Recognizing Compressed Videos: Challenges and Promises

Reza Pourreza[†], Amir Ghodrati[‡], Amirhossein Habibian[‡]
Qualcomm AI Research*
[†]San Diego CA, USA, [‡]Amsterdam, Netherlands
{pourreza, ghodrati, ahabibia}@qti.qualcomm.com

Abstract

This paper studies the effect of quality degradation, caused by lossy video compression, on video recognition. We investigate how the state of the art video enhancement restores the video quality needed for an effective video recognition. Furthermore, we study the impact of various enhancement objectives, namely pixel-level, feature-level, and adversarial, on action recognition performance. Our experiments demonstrate that the models trained on pixel-level loss perform well in terms of visual quality but they hurt the accuracy of action recognition due to over smoothing discriminative features. On the other hand, models trained on perceptual and adversarial loss types not only generate better perceptual quality but also further improve the action recognition performance.

1. Introduction

A huge amount of videos being captured everyday by low power devices *i.e*. IP cameras, dash-cams and drones. High-level understanding of these videos, *i.e*. by classification, detection and segmentation, involve lots of calculations to run deep neural networks, which is often beyond the computational capabilities of low power devices. Hence many of these devices offload the inference to the cloud, which requires transmitting of video streams from sensors to compute servers.

Streaming raw video data would require massive bandwidth and lead to a long response time. Therefore, lossy video compression techniques such as AVC and HEVC are used to greatly reduce the data size before transmission. At encoding time, lossy compression transforms each video into a bit-stream by factoring out the spatio-temporal redundancies and quantizing the signal. The resulting bitstream, which is often much smaller than the raw data, is transmitted to the computing server for inference. At decoding time, video pixels are reconstructed from the bit-stream before being fed into a video recognition network.

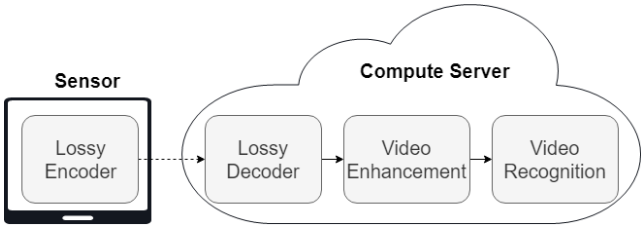


Figure 1: Recognizing compressed videos. Video streams from sensors are compressed then transmitted to a compute server for inference. Decompressed videos are enhanced before being fed into the video recognition network.

Lossy video compression, especially on lower bitrates, can lead to severe artifacts such as blocking and color changes. Decompressed videos might look very different from the uncompressed videos on which the video recognition network is trained. The domain shift between uncompressed and compressed videos, as train and test examples, degrades the recognition performance. Domain adaptation techniques [38] can alleviate this problem, but they require re-training of the recognition model, which is not always possible, *e.g*. if we don't have access to the original training pipeline. Video enhancement can be used as a pre-processing step to restore compressed videos before recognition. A lot of progress has been made on video enhancement by using deep convolutional networks to map noisy decompressed videos into their artifact-free counterparts [6, 33, 9, 40, 21, 23, 1]. These methods aim for generating visually appealing results that look good to the human eye, which may not be necessarily optimal for the recognition network.

This paper studies the challenges involved in recognizing compressed videos using off-the-shelf video recognition networks relevant for remote inference use cases, as illustrated in Figure 1. In particular, we empirically study the following research questions: *i)* What is the impact of lossy compression on video recognition? *ii)* Can video en-

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

hancement compensate the compression effect on recognition? *iii)* What is the relation between visual quality of enhanced videos and recognition performance?

The rest of the paper is organized as follows: Section 2 and Section 3 review the state of the art in video enhancement and recognition as used in this paper for experiments. Section 4 specifies dataset, evaluation metrics, and implementation details of our models. Section 5 discusses the experiments and results studying the three aforementioned research questions. Section 6 concludes the paper.

2. Video Enhancement

Similar to many image and video transformation tasks, *e.g*. style transfer [11, 12], super-resolution [32, 3, 26], and inpainting [39], fully convolutional neural networks are the state of the art in quality enhancement of compressed videos [37, 8]. Networks are trained on pairs of compressed videos (as input) and their corresponding uncompressed videos (as ground-truth), to enhance videos by removing their compression artifacts. Most video enhancement methods can be categorized based on their architecture and loss function.

2.1. Architecture

Enhancement networks can be divided into three architectures: *frame-by-frame*, *multi-frame*, and *frame-recurrent*. The frame-by-frame architecture, originally proposed for still images, applies a 2D CNN independently to each frame [22, 31, 36]. This architecture is simple, but cannot exploit the temporal correlation between frames.

