GENERATIVE ADVERSARIAL NETWORKS AND PERCEPTUAL LOSSES FOR VIDEO SUPER-RESOLUTION

Alice Lucas[1], *Santiago Lopez-Tapia*[2], *Rafael Molina*[2], *Aggelos K. Katsaggelos*[1]

[1]Dept. of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA
[2]Computer Science and Artificial Intelligence Department, Universidad de Granada, Spain

ABSTRACT

Recent research on image super-resolution (SR) has shown that the use of perceptual losses such as feature-space loss functions and adversarial training can greatly improve the perceptual quality of the resulting SR output. In this paper, we extend the use of these perceptual-focused approaches for image SR to that of video SR. We design a 15-block residual neural network, VSRResNet, which is pre-trained on a the traditional mean-squared-error (MSE) loss and later fine-tuned with a feature-space loss function in an adversarial setting. We show that our proposed system, VSRResFeatGAN, produces super-resolved frames of much higher perceptual quality than those provided by the MSE-based model.

Index Terms— Video, Superresolution, Convolutional Neuronal Networks, Generative Adversarial Networks, Perceptual Loss Functions

1. INTRODUCTION

Video super-resolution, namely estimating high-resolution (HR) frames from low-resolution (LR) input sequences, has become one of the fundamental problems in image and video processing. With the popularity of high-definition display devices, such as High-definition television (HDTV), or even Ultra-high-definition television (UHDTV) on the market, there is an avid demand for transferring LR videos into HR videos so that they can be displayed on high resolution TV screens, void of artifacts and noise. Such monitors provide unprecedented details for both entertainment and scientific applications.

Super-resolution (SR) algorithms can be divided into two categories: model-based and learning-based algorithms. Model-based approaches [1, 2, 3, 4] explicitly model the Low Resolution (LR) frames as blurred, subsampled, and noisy versions of the corresponding High Resolution (HR) frames. On the other hand, learning-based algorithms use large training databases of HR and LR videos to learn to solve the video super-resolution problem. Recently, Deep Neural Networks have been the popular tool of choice for such learning-based approach. For example, Liao et al. [5] train a CNN to predict a high-resolution frame from an ensemble of SR solutions obtained from traditional reconstruction methods. Kappeler et al.[6] propose to train a Convolutional Neural Network (CNN) to take bicubically interpolated low-resolution frames as input and learn the direct mapping that reconstructs the central high-resolution frame. Li and Wang [7] show the benefits of residual learning in video super-resolution by predicting only the residuals between the high-frequency and low-frequency frame. Caballero et al. [8] jointly train a spatial transformer network and a super-resolution network to warp the videos frames to one another and benefit from sub-pixel information. Makansi et al. [9] and Tao et al. [10] have found that performing a joint upsampling and motion compensation (MC) operation increases the SR performance of the model. Liu et al. [11] propose to construct a temporal adaptive learning-based framework, in which a neural network is trained to learn the temporal dependency between input frames to increase the quality of the HR prediction.

In this paper, we introduce the use of a deep residual neural network as the basis of our VSR framework. Increasing the depth of the network allows for the training of a more powerful SR system and removes the need for applying the computationally expensive motion-compensation operation to the input LR frames. In addition, we develop a perceptual loss function specifically designed to improve the the quality of the super-resolved video.

The rest of the paper is organized as follows. In section 2.1, we describe the deep residual architecture used as the basis of our framework. Sections 2.2 and 2.3 justify the use of feature-based distance and adversarial loss functions as our proposed perceptual loss model for video super-resolution. We perform experiments to evaluate our VSR system, the VSRResFeatGAN, which are detailed in section 3. The results and discussions are provided in section 3.2.

This work was supported in part by the Sony 2016 Research Award Program Research Project. The work of SLT and RM was supported by the the Spanish Ministry of Economy and Competitiveness through project DPI2016-77869-C2-2-R and the Visiting Scholar program at the University of Granada. SLT received financial support through the Spanish FPU program.

978-1-4799-7061-2/18/$31.00 ©2018 IEEE

2. FROM VSRNET TO VSRRESNET WITH A COMBINED PERCEPTUAL LOSS

Let us consider a set of high resolution and low resolution video sequence pairs, T. We denote by x the images in the high resolution video sequences, y is used to denote the low resolution, interpolated images in a window around x. Our goal is to learn a super-resolving network, $\mathrm{f}_\theta(.)$, which takes y as input and predicts x.

