



Article

Lightweight Reference-Based Video Super-Resolution Using Deformable Convolution

Tomo Miyazaki , Zirui Guo and Shinichiro Omachi * 

Graduate School of Engineering, Tohoku University, Sendai 9808579, Japan; tomo@tohoku.ac.jp (T.M.); guo.zirui.r8@alumni.tohoku.ac.jp (Z.G.)

* Correspondence: shinichiro.omachi.b5@tohoku.ac.jp

Abstract: Super-resolution is a technique for generating a high-resolution image or video from a low-resolution counterpart by predicting natural and realistic texture information. It has various applications such as medical image analysis, surveillance, remote sensing, etc. However, traditional single-image super-resolution methods can lead to a blurry visual effect. Reference-based super-resolution methods have been proposed to recover detailed information accurately. In reference-based methods, a high-resolution image is also used as a reference in addition to the low-resolution input image. Reference-based methods aim at transferring high-resolution textures from the reference image to produce visually pleasing results. However, it requires texture alignment between low-resolution and reference images, which generally requires a lot of time and memory. This paper proposes a lightweight reference-based video super-resolution method using deformable convolution. The proposed method makes the reference-based super-resolution a technology that can be easily used even in environments with limited computational resources. To verify the effectiveness of the proposed method, we conducted experiments to compare the proposed method with baseline methods in two aspects: runtime and memory usage, in addition to accuracy. The experimental results showed that the proposed method restored a high-quality super-resolved image from a very low-resolution level in 0.0138 s using two NVIDIA RTX 2080 GPUs, much faster than the representative method.

Keywords: super-resolution; reference image; texture information; deformable convolution



Citation: Miyazaki, T.; Guo, Z.; Omachi, S. Lightweight Reference-Based Video Super-Resolution Using Deformable Convolution. *Information* **2024**, *15*, 718. <https://doi.org/10.3390/info15110718>

Academic Editors: Nikolaos Mitianoudis and Ilias Theodorakopoulos

Received: 30 September 2024
Revised: 2 November 2024
Accepted: 6 November 2024
Published: 8 November 2024



Copyright: © 2024 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

Super-resolution (SR) is a technique for enhancing the resolution of an image or video to obtain a high-resolution counterpart by predicting natural and realistic texture information not included in the original input. Recently, machine learning-based methods have achieved high accuracy. Convolutional neural networks (CNNs) have contributed to significant improvements in super-resolution. SR is a useful technique in computer vision and has various applications. For example, in medical image analysis [1], converting low-resolution medical images into high-resolution ones enables a more accurate detection of subtle lesions and abnormalities. If applied to surveillance [2], it will enhance identification capabilities in criminal investigations and crime prevention activities. In the field of remote sensing [3], it is helpful in obtaining a more detailed analysis of land use and disaster monitoring. Pan et al. applied super-resolution to the images a UAV taken to detect defects in a transmission line insulator [4].

The typical SR method is the single-image SR, which generates the high-resolution (HR) image using only a single low-resolution (LR) image. However, the generated HR image suffers from blurry noise due to the limited information in the LR image. To tackle this problem, attempts were made using a generative adversarial network (GAN) [5]. The images generated by GAN are sharp; however, the details may have the wrong textures. Another approach is the reference-based SR. Generally, two individual images, LR and HR

images, are used in this approach. The LR image is the target that needs to be converted to the SR image. The HR image is the reference image, which differs from the LR image. High-frequency information is extracted from the reference image and then used to convert the LR image to the SR image. The main challenge is extracting useful high-frequency information to fill a gap between the LR and the reference images. There are attempts to match both images based on an optical flow [6] and patch-matching [7]. Zheng et al. [6] developed an optical flow-based approach. However, due to the structural limitation of its module, which calculates the optical flow called FlowNet [8], the length and width of the input image must be an integer multiple of 16. Zhang et al. [7] developed a patch-matching-based network called SRNTT. However, the patch matching stage is time-consuming as the operator executed in this stage is not implemented on a GPU. Recently, Yang et al. [9] proposed a Transformer-based network, TTSR. However, the Transformer consumes a lot of memory. In summary, the existing methods are inefficient at computational time and memory consumption.

Image-processing technology using machine learning is expected to be used in various situations, including edge computers. Using super-resolution technology in environments other than those with high-speed, large-capacity GPUs is desirable. Therefore, there is a demand for the development of easy-to-use methods. The purpose of this study is to propose a lightweight method for video super-resolution using reference images. The overview of the proposed method is shown in Figure 1. The proposed method inputs a low-resolution image and a high-resolution reference image and outputs a super-resolution image of the low-resolution image. Instead of time-consuming matching algorithms, such as optical flow and patch-matching necessary for feature alignment, we adopt deformable convolution [10,11] to align high-frequency information to the LR image. We call the proposed method Reference-based Super-Resolution with Deformable Convolutional Network, or RSRDCN. Our method accepts the input of arbitrary size and runs relatively faster than current methods with much less memory usage. The experimental results on time and memory usage are described in Section 4.4.2. The proposed method makes the reference-based super-resolution a technology that can be easily used even in environments with limited computational resources.

