A Composite Discriminator for Generative Adversarial Network based Video Super-Resolution

Xijun Wang*, Alice Lucas*, Santiago Lopez-Tapia†,
Xinyi Wu*, Rafael Molina†, and Aggelos K. Katsaggelos*
*Dept. of Electrical Engineering and Computer Science Northwestern University Evanston, IL, USA
†Depto. de Ciencias de la Computación e I.A., University of Granada, Granada, Spain

***Abstract*—Generative Adversarial Networks (GANs) have been used for solving the video super-resolution problem. So far, video super-resolution GAN-based methods use the traditional GAN framework which consists of a single generator and a single discriminator that are trained against each other. In this work we propose a new framework which incorporates two collaborative discriminators whose aim is to jointly improve the quality of the reconstructed video sequence. While one discriminator concentrates on general properties of the images, the second one specializes on obtaining realistically reconstructed features, such as, edges. Experiments results demonstrate that the learned model outperforms current state of the art models and obtains super-resolved frames, with fine details, sharp edges, and fewer artifacts.**

***Index Terms*—Video Super-Resolution, Spatially Adaptive, Generative Adversarial Networks, the Composite Discriminator**

I. Introduction

One of the fundamental problems in image and video processing is Video Super-Resolution (VSR) whose aim is to recover High-Resolution (HR) video sequences from Low-Resolution (LR) ones. Recent Super-Resolution (SR) works seem to indicate that learning-based methods produce more realistic images than model-based methods [1], [2]. Deep Neural Networks have been widely chosen as the tool for such learning-based approaches [3]–[5].

Generative Adversarial Networks (GAN) [6], which are able to learn complex distributions from samples, have recently gained popularity in the VSR literature. Researchers utilize GAN-based training instead of classical (Mean-Squared-Error) MSE to encourage networks to favor solutions that look more like natural videos [7]–[10]. Most of these GAN-based approaches incorporate feature-based perceptual losses to generate frames of higher perceptual quality. In addition, recent GAN-based VSR works improve their performance by incorporating useful information, like image spatial [9] or temporal [7], [8], [10] information, during training. For example, the results reported in [9] show that incorporating spatial information into part of the training objective function (pixel and perceptual losses) helps to generate sharper frames with fewer artifacts and less noise. All these works include this information by either adding or improving loss terms in the training objective function [8], [9] or by improving the generator network [7], [8], [10]. No published results in VSR have tried to incorporate this information by making the discrminator much more powerful. GAN-based VSR methods have so far utilized the traditional two-player GAN framework proposed in [6], that is, an one generator and one discriminator min-max game. Some recent GAN studies [11], [12] have proposed alternative GAN frameworks and shown that these models are able to effectively tackle many modeling limitations of traditional GANs.

In this work we propose a new GAN framework to tackle VSR problems. A composite discriminator is used during training. The model learns useful spatial information and uses it to produce improved generated frames with more accurate edge reconstruction and visually pleasing quality. The convergence of the proposed approach is established. By training with this new GAN framework, we outperform current state-of-art methods [7], [9], [10] which are trained using the traditional GAN framework with no additional spatial information included in the discriminator. The main contribution of this paper is the use of two discriminators with complementary objectives. The second discriminator could potentially enforce other important attributes of an image, other than edges, such as, attributes pertaining to the texture of an image.

The rest of the paper is organized as follows. In section II, we describe our new GAN model with two collaborative discriminators and how it takes useful spatial information into account to generate better super-resolved video sequences. We also provide a proof of convergence of the proposed approach and an intuitive justification of the collaborative model. We then describe the pixel and feature losses used to train. In section III we report on experiments to evaluate our proposed model for VSR, and show that we are more successful at generating more accurate edges and reducing artifacts and noise than models trained with the traditional GAN framework with a single discriminator. We draw our conclusions in section IV.

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.

II. The proposed approach

The model proposed in [7], [10] sometimes produces SR frames with blurred edges and noise in high-frequency areas. A limitation of this model is that the spatial activity of image regions is not specifically taken into account during training. It is, however, clear that edge information plays a very important role in the quality of the reconstruction and that edge regions are more difficult to super-resolve than flat-regions. We propose the use of a composite discriminator which includes a novel edge sharpness enforcing collaborative discriminator to obtain realistic edges. Our model makes use of high frequency information extracted from frames in our training datasets and forces the generator to specifically take into account edge (high spatial activity) areas. By doing so, the generator is forced to produce crisper edges and fewer artifacts.

