





Article

SR-Inpaint: A General Deep Learning Framework for High Resolution Image Inpainting

Haoran Xu ^{1,2} , Xinya Li ³, Kaiyi Zhang ⁴, Yanbai He ⁵, Haoran Fan ^{1,2}, Sijiang Liu ^{3,*} , Chuanyan Hao ^{3,*}  and Bo Jiang ^{3,*} 

- ¹ School of Electronic and Optical Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210049, China; haoranxu2000@gmail.com (H.X.); 1220024206@njupt.edu.cn (H.F.)
² School of Microelectronics, Nanjing University of Posts and Telecommunications, Nanjing 210049, China
³ School of Educational Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210049, China; lixinya.xy@gmail.com
⁴ School of Overseas Education, Nanjing University of Posts and Telecommunications, Nanjing 210049, China; zhangky1999@gmail.com
⁵ State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China; heyantai1999@gmail.com
* Correspondence: liusj@njupt.edu.cn (S.L.); hcy@njupt.edu.cn (C.H.); jiangbo@njupt.edu.cn (B.J.)

Abstract: Recently, deep learning has enabled a huge leap forward in image inpainting. However, due to the memory and computational limitation, most existing methods are able to handle only low-resolution inputs, typically less than 1 K. With the improvement of Internet transmission capacity and mobile device cameras, the resolution of image and video sources available to users via the cloud or locally is increasing. For high-resolution images, the common inpainting methods simply upsample the inpainted result of the shrunk image to yield a blurry result. In recent years, there is an urgent need to reconstruct the missing high-frequency information in high-resolution images and generate sharp texture details. Hence, we propose a general deep learning framework for high-resolution image inpainting, which first hallucinates a semantically continuous blurred result using low-resolution inpainting and suppresses computational overhead. Then the sharp high-frequency details with original resolution are reconstructed using super-resolution refinement. Experimentally, our method achieves inspiring inpainting quality on 2K and 4K resolution images, ahead of the state-of-the-art high-resolution inpainting technique. This framework is expected to be popularized for high-resolution image editing tasks on personal computers and mobile devices in the future.

Keywords: deep learning; image inpainting; super-resolution; high-resolution; high-frequency information reconstruction



Citation: Xu, H.; Li, X.; Zhang, K.; He, Y.; Fan, H.; Liu, S.; Hao, C.; Jiang, B. SR-Inpaint: A General Deep Learning Framework for High Resolution Image Inpainting. *Algorithms* **2021**, *14*, 236. <https://doi.org/10.3390/a14080236>

Academic Editors: Frank Werner and Andres Iglesias Prieto

Received: 15 July 2021

Accepted: 9 August 2021

Published: 10 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Image inpainting or image completion, which involves the automatic recovery of missing pixels of an image according to the known information within the image, is an important research area in computer vision. With the rapid development of digital image editing technology, image inpainting has been widely applied to damaged photo restoration, occlusion removal, intelligent aesthetics and other graphics fields. Inpainting has been an active research area in the past few decades and many studies have been devoted to achieving visual realism and vividness [1–16]. However, due to the complexity of damaged images and the inherent ambiguity of methods, the semantics-continuous and texture-clear inpainting remains a major challenge, especially for High-Resolution (HR) images [17]. Hence, our work is motivated by the issue that most existing image inpainting techniques cannot realize high quality completion of damaged HR images.

Early inpainting methods can be broadly divided into the diffusion methods based on pixel propagation [1–3] and the patching methods based on texture borrowing [4–8], which do a poor job of reconstructing complex details [9]. In recent years, deep learning

approaches have achieved promising success in inpainting. A stream of these methods hallucinates missing pixels using learned data distribution [9–11,18]. Another stream fills the hole using a data-driven manner with the external image sources [12–16]. Though these methods can yield meaningful structure in missing regions, the generated regions are often blurred and accompanied by artifacts. In addition, with the improvement of Internet transmission capacity and mobile device cameras, the resolution of image and video sources available to users via the cloud or locally is increasing [17]. However, for HR images, general image inpainting methods often yield a limited result. In addition the input is even rejected due to the memory limitation [17]. Now, there is an urgent need for methods that can reconstruct the missing high-frequency information in HR images and generate sharp texture details.

Therefore, several inpainting strategies have been proposed for the high-resolution reconstruction of high-frequency information. For example, Ikehata et al. [19] proposed a combined framework of patch-based inpainting and super-resolution to generate a dense high-resolution depth map from a corrupted low-resolution depth map and its corresponding high-resolution texture image. Kim et al. [20] proposed a method called “Zoom-to-Inpaint”, which enhances the high-frequency details of the inpainted area through a zoom-in, refine and zoom-out strategy, combines with high-resolution supervision and progressive learning. These frameworks improve the high-frequency reconstruction of the missing regions in general images. However, for HR images, these methods are not yet perfectly applicable and still face problems such as computational limitation. On the other hand, Yi et al. [17] proposed an HR image inpainting algorithm, which upsamples the Low-Resolution (LR) inpainted result and adds a high-frequency residual image into the blurred image to generate a sharp result through a contextual residual aggregation mechanism. The method effectively suppresses the cost of memory and computing power as well as achieves compelling quality in natural photographs with a monotonic background. However, the realism and semantic continuity of the inpainted results for the images with complex compositions or textures need to be further improved. For now, the visually realistic recovery of high-frequency information for the HR images with complex backgrounds is still a tricky task.