Our proposed architecture is based on the VSRNet model described in [6], which consists of three convolutional layers that learn a mapping from the input low-resolution motion compensated frames to the super-resolved central frame x. We propose major improvements over the framework originally proposed by [6] We train a much deeper residual architecture which outputs high-quality reconstructed SR frames and whose input images are not motion compensated. Furthermore, we append a combined perceptual loss function to the traditional pixel-wise Mean-Squared-Error (MSE)

$$\mathrm{L}_{\mathrm{RMS}}(\mathrm{x}, \mathrm{f}_\theta(\mathrm{y})) = \|\mathrm{x} - \mathrm{f}_\theta(\mathrm{y})\|_2^2 \qquad (1)$$

used in [6], which we show significantly increases the overall quality of the super-resolved image.

2.1. Increasing the capacity with the use of residual blocks

Intuitively, better super-resolution solutions may be obtained by increasing the representation power of our video super-resolution CNN, as an increase in the representation capability of our model results in richer final learned representation of the latent HR frame from the input LR frames. Indeed, current state-of-the-art NN-based architectures for image SR are based on very deep residual neural networks ([12, 13]). This poses the question of whether such increase in performance can also be observed for the video super-resolution problem.

We modify the VSRNet model proposed in [6] by extending its architecture with fifteen residual blocks. Our proposed VSRResNet architecture is shown in Figure 1. The 5 × 5 and 7 × 7 convolutional kernels in VSRNet are replaced by 3 × 3 kernels in VSRResNet. In order to keep the spatial size of the feature maps constant across the neural network architecture, padding is used at each convolution step. We note here that no batch normalization ([14]) operation is used in the residual blocks, as our experiments did not find it to lead to an increase in performance.

2.2. Incorporating image statistics learned by discriminative CNNs into the loss function

When the first CNN-based models for super-resolution were introduced, the de facto standard loss function was the mean squared error between the proposed super-resolved image and the corresponding ground truth high-resolution image, measured in the pixel-space [6]. While using this loss offers advantages, such as easier optimization and favouring larger PSNR values, it has been shown to fail to correlate with the Human Visual System (HVS) characteristics [15]. Indeed, recent studies show that an image with a high PSNR can be significantly less photo-realistic than one with a lower PSNR (see, for example, [16, 17]).

Convolutional neural networks trained for discriminative tasks have been shown to be excellent feature extractors for tasks involving natural images. One successful deep CNN is the VGG-16 classification model proposed in [18]. Our first perceptual loss component utilizes the activations provided by the fourth convolutional layer of VGG-16 (which we denote as VGG(.)) as the space in which we compute the mean-squared-error between the predicted HR frame and the ground truth HR frame, that is, we introduce the term:

$$L_{\mathrm{VGG}}(\mathbf{x}, \mathbf{f}_\theta(\mathbf{y})) = \|\mathrm{VGG}(\mathrm{x}) - \mathrm{VGG}(\mathrm{f}_\theta(\mathrm{y}))\|_2^2 \qquad (2)$$

Using Eq. 2 as a component of our training loss function forces the super-resolved frame $\mathrm{f}_\theta(\mathrm{y})$ to be perceptually close to the ground truth x, where perceptual distance here is measured in the space of the image statistics learned by the VGG-16 network.

2.3. Learning powerful VSR image priors with GANs

The second loss component we introduce forces the network $\mathrm{f}_\theta(\mathrm{y})$ to cheat a discriminator network d_ϕ (which must be learned) capable to discriminate between real and fake SR images. This learning process can be realised with the use of Generative Adversarial Networks (GANs) [19].

