arXiv:1807.07930v1 [cs.CV] 20 Jul 2018

Photorealistic Video Super Resolution

Eduardo Pérez-Pellitero[1], Mehdi S. M. Sajjadi[1], Michael Hirsch[2], and Bernhard Schölkopf[12]

[1] Max Planck Institute for Intelligent Systems
[2] Amazon Research

Abstract. With the advent of perceptual loss functions, new possibilities in super-resolution have emerged, and we currently have models that successfully generate near-photorealistic high-resolution images from their low-resolution observations. Up to now, however, such approaches have been exclusively limited to single image super-resolution. The application of perceptual loss functions on video processing still entails several challenges, mostly related to the lack of temporal consistency of the generated images, i.e., flickering artifacts. In this work, we present a novel adversarial recurrent network for video upscaling that is able to produce realistic textures in a temporally consistent way. The proposed architecture naturally leverages information from previous frames due to its recurrent architecture, i.e. the input to the generator is composed of the low-resolution image and, additionally, the warped output of the network at the previous step. We also propose an additional loss function to further reinforce temporal consistency in the generated sequences. The experimental validation of our algorithm shows the effectiveness of our approach which obtains competitive samples in terms of perceptual quality with improved temporal consistency.

1 Introduction

Advances in convolutional neural networks have revolutionized computer vision and the popular field of super-resolution has been no exception to this rule, as in recent years numerous publications have made great strides towards better reconstructions of high-resolution pictures. A most promising new trend in super-resolution has emerged as the application of *perceptual* loss functions rather than the previously ubiquitous optimization of the mean squared error. This paradigm shift has enabled the leap from images with blurred textures to near-photorealistic results in terms of perceived image quality using deep neural networks. Notwithstanding the recent success in single image super-resolution, perceptual losses have not yet been successfully utilized in the video super-resolution domain as perceptual losses typically introduce artifacts that, while being undisturbing in the spatial domain, emerge as spurious flickering artifacts in videos.

In this paper we propose a neural network model that is able to produce sharp videos with fine details while improving its behaviour in terms of temporal consistency. The contributions of the paper are: (1) We propose a recurrent generative adversarial model that uses optical flow in order to exploit temporal cues across frames, and (2) we introduce a temporal-consistency loss term that reinforces coherent consecutive frames in the temporal domain.

2 Related work

The task of super-resolution can be split into the groups of single image super-resolution and multi-frame or video super-resolution methods.

Single image super-resolution is one of the most relevant inverse problems in the field of generative image processing tasks [1,2]. Since the initial work by Dong et al. [3] which applied small convolutional neural networks to the task of single image super-resolution, several better neural network architectures have been proposed that have achieved a significantly higher PSNR across various datasets [4,5,6,7,8,9,10]. Generally, advances in network architectures for image detection tasks have also helped in super-resolution, e.g., adding residual connections [11] enables the use of much deeper networks and speeds up training [12]. We refer the reader to Agustsson and Timofte [13] for a survey of the state of the art in single image super-resolution.

Since maximizing for PSNR leads to generally blurry images [14], another line of research has investigated alternative loss functions. Johnson et al. [15] and Alexey and Brox [16] replace the mean squared error (MSE) in the image space with an MSE measurement in feature space of large pre-trained image recognition networks. Ledig et al. [17] extend this idea by adding an adversarial loss and Sajjadi et al. [14] combine perceptual, adversarial and texture synthesis loss terms to produce sharper images with hallucinated details. Although these methods produce detailed images, they typically contain small artifacts that are visible upon close inspection. While such artifacts are bearable in images, they lead to flickering in super-resolved videos. For this reason, applying these perceptual loss functions to the problem of video super-resolution is more involved.

Amongst classical video super-resolution methods, Liu et al. [18] have achieved notable image quality using Bayesian optimization methods, but the computational complexity of the approach prohibits use in real-time applications. Neural network based approaches include Huang et al. [19] who use a bidirectional recurrent architecture with comparably shallow networks without explicit motion compensation. More recently, neural network based methods operate on a sliding window of input frames. The main idea of Kappeler et al. [20] is to align and warp neighboring frames to the current frame before all images are fed into a super-resolution network which combines details from all frames into a single image. Inspired by this idea, Caballero et al. [21] take a similar approach but employ a flow estimation network for the frame alignment. Similarly, Makansi et al. [22] use a sliding window approach but they combine the frame alignment and super-resolution steps. Tao et al. [23] also propose a method which operates on a stack of video frames. They estimate the motion in the frames and subsequently map them into high-resolution space before another super-resolution network combines the information from all frames. Liu et al. [24] operate on varying numbers of frames at the same time to generate different high-resolution images and then condense the results into a single image in a final step.