To this end, we propose a novel deep learning framework for HR image inpainting. The framework mainly consists of two deep learning modules: (1) a low-resolution inpainting module for the reconstruction of high-frequency information in the missing region, and (2) a super-resolution module for the enhancement of the resolution of the inpainted region. We input the HR images to the inpainting network by downsampling, hallucinating an LR map with high semantic continuity and coherence, then sending it to the super-resolution network for refinement, and finally obtaining a visually realistic inpainted result at high-resolution. Our method is capable of entering 2K and 4K resolution images and generating results at the same resolution, while ensuring the structural and semantic coherence, which is ahead of the state-of-the-art technology. In summary, our contributions are four-fold:

- A novel deep learning framework for high-resolution inpainting, which allows the input of 2K and 4K resolution images to yield equally sharp results.
- A “degradation and refinement” strategy is proposed to suppress suppressing memory and computational overhead while guaranteeing a high inpainting quality at high-resolution.
- The structural coherence and visual fidelity of the inpainted results are enhanced to be ahead of the state-of-art technology.
- A general high-resolution inpainting pipeline consisting of an independent inpainter and refiner in series that can be trained and modified separately.

2. Related Work

2.1. Image Inpainting

Image inpainting is the fundamental and long-standing problem in computer vision. Traditional inpainting methods can be broadly classified into two categories: (1) diffusion methods [1–3], which propagate neighboring pixels; (2) patch methods [4–8,21,22], which explicitly borrow textures from surroundings. These methods are limited to locally available information and cannot recover meaningful structures in the missing regions, let alone complex details. The development of the image processing field including image synthesis, image super-resolution and image inpainting have been greatly facilitated with the proposal of deep learning and Convolutional Neural Networks (CNN), especially Generative Adversarial Networks (GAN) [15,18,23–28]. For example, Pathak et al. [15] proposed a context encoder that makes a reasonable assumption about the hole in the picture by training a CNN. Furthermore, Yang et al. [18] proposed an optimization method based on GAN that produces more realistic and coherent results. In GAN, higher-order semantic acquisition is trained together with low-order pixel synthesis, which effectively compensates the shortcomings of traditional algorithms. However, due to the complexity and diversity of natural images, it is not enough to only generate new pixels, but also to ensure the visual fidelity and vividness of the inpainted results [15]. Classical single-stage GAN will lead to discontinuities, blurring, artifacts and excessive smoothing defects. Therefore, researchers have improved and innovated the framework based on GAN, such as Iizuka et al. [12] who used global and local two-stage discriminators to judge the semantics of the generated images and improve the consistency of the generated pixels with the original pixels. EdgeConnect [9] is an effective GAN-based inpainting framework inspired by the idea of “lines first, color next” in art creation, which generates complex details through a two-stage GAN, adhering well to the principles of structure-first. The result is impressive. However, general inpainting methods still struggle to remove its inherent blurriness, which is more obvious after zooming in. Hence, the high-resolution recovery of the missing high-frequency information in HR images is a non-negligible problem for HR image inpainting.

2.2. High-Frequency Image Content Reconstruction

For complex HR images, although some current methods can inpaint meaningful contents, they will lead to severe high-frequency information loss due to the input resolution limitation and the inherent ambiguity. For this reason, Yi et al. [17] proposed a contextual residual aggregation mechanism to produce high-frequency residuals for the missing content by weighted aggregating residuals from contextual patches, then add them to the blurry image to yield high-resolution result. However, this mechanism is difficult to ensure the structural and semantic consistency of the inpainted results. If we want to take full advantage of the existing semantic continuous inpainting, we can only downsample the input and thus obtain a low-resolution result. Hence, we propose to solve this contradiction using Super-Resolution (SR) techniques. SR reconstruction allows an HR image to be extrapolated from an LR image and to recover as much high-frequency information as possible, such as texture details. Early SR algorithms were based on image processing in the frequency or space domain, such as the Multiframe Image Restoration proposed by Tsai et al. [29] and Projection onto Convex Sets (POCS) proposed by Stark et al. [30]. In 2014, Dong et al. pioneered the application of deep learning to the super-resolution reconstruction. Since then, a large number of super-resolution models based on deep learning have been proposed, from CNN to GAN. The SRCNN proposed by Dong et al. [31] uses a three-layer CNN, each layer corresponding to the feature extraction, nonlinear mapping, and high-quality reconstruction of the image, respectively. However, the network is too shallow leading to too small perceptual field of the generated images. Compared with the Deep Neural Networks (DNN), the SRCNN has weaker fitting ability and does poor job in complex details. The most direct solution is to increase the network depth. A deeper network inevitably leads to a larger perceptual field [32], which allows the

network to utilize more contextual information and have a more reflective global mapping. For example, Reuben et al. [33] proposed a spatial light field super-resolution method, using deep CNN to restore the entire light field with consistency across all angular views. Kim et al. [34] proposed a very deep network (VDSR) that improves the SR performance in terms of both PSNR and SSIM. However, both SRCNN and VDSR input the LR image to the network by bicubic interpolation, resulting in low efficiency. For this reason, FSRCNN [35] and ESPCN [36] operate the LR input directly and upsample at the end of the network. Although great success has been achieved in high-frequency recovery with bicubic degradation [37–39], for arbitrary blur caused by LR inpainting, these methods perform poorly due to the mismatch of degradation models. Zhang et al. [40] proposed a Plug-and-Play deep framework (DPSR) with a new degradation model that can handle LR images for arbitrary blur kernels, achieving promising results in synthetic and real LR images. Hence, we migrated this framework to HR inpainting task for high-frequency information reconstruction from LR to HR inpainted images.

3. Method

3.1. Framework and Flow

We divide the HR inpainting task into two distinct problems: HR image inpainting and high-frequency information reconstruction. Hence, we propose a novel HR inpainting framework that first downsamples the HR input into a nLR network for inpainting, and the preliminary inpainted result is fed into a SR network for detail refinement. Finally, the inpainted HR result with high-frequency details can be obtained.