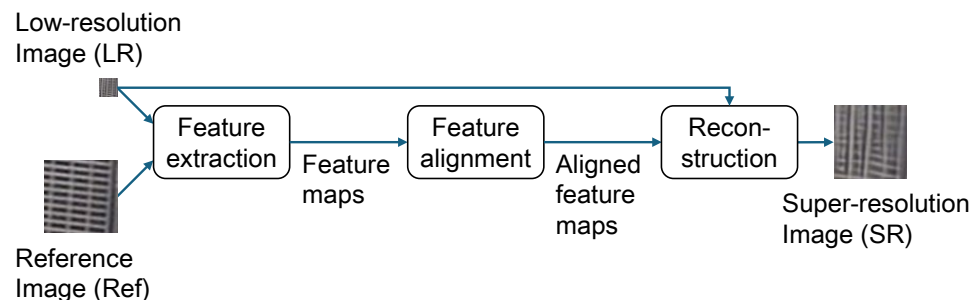


Figure 1. Overview of the proposed RSRDCN. An SR image of the input LR image is generated by aligning texture information in the reference image (Ref) to the LR image. Deformable convolution is adopted for feature alignment.

The rest of this paper is organized as follows. First, we introduce some related works in Section 2. The proposed reference-based super-resolution method is described in detail in Section 3. In Section 4, we describe the experimental settings and results. Finally, Section 5 provides some concluding remarks regarding this study.

2. Related Work

This section introduces some works that are related to our proposed method. First, we discuss works on single-image super-resolution, and then those on video super-resolution. Finally, we introduce the works on reference-based super-resolution that are most related to our method and discuss the similarities and differences between our proposed and existing methods.

2.1. Single-Image Super-Resolution

Single-image SR is a technique for generating an SR image from an LR image, and deep learning has improved performance on SR. The first attempt using deep learning was SRCNN [12]. This is an end-to-end image SR approach. Timofte et al. developed a DIV2K dataset [13] composed of 1000 images in 2 K resolution. The DIV2K dataset is widely used for training neural networks in SR. The residual block [14] was used to develop deeper networks, such as EDSR [15] and RCAN [16]. The aforementioned methods used pixel-based loss functions, such as mean square error, MSE, or mean absolute error, MAE. However, generated SR images based on these criteria tend to be blurred. To tackle this problem, perceptual loss [17] and adversarial loss [18] have been developed to incorporate human perceptions to generate sharp SR images.

Recently, much research has been conducted using the Transformer [19] that has an attention mechanism. Although the Transformer was originally proposed for natural language processing, Kolesnikov et al. applied the Transformer to image recognition tasks [20]. Liang et al. proposed a method for image restoration [21] using Swin Transformer [22]. They showed the effect of the proposed SwinIR on several representative tasks, including image super-resolution. Yao et al. proposed a super-resolution algorithm for omnidirectional images based on the enhanced SwinIR [23]. Zheng et al. proposed the Efficient Mixed Transformer by combining global and local Transformer layers [24].

2.2. Video Super-Resolution

Video SR is a technique to generate SR frames from LR frames. To this end, research has been conducted to align information between LR frames, and the optical flow has often been used for this purpose [25,26]. The optical flow field between a center frame and its neighboring frames is estimated, and then the neighboring frames are warped according to the field. However, accurate optical flow is hard to estimate if large motions occur between the frames. To tackle this problem, a dynamic filter [27] can be used. Jo et al. proposed the dynamic upsampling filters that are generated locally and dynamically according to the spatiotemporal neighborhoods [28].

Another approach is to use deformable convolution [10,11], which can model geometric transformations that were the limitation of the traditional convolution kernel due to its fixed configuration. It has been used to tackle high-level vision tasks such as object detection and semantic segmentation. The deformable convolutional networks have been widely used in video super-resolution tasks. Tian et al. proposed the temporally deformable alignment network that performs temporal alignment [29]. Wang et al. proposed enhanced deformable convolutions for video restoration, EDVR, by introducing a pyramid, cascading and deformable alignment module [30]. Chan et al. [31] showed that deformable alignment can be formulated as a combination of feature-level flow-warping and convolution. They also extracted offsets from the pre-trained EDVR model and compared them with optical flows. After quantitatively studying the correlation between the offsets and optical flows, they found that over 80% of the estimations have a difference smaller than one pixel from the optical flow.

Recent works have revealed the effectiveness of reference frames in video super-resolution. Zhang et al. proposed the use of bidirectional optical flows calculated from intermediate frames [32]. Feng et al. supplemented spatial information in an LR video by combining it with an HR video using a hybrid imaging system [33]. We followed the findings of these works to develop the proposed method. However, we used the deformable convolution instead of the optical flow since the optical flow is time-consuming.

2.3. Reference-Based Super-Resolution

Reference-based SR complements the details in a low-resolution input image using another high-resolution image. A primary challenge of the reference-based SR is the alignment between the LR and reference images. Zheng et al. developed CrossNet using optical flow for alignment [6]. However, calculating optical flow is time-consuming. Furthermore,

the quality of generated SR images largely depends on the preciseness of the alignment. SRNTT [7] uses patch matching to align features between the LR and reference images. The Transformer is also used for reference-based SR. TTSR [9] adopted the Transformer to select useful patches in the reference image for the LR image. Liu et al. proposed a strategy based on dual-view supervised learning and multi-attention mechanism [34]. Their method integrated the supervised signals of feature and image layers to optimize the network.

Shim et al. used the deformable convolution kernels for the reference-based SR [35]. They used predicted alignment parameters and offsets in the deformable convolution kernels from input images. Inspired by this work, the proposed method also uses deformable convolution kernels. In addition, since the proposed method aims to develop a lightweight network with faster speed and lower memory consumption, it introduces a pyramidal feature alignment scheme to predict alignment parameters efficiently.

