Multiple-Degradation Video Super-Resolution with Direct Inversion of the Low-Resolution Formation Model

Santiago Lopez-Tapia *, Alice Lucas†, Rafael Molina*, Aggelos K. Katsaggelos†
*Depto. de Ciencias de la Computación e I.A., University of Granada, Granada, Spain
Email: {sltapia, rms}@decsai.ugr.es
†Dept. of Electrical Engineering and Computer Science Northwestern University Evanston, IL, USA
Email: alicelucas2015@u.northwestern.edu, aggk@eecs.northwestern.edu

***Abstract*—With the increase of popularity of high and ultra high definition displays, the need to improve the quality of content already obtained at much lower resolutions has grown. Since current video super-resolution methods are trained with a single degradation model (usually bicubic downsampling), they are not robust to mismatch between training and testing degradation models, in which case their performance deteriorates. In this work we propose a new Convolutional Neural Network for video super resolution which is robust to multiple degradation models and uses the pseudo-inverse image formation model as part of the network architecture during training. The experimental validation shows that our approach outperforms current state of the art methods.**

***Index Terms*—Video Super-resolution, convolutional neuronal networks, image formation**

I. Introduction

The problem of image super-resolution (SR) is to obtain a high-resolution (HR) image from an observed low-resolution (LR) image. The high to low image formation model can be written as:

$$y = D(x \otimes k) + \epsilon, \tag{1}$$

where y is the LR image, x is the HR image, ϵ is the noise, $x \otimes k$ represents the convolution of x with the blur kernel k and D is a downsampling operator (usually bicubic downsampling). In the case of video SR (VSR), y and x represent frames of the LR and HR sequences, respectively. Recently, the demand for high and ultra high definition displays has been increasing while most of the available content has been obtained at much lower resolutions. Consequently, the need for methods to improve the quality of these LR videos has also increased.

We can distinguish two approaches for image and video SR: model-based and learning-based. Approaches in the first category explicitly define and use the process (blurring, sub-sampling and noise adding) by which an LR image is obtained from the HR image or video sequence [1]–[4]. Learning-based algorithms use large training databases of HR and LR image/sequence pairs to learn to solve the super-resolution problem. Very frequently, they do not explicitly utilize the LR image formation model. Convolutional Neural Networks (CNN) have become a popular tool when using the learning approach. Liao et al. [5] trained a CNN to predict an HR frame from an ensemble of SR solutions obtained by traditional reconstruction methods. In [6], they show the benefits of residual learning for video SR by predicting only the residuals between HR and LR frames. Caballero et al. [7] train a spatial transformer network jointly with an SR network which registers video frames so the network benefits from sub-pixel information. Makansi et al. [8] and Tao et al. [9] found that performing the upsampling and motion compensation (MC) jointly increases the quality of the resulting SR frame. In [10], Liu et al. construct a temporal adaptive learning-based framework. In this framework 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 learns the direct mapping that reconstructs the central HR frame. In [12] we proposed a deeper residual network trained using feature and adversarial losses that significantly increased the perceptual quality of the output when compared with CNNs trained with Mean-Squared-Error based losses.

Although learning-based algorithms that use CNNs have, in general, produced better results than classical SR methods, LR sequences at test time are assumed to have been subjected to the same degradation used during the training phase. In other words, current methods are not robust to mismatch between training and testing degradation models, in which case their performance greatly deteriorates [13]. If they are trained for example with only bicubic downsampling, this lack of robustness against changes in the degradation significantly jeopardizes their application in practice.

In [14], Zhang et al. use Alternating Direction Method of Multipliers (ADMM) for image recovering problems with known linear degradation models, such as image deconvolution, blind image deconvolution, and Super-Resolution

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 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.

(SR). ADMM methods split the recovery problem into two subproblems: a regularized recovery one (subproblem A) and a denoising one (subproblem B). The authors propose to use a CNN for the denoising problem. This allows them to use the same network for multiple ill posed inverse imaging problems. At the same time, some works have been proposed to increase the performance and the flexibility of SR learning-based models by taking into account the image formation model. Sonderby et al. [15] proposed a new approach which estimates and explicitly uses the image formation model to learn the network. The blurring and downsampling process that obtains 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 Networks (GANs). Zhang et al. [16] propose for multiple-degradation SR the use of a CNN that has as input not only the LR image but also the PCA representation of the blur kernel used in the degradation process.

