ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414731

ULTRA-LOW BITRATE VIDEO CONFERENCING USING DEEP IMAGE ANIMATION

Goluck Konuko[†], *Giuseppe Valenzise*[‡], *Stéphane Lathuilière*[†]

[†] LTCI, Télécom Paris, Institut polytechnique de Paris, France
[‡] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, France

ABSTRACT

In this work we propose a novel deep learning approach for ultra-low bitrate video compression for video conferencing applications. To address the shortcomings of current video compression paradigms when the available bandwidth is extremely limited, we adopt a model-based approach that employs deep neural networks to encode motion information as keypoint displacement and reconstruct the video signal at the decoder side. The overall system is trained in an end-to-end fashion minimizing a reconstruction error on the encoder output. Objective and subjective quality evaluation experiments demonstrate that the proposed approach provides an average bitrate reduction for the same visual quality of more than 80% compared to HEVC.

Index Terms— Video compression, Video conferencing, Model-based compression, Deep learning

1. INTRODUCTION

Video conferencing applications represent a substantial share of Internet video traffic, which has significantly increased in the past months due to the global pandemic. Video conferencing relies heavily of the availability of efficient video compression standards, such as H.264, HEVC [1], and, in the next future, VVC [2]. Although these video codecs have been optimized and tuned for over 30 years, they are still unable to provide acceptable performance at very low bitrates. In fact, when bandwidth is extremely limited, e.g., due to a congested network or to poor radio coverage, the resulting video quality becomes unacceptable, degrading the video conferencing experience significantly.

A major limitation of the current video compression paradigm comes from employing elementary models of pixel dependencies in order to increase coding efficiency. For instance, the commonly used discrete cosine transform implicitly assumes that pixels are generated by a Gaussian stationary process [3]. Similarly, spatial and temporal prediction exploit low-level pixel dependencies such as spatial directional smoothness and simple translational motion across time. Furthermore, these tools are often optimized with pixel-based fidelity metrics such as the Mean Squared Error (MSE), which is known to correlate poorly with human perception [4].

In this paper, we consider a very different, *model-based* compression approach for video conferencing that goes beyond simple pixel-based techniques. Specifically, we interpret video frames as points of a low-dimensional manifold living in the higher-dimensional pixel space. We leverage recent advances in image generation and deep generative networks [5], which have made possible to automatically synthesize images or videos of faces with unprecedented quality and realism [6, 7, 8, 9]. These tools have recently reached a great deal of popularity in computer vision and computer graphics applications, such as motion transfer [10], creation of video portraits [11], deep fakes [12], image to image translation [13] etc. However, their potential as a video compression tool is still unexplored. In this work, we describe, for the first time, a video coding pipeline that employs a recently proposed image animation framework [8, 9] to achieve long-term video frame prediction. Our scheme is open loop: we encode the first frame of the video using conventional Intra coding, and transmit keypoints extracted by subsequent frames in the bitstream. At the decoder side, the received keypoints are used to warp the Intra reference frame to reconstruct the video. We also propose and analyze an adaptive Intra frame selection scheme that, in conjunction with varying the quantization of Intra frames, enable to attain different rate-distortion operating points. Our experiments, validated with several video quality metrics and a subjective test campaign, show average bitrate savings compared to HEVC of over 80%, demonstrating the potential of this approach for ultra-low bitrate video conferencing. The source code and additional results are available at https://goluck-konuko.github.io/.

2. RELATED WORK

Model-based compression paradigms are not new in data compression. A prominent example comes from speech coding standards, where vocoders [14] are a broadly used kind of generative models. Unfortunately, generative image and video models turn out to be significantly more complex than in the audio case. Previous work aimed at using generative models for talking heads has failed to provide realistic and appealing video compression performance. The most

notable example was the MPEG-4 Visual Objects standard, which enables embedding of synthetic objects such as faces or bodies and animate them using a sequence of animation parameters [15]. Although the concept of visual object was an innovative idea and inspired follow-up work (even recently, e.g., [16]), the quality of the resulting animated faces was quite unrealistic. As a result, this part of the standard has been rarely used in practice. In this work, we use a recently proposed image animation model which can instead produce highly realistic faces [9].

