GAN-BASED VIDEO SUPER-RESOLUTION WITH DIRECT REGULARIZED INVERSION OF THE LOW-RESOLUTION FORMATION MODEL

Santiago Lopez-Tapia[(a)], Alice Lucas[(b)], Rafael Molina[(a)], Aggelos K. Katsaggelos[(b)]

a) Dpto. de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Spain
b) Dept. of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA

ABSTRACT

While high and ultra high definition displays are becoming popular, most of the available content has been acquired at much lower resolutions. In this work we propose to pseudo-invert with regularization the image formation model using GANs and perceptual losses. Our model, which does not require the use of motion compensation, utilizes explicitly the low resolution image formation model and additionally introduces two feature losses which are used to obtain perceptually improved high resolution images. The experimental validation shows that our approach outperforms current video super resolution learning based models.

Index Terms— Video, Super-resolution, Convolutional Neuronal Networks, Generative Adversarial Networks, Perceptual Loss Functions

1. INTRODUCTION

The image super-resolution (SR) problem consists of obtaining a high-resolution (HR) image from an input set of low-resolution (LR) images. In Video SR (VSR) the input and output are sequences of LR and HR video frames, respectively. While high and ultra high definition displays are becoming popular, most of the available content has been acquired at much lower resolutions. Because of this, the demand for methods to convert LR videos into HR ones has increased.

Current SR methods can be grouped into two categories: model-based and learning-based algorithms. The ones in the first category explicitly define the process (blurring, sub-sampling and noise adding) by which an LR image is obtained from the HR image or video sequence [1, 2, 3, 4]. However, this is not the case for learning-based algorithms. These methods use large training databases of HR and LR image/sequence pairs to learn to solve the super-resolution problem. In this regard, Convolutional Neural Networks (CNN) have become a popular tool for solving this learning problem. Liao et al. [5] trained a CNN to predict an HR frame from an ensemble of SR solutions obtained from traditional reconstruction methods. Li and Wang [6] show the benefits of residual learning in video SR by predicting only the residuals between the HR and LR frames. Caballero et al. [7] jointly train a spatial transformer network and an SR network to warp the videos frames to one another and benefit from sub-pixel information. Makansi et al. [8] and Tao et al. [9] have found that performing a joint upsampling and motion compensation (MC) increases the SR performance of the model. Liu et al. [10] 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. Kappeler et al. [11] propose to train a CNN which takes bicubically interpolated LR frames as input and learn the direct mapping that reconstructs the central HR frame. Following [11], we introduced in [12] a deeper residual network trained using feature and adversarial losses that increased the perceptual quality of the output.

Recently, a new learning based image SR approach has been introduced [13]. This approach estimates and explicitly uses the image formation model to learn the network. The blurring and downsampling process to obtain LR frames from HR ones is estimated and a Maximum a Posteriori (MAP) HR image estimation is approximated with the use of a Generative Adversarial Network (GAN).

In this work, we propose the combination of two networks to be optimized to solve the VSR problem. The first one, which adapts our VSR architecture [12] to the model proposed in [13], obtains very good mean squared error and structural similarity index values. The second one uses feature losses and an internal spatial smoothness constraint to improve the perceptual quality of the estimated HR sequences and overcomes some of the drawbacks that limit the application of GAN to upscaling factors greater than 2.

The rest of the paper is organized as follows. In section 2, we present our model for VSR and the CNN architecture used. In section 3, we detail and discuss our experiments

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.

with the proposed model. Finally, conclusions are drawn in section 4.

2. MODEL DESCRIPTION

In this paper, x is used to denote a high resolution (HR) image in a video sequence, y is used to denote its corresponding observed Low Resolution (LR) version and $\mathbf{y}$ refers to the low resolution (LR) images in a time window around x, that is, $\mathbf{y}$ contains $2K+1$ LR images. We seek to learn a network $\mathrm{f}_\theta(.)$ that given $\mathbf{y}$ predicts x. To train these networks, we are provided with a set of high and low resolution video sequence pairs.

In this work we assume that the image formation noise is negligible and absorbed by the downsampling process, that is,

$$y = Ax \tag{1}$$

where A is the degradation operator, the concatenation of a blur and downsampling operators. In this work, following the previous literature ([5, 6, 7, 8, 9, 10, 11, 12]), we assume that A corresponds to bicubic downsampling.