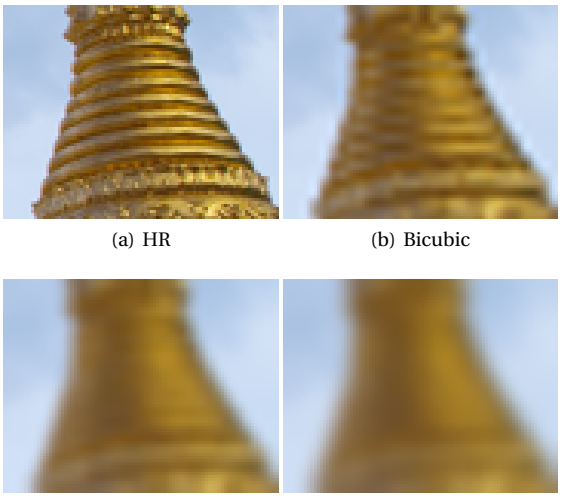
In this work, we propose a new model that adapts the approximation proposed in [15] to Multiple-Degradation Video Super-Resolution using the pseudo-inverse image formation model not only in the image formation model (as proposed in [15]), but also as an input to the network. We show that the proposed model outperforms by far current state of the art methods for bicubic degradation in terms of PSNR and SSIM metrics. Our experiments also show that the proposed model is far more robust to multiple degradations than current approaches.

The rest of the paper is organized as follows. In section II, we present our model for VSR and the CNN architecture used. In section III, we detail and discuss our experiments with the proposed model. Finally, conclusions are drawn in section IV.

II. Model description

For this work, as we have already indicated, we use x to denote an HR image in a video sequence and y its corresponding observed LR image. Furthermore, we use $\mathbf{y}$ to refer to the LR images in a time window around x. This means that $\mathbf{y}$ contains $2l+1$ LR images, that is, if x is indexed by t, $\mathbf{y}$ contains frames $t-l,...,t,...t+l$. In our case, $l=2$.

The process of obtaining an LR image from the HR one, as previously indicated in Section I, is usually modeled using Eq. 1. In this paper, we assume that the image formation noise is negligible ($\epsilon = 0$) and, following previous works in the literature, see, for instance, [16], we assume that the blur k is an isotropic Gaussian kernel and D represents bicubic downsampling. Although more complex blurs can also be considered these models are frequently assumed to be a good representation of the high to low degradation process [16]. The process of obtaining x from $\mathbf{y}$ is now much more challenging than when only bicubic downsampling is considered, which is the modelling used in previous works, such as, [5], [6], [8]–[12].



(a) HR (b) Bicubic

(c) $\sigma = 2.0$ (d) $\sigma = 3.0$

Fig. 1: Example of artifact introduced by the bicubic downsampling interpolation for a scaling factor of 3. (a) shows the original image, (b) corresponds to bicubically downsampling the image in (a), (c) and (d) show bicubically downsampled images which have previously being blurred with $\sigma = 2$ and $\sigma = 3$ Gaussian kernels. The downsampled images have been enlarged to the size of the original one using bicubic upsampling.

Let us now see how we can approach this multi-degradation model. First, we assume that the blur k is known. It can be approximated using any of the techniques proposed in [17]–[19]. Our experiments show that the estimated kernels are in practice accurate enough. Furthermore, estimating k and x at the same time does not work well in practice. A blind approach without a specially designed architecture has poor generalization ability. This goes against our objective which is to obtain a model robust against multiple degradation operators.

In order to make our network capable of dealing with multiple degradations, we try to separate its learning from the degradation as much as possible. To reach this goal, given D and k we define $A = Dk$ and adapt the approach in [15] to our Multiple-Degradation Video Super-Resolution problem by considering 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 the degradation 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) when $\epsilon = 0$.