Recently, deep neural networks have been successfully applied to image and video compression [17]. Learning-based image compression has been generally cast as the problem of learning deep representations, typically by means of deep auto-encoders, which are optimized in an end-to-end fashion by maximizing a rate-distortion criterion [18, 19]. Learning-based image codecs can compress pictures in a more natural way than conventional codecs [20]. Similar ideas have been employed later for the case of video. There, deep learning tools have been mainly used to replace and optimize single stages of the video coding pipeline [21, 22]. In this work, we consider instead a generative modeling perspective, and propose a coding architecture which departs substantially from that of a conventional hybrid video codec.

Deep generative models have also been recently employed in image/video compression schemes [23, 24], to reduce bitrate by hallucinating parts of the video that are outside the region of interest (ROI). In this work, instead, we use deep generative models to synthesize face images, which constitute the main ROI in video communication. ROI-based coding has been previously used for video conferencing [25]. There, coding gains are obtained by varying the bitrate allocation between the ROI/non-ROI regions. We use instead a very different warping tool to animate the ROI with a very low bitrate.

3. PROPOSED CODING METHOD

3.1. Coding scheme based on image animation

The overall pipeline of the proposed codec is illustrated in Fig. 1. It consists of an Intra frame compression module (which could be any state-of-the-art image codec); a sparse keypoint extractor to code Inter frames; a reconstruction (warping) module to motion compensate and reconstruct the Inter frames; and a binarizer and entropy coder to produce a compressed binary bitstream. In the following, we provide a walk-through of the main coding steps of our system:

Intra-frame coding. We first encode the initial frame F_0 using the *BPG* image codec, which essentially implements HEVC Intra coding, using a given QP_0. The coded frame is sent to the decoder. On the decoder side, the initial frame is decoded using the *BPG* decoding procedure. This decoded initial frame is referred to as $\tilde{F}_0$.

Keypoint prediction. Following the recent work of Siarohin *et al.*[9], we encode motion information via a set of 2D keypoints learned in a self-supervised fashion. The keypoints are predicted by a U-Net architecture [26] network that, from its input frame, estimates M heatmaps (one for each keypoint). Each heatmap is then interpreted as detection confidence map and used to compute the expected keypoint locations. In addition, following [9], we compute a (symmetric) 2×2 Jacobian matrix for each keypoint, which encodes the orientation of the face region associated to that keypoint. Each keypoint is then represented by a 5-dimensional vector (2 spatial coordinates, plus the 3 elements to represent the Jacobian matrix).

Entropy coding. Keypoints are represented as floating point values with 16 bits precision. We compress these values using an off-the-shelf LZW encoder (*gzip*). Notice that any other entropy coding solution, e.g., a binary arithmetic codec with contexts, could be equally used. The coded motion information, together with the *BPG* compressed Intra frames, form the bitstream that is sent to the decoder.

Motion compensation and inter-frame reconstruction At the decoder side, we reconstruct the frame F_t at time t from the decoded initial frame $\tilde{F}_0$ and the displacement of a set of M keypoints, $K_0 \in \mathbb{R}^{5\times M}$ and $K_t \in \mathbb{R}^{5\times M}$, extracted from frames $\tilde{F}_0$ and F_t, respectively. These keypoints are detected for each frame at the encoder side and transmitted in the bitstream, as explained above. We then employ the reconstruction network introduced in [9] to reconstruct $\tilde{F}_t$ from $\tilde{F}_0$ and the keypoints K_0 and K_t. This reconstruction network is composed of two sub-networks. The first sub-network predicts the optical flow between the frames at times 0 and t from $\tilde{F}_0$, K_0 and K_t. This optical flow is then used by the second sub-network to predict $\tilde{F}_t$ from $\tilde{F}_0$. Please refer to [9] for the details of the reconstruction network architecture.