In the adversarial setting defined by GANs, a HR frame, either the generated HR frame, noted here $g_\theta(\mathrm{y})$, or sampled from the training data T, is fed to the discriminator d_ϕ, which then outputs the probability that the input frame was real (i.e., $d_\phi(\mathrm{x}) = 1$) or produced by g_θ (i.e., $d_\phi(g_\theta(\mathrm{y}) = 0)$. In a GAN framework, the video SR network and the discriminator are trained by solving:

$$\max_{\theta,\phi} V(\theta, \phi) = \mathbb{E}_{\mathrm{x}}[\log d_\phi(\mathrm{x})] + \mathbb{E}_{\mathrm{y}}[(1 - \log(d_\phi(g_\theta(\mathrm{y}))] \qquad (3)$$

The ability of GANs to indirectly learn powerful image priors makes them particularly attractive in the context of producing images of high-perceptual quality, and has been proved useful in various imaging problems (e.g., [16, 20, 21]). Therefore, our video super-resolution combines the distance in feature-space and pixel-space with a GAN-based adversarial setting, which defines our final optimization problem to train VSRResFeatGAN:

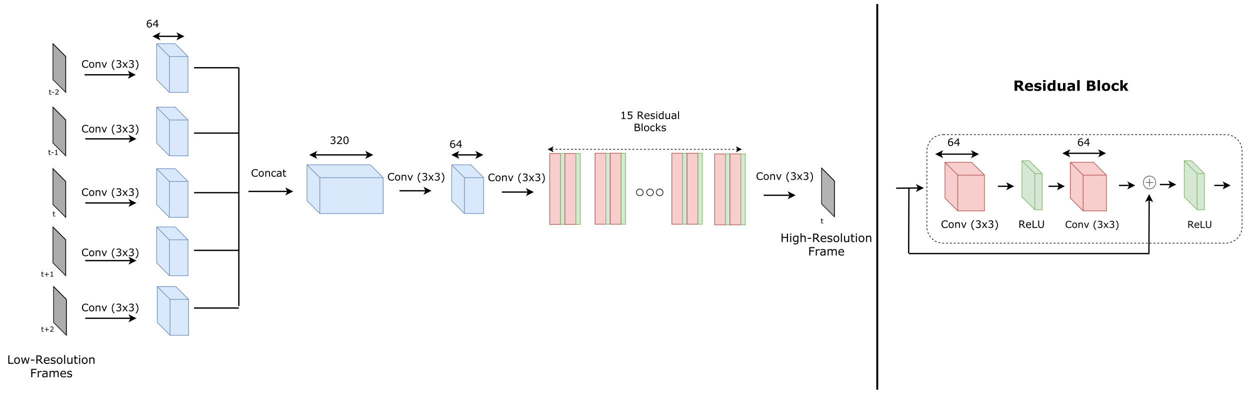

Fig. 1: The VSRResNet architecture. The first convolution extracts spatial information from each frame. The second convolution layer takes a concatenation of the extracted features across the different time steps. The following fifteen residual blocks learn the transformation that provides the final HR solution.

$$\hat{\theta} = \arg\min_{\theta} \sum_{(\mathrm{y},\mathrm{x})\in T} \Big[\alpha \|\mathrm{VGG}(\mathrm{x}) - \mathrm{VGG}(\mathrm{f}_\theta(\mathrm{y}))\|_2^2 + \beta \log(1 - d_{\hat{\phi}}(\mathrm{f}_\theta(\mathrm{y}))) + (1-\alpha-\beta)\|\mathrm{x} - \mathrm{f}_\theta(\mathrm{y})\|_2^2\Big] \tag{4}$$

$$\hat{\phi} = \max_{\phi}\{\, \mathbb{E}_{\mathrm{x}}[\log d_\phi(\mathrm{x})] + \mathbb{E}_{\mathrm{y}}[(1-\log(d_\phi(\mathrm{f}_{\hat{\theta}}(\mathrm{y})))]\,\} \tag{5}$$

where we fix the discriminator architecture to that used in [16], and $\alpha > 0$ and $\beta > 0$ with $\alpha + \beta < 1$ are hyper-parameters which control the contribution of each loss component and are determined experimentally.

3. EXPERIMENTAL RESULTS

3.1. Training Procedure

We use the 4K resolution *Myanmar* video dataset to train our model. Following Kappeler et al.'s [6] approach, we use 53 of the provided 59 scenes as our training and validation set, and use the remaining 6 scenes for testing. The low-resolution frames are obtained from the high-resolution frames by using MATLAB's *imresize* to downscale the frames by a downsampling factor of 3. To match the sizes of the input and output of our network, we bicubically interpolate the low resolution frames. Motion compensation on the input frames is commonly performed in NN-based settings for video SR, and its use has been investigated by ([6], [7], [9]). However, in order to save computational time during inference and to let the neural network learn sub-pixel motion to perform better upsampling, we choose not to apply MC on our input video sequences. The training dataset consists of patches pairs of 36 × 36 pixels formed by one high-resolution patch at time t and the corresponding low-resolution patches at time $t-2$, $t-1$, t, $t+1$, and $t+2$.