Our approach still needs to address a very important problem: the parameters of the network θ should be found as to perform in a satisfactory way for all possible A. To solve this robustness issue, the network $\mathrm{f}_\theta(\cdot)$ needs to know not only the LR observation $\mathbf{y}$ but also the degradation A (notice that this is not necessary when only one degradation

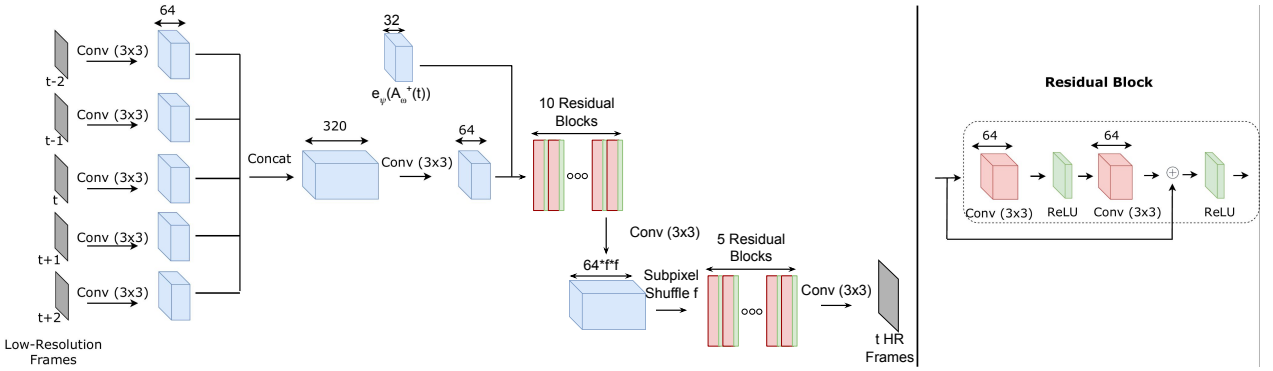


Fig. 2: The MD-AVSR architecture based on VSRResNet [12]. The network consists of a series of convolution operations with 64 kernels of size 3×3, applied on each input frame. The resulting feature maps are then concatenated together to obtain 320 feature maps. This is followed by two convolution operations and 15 residual blocks. Each residual block consists of two convolutional operations with 64 kernels of size 3×3, each followed by a ReLU layer. Following the definition of a residual block, the inputted feature maps are added to the output feature maps to obtain the final output of the residual block. Before the 10th Residual block, we use upscale the feature maps by a factor f using a subpixel shuffle layer [20].

is considered). We solve this problem by adding to the input of the network A^+y encoded by a network $e_\psi(\cdot)$. The goal of this information encoder $e_\psi(\cdot)$ is to extract significant information from the degradation to guide the SR process. Notice that other approaches, as the one used in [16], can be considered. $e_\psi(A^+y)$ consists of three convolutions of 3×3 and 32 filters with zero-padding and followed by ReLU activation. To ensure that the spatial size matches that of $\mathbf{y}$ we use a stride equal to the scaling factor at the last convolution. We train $e_\psi(\cdot)$ and $f_\theta(\cdot)$ simultaneously.

For the architecture of $f_\theta(\cdot)$, we adapt our VSRResNet model [12] to this approach. VSRResNet 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. Instead of using as input the bicubically upsampled frames as in [12], we decided to use a sub-pixel shuffle layer [20] to perform the upscaling. Together with the speed increase, we opted for this approach because bicubic upsampling over smoothes the images and introduces artifacts. Notice that these artifacts differ from one degradation operator to another (see Fig.1). This makes the learning process more difficult. Finally, we experimentally determined that the best way to incorporate $e_\psi(A^+y)$ is by concatenating these feature maps before the first residual block, see Fig.2 for details. We call the resulting model Multiple Degradation Affine Video Super Resolution (MD-AVSR).

Notice that in order to use this approach we need to calculate the A^+ operator prior to training. In [15] this operator is estimated using a convolution operation followed by a subpixel shuffle layer [20]. The network parameters w are estimated by minimizing the following loss function with gradient descent, that is,

$$\begin{aligned}\hat{\omega} =& \arg\min_{\omega} \mathbb{E}_x \|Ax - AA^+_\omega(Ax)\|_2^2 \\ &+ \mathbb{E}_y \|A^+_\omega(y) - A^+_\omega(AA^+_\omega(y))\|_2^2, \end{aligned} \tag{4}$$

where A^+_ω denotes the pseudo-inverse with ω network parameters.