For generative video processing methods, temporal consistency of the output is crucial. Since most recent methods operate on a sliding window [21,22,23,24], it is hard to optimize the networks to produce temporally consistent results as no information of the previously super-resolved frame is directly included in the next step. To accomodate for this, Sajjadi et al. [25] use a frame-recurrent approach where the estimated high-resolution frame of the previous step is fed into the network for the following step. This encourages more temporally consistent results, however the authors do not explicitly employ a loss term for the temporal consistency of the output.

To the best of our knowledge, video super-resolution methods have so far been restricted to MSE optimization methods and recent advancements in perceptual image quality in single image super-resolution have not yet been successfully transferred to video super-resolution. A possible explanation is that perceptual losses lead to sharper images which makes temporal inconsistencies significantly more evident in the results, leading to unpleasing flickering in the high-resolution videos [14].

The style transfer community has faced similar problems in their transition from single-image to video processing. Single-image style-transfer networks might produce very distant images for adjacent frames [26], creating very strong transitions from frame to frame. Several recent works have overcome this problem by including a temporal-consistency loss that ensures that the stylized consecutive frames are similar to each other when warped with the optical flow of the scene [27,28,29].

In this work, inspired by the contributions above, we explore the application of perceptual losses for video super-resolution using adversarial training and temporal consistency objectives.

3 Proposed method

3.1 Notation and problem statement

Video super-resolution aims at upscaling a given LR image sequence $\{Y_t\}$ by a factor of s, so that the estimated sequence $\left\{\tilde{X}_t\right\}$ resembles the original sequence $\{X_t\}$ by some metric. We denote images in the low-resolution domain by $Y \in [0,1]^{h\times w\times 3}$, and ground-truth images in the high-resolution domain by $X \in [0,1]^{sh\times sw\times 3}$ for a given magnification factor s. An estimate of a high-resolution image X is denoted by $\tilde{X}$. We discern within a temporal sequence by a subindex to the image variable, e.g., Y_{t-1}, Y_t. We use a superscript w, e.g. $\tilde{X}^w_{t-1}$, to denote an image $\tilde{X}$ that has been warped from its time step $t-1$ to the following frame X_t.

The proposed architecture is summarized in Figure 1 and will be explained in detail in the following sections. We define an architecture that naturally leverages not only single image but also inter-frame details present in video sequences by

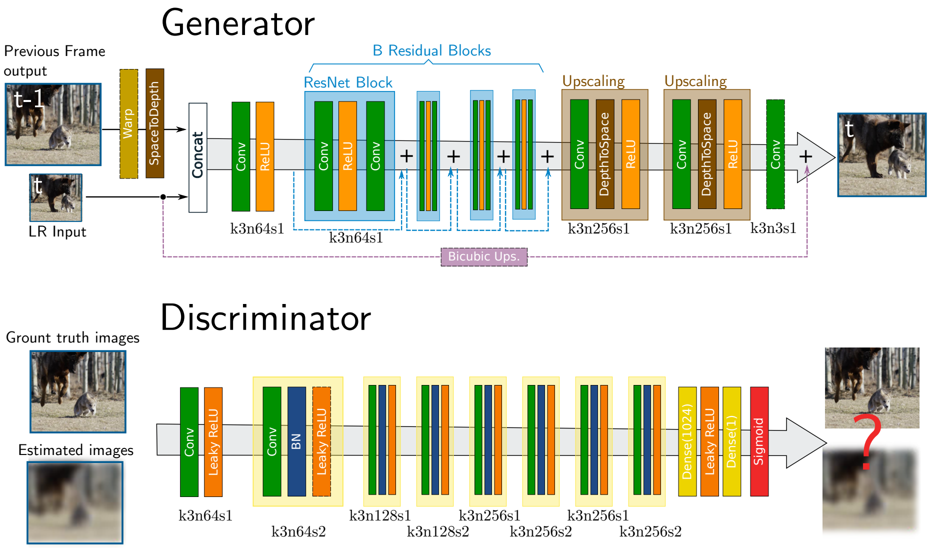

Fig. 1. Network architectures for generator and discriminator. The previous output frame is warped onto the current frame and mapped to LR with the space to depth transformation before being concatenated to the current LR input frame. The generator follows a ResNet architecture with skip connections around the residual blocks and around the whole network. The discriminator follows the common design pattern of decreasing the spatial dimension of the images while increasing the number of channels after each block.

using a recurrent neural network architecture. The previous output frame is warped according to the optical flow estimate given by FlowNet 2.0 [30]. By including a discriminator that is only needed at the training stage, we further enable adversarial training which has proved to be a powerful tool for generating sharper and more realistic images [14, 17].

To the best of our knowledge, the use of perceptual loss functions (i.e. adversarial training in recurrent architectures) for video super-resolution is novel. In a recently published work, Sajjadi et al. [25] propose a similar recurrent architecture for video super-resolution, however, they do not utilize a perceptual objective and in contrast to our method, they do not apply an explicit loss term that enforces temporal consistency.

