Fast Video Visual Quality and Resolution Improvement using SR-UNet

Federico Vaccaro, Marco Bertini, Tiberio Uricchio, Alberto Del Bimbo
MICC - Università degli Studi di Firenze
Firenze, Italy
[name.surname]@unifi.it

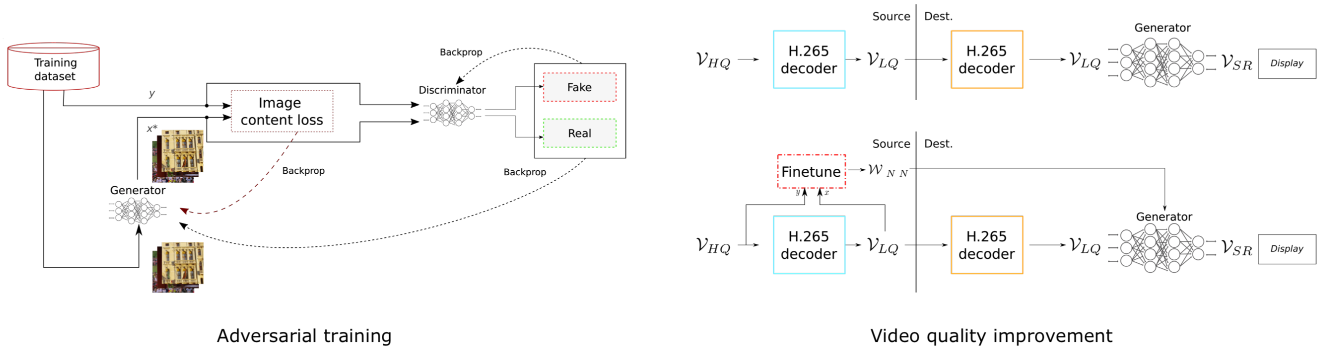

Figure 1: System overview: *left)* SR-UNet is trained using a GAN-based framework, with an image content loss that combines perceptual and signal based metrics; *top right)* the SR-UNet generator is used for visual quality improvement in real-time; *bottom right)* the network can be fine tuned for a specific video, to further improve its performance.

ABSTRACT

In this paper, we address the problem of real-time video quality enhancement, considering both frame super-resolution and compression artifact-removal. The first operation increases the sampling resolution of video frames, the second removes visual artifacts such as blurriness, noise, aliasing, or blockiness introduced by lossy compression techniques, such as JPEG encoding for single-images, or H.264/H.265 for video data.

We propose to use SR-UNet, a novel network architecture based on UNet, that has been specialized for fast visual quality improvement (i.e. capable of operating in less than 40ms, to be able to operate on videos at 25FPS). We show how this network can be used in a streaming context where the content is generated live, e.g. in video calls, and how it can be optimized when video to be streamed are prepared in advance. The network can be used as a final post processing, to optimize the visual appearance of a frame before showing it to the end-user in a video player. Thus, it can be applied without any change to existing video coding and transmission pipelines.

Experiments carried on standard video datasets, also considering the H.265 compression, show that the proposed approach is able to either improve visual quality metrics given a fixed bandwidth budget, or video distortion given a fixed quality goal.

CCS CONCEPTS

• **Computing methodologies** → Video segmentation; **Image compression**; **Neural networks**; **Learning from critiques**.

KEYWORDS

GANs, video super resolution, video quality enhancement

ACM Reference Format:
Federico Vaccaro, Marco Bertini, Tiberio Uricchio, Alberto Del Bimbo. 2021. Fast Video Visual Quality and Resolution Improvement using SR-UNet. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475683

1 INTRODUCTION

Video streaming has become the major source of Internet traffic in the latest years, over desktop and mobile platforms, either for work or entertainment. Videoconferencing has become an important form of communication, especially after the emergence of the COVID-19 pandemic, and video on demand (VOD) streaming services like Netflix, Amazon Prime Video and Disney+ provide an alternative to cable or satellite TV, offering movies and shows along with broadcasters that offer their live programmes through streaming apps. All of these applications require video compression algorithms like H.264 or the more recent H.265, to optimize the available bandwidth and reduce transmission costs. However, these compression algorithms are usually lossy and they introduce visual

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475683

artifacts like blocking, mosquito noise, posterization, etc. that may hinder the user experience.

In this work, we propose to use a novel neural network, called SR-UNet, to improve the visual quality of the decoded video on the device of the end user in real-time, without requiring any change in the video compression and delivery pipeline. This network is designed to improve the visual quality while reducing the bandwidth required to stream a video; it does so by performing both super resolution, i.e. reconstructing high resolution frames from a low resolution stream, and reducing video compression artifacts. Considering the visual quality improved by these operations, in order to reduce bandwidth consumption, videos may be streamed at lower resolution or with a higher compression factor.

The main contributions of this work are: *i)* a novel network that extends the UNet architecture by reducing its size and computational cost; *ii)* a loss that combines signal-based and perceptual-based losses within a Generative Adversarial Network (GAN) framework. Furthermore, we show how the proposed network can be tailored on specific video clips, to further improve the performance of the base SR-UNet model. This operating context is particularly relevant for VOD services. They commonly aim to reduce the bandwidth required to transmit videos, for instance, by looking for the best encoding parameters of each video. Extensive experiments on a standard video dataset encoded with H.265 show that the proposed network outperforms baselines and other state-of-the-art approaches; objective and subjective evaluations show that the network is able to improve the visual appearance of videos at different compression rates.

2 PREVIOUS WORK

Visual quality of images and videos can be improved addressing different aspects, i.e. increasing the resolution (super resolution - SR) and eliminating the compression artifacts or other quality defects, such as noise.