A. A composite discriminator

GANs [6] learn to generate samples from a specific data distribution through an adversarial training procedure. In the traditional GAN approach for image generation, a *generator* network learns to generate an image given a latent random vector z at its input. The learning of the generator is guided by an auxiliary network, a *discriminator*, which is simultaneously trained to distinguish between images generated by the generator and images from the training dataset. Given a generator $G(\mathrm{z})$, with latent variable z to be defined later, the discriminator is trained to distinguish between real and fake images, i.e., it outputs $D(\mathrm{x}) = 1$ when x is sampled from the training dataset of natural images and $D(G(\mathrm{z})) = 0$ when the images are produced by the generator. On the other hand, the generator is trained to make the discriminator believe that its generated images $G(\mathrm{z})$ are real, i.e., trained to assign to the discriminator output a probability $D(G(\mathrm{z})) = 1$. As a result of this adversarial training, the generator eventually converges to a solution which the discriminator fails to identify as "fake", which generally implies successful learning of the image manifold by the generator.

In [7], [10] we propose the use of the powerful generative property of GANs in VSR. Using GAN-based instead of MSE-based training enables the models to obtain frames of much higher perceptual quality. The original GAN setting was modified by inputting the sequence of input low-resolution frames Y to the generator instead of a random vector z. This is similar to the use of GANs in still image super-resolution [13], where a single LR image is provided as input to the generator. The generator is adversarially trained to super-resolve the input LR frames in a way that the discriminator cannot distinguish between the reconstructed HR frames, $\hat{\mathrm{x}} = G(\mathrm{Y})$ and real HR images. The GAN formulation first introduced in [6] was adapted to VSR by solving:

$$\min_{\mu}\max_{\acute{A}} \mathrm{L}_{\mathrm{GAN}}(\acute{A},\mu) = \mathsf{E}_{\mathrm{x}}[\log D_{\acute{A}}(\mathrm{x})] + \mathsf{E}_{\mathrm{Y}}[\log(1 - D_{\acute{A}}(G_{\mu}(\mathrm{Y})))], \quad (1)$$

where x is the center HR frame of dimensions $N \times N$, Y is a short sequence of LR input frames around its LR version y, each of dimensions $N \times N$ (notice that the LR images are bicubically upsampled), $D_{\acute{A}}$ is the discriminator network with trainable parameters $\acute{A}$ and G_{μ} is the generator network with trainable parameters μ, where here these parameters correspond to the learnable convolutional kernels of our networks.

The above model is the basis of all the so far based GAN formulation, but a close look at the optimization function indicates that the better discrimination could be achieved by using a model of the form

$$\min_{\mu}\max_{\acute{A},\acute{A}'} \mathrm{L}_{\mathrm{GAN}}(\acute{A},\acute{A}',\mu) = \mathsf{E}_{\mathrm{x}\sim \mathrm{p}_{\mathrm{x}}(\mathrm{x})}[\log(D_{\grave{1}\acute{A}}(\mathrm{x}) D_{2\acute{A}'}^{1-}(\mathrm{x}))] + \mathsf{E}_{\mathrm{Y}\sim \mathrm{p}_{\mathrm{Y}}(\mathrm{Y})}[\log((1 - D_{1\acute{A}}(G_{\mu}(\mathrm{Y})))^{,} (1 - D_{2\acute{A}'}(G_{\mu}(\mathrm{Y})))^{1-,})], \quad (2)$$

where two different discriminators $D_{1\acute{A}}$ and $D_{2\acute{A}'}$ with parameters $\acute{A}$ and $\acute{A}'$, respectively, are used and $0 < , < 1$.

Notice that following the approach in [6], it can be easily shown that fixing μ

$$\max_{\acute{A},\acute{A}'} \mathrm{L}_{\mathrm{GAN}}(\acute{A},\acute{A}',\mu) = \mathsf{E}_{\mathrm{x}\sim \mathrm{p}_{\mathrm{x}}(\mathrm{x})}\left[\log \frac{\mathrm{p}_{\mathrm{x}}(\mathrm{x})}{\mathrm{p}_{\mathrm{x}}(\mathrm{x}) + \mathrm{p}_{\mathrm{g}}(\mathrm{x})}\right]^{,} + \mathsf{E}_{\mathrm{x}\sim \mathrm{p}_{\mathrm{g}}(\mathrm{x})}\left[\log \frac{\mathrm{p}_{\mathrm{g}}(\mathrm{x})}{\mathrm{p}_{\mathrm{x}}(\mathrm{x}) + \mathrm{p}_{\mathrm{g}}(\mathrm{x})}\right]^{,}, \quad (3)$$

