Y}[\|x - y\|] \tag{1}$$

Of which, $\Pi(P_r, P_g)$ is the set of all possible joint distributions combined by P_r and P_g. For each possible joint distribution Y, we can sample $(x, y) \sim Y$ from it to get a real sample x and a generated sample y. Then we calculate the distance $\|x - y\|$ between the samples. Finally, we calculate the expected value E of the distance between sample pairs in the joint distribution Y. On this basis, the objective function based on WGAN is established:

$$\min_{G} \max_{D \in \mathcal{D}} E_{x\sim P_r}[D(x)] - E_{\widetilde{x}\sim P_g}[D(\widetilde{x})] \tag{2}$$

where, $\mathcal{D}$ is a set of 1-Lipschitz functions, P_g is the model distribution implicitly defined by $\widetilde{x} = G(z)$, and z is the input of the generator. In order to realize the Lipschitz continuity condition, the original WGAN clip the updated parameter of the discriminator to a smaller interval $[-c, c]$, so the parameter gathers at two points of $-c$ and c, which limits the fitting ability to some extent.

WGAN-GP has improved on the basis of WGAN, replacing weight clipping with gradient penalty: $\lambda E_{\widetilde{x}-P_{\widetilde{x}}}(\|\nabla_{\widetilde{x}} D(\widetilde{x})\|_2 - 1)^2$. WGAN-GP uses the penalty limits the value of gradient.

For image inpainting, we only try to predict region A need repaired, therefore gradient penalty should only be applied to pixels in region A. We can achieve it by gradient multiplication and mask m. Format is as follows:

$$\lambda E_{\widetilde{x}\sim P_{\widetilde{x}}}(\|\nabla_{\widetilde{x}} D(\widetilde{x}) \odot (1 - m)\|_2 - 1)^2 \tag{3}$$

where, the mask value is 0 for missing pixels and 1 for pixels at other locations. λ is set to 10 at all LABS.

The improvement also addresses the problem of the disappearance of training gradient and gradient explosion. Moreover, it has a faster convergence speed in deep learning than the original WGAN. It can also generate higher quality images and reduce the time of parameter adjustment in the training process.

3.3. Generation of Mask Data Sets

In WeChat or other social networks, some photos of scenic spots with a comfortable and clean background are shared by tourists. However, there will be more redundant pedestrians in background destroying the beauty and artistic conception of the images, especially in popular tourist resorts. So we propose an image inpainting task of retaining the target character in the image while filtering out the redundant pedestrians in background. In order to complete the image inpainting task described above, we construct the relevant mask dataset of image inpainting, which must contain various pose to produce a better inpainting result on the character filter task. This paper uses the COCO dataset to construct the mask dataset, which is a large and rich dataset of object detection, segmentation, and caption. The dataset includes 91 types of targets, which contains more than 30,000 images of human, mainly from the complex daily scenes to meet the needs of various pose.

We select images with multiple pedestrians in COCO dataset to construct the irregular mask dataset. As shown in Figure 3, first we take the picture C through the Mask-R-CNN network to find all the people in the image. Mask-R-CNN is a general instance segmentation framework, which can not only find all the target objects in the image but also accurately segment them. We could segment them after finding all the people, and the instance segmentation result is expressed as M. So mask $\mathrm{C_{mask}}$ can be expressed as:

$$\mathrm{C_{mask}}(x,y) = \begin{cases} 255 & \mathrm{C}(x,y) \in \mathrm{M} \\ 0 & \textit{otherwise} \end{cases} \tag{4}$$

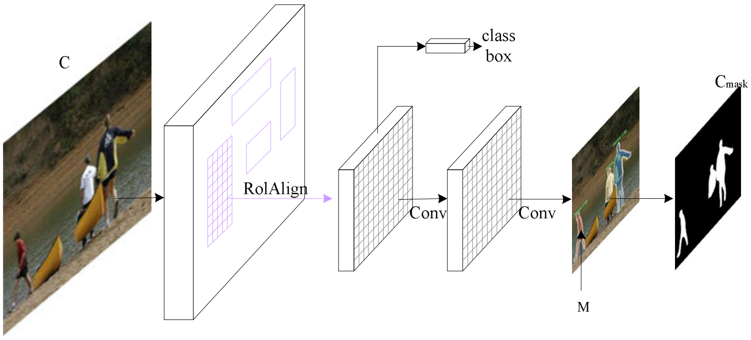

Figure 3. Construction of irregular mask dataset.