2.1 Super resolution

In [30] has been proposed ESPCN, an architecture that operates in the low-resolution feature space and performs the upsample just in the last layer, reducing computational cost and yielding better results. In [2], Caballero *et al.* introduce a novel technique which consists in first training a small network for computing the optical flow between two adjacent frames, and then applying the predicted flow to “align” the low-res adjacent frames into the current one, feeding them into the ESPCN SR-network. In [20], Ledig *et al.* train an SR-ResNet as a Generative Adversarial Network, with also the support of the perceptual VGG-loss. After noticing how their models yielded low similarity with signal-based metrics, they performed extensive subjective tests proving how their perceptual-oriented training was capable of producing more realistic upsamples, compared to the former methodologies. In [29], Sajjadi *et al.* address video SR proposing a RNN-based technique. A flow-net is used for motion compensation between previous and current frame, and they also train their architecture using the SR-frame generated at the previous time-step along with the current LR-frame. Frames are aligned with a predicted optical flow, similarly to [2], then spatially compressed with a Space-to-Depth operation. In [3], Chu *et al.* proposed a novel GAN-based technique for VSR. Starting from the recurrent framework proposed in [29], they train the network through a novel temporal-discriminator. The discriminator is fed with a sequence of generated and ground-truth frames, to better understand the temporal consistencies (and inconsistencies) between frames. The “Ping-Pong" loss is introduced: considering one frame at a certain timestep, the result should not change if coming from the previous or next timestep. This provides a powerful training constraint but could also be intended as a data-augmentation technique.

2.2 Artifact removal and quality restoration

In [23], Maleki *et al.* propose *Block CNN* for JPEG artifact removal and also for image compression. This architecture operates on the typical 8 × 8-block JPEG artifacts: each block is restored by the CNN separately, but considering also its adjacent blocks. Block-CNN layers are structured on residual blocks since artifacts can be modeled as a residual added on the original image. In [21], Li *et al.* apply multiple context-based channel attentions to capture features from different resolutions. The entire architecture is trained progressively from the image space of low quality factor to that of high quality factor. The employed architecture is a stack of 4 hourglass (UNet) networks. In [35], Xu *et al.* address video artifact removal, proposing a novel end-to-end deep neural network called “non-local ConvLSTM” (NL-ConvLSTM) that exploits multiple consecutive frames. This architecture is structured as an autoencoder. However, between the two components, they collocated a ConvLSTM to capture temporal information. Since ConvLSTM is not good at handling large motions and blur, they also embed the non-local (NL) mechanism, which can be seen as a special attention.

Several methods have used the Generative Adversarial Network (GAN) approach. In [8], Galteri *et al.* apply adversarial training methods for removal of artifact generated by lossy compression algorithm for images and videos. A relevant novelty of the work is the idea of learning the discriminator over sub-patches of a single generated patch to reduce high frequency noise, such as mosquito noise which is hard to remove using a full-patch discriminator, and to train the models using larger batches. To tackle variable compression factor, an ensemble of N networks is adopting, to avoid mode collapse phenomenon when using a single model. In [10] and [11], Galteri *et al.* address the problem of artifact removal in real-time. Respect to [8], the generator is inspired from the blocks of *MobileNetV2*, after replacing the standard convolutional layer with lighter depth-wise separable convolutions. In [24] Mameli *et al.* applied the *no-GAN* approach for compression artifact removal and also for super-resolution. The no-GAN approach is characterized by an initial pre-training of generator and discriminator, followed by fine-tuning the generator with very few GAN training iterations. In [17], Kaneko *et al.* propose an architecture capable of removing noise (NR-GAN), compression artifacts (CA-GAN) and blurring artifacts (BR-GAN) from images. Since these issues may occur in combination, the three models are merged into one single BNCR-GAN architecture; in particular, they introduce an adaptive consistency losses to handle the uncertainty caused by the combination. In [26], Pourreza *et al.* study how the quality loss due to video compression causes a loss of accuracy in the action recognition task, and propose a GAN-based quality enhancement method to alleviate the issue. Their architecture employs a frame-recurrent strategy to gather the temporal information in the enhancement process: to improve the current frame, they add to the input also the previous frame, warped (with spatial transformers [14]) in the direction of the optical flow, estimated with a secondary network. In [36], Yu *et al.* address HEVC compression-generated artifact removal, proposing VR-GAN, a GAN-based architecture, that operates at inter-frame level. To maintain the coherence of the generated frames, they use flow estimation based on the current and the previous frame. Location residuals, estimated by the flow net, are added to the location of the previous frame as motion compensation. Then, the prediction result and the current low-quality frame are input to the generator. The estimated location is used to warp the image reconstructed from the previous frame, and to produce an initial estimation of the current frame, in a manner similar to [26]. In [33], He *et al.* uses an adversarial loss combined with L1 loss to enhance HEVC compressed videos. The first part of the loss lets the generator to add missing high frequency details, the second is the reconstruction term. The generator is a residual network, to speed up convergence during training. During training, the image is split in blocks of dimension 32 × 32. In [7], Galteri *et al.* propose a full pipeline architecture composed of semantic deep encoding and decoding. Semantic video encoding allocates more bits to the regions that depict semantically interesting content, using a semantic mask constructed via deep network for each frame. User studies on videos crafted in this way have shown little or no damage on the user experience [34]. On the decoding side, a Relativistic GAN and a loss that accounts for the segmentation are used.

2.3 Quality metrics

In recent years, new perceptual quality metrics, based on deep learning, have been proposed, to complement signal-based metrics such as SSIM and PSNR. These metrics are particularly relevant when evaluating generative models since. In fact, as reported in [18] following a subjective study, images generated by GANs may appear quite realistic and similar to an original, yet they may match it poorly based on simple pixel comparisons; metrics based on "naturalness" are thus more suitable in this case.

In [38], Zhang *et al.* collected a very large dataset of human judgements about similarity between distorted images (photometric, noise, blur, spatial and compression distortion), and proposed a novel full-reference metric, called LPIPS (Learned Perceptual Image Patch Similarity), that evaluates the distance between image patches based on deep features; LPIPS outperforms traditional metrics like SSIM by a large margin in a two alternative forced choice (2AFC) test. This metric can be used as loss for training purposes.

In [1], Blau and Michaeli propose a generalization of rate-distortion theory which takes perceptual quality into account. Incorporating generative adversarial losses has been shown to lead to significantly better perceptual quality, but at the cost of increased distortion: their theoretical characterization leads to the fact that, to obtain good perceptual quality, it is necessary to make a sacrifice in either the distortion or the rate of the algorithm.

3 THE PROPOSED METHOD

Generative Adversarial Networks (GANs) work by putting in competition two networks, a Generator, which produces fake data, and a Discriminator, which is trained for discerning fake data (obtained from the generator) from real data (picked from the training dataset). As discussed in Sect. 2, GANs have been used for super-resolution [19] or compression artifact removal, and although they are not specialized in maximizing signal-based metrics such as PSNR or SSIM about the enhanced image, the outputs are generally perceived as better quality than those of architectures that do not use perceptual-driven losses.