The entire framework is depicted in Figure 1. It mainly consists of two networks in series: (1) an LR inpainting network and (2) an SR network. As shown in Figure 1, a damaged HR image (2K or 4K) with a mask are used as input. Firstly, the input image and mask are bicubically degraded in the input layer to obtain the LR map, avoiding the memory overflow caused by a too large input size. Next, the LR maps are fed into the LR inpainting network to yield a structure-coherent and detail-rich result in the LR field-of-view. The LR inpainted map is then sent to the SR network and scaled up to the original resolution by nonlinear mapping. This process realizes the high frequency information reconstruction at high resolution. Finally, the generated content is fused with the remaining part of the ground-truth image to obtain the HR inpainted image. The algorithm flowchart of our HR image inpainting method is shown in Figure 2. The following subsections depict the technical details of the deep learning networks used in our method.

3.2. LR Inpainting Network

The LR inpainting network aims to characterize variations across the entire image in the LR field-of-view and to recover missing information. As the fundamental quality of HR reconstruction, LR inpainting must ensure structural consistency, semantic continuity, and sufficient details of filling content in the LR field-of-view. Hence, we adopt a two-stage GAN framework [9] to realize high-quality image inpainting in LR field-of-view. The LR inpainting network consists of an edge generator and an image completion network, each stage of which follows an adversarial model consisting of a pair of generator–discriminator. Specifically, the generator follows the architecture proposed by Johnson et al. [41] and consists of two downsampling encoders, eight residual blocks [42], and two upsampling decoders. The discriminator uses the 70×70 PatchGAN [43,44] architecture, which discriminates whether the 70×70 overlapping blocks are true or not.

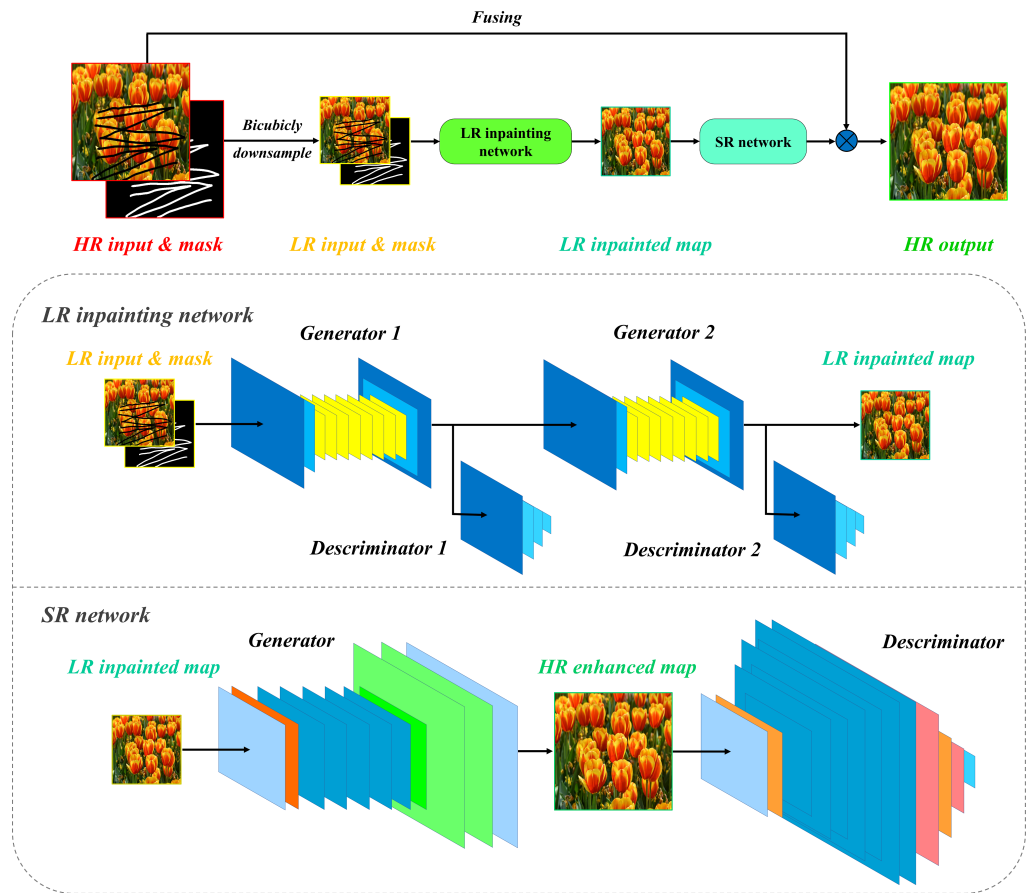


Figure 1. The overall pipeline of our method (SR-Inpaint): (top) the pipeline of SR-Inpaint, (bottom) the architectures of networks.

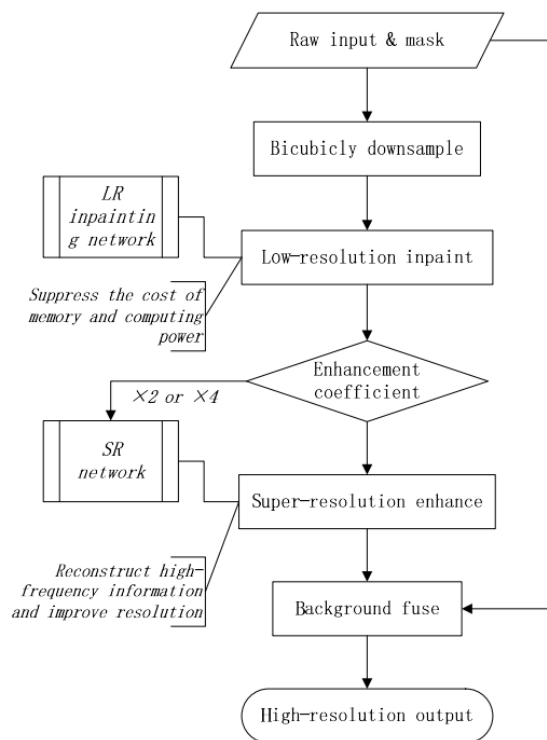


Figure 2. Algorithm flowchart of SR-inpaint.

