
iSeeBetter: Spatio-Temporal Video Super Resolution using Recurrent-Generative Back-Projection Networks

Aman Chadha*

Stanford University

amanc@stanford.edu

Abstract

Recently, learning-based models have enhanced the performance of Single-Image Super-Resolution (SISR). However, applying SISR successively to each video frame leads to lack of temporal coherency. On the other hand, Video Super Resolution (VSR) models based on Convolutional Neural Networks (CNNs) outperform traditional approaches in terms of image quality metrics such as Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM). However, Generative Adversarial Networks (GANs) offer a competitive advantage in terms of being able to mitigate the issue of lack of finer texture details when super-resolving at large upscaling factors which is usually seen with CNNs. We present iSeeBetter, a novel spatio-temporal approach to VSR. iSeeBetter seeks to render temporally consistent Super Resolution (SR) videos by extracting spatial and temporal information from the current and neighboring frames using the concept of Recurrent Back-Projection Networks (RBPN) as its generator. Further, to improve the "naturalness" of the super-resolved image while eliminating artifacts seen with traditional algorithms, we utilize the discriminator from Super-Resolution Generative Adversarial Network (SRGAN). Mean Squared Error (MSE) as a primary loss-minimization objective improves PSNR and SSIM, but these metrics may not capture fine details in the image leading to misrepresentation of perceptual quality. To address this, we use a four-fold (adversarial, perceptual, MSE and Total-Variation (TV)) loss function. Our results demonstrate that iSeeBetter offers superior VSR fidelity and surpasses state-of-the-art performance.

1 Introduction

The goal of Super Resolution (SR) is to enhance a Low Resolution (LR) image to a Higher Resolution (HR) image by filling in missing fine-grained details in the LR image. This domain can be divided into three main areas: Single Image-SR (SISR) (1), (2), (3), (4), Multi Image SR (MISR) (5), (6) and Video SR (VSR) (7), (8), (9), (10), (11). The idea behind SISR is to to super-resolve an LR frame LR_t , independently of other frames in the video sequence. While this technique takes into account spatial information, it fails to exploit the temporal details inherent in a video sequence. MISR seeks to address just that - it utilizes the missing details available from neighboring frames and fuses them for super-resolving LR_t . After spatially aligning frames, missing details are extracted by separating differences between the aligned frames from missing details observed only in one or some of the frames. However, in MISR, the alignment of the frames is done without any concern for temporal smoothness, while in VSR, frames are typically aligned in temporal smooth order.

Traditional VSR methods upscale based on a single degradation model (usually bicubic interpolation), followed by reconstruction. This is sub-optimal and adds computational complexity (12). Recently, learning-based models based on Convolutional Neural Networks (CNNs) have outperformed traditional approaches in terms of widely-accepted image reconstruction metrics such as Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM). A crucial aspect of an effective VSR system is its ability to handle motion sequences since those are often important components of videos (7), (13).

The proposed method, iSeeBetter, is inspired by Recurrent Back-Projection Networks (RBPNS) (10), which utilize "back-projection" as their underpinning approach which was originally introduced in (14), (15). The basic concept behind back-projection is to iteratively calculate residual images as reconstruction error between a target image and a set of a neighboring

*Aman Chadha is an engineer in the System Performance and Architecture team at Apple Inc., but did this work at Stanford.

images. The residuals are then back-projected to the target image for improving super-resolution accuracy. The multiple residuals enable representing subtle and significant differences between the target frame and other frames and thus exploit temporal relationships between adjacent frames as shown in Figure 1. This results in superior SR accuracy.

To mitigate the issue of lack of finer texture details when super-resolving at large upscaling factors, which is usually seen with CNNs (16), iSeeBetter utilizes GANs with a loss function that weighs adversarial loss, perceptual loss (16), Mean Square Error (MSE)-based loss and Total-Variation (TV) loss (17). Our approach combines the merits of RBPN and SRGAN (16) - it is based on RBPN as its generator which is complemented by SRGAN’s discriminator architecture. Blending these techniques yields iSeeBetter, a state-of-the-art system that is able to recover photo-realistic textures and motion-based scenes from heavily down-sampled videos.