3.1 The network architecture

*UNet*s were firstly proposed in [28] for fast and semantic segmentation of biomedical images. The idea was modelling a fully convolutional neural network capable of per-pixel classification or, more in general, to address *image-to-image* tasks. The UNet architecture can be subdivided in three section:

- *Encoder*: the input is first progressively encoded to smaller spatial dimension but deeper channel dimension, to extract higher-level features.
- *Decoder*: it follows the inverse path, transforming the encoded image progressively upsampling by enlarging the spatial resolution and compressing the channel dimension.
- *Skip connections*: features at the same depth level are concatenated channel-wise, creating a direct connection between the encoder and the decoder, enabling a better information flow at training time, and providing also a better context to the decoder, *i.e.* keeping lower lever features that could be lost along the contracting path.

We selected UNet as the base for our image-enhancement method not only because of the success of the architecture in other image-to-image tasks beyond semantic segmentation, but also for engineering purposes. The network processes the input at multiple (lower) scales with respect to the original size, enabling an improved feature detection, but also implying that this processing will occur at smaller computational cost, since the convolution operation complexity is squarely dependent to the spatial resolution (and linearly to the channel depth). In fact, other more complex variants have been proposed [39], but we base our work on the original version, since it is simpler and more suited for our real-time constraint that is particularly stringent when operating on high resolution frames.

Our proposed SR-UNet, an adaptation of UNet for Super-Resolution and compression artifact removal, is based on a series of modification: starting from a fixed number of filters $\frac{F}{2}$, at each level of depth we double the number of filters until it reaches $F = 64$, meaning that from the second to the fourth blocks, the number of filters is limited to F and does not increase. This is motivated by the fact that typically SR models does not need a huge number of filters (differently from more abstract tasks *e.g.* classification). This allowed to reduce the total number of parameters (see Tab. 1), and stacking more Conv-ReLU blocks, mostly compensating the lower number of parameters. The difference between our architecture and the one proposed in [12] are many: since we focused on producing a fast model for video SR and not for image SR, the model only processes the features from the low-resolution space to below, and also we adopted different residual layers and upsampling techniques. We also will see that we both train our model for robustness, but we achieve this in different way: [12] randomly blur the training images, we encode the videos obtaining a series of artifacts, which could even introduce some at high-frequency.

To perform the upscaling we employ pixel-shuffle (also known as sub-pixel convolutional layer [30]), since it is the fastest upsample layer available: it comprises a depth-compression of the output tensor into 12-channels via-convolution operation, and then these features are reshuffled into an RGB image but at double resolution. Alternatives, such as bilinear upsample with convolution, the transposed convolution, or even the reshuffling to an higher dimension with same depth (as in [19]), would add an exaggerated overhead, since they would work in the high-resolution space. We also added a direct residual connection between the input image and the high-resolution output as in Eq. 1.

$$x^{SR} = Hard\tanh(U(x^{LR}) + upsample(x^{LR})) \tag{1}$$

where x^{SR} is the super-resolved output, x^{LR} is the low-resolution input, U is the convolutional SR-network, $upsample$ is an upsample filter such as bilinear or bicubic interpolation, and $Hard\tanh$ is for clipping the output between the interval $[-1, 1]$.

Modelling the problem as producing a residual on the top of the upsampled image is particularly convenient. This forces the model to focus on the high frequency patterns sharpening edges or increasing texture details, since the low frequency patterns are still from the upsampled image. Furthermore, a faster convergence of the training process is ensured.

Table 1: Parameter comparison between UNet and the proposed SR-UNet.

Architecture	# parameters	Binary size (MB)
SR-UNet (ours)	**740,975**	**2.96**
UNet	2,164,911	8.66

Fig. 2 shows the architecture of the proposed SR-UNet. As shown in Tab. 1 it has a reduced number of parameters vs. the standard UNet and its memory footprint is much smaller.

One further modification aimed to increase model capacity without adding computational cost is how we modelled the in-block skip connections. Skip connections are known to provide many benefits during training, but cost both in terms of memory and processing time, making them less attractive during inference phase. However, in DiracNets [37] and RepVGG [5] has been proposed a structural reparameterization for merging the skip connections (and also batch-normalizations) into one single convolutional 3 × 3 layer. The formula in Eq. 2 represents the basic block employed in our model.

$$x' = ReLU(\mathbf{W}_{3\times3} * x + \mathbf{W}_{1\times1} * x + x) \tag{2}$$

where x' is the output tensor, x is the input, $\mathbf{W}_{n\times n}$ are the weights of a convolutional layer with kernel size $n \times n$. For simplicity, biases are omitted. The arguments of the non-linear function can be easily refactored into one single 3 × 3 layer, which filters are computed as in Eq. 3:

$$\hat{\mathbf{W}} = \mathbf{W}_{3\times3} + \mathbf{W}^{pad}_{1\times1} + diag(\mathbf{Id})_{3\times3} \tag{3}$$

where $\mathbf{W}_{3\times3}$ are the 3 × 3 filters as before; to transform a layer with 1 × 1 kernels into 3 × 3, it is enough to add a zero-padding around the filters, and the identity skip connection can be easily modelled as a layer containing diagonal identity. After reshaping the skip-connections, the layers can be refactored via weight-sum.

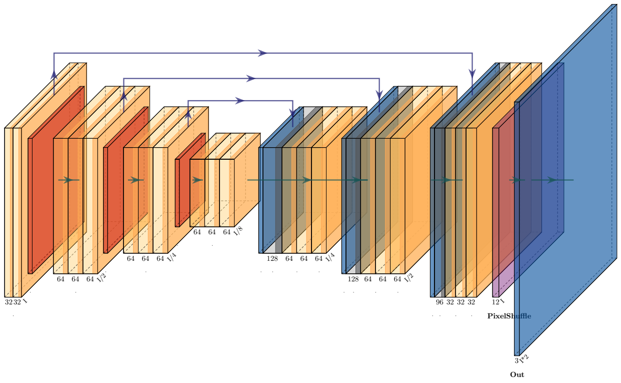

Figure 2: Architecture of the proposed SR-UNet, with PixelShuffle upsampling. The lower parameters count w.r.t. the standard UNet is compensated with more layers stacked into each block.

3.2 The proposed loss

For training our models, we employed a multiple loss which comprises a weighted sum of LPIPS loss [38] for perceptual similarity, L1 loss as signal based loss and the adversarial loss. The LPIPS loss is intended as an improvement of the [16] perceptual loss.