3.2 Recurrent generator and discriminator

Following recent super-resolution state of the art methods for both classical and perceptual loss functions [9,14,17,31], we use deep convolutional neural networks with residual connections. This class of networks facilitates learning the identity mapping and leads to better gradient flow through deep networks. Specifically, we adopt a ResNet architecture for our recurrent generator that is similar to the ones introduced by [14, 17] with some modifications.

Each of the residual blocks is composed by a convolution, a Rectified Linear Unit (ReLU) activation and another convolutional layer following the activation. Previous approaches have applied batch normalization layers in the residual blocks [17], but we choose not to add batch normalization to the generator due to the comparably small batch size, to avoid potential color shift problems, and also taking into account recent evidence hinting that they might be problematic for generative image models [32]. In order to further accelerate and stabilize training, we create an additional skip connection over the whole generator. This means that the network only needs to learn the residual between the bicubic interpolation of the input and the high-resolution ground-truth image rather than having to pass through all low frequencies as well [7, 14].

We perform most of our convolutions in low-resolution space for a higher receptive field and higher efficiency. Since the input image has a lower dimension than the output image, the generator needs to have a module that increases the resolution towards the end. There are several ways to do so within a neural network, e.g., transposed convolution layers, interpolation, or depth to space units *(pixelshuffle)*. Following Shi et al. [4], we reshuffle the activations from the channel dimension to the height and width dimensions so that the effective spatial resolution is increased (and, consequently, this operation decreases the number of channels). The upscaling unit is divided into two stages with an intermediate magnification step r (e.g. two times $\times 2$ for a magnification factor of $\times 4$). Each of the upscaling stages is composed of a convolutional layer that increments the number of channels by a factor of r^2, a depth to space operation and a ReLU activation.

Our discriminator follows common design choices and is composed by strided convolutions, batch normalization and leaky ReLU activations that progressively decrease the spatial resolution of the activations and increase the channel count [14, 17, 33]. The last stage of the discriminator is composed of two dense layers and a sigmoid activation function.

In contract to general generative adversarial networks, the input to the proposed generative network is not a random variable but it is composed of the low-resolution image Y_t (corresponding to the current frame t) and, additionally, the warped output of the network at the previous step $\tilde{X}^w_{t-1}$. The difference in resolution of these two images is adapted through a space to channel layer which decreases the spatial resolution of $X^{\tilde{w}}_{t-1}$ without loss of data.

For warping the previous image, a dense optical flow field is estimated with a flow estimation network as described in the following section. As described in Section 3.4, the warped frames are also used in an additional loss term that enforces higher temporal consistency in the results.

3.3 Flow estimation

Accurate dense flow estimation is crucial to the success of the proposed architecture. For this reason, we opt to use one of the best available flow estimation methods, FlowNet 2.0 [30]. We use the pre-trained model supplied by the authors [34] and run optical flow estimation on our whole dataset. FlowNet 2.0 is

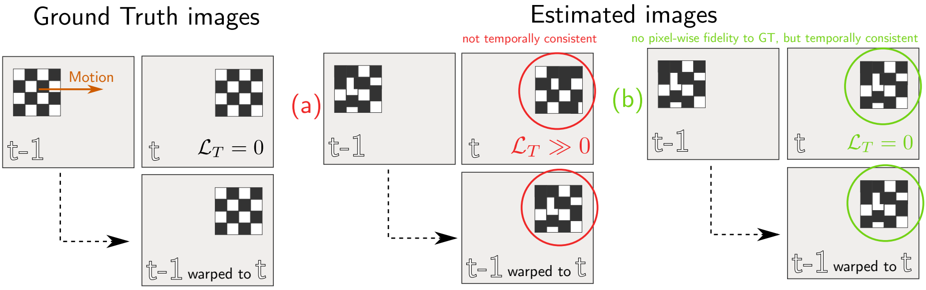

Fig. 2. Behavior of the proposed temporal-consistency loss. The sequence depicts a checkerboard pattern which moves to the right. In the first group of GT images, the warped images are exactly similar, and thus the loss is 0. In (a), the results are not temporally consistent, so the warped image is different from the current frame which leads to a loss that is higher than 0. In the example in (b), the estimated patterns are temporally consistent despite not being the same as the GT images, so the loss is 0.

the successor of the original FlowNet architecture which was the first approach that used convolutional neural networks to predict dense optical flow maps. The authors show that it is both faster and more accurate than its predecessor. Besides a more sophisticated training procedure, FlowNet 2.0 relies on an arrangement of stacked networks that capture large displacements in coarse flow estimates which are then refined by the next network in a subsequent step. In a final step, the estimates are fused by a shallow fusion network. For details, we refer the reader to the original publication [30].

3.4 Losses

We train our model with three different loss terms, namely: pixel-wise mean squared error (MSE), adversarial loss and a temporal-consistency loss.

Mean Squared Error MSE is by far the most common loss in the super-resolution literature as it is well-understood and easy to compute. It accurately captures sharp edges and contours, but it leads to over-smooth and flat textures as the reconstruction of high-frequency areas falls to the local mean rather than a realistic mode [14].