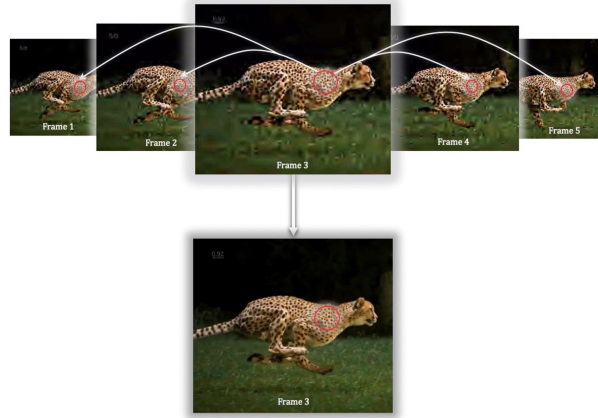


Figure 1: Adjacent frame similarity

Our contributions include the following key innovations.

Combining the state-of-the-art in SR: We propose a model that leverages two superior SR techniques - RBPN and SRGAN. RBPN enables iSeeBetter to extract details from neighboring frames, while the generator-discriminator architecture pushes iSeeBetter to generate more realistic frames and eliminate artifacts.

"Optimizing" the loss function: Minimizing MSE encourages finding pixel-wise averages of plausible solutions which are typically overly-smooth and thus have poor perceptual quality (18) (19) (20) (21). To address this, we adopt a four-fold (adversarial, perceptual, MSE and TV) loss for superior results.

Extended evaluation protocol: To evaluate iSeeBetter, we used standard datasets: Vimeo90K (22), Vid4 (23) and SPMCS (8). To expand the spectrum of data diversity, we wrote scripts to collect additional data from YouTube and augment our dataset to 170,000 clips.

User-friendly script infrastructure: We built several tools to download and structure datasets, visualize temporal profiles and run benchmarks to be able to iterate on different models quickly. Further, we also built a video-to-frames tool to enable directly input videos to iSeeBetter, rather than frames.

2 Related work

Learning-based methods have emerged as superior VSR techniques compared to traditional statistical methods. We thus focus our discussion in this section solely on learning-based methods that are trained end-to-end.

Deep VSR can be primarily divided into three types based on the approach to preserving temporal information.

(a) Temporal Concatenation. The most popular approach to retain temporal information in VSR is by concatenating the frames as in (24), (7), (11), (25). Essentially, this approach can be seen as an extension of SISR to accept multiple input images.

(b) Recurrent Networks. A many-to-one architecture is used in (26), (8) where a sequence of LR frames is mapped to a single target HR frame. A many-to-many RNN has recently been used in VSR by (9), to map the current LR frame and previous HR estimate to the target HR frame.

(c) Optical Flow-Based Methods. To reduce unwanted flickering artifacts in the output frames (17), (9) proposed a method that utilizes a network that is trained on estimating optical flow along with the SR network. Optical flow methods allow estimation of the trajectories of a moving objects, thereby assisting in VSR.

3 Datasets

To train iSeeBetter, we amalgamated diverse datasets with differing video lengths, resolutions, motion sequences and number of clips. Table 1 presents a summary of the datasets used. When training our model, we generated the corresponding LR frame for each HR input frame by performing $4\times$ down-sampling using bicubic interpolation. To extend our dataset further, we wrote scripts to collect additional data from YouTube. The dataset was shuffled for training and testing. Our training/validation/test split was 80%/10%/10%.

Dataset	Resolution	# of clips	# of frames/clip	# of frames
Vimeo90K	448×256	13,100	7	91,701
SPMCS	240×135	30	31	930
Vid4	$(720 \times 576 \times 3), (704 \times 576 \times 3), (720 \times 480 \times 3), (720 \times 480 \times 3)$	4	41, 34, 49, 47	684
Augmented	960×720	7,000	110	77,000
Total	-	46,034	-	170,315

Table 1. Datasets used for training and evaluation

4 Methods

4.1 Implementation²

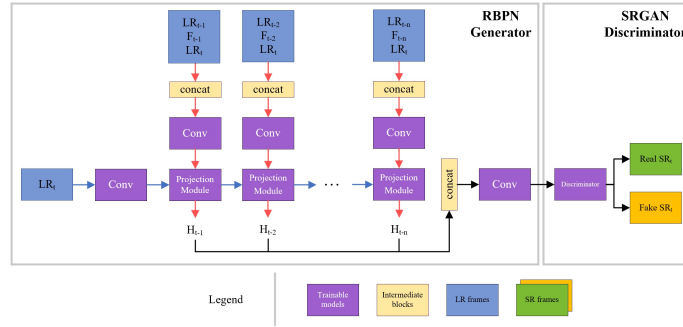


Figure 2: Overview of iSeeBetter

Figure 2 shows the iSeeBetter architecture which uses RBP (10) and SRGAN (16) as its generator and discriminator respectively. RBP has two approaches that extract missing details from different sources, namely SISR and MISR. Figure 3 shows the horizontal flow (blue arrows in Figure 2) that enlarges LR_t using SISR. Figure 4 shows the vertical flow (red arrows in Figure 2) which is based on MISR that computes residual features from a pair of LR_t to neighbor frames ($LR_{t-1}, \dots, LR_{t-n}$) and the flow maps (F_{t-1}, \dots, F_{t-n}). At each projection step, RBP observes the missing details from LR_t and extracts residual features from neighboring frames to recover details. Within the projection models, RBP utilizes a recurrent encoder-decoder mechanism for incorporating details extracted in SISR and MISR through back-projection.