In the process of obtaining an HR image x from the LR video sequence $\mathbf{y}$ we adapt the approach in [13] to our VSR problem and consider the function

$$\mathrm{g}_\theta(\mathbf{y}) = (I - A^+A)\mathrm{f}_\theta(\mathbf{y}) + A^+y, \tag{2}$$

where A^+ denotes the Moore-Penrose pseudoinverse of A. Since $AA^+A = A$ and $A^+AA^+ = A^+$, and because the rows of A are independent $AA^+ = I$, we have

$$A\mathrm{g}_\theta(\mathbf{y}) = A(I - A^+A)\mathrm{f}_\theta(\mathbf{y}) + AA^+y = y \tag{3}$$

and so $\mathrm{g}_\theta(\mathbf{y})$ is an HR image which satisfies eq. (1).

Taking into account that the transformation $g_\theta(\mathbf{y})$ defines (from the distribution on $\mathbf{y}$) a probability distribution function $\mathrm{q}_\theta(.)$ on the set of HR images the Kullback-Leibler divergence between $\mathrm{q}_\theta(.)$ and the distribution of real images $p_X(.)$

$$\mathrm{KL}(q_\theta||p_X) = \int \mathrm{q}_\theta(x)\log\frac{q_\theta(x)}{p_X(x)}dx \tag{4}$$

is minimized using a GAN approach. This model has a maximum a posteriori approximation interpretation which will not be discussed here, see [13].

Together with the generative network $g_\theta(\mathbf{y})$, we learn a discriminative one $d_\phi(x)$ using the following two functions on ϕ and θ

$$\mathrm{L}(\phi;\theta) = -\mathbb{E}_{x\sim\mathrm{X}}[\log \mathrm{d}_\phi(x)] - \mathbb{E}_{\mathbf{y}\sim\mathbf{Y}}[\log(1-\mathrm{d}_\phi(\mathrm{g}_\theta(\mathbf{y})))] \tag{5}$$

$$\mathrm{L}(\theta;\phi) = -\mathbb{E}_{\mathbf{y}\sim\mathbf{Y}}[\log\frac{\mathrm{d}_\phi(\mathrm{g}_\theta(\mathbf{y}))}{1-\mathrm{d}_\phi(\mathrm{g}_\theta(\mathbf{y}))}]. \tag{6}$$

Iteratively, the algorithm updates ϕ by lowering $\mathrm{L}(\phi;\theta)$ while keeping θ fixed, and updates θ by lowering $\mathrm{L}(\theta;\phi)$ while keeping ϕ fixed.

The still image SR method which uses this approach was named AffGAN. For consistency, the video SR model, which adapts our VSRResGAN model, [12] to this approach will be henceforth referred to as VSRResAffGAN. VSRResAffGAN is a deep residual CNN that consists of 3 3×3 convolutional layers followed by a ReLU activation, 15 Residuals Blocks with no batch normalization and a final 3×3 convolutional layer. Padding is used at each convolution step in order to keep the spatial extent of the feature maps fixed across the network. However, instead of using as input the bicubic interpolated frames as in [12], we used $A^+\mathbf{y}$, since this input is better suited for other degradation operators besides bicubic downsampling. Finally, we fix the discriminator architecture d_ϕ to the one used in [14].

As we will show in the experimental section, our VSRResAffGAN model performs well for scale factor 2; however, it becomes unstable for larger scale factors such as 3 and 4, where the VSRResAffGAN failed to converge (see fig. 1). This is most likely caused by the discriminator's high power which easily distinguishes between real and generated frames (the generator has to produce 16 pixels for each pixel in the LR frame for a scale factor of 4).

In order to stabilize the training, the GAN must be regularized, see [15]. However, the mean squared error (MSE) between the real and predicted pixels proposed in [15] to regularize the GAN is not well suited for our model. Notice that it is redundant (the AffGAN framework takes into account the pixel likelihood), furthermore, this loss has been shown to fail to correlate with the Human Visual System (HVS) characteristics [16]. Instead of the MSE, we propose to incorporate to our loss function the feature based loss proposed in [12], which consists of the use of Charbonnier loss between two images u and v in a feature space, that is,

$$\gamma(u,v) = \sum_k\sum_i\sum_j\sqrt{(u_{k,i,j}-v_{k,i,j})^2+\epsilon^2}\,. \tag{7}$$

The features used correspond to the third and fourth convolution convolutional layers of VGG-16 (denoted by VGG(.)).