The pixel-wise MSE is defined as the Frobenius norm of the difference of two images:

$$\mathcal{L}_E = \left\| \tilde{X}_t - X_t \right\|_2^2, \tag{1}$$

where $\tilde{X}_t$ denotes the estimated image of the generator for frame t and X_t denotes the ground-truth HR frame t.

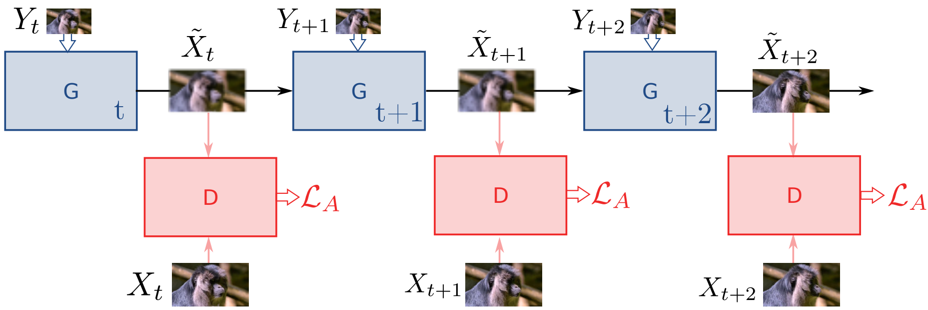

Fig. 3. Unfolded recurrent generator G and discriminator D during training for 3 temporal steps. The output of the previous time step is fed into the generator for the next iteration. Note that the weights of G and D are shared across different time steps. Gradients of all losses during training pass through the whole unrolled configuration of network instances. The discriminator is applied independently for each time step.

Adversarial Loss Generative Adversarial Networks (GANs) [35] and their characteristic adversarial training scheme have been a very active research field in the recent years, defining a wide landscape of applications. In GANs, a generative model is obtained by simultaneously training an additional network. A generative model G (i.e. generator) that learns to produce samples close to the data distribution of the training set is trained along with a discriminative model D (i.e. discriminator) that estimates the probability of a given sample belonging to the training set or not, i.e., it is generated by G. The objective of G is to maximize the errors committed by D, whereas the training of D should minimize its own errors, leading to a two-player minimax game.

Similar to previous single-image super-resolution [14, 17], the input to the generator G is not a random vector but an LR image (in our case, with an additional recurrent input), and thus the generator minimizes the following loss:

$$\mathcal{L}_A = -\log(D(G(Y_t||\tilde{X}_{t-1})), \tag{2}$$

where the operator $||$ denotes concatenation. The discriminator minimizes:

$$\mathcal{L}_D = -\log(D(X_t)) - \log(1 - D(G(Y_t||\tilde{X}_{t-1})). \tag{3}$$

Temporal-consistency Loss Upscaling video sequences has the additional challenge of respecting the original temporal consistency between adjacent frames so that the estimated video does not present unpleasing flickering artifacts.

When minimizing only $\mathcal{L}_E$ such artifacts are less noticeable for two main reasons: because (1) MSE minimization often converges to the mean in textured regions, and thus flickering is reduced and (2) the pixel-wise MSE with respect to the GT is up to a certain point enforcing the inter-frame consistency present in the training images. However, when adding an adversarial loss term, the difficult to maintain temporal consistency increases. Adversarial training aims at generating samples that lie in the manifold of images, and thus it generates high-frequency content that will hardly be pixel-wise accurate to any ground-truth image. Generating video frames separately thus introduces unpleasing flickering artifacts.

In order to tackle the aforementioned limitation, we introduce the temporal-consistency loss, which has already been successfully used in the style transfer community [27, 28, 29]. The temporal-consistency loss is computed from two adjacent estimated frames (without need of ground-truth), by warping the frame $t-1$ to t and computing the MSE between them. We show an example of the behavior of our proposed temporal-consistency loss in Figure 2.

Let $W(X, O)$ denote a image warping operation and O an optical flow field mapping $t-1$ to t. Our proposed loss reads:

$$\mathcal{L}_T = \left\| \tilde{X}^w_{t-1} - \tilde{X}_t \right\|_2^2, \qquad \text{for } \tilde{X}^w_{t-1} = W(\tilde{X}_{t-1}, O). \tag{4}$$

4 Results

4.1 Training and parameters

Our model falls in the category of recurrent neural networks, and thus must be trained via Backpropagation Through Time (BPTT) [36], which is a finite approximation of the infinite recurrent loop created in the model. In practice, BPTT unfolds the network into several temporal steps where each of those steps is a copy of the network sharing the same parameters. The backpropagation algorithm is then used to obtain gradients of the loss with respect to the parameters. An example of unfolded recurrent generator and discriminator can be visualized in Figure 3. We select 10 temporal steps for our training approximation. Note that our discriminator classifies each image independently and is not recurrent, thus the different images produced by G can be stacked in the batch dimension (i.e. the discriminator does not have any connection between adjacent frames).