We can generate 34,980 irregular masks $\mathrm{C_{mask}}$ through the method with COCO dataset. Among the 34,980 irregular masks, we randomly selected 23,500 masks used for training and 11,480 masks for testing. To confirm that our mask dataset is true and reliable, we designed an experimental framework for the image inpainting task. As the entire network is shown in the following figure, the input used for training is a real image named IMAG with one character at most, and the size of the input image is 256 × 256. Image IMAG first detects the target pedestrian through the Mask-R-CNN network. Then we randomly select a mask $\mathrm{C_{mask}}$ applied to image IMAG from the 23,500 masks for training. Finally, image IMAG and Mask $\mathrm{C_{mask}}$ can be used as input repair to train the image inpainting network.

However, in the actual application scene, the target character of the street photography is often prominent and unobtrusive. So we should try our best not to destroy the structure of the target character, and simply perform in background. In order to make the training more consistent with the actual application scene, we need to "protect" the target pedestrian when applying the binary mask to the target image IMAG to simulate real pedestrians. The protection mechanism can be defined as:

$$\mathrm{A}(x,y) = \begin{cases} 255 & \mathrm{A}(x,y) \notin \ \mathrm{P} \quad \text{and} \quad \mathrm{C_{mask}}(x,y) = 255 \\ \mathrm{A}(x,y) & \textit{otherwise} \end{cases} \tag{5}$$

where, $\mathrm{A}(x,y)$ represents an image with the area need repaired, and P is the area of detected target pedestrian.

In this way, the experimental method to test the performance of our irregular mask dataset is complete. The entire network to train is shown in Figure 4. The network integrates the existing research of instance segmentation and image inpainting. It can solve the common problem of more unnecessary pedestrians in image background destroying the beauty and artistic conception of the images in daily life.

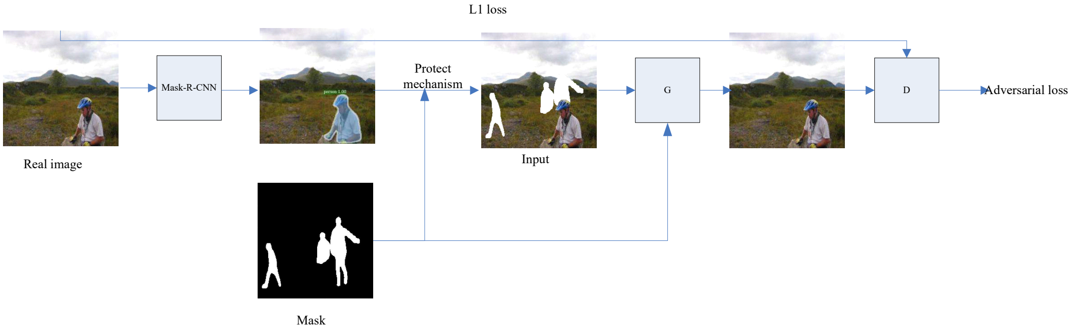

Figure 4. Network framework for pedestrian removal.

4. Result

4.1. Improved Inpainting Network

We evaluate our proposed image inpainting model on VOC2017 and COCO dataset without using tags or other information related to these images. The COCO dataset contains 118,288 images for training and 100 test images. VOC dataset contains 17,126 images for training and 100 test images. These test images are randomly selected from the validation dataset.

We compared the experimental results with PatchMatch [25] and contextual attention (Yu J [14]). PatchMatch [25] is one of the most advanced methods in patch synthesis, and contextual attention (Yu J [14]) is currently a relatively advanced image inpainting network based on deep learning. To be fair, we use all the methods to train on our dataset. Yu J [14] trained the model to handle the fixed hole. Therefore, we used fixed holes on the testing dataset to make it easy to compare the results with PatchMatch [25] and contextual attention (Yu J [14]). The fixed hole is located in the center of the input image, with the size 128 × 128. All results are generated from directly exported training models, and no post-processing is performed.