3. Proposed Method

The proposed reference-based SR method, RSRDCN, mainly consists of three processes as shown in Figure 1: feature extraction, feature alignment, and reconstruction. The RSRDCN receives a low-resolution input image (LR) and a high-resolution reference image (Ref) and learns to generate an SR image of the input LR image by aligning texture information in the reference image to the LR image. As mentioned above, the purpose of this study is to develop a lightweight model. To achieve this, we introduce feature alignment at multiple scales using deformable convolution.

3.1. Feature Extraction

Feature extraction from the reference image is crucial in the reference-based SR because appropriate features help to restore SR images accurately. We adopted VGG19 [36] for feature extraction since it has shown the ability to extract high-level perceptual information in various image-processing tasks, such as style transfer [37], which also concerns transferring textures between two images.

The feature extraction process is displayed in Figure 2. Reference-based super-resolution needs to find the regions in the Ref image corresponding to each LR region. However, it is difficult to compare the low-resolution and high-resolution images directly. Inspired by TTSR [9], we used features extracted from LR \uparrow and Ref $\downarrow\uparrow$ in addition to the Ref image. LR \uparrow was obtained by resizing an LR image into a 4 \times larger image using bicubic-upsampling. Ref $\downarrow\uparrow$ was obtained by applying bicubic-downsampling and then upsampling to the reference image Ref with the scale factor 4 \times . Then, VGG19 extracted three feature maps at the 10th, 5th, and 2nd layers. These feature maps were the same size, half the size, and quarter the size of the reference image, denoted by subscripts 1, 2, and 3, respectively. The feature maps obtained from LR \uparrow are denoted as $F_{LR\uparrow_1}$, $F_{LR\uparrow_2}$, and $F_{LR\uparrow_3}$. The feature maps of Ref $\downarrow\uparrow$ and Ref are denoted the same way.

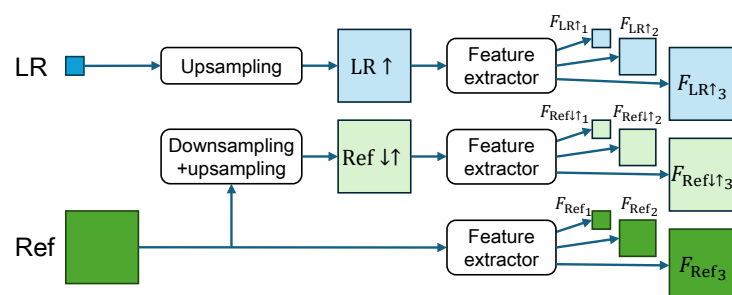


Figure 2. Feature extraction. LR \uparrow represents the upsampled LR image. Ref $\downarrow\uparrow$ is the downsampled and then upsampled reference image.

3.2. Feature Alignment

The proposed feature alignment module is displayed in Figure 3. We introduced deformable convolution [10,11] for feature alignment because traditional optical flow-based methods are time-consuming, and the latest Transformer-based methods [9] consume a huge amount of memory while calculating relevance embedding. The offset for deformable convolution was estimated from LR \uparrow and Ref $\downarrow\uparrow$, and the feature map of the reference image was aligned according to the offset. We introduced a pyramidal feature alignment module inspired by EDVR [30] to achieve an efficient alignment.

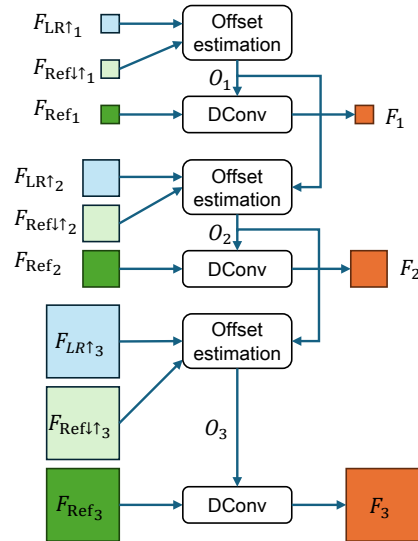


Figure 3. Feature alignment. The offset for deformable convolution is estimated from low resolution to high resolution using LR \uparrow and Ref $\downarrow\uparrow$, and the feature map of the reference image is aligned according to the offset.

The deformable convolution kernel is defined as Equation (1):

$$y(p) = \sum_{k=1}^{n^2} w_k \cdot x(p + m_k + o_k). \tag{1}$$

The kernel produces the output $y(p)$ by applying the weights w_k to an input x at a location p . m_k represents a set of movements in the kernel such as $m_k \in \{(-1, -1), (-1, 0), \dots, (1, 1)\}$. The input x is deformed by the offset o_k . The offset has the same spatial size as the feature map, and the number of channels is $C_{\text{offset}} = n^2 \times C_{\text{feature}}$, where C_{feature} denotes the number of channels of the features and n denotes the kernel size usually set to 3.

We used the deformable convolution kernels to align the feature maps extracted from the LR and Ref images. Firstly, We integrated features extracted from LR \uparrow and Ref $\downarrow\uparrow$ to predict the offset

$$O_i = f([F_{LR\downarrow\uparrow i}, F_{Ref\uparrow\downarrow i}]), \tag{2}$$

where f is a general function consisting of several convolution layers, and $[\cdot, \cdot]$ denotes the concatenation operation.