The core task of edge generator is to predict the edge map for the masked region, as shown in Equation (1). Let \mathbf{I}_{gt} be the ground truth image. \mathbf{C}_{gt} and \mathbf{I}_{gray} denote its edge map and grayscale counterpart respectively. In this stage, we use the masked grayscale image $\tilde{\mathbf{I}}_{gray} = \mathbf{I}_{gray} \odot (1 - \mathbf{M})$ as input. Its edge map and image mask are denoted as $\tilde{\mathbf{C}}_{gt} = \mathbf{C}_{gt} \odot (1 - \mathbf{M})$ and \mathbf{M} , respectively, and used as pre-condition. Here, \odot denotes the Hadamard product.

$$\mathbf{C}_{pred} = G_1(\tilde{\mathbf{I}}_{gray}, \tilde{\mathbf{C}}_{gt}, \mathbf{M}) \quad (1)$$

The network is trained with both the adversarial loss $\mathcal{L}_{adv,1}$ and feature-matching loss \mathcal{L}_{FM} , as shown in Equation (2), where $\lambda_{adv,1}$ and λ_{FM} are regularization parameters. The feature-matching loss is very similar to perceptual loss, and compare the activation maps in the intermediate layers of the discriminator, which further stabilize the training process by forcing the similarities of both the results of the generator and the real images.

$$\min_{G_1} \max_{D_1} \mathcal{L}_{G_1} = \min_{G_1} \left(\lambda_{adv,1} \max_{D_1} (\mathcal{L}_{adv,1}) + \lambda_{FM} \mathcal{L}_{FM} \right) \quad (2)$$

For the second stage, i.e., the image completion network, the incomplete color image $\tilde{\mathbf{I}}_{gt} = \mathbf{I}_{gt} \odot (1 - \mathbf{M})$ are used as input, conditioned using a composite edge map \mathbf{C}_{comp} , which is constructed by combining the edges inferred with the first stage and ground truth edges in the remaining part of the original image, i.e., $\mathbf{C}_{comp} = \mathbf{C}_{gt} \odot (1 - \mathbf{M}) + \mathbf{C}_{pred} \odot \mathbf{M}$. The network infers a color image \mathbf{I}_{pred} , with missing regions inpainted. This procedure is denoted as Equation (3).

$$\mathbf{I}_{pred} = G_2(\tilde{\mathbf{I}}_{gt}, \mathbf{C}_{comp}) \quad (3)$$

This network is trained over a joint loss representation, as shown in Equation (4), containing ℓ_1 loss \mathcal{L}_{ℓ_1} , adversarial loss $\mathcal{L}_{adv,2}$, perceptual loss \mathcal{L}_{perc} and style loss \mathcal{L}_{style} . λ_{ℓ_1} , $\lambda_{adv,2}$, λ_p and λ_s are all regularization parameters.

$$\min_{G_2} \max_{D_2} \mathcal{L}_{G_2} = \min_{G_2} \left(\lambda_{\ell_1} \mathcal{L}_{\ell_1} + \lambda_{adv,2} \max_{D_2} (\mathcal{L}_{adv,2}) + \lambda_p \mathcal{L}_{perc} + \lambda_s \mathcal{L}_{style} \right) \quad (4)$$

3.3. SR Network

Since only LR inpainted results can be obtained from LR inpainting network, in addition to the unavoidable blur and noise in the inpainting process, it is necessary to address the problem of high-frequency information reconstruction at high resolution. Hence, the SR network aims to recover the missing high-frequency information in the HR field-of-view and enhance the resolution of LR inpainted results. Since the blurring pattern of the generated LR content is unknown, we adopt a deep plug-and-play SR framework for arbitrary blur kernels (DPSR) [40].

Most existing SR methods assume some degradation model. A widely used general degradation model for SR is depicted as Equation (5).

$$\mathbf{y} = (\mathbf{x} \otimes \mathbf{k}) \downarrow_s + \mathbf{n} \quad (5)$$

where $\mathbf{x} \otimes \mathbf{k}$ means the convolution between blur kernel k and HR image x . \downarrow_s is a subsequent downsampling operation with scale factor s , and n is additive white Gaussian noise (AWGN).

However, DPSR employs a new degradation model that supports blur kernel estimation using existing deblurring methods. As shown in Equation (6), the degradation model of DPSR made a modification to the general degradation model by first bicubic

downsample the full size image and then convolution with kernel K , rather than the convolution-downsample order, which is effective in dealing with blurry LR image.

$$\mathbf{y} = (\mathbf{x} \downarrow_s) \otimes \mathbf{k} + \mathbf{n} \quad (6)$$

Both models are then plus a noise term n . Once the model is defined, an energy function is formulated according to Maximum A Posteriori (MAP) probability, which contains two terms: data fidelity (likelihood) term and a regularization term. This optimization problem is solved with a quadratic splitting (HQS) algorithm.

Later, a super-resolver needs to be specified, which should also take the noise level as input. Here, we only need to modify the existing DNN-based super-resolver by adding a noise map as input. Methods such as SRMD can also be adopted as they already contain the noise level map.

4. Training Configuration and Strategy

We train both the Edge-Connect and DPSR model on a single NVidia Geforce GTX 1080 Ti GPU, with the PyTorch framework.

For the Edge-Connect model, the size of input image is 256×256 . The batch size is set to 8. An Adam algorithm is adopted to optimize the model. The parameter β_1 is set to 0 and β_2 is set to 0.9. First, the Generator G_1 and Generator G_2 are trained separately using Canny edges. The learning rates are set to 10^{-4} until the training reaches the plateau. Then, the learning rate is reduced to 10^{-5} . Generator G_1 and Generator G_2 continue to train until convergence. Finally, the networks are fine-tuned by removing D_1 . Generator G_1 and Generator G_2 are trained end-to-end with learning rate 10^{-6} until convergence. The learning rate for the training of Discriminators are $\frac{1}{10}$ of the generators.

