DCNGAN: A DEFORMABLE CONVOLUTION-BASED GAN WITH QP ADAPTATION FOR PERCEPTUAL QUALITY ENHANCEMENT OF COMPRESSED VIDEO

Saiping Zhang[1]*, Luis Herranz*[2] *, Marta Mrak*[3]*, Marc Górriz Blanch*[3]*, Shuai Wan*[4]*, Fuzheng Yang*[1]

[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
[2]Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain
[3]BBC Research & Development, The Lighthouse, White City Place, 201 Wood Lane, London, UK
[4]School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China

ABSTRACT

In this paper, we propose a deformable convolution-based generative adversarial network (DCNGAN) for perceptual quality enhancement of compressed videos. DCNGAN is also adaptive to the quantization parameters (QPs). Compared with optical flows, deformable convolutions are more effective and efficient to align frames. Deformable convolutions can operate on multiple frames, thus leveraging more temporal information, which is beneficial for enhancing the perceptual quality of compressed videos. Instead of aligning frames in a pairwise manner, the deformable convolution can process multiple frames simultaneously, which leads to lower computational complexity. Experimental results demonstrate that the proposed DCNGAN outperforms other state-of-the-art compressed video quality enhancement algorithms.

Index Terms— Compressed video perceptual quality enhancement, Deformable convolution, GAN, QP adaptation

1. INTRODUCTION

Recent years have witnessed the tremendous development of video compression algorithms [1] [2]. However, compressed videos, especially at low bit rate, still suffer from the degraded quality due to compression artifacts. In this case, it is crucial to enhance the quality of compressed videos.

Previous works focus on enhancing the objective quality of compressed videos [3] [4] [5]. Yang et al. [3] proposed a compressed video quality enhancement algorithm which aggregated information from neighboring high quality frames, named MFQE. By further optimizing the network in [3], Guan et al. [4] proposed MFQE 2.0 and achieved better performance. Deng et al. [5] incorporated deformable convolutions [6] to efficiently enhance the PSNR of compressed videos. However, sometimes the objective quality is inconsistent with the perceptual quality [7]. To achieve higher quality of experience (QoE) [8], many works have aimed at enhancing the perceptual quality of compressed videos [9] [10].

Wang et al. [9] proposed a generative adversarial network (GAN) based on the multi-level wavelet packet transform to enhance compressed videos. Wang et al. [10] removed visual artifacts by an enhancement network with residual blocks.

However, those algorithms require training and storing various models to enhance videos compressed at different QPs, which sets high demand on the memory. Adapting to specific QPs, Huang et al. [11] achieved QP adaptive CNN-filters by feeding the coded QP to the network. Liu et al. [12] embedded the quantization step (Qstep) into the network to adapt QP and improve coding performance.

In this paper, we propose a QP-adaptive deformable convolution-based GAN to enhance the perceptual quality of compressed videos. Thanks to QP adaptation, our method can enhance videos displayed on TV, where QP values vary greatly from frame to frame. Moreover, our method can be used for historical content enhancement to meet higher quality requirements. Our main contributions are:

1. A GAN framework based on deformable convolutions to enhance the perceptual quality of compressed videos.

2. A single adaptive model to enhance videos compressed at various QPs.

3. We compare the proposed DCNGAN with state-of-the-art compressed video quality enhancement networks, showing its superior performance.

2. PROPOSED FRAMEWORK

Our framework is illustrated in Fig. 1, and consists of a generator $\hat{\mathbf{x}}_t = G(\mathbf{y}_t, \mathbf{q})$ and a discriminator $D(\hat{\mathbf{x}}_t)$. At a particular frame index t, the generator processes a sequence of three consecutive frames $\mathbf{y}_t = (\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1})$ and the corresponding QP value $\mathbf{q}$ to output the enhanced frame $\hat{\mathbf{x}}_t$.