Our training set is composed by 4k videos downloaded from *youtube.com* and downscaled to 720×1280, from which we extract around 300.000 128×128 HR crops that serve as ground-truth images, and then further downsample them by a factor of $s = 4$ to obtain the LR input of size 32×32. The training dataset thus is composed by around 30.000 sequences of 10 frames each (i.e. around 30.000 data-points for the recurrent network). We precompute the optical flows with FlowNet 2.0 and load them both during training and testing, as GPU memory becomes scarce specially when unfolding generator and discriminator. We compile a testing set, larger than other previous testing sets in the literature, also downloaded from *youtube.com*, favoring sharp 4k content that is further downsampled to 720×1280 for GT and 180×320 for the LR input. In this dataset there are 12 sequences of diverse nature scenes (e.g. landscapes, natural life, urban scenes) ranging from very little to fast motion. Each sequence is composed of roughly 100 to 150 frames, which totals 1281 frames.

We use a batch size of 8 sequences, i.e., in total each batch contains $8 \times 10 = 80$ training images. All models are pre-trained with $\mathcal{L}_E$ for about 100k training iterations and then trained with the adversarial and temporal loss for about 1.5M iterations. Training was performed on Nvidia Tesla P100 and V100 GPUs, both of which have 16 GB of memory.

4.2 Evaluation

Seq.	**Methods**					
	bicubic	$\mathcal{L}_E$	$\mathcal{L}_{EA}$	$\mathcal{L}_{EAT}$	$\mathcal{L}_A^{SI}$	ENet
bear	0.11527	**0.0531**	0.07284	0.06416	0.06259	0.10279
bird	0.00151	**0.00056**	0.00214	0.00201	0.00102	0.00376
elephant	0.08537	**0.01122**	0.02527	0.01768	0.02925	0.06607
monkey	**0.00006**	0.00007	0.00014	0.00012	0.00008	0.00125
mountain	**0.00008**	**0.00008**	0.00036	0.00014	0.00011	0.00112
newyork	0.23892	0.13093	0.10198	0.1126	0.11511	**0.07888**
newyork2	0.20215	0.06112	0.06522	0.07918	**0.05806**	0.04399
penguin	0.3124	0.17488	0.09841	0.16076	0.16188	**0.07866**
tikal1	0.29802	0.17697	0.09933	0.14472	0.12382	**0.0659**
tikal2	0.08847	0.05812	**0.03569**	0.05089	0.04226	0.06508
tikal3	0.23665	0.09727	**0.09650**	0.12287	0.11457	0.09652
wolf	0.03453	0.01589	0.01985	**0.01534**	0.01573	0.03085
average	0.13445	0.06502	**0.05148**	0.06421	0.06037	0.05291

Table 1. LPIPS scores (AlexNet architecture with linear calibration layer) for 12 sequences. Best performers in bold font, and runner's-up in blue color. The best performer on average is our proposed model trained with $\mathcal{L}_{EA}$ followed closely by ENet.

Models We include in our validation three loss configuration: (1) $\mathcal{L}_E$ is trained only with MSE loss as a baseline, (2) $\mathcal{L}_{EA}$ is our adversarial model trained with $\mathcal{L}_E + 3 \times 10^{-3}\mathcal{L}_A$ and (3) $\mathcal{L}_{EAT}$ is our adversarial model with temporal-consistency loss $\mathcal{L}_E + 3 \times 10^{-3}\mathcal{L}_A + 10^{-2}\mathcal{L}_T$.

We also include two other state-of-the art models in our benchmarking: EnhanceNet (ENet) as a perceptual single image super-resolution baseline [14] (code and pre-trained network weights obtained from the authors' website), which minimizes an additional loss term based on the Gram matrix of VGG activations; and lastly our model without flow estimation or recurrent connections that we denote in the tables by $\mathcal{L}_A^{SI}$. This last model is very similar to the network used in SRGAN from Ledig et al. [17].

