Deep Learning For Super-Resolution Of Unregistered Multi-Temporal Satellite Images

DEEP LEARNING FOR SUPER-RESOLUTION OF UNREGISTERED MULTI-TEMPORAL SATELLITE IMAGES

Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, Enrico Magli

Politecnico di Torino

ABSTRACT

Recently, convolutional neural networks (CNN) have been successfully applied to many remote sensing tasks. However, deep learning for multi-image superresolution from multi-temporal imagery has received little attention so far. We propose a residual CNN that exploits both spatial and temporal correlations in the low-resolution image set by using 3D convolutional layers to combine multiple images from the same scene. The experiments have been carried out using a dataset of PROBA-V satellite ground images, composed of several low-resolution and high-resolution images taken at different times from instruments on the same platform, in the context of a challenge issued by the European Space Agency.

***Index Terms*—** Multi-image superresolution, convolutional neural networks, multi-temporal images

1. INTRODUCTION

Super-resolution (SR) techniques reconstruct a high resolution (HR) image from one or more low resolution (LR) images. Despite the continuous development of ever more advanced optical devices, the limitations and the high cost of hardware technology still highlight the importance of developing post-processing techniques to achieve high spatial resolution. This is especially important in fields such as video surveillance, medical diagnosis, and remote sensing. Remote sensing is playing an increasingly important role in mapping and monitoring the Earth. Increasing the availability of high spatial resolution remote sensing data is crucial for many applications such as urban mapping, military surveillance, intelligence gathering, disaster and vegetation growth monitoring.

The approaches to image super-resolution can be framed into two main categories: single-image SR (SISR) and multi-image SR (MISR). SISR exploits spatial correlation in a single image to recover the HR version. The literature on SISR approaches is extensive and includes classic interpolation methods such as bicubic and Lanczos filters, optimization-based methods explicitly modeling prior knowledge about natural images such as low total variation [1] or sparsity [2, 3] and data-driven approaches learning such a prior in the form of convolutional neural networks (CNNs) [4, 5, 6].

This research has been funded by the Smart-Data@PoliTO center for Big Data and Machine Learning technologies.

However, the amount of information available in a single image is quite limited as some information has inevitably been lost in the LR image formation process. Certain applications provide multiple LR versions of the same scene to be combined by means of MISR techniques, where the reconstruction of high spatial-frequency details takes full advantage of the complementary information coming from different observations of the same scene. For remote sensing problems, multiple images of the same scene can typically be acquired by a spacecraft during multiple orbits, or may be obtained by multiple satellites imaging the same scene. In the context of remote sensing, MISR was first explored by Tsai and Huang [7] who used multiple under-sampled images with sub-pixel displacements to improve the spatial resolution of Landsat TM images. Many other classical MISR algorithms were proposed over the years. Irani and Peleg [8] introduced the iterative back-projection approach (IBP) which aims to improve an initial guess of the super-resolved image by iteratively inverting the forward imaging process. Elad et al. [9] proposed to reconstruct an HR image from LR images by using projection onto convex sets (POCS), maximum likelihood (ML), and maximum a posteriori (MAP) estimators for a linear noisy forward model. Zhang et al. [10] presented a reconstruction algorithm which adaptively weights the contributions of images acquired from multiple angles. Ma et al. [11] proposed an operational SR approach for multi-angle WorldView-2 remote sensing images, which consists of image registration and super-resolution reconstruction. The former corrects for the local geometric distortion and photometric disparity. The latter uses an ℓ_1 norm data fidelity term and a total variation regularizer. To the best of our knowledge, deep learning has not yet been employed for MISR in the remote sensing field, where important issues such as image registration, invariance to absolute brightness variability and unreliable data (e.g., due to cloud coverage) must be handled in order to develop a successful MISR model.

In this paper we present a deep learning architecture aiming to tackle the problem of MISR applied to a novel dataset provided by ESA's Advanced Concept Team in the context of a challenge [12]. The goal is to super-resolve images from the PROBA-V satellite. The unique feature of this dataset is that both LR and HR images have been acquired by the same spacecraft, as opposed to previous works where LR images

are artificially down-scaled, degraded and shifted versions of an HR image. The images are not simultaneously acquired so temporal variations exist and have to be handled as well in the SR process.