The overall loss function is:

$$LPIPS(y, \hat{x}) - SSIM(y, \hat{x}) - \lambda \log(D(\hat{x})) \tag{4}$$

where y is the high-resolution ground-truth, $\hat{x}$ is the generated image, λ is a real parameter equal to 10^{-3}. The LPIPS backbone is the VGG16 network (thus comparable to the common Perceptual Loss). We can consider both LPIPS and SSIM losses equally weighted with 1. We opted for equally weighting LPIPS and SSIM losses, to obtain a good balance between signal reconstruction and perceptual quality (*i.e.* intended as similarity between the distribution of the high-frequency patterns). We obtained these weights investigating different combinations of weights for the loss components, and evaluating their impact in terms of LPIPS and VMAF (for perceptual quality) and SSIM (for signal-based quality) scores. The λ weight for the Adversarial loss is experimentally standard and stable in literature. Indeed, we tried changing this value, with $2 * 10^{-3}$ and $5 * 10^{-3}$, but the result was to drive away the network from the convergence, making harder the entire training process.

3.3 Training details

For training our models, we employed the Adam optimizer with learning rate 10^{-4}, and trained the model for 80 epochs with the batch composed of 32 image patches (randomly sampled from the train set), which was large enough for stabilizing the entire training; each epoch consists of 1280 iterations, thus the model weights received slightly more than 100,000 updates. The train patch size was 96 × 96, and each patch was obtained from one random-crop per image. The only data-augmentation strategy we applied was horizontal reflection. We avoided any warping or rotation transform for maintaining the consistency with the real frames encoded in H.265 that do not suffer from such distortions. Training a model took about 24 hours on a single NVIDIA Titan Xp.

We trained the proposed model using the *BVI-DVC* dataset [22], a dataset designed for deep video compression tasks. The dataset is made of 200 frame sequences, truncated at the 64^{th} frame regardless of the frame-rate, that ranges from 25 to 120. The sequences include a variety of content, from natural scenes to man-made objects and city scenes, as shown in Fig. 3. The sequence resolution is 2160p, but downscaled versions have been created by the authors of the dataset, with resolutions of 1080p, 540p and 270p, leading to a total of 800 sequences and a total of 51,200 frames. The original BVI-DVC dataset constitutes the ground-truth for our training dataset. The train set is randomly split for the 80% for training and 20% for validating the models. To generate our input data, we compressed each sequence with the H.265 codec with Constant Rate Factor (CRF) 23 at half of the resolution. We fixed the CRF in the attempt to avoid the mode-collapse issue described in [9]. However, it should be kept in mind that even a fixed CRF does not imply that all the frames have the same quality: *e.g.* frames presenting motion will present lower quality than steady ones. Since we want to apply super-resolution on compressed videos (thus perform both Artifact Reduction and Super Resolution), it is fundamental to train the models on the compressed videos rather than just on the downscaled version of the HQ videos. Training the model only for SR would cause the model to fail to detect the features to super-resolve, or to even enlarge the compression artifacts and reducing the overall quality.

Figure 3: Example frames from the BVI-DVC dataset used to train the proposed SR-UNet.

4 EXPERIMENTAL RESULTS

4.1 Dataset

Our test set is composed from clips downloaded from 14 non-compressed clips of the *Derf's Collection* [6], that is commonly used to evaluate video coding and streaming [13], super resolution [15], compression artifact reduction [11] and visual quality improvement tasks [4]. The clips were compressed from 1080p to 540p with the H.265 codec with CRF 23, preparing an analogue setting as the train set. The clips are: *Ducks take-off*, *Crowd run*, *Controlled burn*, *Aspen*, *Snow mountain*, *Touchdown pass*, *Station 2*, *Rush hour*, *Blue sky*, *Riverbed*, *Old town cross*, *Rush field*, *Into tree* and *Sun flower*.

4.2 Video quality metrics

To test the quality of our models we employed two signal based metrics and one perceptual based: Structural Similarity Index Measure (SSIM) [40], Video Multimethod Assessment Fusion (VMAF) [25, 27] and LPIPS with AlexNet backbone [38], which is reported to work better to evaluate the enhancement after compression and/or super resolution algorithms than using the VGG backbone. SSIM is an objective full-reference metric and its index varies between 0 and 1, where 1 indicates perfect structural similarity, while 0 indicates no structural similarity. VMAF is an objective full-reference metric, originally proposed by Netflix, that fuses several existing quality metrics and other features to predict video quality using a SVM-based regression to provide a single output score in the range of 0-100 per video frame; a score of 100 means that the quality is identical to the reference video. LPIPS is a perceptual full-reference metric that evaluates the distance between image patches; a higher score means that two patches are more different perceptually, a lower score means they are more similar.

4.3 Video quality improvement

In the first experiment we compare our SR-UNet with a UNet baseline to assess the performance of our proposed changes and, with a H.265+bicubic interpolation to assess the improvement of the methods. We compare our method also with two other competing approaches. In particular, both the base UNet, and a 6-layer ESPCN [30] network implement Rep-VGG residual layers calibrated for processing at the same frame-rate; the last competing state-of-the-art approach is an 8-layer SR-ResNet [19], which is much slower than the other architectures. All the models have been trained with the same methodology; frames are rescaled from 540p to 1080p (Full-HD).

Table 2 reports the results in terms of SSIM, LPIPS and VMAF, reporting also the frames per second processed by each method, as obtained on a NVIDIA GTX 1080Ti. We notice that our architecture largely improves the perceptual metric (LPIPS) over H.265, thanks to the compression artifacts reduction and to the increase of frequencies in the high-frequency spectrum, and the quality improvement is also notable by the increase in VMAF. The SSIM metric, although being more perception-oriented than PSNR, is still based on the original signal, thus it is somehow predictable how its score is reduced by the adversarial and perceptual-driven training, as reported also in [18]. SR-UNet obtains a large speed-up over the SR-ResNet while maintaining the same quality. This is obtained since our model optimizes residual layers, compresses the up-scaling layer, and removes non-useful batch-normalizations, exploiting the particular U-Net architecture. An example of the results is shown in Fig. 9.

In the second experiment we evaluate rate/distortion at different CRF values, comparing the results of using our SR-Unet to scale from a 720p (HD) source to 1080p (Full-HD), with that of the source

Table 2: Comparison between models performances. ↑ indicates that higher values are better, ↓ indicates that lower values are better. Best results are highlighted in bold, second best are underlined.