Multi-frame architectures enhance each frame, using a batch of previous and next frames as context [14, 35, 37, 8, 13, 17]. They often rely on optical flow estimation and warping to align the frames. Although multi-frame methods are generally more effective than frame-by-frame ones, they suffer from two major drawbacks: *i)* There is a lot of computational redundancy. They process frames in a moving-window fashion in the temporal direction where every frame is processed multiple times. *ii)* Multi-frame architectures are able to utilize the temporal correlation within the batch, but the batch and consequently the temporal memory is usually limited to only a couple of frames.

The third approach, frame-recurrent architecture, has been proposed to address aforementioned drawbacks [24, 26, 3]. It employs recurrent structures to capture the spatiotemporal information across frames. The convolutional layers capture the spatial information within the frames while the recurrent structure captures the inter-frame information. As a result, the frame-recurrent architecture has a long temporal memory and processes each frame only once. While this method is proposed for the task of video super-resolution, in this paper we employ the frame-recurrent architecture for the first time to enhance the quality of compressed videos. We conducted an ablation study and compared the performance of three architectures on enhancing compressed videos while keeping the complexities of the networks about the same. Our study revealed that the frame-recurrent architecture in comparison to the multi-frame architecture performs better in terms of spatial quality and performs on-par in terms of temporal coherence while being a lot more computationally efficient. Hence, the frame-recurrent structure was used in our experiments.

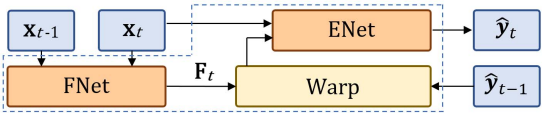


Figure 2: Frame-recurrent architecture for one time step. Blocks in orange are trainable while the yellow blocks are not. Blue blocks indicate input/output images.

The block-diagram of a recurrent structure for video enhancement is shown in Figure 2. $\mathbf{x}_t \in [0,1]^{H\times W\times C}$ and $\hat{\mathbf{y}}_t = \mathrm{G}(\mathbf{x}_t) \in [0,1]^{H\times W\times C}$ denote the compressed frame and the corresponding enhanced frame at time t and G represents the whole enhancement network shown in dashed-lines in Figure 2. At each time step, the previously enhanced frame $\hat{\mathbf{y}}_{t-1}$ is aligned to the current noisy frame $\mathbf{x}_t$ in Warp block and fed to the enhancement block ENet together with $\mathbf{x}_t$ to predict the current enhanced frame $\hat{\mathbf{y}}_t$. Alignment is done by first estimating dense optical flow $\mathbf{F}_t$ using the flow network FNet, which takes as input $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, and then warping $\hat{\mathbf{y}}_{t-1}$ using $\mathbf{F}_t$ in Warp block. Warp block, which is based on spatial transformer [15], shifts the pixels of $\hat{\mathbf{y}}_{t-1}$ in both horizontal and vertical directions based on $\mathbf{F}_t$ using bi-linear interpolation. The above steps are summarized in Eq. 1.

$$\begin{aligned} \mathbf{F}_t &= \mathrm{FNet}(\mathbf{x}_t, \mathbf{x}_{t-1}) \in [-1,1]^{H\times W\times 2} \\ \hat{\mathbf{y}}_t &= \mathrm{ENet}(\mathbf{x}_t, \mathrm{Warp}(\hat{\mathbf{y}}_{t-1}, \mathbf{F}_t))) \in [0,1]^{H\times W\times C} \end{aligned} \quad (1)$$

The optical flow estimation and warping are crucial to mitigate the miss-alignment between consecutive frames caused by scene dynamics and camera movements. For training, the network is unrolled through time for multiple steps due to the recurrent structure. We unroll the network for 10 time steps chosen by performance and memory.

2.2. Loss function

Enhancement networks rely on three types of loss functions to compare an enhanced frame $\hat{\mathbf{y}}_t$ to an uncompressed ground-truth frame $\mathbf{y}_t$: *pixel-level*, *perceptual*, and *adversarial* loss.

