Perceptual Learned Video Compression with Recurrent Conditional GAN

Ren Yang\textsuperscript{1}, Radu Timofte\textsuperscript{1,2}, Luc Van Gool\textsuperscript{1,3}

\textsuperscript{1}ETH Zürich, Switzerland
\textsuperscript{2}Julius Maximilian University of Würzburg, Germany
\textsuperscript{3}KU Leuven, Belgium

ren.yang@vision.ee.ethz.ch, radu.timofte@uni-wuerzburg.de, vangool@vision.ee.ethz.ch

Abstract
This paper proposes a Perceptual Learned Video Compression (PLVC) approach with recurrent conditional GAN. We employ the recurrent auto-encoder-based compression network as the generator, and most importantly, we propose a recurrent conditional discriminator, which judges raw vs. compressed video conditioned on both spatial and temporal features, including the latent representation, temporal motion and hidden states in recurrent cells. This way, the adversarial training pushes the generated video to be not only spatially photo-realistic but also temporally consistent with the groundtruth and coherent among video frames. The experimental results show that the learned PLVC model compresses video with good perceptual quality at low bit-rate, and that it outperforms the official HEVC test model (HM 16.20) and the existing learned video compression approaches for several perceptual quality metrics and user studies. The project page is available at https://github.com/RenYang-home/PLVC.

1 Introduction
Recent years have witnessed the increasing popularity of end-to-end learning-based image and video compression. Most existing models are optimized towards distortion, i.e., PSNR and MS-SSIM. Recently, Generative Adversarial Networks (GANs) have been used in image compression, to boost perceptual quality [Agustsson et al., 2019; Mentzer et al., 2020]. However, efficient perceptual video compression still remains limited.

In this paper, we propose a Perceptual Learned Video Compression (PLVC) approach with recurrent conditional GAN, in which a recurrent generator and a recurrent conditional discriminator are trained in an adversarial way for perceptual compression. The recurrent generator contains recurrent auto-encoders to compress video and generate visually pleasing reconstructions in the adversarial training. More importantly, we propose a recurrent conditional discriminator, which judges between raw and compression video conditioned on the spatial-temporal features, including latent representations, motion and the hidden states transferred through recurrent cells. Therefore, in the adversarial training, the discriminator is able to force the recurrent generator to reconstruct both photo-realistic and temporally coherent video, thus achieving good perceptual quality.

Fig. 1 shows the visual result of the proposed PLVC approach on BasketballDrill (bpp = 0.0730) in comparison with the official HEVC test model (bpp = 0.0813). The top of Fig. 1 shows that our PLVC approach achieves richer and more photo-realistic textures than HEVC. At the bottom of Fig. 1, we show the temporal profiles by vertically stacking a specific row (marked as green) along time steps. It can be seen that the result of our PLVC approach exhibits a temporal coherence comparable with that of HEVC, while offering more detailed textures. As a result, we outperform HEVC on perceptual quality at lower bit-rates. The contributions of this paper are summarized as:

- We propose a novel perceptual video compression approach with recurrent conditional GAN, which learns to compress video and generate photo-realistic and temporally coherent compressed frames.
• We propose adversarial loss functions for perceptual video compression to balance bit-rate, distortion and perceptual quality.
• The experiments (including user study) show the outstanding perceptual performance of our PLVC approach in comparison with the latest learned and traditional video compression approaches.

2 Related Work

Learned image compression. In the past years, learned image compression has gained in popularity. For instance, Ballé et al. proposed utilizing the variational auto-encoder for image compression and proposed the factorized [Ballé et al., 2017] and hyperprior [Ballé et al., 2018] entropy models. Later, the auto-regressive entropy models [Minnen et al., 2018; Lee et al., 2019; Cheng et al., 2019] were proposed to improve the compression efficiency. Recently, the coarse-to-fine model [Hu et al., 2020a] and the wavelet-like deep autoencoder [Ma et al., 2020] were designed to further advance the rate-distortion performance. Besides, there are also the methods with RNN-based auto-encoder [Toderici et al., 2017; Johnston et al., 2018] or conditional auto-encoder [Choi et al., 2019] for variable rate compression. Moreover, Agustsson et al. [2019] and Menzter et al. [2020] proposed applying GANs for perceptual image compression to achieve photo-realistic compressed images at low bit-rate.