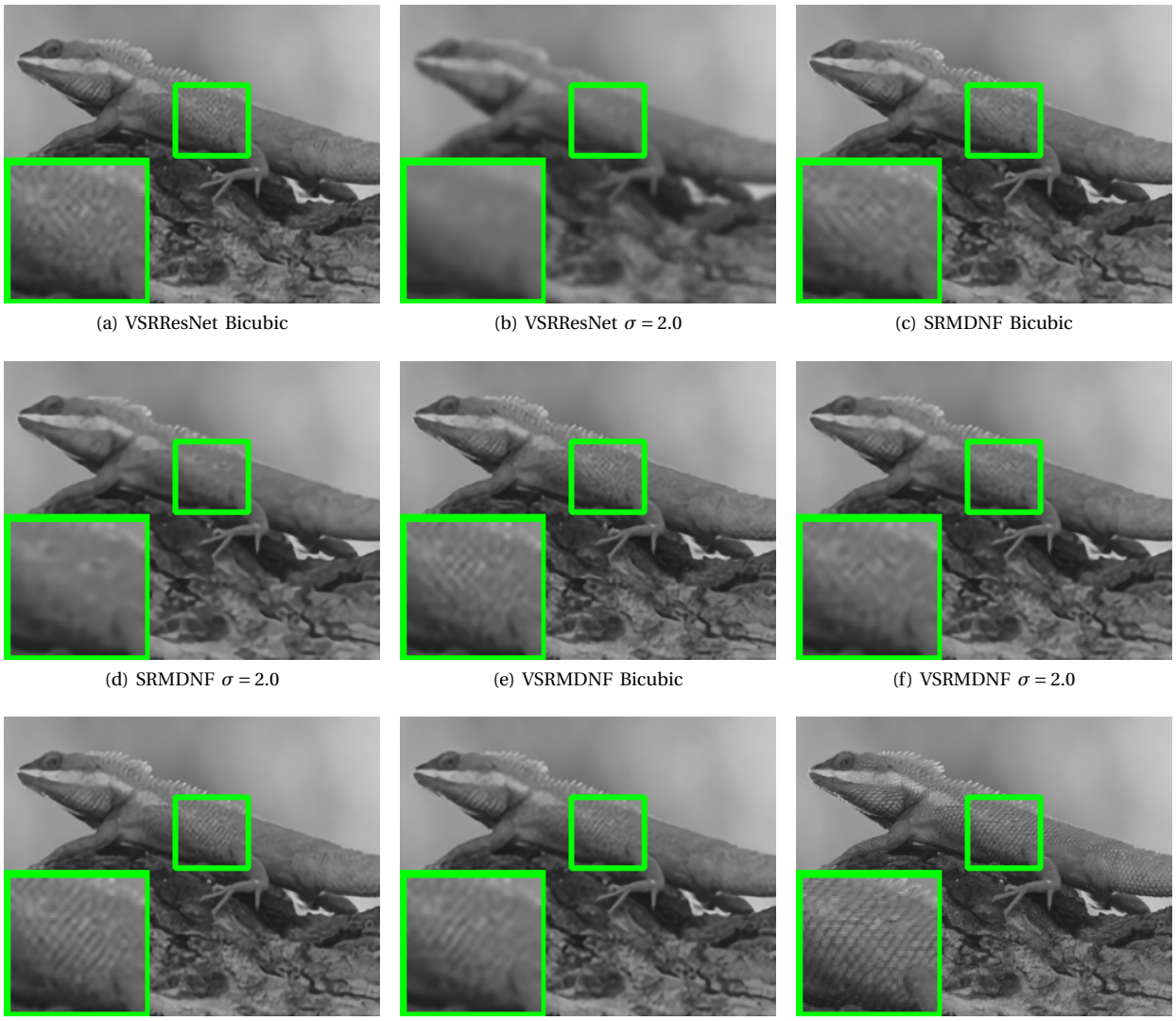
Instead of learning for each A a network that learns the corresponding $A^+_{\hat{w}}$, we have implemented a network that given A predicts its corresponding $\hat{\omega}$. The input to this network is the Principal Component Analysis (PCA) representation of the kernel k of A. The network is trained so the predicted $\hat{\omega}$ solves Eq. 4. We have not found any significant loss in performance by doing this instead of using Eq. 4 to calculate $\hat{\omega}$ for each A.