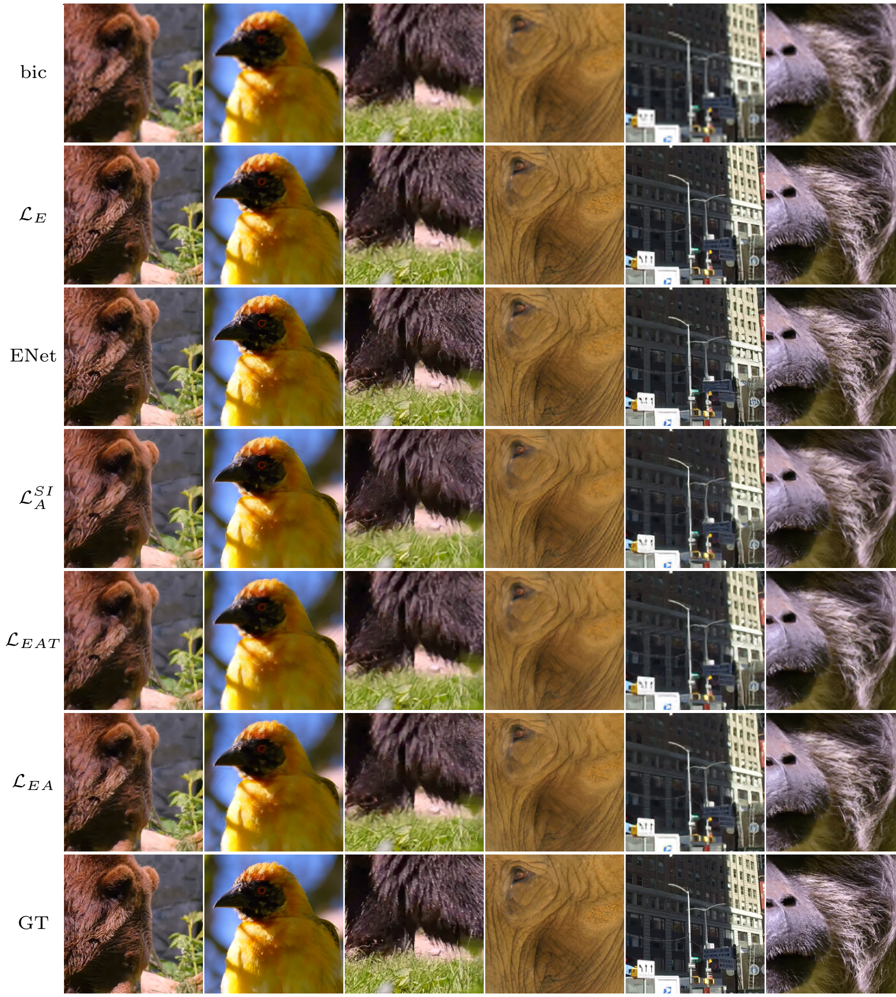

Fig. 4. Image close-ups for visual inspection and quantitative visual assessment. The close-ups have been extracted from the following sequences (left to right): *bear, bird, bear, elephant, newyork2, monkey.*

Intra-frame quality Evaluation of images trained with perceptual losses is still an open problem. Even though it is trivial for humans to evaluate the perceived similarity between two images, the underlying principles of human perception are still not well-understood. Traditional metrics such as PSNR (based on MSE), Structural Self-Similarity (SSIM) or the Information Fidelity Criterion (IFC) still rely on well-aligned, more or less pixel-wise accuracy estimates, and minor artifacts in the images can cause great perturbations in them. In order to evaluate image samples from models that deviate from the MSE minimization scheme other metrics need to be considered.

Seq.	**Methods**					
	bicubic	$\mathcal{L}_E$	$\mathcal{L}_{EA}$	$\mathcal{L}_{EAT}$	$\mathcal{L}_A^{SI}$	ENet
bear	0.01398	0.01832	0.01883	**0.01748**	0.02343	0.03661
bird	0.0138	0.01673	0.01706	**0.01625**	0.01997	0.02727
elephant	0.01055	0.01264	0.01324	**0.01247**	0.01535	0.02126
monkey	0.00983	0.01165	0.0124	**0.01163**	0.01417	0.02118
mountain	0.00839	**0.00988**	0.01069	0.00993	0.01209	0.01824
newyork	0.00962	0.01169	0.01222	**0.01144**	0.01402	0.0201
newyork2	0.00936	0.01124	0.01188	**0.01112**	0.01399	0.02003
penguin	0.00818	0.00978	0.01045	**0.00974**	0.01221	0.01802
tikal1	0.00831	**0.01004**	0.01095	0.01006	0.01298	0.02048
tikal2	0.00819	**0.00982**	0.01076	0.00987	0.01264	0.0201
tikal3	0.00845	**0.00996**	0.0109	0.01005	0.01269	0.01986
wolf	0.01039	**0.01191**	0.01276	0.01199	0.01443	0.02147
average	0.00992	0.01197	0.01268	**0.01184**	0.01483	0.02205

Table 2. Temporal Consistency Loss for adjacent frames (initial frame has not been included in the computation). Best performer in bold and runner-up in blue color (omitting bicubic). The best performer on average is our proposed method trained with $\mathcal{L}_{EAT}$ followed by $\mathcal{L}_E$.

The recent work of Zhang et al. [37] explores the capabilities of deep architectures to capture perceptual features that are meaningful for similarity assessment. In their exhaustive evaluation they show how deep features of different architectures outperform other previous metrics by substantial margins and correlate very well with subjective human scores. They conclude that deep networks, regardless of the specific architecture, capture important perceptual features that are well-aligned with those of the human visual system. Consequently, they propose the Learned Perceptual Image Patch Similarity metric (LPIPS).