Learned video compression. Inspired by the success of learned image compression, the first end-to-end learned video compression method DVC [Lu et al., 2019] was proposed in 2019. Later, M-LVC [Lin et al., 2020] extended the range of reference frames and beats the DVC baseline. Meanwhile, a plenty of learned video compression methods with bi-directional prediction [Djelouah et al., 2019], one-stage flow [Liu et al., 2020], hierarchical layers [Yang et al., 2020a] and scale-space flow [Agustsson et al., 2020] were proposed. Besides, methods like the content adaptive model [Lu et al., 2020], resolution-adaptive flow coding [Hu et al., 2020b], the recurrent framework [Yang et al., 2021] and feature-domain compression [Ha et al., 2021] were employed to further improve compression efficiency. All these methods are optimized for PSNR or MS-SSIM. Veerabadrpan et al. [2020] proposed a GAN-based method for perceptual video compression, but has not outperformed x264 and x265 on LPiPS. Therefore, video compression aimed at perceptual quality still offers room for improvement.

3 Proposed PLVC Approach

3.1 Motivation

A GAN was first introduced by Goodfellow et al. [2014] for image generation. It learns to generate photo-realistic images by optimizing the adversarial loss

$$\min_G \max_D \mathbb{E}[f(D(x))] + \mathbb{E}[g(D(y))]$$

where $f$ and $g$ are scalar functions, and $G$ maps the prior $y$ to $p_y$. We define $x$ as the groundtruth image and $\hat{x} = G(y)$ as the generated image. The discriminator $D$ learns to distinguish $\hat{x}$ from $x$. In the adversarial training, it pushes the distribution of generated samples $p_\hat{x}$ to be similar to $p_x$. As a result, $G$ learns to generate photo-realistic images.

Later, the conditional GAN [Mirza and Osindero, 2014] was proposed to generate images conditioned on prior information. Defining the conditions as $c$, the loss function can be expressed as

$$\min_G \max_D \mathbb{E}[f(D(x | c))] + \mathbb{E}[g(D(\hat{x} | c))]$$

with $\hat{x} = G(y)$. The goal of employing $c$ in (2) is to push $G$ to generate $\hat{x} \sim p_{x|c}$ with the conditional distribution tending to be the same as $p_{x|c}$. In other words, it learns to fool $D$ to believe that $\hat{x}$ and $x$ correspond to a shared prior $c$ with the same conditional probability. By properly setting the condition prior $c$, the conditional GAN is expected to generate frames with desired properties, e.g., rich texture, temporal consistency and coherence, etc. This motivated us to propose a conditional GAN for perceptual video compression.

3.2 Framework

Fig. 2 shows the framework of the proposed PLVC approach with recurrent conditional GAN. We define the raw and compressed frames as $\{x_i\}_{i=1}^T$ and $\{\hat{x}_i\}_{i=1}^T$, resp. In our PLVC approach, we first estimate the motion between the current frame $x_i$ and its previous compressed frame $\hat{x}_{i-1}$ by the pyramid optical flow network [Ranjan and Black, 2017]. The motion is denoted as $m_i$, which is to be used in the recurrent generator ($G$) for compressing $x_i$, and also to serve as one of the temporal conditions in the recurrent conditional discriminator ($D$). The architectures of $G$ and $D$ in our PLVC approach are introduced next.

Recurrent generator $G$. The recurrent generator $G$ plays the role as a compression network. That is, given the reference frame $\hat{x}_{i-1}$ and the motion $m_i$, the target frame $x_i$ can be compressed by an auto-encoder-based DNN, which compresses $x_i$ to a latent representation $y_i$ and outputs the compressed frame $\hat{x}_i$. In our PLVC approach, we employ the recurrent compression model RLVC [Yang et al., 2021] as the backbone of $G$, which contains the recurrent auto-encoders to compress the motion and the residual, and the recurrent probability model $P$ for the arithmetic coding of $y_i$. The detailed architecture of $G$ is shown in the Supplementary Material.¹

Recurrent conditional discriminator $D$. The architecture of the proposed recurrent conditional discriminator $D$ is shown in Figure 2. First, we follow [Menzter et al., 2020] to feed $y_i$ as the spatial condition to $D$. This way, $D$ learns to distinguish the groundtruth and compressed frames based on the shared spatial feature $y_i$, and therefore pushes $G$ to generate $\hat{x}_i$ with spatial features similar to $x_i$.