The overall system is trained in an end-to-end fashion on a large training set of videos depicting talking faces. We minimize a reconstruction loss between the input and the decoded frames. In this way, we force the keypoint detector network to predict 2D keypoints that describe motion in such a way that the reconstruction network can correctly reconstruct each frame. The reconstruction loss is based on the perceptual loss of Johnson *et al.* [27] using a VGG-19 network pre-trained on ImageNet. This loss is completed with a GAN loss [28] and an equivariance loss that enforces that network to be equivariant to random geometric transformations [9]. Notice that Intra frame coding with *BPG*, as well as *gzip*, are not differentiable, and as a consequence, cannot be included in the training process. While differentiable approaches to estimate entropy have been proposed [18], in this work we opt for the simpler solution to ignore this step during training.

3.2. Adaptive Intra frame selection

The coding scheme described in the previous section adopts an open GOP structure, i.e., only the first frame is coded as Intra. While this leads to a very high video compression rate,

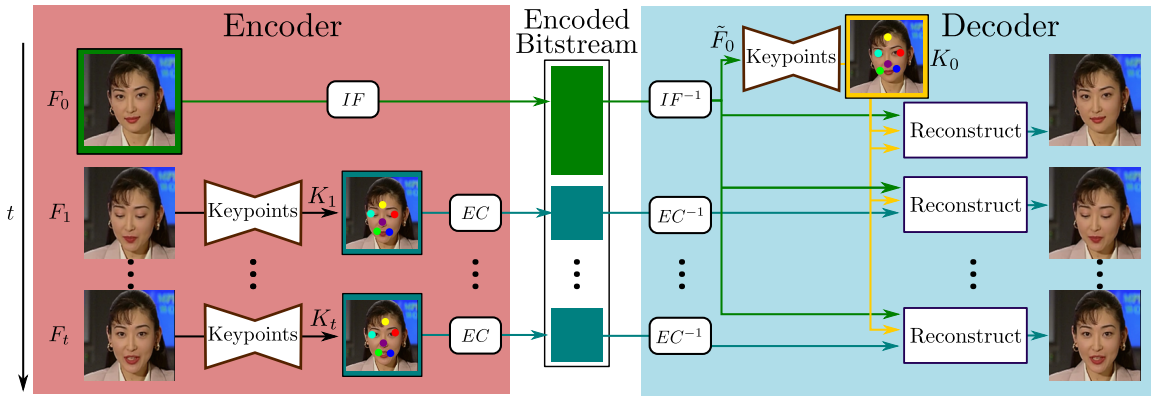


Fig. 1: Basic scheme of our proposed baseline codec. The encoder (in red) encodes the first frame as Intra (IF) using a *BPG* codec. The following frames ($t > 1$) are encoded as keypoints predicted by a neural network. Keypoints are entropy coded (EC) and transmitted in the bitstream. The decoder (in blue) includes a neural network that combines the reconstructed initial frame $\tilde{F}_0$ and its keypoints K_0, with the keypoints K_t at the current time t, to reconstruct F_t. The overall system is trained in an end-to-end manner via reconstruction loss minimization.

the reconstruction quality might rapidly degrade due to the loss of temporal correlation as frames get farther away from the initial one. This is particularly evident in case of occlusions/disocclusions caused by a person's pose change.

To overcome this issue, we introduce in the following an adaptive Intra-Refresh scheme using multiple source frames. The procedure is described in Algorithm 1. The proposed multi-source Intra frame selection is parametrized by a threshold parameter τ that allows quality (and, implicitly, bitrate) adjustment. We employ a buffer $\mathcal{B}$ containing the key frames that are required to reconstruct the input video, together with the corresponding keypoints. At every time step, this buffer is synchronized between the encoder and the decoder. To this end, every frame added to $\mathcal{B}$ by the encoder is sent to the decoder that, in turn, adds the frame to its buffer.