Finally, we have observed that the frames produced by the model in [12] show some noise artifacts (see Fig. 2(b)). Although our Affine GAN model shows significantly fewer of these artifacts in our experiments, we have introduced an internal spatial smoothness constraint to reduce them even further. Thus the new combined loss used to train our proposed generator network which replaces $\mathrm{L}(\theta;\phi)$ in Eq. (6) is:

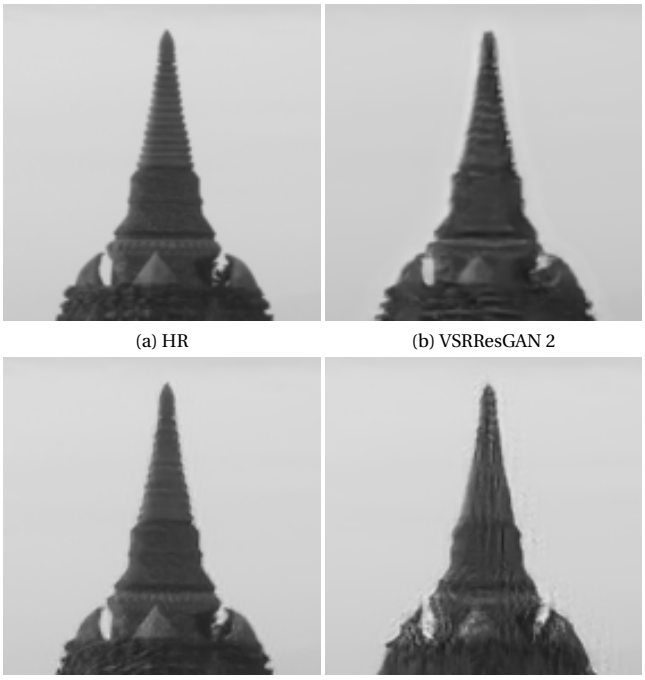
$$\mathrm{L}_{\mathrm{total}}(\theta;\phi) = \alpha\sum_{(x,\mathbf{y})\in T}\gamma(\mathrm{VGG(x)},\mathrm{VGG}(\mathrm{g}_\theta(\mathbf{y}))) \tag{8}$$
$$+(1-\alpha)[\mathbb{E}_{\mathbf{y}}[\log\frac{1-\mathrm{d}_\phi(\mathrm{g}_\theta(\mathbf{y}))}{\mathrm{d}_\phi(\mathrm{g}_\theta(\mathbf{y}))}]] + \beta\sum_{\mathbf{y}\in T}\parallel\mathbf{C}\mathrm{g}_\theta(\mathbf{y})\parallel_2^2$$

where α is in the range [0, 1], β is greater than zero, and $\mathbf{C}$ is the linear image transformation associated with the kernel

$$\mathbf{c} = \begin{bmatrix} 0 & -1/4 & 0 \\ -1/4 & 1 & -1/4 \\ 0 & -1/4 & 0 \end{bmatrix}, \tag{9}$$

.

We refer to this model as VSRResFeatAffGAN.



(a) HR (b) VSRResGAN 2

(c) VSRResAffGAN 2 (d) VSRResAffGAN 4

Fig. 1: Qualitative comparison between VSRResAffGAN and VSRResGAN[12]. We can see how our VSRResAffGAN is able to recover the frame for factor 2 while VSRResGAN fails producing a lot of artifacts and a blur frame.