More importantly, the temporal coherence is essential for visual quality. We insert a ConvLSTM layer in $D$ to recurrently transfer the temporal information along time steps. The hidden state $h^D_i$ can be seen as a long-term temporal condition fed to $D$, facilitating $D$ to recurrently discriminate raw

¹https://arxiv.org/abs/2109.03082
and compressed video taking temporal coherence into consideration. Furthermore, it is necessary to consider the temporal fidelity in video compression, i.e., we expect the motion between compressed frames to be consistent with that between raw frames. For example, the ball moves along the same path in the videos before and after compression. Hence, we propose \( D \) to take as inputs the frame pairs \((x_i, x_{i-1})\) and \((\hat{x}_i, \hat{x}_{i-1})\) and make the judgement based on the same motion vectors \( m_i \) as the short-term temporal condition. Without the \( m_i \)-condition, \( G \) may learn to generate photo-realistic \((\hat{x}_i, \hat{x}_{i-1})\) but with incorrect motion between the frame pair. This leads to poor temporal fidelity when compared to the groundtruth video. This motivates us to include the temporal condition \( m_i \) as an input to the discriminator.

In summary, we propose to condition the discriminator \( D \) on spatial features \( y_i \), short-term temporal features \( m_i \) and long-term temporal features \( h_{D}^{i-1} \). Given these conditions, we have \( c = \{y_i, m_i, h_{D}^{i-1}\} \) in (2) in our PLVC approach. The output of \( D \) can be formulated as \( D(x_i, x_{i-1} | y_i, m_i, h_{D}^{i-1}) \) and \( D(\hat{x}_i, \hat{x}_{i-1} | y_i, m_i, h_{D}^{i-1}) \) for the raw and compressed samples, resp. When optimizing the adversarial loss on sequential frames, the compressed video \( \{x_i\}_{i=1}^{T} \) tends to have the same spatial-temporal features as the raw video \( \{x_i\}_{i=1}^{T} \). Therefore, we achieve perceptual video compression with spatially photo-realistic as well as temporally coherent frames. Please refer to the Supplementary Material for the detailed architecture of \( D \).

It is noteworthy that the effectiveness of \( D \) is not restricted to a specific \( G \), but generalizes to various compression networks. Please refer to the analyses and results in Section 4.6.

### 3.3 Training Strategies

Our PLVC model is trained on the Vimeo-90k [Xue et al., 2019] dataset. Each clip has 7 frames, in which the first frame is compressed as an I-frame, using the latest generative image compression approach [Mentzer et al., 2020], and other frames are P-frames. To train the proposed network, we first warm up \( G \) on by the rate-distortion loss

\[
\mathcal{L}_w = \sum_{i=1}^{N} R(y_i) + \lambda \cdot d(\hat{x}_i, x_i).
\]

In (3), \( R(\cdot) \) denotes the bit-rate, and we use the Mean Square Error (MSE) as the distortion term \( d \). Besides, \( \lambda \) is a hyper-parameter to balance the rate and distortion terms.

Then, we train \( D \) and \( G \) alternately, with the loss function combining the rate-distortion loss and the non-saturating [Lucic et al., 2018] adversarial loss. Specifically, the loss func-

---

**Table 1: The hyper-parameters for training PLVC models.**

| Quality    | \( R_f \) (bpp) | \( \lambda \) | \( \alpha_1 \) | \( \alpha_2 \) | \( \lambda' \) | \( \beta \) |
|------------|-----------------|---------------|---------------|---------------|---------------|---------------|
| Low        | 0.025           | 256           | 3.0           | 0.010         | 100           | 0.1           |
| Medium     | 0.050           | 512           | 1.0           | 0.010         | 100           | 0.1           |
| High       | 0.100           | 1024          | 0.3           | 0.001         | 100           | 0.1           |

---

**Figure 3: Convergence curves of bit-rates during training.**
We use the JCT-VC [Bossen, 2013] (Classes B, C and D) and the UVG [Mercat et al., 2020] datasets as the test sets. On perceptual metrics, we compare with the Low-Delay P mode of the official HEVC model (HM 16.20) and the latest open-sourced\textsuperscript{2} learned compression approaches DVC [Lu et al., 2019; Yang et al., 2020b], HLVC [Yang et al., 2020a], M-LVC [Lin et al., 2020] and RLVC [Yang et al., 2021], and the GAN-based method [Veerabadran et al., 2020]. We also report the MS-SSIM and PSNR results in comparison with several existing approaches. Besides, we conduct MOS experiments as user studies to evaluate perceptual quality. There

\textsuperscript{2}We need the codes to reproduce the videos for the user study.