2.1. Generator

The generator $G(\mathbf{y}_t)$ further consists of two modules: the frame alignment module $\mathbf{z}_t = A(\mathbf{y}_t)$ and the quality en-
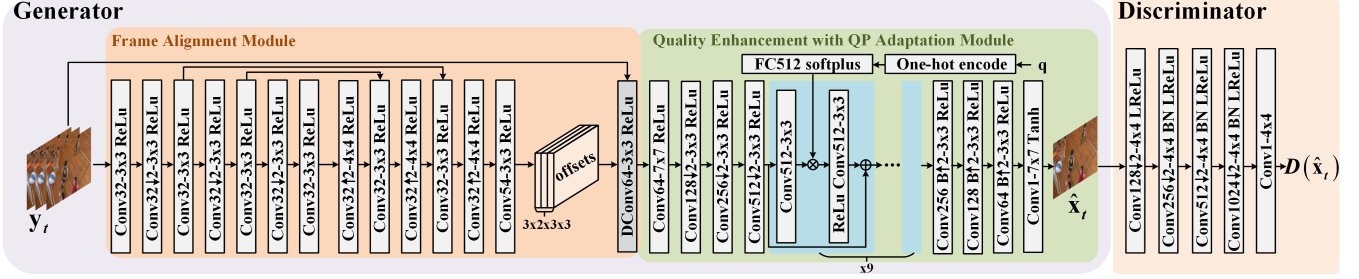
Fig. 1: The proposed DCNGAN. Conv$N$-$K \times K$ is a convolution with N channels, $K \times K$ filters. $\downarrow 2$ and $\uparrow 2$ are strided down and up by a factor 2, respectively. DConv is the deformable convolution. FCN is the fully connected layer with N dimensions. B$\uparrow 2$ is Bilinear upsampling by a factor 2. LReLu is Leaky ReLu with $\alpha = 0.2$. BN is Batch Normalization.

hancement module $\hat{\mathbf{x}}_t = E(\mathbf{z}_t)$. In order to leverage temporal information, the former takes previous, current and next frames and aligns them using a deformable convolution. The output of the deformable convolution is a single representation $\mathbf{z}_t$ which already integrates information from the three frames. The key to the correct alignment is the computation of the corresponding offsets, which are predicted by a network based on U-net [13]. The number of channels of offsets is determined by the number of the input frames (i.e. three in our case), the number of spatial dimensions (i.e. two) and the size of the convolutional kernels (i.e. 3×3 in our case). This approach is more efficient than aligning them using optical flows (generally conducted in a pairwise manner).

The quality enhancement module is based on an encoder-decoder structure with nine residual blocks. To avoid storing multiple sets of model parameters for enhancing videos compressed at different QPs, encoded QP information $\mathbf{q}$ is embedded into each residual block to make the network modulated by QP values. Specifically, $\mathbf{q}$ is represented using one-hot encoding [11] and fed into a fully connected layer. Softplus is selected as the activation function since it ensures positive outputs which we found beneficial in training. Finally, channel-wise multiplication is performed between the feature maps which are the output of the first convolution layer before ReLU in the residual block and the encoded QP information.

2.2. Discriminator

We use a patch discriminator [14] which outputs the probability of each patch being real or fake. It is implemented in a fully convolutional fashion, and the final output is the average real/fake classification probability across patches.

2.3. Objective

Our model optimizes the following adversarial loss [15] (hereinafter we omit the subindex t for simplicity)

$$L_{gan}(G, D) = \mathbb{E}_{(\mathbf{y}, \mathbf{q})}\left[(D(G(\mathbf{y}, \mathbf{q})) - 1)^2\right] + \mathbb{E}_{\mathbf{x}}\left[D(\mathbf{x})^2\right],$$
(1)

where $\mathbf{y}$ are sequences of three consecutive decoded frames and $\mathbf{q}$ is the corresponding QP value, and $\mathbf{x}$ are raw frames, both obtained from a video dataset. The generator is trained to minimize the value of Eq. (1) while the discriminator is trained to maximize it.

We use a perceptual loss based on VGG features [16] that enforces that the features of a given pair $(\mathbf{y}, \mathbf{q})$ match the features of the corresponding target raw frame $\mathbf{x}$.

$$L_{vgg}(G) = \mathbb{E}_{(\mathbf{y}, \mathbf{q}, \mathbf{x})} \sum_{i=1}^{N_f} \frac{1}{M_i} \sum_{j=1}^{M_i} \left\| f_j^i(\mathbf{x}) - f_j^i(G(\mathbf{y}, \mathbf{q})) \right\|_1,$$
(2)

where $f_j^i(\cdot)$ represents the j-th spatial element of the output tensor of the i-th layer from a pre-trained VGG-19 model.