Pixel-level loss compares two frames based on their individual pixel values using a norm distance. Pixel-wise loss functions, ℓ_2 in this work as defined in Eq 2, enjoy stable

training and are widely used in the literature [37, 8]. However if two images are perceptually similar but different in pixel values, *e.g*. shifted by one pixel, then their pixel-level loss functions could be high. Moreover, it is well known that pixel-level loss often yields smooth enhancements, where texture information might be lost.

$$\ell_{pixel} = \mathbb{E}[\|\hat{\mathbf{y}} - \mathbf{y}\|_2^2] \tag{2}$$

Perceptual loss (ℓ_{perc}) compares two frames based on high-level features computed by a pre-trained network. Rather than encouraging the pixels to be similar, this loss encourages the enhanced and uncompressed frames to have similar activation maps in a network ϕ trained for a recognition task (*e.g*. ImageNet classification [29]). More specifically, perceptual loss is defined as the difference between $\phi(\mathbf{y}_t)$ and $\phi(\hat{\mathbf{y}}_t)$ using a norm distance, ℓ_1 here as defined in Eq. 3. Perceptual loss often generates perceptually more convincing enhancements.

$$\ell_{perc} = \mathbb{E}[\|\phi(\hat{\mathbf{y}}) - \phi(\mathbf{y})\|_1] \tag{3}$$

Adversarial loss (ℓ_{adv}) compares two frames using a discriminator network learned to distinguish the enhanced and uncompressed frames. Similar to generative adversarial networks (GAN), adversarial loss encourages the enhancement network to generate frames which reside on the manifold of uncompressed frames by fooling the discriminator. In this work, we use the adversarial loss term proposed in RaGAN [16] that takes two inputs and determines which one looks more realistic. The architecture of our discriminator D is borrowed from ESRGAN [32]. D loss and ℓ_{adv} are shown in Eq 4.

$$\begin{aligned} \ell_{\mathrm{D}} &= -\mathbb{E}_{\mathbf{y}}[\log(1 - \mathrm{D}(\mathbf{y}, \mathrm{G}(\mathbf{x})))] - \mathbb{E}_{\mathbf{x}}[\log(\mathrm{D}(\mathrm{G}(\mathbf{x}), \mathbf{y}))] \\ \ell_{adv} &= -\mathbb{E}_{\mathbf{y}}[\log(\mathrm{D}(\mathbf{y}, \mathrm{G}(\mathbf{x})))] - \mathbb{E}_{\mathbf{x}}[\log(1 - \mathrm{D}(\mathrm{G}(\mathbf{x}), \mathbf{y}))] \end{aligned} \tag{4}$$

Adversarial loss generates visually appealing enhancements by recovering the textures lost thought the compression. However it may cause training instabilities, which require careful selection of learning hyper parameters. So, it is often used in a combination with pixel-wise and perceptual losses as a weighted sum [20, 7, 32]. Following ESRGAN, we define ℓ_{GAN} as a weighted combination of the above loss types as denoted in Eq 5 where $\beta = 0.01$ and $\gamma = 0.005$.

$$\ell_{GAN} = \ell_{perc} + \beta\mathbb{E}[\|\hat{\mathbf{y}} - \mathbf{y}\|_1] + \gamma\ell_{adv} \tag{5}$$

3. Video Recognition

Modern action recognition models are based on 2D and 3D CNN architectures [28, 2]. The seminal work two-stream networks [28] employ two CNNs to model spatial and temporal features disjointedly. Each stream is a 2D CNN separately trained on RGB and optical flow frames. The spatial CNN learns to recognize the actions from its appearance such as foreground and background objects, while the temporal CNN classifies the actions based on motion clues. Despite its simplicity this model outperforms various more complicated alternatives that employ recurrent layers for temporal modeling [5]. Recently 3D CNNs with spatio-temporal kernels has been effectively applied to action classification [10]. These models usually have an immense number of parameters, because of their cubic kernels, whose training has been feasible only recently with the availability of huge video collections such as Kinetics [18].

Since in this work we are investigating the effect of compression and enhancement on recognition accuracy, we train recognition networks on uncompressed videos only and then study how they perform on compressed and enhanced videos.

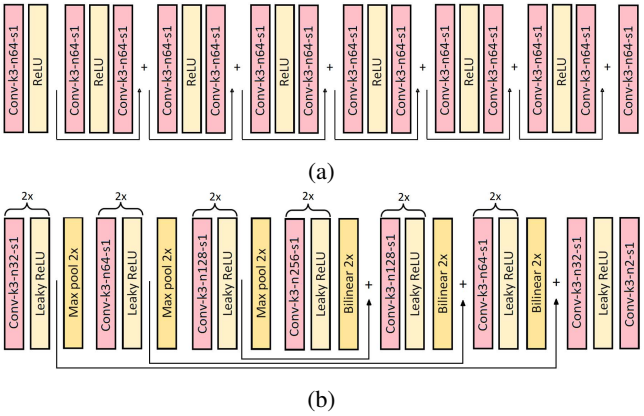


Figure 3: (a): architecture of ENet, (b): architecture of FNet. k, n, and s, denote kernel size, number of filters, and stride.