We wanted to predict the offset using LR \uparrow and Ref $\downarrow\uparrow$. However, the prediction was not straightforward since the viewpoints of the LR image and the Ref image were different. To mitigate this problem and achieve an efficient prediction, we introduced the same pyramid strategy as EDVR [30] to gradually predict the offset and perform alignment. The aligned feature map F_i is generated as

$$F_i = \text{DConv}(F_{\text{Ref}_i}, g([O_i, O_{i-1}^{\uparrow 2}])), \tag{3}$$

where $\text{DConv}(\cdot)$ is the deformable convolution operator, $(\cdot)^{\uparrow 2}$ refers to upsampling by a factor 2 using bilinear interpolation, and g is a general function with several convolution layers. The output of the alignment module is a pyramid of spatial size-aligned feature maps.

3.3. Reconstruction

In the reconstruction process, an HR image was generated by applying a deep generative network, the reconstructor, to the input LR image and the aligned feature maps at multiple scales. The overview of our reconstruction process is shown in Figure 4a. After passing through one reconstructor, the feature map was magnified to a $2 \times$ larger spatial size. In this study, we used three reconstructors to generate HR images.

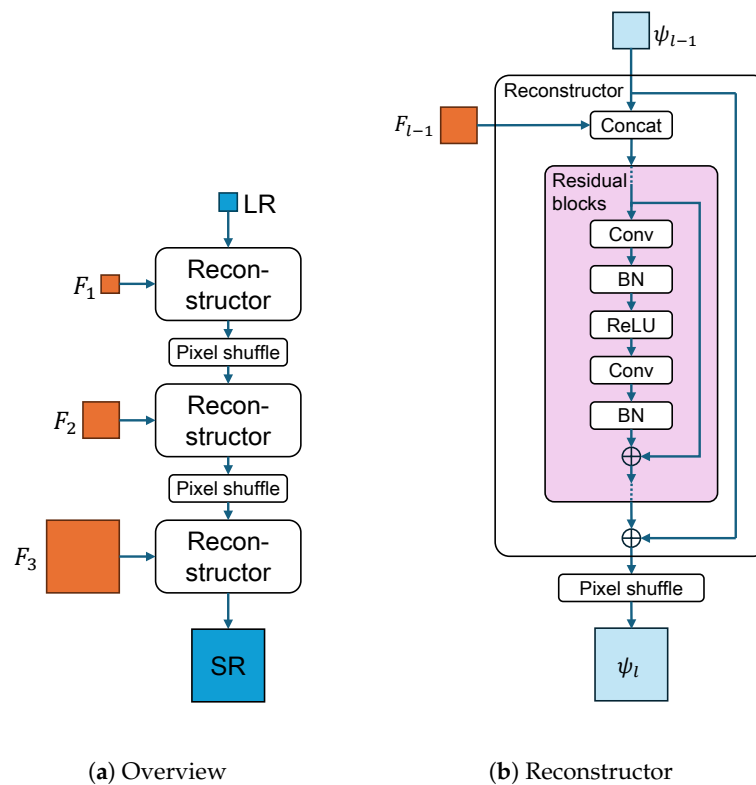


Figure 4. Reconstruction process. The input image is progressively super-resolved using multi-scale aligned feature maps.

The architecture of the reconstructor is illustrated in Figure 4b. We used residual blocks to reconstruct the images since a very deep trainable network helps to recover images [16]. For the l th reconstructor, we firstly concatenated the aligned feature map F_{l-1} and a feature map ψ_{l-1} that was the output feature map from the $(l - 1)$ th reconstructor. We used the input LR image as ψ_1 . Then, we sent the concatenated feature map into residual blocks. Finally, the pixel shuffle block [38] was used to magnify the output.

3.4. Loss Function

We used two loss components: reconstruction loss \mathcal{L}_{rec} and perceptual loss \mathcal{L}_{per} . The loss function is defined as Equation (4). We set λ_{rec} to 1 and λ_{per} to 0.1.

$$L = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{per}} \mathcal{L}_{\text{per}}. \tag{4}$$

As the reconstruction loss, the Charbonnier penalty function [39], defined as Equation (5), was used for achieving a higher PSNR. It is widely accepted as a loss function in the SR tasks for its robustness and ability to handle outliers. ε is set to 1×10^{-3} .

$$\mathcal{L}_{\text{rec}} = \sqrt{\|I^{GT} - I^{SR}\|^2 + \varepsilon^2}. \quad (5)$$

Perceptual loss has shown its ability in [17] to achieve better visual quality. With perceptual loss, we can enhance the similarity in feature space. The perceptual loss is defined as Equation (6). $\phi_i^{\text{vgg}}(\cdot)$ denotes the i th layer's feature map of VGG19. We used the relu1_1, relu2_1, relu3_1 layer. (C_i, H_i, W_i) denotes the shape of the feature map at the i th layer.

$$\mathcal{L}_{\text{per}} = \sum_I \frac{1}{C_i H_i W_i} \left\| \phi_i^{\text{vgg}}(I^{SR}) - \phi_i^{\text{vgg}}(I^{HR}) \right\|_2^2. \quad (6)$$

4. Experimental Results

In this section, we present experimental results to verify the effectiveness of the proposed method. First, we describe the dataset used in the experiments and the baseline models. Then, we clarify the implementation details. Finally, we present the experimental results and discussions.

4.1. Datasets

We used the datasets commonly used in super-resolution research. The training dataset is the REalistic and Dynamic Scenes dataset, REDS [40], which contains high-quality videos. The REDS dataset contains 240 training clips and 30 validation clips. Each clip consists of 100 consecutive frames. We re-grouped them to obtain 270 datasets, from which 4 clips were selected as validation datasets. The validation dataset is denoted as REDS4, as in the paper [30], in which EDVR was proposed. Specifically, REDS4 contains 000, 010, 015, and 020 clips. We let the $(10i + 4)$ th frame in a set to be the reference frame for the $[10i, 10i + 9]$ frame, i.e., 10 frames share a single reference frame. So, we have 23,940 sets of images for one epoch. We crop the LR images to 128×128 during training to save memory. Here, we use random cropping for data augmentation.