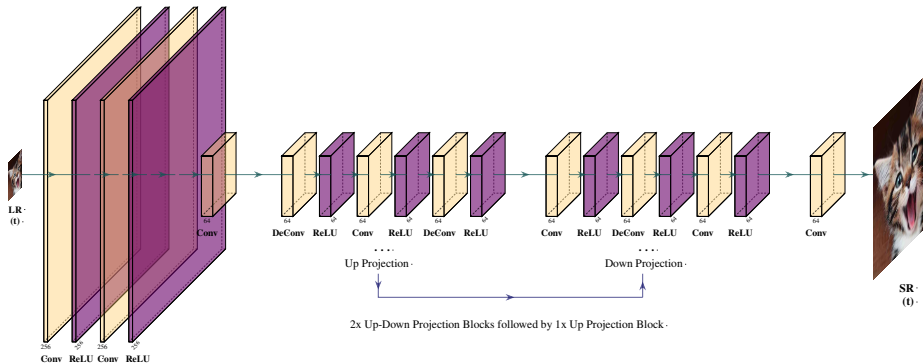


Figure 3: DBPN (2) architecture for SISR, where we perform up-down-up sampling using 8×8 kernels with stride of 4, padding of 2. Similar to the ResNet architecture above, the DBPN network also uses Parametric ReLUs (27) as its activation functions.

²Code and samples for the implementation are available at github.com/amanchadha/iSeeBetter

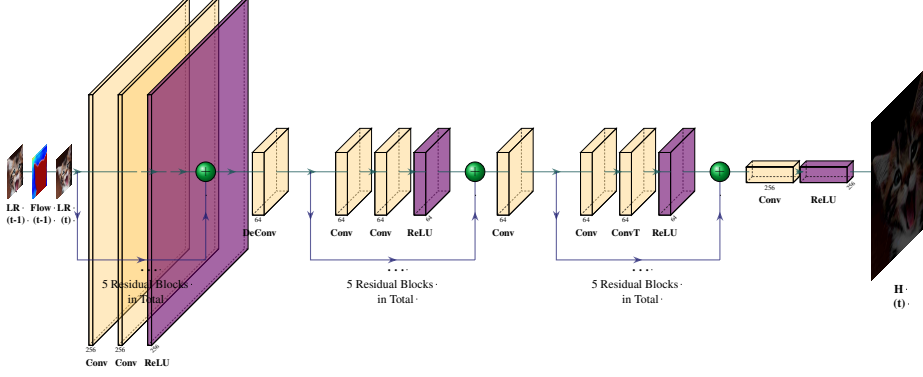


Figure 4: ResNet architecture for MISR that is composed of three tiles of five blocks where each block consists of two convolutional layers with 3×3 kernels, stride of 1 and padding of 1. The network uses Parametric ReLUs (27) for its activations.

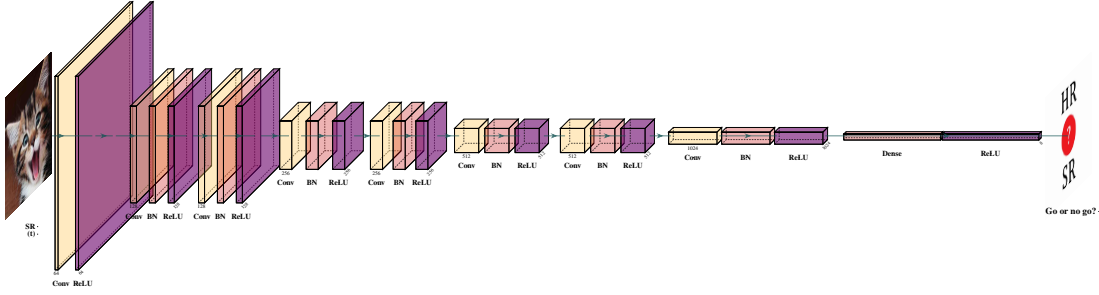


Figure 5: Discriminator Architecture from SRGAN (16). The discriminator uses Leaky ReLUs for computing its activations.