4. Experimental Setup

4.1. Dataset

Kinetics-600 [18] consists of approximately 500k video clips from YouTube with an average length of 10 seconds, which we use for pre-training of the enhancement network. Since the videos of this dataset are already compressed and suffer from compression artifacts, it is not directly applicable for enhancement and cannot serve as ground-truth. Following [35], we select high-quality 1080p videos and downsample them to 640×360. The down-sampled videos are treated as ground-truth uncompressed videos. This leads to a total of 32,000 video clips divided into subsets with 28,000 and 4,000 clips for training and validation. The first 10 frames of each video are used for training of the enhancement network given that the recurrent structure is unfolded

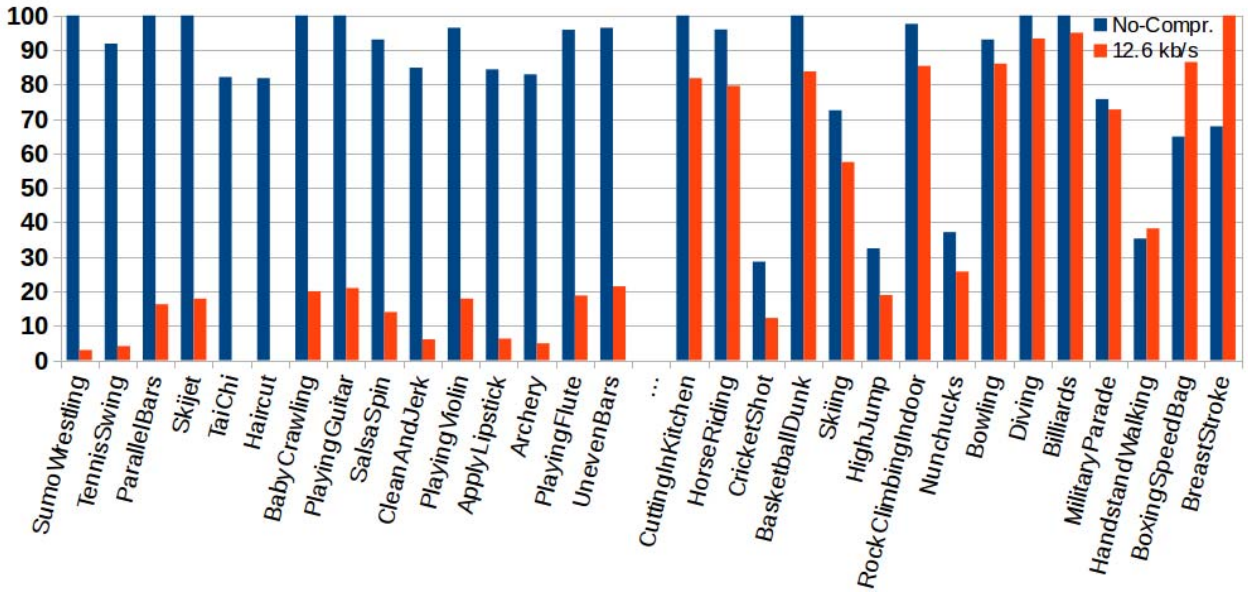


Figure 4: Per class accuracy of Resnet-S model for uncompressed frames and 12.6 kb/s compressed frames of the action recognition dataset. The categories are sorted by accuracy degradation between uncompressed and compressed data. In 98 out of 101 classes recognition performance is higher for uncompressed frames. On three action categories of *HandstandWalking, BoxingSpeedBag*, and *BreastStroke*, the performance of compressed frames is higher than uncompressed frames. After looking into these cases, we observed that many videos are assigned to these categories, leading to having highest false positive rate for these classes.

10 times.

Action recognition dataset: we use a widely used dataset for action recognition containing 13,320 videos from 101 action classes. We follow the standard partitioning (split-1) and use 9,537 videos for training and 3,783 videos for testing. We use this dataset for fine-tuning and evaluation of the enhancement network as well as the evaluation of the video recognition network. For the video enhancement task, the original video frames are used as ground truth and similar to the Kinetics-600 dataset, the first 10 frames of each video are used for training.

4.2. Evaluation metrics

Rate-Accuracy. Inspired by rate-distortion curves we introduce rate-accuracy to measure how recognition accuracy responds to reducing the bitrate by further compressing videos. Recognition accuracy and bitrate are measured in terms of top-1 accuracy and bits-per-second, respectively.