The buffer is initialized with the first frame $\tilde{F}_0$ and the keypoints K_0, as described in Section 3.1. Note that the keypoints K_0 need not be sent to the decoder since K_0 can be re-estimated at the decoder side from $\tilde{F}_0$. Then, for every frame F_t, we apply the following procedure: We estimate the keypoints K_t. Then, we identify the best frame in the buffer that can be used to reconstruct the current frame. To this aim, we reconstruct $\tilde{F}_t$ using all the frames $\tilde{F}_b$ and their keypoints K_b. We select the buffer frame index b^* that leads to the lowest PSNR. We use PSNR here, rather than perceptual loss, because of its lower computational cost. At this point, two cases can happen: i) if the best source frame yields a PSNR higher than the threshold τ, we use the frame F_{b^*} as source frame to reconstruct F_t. Therefore, the index b^* and the keypoints K_t are sent to the decoder; ii) the PSNR is lower than threshold τ. In this case, none of the frames in the buffer is suitable to reconstruct the current frame and F_t needs to be added to the buffer as a new Intra frame and sent to the decoder. To this end, we encode the current frame F_t with QP_0 used for the first frame.

Algorithm 1: Adaptive Intra frame selection (encoder side)

Input : Frames: $F_0, \ldots, F_T$, Threshold: $\tau > 0$

$K_0 = \texttt{Keypoint}(\tilde{F}_0)$ `// Estimate keypoints`

$\mathcal{B} = \{(\tilde{F}_0, K_0)\}$ `// Initialize the buffer`

$\texttt{Send}(F_0)$

for $t \in \{1, ..., T\}$ **:** `// For every frame`

  $K_t = \texttt{Keypoint}(F_t)$ `// Estimate keypoints`

  $\tilde{F}_t^{\mathcal{B}} = \{\texttt{Reconstruct}(\tilde{F}_b, K_b, K_t); \forall (\tilde{F}_b, K_b) \in \mathcal{B}\}$

  $b^* = \underset{\tilde{F}_t^b \in \tilde{F}_t^{\mathcal{B}}}{\text{argmin}} \left(\texttt{PSNR}(\tilde{F}_t^b, F_t)\right)$ `// Best frame in` $\mathcal{B}$

  if $PSNR(\tilde{F}_t^{b^*}, F_t) > \tau$ **:**

    $\texttt{Send}(b^*, K_t)$ `// Signal the best frame`

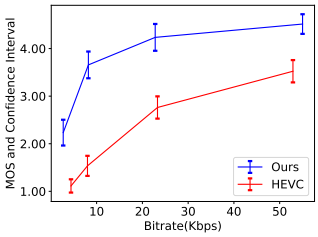
  else:

    $\texttt{Set\_QP}(F_t, \tau)$ `// Set QP for IC`

    $\mathcal{B} \leftarrow \mathcal{B} \cup (\tilde{F}_t, K_t)$ `// Add to buffer`

    $\texttt{Send}(F_t)$ `// Send a new source frame`

If the resulting PSNR is still above τ, we reduce the QP by one unit, and repeat the procedure till the quality constraint is satisfied. This provides a set of operating rate-distortion curves for the codec. By taking the convex hull of these curves, we can effectively attain different rate-distortion trade-offs. We leave the study of efficient rate-distortion optimization strategies for adaptive Intra frame selection to future work. Signaling the best source frame index is a good trade-off between compression rate and computational cost. We implement $\mathcal{B}$ as a first-in, first-out buffer. When a new source frame is added, the oldest source is popped assuming that reconstruction error increases with time. In all our experiments, we use a buffer of size 5. In our preliminary experiments, we observed that a larger buffer size does not bring significant gain but increases memory requirements. Additionally, the source frame keypoints, $K_b^{\mathcal{B}}$ can be signalled directly in the bitstream to eliminate the need for a keypoint detector at the decoder.