2. THE PROBA-V SR DATASET

At present, it is difficult to find a dataset collecting both a set of real-world LR observations and the corresponding HR image for the same scene, captured from the same platform. Many of the works found in literature are based on simulated data, where LR observations for a specific scene are obtained through a degradation and down-sampling process of the HR images by assuming a sensor imaging model. This is a simplified scenario as it either assumes a non-blind problem, i.e., the degradation model can be characterized to some extent, or has the limitation that a too simple degradation model may not accurately match the real one.

The Advanced Concepts Team of the European Space Agency has issued a competition to perform MISR for the images acquired by the PROBA-V satellite. The PROBA-V satellite is an Earth observation satellite designed to map land cover and vegetation growth across the entire globe. It was launched in 2013 into a sun-synchronous orbit at an altitude of 820km. Its payload provides an almost global coverage with 300m LR images and 100m HR images. However, the HR images are acquired with a higher revisit time, roughly one every 5 days, instead of one per day. The dataset gathers satellite data from 74 regions located around the world from the PROBA-V mission. Images are provided as level 2A products composed of radiometrically and geometrically corrected Top-of-Atmosphere reflectance in Plate Carre projection for the RED and NIR spectral bands. The size of the collected images is 128×128 and 384×384 for the LR and HR data respectively. The images have a single channel with a bit-depth of 14 bits. Each data point consists of one HR image and several LR images (ranging from a minimum of 9 to a maximum of 30) from the same scene. In total, the dataset contains 1160 scenes, 566 are from NIR spectral band and 594 are from RED band. The images of a specific scene are captured multiple times over a maximum period of 30 days. Weather and changes in the landscape pose a limitation in the similarity of the images. Clouds, cloud shadows, ice, water, missing regions, presence of agricultural activities and, in general, human activity are the main sources of inconsistency across these images, thus posing a major challenge for any image fusion method. Moreover, each image comes with a mask, indicating which pixels in the image can be reliably used for reconstruction (e.g., they are not covered by clouds). Subpixel shifts in the content do occur and are indeed important for the MISR task. The geometric disparity among the images can be considered as translational only.

The unique nature of this dataset (with real LR and HR images captured by the same platform at multiple times) makes for an interesting case study for SR techniques. Developing SR products from multiple, more frequent LR images could simultaneously provide enhanced resolution and higher temporal availability and is therefore an interesting application of MISR. Moreover, having real images of the same scene for both the low and high resolutions enables data-driven methods such as CNNs to learn the inversion of possibly complex degradation models and the best feature fusion strategy to handle temporal variations.

3. PROPOSED METHOD

The proposed method aims to reconstruct a high-resolution image I^{HR} given a set of N LR images $I^{\text{LR}}_{[0,N-1]}$ representing the same scene:

$$I^{\text{SR}} = f(I^{\text{LR}}_{[0,N-1]}, \theta),$$

where θ represents the model parameters and f represents the mapping function from LR to HR. $I^{\text{LR}}_{[0,N-1]}$ and I^{HR} are represented as real-valued tensors with shape $N \times H \times W \times C$ and $1 \times rH \times rW \times C$ respectively, where H and W are the height and the width of the input LR frames, C is the number of channels and r is the scale factor. While the LR images roughly represent the same scene as the HR image, there are several factors to be considered:

- the LR images are not registered with each other;
- the LR images and the HR image are not registered;
- the brightness of the HR image may be different from that of any LR image;
- LR and HR images may be covered by clouds and cloud shadows or affected by corrupted pixels.