\textsuperscript{3}In this paper, we use OpenDVC [Yang et al., 2020b] to reproduce the results of DVC.

\begin{equation}
\begin{aligned}
\mathcal{L}_D = & \sum_{i=1}^{N} \left( - \log \left( 1 - D(\hat{x}_i, \hat{x}_{i-1} | y_i, m_i, h_i^D) \right) \\
& - \log D(x_i, x_{i-1} | y_i, m_i, h_i^D) \right), \\
\mathcal{L}_G = & \sum_{i=1}^{N} \left( \alpha \cdot R(y_i) + \lambda \cdot d(\hat{x}_i, \hat{x}_1) - \beta \cdot \log D(\hat{x}_i, \hat{x}_{i-1} | y_i, m_i, h_i^D) \right). 
\end{aligned}
\end{equation}
are 12 non-expert subjects participating in the MOS experiments. They are asked to rate the compressed videos with a score from 0 to 100 according to subjective quality. Higher score indicates better perceptual quality.

4.2 Numerical Performance
To numerically evaluate the perceptual quality, we calculate the Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al., 2018], Fréchet Inception Distance (FID) [Heusel et al., 2017] and Kernel Inception Distance (KID) [Bülow et al., 2018] on our PLVC and other approaches. In Fig. 4, we observe that the proposed PLVC approach outperforms all compared methods in terms of all three perceptual metrics. Especially, our PLVC approach reaches comparable or even better LPIPS, FID and KID values than other approaches with $2 \times$ to $4 \times$ the bit-rates of ours. This indicates that PLVC is able to compress video at low bit-rates but with good perceptual quality, showing the efficiency of PLVC for perceptual video compression.

Moreover, Fig. 5 shows that the PSNR of our PLVC approach is comparable with DVC. Our MS-SSIM result beats DVC and is competitive with Lu et al. [2020]. This shows that PLVC is able to maintain the fidelity to an acceptable degree when optimized towards perceptual quality.

4.3 Visual Results and User Study
Fig. 6 shows the visual results of our PLVC approach in comparison with HM 16.20 and the latest learned approach RLVC (MS-SSIM optimized). The top of Fig. 6 illustrates spatial textures, and the bottom shows temporal profiles by vertically stacking the rows along time. It can be seen from Fig. 6 that PLVC yields richer and more photo-realistic textures at lower bit-rates than the methods compared against. Moreover, the temporal profiles indicate that PLVC maintains the temporal coherence comparable to the groundtruth. We provide more visual examples in the Supplementary Material.

In the user study, the rate-MOS curves and the standard deviation of MOS for the JCT-VC dataset are shown in Fig. 7. As we can see from Fig. 7-(a), PLVC successfully outperforms the official HEVC test model HM 16.20, especially at low bit-rates, and we are significantly better than the MS-SSIM-optimized RLVC. Also, Fig. 7-(b) shows that the standard deviation of MOS values of our PLVC approach is relatively lower than for HM 16.20 and RLVC, indicating that PLVC achieves a more stable perceptual quality. In conclusion, these results confirm that people tend to prefer the perceptual quality of PLVC over that of HM 16.20 and RLVC.

4.4 Ablation Studies
In our ablation studies, we analyze the effectiveness of the generative adversarial network, the recurrent structure of the proposed GAN, and the spatial-temporal conditions $m_i$ and $y_i$. We provide ablation analysis in Fig. 8 in terms of both...
Generative adversarial network. We illustrate the results of the distortion-optimized PLVC model, denoted as PLVC (w/o GAN) in Fig. 9, compared with the proposed PLVC approach. The model of PLVC (w/o GAN) is trained only until the convergence of (3) without the adversarial loss in (4). It can be seen from Fig. 9 that the proposed PLVC approach achieves richer textures and more photo-realistic frames than PLVC (w/o GAN), even when PLVC (w/o GAN) consumes $2 \times$ to $3 \times$ bit-rates. This is also verified by the LPIPS and MOS performance in Fig. 8.

Recurrent GAN. We first remove the recurrent cells in the proposed GAN (PLVC (w/o $h_i$)). As we can see from Fig. 10, the temporal profile of PLVC (w/o $h_i$) (third column) is obviously distorted in comparison with the proposed PLVC approach and the groundtruth. As shown in Fig. 8, the LPIPS and MOS performances of PLVC (w/o $h_i$) also degrade in comparison with the proposed PLVC model.