Architecture	SSIM ↑	LPIPS ↓	VMAF ↑	FPS ↑
SR-UNet (ours)	0.7190	**0.2067**	84.30	**46.1**
UNet	0.7273	0.2193	**85.32**	45.4
SR-ResNet-8 [19]	**0.7278**	0.2130	84.24	6.4
ESPCN-6 [30]	0.7159	0.2125	82.29	45.0
H.265 + bicubic	0.7209	0.2821	79.15	-

720p resolution and the target 1080p resolution. In this case, the source 720p is upscaled by SR-UNet to 1440p, then bicubic sampling is used to downscale to 1080p; this approach is similar in spirit to that of supersampling anti-aliasing (SSAA), used in computer graphics to improve the visual quality of renderings. To further reduce the size and computational cost of the network we reduce filters and layers by 1/4, resulting in a network size of only 1.1MB. Fig. 4 reports the distortion in terms of VMAF, while Fig. 5 reports distortion in terms of LPIPS. Table 3 reports a selection of visual quality metrics and bitrate for some CRF values. Observing the curves in Fig. 4 and 5 shows that using SR-UNet results in a visual quality that is similar to, or better than, that of the 1080p resolution but a much lower bitrate (20%-33% less bandwidth for the same quality). This is clearly visible in the table: the bitrate is the same of the 720p, since the network is applied as a filter before showing the frame to the user there's no change in the video stream that is transmitted, but visual quality in terms of VMAF and LPIPS is typically better. The table shows, again, that SSIM score is penalized by the generative approach of our method.

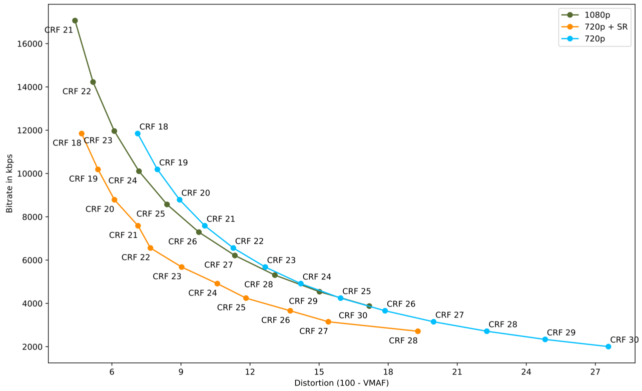

Figure 4: Rate/VMAF-distortion curve the varying of the compression CRF. We observe that after CRF 23-24, the 720p and 1080p curves merge. In this zone, the lower resolution of 720p is compensated from the lesser presence of compression artifacts, which instead may be visible on the original resolution.

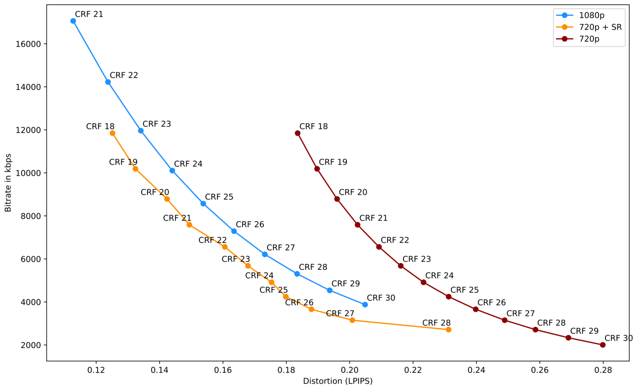

Figure 5: Rate/LPIPS-distortion curve at the varying of the compression CRF. The LPIPS distance, is way more sensitive to the higher frequencies sampled at higher resolution, however has a tendency to ignore some types of artifacts.

Table 3: Comparison of encoding methods performances. Each triplet compares (from top to bottom): the super-resolved video metrics, a 1080p reference of similar quality and the low-quality video fed to the network. After enhancement, the video has similar perceptual quality to the higher resolution one, while preserving the bitrate fo the low resolution.

Method	SSIM ↑	LPIPS ↓	VMAF ↑	Bitrate (*kb/s*) ↓
720p CRF 18 + SR-UNet	0.7611	0.1251	95.3133	**11,846**
1080p CRF 22	0.8181	0.1237	94.8185	14,225
720p CRF 18	0.7987	0.1835	92.8792	11,846
720p CRF 21 + SR-UNet	0.7711	0.1494	92.864	**7,585**
1080p CRF 24	0.8016	0.1440	92.824	10,105
720p CRF 21	0.7793	0.2024	89.964	7,585
720p CRF 23 + SR-UNet	0.7611	0.16790	90.9682	**5,678**
1080p CRF 26	0.7836	0.1634	90.2208	7,290
720p CRF 23	0.7639	0.2161	87.33	5,678
720p CRF 25 + SR-UNet	0.7402	0.1798	88.174	**4,243**
1080p CRF 28	0.7737	0.1834	86.920	5,306
720p CRF 25	0.7461	0.2312	84.065	4,243

In the third experiment we fine tune our SR-UNet for each specific video clip. The idea is that of specializing the SR-UNet for each video or part of a video that is streamed, sending the model weights to the receiver along with the H.265 stream. This approach is feasible for streaming services like Netflix or Prime Video, that already encode videos at different bitrates to account for different available bandwidths; similarly to the creation of multiple encoded streams, there's necessity to overfit the network only once. The advantage with respect to the previous method is we can obtain a better quality (compared to the previous method) for the about the same bitrate, maintaining the real-time requirements; the disadvantage is that this approach can not be applied to live streams. This is possible thanks to the reduced size of our SR-UNet, that can be compressed to 1.1MB. It is important to note that the reported bitrate includes the weights of the network. This approach based on overfitting the network is similar in spirit to [31], by van Rozendaal *et al.* , but differently from them, we do not employ a deep-compression model as codec, relying instead on the industry standard H.265. Also, we do not focus on minimizing a Rate-Distortion loss, but we train our models and assess the results with perceptual metrics and losses.