3. EXPERIMENTAL RESULTS

Prior to training, it is necessary to calculate the A^+ operator. We implement this operator using a convolution operation with a 5 kernel followed by a subpixel shuffle layer [17]. Following the approach in [13], we optimize the weights of the convolutional layer by minimizing the following loss function with gradient descent, that is,

$$\hat{\omega} = \arg\min_{\omega} \mathbb{E}_x \|Ax - AA^+_\omega(Ax)\|_2^2 + \mathbb{E}_y \|A^+_\omega(y) - A^+_\omega(AA^+_\omega(y))\|_2^2, \quad (10)$$

where A is the degradation operator (in this case, bicubic downsampling) and A^+_ω the pseudo inverse. We stop the optimization when the loss value is under 10^{-7}. During the training of the VSR network, we keep ω fix.

The training dataset was constructed by extracting 10^6 patches of size 36×36 pixels from the Myanmar training sequences. From each HR patch at time t, we obtain the corresponding LR sequence of patches at time $t-2$, $t-1$, t, $t+1$, and $t+2$. To remove uninformative patches from our training dataset, patches with variance less than 0.0035 were not considered.

We first train the network for 100 epochs using the MSE loss ($\mathbb{E}_{x,\mathbf{y}}[\|x - g_\theta(\mathbf{y})\|^2]$) with the Adam optimizer [18]. The learning rate was set to 10^{-3} for the first 50 epochs and then divided by 10 at the 50th and 75th epoch. The weight decay parameter was set to 10^{-4}. We refer to this first model as VSRResAffNet.

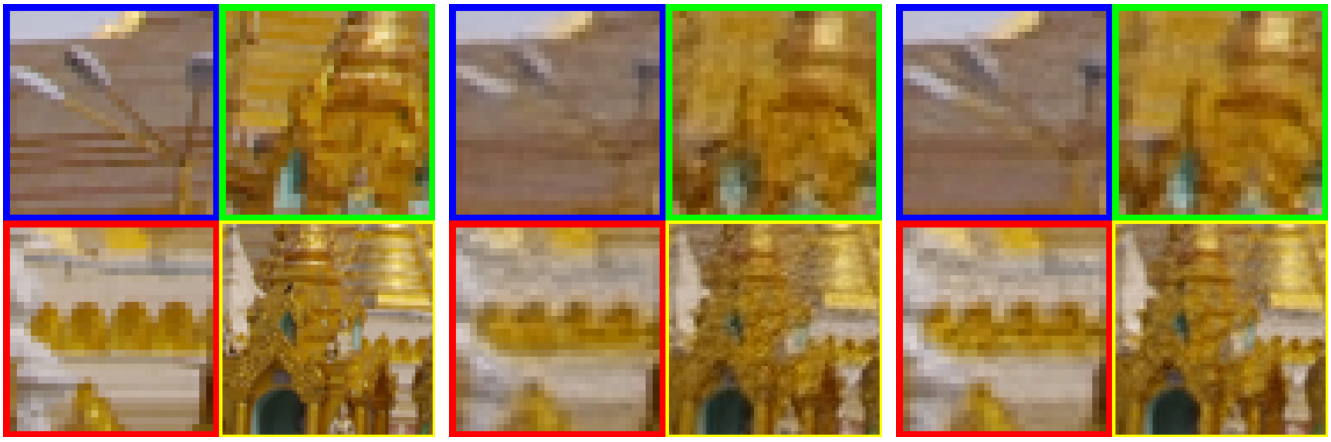
To determine the contribution of the affine projection, we first run an experiment with the adversarial loss only (i.e., no feature or smooth loss is used). With the generator architecture pre-trained using the VSRResAffNet, we fine-tuned the weights of the model for 30 epochs with the losses defined in Eqs. (5) and (6). The learning rate and weight decay for the discriminator is set to 10^{-4} and 10^{-3}, respectively. Both of the generator's learning rate and weight decay are set to 10^{-4}. We call the resulting model the VSRResAffGAN model. Fig. 1 (a-c) shows a comparison between the VSRResGAN [12] and our resulting VSRResAffGAN for factor 2. This figure clearly demonstrates that the affine MAP approximation is able to recover the frame with high fidelity and without creating artifacts, while the VSRResGAN [12] fails to do so. However, as seen in Fig. 1 (d), the VSRResAffGAN model fails to accurately recover the frame for factor 4. As we explained in the previous section, we argue that this is due to the instability of the GAN training, caused by the solutions of the Generator being too far away from the HR frames.