We evaluate our testing set with LPIPS using the AlexNet architecture with an additional linear calibration layer as the authors propose in their manuscript.

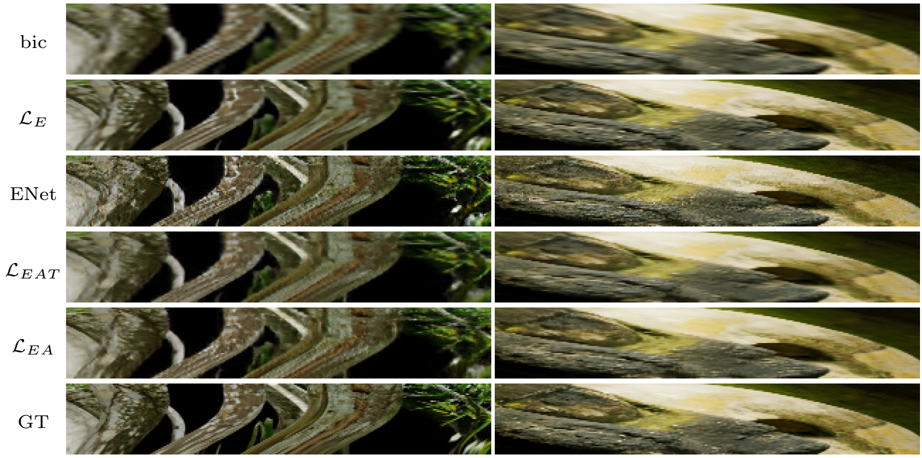

Fig. 5. Temporal profiles (50 frames, central line of the original image) of two video sequences. From top to bottom: Bicubic, ENet, Ground Truth, $\mathcal{L}_{EA}$, $\mathcal{L}_{EAT}$, $\mathcal{L}_E$. Whereas ENet produces very sharp images, the profiles show how it is temporally unconsistent (i.e. profiles are often cut or jaggy). Our proposed models present cleaner profiles, while still retaining photorealistic textures. Better viewed zoomed in.

We show our LPIPS scores in Table 1. These scores are in line with what we show in the quantitative visual inspection in Figure 4: The samples obtained by ENet are together with $\mathcal{L}_{EA}$ the most similar to the GT, with ENet producing slightly sharper images than our proposed method. $\mathcal{L}_{EAT}$ comes afterwards, as a more conservative generative network (i.e. closer to the one obtained with the MSE loss).

Temporal Consistency To evaluate the temporal consistency of the estimated videos, we compute the temporal-consistency loss as described in Equation 4 and Fig. 2 between adjacent frames for all the methods in the benchmark.

We show the results in Table 2. We note that all the configurations that are recurrent perform well in this metric, even when we do not minimize $\mathcal{L}_T$ directly (e.g. $\mathcal{L}_E$, $\mathcal{L}_{EA}$). In contrast, models that are not aware of the temporal dimensions (such as ENet or $\mathcal{L}_A^{SI}$) obtain higher errors, validating that the recurrent network learns inter-frame relationships. Not considering the bicubic interpolation which is very blurry, the best performer is $\mathcal{L}_{EAT}$ followed closely by $\mathcal{L}_E$. Our model $\mathcal{L}_{EA}$ indeed performs reasonably well, especially taking into consideration that it is the best performer in the quality scores shown in Table 1.

Evaluating the temporal consistency over adjacent frames in a sequence where the ground-truth optical flow is not known poses several problems, as errors present in the flow estimation will directly affect this metric. Additionally, the bilinear resampling performed for the image warping is, when analyzed in the frequency domain, a low-pass filter that can potentially blur high-frequencies and thus, result in an increase of uncertainty in the measured error. In order to ensure the reliability of the temporal consistency validation, we perform further testing with the MPI Sintel synthetic training sequence (which includes ground-truth optical flow). This enables us to asses the impact of using estimated flows in the temporal consistency metric. We show in Table 3 the results in terms of temporal consistency of the MPI Sintel training 23 sequences using the GT and also FlowNet2 estimated optical flows for the warping in the metric. The error from estimated flows to GT is not significant, and the relationship among methods is similar: $\mathcal{L}_{EA}$ improves greatly over the non-recurrent SRGAN $\mathcal{L}_{A}^{SI}$ or EnhanceNet, and $\mathcal{L}_{EAT}$ (with temporal consistency loss) further improves over $\mathcal{L}_{EA}$.

Seq.	**Methods**					
	bicubic	$\mathcal{L}_E$	$\mathcal{L}_{EA}$	$\mathcal{L}_{EAT}$	$\mathcal{L}_A^{SI}$	ENet
GT flow	0.07784	0.02823	0.02879	0.02865	0.64735	0.08927
FlowNet2 flow	0.07563	0.03112	0.03094	0.03050	0.65836	0.08694

Table 3. Temporal Consistency Loss for adjacent frames for MPI Sintel Dataset using ground-truth and estimated optical flows.