To tackle this problem we propose to employ a supervised deep learning approach, where a CNN learns the residual between bicubic interpolation and the ground truth HR image of the scene. First, a bicubic interpolator scales the LR images to the desired size; such images are then registered with respect to a reference image. The registration is handled by the phase correlation algorithm translating the images by integral shifts at the high resolution (or equivalently, sub-pixel shifts at low resolution). After these preprocessing operations, the interpolated registered LR (IRLR) images are fed into a CNN composed of two main building blocks. An overview of the network is shown in Fig. 1. The first block, referred as SISRNet, is a CNN aiming to perform a SISR task where each of the N images is processed independently by a sequence of 2D convolutional layers. The convolutional filters are shared along the temporal dimension, i.e., all the N LR images go through the same set of filters. The second block, referred as FusionNet, aims to merge the representations of the images in the feature space in a "slow" fashion, i.e., by exploiting a sequence of 3D convolutions with small kernels. The N outputs of the SISRNet are progressively fused by $N/2 - 1$ 3D convolutional layers until the network's depth reduces to 1.

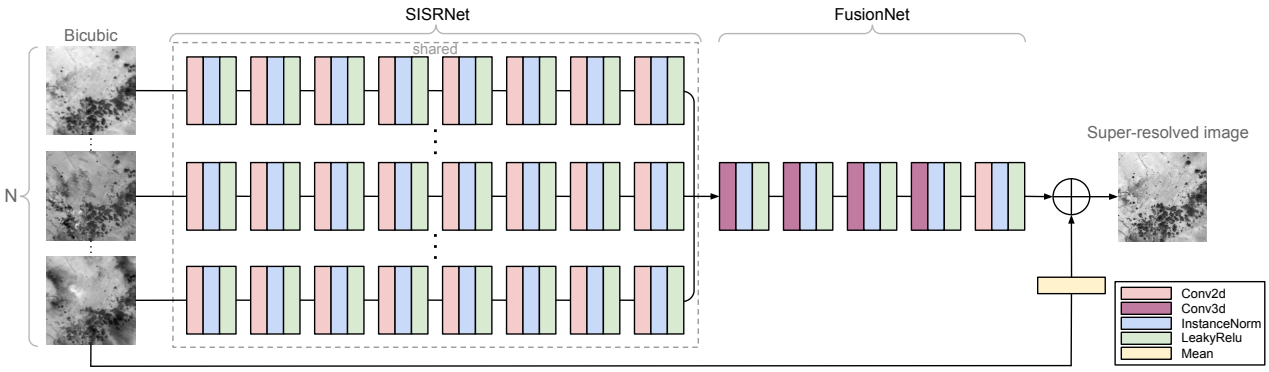

Fig. 1. MISR fusion network. The input N bicubic upsampled and registered images are processed by SISR and fusion subnetworks to produce a residual image. The residual image is then added to the average of the input to obtain the SR image.

Slow fusion in the feature space allows the network to learn the best space to decouple image features that are relevant to the fusion from irrelevant variations and to construct the best function to exploit spatio-temporal correlations [13].

The proposed architecture employs an input-output residual connection. The network estimates the high-frequency residual details to be added to the input, which carries most of the low-frequency information, in order to obtain the high-resolution image. This is an established technique for image restoration problems using deep learning [4], including SISR. However, with respect to SISR, the network proposed in this work implements a many-to-one mapping, so the residual is actually added to a basic merge of the IRLR images in the form of their average:

$$\bar{I}^{\text{IRLR}} = \frac{1}{N} \sum_{i \in [0,N-1]} I_i^{\text{IRLR}},$$

$$I^{\text{SR}} = \bar{I}^{\text{IRLR}} + R,$$

being R the residual estimated by the CNN.

3.1. Loss Function

Model parameters are optimized minimizing a loss function computed as a modified version of the Euclidean distance between the SR image and the HR target. Minimizing the Euclidean distance is optimal in terms of the mean-squared error metric. Some deep learning works on SISR attempted to use an adversarial loss [14]. While this approach produces visually pleasing results, it tends to hallucinate information, resulting in lower MSE scores and less reliable products in the context of remote sensing; hence, the adversarial approach has not been followed in the present work. As we mentioned in Sec. 2, since the PROBA-V satellite does not capture LR images and HR images of a specific ground scene simultaneously, there are discrepancies coming from different weather conditions, changes in the landscape and variable absolute brightness due to the large inter-time interval between satellite shots. The LR images could be quite different from one another and from the corresponding HR image as well. For this reason, we must make the training objective as invariant as possible to such conditions. In particular, in order to build invariance to absolute brightness differences between I^{SR} and I^{HR}, the modified loss function equalizes the intensities of the SR and HR images so that the average pixel brightness of both images match. Moreover, since the I^{SR} and I^{HR} could be shifted by a maximum of 3 pixels, the loss embeds a shift correction. I^{SR} is cropped by a 3 pixel border, then all possible patches ${I_{u,v}}^{\text{HR}}$ of size $(rH-3)\times(rW-3)$ are extracted from the target I^{HR}. All possible Euclidean distances are computed and the lowest one is taken as loss to optimize. Our loss takes three inputs: residual estimate, network input (averaged IRLR image) and ground truth HR image:

$$L = \min_{u,v \in [0,6]} \parallel I_{u,v}^{\text{HR}} - (I_{\text{crop}}^{\text{SR}} + b) \parallel^2,$$

where b represents the brightness correction and $I_{\text{crop}}^{\text{SR}}$ the cropped network output:

$$b = \frac{1}{(rW-3)(rH-3)} \sum_{x,y} \left(I_{u,v}^{\text{HR}} - I_{\text{crop}}^{\text{SR}} \right),$$

$$I_{\text{crop}}^{\text{SR}} = \bar{I}^{\text{IRLR}} + R.$$

The loss is computed by utilizing only the HR image pixels that are marked as reliable by the mask provided with the dataset and the SR image pixels for which at least one out of N LR images were clear. The reason for this is that a cloud in the HR image can never be predicted from terrain data in the IRLR images, so its pixels should not contribute to the loss function. Viceversa, it is also impossible to predict HR terrain if all the IRLR images have concealed regions.

4. RESULTS

In this section we perform an experimental evaluation of the proposed method, comparing it with several alternative approaches. Such approaches include a CNN-based SISR method, a MISR baseline consisting of averaging bicubic-interpolated and registered images, and a MISR method where a CNN performs SISR on each IRLR image and then the obtained SR images are averaged.

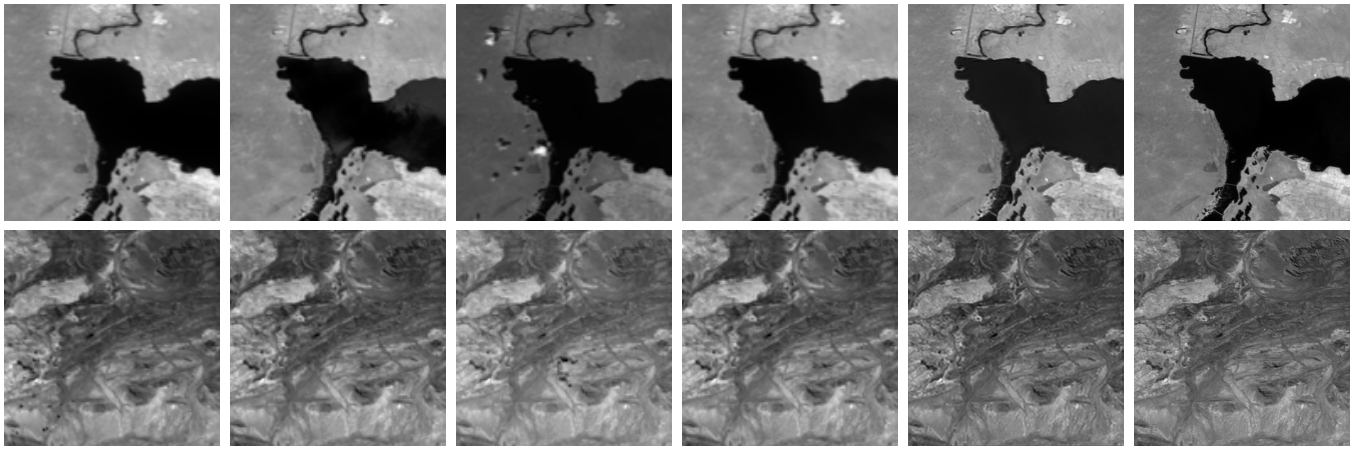

Fig. 2. Left to right: 4 random LR images, SR image reconstructed by our method and HR image