For the DPSR model, we trained an enhanced version of SRResNet, namely SRResNet+ as Zhang et al. [40]. The Adam algorithm [45] is again adopted to optimize the SRResNet+ model. The learning rate is first set to 10^{-4} . Then, for every 5×10^5 iterations, the learning rate decreases by half and finally be fixed when reach 10^{-7} . The batch size for training procedure is 16. The patch size of LR input is 48×48 . Data augmentation is performed, by image rotation and flip.

5. Experimental Results and Discussion

Our proposed method (SR-Inpaint) is evaluated on the DIV2K dataset [46] and 200 2K-images as well as 200 4K-images of people, animals, nature, cities, objects, etc. Results are compared against the state-of-the-art HR image inpainting technology (HiFill by Yi et al. [17], CVPR 2020) both qualitatively and quantitatively. In the experiment, the damaged pixels were 10.612% of the total pixels for the 2K-image test set, and the damaged pixels were 12.799% of the total pixels for the 4K-image test set.

5.1. Implementation Details

Figure 3 displays the HR image inpainting pipeline in our framework. Firstly, the damaged image with mask is bicubically downsampled to 1K resolution. The LR damaged map is then fed to the LR inpainting network for inpainting to yield an LR inpainted map. Subsequently, the LR inpainted map is fed into the SR network for frame inference to reconstruct the high frequency details at high resolution. Notice that a gate exists here to match the corresponding SR networks for input images of different resolutions. For 2K input, the “ $\times 2$ ” network is matched to recover the original resolution; for 4K input, the “ $\times 4$ ” network is matched to recover the original resolution. Finally, the SR-enhanced generated content is fused with the real background by masking to obtain the completed HR inpainting result.

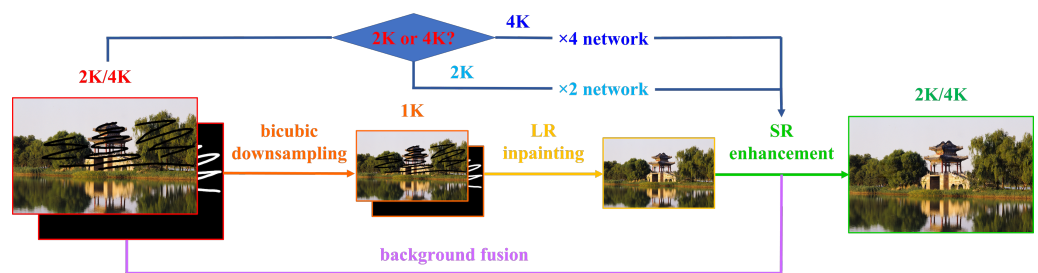


Figure 3. Implementation diagram of our high-resolution inpainting pipeline.

5.2. Qualitative Evaluation

Firstly, our approach is compared with the state-of-the-art technology in terms of visual results. Figures 4 and 5 show the examples of the images generated by the our model and the comparison model under 2K and 4K inputs. For visualization, we replace the damaged area with black color. It is clearly visible that our model is able to generate semantically continuous results that are closer to ground-truth. In addition, most of the image structures remain coordinated. In contrast, the HiFill model does a poor job in terms of structure and semantics. Particularly, for complex background, the results of the HiFill model suffer from deformation and semantic incoherence.

We think it is explained by the fact that the HiFill algorithm borrows the surrounding texture to fill the holes. If the structure and semantics of the missing region are completely different from the surrounding, then it is difficult to guarantee a meaningful structure. In contrast, our inpainting is based on edge connection, following the principle of “lines first, color next”, generating coherent structures through a two-stage GAN to achieve visual realism.

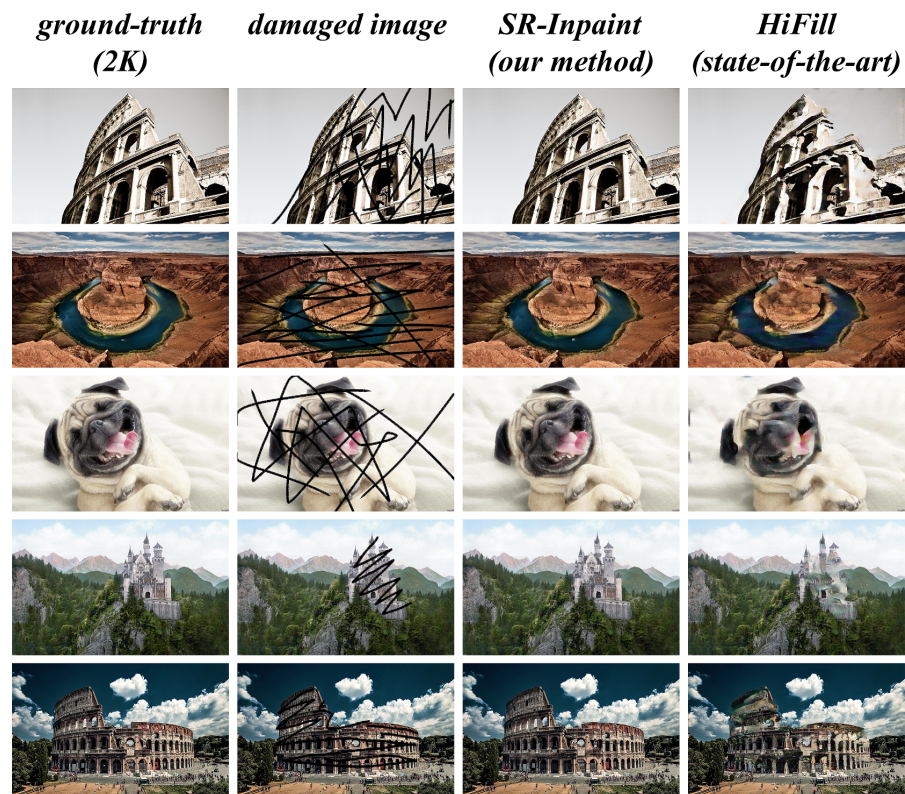


Figure 4. Example of inpainting results for 2K resolution: (left to right) ground-truth, input damaged image, inpainted image by our method, inpainted image by state-of-the-art technology.