Prior to training VSRResFeatGAN with Equation 4, we pre-train our deep residual network VSRResNet using the Adam optimizer [22] for 100 epochs and the traditional pixel-space mean-squared-error as our loss function. We use an initial learning rate of 0.001, which is then divided by 10 at the 50th and 75th epoch of the training.

To utilize the perceptual components, we fine-tune the weights of VSRResNet for 15 epochs with the combined loss defined in Eq 4 to obtain VSRResFeatGAN. The learning rate γ for $\mathrm{f}_\theta(\mathrm{y})$ was fixed to 10^{-5}. The discriminator d_ϕ was trained with a learning rate of $\gamma_d = 10^{-6}$ and was pre-trained for 5 epochs before starting the fine-tuning of the generator. We use l_2-weight decay of strength 0.0001 on the parameters of both $\mathrm{f}_\theta(\cdot)$ and $d_\phi(\cdot)$.

The α and β parameters in Equation 4, which determine the contribution of each loss component, were set to $\alpha = 0.998$ and $\beta = 0.001$. These hyper-parameters were determined experimentally through the use of a small fraction of our training dataset.

3.2. Results and Discussion

Table 1 provides a quantitative comparison between VSRNet [6], VSRResNet, and VSRResFeatGAN in terms of PSNR and SSIM.

	Bicubic	**VSRNet** [6]	**VSRResNet**	**VSRResFeatGAN**
PSNR	31.59	34.42	35.86	28.53
SSIM	0.8957	0.9247	0.9478	0.9216

Table 1: Average PSNR and SSIM values for the Myanmar test sequences for a scale factor of three. The reported PSNR for the VSRNet [6] model was obtained by testing on a motion-compensated video sequence.

Our results in Table 1 show that training a deeper model (the VSRResNet model) results in a significant increase in

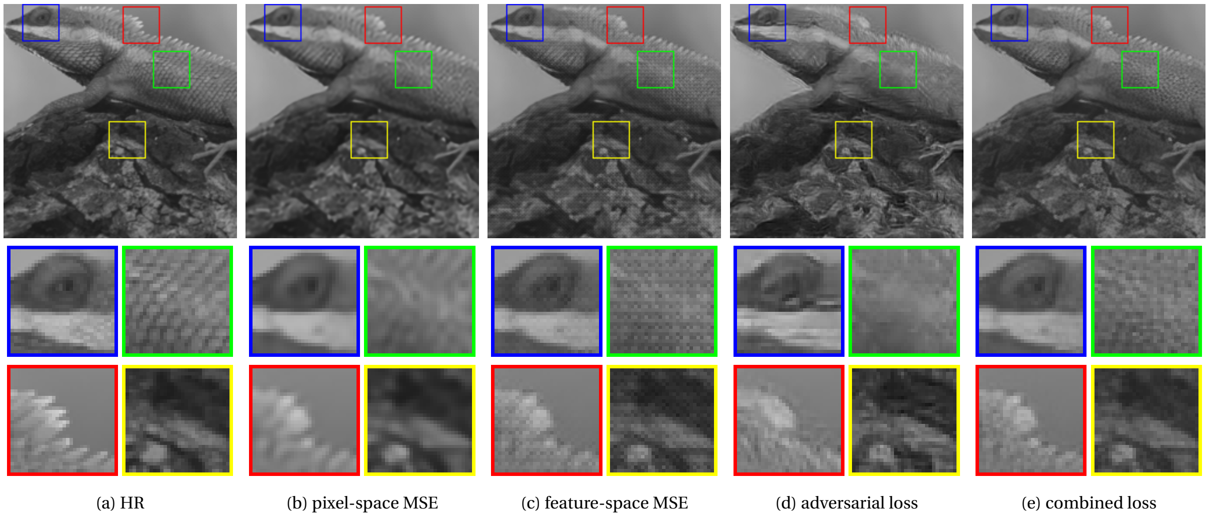

Fig. 2: Qualitative results of our video super-resolution system: (a) original HR frame, (b) VSRResNet, (c) model obtained with $\alpha = 0.99$ and $\beta = 0$, (d) model obtained by setting $\alpha = 0$ and $\beta = 0.001$ in Equation 4, and (e) VSRResFeatGAN with $\alpha = 0.998$ and $\beta = 0.001$ in Equation 4. The patches in the second and third rows correspond to the patches highlighted in the correspond first row image. Results for the full test dataset are available at this url: `https://goo.gl/wKe9Rx`