Similarly, we also enforce matching the features of enhanced and of raw images in the discriminator.

$$L_{fm}(G, D) = \mathbb{E}_{(\mathbf{y}, \mathbf{q}, \mathbf{x})} \sum_{i=1}^{N_g} \frac{1}{M_i^g} \sum_{j=1}^{M_i^g} \left\| g_j^i(\mathbf{x}) - g_j^i(G(\mathbf{y}, \mathbf{q})) \right\|_1,$$
(3)

where $g_j^i(\cdot)$ represents the j-th spatial element of the output tensor of the i-th layer selected from the discriminator.

Finally, during training we optimize

$$\min_G \max_D \left(L_{gan}(G, D) + L_{vgg}(G) + L_{fm}(G, D)\right).$$
(4)

3. EXPERIMENTAL RESULTS

3.1. Dataset and Training Strategy

We employ 106 sequences collected by [4] for training and the Joint Collaborative Team on Video Coding (JCT-VC) standard test sequences for testing. All sequences are compressed by H.265/HEVC software HM16.5 under Low Delay P configuration at QPs 22, 27, 32 and 37. Note that the 106 raw sequences and their compressed versions are randomly cropped into 128×128 clips as training samples. The training dataset includes mixed and shuffled samples at the four QPs.

Table 1: Overall performance on LPIPS and DISTS of JCT-VC standard test sequences at four QPs

QP	Sequences	Compressed LPIPS	Compressed DISTS	MFQE 2.0 [4] LPIPS	MFQE 2.0 [4] DISTS	STDF [5] LPIPS	STDF [5] DISTS	MW-GAN [9] LPIPS	MW-GAN [9] DISTS	VPE-GAN [10] LPIPS	VPE-GAN [10] DISTS	Proposed LPIPS	Proposed DISTS
	Class A Traffic	0.170	0.014	0.184	0.014	0.094	0.009	0.138	—	0.179	0.029	**0.070**	**0.006**
	PeopleOnStreet	0.150	0.018	0.167	0.018	0.133	0.010	0.130	—	0.135	0.015	**0.086**	**0.008**
	Class B Kimono	0.258	0.043	0.294	0.046	0.160	0.026	0.189	—	0.180	0.034	**0.108**	**0.023**
	ParkScene	0.276	0.044	0.286	0.045	0.182	0.027	0.244	—	0.196	0.037	**0.123**	**0.023**
	Cactus	0.260	0.022	0.288	0.022	0.136	0.012	0.151	—	0.126	0.017	**0.096**	**0.010**
	BQTerrace	0.215	0.032	0.241	0.034	0.152	0.021	0.116	—	0.140	0.040	**0.113**	**0.018**
	BasketballDrive	0.247	0.028	0.279	0.031	0.166	0.022	0.141	—	0.132	0.025	**0.099**	**0.015**
	Class C RaceHorses	0.147	0.066	0.174	0.075	0.120	0.061	0.126	—	0.101	0.055	**0.089**	**0.042**
32	BQMall	0.124	0.066	0.145	0.071	0.089	0.050	0.091	—	0.112	0.063	**0.072**	**0.038**
	PartyScene	0.101	0.057	0.126	0.060	0.067	0.042	**0.026**	—	0.091	0.045	0.075	**0.029**
	BasketballDrill	0.156	0.073	0.181	0.079	0.126	0.068	0.109	—	0.105	0.060	**0.072**	**0.040**
	RaceHorses	0.122	0.121	0.143	0.132	0.098	0.113	0.117	—	0.093	0.126	**0.072**	**0.091**
	Class D BQSquare	0.110	0.150	0.121	0.160	0.084	0.130	0.073	—	**0.066**	**0.112**	0.104	0.123
	BlowingBubbles	0.102	0.117	0.111	0.128	0.068	0.104	**0.063**	—	0.072	0.096	0.065	**0.084**
	BasketballPass	0.116	0.135	0.135	0.150	0.099	0.127	0.095	—	0.085	0.116	**0.067**	**0.099**
	Class E FourPeople	0.120	0.037	0.128	0.038	0.089	0.022	0.080	—	0.103	0.028	**0.054**	**0.016**
	Johnny	0.148	0.035	0.159	0.035	0.111	0.021	0.083	—	0.178	0.059	**0.063**	**0.014**
	KristenAndSara	0.134	0.038	0.148	0.039	0.106	0.025	0.108	—	0.136	0.046	**0.062**	**0.019**
	Average	0.164	0.061	0.184	0.065	0.116	0.049	0.115	—	0.124	0.056	**0.083**	**0.039**
22	*Average*	0.077	0.020	0.087	0.022	0.050	**0.014**	—	—	0.097	0.047	**0.042**	0.017
27	*Average*	0.116	0.037	0.130	0.040	0.077	0.029	—	—	0.103	0.054	**0.059**	**0.026**
37	*Average*	0.223	0.089	0.232	0.086	0.168	0.080	0.177	—	0.148	0.070	**0.120**	**0.058**

We employ Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. Batch size is set to be 32. Learning rate remains unchanged at 10^{-4}. For a fair comparison with previous work, we only enhance the luminance component.