where $\mathrm{p}_g(\mathrm{x})$ is the probability distribution induced on x by Y. Furthermore, following the approach in [6] it can be shown that

$$D_{1\acute{A}}(\mathrm{x}) = \mathrm{D}_{2\acute{A}'}(\mathrm{x}) = \frac{\mathrm{p}_{\mathrm{x}}(\mathrm{x})}{\mathrm{p}_{\mathrm{x}}(\mathrm{x}) + \mathrm{p}_{\mathrm{g}}(\mathrm{x})}. \quad (4)$$

and also that the global minimum on μ in Eq. (2) is achieved when and only when $\mathrm{p}_g(\cdot) = \mathrm{p}_{\mathrm{x}}(\cdot)$. However, the network architecture of $D_{1\acute{A}}(\mathrm{x})$ will prevent it from always satisfying Eq. (4). For instance, this discriminator may concentrate on detecting important image properties but may not be capable of detecting all of them. This is similar to the case when prior models are used to restore images. For instance, horizontal filters are used to regularize horizontal differences but these filers are insensitive to vertical ones. In GAN terminology, a discriminator may be good at detecting certain fake note properties (or even be good on some notes) but not all of them. In this paper we approach the possible limited capabilities of $D_{1\acute{A}}$ by introducing a second discriminator that, in our case, concentrates on the quality of high frequency areas and, in particular, on producing realistic edges. To do so we redefine $D_{2\acute{A}'}(\cdot) = D_{2\acute{A}'}(W\cdot)$, where W denotes a high pass filter. Notice that other definitions of $D_{2\acute{A}'}(\cdot)$ are possible, but we will concentrate here on recovering high spatial activity areas.

Let us now provide a graphical description of our collaborative model. In Fig. 1, the W operator is a spatial information extractor. We use the method illustrated in [14] [9] to define the W operator, which is consistent with the masking property of the human visual system, according to which noise is visible in flat regions but not visible at edges.

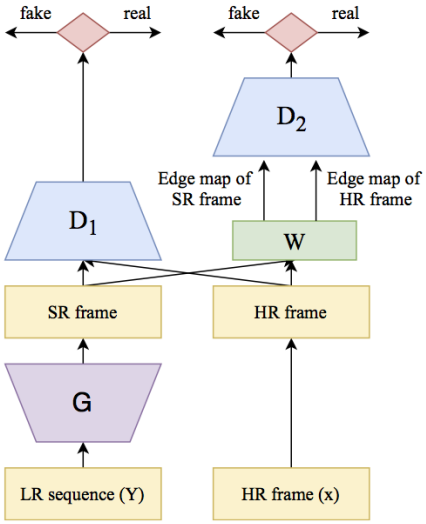

Fig. 1: The proposed model.

The output of the W operator is a weighted image (edge focusing map), where pixels in areas of high spatial activity, like edge regions, have much larger values than those in flat regions (in this work, all the pixel values are normalized to the range [0,1] during training). In the rest of the paper, the weighted image generated after the application of operator W will be simply referred to as the edge map.

We adopt the VSRResNet architecture proposed in [7], [10] as our generator. The architecture is shown in Fig. 2. It is based on 15 residual blocks, each block containing two convolutional layers with kernels of size 3 by 3, with a Rectified Linear Unit (ReLU) activation function after each convolution step.

We adopt the same architecture for both discriminators $D_{1Á}$ and $D_{2Á'}$ (used in [10]), shown in Fig. 2. The network is composed of three convolution layers followed by a fully connected layer and a sigmoid operation. However, they are provided with different inputs. The input to discriminator $D_{1Á}$ are super resolved and HR frames, while the input to discriminator $D_{2Á'}$ are the corresponding edge maps of the super resolved and HR frames. Besides of fooling the original discriminator $D_{1Á}$, the generator has also to fool the discriminator $D_{2Á'}$, so the edge map of the generated super resolved frame has to be realistic, close to the edge map of its corresponding HR frame. As a result, generated frames have more accurate edges with less noise and fewer artifacts and both discriminators collaborate to obtain better images. We name our proposed model the Collaborative Discriminator GAN (CoDiGAN), when applied to VSR we will denote it by VSRCoDiGAN.