	PCC	SROCC
PSNR	0.92/0.95	0.77/0.92
SSIM	0.91/0.74	0.80/0.67
MS-SSIM	0.94/0.95	0.84/0.95
VIF	0.38/0.75	0.28/0.75
VMAF	0.94/0.96	0.80/0.89

Fig. 2: User evaluation: Mean Opinion Score (MOS) for our codec compared to HEVC (average over all sequences). We study correlation between MOS and each metric for videos compressed with our approach/HEVC.

4. EXPERIMENTS AND RESULTS

Datasets. We employ two datasets suited for a video conferencing scenario:

• **Voxceleb** is a large audio-visual dataset of human speech of 22,496 videos, extracted from YouTube videos. We follow the pre-processing of [9] and filter out sequences that have resolution lower than 256×256 and resize the remaining videos to 256×256 preserving the aspect ratio. This dataset is split into 12,331 training and 444 test videos. For evaluation, in order to obtain high quality videos, we select the 90 test videos with the highest resolution before resizing.

• **Xiph.org** is a collection of videos we downloaded from Xiph.org [1]. This repository includes video sequences widely used in video processing ("News", "Akiyo", "Salesman", etc.). We select 16 sequences of talking heads that correspond to the targeted video conferencing scenario. The full list of videos used in the experiments is available on the GitHub page of the paper [2]. The region of interest is cropped with a resolution of 256×256 with speakers' faces comprising 75% of the frame. Notice that we use this dataset for testing *only*, while the system is trained on the Voxceleb training set.

Comparison with HEVC. We compare our approach with the standard HM 16 (HEVC) low delay configuration[3] by performing subjective test on 10 videos (8 from Voxceleb and 2 from Xiph.org) using Amazon Mechanical Turk (AMT). Each sequence is encoded using 8 different bit-rate configurations (4 with HEVC, 4 with proposed method) ranging from 1.5Kbps to 55Kbps. The different points in the curve are obtained by changing QP_0 in BPG, as well as τ. We implement a simple Double Stimulus Impairment Scale (DSIS) test [29] interface in AMT. Users are invited to follow a brief training on a sequence not used for test. The extreme quality levels ("very annoying", "imperceptible") are used as gold units to check the reliability of voters. We collected 30 subjective evaluations per stimulus. After this screening, no further outliers were found. The Mean Opinion Scores (MOS) with 95% confidence intervals are shown in Fig. 2-left. It shows that our approach clearly outperforms HEVC (average BD-MOS = 1.68, BD-rate = -82.35%). In Fig 2-right, we report the Pearson Correlation Coefficient (PCC) and the Spearman Rank-Order Correlation Coefficient (SROCC) between MOS and five commonly used quality metrics, for our codec and HEVC. We observe that, except VIF, these metrics correlate well with human judgment, at least on the tested data. Based on these preliminary results, we proceed with an extensive performance evaluation of the proposed method using these quality metrics. In Table 1, we report the average Bjontengaard-Delta performance for VoxCeleb test and Xiph.org images. The gains are significant (over 80% BD-rate savings) on the two datasets.

[1] https://media.xiph.org/video/derf/

[2] https://github.com/Goluck-Konuko/dac.git

[3] https://github.com/listenlink/HM

Table 1: Bjontengaard-Delta Performance over HEVC

	VoxCeleb BD quality / BD rate	**Xiph.org** BD quality / BD rate
PSNR	2.88 / -65.50	3.14 / -72.44
SSIM	0.122/ -83.96	0.02/ -65.79
MS-SSIM	0.070 / -83.60	0.075 / -86.41
VIF	0.027 / -72.29	0.021 / -68.02
VMAF	37.43 / -82.29	31.04 / -83.44

Fig. 3: Qualitative evaluation against HEVC on *Deadline* sequence in a low bitrate setting.