Following the example of [21], we also show in Figure 5 temporal profiles for qualitative evaluation of temporal consistency. A temporal profile shows an image where each row is a fixed line over a set of consecutive time frames, creating a 2-dimensional visualization of the temporal evolution of the line. In this figure, we can corroborate the objective assessment performed in Table 1 and Table 2. ENet produces very appealing sharp textures, to the point that it is hallucinating frequencies not present in the GT image (i.e. over-sharpening). This is not necessarily unpleasing with static images, but temporal consistency then becomes very challenging, and some of those textures resemble noise in the temporal profile. Our model $\mathcal{L}_A$ is hardly distinguishable from the GT image, generating high-quality plausible textures, while showing a very clean temporal profile. $\mathcal{L}_{EAT}$ has fewer hallucinated textures than $\mathcal{L}_A$, but on the other hand we also see an improved temporal behavior (i.e. less flickering).

5 Conclusions

We presented a novel generative adversarial model for video upscaling. Differently from previous approaches to video super-resolution based on MSE minimization, we used adversarial loss functions in order to recover videos with photorealistic textures. To the best of our knowledge, this is the first work that applies perceptual loss functions to the task of video super-resolution.

In order to tackle the problem of lacking temporal consistency due to perceptual loss functions, we propose two synergetic contributions: (1) A recurrent generator and discriminator model where the output of frame $t-1$ is passed on to the next iteration in combination with the input of frame t, enabling temporal cues during learning and inference. Models trained with adversarial and MSE losses show improved behavior in terms of temporal consistency and a competitive quality when compared to SISR models. (2) Additionally, we introduce the temporal-consistency loss to video super-resolution, in which deviations from the previous warped frame are punished when estimating a given frame. We conducted evaluation by means of the LPIPS and temporal-consistency loss on a testing dataset of more than a thousand *4k* video frames, obtaining promising results that open new possibilities within video upscaling.

References

1. Milanfar, P.: Super-resolution Imaging. CRC press (2010)
2. Nasrollahi, K., Moeslund, T.B.: Super-resolution: A comprehensive survey. Machine Vision and Applications (2014)
3. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV. (2014)
4. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR. (2016)
5. Kim, J., Kwon Lee, J., Mu Lee, K.: Deeply-recursive convolutional network for image super-resolution. In: CVPR. (2016)
6. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: ECCV. (2016)
7. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: CVPR. (2017)
8. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: CVPR. (2017)
9. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPR workshop. (2017)
10. Chen, C., Tian, X., Wu, F., Xiong, Z.: UDNet: Up-down network for compact and efficient feature representation in image super-resolution. In: ICCV. (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)
12. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: CVPR. (2016)
13. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: CVPR workshop. (2017)
14. Sajjadi, M.S.M., Schölkopf, B., Hirsch, M.: EnhanceNet: Single image super-resolution through automated texture synthesis. In: ICCV. (2017)
15. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. (2016)
16. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: NIPS. (2016)
17. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR. (2017)

18. Liu, C., Sun, D.: A bayesian approach to adaptive video super resolution. In: CVPR. (2011)
19. Huang, Y., Wang, W., Wang, L.: Bidirectional recurrent convolutional networks for multi-frame super-resolution. In: NIPS. (2015)
20. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. In: IEEE Transactions on Computational Imaging. (2016)
21. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: CVPR. (2017)
22. Makansi, O., Ilg, E., Brox, T.: End-to-end learning of video super-resolution with motion compensation. In: GCPR. (2017)
23. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: ICCV. (2017)
24. Liu, D., Wang, Z., Fan, Y., Liu, X., Wang, Z., Chang, S., Huang, T.: Robust video super-resolution with learned temporal dynamics. In: CVPR. (2017)
25. Sajjadi, M.S.M., Vemulapalli, R., Brown, M.: Frame-Recurrent Video Super-Resolution. In: CVPR. (2018)
26. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR. (2016)
27. Gupta, A., Johnson, J., Alahi, A., Fei-Fei, L.: Characterizing and improving stability in neural style transfer. In: ICCV. (2017)
28. Ruder, M., Dosovitskiy, A., Brox, T.: Artistic style transfer for videos. GCPR (2016)
29. Huang, H., Wang, H., Luo, W., Ma, L., Jiang, W., Zhu, X., Li, Z., Liu, W.: Real-time neural style transfer for videos. In: CVPR. (2017)
30. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: CVPR. (2017)
31. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR. (2016)
32. Xiang, S., Li, H.: On the effects of batch and weight normalization in generative adversarial networks. arXiv:1704.03971v4 (2017)
33. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR. (2016)
34. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Caffe for flownet2. https://github.com/lmb-freiburg/flownet2 (2017) (last commit on Oct 27, 2017).
35. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014)
36. Werbos, P.J.: Backpropagation through time: what it does and how to do it. Proceedings of the IEEE (1990) 1550–1560
37. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. arXiv:1801.03924 (2018)