Peak Signal to Noise Ratio. PSNR is a common metric to measure the quality of enhanced images and video. It is defined as a normalized ℓ_2 distance between the enhanced and ground-truth frames in a logarithmic decibel scale. PSNR measure the video quality only in the spatial domain ignoring temporal consistencies.

Temporal Consistency. TC measure the temporal consistency between frames in an enhanced video [24]. It is measured as the PSNR between the current and the warped previous frame averaged over all frames. The higher TC score an enhanced video has, the less flickering exist between the frames.

Learned Perceptual Image Patch Similarity [41]. LPIPS is an image quality metric proposed to better reflect the human perception. It is calculated over high level features from a pre-trained CNN. Subjective studies confirm that LPIPS better reflects human image quality measures compared to classical metrics such as PSNR and MS-SSIM [34].

4.3. Implementation details

Video Compression We compress videos using the latest High Efficiency Video Coding (HEVC) standard. We use the *ffmpeg* implementation [4] with default settings. We control the compression rate by Constant Rate Factor (CRF) parameter ranging from 20 to 50 in steps of 5 that correspond to bitrates 364.9, 192.8, 96.6, 48.1, 24.8, 15.2, and 12.6 kb/s, respectively.

Video enhancement network The architectures of ENet and FNet are shown in Figure 3. ENet is a stack of residual layers. FNet is a U-net [25] with 3 downsampling/upsampling steps with skip connections. The U-net architecture makes the network capture both small and large displacements in the optical flow accurately. We train the

Bitrate	Resnet-S				Resnet-ST			
	Compressed	ℓ_{pixel}	ℓ_{GAN}	ℓ_{perc}	Compressed	ℓ_{pixel}	ℓ_{GAN}	ℓ_{perc}
Uncompressed	82.3	82.3	82.3	82.3	87.9	87.9	87.9	87.9
364.9	80.6	79.3	**81.4**	80.8	**87.7**	87.4	87.6	87.6
192.8	79.3	76.2	**80.4**	79.8	87.4	87.2	87.4	**87.6**
96.6	77.2	70.1	**79.7**	78.2	**87.4**	86.8	**87.4**	87.3
48.1	75.3	66.8	**77.0**	75.4	**86.8**	86.2	86.6	**86.8**
24.8	71.2	57.1	**72.1**	70.8	**85.6**	84.1	85.2	85.3
15.2	55.9	43.7	54.9	**56.5**	79.0	75.9	**80.4**	79.3
12.6	36.8	32.2	35.0	**41.7**	68.1	64.1	69.6	**69.8**

Table 1: Recognition results

enhancement network using a batch size of 16, where each sample is a sequence of 96×96 patches randomly cropped over a video clip. We use the Adam optimizer [19] with initial learning rate of 10^{-4}. We first train the enhancement network on Kinetics-600 dataset for 150 epochs using ℓ_2 loss. We train individual networks for each of the above bitrates. Then, each pre-trained network is fine-tuned on the action recognition dataset using ℓ_{pixel}, ℓ_{perc}, and ℓ_{GAN} separately. As Perceptual loss, we use the features from the last convolutional layer of a VGG-19 [29] before activation, which could provide stronger supervision for brightness consistency and texture recovery [32]. It is also worth mentioning that we only enhance the luminance channels of the videos while the chrominance channels remain unchanged.

Video classification network Our 2D action recognition network is a ResNet-101 pre-trained on ImageNet then fine-tuned on RGB frames from the action recognition dataset training set as in [28]. Following [30], we calculate the loss over a video-level prediction averaged over 3 randomly selected frames per video. The network is trained using a mini batch of size 25, using a SGD optimizer with an initial learning rate of 5×10^{-4}. We evaluate the network on 224×224 center crops from 25 frames uniformly sampled per video.

Our 3D action recognition network is a ResNet-34 with cubic 3D kernels as in [10]. The network is pre-trained on Kinetics and fine-tuned on the action recognition dataset training set. We follow [10] for training and evaluation. The network is trained with mini batch of size 128, where each sample is a $16 \times 112 \times 112$ clip cropped over 16 consecutive frames. At train time, we use random scaling, flipping, and cropping. For evaluation, we split each video as non-overlapped 16-frame clips and feed their center crops into the network. Video-level predictions are calculated by averaging the classification scores over clips. The network is trained using a SGD optimizer with an initial learning rate of 0.1.

5. Experiments

We first study the impact of video compression on recognition performance. Then, we apply the enhancement method described in section 2 on compressed videos and show the effect of enhancement in terms of video quality metrics. Finally we measure the effect of enhancements on recognition and relation between common enhancement metrics like PSNR and recognition metrics like rate-accuracy.