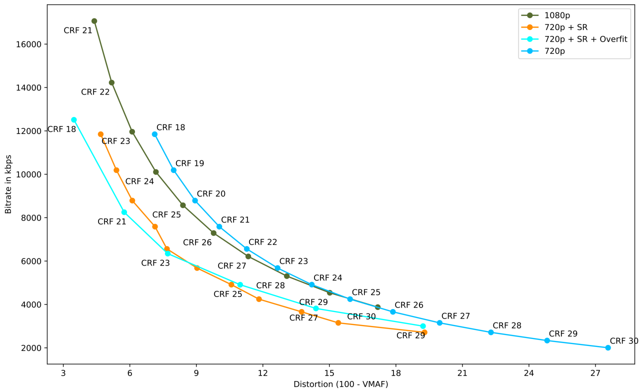

Figure 6: Rate/VMAF-distortion curve at the varying of the compression CRF, showing that before a certain threshold it is possible to further improve the rate/distortion curve.

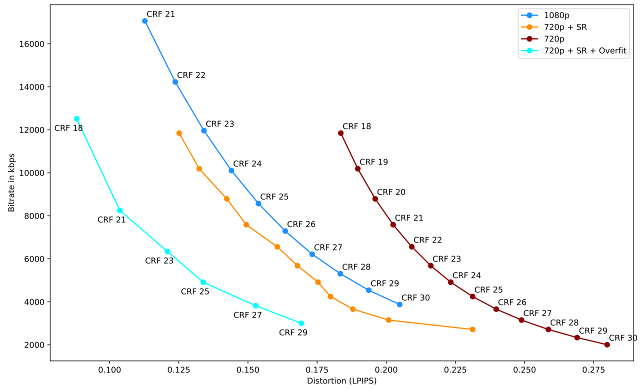

Figure 7: Rate/LPIPS-distortion curve for various techniques at the varying of the compression CRF. The LPIPS distance, is more sensitive to the lack of higher frequencies, however has a tendency to ignore some types of artifacts.

In the curves represented in Fig. 7 and 6 is reported a comparison between the overfit SR-UNet, the base SR-UNet and the source and target resolutions (720p and 1080p). The LPIPS distortion curve shows that overfitting is always beneficial in terms of quality and bitrate, while the VMAF distortion curve shows this effect is true up to CRF 23.

Tab. 4 shows a bitrate comparison between our proposed SR-UNet overfitting technique, the base SR-Unet and a target 1080p resolution, to achieve a specific quality. Both LPIPS and VMAF score are improved and bitrate is further reduced w.r.t. base SR-Unet. An example is shown in Fig. 10.

In the last experiment we conducted a subjective test based on the two-alternative forced choice (2AFC) methodology, using the *Versus* tool [32]. The proposed test included the inspection of 22 pairs of frames, structured as follows:

- 17 pairs were meant to validate the fidelity of the reconstruction. For this purpose, we first selected a pair composed of a 720p low resolution (LR) frame and a 1080p high resolution (HR) one, in a such way that after the SR process the two would have been very similar (according to the Fig. 4). Each couple was composed from the super-resolved (SR) frame (starting from the LR) and the related HR version. The frame were taken at various configurations (*e.g.* CRF 21/24, CRF 24/27, CRF 23/26) to generalize the results on several rate/distortion settings.
- 5 "trap" pairs were composed by the HR frame and a LR frame sampled as just explained.

Table 4: Comparison of various encoding methods performances. The bitrate related to the CRF 18 and 21 with overfit depends from the video length in seconds. The weights transmission overhead is already considered into the computation of the bitrate.

Method	SSIM ↑	LPIPS ↓	VMAF ↑	Bitrate (*kb/s*) ↓
720p CRF 21 + SR overfitted	0.7672	**0.1036**	**94.25**	**8,250**
720p CRF 20 + SR	0.7769	0.1423	93.88	8,786
1080p CRF 23	0.8099	0.1340	93.895	11,961
720p CRF 18 + SR overfitted	0.7840	**0.0880**	**96.52**	**12,511**
1080p CRF 20	0.8349	0.1012	96.27	20,617

The goal is to evaluate if HR/SR pairs are practically indistinguishable, and thus that our proposed reconstruction method is reasonably effective. The purpose of the "trap" pairs was in fact to validate that the user was actually able at least to discern the HR frames from the LR ones, and also this would have implicitly proven that the enhancement operation was effective. Conducting a 2AFC was chosen to force people to declare a preference for one of the high resolution (whether HR or SR). The presented frames were randomly cropped with a window size of 600x768, to avoid the automatic downscaling operated by the browser. Before starting the test, we recommended to use at least a Full HD screen of a reasonable display size (more than 20 inches) and to carefully observe the pictures. For each pair, the question asked was "Which image is the sharper?". 42 people participated in the survey.

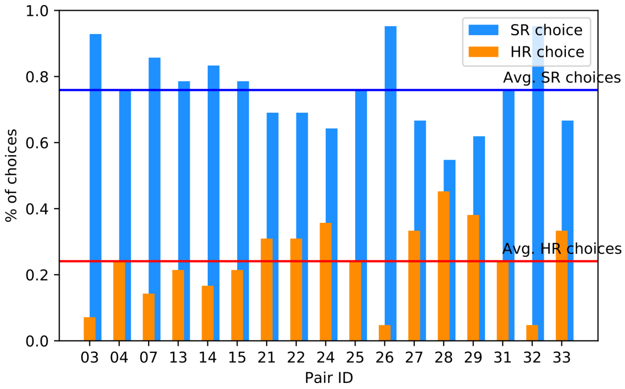

Figure 8: Results from the 2AFC subjective test, in percentage for each HR/SR pair.

The results, summarized in Fig. 8, show that the generated images were very competitive with the real one, and a number of participant in fact reported difficulties at choosing between the SR and the HR, and surprisingly there was a general attitude towards the SR generated images with 75.91% (vs. the 24.08% for the HR pictures)