First of all, the display comparison between our results and PatchMatch [25], contextual attention (Yu J [14]) in high-resolution images is shown in Figure 5. It can be seen, the inpainting results of our model are more realistic, smoother and more similar to the texture of the surrounding area than the other two methods. Next, the quantitative comparison in Table 1 also shows the results of the comparison between our method and PatchMatch [25], contextual attention (Yu J [14]). We use three evaluation indexes: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and average error (L1 loss). The unit of PSNR is dB, and the larger the value is, the smaller the image distortion is. The value range of SSIM ranges from 0 to 1, and the larger the value is, the smaller the image distortion is. L1 loss is the sum of the absolute difference between input and output, and the smaller the value is, the smaller the image distortion is. As you can see from the table, the methods based on deep learning have a better performance than the traditional methods based on patch in three indexes including PSNR, SSIM and L1 loss. Our model has improved in terms of data compared with contextual attention (Yu J [14]). And it is obvious in Figure 5 that our model can effectively reduce the blurred texture and the unpleasant boundaries that inconsistent with the surrounding area. Our results are superior to contextual attention, which prove the effectiveness of our model in recovering texture details in image inpainting.

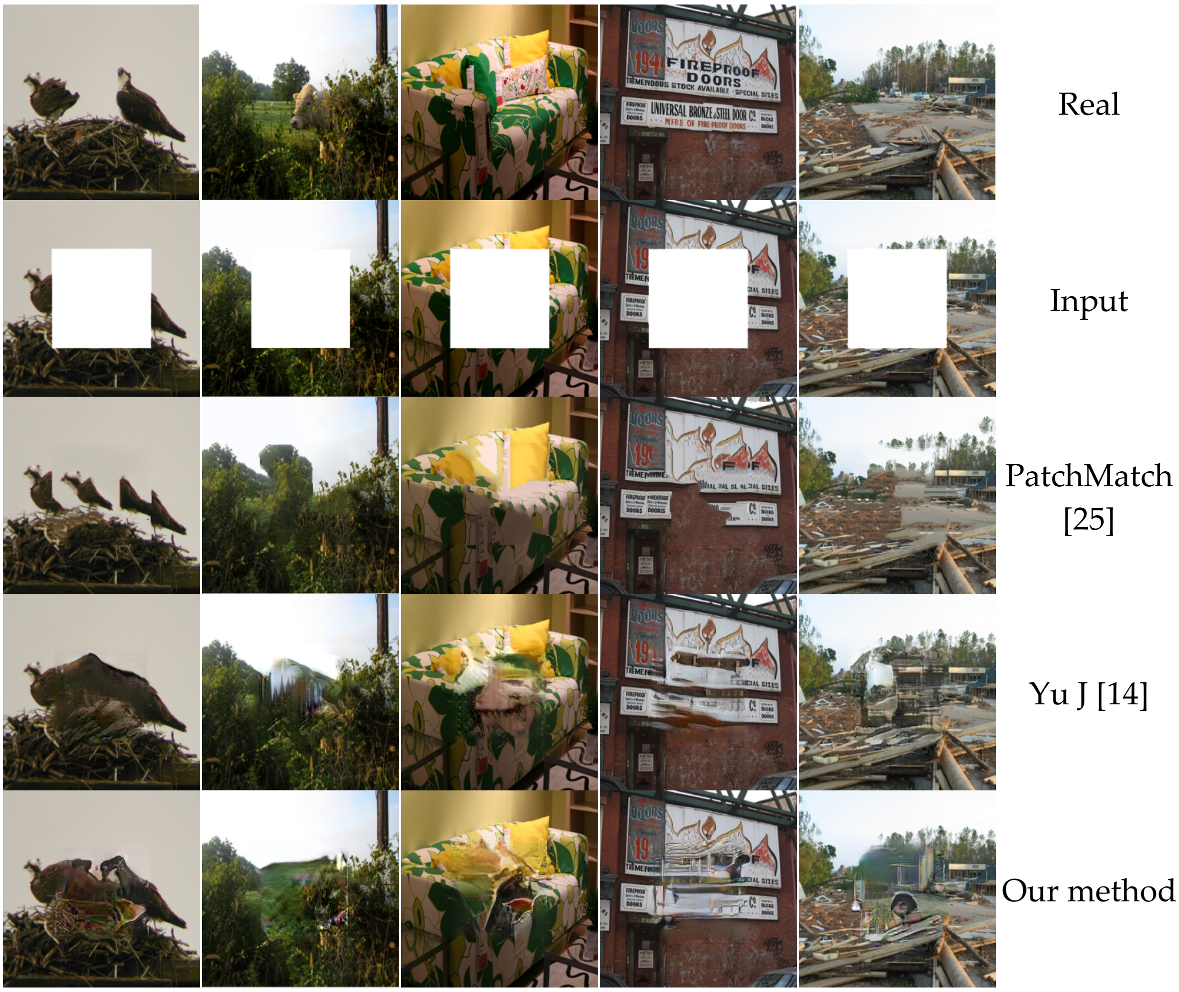

Figure 5. Comparison diagram of the results of our algorithm, PatchMatch [25] algorithm and Yu J [14] algorithm.