Temporal condition $m_i$ in $D$. We further remove the temporal condition $m_i$ from $D$ and denote it as PLVC (w/o $h_i$, w/o $m_i$ in $D$). As such, $D$ becomes a normal discriminator independent along time steps. It can be seen from the right column of each example in Fig. 10 that the temporal coherence becomes even worse when further removing the $m_i$ condition in $D$. Similar result can also be observed from the quantitative and MOS results in Fig. 8. These results indicate that the long-term and short-term temporal conditions $h_i$ and $m_i$ are effective in enabling $D$ to judge raw and compressed videos according to temporal coherence, in addition to spatial texture. This way, it is able to force $G$ to generate temporally
coherent and visually pleasing videos, thus resulting in good perceptual quality.

Note that in Fig. 8 the MOS values of PLVC (w/o \( h_i \)) and PLVC (w/o \( h_i \), w/o \( m_i \) in \( D \)) are even lower than those of PLVC (w/o GAN) at some bit-rates. This is probably because the incoherent frames generated by PLVC without temporal conditions \( h_i \) and/or \( m_i \) severely degrade the perceptual quality, making it worse than the distortion-optimized model.

**Spatial condition \( y_i \) in \( D \).** Fig. 8 shows that when further removing \( y_i \) from the conditions of \( D \) (denoted as PLVC (w/o \( h_i \), w/o \( m_i \), w/o \( y_i \) in \( D \))), the performance in terms of LPIPS degrades further. We also show visual results in Fig. 11. It can be seen that, compared to PLVC (w/o \( h_i \), w/o \( m_i \) in \( D \)), the compressed frames generated by PLVC (w/o \( h_i \), w/o \( m_i \), w/o \( y_i \) in \( D \)) exhibit more artifacts and color shifts (the right example in Fig. 11). This suggests that the spatial condition in \( D \) is effective for pushing the generator \( G \) to generate compressed frames with spatial features closer to the groundtruth. It boosts the fidelity of the compressed video and improves the visual quality.

### 4.5 GOP Structure and Error Propagation

In PLVC, we follow RLVC to use the bi-IPPP [Yang et al., 2021] structure with the GOP size of nine frames (one I-frame and eight P-frames). Fig. 13 shows the average LPIPS-bpp curve for each of the eight P-frames in GOPs. The performance is averaged among all GOPs in the JCT-VC dataset. There is no obvious error propagation in the P-frames for PLVC. The possible reasons may be two-fold. On the one hand, the bi-IPPP structure shortens the distance between I-frames and P-frames, resulting in less error propagation. On the other hand, thanks to the recurrent structure of PLVC, the hidden states contain richer temporal priors when the distance to I-frames increases. This is beneficial for compression performance, and hence it can combat error propagation.

### 4.6 Generalizability of the Proposed \( D \)

Recall that in PLVC, we utilize RLVC [Yang et al., 2021] as the backbone of \( G \). However, the proposed recurrent conditional discriminator \( D \) generalizes to various compression networks. For instance, Fig. 12 shows the results of the DVC trained with the proposed \( D \) and GAN loss (blue curve) compared with the original DVC (green curve). It can be seen that the proposed \( D \) and GAN loss significantly improve the perceptual performance from the original DVC and make it outperform HM 16.20 on FID and KID. This shows that the proposed approach is not restricted to a specific compression network, but can be generalized to different backbones. It is worth mentioning that DVC trained with the proposed \( D \) performs worse than PLVC, because DVC does not contain recurrent cells, while PLVC has a recurrent \( G \) which works with the recurrent conditional \( D \), and they together make up the proposed recurrent GAN. As a result, PLVC outperforms the previous traditional and learned video compression approaches on LPIPS, FID, KID and MOS.

### 5 Conclusion

This paper proposes a GAN-based perceptual video compression approach. Its recurrent generator learns to compress video with coherent and visually pleasing frames to fool the recurrent discriminator, which learns to assess the raw vs. compressed videos conditioned on spatial-temporal features. An adversarial loss function is designed to train the proposed model towards the trade-off among bit-rate, distortion and perceptual quality. The numerical results and user studies both validate the good perceptual performance of the proposed method, in comparison with the traditional video coding standard HM 16.20 and the learned video compression methods.