We use Vid4 for the validation dataset. Vid4 [41] contains four clips of video. However, as the CrossNet [6] requires the height and weight to be a multiple of 16, we cropped the dataset. The *calendar* and *city* clips were cropped to 144×176 , and *foliage* and *walk* clips were cropped to 112×176 . The center frame in each clip serves as the reference image for the whole clip. We also used the CUFED5 dataset [7] as one of our validation datasets. CUFED5 defines four similarity levels from high to low, i.e., L1, L2, L3, and L4, according to the number of best matches of SIFT features. The testing set contains 126 groups of samples.

4.2. Baseline Models

We compared the proposed method, RSRDCN, with several existing methods. These are representative methods often used for benchmarking super-resolution methods. We also took Bicubic interpolation into account. The baseline methods are described below.

Single-Image Super-Resolution Method

We used the single image super-resolution architecture, RCAN [16], as one of our baselines. Because its pre-trained model is not trained on the REDS dataset, we trained this model on the REDS dataset for 50 epochs.

Reference-based Super-Resolution Method

We used CrossNet [6] and SRNTT [7] as our reference-based SR baselines. We used the pre-trained model provided on GitHub.

4.3. Implementation Details

Each mini-batch contains four LR images with size 128×128 along with four reference images with size 512×512 . We used the gradient accumulation training strategy to obtain a bigger mini-batch size. The accumulation step was set to four, and the actual mini-batch size was 16. We trained our model with Adam optimizer by setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was 4×10^{-4} . We implemented our model with the PyTorch framework and trained them using two NVIDIA RTX 2080 GPUs.

4.4. Results

To show the effect of the proposed method, we first present the results of a quantitative comparison between the proposed method and baseline methods. Then, the proposed method is compared with the reference-based SR methods regarding runtime and memory usage.

4.4.1. Quantitative Evaluation

To evaluate the proposed method, we reported the PSNR and SSIM scores of every validation dataset in Tables 1 and 2. The bold and underlined scores represent the best and second-best scores, respectively. The proposed method (RSRDCN) obtained the highest PSNR and SSIM on REDS4 and CUFED5. Also, RSRDCN achieved the second-best on Vid4. Specifically, the PSNR of RSRDCN surpassed SRNTT by 0.66 points on REDS4. When compared with CrossNet, RSRDCN improved PSNR by 5.99 points on REDS4. Likewise, the SSIM of RSRDCN outperformed SRNTT and CrossNet by 0.036 and 0.311 on REDS4, respectively.

Table 1. Average PSNR over the validation datasets. Bold and underlined scores represent the best and second-best scores, respectively.

| Method | Vid4 | REDS4 | CUFED5 Levels | | | | |
|---------------|--------------|--------------|---------------|--------------|--------------|--------------|--------------|
| | | | 1 | 2 | 3 | 4 | 5 |
| Bicubic | 20.41 | 25.93 | 22.92 | 22.92 | 22.92 | 22.92 | 22.92 |
| RCAN | <u>20.00</u> | <u>28.15</u> | <u>24.21</u> | <u>24.21</u> | <u>24.21</u> | <u>24.21</u> | <u>24.21</u> |
| CrossNet | 18.63 | 22.28 | 20.18 | 20.06 | 20.09 | 20.04 | 20.07 |
| SRNTT | 19.04 | 27.61 | 24.12 | 24.09 | 24.09 | 24.06 | 24.09 |
| RSRDCN (ours) | <u>20.00</u> | 28.27 | 24.24 | 24.24 | 24.23 | 24.23 | 24.24 |

Table 2. Average SSIM over the validation datasets. Bold and underlined scores represent the best and second-best scores, respectively.

| Method | Vid4 | REDS4 | CUFED5 Levels | | | | |
|---------------|--------------|--------------|---------------|--------------|--------------|--------------|--------------|
| | | | 1 | 2 | 3 | 4 | 5 |
| Bicubic | <u>0.520</u> | 0.724 | 0.632 | 0.632 | 0.632 | 0.632 | 0.632 |
| RCAN | 0.513 | <u>0.802</u> | 0.712 | 0.712 | 0.712 | 0.712 | 0.712 |
| CrossNet | 0.372 | 0.504 | 0.465 | 0.451 | 0.454 | 0.448 | 0.448 |
| SRNTT | 0.488 | 0.779 | <u>0.717</u> | <u>0.715</u> | <u>0.715</u> | <u>0.714</u> | <u>0.714</u> |
| RSRDCN (ours) | 0.530 | 0.815 | 0.727 | 0.727 | 0.727 | 0.727 | 0.727 |

On the Vid4 validation dataset, Bicubic achieved the best and second-best performance at PSNR and SSIM. However, the reconstructed images of Bicubic are deteriorated visually. For example, as shown in Figure 5, Bicubic was better than SRNTT at PSNR and SSIM, whereas SRNTT has restored clear texture. The result of RSRDCN was sharper than that of Bicubic. Also, RSRDCN outperformed Bicubic and SRNTT at PSNR and SSIM. Figure 6 shows another result. The proposed RSRDCN achieved the best scores for both PSNR and SSIM among all methods except for Bicubic. Figure 7 displays the cropped images of Figure 6. RSRDCN reconstructed the horizontal frames of the building's windows, while

RCAN and SRNTT failed to recover the correct texture. Although the PSNR and SSIM of Bicubic were the best, the image is blurry and undesirable as a high-resolution image.