Table 1. The results of image inpainting using three methods.

Method	PSNR	SSIM	L1 Loss (%)
PatchMatch [25]	17.36	0.5908	8.78
Yu J [14]	19.14	0.7090	5.06
Our method	19.78	0.7205	5.52

Our full model is implemented on TensorFlow v1.3, CUDNN v7.0, CUDA v9.0 and run on hardware of CPU Intel(R) Xeon(R) gold 5117 (2.00 GHz) and GPU GTX 1080 Ti. We introduced 16 residuals into the texture detail repair network. However, in the training, these 16 residual blocks consume a lot of memory and slow down the training speed. After trying to lessen 16 residual blocks to 5 residual blocks, we found that our full model run 0.2 s per frame on the GPU, with significant improvement in speed and no significant change in performance.

In addition, the proposed inpainting framework can also be applied to conditional image generation, image editing, and computational photography tasks, including image-based rendering, image super-resolution, boot editing, and so on

4.2. Mask Experiment

We also evaluated our proposed mask dataset on two dataset including VOC2017 and COCO dataset. In the previous section, our model is trained to handle fixed holes to make it easy to compare.

While the model in [13] was trained to handle random holes and used the irregular mask dataset proposed by Liu [12], which can meet our experimental requirements. So we used the inpainting model proposed in [13] to prove the reliability of proposed mask dataset in removing redundant pedestrians in the image background.

We compared our mask dataset with the mask dataset proposed by Liu [12], the mixed mask dataset by training with the inpainting model [13]. Our mask dataset contains 23,500 masks for training and 11,480 masks for testing which are randomly generated from COCO dataset. The mask dataset proposed by Liu [12] contains 55,115 masks for training. The mixed mask dataset has a total of 78,615 masks (including 23,500 training masks that we randomly generated and 55,115 masks for training in Liu [12]).

Our comparison results are shown in the Table 2, from which we can see that we have improved the data in processing the image inpainting in a special application scene. We preserved the target character while filtering out redundant pedestrians by using our randomly generated masks. Although using our mask dataset and mixed mask dataset have similar results, it performes poorly without using our mask dataset. The comparison proves the reliability of our mask dataset in removing redundant pedestrians from image background.

Table 2. The results of image inpainting using different mask datasets.

COCO Dataset	PSNR	SSIM	L1 Loss (%)
Liu mask [12]	26.59	0.9146	2.31
Liu mask [12] + our mask	27.55	0.9233	2.01
our mask	27.54	0.9230	2.02

Figure 6 shows the intermediate results of our test, from top to bottom, which are the original picture, the instance segmentation result about people, the result of removing redundant pedestrians, and the result of image inpainting. The visualization results show that our experiment can easily screen out one or more redundant pedestrians in background and remove them.