III. Experimental Results

The training dataset consists of 10^6 patches of size 48×48 pixels extracted 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 $t+2$. Patches with variance less than 0.0035 were removed due to being uninformative. Our models are compared on the test sequences of the Myammar dataset.

The discussed architecture was trained using the MSE loss ($\mathbb{E}_{x,\mathbf{y}}[\|x - g_\theta(\mathbf{y})\|^2]$) with the Adam optimizer [21] for 100 epochs. The learning rate was set to 10^{-3} for the first 50 epochs and then divided by 10 at the 50th and 75th epochs. The weight decay parameter was set to 10^{-5} for all the models.

In order to determine the contribution of the proposed architecture in conjunction with the pseudo-inverse input, we first train the network with only bicubic downsampling of factor 3 as the degradation. This model uses $f_\theta(\mathbf{y})$, not $f_\theta(\mathbf{y}, e_\psi(A^+y))$. In other words, it does not incorporate the information of the degradation inside the network that calculates the residual. However, it uses $g_\theta(\cdot)$. We refer to



(a) VSRResNet Bicubic (b) VSRResNet $\sigma = 2.0$ (c) SRMDNF Bicubic

(d) SRMDNF $\sigma = 2.0$ (e) VSRMDNF Bicubic (f) VSRMDNF $\sigma = 2.0$

(g) MD-AVSR Bicubic (h) MD-AVSR $\sigma = 2.0$ (i) HR

Fig. 3: Qualitative results of our video super-resolution system compared to other state of the art methods for bicubic downsampling and Gaussian blur with $\sigma = 2.0$ and bicubic downsampling. Notice how MD-AVSR is able to recover more details compared to others.

this model as AVSR. We call NoAVSR the model that uses the same architecture as AVSR (see Fig. 2) without the affine approach, that is, it minimizes $\mathbb{E}_{x,\mathbf{y}}\|x - f_\theta(\mathbf{y})\|^2$. To determine the contribution of the subpixel shuffle layer, we train an architecture similar to AVSR but using bicubic upsampling at the input instead of using the subpixel shuffle layer. We refer to this model as B-AVSR.

We call the model that uses $\mathrm{f}_\theta(\mathbf{y}, \mathrm{e}_\psi(A^+ y))$ MD-AVSR which is trained with multiple degradations. The degradations considered here are a combination of a Gaussian blur with different kernels k and bicubic downsampling of factor 3. For each training sample, we generated random Gaussian kernels with σ in the range [0.2, 3.0]. Then, we blurred the HR video sequences and applied bicubic downsampling to generate the LR samples. To determine the contribution of using $\mathrm{f}_\theta(\mathbf{y}, \mathrm{e}_\psi(A^+ y))$ instead of $\mathrm{f}_\theta(\mathbf{y})$, we trained $\mathrm{f}_\theta(\mathbf{y})$ following the same procedure used for MD-AVSR. We call this new model BMD-AVSR. Our model was also trained to incorporate the degradation operator following the SRMDNF approach in [16], that is, using PCA(k). We have experimentally determined that the optimal place to add this information is before the first residual block. We trained this network as we did with MD-AVSR. We call this model VSRMDNF.