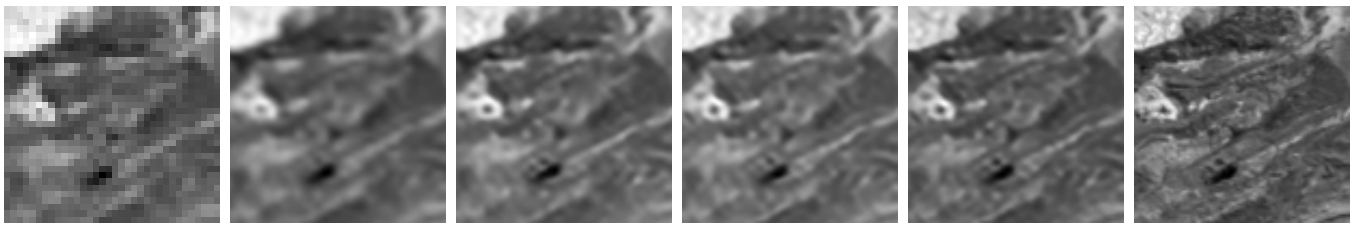

Fig. 3. Left to right: one among the LR images, baseline, SISR only, SISR+Mean, our method, HR image

Table 1. Average mPSNR (dB) - PROBA-V Test Set

	Proposed	SISR+Mean	Baseline	SISR
NIR	**47.21**	46.48	45.63	45.56
RED	**49.52**	48.79	47.88	48.21

4.1. Experimental setting and training process

In the following experiments, we employed both NIR and RED band datasets described in Sec. 2 and the same training procedure has been applied separately for them. We used 396 scenes for training and 170 for testing from NIR band dataset and 415 for training and 176 for testing from RED band dataset. Since the proposed network is devised to work with a fixed size temporal dimension, we decided to train the network using the minimum number of images available for each scene, i.e., 9 images. When more images are available we select the 9 clearest images according to the masks. For each scene the clearest image is considered as the reference to which all the other upsampled LR images are registered. Each scene is a data-cube of size $9 \times 384 \times 384$, from which we extracted 200 random patches, resulting in a total of 79200 and 83000 samples of size $9 \times 48 \times 48$ for NIR and RED band respectively. The network has been trained for around 80 epochs with a batch size of 32. The network has a first block composed by 8 2D convolutional layers with 64 filters each (see Fig. 1) and a second block having 4 3D convolutional layers and a last 2D convolutional layer. Each layer is followed by an Instance Norm layer and a Leaky ReLU nonlinearity, except for the last. Instance Normalization is used in place of Batch Norm layer to make the network training as independent as possible of the contrast and brightness differences among the input images. Finally, since the network produces a residual estimate R, we want $\bar{I}^{\text{IRLR}}$ and I^{HR} to be normalized so that their difference gives a unit variance residual R, thus avoiding any scaling to be performed by the last layer of the network and improving convergence speed.

4.2. Quantitative and qualitative results

We want to compare the proposed MISR technique, where image fusion is performed by 3D convolutions in the feature space, to a fusion method based on averaging SISR images, referred as SISR+Mean, as well as to a SISR only solution. A MISR baseline consisting in the averaged bicubic IRLR images is also included in the comparison. For all these methods we followed the same steps for the data preparation: bicubic interpolation and registration by phase correlation algorithm. The metric chosen for evaluation is a modified version of the PSNR (mPSNR) used also in the ESA challenge.

$$mPSNR = \max_{u,v \in [0,6]} 20 \log \frac{2^{16}-1}{\parallel I^{\text{HR}}_{u,v} - (I^{\text{SR}}_{\text{crop}} + b) \parallel^2}$$

The mPSNR computation is meant only for pixels that are not concealed both in the target HR image and in the reconstructed image. This metric has been devised to cope with the high sensitivity of the PSNR to biases in brightness and with the relative translation the reconstructed image might

have with respect to the target HR image, in the same way the loss function does during training. In this case the maximum mPSNR over all possible shifts is considered for evaluation. Note that by design of the dataset, the maximum shift in the horizontal and vertical directions is equal to 6 pixels.