Finally, we provide a qualitative comparison in Fig. 3 on a sequence from Xiph.org. In this example, we observe that at a similar bitrate, our approach produces much fewer artifacts. The difference is clearly visible in the eye region. Even at a much lower bitrate (2.3 Kbps) – one lower than the minimum rate achievable by HEVC – our approach generates better quality images.

5. CONCLUSIONS

In this paper, we propose the first video conferencing codec that employs a deep generative frame animation scheme and drastically improve coding performance at ultra-low bitrate. We encode the initial frame using a state-of-the-art image codec. Motion information is then encoded via moving keypoints predicted by a deep network. We propose an adaptive Intra frame selection mechanism to improve reconstruction quality over long sequences. Our experiments show that our approach outperforms HEVC with a large margin. As future work, we plan to handle high-resolution videos and more complex scenarios with multiple speakers in the same video.

6. REFERENCES

[1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE TCSVT*, 2012.

[2] B. Bross, J. Chen, and S. Liu, "Jvet-m1001: Versatile video coding (draft 5)," in *Joint Video Experts Team (JVET), 14th Meeting: Geneva, SW, Tech. Rep*, 2019.

[3] P. Pad and M. Unser, "On the optimality of operator-like wavelets for sparse AR (1) processes," in *IEEE ICASSP*, 2013.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, 2004.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, Warde-Farley, S. D., Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets.," in *NIPS*, 2014.

[6] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE CVPR*, 2019.

[7] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *IEEE ICCV*, 2019.

[8] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in *IEEE/CVF CVPR*, 2019.

[9] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Neurips*, 2019.

[10] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," *CoRR*, vol. abs/1808.07371, 2018.

[11] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *ACM TOG*, 2018.

[12] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *IEEE ICASSP*, 2019.

[13] T.-P. Nguyen, S. Lathuilière, and E. Ricci, "Multi-domain image-to-image translation with adaptive inference graph," *ICPR*, 2020.

[14] A. S. Spanias, "Speech coding: A tutorial review," *Proceedings of the IEEE*, 1994.

[15] M. Preda and F. Preteux, "Critic review on MPEG-4 face and body animation," in *ICIP*, 2002.

[16] M. Wijnants, S. Coppers, G. R. Ruiz, P. Quax, and W. Lamotte, "Talking video heads: Saving streaming bitrate by adaptively applying object-based video principles to interview-like footage," in *27th ACM International Conference on Multimedia (MM'19)*, 2019.

[17] S. Ma, X. Zhang, Z. Jia, C., Z., and S. Wang, "Image and video compression with neural networks: A review," in *IEEE TCSVT*, 2019.

[18] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[19] O. Rippel and L. Bourdev, "Real-time adaptive image compression," *arXiv preprint arXiv:1705.05823*, 2017.

[20] G. Valenzise, A. Purica, V. Hulusic, and M. Cagnazzo, "Quality assessment of deep-learning-based image compression," in *IEEE MMSP*, 2018.

[21] O. Rippel, S. Nair, C. Lew, S. Branson, A. G. Anderson, and L. Bourdev, "Learned video compression," in *IEEE Xplore/Computer Vision Foundation*, 2019.

[22] L. Wang, A. Fiandrotti, A. Purica, G. Valenzise, and M. Cagnazzo, "Enhancing hevc spatial prediction by context-based learning," in *IEEE ICASSP*, 2019.

[23] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *IEEE ICCV*, 2019.

[24] A. S. Kaplanyan, A. Sochenov, T. Leimkühler, M. Okunev, T. Goodall, and G. Rufo, "Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos," *ACM TOG*, 2019.

[25] M. Meddeb, M. Cagnazzo, and B. Pesquet-Popescu, "Region-of-interest-based rate control scheme for high-efficiency video coding," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 2014.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[28] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *IEEE ICCV*, 2017.

[29] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," ITU-R Recommendation BT. 500-13, Jan 2012.