The second column of table I shows the comparison of our models to current state of the art VSR methods for only bicubic downsampling testing, that is, given an HR video sequence, we obtain an LR one by only bicubically downsampling it. As we can see, AVSR shows a significant increase in PSNR of more than 0.5dB compared to NoAVSR

and VSRResNet, that is, the benefits of using $g_\theta(\cdot)$ instead of $f_\theta(\cdot)$ are obvious. Notice also that there is no different between NoAVSR and VSRResNet, which indicates that a deep residual network which does not use an affine projection does not benefit from learning the up-scaling operation, that is, there is no difference between using the bicubically upsampled input or a subpixel shuffle layer. However, as can be seen in table I (second column), B-AVSR performs significantly worse than AVSR, which suggests that artifacts introduced by bicubic interpolation harm the training of the affine network. Finally, MD-AVSR outperforms all other models on bicubic degradation even when it was not trained to specialize on it.

Table I also shows the results of our methods compared against current state of the art for multiple degradations. We can see that the proposed MD-AVSR surpasses all other models for all values of σ considered. As expected, AVSR performs poorly for degradations different from bicubic downsampling. This probes the importance of training with all the types of degradation that can be expected in real applications. Notice that BMD-AVSR suffers from a sharp decrease in performance which indicates the need to use the degradation information in $f_\theta(\cdot)$ if we expect to utilize the same network with multiple degradations. Finally, MD-AVSR outperforms VSRMDNF as much as AVSR does NoAVSR and VSRResNet, which shows that the benefits of using the pseudo-inverse are carried over to the multiple degradation setting. Figure 3 shows a qualitative comparison of VSRResNet, VSRMDNF and MD-AVSR.

	Bicubic PSNR/SSIM	$\sigma = 1.0$ PSNR/SSIM	$\sigma = 2.0$ PSNR/SSIM	$\sigma = 3.0$ PSNR/SSIM
VSRResNet	35.97/0.9481	33.39/0.9210	29.09/0.8365	27.18/0.7680
NoAVSR	35.92/0.9474	33.39/0.9156	29.16/0.8362	27.18/0.7678
B-AVSR	36.09/0.9487	33.41/0.9214	29.13/0.8374	27.20/0.7685
AVSR	36.35/0.9522	33.52/0.9279	29.23/0.8454	27.42/0.7836
IRCNN [22]	34.41/0.8937	34.44/0.8937	33.58/0.8937	29.92/0.8937
SRMDNF [16]	35.08/0.9299	35.14/0.9298	34.78/0.9224	33.20/0.8937
BMD-AVSR	34.97/0.9330	34.59/0.9275	34.27/0.9200	34.62/0.9270
VSRMDNF	35.95/0.9471	35.82/0.9439	35.28/0.9365	35.01/0.9319
MD-AVSR	**36.52/0.9525**	**36.27/0.9494**	**35.60/0.9406**	**35.22/0.9352**

TABLE I: Comparison of the proposed and state of the art models for Myanmar dataset for factor 3. σ refers to the Gaussian blur deviation used.

IV. Conclusions

We have introduced a Multiple-Degradation Video Super-Resolution model that explicitly utilizes the LR image formation model as an input to the network: MD-AVSR. The model is trained with MSE only. Experiments show that MD-AVSR outperforms current state of the art methods in terms of PSNR and SSIM for both multiple degradation and bicubic degradation only settings. In the future, we will further improve the perceptual quality of the SR frames by incorporating perceptual losses such as Adversarial and Feature losses.

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] G. Riegler, S. Schulter, M. RÃijther, and H. Bischof, "Conditioned regression models for non-blind single image super-resolution," in *IEEE International Conference on Computer Vision*, pp. 522–530, Dec 2015.

[14] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," *CoRR*, vol. abs/1704.03264, 2017.

[15] 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.

[16] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[17] S. Harmeling, S. Sra, M. Hirsch, and B. SchÃűlkopf, "Multiframe blind deconvolution, super-resolution, and saturation correction via incremental em," in *IEEE International Conference on Image Processing*, pp. 3313–3316, Sep. 2010.

[18] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," in *IEEE International Conference on Computer Vision*, pp. 945–952, Dec 2013.

[19] Q. Wang, X. Tang, and H. Shum, "Patch based blind image super resolution," in *IEEE International Conference on Computer Vision*, vol. 1, pp. 709–716 Vol. 1, Oct 2005.

[20] 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," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

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

[22] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep cnn denoiser prior for image restoration," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2808–2817, July 2017.