4.2 Loss functions

To evaluate the quality of an image, a commonly used loss function is MSE which aims to improve the PSNR of an image (28). While optimizing MSE during training improves PSNR and SSIM, these metrics may not capture fine details in the image leading to misrepresentation of perceptual quality and can cause the resulting video frames to be too smooth (29). In a series of experiments, it was found that even manually distorted images still had an MSE score comparable to the original image (30).

To address this, we use a four-fold (adversarial, perceptual, MSE and TV) loss. (19) introduced a new loss function called perceptual loss, which relies on features extracted from a pre-trained VGG network instead of low-level pixel-wise error measures. Per (16), we use adversarial loss along with content loss which focuses on perceptual similarity instead of similarity in pixel space to limit model “fantasy”. Further, we use a de-noising function called TV loss (19). We weigh these losses together as a final evaluation standard for training iSeeBetter.

We define our loss function for each frame as follows. The total loss of a sample is the average of all frames.

$$\begin{aligned}
 Loss_{G_{\theta_G}}(t) = & \alpha \times MSE(I_t^{est}, I_t^{HR}) \\
 & - \beta \times \log(D_{\theta_D}(I_t^{est})) \\
 & + \gamma \times PercepLoss(I_t^{est}, I_t^{HR}) \\
 & + \delta \times TVLoss(I_t^{est}, I_t^{HR})
 \end{aligned} \tag{1}$$

$$Loss_{D_{\theta_D}}(t) = 1 - D_{\theta_D}(I_t^{HR}) + D_{\theta_D}(I_t^{est}) \tag{2}$$

5 Results

To train the model, we used the Amazon EC2 P3.2xLarge instance with an NVIDIA Tesla V100 GPU with 16GB VRAM, 8 vCPUs and 61GB of host memory. We used the hyperparameters from RBPN and SRGAN. Table 2 and 3 compare iSeeBetter with six state-of-the-art VSR algorithms.

Dataset	Clip Name	Flow	Bicubic	DBPN (2)	B ₁₂₃ + T (31)	DRDVR (8)	FRVSR (9)	VSR-DUF (11)	RBP/6-PF (10)	iSeeBetter
Vid4	Calendar	1.14	19.82/0.554	22.19/0.714	21.66/0.704	22.18/0.746	-	24.09/0.813	23.99/0.807	24.13/0.817
	City	1.63	24.93/0.586	26.01/0.684	26.45/0.720	26.98/0.755	-	28.26/0.833	27.73/0.803	28.34/0.841
	Foliage	1.48	23.42/0.575	24.67/0.662	24.98/0.698	25.42/0.720	-	26.38/0.771	26.22/0.757	26.27/0.773
	Walk	1.44	26.03/0.802	28.61/0.870	28.26/0.859	28.92/0.875	-	30.50/0.912	30.70/0.909	30.68/0.908
Vimeo90K	Fast Motion	8.30	34.05/0.902	37.46/0.944	-	-	-	37.49/0.949	40.03/0.960	40.17/0.971
Average		1.42	23.53/0.629	25.37/0.737	25.34/0.745	25.88/0.774	26.69/0.822	27.31/0.832	27.12/0.818	27.36/0.835

Table 2. PSNR/SSIM evaluation of state-of-the-art VSR algorithms using Vid4 and Vimeo90K for 4×. Bold numbers indicate best performance.







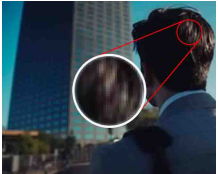


Dataset	Clip Name	VSR-DUF (11)	iSeeBetter	Ground Truth
Vid4	Calendar			
				
				

Table 3. Visually inspecting examples from Vid4, SPMCS and Vimeo-90k comparing RBP and iSeeBetter. We chose VSR-DUF for comparison because it was the state-of-the-art at the time of publication. Top row: fine-grained textual features that help with readability; middle row: intricate high-frequency image details; bottom row: camera panning motion.

6 Conclusion

We proposed iSeeBetter, a novel spatio-temporal approach to VSR that uses recurrent-generative back-projection networks. iSeeBetter couples the virtues of RBP and SRGAN. RBP enables iSeeBetter to generate superior SR images by combining spatial and temporal information from the input and neighboring frames. In addition, SRGAN’s discriminator architecture fosters generation of photo-realistic frames. We used a four-fold loss function that helps emphasize perceptual quality. Further, we proposed a new evaluation protocol for video SR by collating diverse datasets. With extensive experiments, we assessed the role played by various design choices in the ultimate performance of iSeeBetter, and demonstrate that on a vast majority of test video sequences, iSeeBetter shows better results compared to the state-of-the-art VSR systems.