Let us see how the use of the proposed loss function in Eq. (8) alleviates this problem. With the generator architecture pre-trained using the VSRResAffNet model, we also fine-tuned the weights of the model with combined features losses defined in Eqs. (5) and (8) for 30 epochs. The learning rate and weight decay parameters of the discriminator and the generator are the same used in the training of the VSRResAffGAN. The optimal values of the α and β hyper-parameters were determined experimentally using a small fraction of our training dataset. As a result of this, we set these to 0.995 and 10.0, respectively. We note here that α is set to high value because the scale of the feature loss is several orders of magnitude smaller than that of the adversarial loss, not because it is more relevant. As we have already indicated, we refer to the resulting model as VSRResFeatAffGAN.

	VSRResNet PSNR/SSIM	VSRResFeatGAN PSNR/SSIM	VSRResAffNet PSNR/SSIM	VSRResFeatAffGAN PSNR/SSIM
2	40.58/0.9807	39.08/0.9744	**40.69/0.9811**	39.80/0.9787
3	35.97/0.9481	34.41/0.9312	**36.09/0.9487**	35.20/0.9406
4	32.85/0.9075	31.44/0.8694	**33.38/0.9077**	32.15/0.8866
	PercepDist	PercepDist	PercepDist	PercepDist
2	0.0122	0.0053	0.0118	**0.0052**
3	0.0523	0.0244	0.0496	**0.0234**
4	0.0904	0.0618	0.0844	**0.0612**

Table 1: Comparison with VSRResNet and VSRResFeatGAN [12] for Myanmar dataset.

Our VSRResAffNet and VSRResFeatAffGAN models are now compared on the test sequences of the Myammar dataset. Table 1 shows the performance of both models compared against the VSRResNet and VSRResFeatGAN net-



(a) HR (b) VSRResFeatGAN[12] (c) VSRResFeatAffGAN

Fig. 2: Qualitative results of our video super-resolution system for factor 3. Results for the full test dataset are available at this url: `https://goo.gl/wKe9Rx`

works, which do not explicitly use the observation model. As seen in the table, our approach quantitatively surpasses all the other models. Particularly, the table reveals a significant improvement in the PSNR metric obtained by the VSRResAffNet model, with an increase of more than 0.7dB for all scales. Similarly, we observe a sharp increase in SSIM for scale factor 4.

A careful examination of the produced frames reveals that the increase in PSNR and SSIM comes with the removal of high frequency components in the reconstructed frames. This is reflected in the Perceptual Distance [19] metric. For this metric, the VSRResFeatGAN model outperforms our VSRResAffNet model, with VSRResFeatAffGAN being the best one. Furthermore, the VSRResFeatAffGAN produces structures that are more similar to the ones in the original HR frame. Fig. 2 shows a comparison between the two perceptually best performing models where the above discussed events are particularly noticeable for scale factor 3. As seen in the figure, the dot-like noise has been removed in the images produced by the VSRResFeatAffGAN. Furthermore, details in the resulting frames are more defined and closer to the original HR frame, although this effect is more subtle. Notice also that the VSRResFeatAffGAN surpasses VSRResFeatGAN in terms of PSNR and SSIM by a significant amount.

4. CONCLUSIONS

We have introduced two new VSR models that explicitly utilize the LR image formation model: VSRResAffNet, trained with MSE only, and VSRResFeatAffGAN, trained with a combination of perceptual losses. The experiments show that VSRResAffNet outperforms current state of the art methods in terms of PSNR and SSIM. In terms of perceptual quality, VSRResFeatAffGAN surpasses VSRResAffNet and the state of the art while achieving a high PSNR and SSIM. In the future, both models will be extended to handle multiple degradation operators.

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

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

[8] 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.

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

[10] 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.

[11] 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.

[12] A. Lucas, S. Lopez Tapia, R. Molina, and A. K. Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," *IEEE Trans. on Image Processing*, 2019 (accepted for publication).

[13] C. Sonderby, J. Caballero, L. Theis, W. Shi, and F. Huszar, "Amortised MAP inference for image super-resolution," in *International Conference on Learning Representations*, 2017.

[14] 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.

[15] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[16] 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.

[17] 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," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1874–1883, 2016.

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

[19] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep networks as a perceptual metric," in *CVPR*, 2018.