### Acknowledgments

This work was supported by the ETH Zürich General Fund and by Humboldt Foundation.
References

[Agustsson et al., 2019] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In ICCV, 2019.

[Agustsson et al., 2020] Eirikur Agustsson, David Minnen, Nick Johnston, et al. Scale-space flow for end-to-end optimized video compression. In CVPR, 2020.

[Ballé et al., 2017] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In ICLR, 2017.

[Ballé et al., 2018] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In ICLR, 2018.

[Bińkowski et al., 2018] Mikolaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In ICLR, 2018.

[Bosson, 2013] Frank Bosson. Common test conditions and software reference configurations. JCTVC-L1100, 2013.

[Cheng et al., 2019] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learning image and video compression through spatial-temporal energy compaction. In CVPR, 2019.

[Choi et al., 2019] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In ICCV, 2019.

[DJelouah et al., 2019] Abdelaziz DJelouah, Joaquim Campos, et al. Neural inter-frame compression for video coding. In ICCV, 2019.

[Goodfellow et al., 2014] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. Generative adversarial nets. In NeurIPS, 2014.

[Heusel et al., 2017] Martin Heusel, Hubert Ramsauer, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In NeurIPS, 2017.

[Hu et al., 2020a] Yueyu Hu, Wenhan Yang, and Jiaying Liu. Coarse-to-fine hyper-prior modeling for learned image compression. In AAAI, 2020.

[Hu et al., 2020b] Zhihao Hu, Zhenghao Chen, Dong Xu, et al. Improving deep video compression by resolution-adaptive flow coding. In ECCV, 2020.

[Hu et al., 2021] Zhihao Hu, Guo Lu, and Dong Xu. FVC: A new framework towards deep video compression in feature space. In CVPR, 2021.

[Johnston et al., 2018] Nick Johnston, Damien Vincent, David Minnen, et al. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In CVPR, 2018.

[Lee et al., 2019] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. In ICLR, 2019.

[Lin et al., 2020] Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu. M-LVC: multiple frames prediction for learned video compression. In CVPR, 2020.

[Liu et al., 2020] Haojie Liu, Lichao Huang, Ming Lu, Tong Chen, and Zhan Ma. Learned video compression via joint spatial-temporal correlation exploration. In AAAI, 2020.

[Lu et al., 2019] Guo Lu, Wanli Ouyang, Dong Xu, et al. DVC: An end-to-end deep video compression framework. In CVPR, 2019.

[Lu et al., 2020] Guo Lu, Chunlei Cai, Xiaoyun Zhang, et al. Content adaptive and error propagation aware deep video compression. In ECCV, 2020.

[Lucie et al., 2018] Mario Lucie, Karol Kurach, Marcin Michalski, et al. Are GANs created equal? A large-scale study. In NeurIPS, 2018.

[Ma et al., 2020] Haichuan Ma, Dong Liu, Ning Yan, et al. End-to-end optimized versatile image compression with wavelet-like transform. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[Mentzer et al., 2020] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. NeurIPS, 33, 2020.

[Mercat et al., 2020] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. UVG dataset: 50/120fps 4K sequences for video codec analysis and development. In ACM MM-Sys, 2020.

[Minnen et al., 2018] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In NeurIPS, 2018.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[Ranjan and Black, 2017] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In CVPR, 2017.

[Toderici et al., 2017] George Toderici, Damien Vincent, Nick Johnston, et al. Full resolution image compression with recurrent neural networks. In CVPR, 2017.

[Veerabadran et al., 2020] Vijay Veerabadran, Reza Pourreza, et al. Adversarial distortion for learned video compression. In CVPR Workshops, 2020.

[Xue et al., 2019] Tianfan Xue, Baian Chen, Jiajun Wu, et al. Video enhancement with task-oriented flow. IJCV, 127(8):1106–1125, 2019.

[Yang et al., 2020a] Ren Yang, Fabian Mentzer, et al. Learning for video compression with hierarchical quality and recurrent enhancement. In CVPR, 2020.

[Yang et al., 2020b] Ren Yang, Luc Van Gool, and Radu Timofte. OpenDVC: An open source implementation of the dvc video compression method. arXiv preprint arXiv:2006.15862, 2020.

[Yang et al., 2021] Ren Yang, Fabian Mentzer, et al. Learning for video compression with recurrent auto-encoder and recurrent probability model. IEEE J-STSP, 2021.

[Zhang et al., 2018] Richard Zhang, Phillip Isola, Alexei A Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.