B. Pixel and Feature Losses

To regularize undesired artifacts that may escape the collaborative model, we use two distance-based regularizers, defined in pixel and feature spaces, respectively.

Let us consider the Charbonnier loss, defined as

$$°(u,v) = \sum_k \sum_i \sum_j \sqrt{(u_{k,i,j} - v_{k,i,j})^2 + ²^2}, \quad (5)$$

where u and v are multichannel images with elements $u_{k,i,j}$ and $v_{k,i,j}$, respectively, where k denotes channel (for instance, $k=1$ for a gray-scale image and $k=1,2,3$ for a color image), i,j denotes pixel location and $²>0$. The pixel-wise loss only depends on low-level pixel information, and it is defined as the Charbonnier loss of the difference of two frames in pixel space, that is, $\sum_{(x,Y)\in T} °(x, G_µ(Y))$, where x and Y are sampled from the training dataset T. The perceptual loss in feature space computes differences between high-level image feature representations extracted from pre-trained convolutional neural networks. In this paper, we choose our feature space to be the representation space obtained from extracting the feature maps from the third and fourth convolution layer of the VGG network defined in [15], denoted as $VGG(\cdot)$ in this paper. Therefore, the feature loss is defined as $\sum_{(x,Y)\in T} °(VGG(x), VGG(G_µ(Y)))$. In the next section, we show the final loss for training the generator.

C. Final Loss for Generator

Combining the losses defined in the previous sections, our generator has to effectively minimize adversarial losses together with pixel and feature losses, thus our final loss function becomes:

$$\begin{aligned} L_{\text{final}}(µ) = ®_1 \sum_Y -\log D_{1Á}(G_µ(Y)) - \log D_{2Á'}(W(G_µ(Y))) \\ + ®_2 \sum_{(x,Y)\in T} °(x, G_µ(Y)) \\ + (1-®_1-®_2) \sum_{(x,Y)\in T} °(VGG(x),\ VGG(G_µ(Y))), \end{aligned} \quad (6)$$

where $0<®_1, ®_2$ and $®_1+®_2<1$, We have fixed ¸ to 1/2. In the next section, we show that this model improves the quality of the super resolved video.

III. Experiments

A. Training and Parameters

We synthesized the training dataset of HR/LR-sequence pairs from the Myanmar video sequence. Our training dataset consists of nearly 1 million pairs, where each sample in the training dataset is composed of five extracted 36×36 LR patches at times $t-2$, $t-1$, t, $t+1$, and $t+2$, and its corresponding 36×36 HR patch at time t. The LR frames were computed using bicubic downsampling followed by bicubic upsampling to bring them to the same spatial extent as the original HR patch.

To ensure convergence of generator and discriminator loss functions, it is critical for the generator network to start at a reasonable $µ$ at the beginning of the training [10]. Thus, prior to beginning the adversarial training, we first trained the generator network for 100 epochs with the traditional pixel based MSE loss using the ADAM [16] optimizer and a batch size of 64. For this pre-training, the initial learning