5.1. Impact of compression on recognition

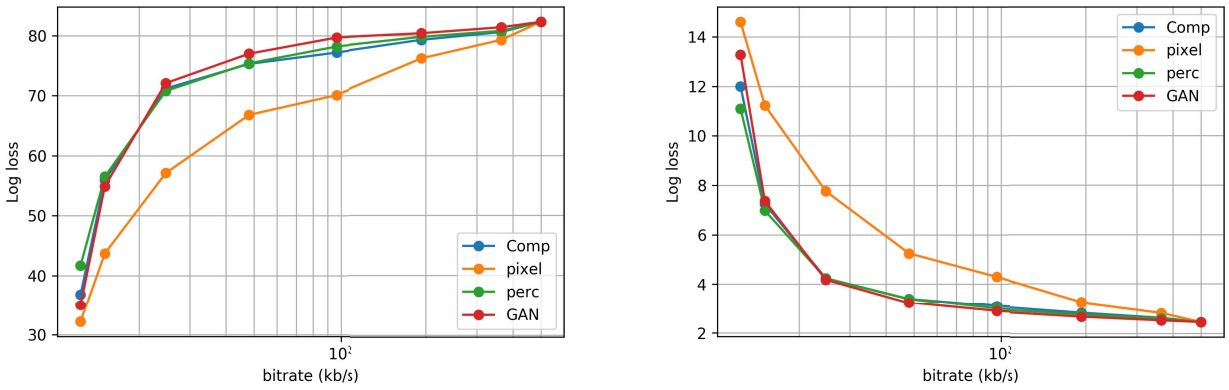
We employ two action recognition models, both based on Resnet architecture. Resnet-S is an appearance-based architecture that receives an RGB frame at a time and just captures spatial information. Resnet-ST receives 16 frames at a time and takes into account both spatial and temporal information. Both models are kept fixed after fine-tuning on uncompressed frames of the action recognition dataset.

We start with evaluating the models on uncompressed frames. As shown in Table 1, the performance of uncompressed frames is 82.3% and 87.9% for Resnet-S and Resnet-ST respectively. Not surprisingly, as compression rate increases, the drop in performance becomes higher. With both models, recognition degradation is exponential on highly compressed videos. However, performance drop is much higher on Resnet-S compared to Resnet-ST particularly for high compression rates (45.5 vs. 19.8 for 12.6 kb/s). This is mainly because compression operates on spatial domain while keeps motion information relatively intact. Resnet-ST can exploit such temporal cues to compensate spatial degradation on highly compressed videos.

To have a more in-depth analysis, we compute per-category accuracy using Resnet-S model for uncompressed frames and 12.6 kb/s compressed frames. Figure 4 shows top 15 and bottom 15 classes with highest recognition accuracy degradation. As expected, in 98 out of 101 classes recognition performance is higher for uncompressed frames. Among top 15 classes, for compressed frames, in some cases like *SkiJet* the model is confused with similar classes of *Rowing* and *Surfing*, while in some cases like

Bitrate	PSNR (dB)				TC (dB)				LPIPS			
	Comp	ℓ_{pixel}	ℓ_{GAN}	ℓ_{perc}	Comp	ℓ_{pixel}	ℓ_{GAN}	ℓ_{perc}	Comp	ℓ_{pixel}	ℓ_{GAN}	ℓ_{perc}
364.9	27.91	**28.87**	27.84	27.92	27.80	**28.96**	27.78	27.91	**0.068**	0.083	0.068	0.070
192.8	27.73	**28.57**	27.55	27.22	27.94	**29.01**	27.73	28.04	**0.084**	0.108	0.083	0.085
96.6	27.35	**27.98**	26.96	26.91	28.04	**29.16**	27.84	28.11	0.096	0.131	0.097	**0.095**
48.1	26.58	**27.20**	26.29	26.25	28.14	**29.05**	27.69	28.36	0.136	0.168	**0.132**	0.139
24.8	25.34	**25.92**	24.64	25.17	28.38	**29.45**	27.21	28.47	0.207	0.248	0.204	**0.202**
15.2	23.68	**24.22**	22.58	23.63	28.79	**29.82**	27.31	28.64	0.311	0.331	**0.262**	0.281
12.6	22.50	**23.02**	21.07	22.35	28.88	**29.92**	26.03	28.50	0.415	0.422	**0.353**	0.368

Table 2: Enhancement results for the action recognition dataset test-set.



(a) (b)

Figure 5: (a): Accuracy vs. bitrate on the action recognition dataset for Resnet-S. (b) log loss vs bitrate for the action recognition dataset.