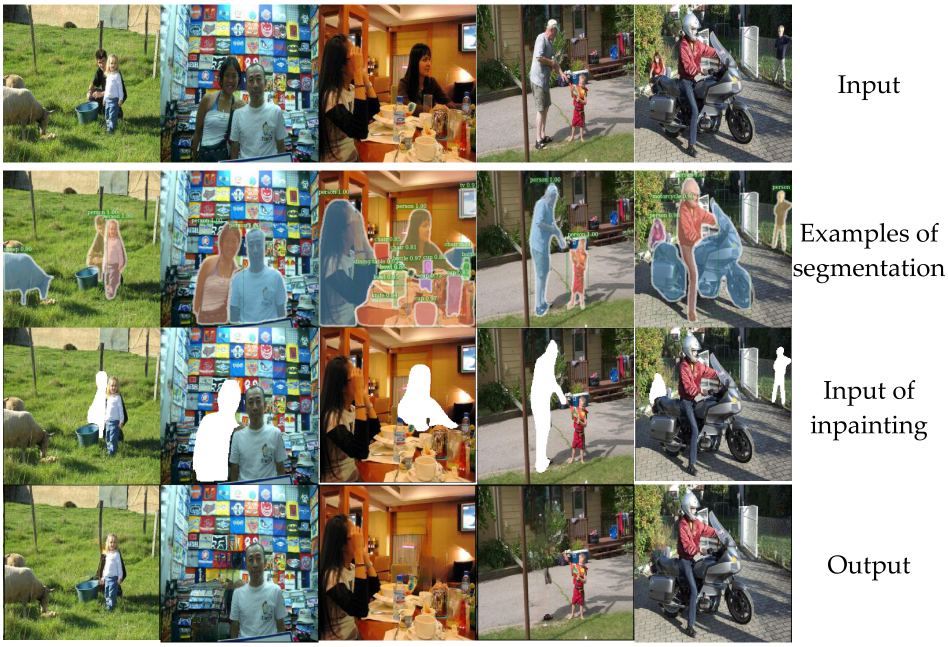

Figure 6. The result with our proposed mask dataset.

But our experiment still has some limitations. (1) When the target character is "glued" to the other characters, as shown in the left-most figure, it produces poor results even if the redundant pedestrians can be detected. (2) Although the instance segmentation can segment the redundant pedestrians in background, it is not accurate enough to leave the hands or shoes of the redundant pedestrians, affecting the visual result. This requires further study of the experiment.

Finally, we randomly downloaded some travel photos from the internet for testing. The photos contain mountain scenery, buildings, streets, coast, and other areas. As we can see from Figure 7, the method proposed in the paper also has a high visual result in real life. In the future, we can apply it to mobile phone application to detect pedestrians in background of personal travel photos, wedding photos, and other photos. At the same time users filter unnecessary pedestrians with one key and share the beautiful travel photos in real time.