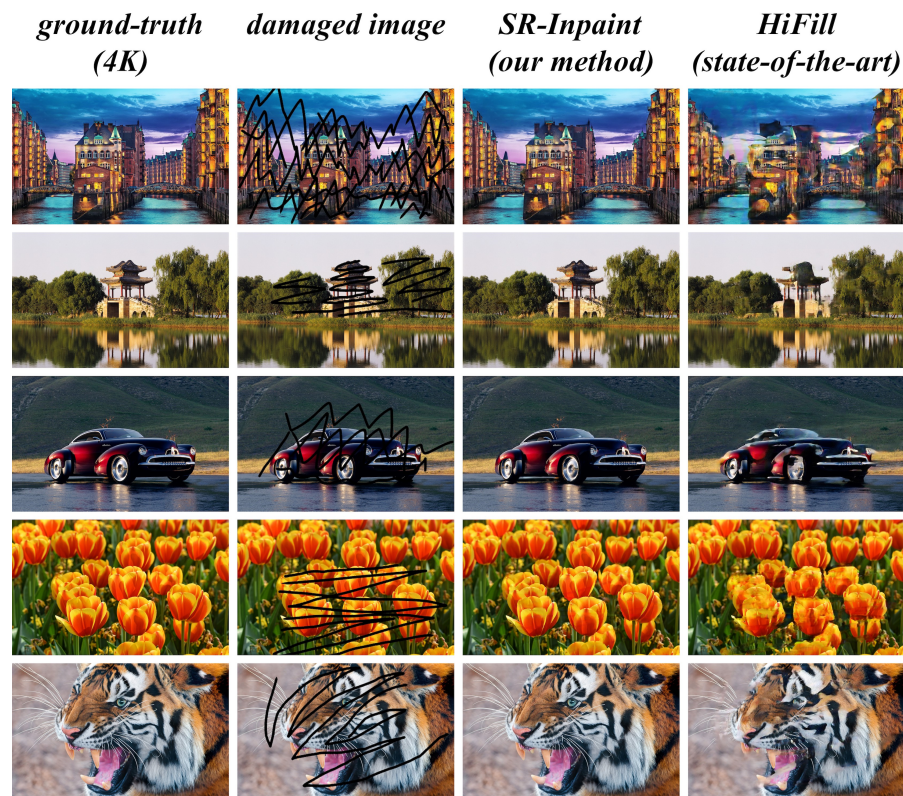


Figure 5. Example of inpainting results for 4K resolution: (left to right) ground-truth, input damaged image, inpainted image by our method, inpainted image by state-of-the-art technology.

Compared to general image inpainting, for HR inpainting, the sharpness of the inpainted area is as important as the global picture coherence. Therefore, Figure 6 shows a zoomed-in comparison of the inpainting results. It can be seen that the areas generated after bicubicly upsampling and Gaussian pyramid-up are blurred and low resolution. The areas enhanced by the wavelet method have a non-negligible color difference with the original images. Meanwhile, the areas enhanced by the Super-Resolution (SR) enhancement mechanism show minimized ambiguity. Compared with the traditional techniques, the SR enhancement mechanism based on deep learning achieves the reconstruction of high frequency details at high resolution, which significantly improves the sharpness of the generated image.

Although the HiFill model can also generate high frequency details at high resolution through the Contextual Residual Aggregation (CRA) mechanism. The CRA mechanism aggregates high-frequency details using background residuals. However, if the high-frequency information in the background is not relevant to the high-frequency information in the missing region, then the generated high-frequency details are meaningless. Compared with the CRA mechanism, SR enhancement mechanism generates high-frequency details through global picture inference based on deep learning of big-data, which is guaranteed to be meaningful in most cases.

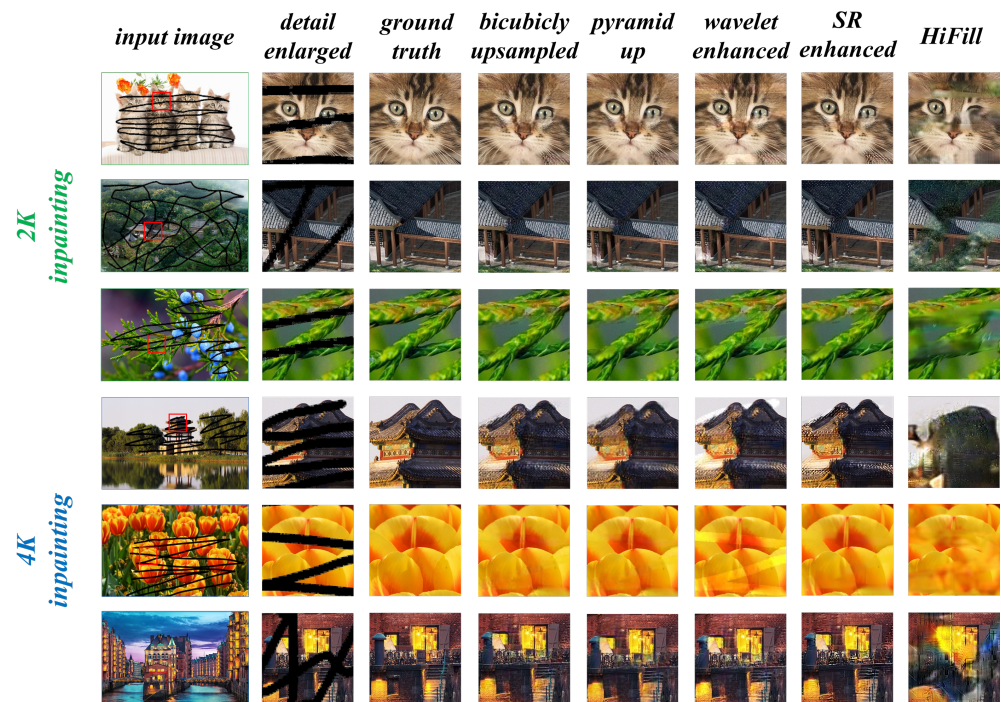


Figure 6. Zoomed-in comparison of 2K and 4K inpainted details: (left to right) input image, local details of ground-truth, local details of LR inpainting with bicubicy upsampling, local details of LR inpainting upsampled on Gaussian pyramid, local details of LR inpainting with wavelet enhancement, local details of LR inpainting with SR enhancement, local details of state-of-the-art technology.