Taichi and *Haircut*, videos are assigned to high false positive rate categories of *HandstandWalking, BoxingSpeedBag*. Among bottom 15 classes, surprisingly, on three action categories of *HandstandWalking, BoxingSpeedBag*, and *BreastStroke*, the performance of compressed frames is higher than uncompressed frames. After looking into these cases, we observed that many videos are assigned to these categories, leading to having highest false positive rate for them as well.

5.2. Impact of enhancement on video quality

We study how enhancements introduced in section 2 improve video quality metrics introduced in section 4. Table 2 compares the different configurations when applied to the action recognition dataset test-set, in terms PSNR, Temporal Consistency (TC), and LPIPS. Also Figure 6 provides a side-by-side visual comparison of the networks trained on ℓ_{pixel}, ℓ_{perc}, and ℓ_{GAN} for bitrate 15.2 kb/s. The results confirm that the ℓ_{pixel} model removes the compression artifacts to a high extent, but the enhanced video looks oversmooth. The ℓ_{perc} and ℓ_{GAN} models generate better results in terms of perceptual quality but a closer comparison reveals that the ℓ_{GAN} results look sharper and more realistic. Based on the results in Table 2, the ℓ_{pixel} models generate the highest PSNR and TC. The best PSNR is expected as the models are trained on ℓ_2. Highest TC could be due to the over-smoothness of the enhanced frames. The ℓ_{GAN} models perform very well in terms of LPIPS and visual quality but they deliver very poor TC results. That could be because GAN-based models might add some content to the frames to make them look more realistic. The added content could vary across frames and lead to poor temporal consistency. The performance and visual quality of the ℓ_{perc} models are somewhat between the ℓ_{pixel} and ℓ_{GAN} models.

5.3. Impact of enhancement on recognition

In this section, we show how enhancement can influence action classification accuracy. To do this, we compare three different enhancement objectives. The recognition performance for the three different enhancement objectives is shown in Table 1 and Figure 5. An interesting observation is that ℓ_{pixel} enhancement hurts the accuracy on both Resnet-S and Resnet-ST. This attributes to high blurriness of enhanced frames due to using ℓ_2 loss for pixel

Figure 6: Visual comparison of the studied loss functions. The uncompressed video is compressed using bitrate 15.2 kb/s and then restored using the networks trained on ℓ_{pixel}, ℓ_{perc}, and ℓ_{GAN}.
[Video by Gari Gonzalez, licensed under Creative Commons Attribution license (reuse allowed) via YouTube]

reconstruction. Also, we observe that ℓ_{GAN} and ℓ_{perc} based enhancements perform better particularly on highly compressed videos. This is expected as ℓ_{perc} optimizes the enhancement on feature space rather than pixel space. ℓ_{GAN} also generates sharper frames which might be helpful for recognition.

One drawback of rate-accuracy as metric is that is does not take into account the confidence scores generated by the model. To take into account uncertainty of the predictions based on how much it varies from the actual label, we use Log Loss metric as another recognition metric. It is defined as $-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} y_{ij}\log(p_{ij})$ where N is number of samples, L is number of labels, y is true label and p is confidence score. A perfect classifier (with confidence 1 on each prediction) will have log loss of 0 and a completely random classifier will have log loss $\log(L)$. Figure 5 shows the classification log loss for both compressed and enhanced videos using $\mathrm{Resnet\text{-}S}$ predictions. On low-rate compression, all models expect ℓ_{pixel} have almost same prediction confidence, however, we observe that the ℓ_{perc} model is more confident about its predictions on high-rate compression while predicted probability of other methods diverges more from actual label.

We previously showed in Figure 6 how the enhancements appear in video frames. Here, we show how the action recognition network "sees" these transformations. To visualize how the trained model "sees" the frames, we use Grad-CAM [27] to highlight the important regions in the input image with respect to the predicted category. As shown in Figure 7, model focuses on interested regions better with ℓ_{GAN} and ℓ_{perc} enhancement. This is consistent with quantitative results where ℓ_{GAN} and ℓ_{perc} performs better in highly compressed frames.

6. Conclusion

We study the impact of quality degradation of videos, caused by lossy compression, on action recognition performance. We investigate how the state of the art video enhancement methods, trained by various loss types, restore the video quality needed for an effective action recognition. Our experiments demonstrate that the models trained on pixel-level loss, which is a popular loss in enhancement domain, perform well in terms of PSNR and temporal consistency but they hurt the accuracy of action recognition. On the other hand, models trained on perceptual and adversarial loss types not only generate better perceptual quality *i.e*. in terms of LPIPS, but also further improve the action recognition performance.