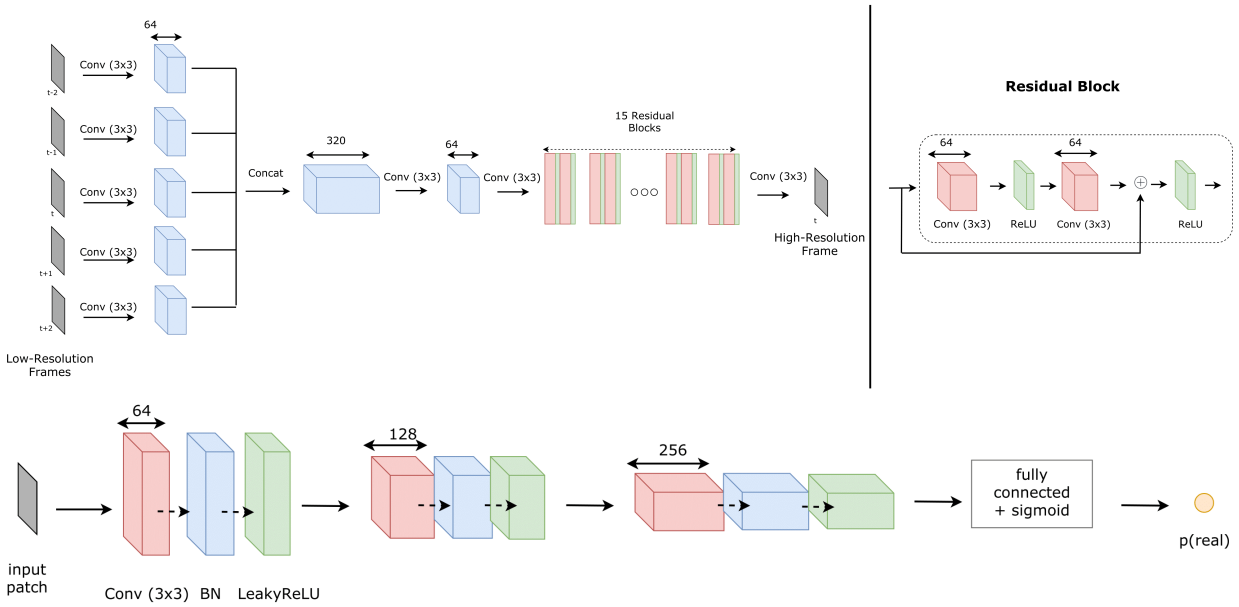

Fig. 2: The proposed architecture for the generator (first row) and discriminator (second row) [10]

rate was set to 10^{-3} and it was then further divided by a factor of 10 at the 50th and 75th epochs of the training. We train our generator for the scale factor 3. Using the weights of this pre-trained generator as initial weights, we trained our spatially adaptive collaborative GAN model with the loss functions defined in (6) for 30 epochs, setting the learning rate to 10^{-4} for the generator and both discriminators. The weight decay was set to 10^{-3} for the discriminators and 10^{-4} for the generator. We use the ADAM [16] optimizer and a batch size of 64. The values of $®_1$ and $®_2$ used in (6) were determined experimentally. We found their optimal values to be: $®_1 = 0.0005$, $®_2 = 0.001$. The parameter ^{2} in (5) is set to 0.001. We found out that 30 epochs were appropriate for our model to converge.

B. Evaluation Results

As we have already indicated, we trained our model on the Myanmar dataset. In order to check whether our model could also work well in different datasets, we tested it on the VidSet4 dataset [17], a commonly used dataset for testing VSR models, which contains 4 scenarios.

We compared our proposed VSRCoDiGAN model with the current state-of-the-art video super-resolution models for VidSet4 test dataset. More specifically, we compared it with VSRResFeatGAN [10] which uses a similar adversarial training approach as ours but with a single discriminator, that is, without explicitly using spatial information into account. Our second model used for comparison is the spatially adaptive GAN (SA-GAN) for the VSR problem [9], which incorporates spatial information into pixel and feature losses and is trained using the traditional single discriminator GAN framework.

Table I reveals that both VSRCoDiGAN and SA-GAN outperform the state-of-the-art VSRResFeatGAN model in all scenarios (calendar, city, foliage, walk). This suggests that including spatial information into training helps to improve the quality of generated frames. Furthermore, we observe that our VSRCoDiGAN model performs better than SA-GAN in most cases, which indicates that it is beneficial to formulate the training with collaborative discriminators to explictly incorporate high frequency spatial information in the learning process. Also, as table I indicates, our method improves the visual quality of the generated frames for all scenarios. A qualitative comparison is shown in Fig. 3. Considering the zoomed in regions in the frames (numbers and letters in the first column and the star in the second column), we can clearly observe that our VSRCoDiGAN model generates more accurately super-resolved edges (closer to those in ground truth frames) with fewer artifacts and less noise compared to VSRResFeatGAN and SA-GAN. We conclude from these quantitative and qualitative results that including a spatially adaptive discriminator to incorporate spatial information into the GAN training has a significant positive impact on the quality of the resulting frames.

TABLE I: PSNR and SSIM comparison with state-of-the-art VSR models on the VidSet4 dataset for scale factor 3.

	VSRResFeatGAN PSNR/SSIM	SA-GAN PSNR/SSIM	VSRCoDiGAN PSNR/SSIM
calendar	23.40/0.8033	23.59/**0.8130**	**23.63**/0.8086
city	27.23/0.7832	27.48/**0.7925**	**27.57**/0.7869
foliage	25.29/0.7544	25.74/0.7754	**26.37/0.7974**
walk	30.20/0.9182	30.40/0.9213	**30.64/0.9230**
Average	26.53/0.8148	26.80/0.8256	**27.06/0.8289**