3.2. Quantitative and Qualitative Comparison

We compare the proposed DCNGAN with state-of-the-art video quality enhancement networks in [4] (MFQE 2.0), [5] (STDF), [9] (MW-GAN) [1] and [10] (VPE-GAN). LPIPS [17] and DISTS [18] are employed to quantitatively evaluate the perceptual quality of enhanced videos. For better illustration, LPIPS and DISTS of "Compressed" (i.e. the input videos) are also shown in Table 1. Smaller values indicate better perceptual quality. It should be noted that we only train one model for testing sequences compressed at four different QPs, while the other networks need to train four models since their proposed network cannot achieve QP adaptation. As shown in Table 1, the proposed DCNGAN achieves the advanced performance and alleviates memory requirements

at the enhancement stage.

Besides, three examples are randomly selected to qualitatively illustrate the performance of the proposed DCNGAN compared with the other video quality enhancement networks. Specifically, as shown in Fig. 2, MFQE 2.0 and STDF, which were designed to improve the PSNR of compressed videos, still tend to produce blurred results and penalize the perceptual quality. As for VPE-GAN, it alleviates blurring to some extent, but sometimes generates artifacts (e.g. the artifacts on the horse and the wall in Fig. 2). These artifacts also result in degraded perceptual quality, while the proposed DCNGAN can generate more realistic high-frequency details to combat with blurring and greatly improve the perceptual quality of compressed videos.

3.3. QP Adaptation Performance Evaluation

To evaluate the ability of QP adaptation of the proposed DCN-GAN, we separately train four models at the four QPs, represented by "Trained_QP22", "Trained_QP27", "Trained_QP32" and "Trained_QP37", and compare these four models with the single model trained to adapt to various QPs, represented by "Trained_4QPs". The performance comparison is shown in Fig. 3, where LPIPS is averaged on all sequences in Table 1.
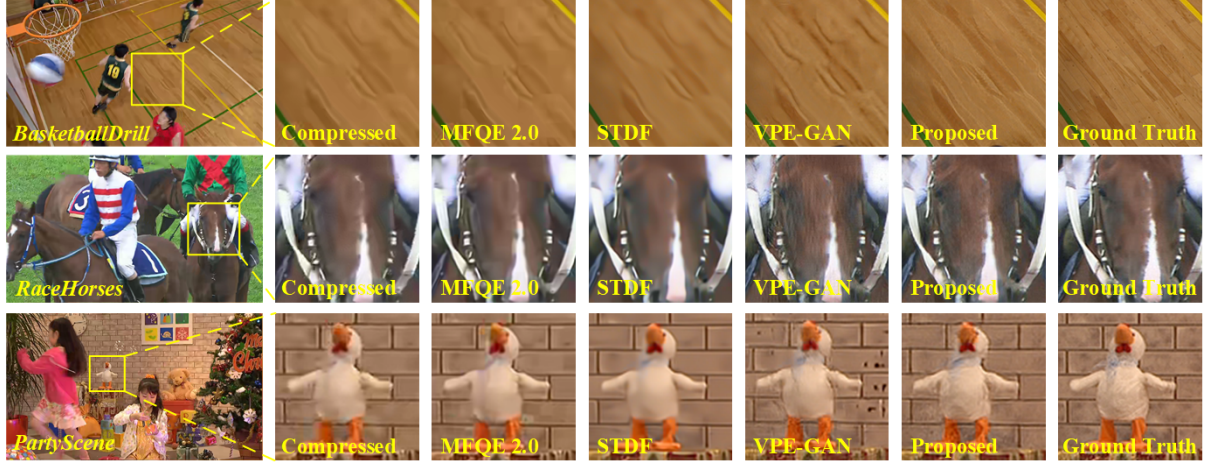
---

[1]For fair comparison, the performance of MW-GAN shown in Table 1 is from their published paper since our retrained model has not achieved the same good performance and their pretrained models at the four QPs have not been published.