References

[1] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2848–2857, July 2017.

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the*

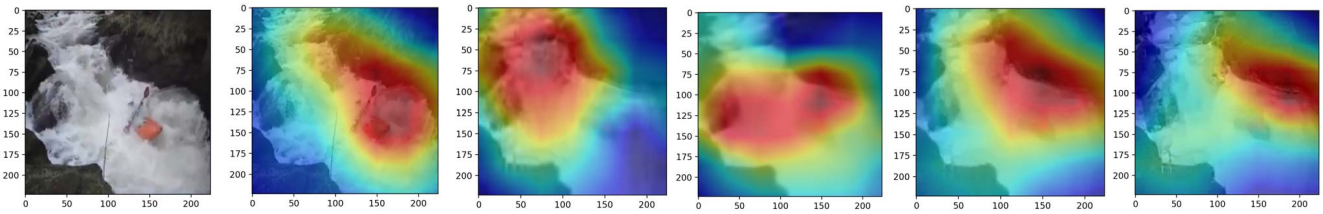


Figure 7: model visualization using Grad-CAM method for class "Kayaking". The video is compressed at 12.6 kb/s and then restored using the networks trained on ℓ_{pixel}, ℓ_{perc}, and ℓ_{GAN} loss types. Numbers in brackets are confidence of the network.
[Video by Stefan Senk, licensed under Creative Commons Attribution license (reuse allowed) via YouTube]

IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.

[3] M. Chu, Y. Xie, L. Leal-Taixé, and N. Thuerey. Temporally coherent gans for video super-resolution (tecogan). *CoRR*, abs/1811.09393, 2018.

[4] F. Developers. ffmpeg tool (version 4.1 "al-khwarizmi"), 2018. [Software]. Available from http://ffmpeg.org.

[5] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[6] C. Dong, Y. Deng, C. C. Loy, and X. Tang. Compression artifacts reduction by a deep convolutional network. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 576–584, Dec 2015.

[7] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo. Deep generative adversarial compression artifact removal. *CoRR*, abs/1704.02518, 2017.

[8] Z. Guan, Q. Xing, M. Xu, R. Yang, T. Liu, and Z. Wang. MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video. *CoRR*, abs/1902.09707, 2019.

[9] J. Guo and H. Chao. Building dual-domain representations for compression artifacts reduction. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 628–644, Cham, 2016. Springer International Publishing.

[10] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[11] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] Y. Huang, W. Wang, and L. Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):1015–1028, April 2018.

[14] T. Hyun Kim, M. S. M. Sajjadi, M. Hirsch, and B. Scholkopf. Spatio-temporal transformer network for video restoration. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[15] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015.

[16] A. Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. In *International Conference on Learning Representations*, 2019.

[17] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, June 2016.

[18] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 12 2014.

[20] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, July 2017.

[21] D. Li and Z. Wang. Video superresolution via motion compensation and deep residual learning. *IEEE Transactions on Computational Imaging*, 3(4):749–762, Dec 2017.

[22] K. Li, B. Bare, and B. Yan. An efficient deep convolutional neural networks model for compressed image deblocking. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1320–1325, July 2017.

[23] O. Makansi, E. Ilg, and T. Brox. End-to-end learning of video super-resolution with motion compensation. In *German Conference on Pattern Recognition (GCPR) 2017*, 2017.

[24] E. Prez-Pellitero, M. S. M. Sajjadi, M. Hirsch, and B. Schlkopf. Photorealistic video super resolution. In *ECCV Workshop (PIRM)*, 2018.

[25] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[26] M. S. M. Sajjadi, R. Vemulapalli, and M. Brown. Frame-recurrent video super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Oct 2017.

[28] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[30] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[31] T. Wang, M. Chen, and H. Chao. A novel deep learning-based method of improving coding efficiency from the decoder-end for hevc. In *2017 Data Compression Conference (DCC)*, pages 410–419, April 2017.

[32] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018.

[33] Z. Wang, D. Liu, S. Chang, Q. Ling, Y. Yang, and T. S. Huang. D3: Deep dual-domain based fast restoration of jpeg-compressed images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2764–2772, June 2016.

[34] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, Nov 2003.

[35] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman. Video enhancement with task-oriented flow. *arXiv*, 2017.

[36] R. Yang, M. Xu, and Z. Wang. Decoder-side hevc quality enhancement with scalable convolutional neural network. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 817–822, July 2017.

[37] R. Yang, M. Xu, Z. Wang, and T. Li. Multi-frame quality enhancement for compressed video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[38] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[39] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[40] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, July 2017.

[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018.