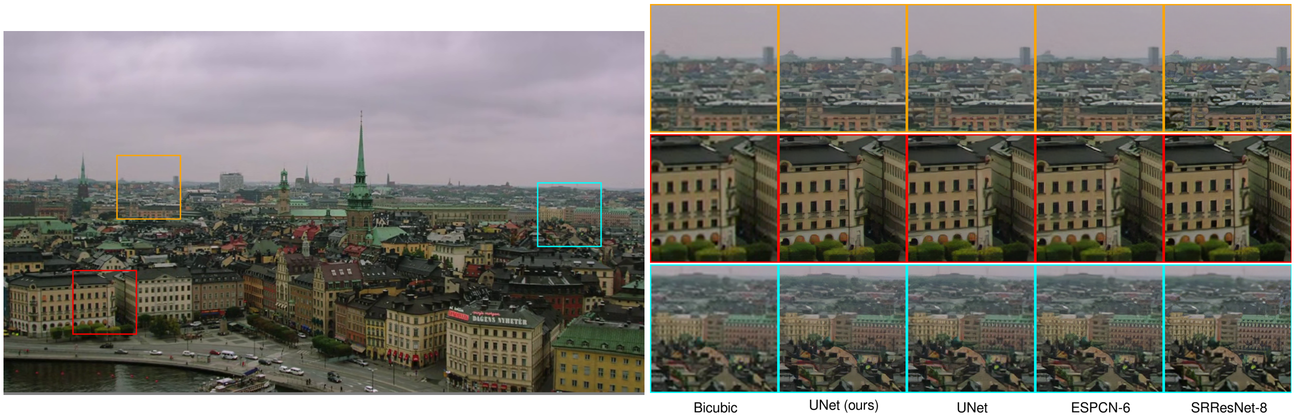

Figure 9: Qualitative comparison between models. (a) bicubic filter, (b) our SR-UNet, (c) UNet, (d) ESPCN [30], (e) SR-ResNet-8 [19]. From *Old Town Cross* clip.

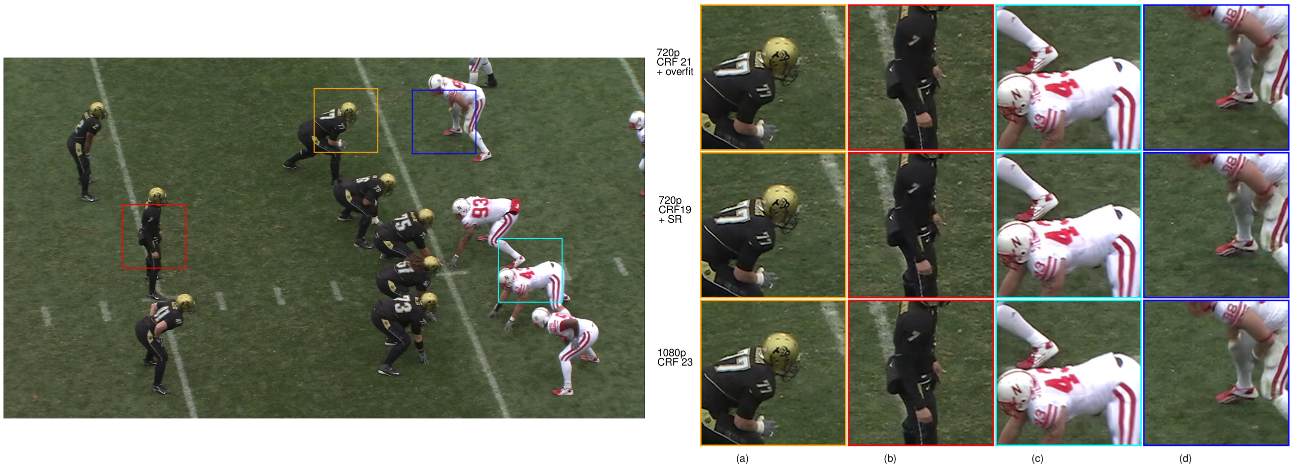

Figure 10: Qualitative comparison between encoded patches and upsampled. Row (1) 720p CRF 21 + overfitting SR, (2) 720p CRF 19 with Super Resolution, (3) 1080p H265 CRF 23. The three configurations have similar distortion factors, but different bitrate requirements.

of overall preferences; no one had problems at correctly choosing the HR image over the LR when the “trap" pairs occurred, so we also proved the perceptual effectiveness of the enhancement. Must be noticed that, given how we sampled the frames, also the slightly higher VMAF similarity of the SR frame (see Fig. 4 and Tab. 3) confirms the subjective results (and viceversa), while the LPIPS metric would suggest the opposite (Fig. 5).

4.3.1 Qualitative examples. In the following are shown a few examples of SR-UNet. Fig. 9 compares the proposed network w.r.t. a bicubic upsample baseline, a UNet baseline and two competing approaches [19, 30]. Fig. 10 shows a comparison of the fine tuned SR-Unet vs. the base SR-Unet and the 1080p baseline. The three configurations have similar visual quality, but using SR-UNet results in a bandwidth reduction (with respect to the 1080p version) of 20%, while the finetuned SR-Unet further reduces bandwidth requirements by 35%.

5 CONCLUSIONS

In this work we have presented a novel network architecture, called SR-UNet that can be used to perform super resolution and compression artifact removal in videos, thanks to its reduced computational cost; we have proposed also a loss that combines perceptual and signal-based losses, within a Generative Adversarial Network framework. The network can be used to improve the visual quality of videos compressed with H.265 codec, or to reduce the required bandwidth while maintaining a specified visual quality. The effectiveness has been demonstrated using both subjective and objective metrics, considering both signal-based scores (like VMAF) or perceptual ones like LPIPS. Its performance improves w.r.t. a UNet baseline and other competing state-of-the-art approaches.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 951911 - AI4Media.

REFERENCES

[1] Yochai Blau and Tomer Michaeli. 2019. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *Proc. of International Conference on Machine Learning (ICML).*

[2] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation. In *Proc. of IEEE Computer Vision and Pattern Recognition.*

[3] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. 2020. Learning temporal coherence via self-supervision for GAN-based video generation. *ACM Transactions on Graphics* 39, 4 (Jul 2020). https://doi.org/10.1145/3386569.3392457

[4] Valery Dewil, Jeremy Anger, Axel Davy, Thibaud Ehret, Gabriele Facciolo, and Pablo Arias. 2021. Self-Supervised Training for Blind Multi-Frame Video Denoising. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV).* 2724–2734.

[5] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. 2021. RepVGG: Making VGG-style ConvNets Great Again. arXiv:2101.03697 [cs.CV]

[6] Xiph.Org foundation. [n.d.]. Derf's collection. https://media.xiph.org/video/derf/. [Online; last time accessed 12-March-2021].

[7] Leonardo Galteri, Marco Bertini, Lorenzo Seidenari, Tiberio Uricchio, and Alberto Del Bimbo. 2020. Increasing Video Perceptual Quality with GANs and Semantic Coding. In *Proc. of ACM Multimedia (MM '20).* 862–870. https://doi.org/10.1145/3394171.3413508

[8] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. 2017. Deep Generative Adversarial Compression Artifact Removal. In *Proc. of International Conference on Computer Vision.*

[9] L. Galteri, L. Seidenari, M. Bertini, and A. Del Bimbo. 2019. Deep Universal Generative Adversarial Compression Artifact Removal. *IEEE Transactions on Multimedia* 21, 8 (2019), 2131–2145.