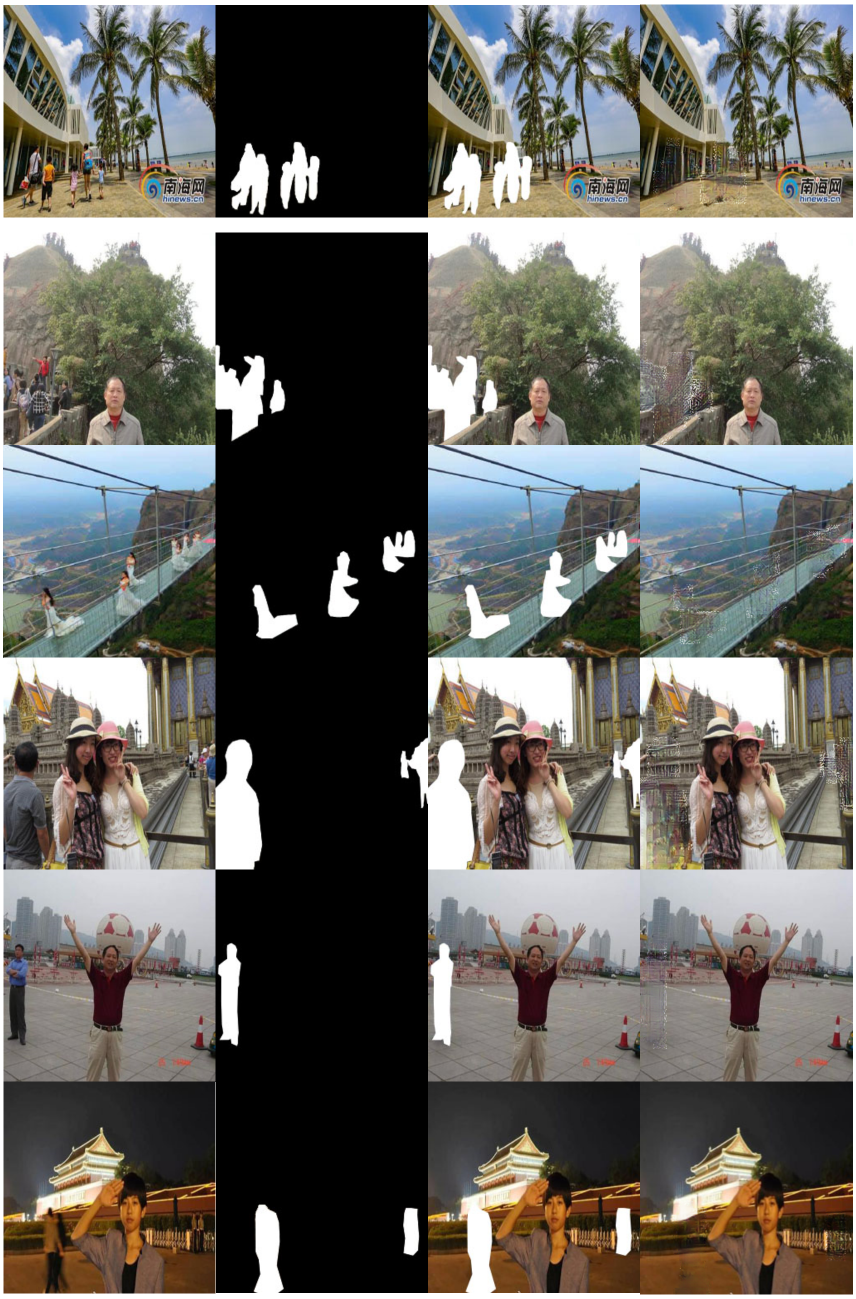

Figure 7. *Cont.*

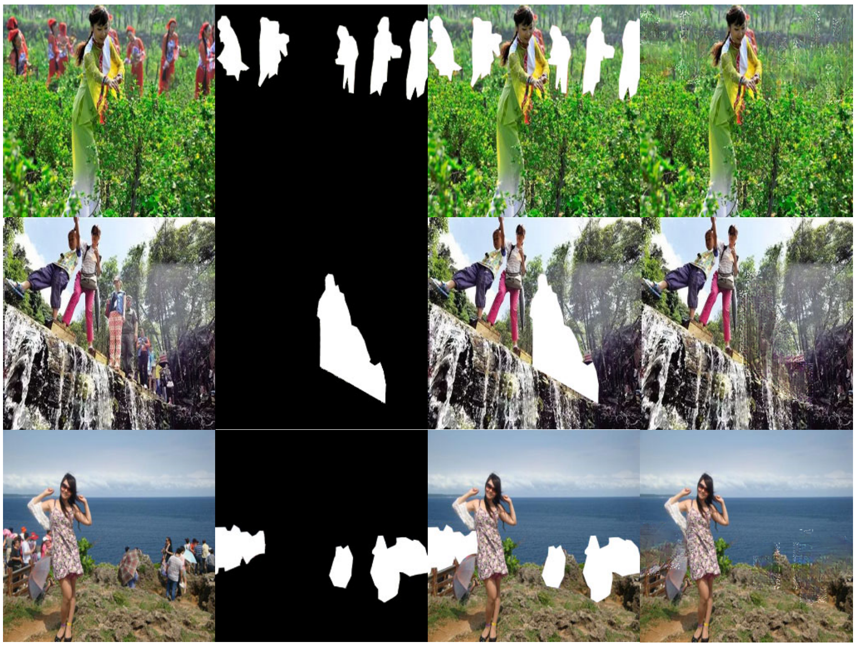

Figure 7. The actual application effect in network image.

5. Conclusions

Image inpainting based on GAN has made great progress in recent years, but they can only process low-resolution images because of memory limitations and difficulty in training. For high-resolution images, the inpainted regions become blurred and the unpleasant boundaries become visible. Many researchers are committed to improve the existing image inpainting network framework. We propose a novel high-resolution image inpainting method based on deep generative model. It is a two-stage network including content reconstruction and texture detail restoration. After obtaining the visually believable fuzzy texture, we further restore the finer texture details improve the image inpainting quality. Meanwhile, we integrate the existing research of instance segmentation and image inpainting to delete the unnecessary pedestrians in background and ensure the reality of background restoration. To improve the accuracy of image inpainting in the special application scene, we proposed a new mask dataset, which collected the characters in COCO dataset as a mask, and could produce better inpainting results for the special application scene.

In our future work, we will experiment with convolutional deep belief network (CDBN) [37] and PCANET [38] based on the paper. Like the CNN, CDBN can extract the high-frequency features of images. According to the latest research, CDBN performs better than CNN in the classification task of large-size images, so it may be better to use CDBN instead of CNN for high-resolution images. In addition, PCANET can conduct feature fusion of feature maps of different sizes in encoder, which strengthens the correlation between input and output. However, the better result of PCANET may come at the cost of speed.

Author Contributions: Conceptualization, T.S. and W.C.; methodology, T.S., W.C. and W.F.; validation, F.B. and B.W.; formal analysis, T.S. and W.C.; investigation, T.S.; resources, W.C.; writing—original draft preparation, T.S.; writing—review and editing, T.S., W.C. and Y.Y.; visualization, T.S.; supervision, W.C.; project administration, W.C.; funding acquisition, W.C.