Figure 5. The reconstructed images of the first frame of Vid4's calendar clip. The values are PSNR and SSIM.

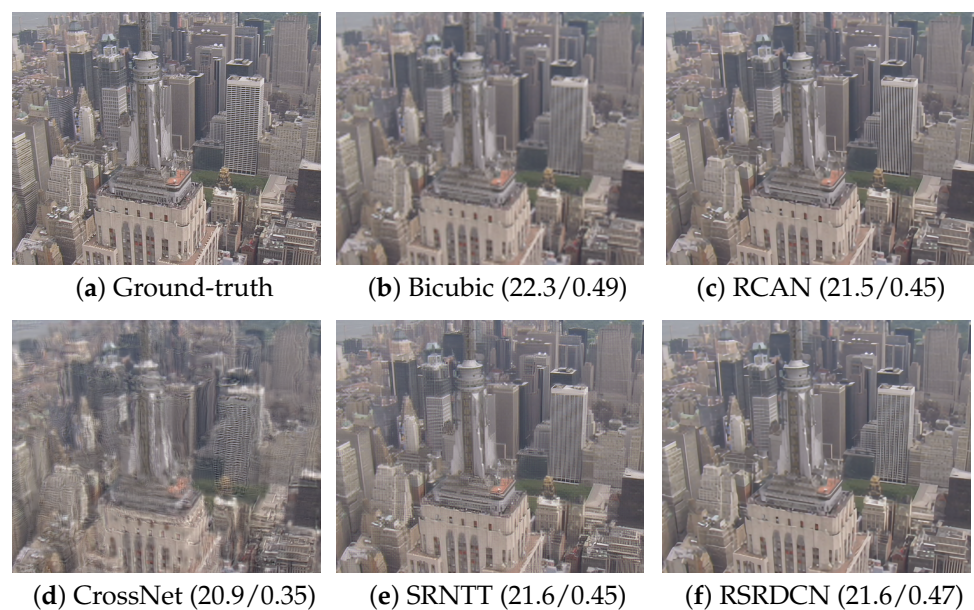


Figure 6. The reconstructed images of the first frame of Vid4's city clip. The values are PSNR and SSIM.

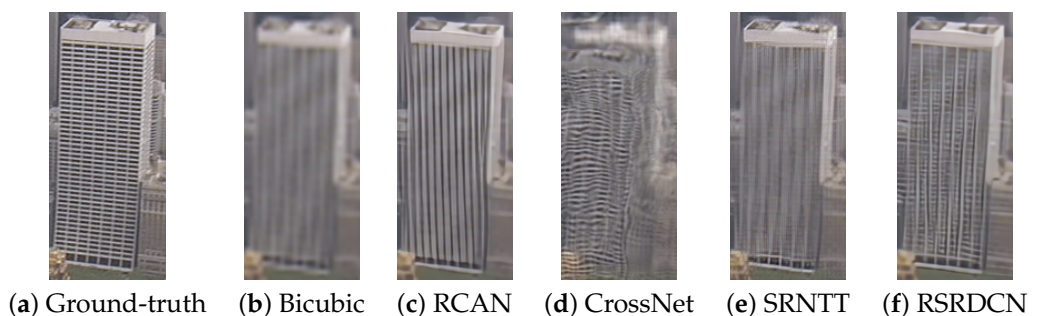


Figure 7. Cropped results of the first frame of Vid4's city clip.

Figures 8 and 9 show results of the REDS and CUFED5's level 1, respectively. In Figure 8, SRNTT recovered the texture of the stone but generated a fault pattern of bricks. In Figure 9, SRNTT recovered the shape of chairs while the others did not. From these examples, SRNTT mostly recovered fine textures from reference images but sometimes recovered fault textures. Our RSRDCN architecture achieved smoother images than SRNTT and could recover correct patterns when significant information was lost.

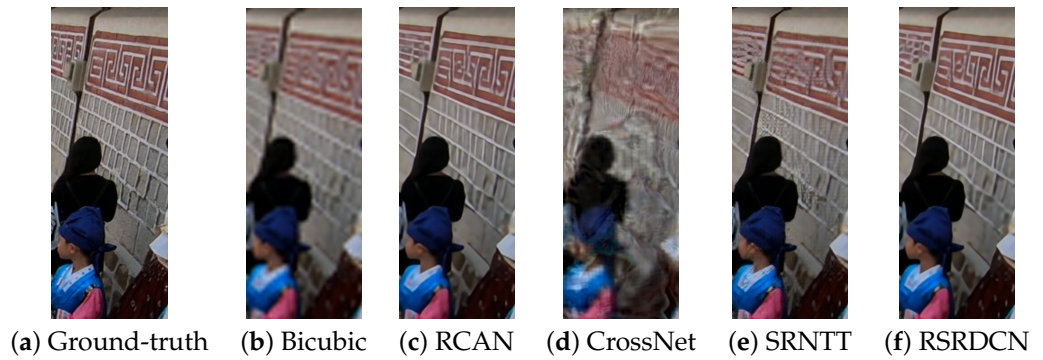


Figure 8. Cropped results of the 35th frame of REDS' 010 clip.