7 Error Analysis

Table 4 takes a deeper look into the Walk scene from Vid4 where iSeeBetter showed room for improvement. We noticed that the scene had a very different composition compared to other Vid4 scenes - it consists of 10+ faces which is in stark contrast to the other scenes which mostly consist of non-human imagery.

RBP/6-PF (11)	iSeeBetter	Ground Truth
		

Table 4. Investigating the characteristics of the Walk scene from Vid4 to understand what is leading RBP to perform better than iSeeBetter.

8 Future Work

To improve iSeeBetter, a couple of ideas come to mind. First, train iSeeBetter with more faces to improve performance in scenes containing humans. This is especially important if the intended application is VSR for human-centric scenes such as for

high-resolution TVs. Second, in visual imagery, most of the attention is on the foreground which typically includes humans, objects etc. To improve perceptual quality, we can segment the foreground and background, and make iSeeBetter perform "intelligent VSR" by adopting different policies for the foreground and background. Third, another way to further improve iSeeBetter would be to make it assign weights to the adjacent frames (for e.g., adjacent frames from a different scene can be weighed lower, compared to frames from the same scene) - à la the concept of attention in NLP, but applied to VSR.

References

- [1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.
- [2] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1664–1673, 2018.
- [3] M. Haris, M. R. Widyanto, and H. Nobuhara, "Inception learning super-resolution," *Applied optics*, vol. 56, no. 22, pp. 6043–6048, 2017.
- [4] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016.
- [5] E. Faramarzi, D. Rajan, and M. P. Christensen, "Unified blind method for multi-image super-resolution and single/multi-image blur deconvolution," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2101–2114, 2013.
- [6] D. C. Garcia, C. Dorea, and R. L. de Queiroz, "Super resolution for multiview images using depth information," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1249–1256, 2012.
- [7] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4778–4787, 2017.
- [8] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4472–4480, 2017.
- [9] M. S. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6626–6634, 2018.
- [10] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3897–3906, 2019.
- [11] Y. Jo, S. Wug Oh, J. Kang, and S. Joo Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3224–3232, 2018.
- [12] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.
- [13] O. Makansi, E. Ilg, and T. Brox, "End-to-end learning of video super-resolution with motion compensation," in *German conference on pattern recognition*, pp. 203–214, Springer, 2017.
- [14] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.
- [15] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *Journal of Visual Communication and Image Representation*, vol. 4, no. 4, pp. 324–335, 1993.
- [16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *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*, pp. 4681–4690, 2017.
- [17] H. Ren and X. Fang, "Recurrent back-projection network for video super-resolution," in *Final Project for MIT 6.819 Advances in Computer Vision*, pp. 1–6, 2018.
- [18] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

- [19] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European conference on computer vision*, pp. 694–711, Springer, 2016.
- [20] A. Dosovitskiy and T. Brox, “Generating images with perceptual similarity metrics based on deep networks,” in *Advances in neural information processing systems*, pp. 658–666, 2016.
- [21] J. Bruna Estrach, P. Sprechmann, and Y. LeCun, “Super-resolution with deep convolutional sufficient statistics,” 1 2016. 4th International Conference on Learning Representations, ICLR 2016 ; Conference date: 02-05-2016 Through 04-05-2016.
- [22] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video enhancement with task-oriented flow,” *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.
- [23] C. Liu and D. Sun, “A bayesian approach to adaptive video super resolution,” in *CVPR 2011*, pp. 209–216, IEEE, 2011.
- [24] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, “Video super-resolution with convolutional neural networks,” *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, 2016.
- [25] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, “Video super-resolution via deep draft-ensemble learning,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 531–539, 2015.
- [26] Y. Huang, W. Wang, and L. Wang, “Bidirectional recurrent convolutional networks for multi-frame super-resolution,” in *Advances in Neural Information Processing Systems*, pp. 235–243, 2015.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.
- [28] A. Hore and D. Ziou, “Image quality metrics: Psnr vs. ssim,” in *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369, IEEE, 2010.
- [29] M.-H. Cheng, N.-W. Lin, K.-S. Hwang, and J.-H. Jeng, “Fast video super-resolution using artificial neural networks,” in *2012 8th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP)*, pp. 1–4, IEEE, 2012.
- [30] Z. Wang and A. C. Bovik, “A universal image quality index,” *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002.
- [31] D. Liu, Z. Wang, Y. Fan, X. Liu, Z. Wang, S. Chang, and T. Huang, “Robust video super-resolution with learned temporal dynamics,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2507–2515, 2017.