Table 1 shows the mPSNR results on both NIR and RED test sets for various methods. It can be noticed that the proposed method outperforms all the other methods, confirming the effectiveness of the 3D fusion network at exploiting the correlation among the images along the temporal dimension. As a term of comparison, the SISR+Mean method uses the same SISR network (with the addition of a final layer projecting from the feature space to the image space) of our method but replaces the whole 3D fusion network with a simple average across the images. SISR+Mean is outperfomed by our method by 0.93 dB and 0.73 dB for NIR and RED test sets respectively. The last MISR method included in the evaluation is the baseline solution that is a simple average across the bicubic IRLR images, giving worst results than our method. Indeed, this baseline is the input-output connection in the proposed architecture. The comparison between our method and the SISR only method is meant to highlight the huge gain brought by exploiting both the spatial and temporal correlations, even if the LR images of a specific scene are taken under different conditions and might be wildly different from one another in terms of contrast, brightness and landscape due to temporal variations. The SISR only solution is based on a residual network with 2D convolutional layers and roughly the same number of learnable parameters with respect to our method for fair comparison. The choice of a single scene image was made using the masks, taking the clearest image. The SISR only method is outperfomed by our method by 1.65 dB and 1.31 dB for NIR and RED testset respectively. This result shows that exploiting temporal correlation on sub-pixel shifted images improves the quality of the reconstructed HR image. These quantitative results are accompanied by a qualitative comparison of the super-resolution methods in Fig. 3. It can be noticed that our proposed method produces visually more detailed images, recovering finer texture and sharper edges. All methods are able to reconstruct images with minimal presence of artifacts.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have introduced a CNN architecture to deal with super-resolution applied to multi-temporal images . We showed that fusing multiple images from the same ground scene allows better accuracy at increasing spatial resolution, even when significant temporal variations are present. In the present work the registration is handled in the preprocessing stage by the phase correlation algorithm. Future work will investigate an enhanced solution which account for an integration of the registration task, as a trainable block, directly in the proposed CNN architecture in order to exploit its powerful feature representations.

6. REFERENCES

[1] Michael K Ng, Huanfeng Shen, Edmund Y Lam, and Liangpei Zhang, "A total variation regularization based super-resolution reconstruction algorithm for digital video," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 074585, 2007.

[2] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, Nov 2010.

[3] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*, Berlin, Heidelberg, 2012, pp. 711–730, Springer Berlin Heidelberg.

[4] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. PP, 08 2016.

[5] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang, "Non-local recurrent network for image restoration," in *Advances in Neural Information Processing Systems*, 2018, pp. 1673–1682.

[6] Diego Valsesia, Giulia Fracastoro, and Enrico Magli, "Image denoising with graph-convolutional neural networks," in *2019 26th IEEE International Conference on Image Processing (ICIP)*, 2019.

[7] R.Y. Tsai and T.S. Huang, "Multiframe image restoration and registration," in *In Advances in Computer Vision and Image Processing*. JAI Press: Greenwich, CT, USA, 1984, vol. I, pp. 317–339.

[8] Michal Irani and Shmuel Peleg, "Super resolution from image sequences," 07 1990, vol. ii, pp. 115 – 120 vol.2.

[9] Michael Elad and Arie Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 6, pp. 1646–58, 02 1997.

[10] Hongyan Zhang, Zeyu Yang, Liangpei Zhang, and Huanfeng Shen, "Super-resolution reconstruction for multi-angle remote sensing images considering resolution differences," *Remote Sensing*, vol. 6, 12 2013.

[11] J. Ma, J. Cheung-Wai Chan, and F. Canters, "An operational superresolution approach for multi-temporal and multi-angle remotely sensed imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 110–124, Feb 2012.

[12] Kelvins ESA's Advanced Concepts Team, "PROBA-V Super Resolution," `https://kelvins.esa.int/proba-v-super-resolution`.

[13] Jose Caballero, Christian Ledig, Andrew P. Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," *CoRR*, vol. abs/1611.05250, 2016.

[14] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.