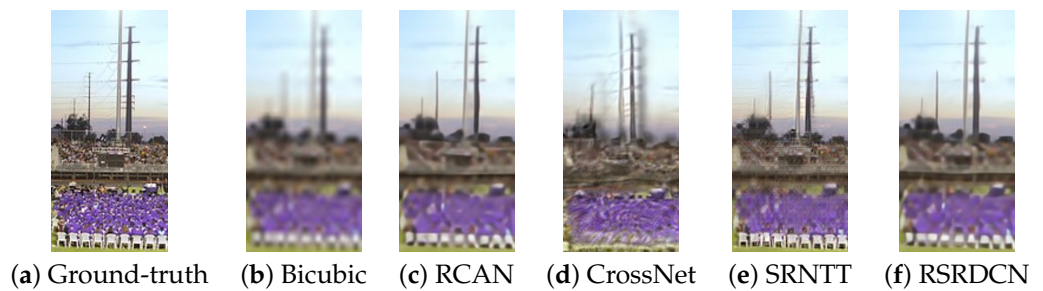


Figure 9. Cropped results of the first frame of CUFED5's level 1.

4.4.2. Analysis on Running Time and Memory Usage

We evaluated the computational time and memory usage of each model for one frame on an NVIDIA GTX 1080 Ti using the Vid4 dataset. The results are shown in Figure 10. RSRDCN's average runtime was 0.0138 s, while CrossNet's average runtime was 0.146 s. SRNTT's average runtime was 4.666 s, which means SRNTT took 30 times longer than RSRDCN. CrossNet took 1.5 times longer than RSRDCN.

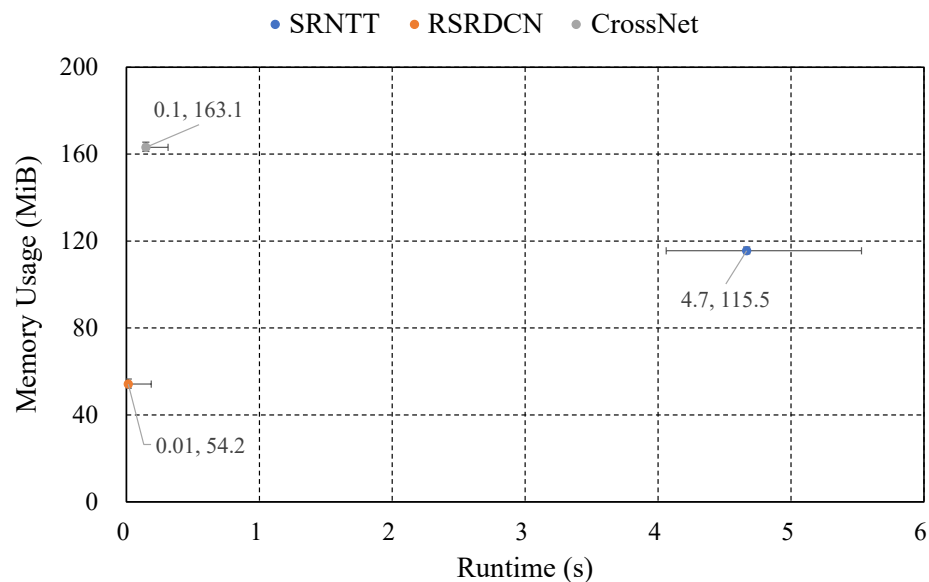


Figure 10. The runtime and memory usage for reference-based super-resolution methods on the Vid4 dataset.

As for memory usage, RSRDCN's average memory usage was 54.186 MiB, CrossNet's average memory was 163.132 MiB, and SRNTT's average memory was 115.533 MiB. SRNTT required twice as much memory as RSRDCN. CrossNet required three times as much

memory as RSRDCN. Without bells and whistles, RSRDCN outperformed CrossNet and SRNTT on runtime and memory usage.

Methods that consume large amounts of computational resources have been proposed to improve the accuracy of reference-based super-resolution. Still, this study has confirmed that reference-based super-resolution can be achieved with a certain degree of accuracy with much less computational effort than conventional methods.

5. Conclusions

We have proposed a reference-based super-resolution method using deformable convolutional networks. This study aimed to develop a lightweight model to make the reference-based super-resolution applicable in practical use. By employing deformable convolution, we could exploit high-frequency information in the reference frame with little time and memory usage. Our proposed model consists of three processes. First, the multi-scale feature maps are extracted from the LR and reference images. Then, the aligned feature maps are obtained using deformable convolution. Finally, the SR image is reconstructed from the input LR image using the aligned feature maps.

We compared the proposed model with other reference-based models in the experiments. Our model achieved the best scores in PSNR and SSIM, while it ran much faster and required less memory than optical flow-based methods. The experimental result showed that the average runtime of the proposed method was 0.0138 s, which is more than 30 times faster than that of a representative method. We believe that our proposed method enables reference-based super-resolution to be used in various environments, including edge computers.

The importance and effect of deformable convolutional networks in super-resolution have been demonstrated in this study. However, the role of deformable convolutional networks in texture migration has not been verified. In the future, we may need to conduct extra experiments to address deformable convolutional networks' role in texture migration.

Author Contributions: Conceptualization, T.M. and Z.G.; methodology, Z.G.; software, Z.G.; validation, T.M. and Z.G.; formal analysis, Z.G.; investigation, Z.G. and T.M.; resources, Z.G.; data curation, Z.G.; writing—original draft preparation, T.M. and Z.G.; writing—review and editing, S.O.; visualization, Z.G.; supervision, S.O.; project administration, S.O.; funding acquisition, S.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grants JP22H00540 and JP23K11176.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data were derived from public domain resources.