[10] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. 2019. Towards Real-Time Image Enhancement GANs. In *Proc. of International Conference on Analysis of Images and Patterns (CAIP).*

[11] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2019. Fast Video Quality Enhancement Using GANs. In *Proc. of ACM Multimedia (MM '19).* 1065–1067. https://doi.org/10.1145/3343031.3350592

[12] X. Hu, M. A. Naiel, A. Wong, M. Lamm, and P. Fieguth. 2019. RUNet: A Robust UNet Architecture for Image Super-Resolution. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* 505–507. https://doi.org/10.1109/CVPRW.2019.00073

[13] Q. Huang, H. Wang, S. C. Lim, H. Y. Kim, S. Y. Jeong, and C. . J. Kuo. 2017. Measure and Prediction of HEVC Perceptually Lossy/Lossless Boundary QP Values. In *2017 Data Compression Conference (DCC).* 42–51. https://doi.org/10.1109/DCC.2017.17

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2016. Spatial Transformer Networks. arXiv:1506.02025 [cs.CV]

[15] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. 2018. Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of European Conference on Computer Vision.*

[17] Takuhiro Kaneko and Tatsuya Harada. 2020. Blur, Noise, and Compression Robust Generative Adversarial Networks. arXiv:2003.07849 [cs.CV]

[18] H. Ko, D. Y. Lee, S. Cho, and A. C. Bovik. 2020. Quality Prediction on Deep Generative Images. *IEEE Transactions on Image Processing* 29 (2020), 5964–5979.

[19] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proc. of IEEE Computer Vision and Pattern Recognition*, Vol. abs/1609.04802.

[20] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. arXiv:1609.04802 [cs.CV]

[21] B. Li, J. Liang, and Y. Wang. 2019. Compression Artifact Removal with Stacked Multi-Context Channel-Wise Attention Network. In *Proc. of IEEE International Conference on Image Processing (ICIP).* 3601–3605.

[22] Di Ma, Fan Zhang, and David R. Bull. 2020. BVI-DVC: A Training Database for Deep Video Compression. arXiv:2003.13552 [eess.IV]

[23] Danial Maleki, Soheila Nadalian, Mohammad Mahdi Derakhshani, and Mohammad Amin Sadeghi. 2018. BlockCNN: A Deep Network for Artifact Removal and Image Compression.. In *Proc. of CVPR Workshops.* 2555–2558.

[24] Filippo Mameli, Marco Bertini, Leonardo Galteri, and Alberto Del Bimbo. 2020. Image and Video Restoration and Compression Artefact Removal Using a NoGAN Approach. In *Proc. of ACM Multimedia (MM '20).* 4539–4541. https://doi.org/10.1145/3394171.3414451

[25] Netflix. [n.d.]. VMAF Github repository. https://github.com/Netflix/vmaf. [Online; last time accessed 12-March-2021].

[26] R. Pourreza, A. Ghodrati, and A. Habibian. 2019. Recognizing Compressed Videos: Challenges and Promises. In *Proc. of IEEE International Conference on Computer Vision Workshop (ICCV-W).* 999–1007.

[27] R. Rassool. 2017. VMAF reproducibility: Validating a perceptual practical video quality metric. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB).* 1–2. https://doi.org/10.1109/BMSB.2017.7986143

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proc. of International Conference on Medical image computing and computer-assisted intervention.*

[29] Mehdi S. M. Sajjadi, Raviteja Vemulapalli, and Matthew Brown. 2018. Frame-Recurrent Video Super-Resolution. In *Proc. of IEEE Computer Vision and Pattern Recognition.* arXiv:1801.04590 http://arxiv.org/abs/1801.04590

[30] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *Proc. of IEEE Computer Vision and Pattern Recognition.* arXiv:1609.05158 [cs.CV]

[31] Ties van Rozendaal, Iris A. M. Huijben, and Taco S. Cohen. 2021. Overfitting for Fun and Profit: Instance-Adaptive Data Compression. arXiv:2101.08687 [cs.LG]

[32] Jenny Vuong, Sandeep Kaur, Julian Heinrich, Bosco K. Ho, Christopher J. Hammang, Benedetta F. Baldi, and Seán I. O'Donoghue. 2018. Versus—A tool for evaluating visualizations and image quality using a 2AFC methodology. *Visual Informatics* 2, 4 (2018), 225–234. https://doi.org/10.1016/j.visinf.2018.12.003

[33] T. Wang, J. He, S. Xiong, P. Karn, and X. He. 2020. Visual Perception Enhancement for HEVC Compressed Video Using a Generative Adversarial Network. In *Proc. of International Conference on UK-China Emerging Technologies (UCET).* 1–4.

[34] Maarten Wijnants, Sven Coppers, Gustavo Rovelo Ruiz, Peter Quax, and Wim Lamotte. 2019. Talking Video Heads: Saving Streaming Bitrate by Adaptively Applying Object-Based Video Principles to Interview-like Footage. In *Proc. of ACM Multimedia (MM '19).* 2449–2458. https://doi.org/10.1145/3343031.3351045

[35] Yi Xu, Longwen Gao, Kai Tian, Shuigeng Zhou, and Huyang Sun. 2019. Non-Local ConvLSTM for Video Compression Artifact Reduction. In *Proc. of IEEE International Conference on Computer Vision.* arXiv:1910.12286 [eess.IV]

[36] S. Yu, B. Chen, Y. Xu, W. Chen, Z. Chen, and T. Zhao. 2019. HEVC Compression Artifact Reduction with Generative Adversarial Networks. In *Proc. of International Conference on Wireless Communications and Signal Processing (WCSP).* 1–6.

[37] Sergey Zagoruyko and Nikos Komodakis. 2018. DiracNets: Training Very Deep Neural Networks Without Skip-Connections. arXiv:1706.00388 [cs.CV]

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proc. of IEEE Computer Vision and Pattern Recognition.*

[39] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Proc. of Deep learning in medical image analysis and multimodal learning for clinical decision support (DLMIA, ML-CDS).* arXiv:1807.10165 [cs.CV]

[40] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861