PSNR, observing a 1.6 dB increase relative to that of the VSRNet model. On the other hand, the use of losses that depart from the traditional pixel-wise difference such as the feature-space and the adversarial losses used in VSRResFeatGAN leads to a consequent decrease in the PSNR, resulting in a drop from 35.86 dB to 28.53 dB. This is consistent with the SR literature (e.g., [16]) which show that the use of perceptual loss disagrees with the PSNR measure, yet on the other hand still improves the image quality. Figure 2(e) supports this claim, as it is clear that the VSRResFeatGAN system increases the overall sharpness and perceptual quality of the output frame compared with that shown in Figure 2(b) which shows the results of training with pixel-wise MSE.

Looking closely at the patches in Figure 2 reveals that the feature-space and adversarial losses (See Figure 2(c) and Figure 2(d), respectively), when used individually, introduce a mild form of artifacts in the super-resolved frames. The output frames when training with $\alpha = 0.99$ and $\beta = 0$ in Eq. 4 (i.e., removing the contribution of the adversarial loss), for example, contain grid-like patterns (also referred to as checkerboard artifacts in the literature). The artifacts observed with the adversarial loss training (this time setting $\alpha = 0$ and $\beta = 0.001$ in Eq. 4) are of a different nature, ressembler ringing patterns at edges in the frames.

Using a combination of the pixel-space, feature-space, and adversarial losses as in Equation 4 leads to a more visually pleasing image, as observed in Figure 2(e). The grid-like pattern originally generated by the feature loss is removed in the predictions obtained by the combined loss. This can be explained by the fact that the generator eventually learns to "undo" the grid-like pattern generated by the feature-loss, as these patterns are easily detectable by the discriminator. Similarly, incorporating the adversarial loss tames the strong edge ringing effect seen in Figure 2(d) produced by this loss; which produces an overall perceptually pleasing HR frame. Therefore, we conclude that using these individual loss components as a combination instead of individually leads to results of much higher quality.

4. CONCLUSIONS

In this paper, we have introduced VSRResFeatGAN, a new learning-based system for video super-resolution. Our method introduces the use of a very deep residual neural network with high learning capacity, trained with a feature-based distance metric within a GAN framework. These losses allow our SR system to learn powerful on the super-resolved video frames. The experimental validation shows that this perceptual-based setup allows to estimate SR frames of greater quality than those obtained with the standard MSE estimate.

5. REFERENCES

[1] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Variational bayesian super resolution," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 984–999, 2011.

[2] S. P. Belekos, N. P. Galatsanos, and A. K. Katsaggelos, "Maximum a posteriori video super-resolution using a new multichannel image prior," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1451–1464, 2010.

[3] C. Liu and D. Sun, "On bayesian adaptive video super resolution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 346–360, 2014.

[4] A. K. Katsaggelos, R. Molina, and J. Mateos, "Super resolution of images and video," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 1, no. 1, pp. 1–134, 2007.

[5] 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.

[6] 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.

[7] D. Li and Z. Wang, "Video super-resolution via motion compensation and deep residual learning," *IEEE Transactions on Computational Imaging*, 2017.

[8] 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," *arXiv preprint arXiv:1611.05250*, 2016.

[9] 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.

[10] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," *arXiv preprint arXiv:1704.02738*, 2017.

[11] 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 Conference on Computer Vision and Pattern Recognition*, pp. 2507–2515, 2017.

[12] 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.

[13] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, vol. 1, p. 3, 2017.

[14] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, 2015.

[15] P. Gupta, P. Srivastava, S. Bharadwaj, and V. Bhateja, "A hvs based perceptual quality estimation measure for color images," *ACEEE International Journal on Signal & Image Processing (IJSIP)*, vol. 3, no. 1, pp. 63–68, 2012.

[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," *arXiv preprint arXiv:1609.04802*, 2016.

[17] 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.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[20] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2016.

[21] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised map inference for image super-resolution," *arXiv preprint arXiv:1610.04490*, 2016.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.