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

References

1. Gupta, R.; Sharma, A.; Kumar, A. Super-resolution using GANs for medical imaging. *Procedia Comput. Sci.* **2020**, *173*, 28–35. [[CrossRef](#)]
2. Zhang, L.; Zhang, H.; Shen, H.; Li, P. A super-resolution reconstruction algorithm for surveillance images. *Signal Process.* **2010**, *90*, 848–859. [[CrossRef](#)]
3. Wang, P.; Bayram, B.; Sertel, E. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Sci. Rev.* **2022**, *232*, 104110. [[CrossRef](#)]
4. Pan, L.; Chen, L.; Zhu, S.; Tong, W.; Guo, L. Research on small sample data-driven inspection technology of UAV for transmission line insulator defect detection. *Information* **2022**, *13*, 276. [[CrossRef](#)]
5. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*, Montreal, QC, Canada, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: New York, NY, USA, 2014; Volume 27.

6. Zheng, H.; Ji, M.; Wang, H.; Liu, Y.; Fang, L. CrossNet: An end-to-end reference-based super resolution network using cross-scale warping. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 87–104. [[CrossRef](#)]
7. Zhang, Z.; Wang, Z.; Lin, Z.; Qi, H. Image super-resolution by neural texture Transfer. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7974–7983. [[CrossRef](#)]
8. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Häusser, P.; Hazirbas, C.; Golkov, V.; Smagt, P.V.D.; Cremers, D.; Brox, T. FlowNet: Learning optical flow with convolutional networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2758–2766. [[CrossRef](#)]
9. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning texture Transformer network for image super-resolution. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, DC, USA, 14–19 June 2020; pp. 5790–5799. [[CrossRef](#)]
10. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773. [[CrossRef](#)]
11. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets V2: More deformable, better results. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9300–9308. [[CrossRef](#)]
12. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 184–199. [[CrossRef](#)]
13. Agustsson, E.; Timofte, R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1122–1131. [[CrossRef](#)]
14. 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](#)]
15. 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](#)]
16. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 294–310. [[CrossRef](#)]
17. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 694–711. [[CrossRef](#)]
18. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 105–114. [[CrossRef](#)]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in 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.: New York, NY, USA, 2017; Volume 30.
20. Kolesnikov, A.; Dosovitskiy, A.; Weissenborn, D.; Heigold, G.; Uszkoreit, J.; Beyer, L.; Minderer, M.; Dehghani, M.; Houtsby, N.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR) 2021, Virtual, 3–7 May 2021. [[CrossRef](#)]
21. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. SwinIR: Image restoration using Swin Transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844. [[CrossRef](#)]
22. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002. [[CrossRef](#)]
23. Yao, X.; Pan, Y.; Wang, J. An omnidirectional image super-resolution method based on enhanced SwinIR. *Information* **2024**, *15*, 248. [[CrossRef](#)]
24. Zheng, L.; Zhu, J.; Shi, J.; Weng, S. Efficient mixed transformer for single image super-resolution. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108035. [[CrossRef](#)]
25. Haris, M.; Shakhnarovich, G.; Ukita, N. Recurrent back-projection network for video super-resolution. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3892–3901. [[CrossRef](#)]

26. Chan, K.C.; Wang, X.; Yu, K.; Dong, C.; Loy, C.C. BasicVSR: The search for essential components in video super-resolution and beyond. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021. [[CrossRef](#)]
27. De Brabandere, B.; Jia, X.; Tuytelaars, T.; Van Gool, L. Dynamic filter networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 667–675.
28. Jo, Y.; Oh, S.W.; Kang, J.; Kim, S.J. Deep Video super-resolution network using dynamic upsampling filters without explicit motion compensation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3224–3232. [[CrossRef](#)]
29. Tian, Y.; Zhang, Y.; Fu, Y.; Xu, C. TDAN: Temporally-deformable alignment network for video super-resolution. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, DC, USA, 14–19 June 2020; pp. 3357–3366. [[CrossRef](#)]
30. Wang, X.; Chan, K.C.; Yu, K.; Dong, C.; Loy, C.C. EDVR: Video restoration with enhanced deformable convolutional networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 1954–1963. [[CrossRef](#)]
31. Chan, K.C.; Wang, X.; Yu, K.; Dong, C.; Loy, C.C. Understanding deformable alignment in video super-resolution. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021.
32. Zhang, Y.; Wang, H.; Zhu, H.; Chen, Z. Optical flow reusing for high-efficiency space-time video super resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 2116–2128. [[CrossRef](#)]
33. Feng, Z.; Zhang, W.; Liang, S.; Yu, Q. Deep video super-resolution using hybrid imaging system. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 4855–4867. [[CrossRef](#)]
34. Liu, X.; Li, J.; Duan, T.; Li, J.; Wang, Y. DSMA: Reference-based image super-resolution method based on dual-view supervised learning and multi-attention mechanism. *IEEE Access* **2022**, *10*, 54649–54659. [[CrossRef](#)]
35. Shim, G.; Park, J.; Kweon, I.S. Robust reference-based super-resolution with similarity-aware deformable convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8425–8434. [[CrossRef](#)]
36. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR) 2015, Diego, CA, USA, 7–9 May 2015. [[CrossRef](#)]
37. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423. [[CrossRef](#)]
38. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883. [[CrossRef](#)]
39. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep Laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 624–632. [[CrossRef](#)]
40. Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; Lee, K.M. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 1996–2005. [[CrossRef](#)]
41. Liu, C.; Sun, D. On Bayesian adaptive video super resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 346–360. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.