Funding: This word was funded by National Natural Science Foundation of China, grant number 51874300, National Natural Science Foundation of China and Shanxi Provincial People's Government Jointly Funded Project of China for Coal Base and Low Carbon, grant number U1510115, and the Open Research Fund of Key Laboratory of Wireless Sensor Network & Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, grant numbers 20190902 and 20190913. The APC was funded by 51874300, U1510115, 20190902, and 20190913.

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

References

1. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ECO: Efficient Convolution Operators for Tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6931–6939.
2. Mao, J.; Xiao, T.; Jiang, Y.; Cao, Z. What Can Help Pedestrian Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6034–6043.
3. Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. Joint Detection and Identification Feature Learning for Person Search. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3376–3385.
4. Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1302–1310.
5. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823.
6. Yoo, S.; Park, R.-H. Red-eye detection and correction using inpainting in digital photographs. *IEEE Trans. Consum. Electron.* **2009**, *55*, 1006–1014. [CrossRef]
7. Dolhansky, B.; Ferrer, C.C. Eye In-painting with Exemplar Generative Adversarial Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7902–7911.
8. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), Montreal, Canada, 7–12 December 2015; MIT: Cambridge, MA, USA, 2015; pp. 91–99.
10. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2536–2544.
11. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* **2017**, *36*, 107. [CrossRef]
12. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.-C.; Tao, A.; Catanzaro, B. Image Inpainting for Irregular Holes Using Partial Convolutions. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 89–105.
13. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; Ebrahimi, M. EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. *arXiv* **2019**, arXiv:1901.00212.
14. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative Image Inpainting with Contextual Attention. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5505–5514.
15. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; Shi, W. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
17. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
18. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3296–3297.

19. Shrivastava, A.; Gupta, A.; Girshick, R. Training Region-Based Object Detectors with Online Hard Example Mining. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 761–769.
20. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully Convolutional Instance-Aware Semantic Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4438–4446.
21. Dai, J.; He, K.; Li, Y.; Ren, S.; Sun, J. Instance-Sensitive Fully Convolutional Networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 534–549.
22. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In Proceedings of the Advances in Neural Information Processing Systems; Barcelona, Spain, 5–10 December 2016; pp. 379–387.
23. Fedorov, V.V.; Facciolo, G.; Arias, P. Variational Framework for Non-Local Inpainting. *Image Process. Line* **2015**, *5*, 362–386. [CrossRef]
24. Newson, A.; Almansa, A.; Gousseau, Y.; Pérez, P. Non-Local Patch-Based Image Inpainting. *Image Process. Line* **2017**, *7*, 373–385. [CrossRef]
25. Barnes, C.; Shechtman, E.; Finkelstein, A.; Dan, B.G. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. *ACM Trans. Graph.* **2009**, *28*, 24. [CrossRef]
26. Goodfellow, I.J.; Pougetabadie, J.; Mirza, M.; Xu, B.; Wardefarley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
27. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434.
28. Zhao, J.; Mathieu, M.; LeCun, Y. Energy-based Generative Adversarial Network. *ArXiv* **2016**, arXiv:1609.03126.
29. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved Techniques for Training GANs. *ArXiv* **2016**, arXiv:1606.03498.
30. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1704.00028.
31. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved Training of Wasserstein GANs. *ArXiv* **2017**, arXiv:1704.00028.
32. Berthelot, D.; Schumm, T.; Metz, L. BEGAN: Boundary Equilibrium Generative Adversarial Networks. *ArXiv* **2017**, arXiv:1703.10717.
33. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Smolley, S.P. Least Squares Generative Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2813–2821.
34. Isola, P.; Zhu, J.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv* **2016**, arXiv:1611.07004.
35. Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. *ArXiv* **2017**, arXiv:1711.11585.
36. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
37. Lee, H.; Grosse, R.; Ranganath, R.; Ng, A.Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th Annual International Conference, Montreal, QC, Canada, 14–18 June 2009; pp. 609–616.
38. Chan, T.-H.; Jia, K.; Gao, S.; Lu, J.; Zeng, Z.; Ma, Y. PCANet: A Simple Deep Learning Baseline for Image Classification? *IEEE Trans. Image Process.* **2015**, *24*, 5017–5032. [CrossRef] [PubMed]