In summary, although both generate HR results, our method does significantly better than the current state-of-the-art HR inpainting method in terms of semantic consistency as well as structural continuity. Our strategy maximizes the visual realism of the inpainting.

5.3. Quantitative Evaluation

For a more objective comparison between our method and state-of-the-art method in terms of high-resolution inpainting, we tested our method against state-of-the-art method on 2K and 4K image testsets and calculated the numerical metrics. The quality of our results are evaluated using the following metrics: Peak Signal-to-Noise Ratio (PSNR) [47], Structural SIMilarity (SSIM) [48], Normalized Root Mean Square Error (NRMSE), and Fréchet Inception Distance (FID) [49]. Among them, PSNR is used to measure the degree of deformation and noise; SSIM is used to describe the degree of similarity of the graphics structure; NRMSE is used to measure the pixel error; FID is used to measure the perceptual error based on deep features, using a pre-trained Inception-V3 model [50].

Table 1 presents the numerical results of our model and current state-of-the-art model on the 2K and 4K image testsets. It can be seen that our model performs better on both 2K and 4K testsets for all numerical metrics. It indicates that our framework is ahead of the state-of-the-art method in the quality of inpainting at 2K and 4K resolutions, better adding pixel-level details, better recovering the global structure, and obtaining more realistic results on perception.

Table 1. Quantitative comparison of our model and the state-of-the-art model.

Method	2K Inpainting				4K Inpainting			
	PSNR↑	SSIM↑	NRMSE↓	FID↓	PSNR↑	SSIM↑	NRMSE↓	FID↓
HiFill	21.386	0.810	0.175	1.193	20.503	0.813	0.239	2.082
SR-Inpaint	27.364	0.923	0.092	0.097	26.065	0.910	0.130	0.138

6. Conclusions

We propose a general deep learning framework for the reconstruction of missing high-frequency information in high-resolution image through a super-resolution enhancement mechanism. Compared with traditional deep learning inpainting techniques, our model can handle both 2K and 4K images. Since our model adopts a “degradation and refinement” strategy, the computational overhead is well suppressed, while the inpainting quality is guaranteed. In addition, compared with the current state-of-the-art high-resolution inpainting model, our model leads in both visual results and numerical metrics, achieving semantic continuity, texture clarity, and visual fidelity. In the future, we will further optimize the network structure and training strategy to achieve better results as well as higher efficiency.

Author Contributions: Conceptualization, B.J., C.H. and S.L.; Data curation, H.X., Y.H., K.Z. and X.L.; funding acquisition, B.J., C.H. and S.L.; investigation, C.H.; methodology, B.J., H.X. and Y.H.; project administration, B.J.; resources, C.H. and K.Z.; software, H.X., Y.H. and B.J.; visualization, Y.H., X.L., K.Z. and H.F.; writing—original draft, H.X., Y.H., X.L. and H.F.; writing—review and Editing, H.X. and B.J. 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. 61907025, 61807020, 61702278), the Natural Science Foundation of Jiangsu Higher Education Institutions of China (Grant No. 19KJB520048) and NUPTSF (Grant No. NY219069).

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank all the anonymous reviewers for their valuable suggestions to improve this work.

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

Abbreviations

The following abbreviations are used in this paper:

HR	High-Resolution
LR	Low-Resolution
SR	Super-Resolution
CNN	Convolutional Neural Networks
GAN	Generative Adversarial Networks

References

1. Ballester, C.; Bertalmio, M.; Caselles, V.; Sapiro, G.; Verdera, J. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Trans. Image Process.* **2001**, *10*, 1200–1211. [[CrossRef](#)] [[PubMed](#)]
2. Esedoglu, S.; Shen, J. Digital inpainting based on the Mumford–Shah–Euler image model. *Eur. J. Appl. Math.* **2002**, *13*, 353–370. [[CrossRef](#)]
3. Liu, D.; Sun, X.; Wu, F.; Li, S.; Zhang, Y.Q. Image compression with edge-based inpainting. *IEEE Trans. Circuits Syst. Video Technol.* **2007**, *17*, 1273–1287.
4. Waykule, M.; Patil, M. Region filling and object removal by exemplar-based image inpainting. *Int. J. Sci. Eng. Res.* **2012**, *3*, 2229–5518.
5. He, K.; Sun, J. Statistics of patch offsets for image completion. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 16–29.
6. Drori, I.; Cohen-Or, D.; Yeshurun, H. Fragment-based image completion. In *ACM SIGGRAPH 2003 Papers*; ACM: New York, NY, USA, 2003; pp. 303–312.
7. Wilczkowiak, M.; Brostow, G.J.; Tordoff, B.; Cipolla, R. Hole filling through photomontage. In Proceedings of the BMVC 2005-Proceedings of the British Machine Vision Conference, Oxford, UK, 5–8 September 2005.
8. Xu, Z.; Sun, J. Image inpainting by patch propagation using patch sparsity. *IEEE Trans. Image Process.* **2010**, *19*, 1153–1165.
9. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv* **2019**, arXiv:1901.00212.
10. Yeh, R.A.; Chen, C.; Lim, T.Y.; Schwing, A.G.; Hasegawa-Johnson, M.; Do, M.N. Semantic Image Inpainting with Deep Generative Models. *arXiv* **2017**, arXiv:1607.07539.

11. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative Image Inpainting with Contextual Attention. *arXiv* **2018**, arXiv:1801.07892.
12. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* **2017**, *36*, 1–14. [[CrossRef](#)]
13. Oord, A.V.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel Recurrent Neural Networks. In *Machine Learning Research, Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016*; Balcan, M.F., Weinberger, K.Q., Eds.; PMLR: New York, New York, USA, 2016; Volume 48, pp. 1747–1756.
14. Liao, L.; Hu, R.; Xiao, J.; Wang, Z. Edge-Aware Context Encoder for Image Inpainting. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 3156–3160. [[CrossRef](#)]
15. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
16. Xiong, W.; Yu, J.; Lin, Z.; Yang, J.; Lu, X.; Barnes, C.; Luo, J. Foreground-Aware Image Inpainting. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5833–5841. [[CrossRef](#)]
17. Yi, Z.; Tang, Q.; Azizi, S.; Jang, D.; Xu, Z. Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 7505–7514. [[CrossRef](#)]
18. Yang, C.; Lu, X.; Lin, Z.; Shechtman, E.; Wang, O.; Li, H. High-Resolution Image Inpainting Using Multi-scale Neural Patch Synthesis. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4076–4084. [[CrossRef](#)]
19. Ikehata, S.; Cho, J.H.; Aizawa, K. Depth map inpainting and super-resolution based on internal statistics of geometry and appearance. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, Australia, 15–18 September 2013; pp. 938–942. [[CrossRef](#)]
20. Kim, S.Y.; Aberman, K.; Kanazawa, N.; Garg, R.; Wadhwa, N.; Chang, H.; Karnad, N.; Kim, M.; Liba, O. Zoom-to-Inpaint: Image Inpainting with High-Frequency Details. *arXiv* **2021**, arXiv:2012.09401.
21. Efros, A.A.; Freeman, W.T. Image Quilting for Texture Synthesis and Transfer. In Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 12–17 August 2001; pp. 341–346. [[CrossRef](#)]
22. Efros, A.; Leung, T. Texture synthesis by non-parametric sampling. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Corfu, Greece, 20–27 September 1999; Volume 2, pp. 1033–1038. [[CrossRef](#)]
23. Wang, L.; Chen, W.; Yang, W.; Bi, F.; Yu, F.R. A State-of-the-Art Review on Image Synthesis With Generative Adversarial Networks. *IEEE Access* **2020**, *8*, 63514–63537. [[CrossRef](#)]
24. Pan, X.; Zhan, X.; Dai, B.; Lin, D.; Loy, C.C.; Luo, P. Exploiting Deep Generative Prior for Versatile Image Restoration and Manipulation. *arXiv* **2020**, arXiv:2003.13659.
25. Brock, A.; Donahue, J.; Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv* **2019**, arXiv:1809.11096.
26. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv* **2018**, arXiv:1710.10196.
27. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434.
28. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5908–5916. [[CrossRef](#)]
29. Huang, T.S. Advances in Computer Vision and Image Processing: A Research Annual: Image Enhancement and Restoration, v. 2. Available online: <http://a.xueshu.baidu.com/usercenter/paper/show?paperid=ff63d1c895dbe3a66d889dbc93368fad> (accessed on 15 July 2021).
30. Stark, H.; Yang, Y. *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics*; Wiley-Interscience: Hoboken, NJ, USA, 1998.
31. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [[CrossRef](#)]
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034. [[CrossRef](#)]
33. Farrugia, R.A.; Guillemot, C. Light Field Super-Resolution Using a Low-Rank Prior and Deep Convolutional Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 1162–1175. [[CrossRef](#)]
34. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. *arXiv* **2016**, arXiv:1511.04587.
35. Dong, C.; Loy, C.C.; Tang, X. Accelerating the Super-Resolution Convolutional Neural Network. *arXiv* **2016**, arXiv:1608.00367.

36. Talab, M.A.; Awang, S.; Najim, S.A.d.M. Super-Low Resolution Face Recognition using Integrated Efficient Sub-Pixel Convolutional Neural Network (ESPCN) and Convolutional Neural Network (CNN). In Proceedings of the 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), Shah Alam, Malaysia, 29 June 2019; pp. 331–335. [[CrossRef](#)]
37. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1132–1140. [[CrossRef](#)]
38. Plötz, T.; Roth, S. Neural Nearest Neighbors Networks. *arXiv* **2018**, arXiv:1810.12575.
39. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual Dense Network for Image Super-Resolution. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2472–2481. [[CrossRef](#)]
40. Zhang, K.; Zuo, W.; Zhang, L. Deep Plug-and-Play Super-Resolution for Arbitrary Blur Kernels. *arXiv* **2019**, arXiv:1903.12529.
41. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *arXiv* **2016**, arXiv:1603.08155.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
43. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv* **2018**, arXiv:1611.07004.
44. 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. [[CrossRef](#)]
45. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
46. Timofte, R.; Gu, S.; Wu, J.; Van Gool, L.; Zhang, L.; Yang, M.H.; Haris, M.; Shakhnarovich, G.; Ukita, N.; Hu, S.; et al. NTIRE 2018 Challenge on Single Image Super-Resolution: Methods and Results. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 965–96511. [[CrossRef](#)]
47. Sara, U.; Akter, M.; Uddin, M.S. Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study. *J. Comput. Commun.* **2019**, *7*, 8–18. [[CrossRef](#)]
48. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003; IEEE: Manhattan, NY, USA, 2003; Volume 2, pp. 1398–1402.
49. 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. *arXiv* **2017**, arXiv:1706.08500.
50. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [[CrossRef](#)]