Fig. 2: Comparison results of the proposed DCNGAN with the other video enhancement algorithms at QP 37.
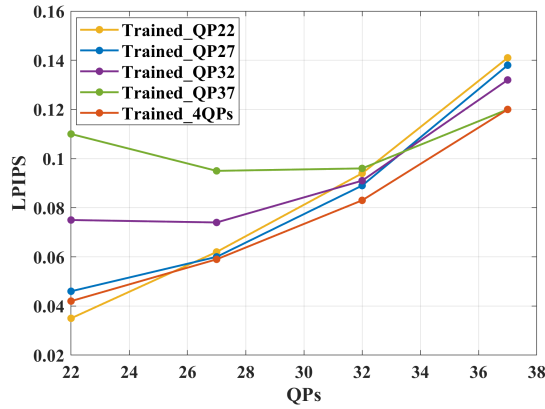

Fig. 3: LPIPS performance of models trained at different QPs.
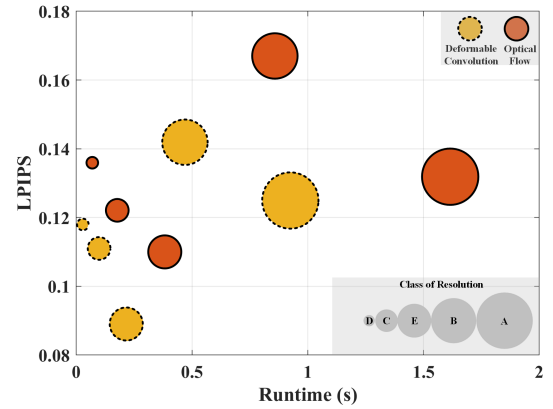

Fig. 4: LPIPS performance vs. runtime.

Overall, the model trained at a certain QP achieves the best performance tested at that QP because it has fully learned the characteristics of videos compressed at the corresponding QP in the training process, while the model can hardly achieve satisfactory performance in enhancing the perceptual quality of videos compressed at other QPs due to different characteristics. By feeding encoded QP information [2] into the proposed DCNGAN, the model "Trained_4QPs" can be employed to enhance the perceptual quality of compressed videos at various QPs with advanced performance.

3.4. Deformable Convolutions vs. Optical Flows

An alternative to deformable convolutions to align frame information is optical flow. For comparison, we replace deformable convolutions with optical flows estimated by a pretrained SpyNet model [19], in order to achieve frame align-

ments. The other modules of the proposed DCNGAN (i.e., the quality enhancement with QP adaptation module, the discriminator and the loss functions) are kept unchanged. Finally, the average LPIPS performance and runtime of two approaches are compared. As shown in Fig. 4, the network using deformable convolutions is much faster and achieves better performance than that using optical flows, which highlights the accuracy and efficiency of deformable convolutions.

4. CONCLUSION

In this paper, DCNGAN is proposed to enhance the perceptual quality of videos compressed at various QPs. By incorporating the deformable convolution to align temporal neighboring frames and feeding encoded QP to modulate the network, the proposed DCNGAN achieves the advanced performance and saves model parameters. Experimental results showed the superiority of the proposed DCNGAN compared with state-of-the-art compressed video quality enhancement networks.

---

[2]HM 16.5 compresses videos with a small QP variation. We have tried to feed QP of each frame to the network, but haven't see benefit of retraining with small delta QP. Hence, we only feed the QP of I frame to the network.

5. REFERENCES

[1] Thomas Wiegand, Gary J. Sullivan, Gisle Bjontegaard, and Ajay Luthra, "Overview of the h.264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] Ren Yang, Mai Xu, Zulin Wang, and Tianyi Li, "Multi-frame quality enhancement for compressed video," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6664–6673.

[4] Zhenyu Guan, Qunliang Xing, Mai Xu, Ren Yang, Tie Liu, and Zulin Wang, "Mfqe 2.0: A new approach for multi-frame quality enhancement on compressed video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 949–963, 2021.

[5] Jianing Deng, Li Wang, Shiliang Pu, and Cheng Zhuo, "Spatio-temporal deformable convolution for compressed video quality enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 10696–10703.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, "Deformable convolutional networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 764–773.

[7] Yochai Blau and Tomer Michaeli, "The perception-distortion tradeoff," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6228–6237.

[8] Christos George Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik, "Study of temporal effects on subjective video quality of experience," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5217–5231, 2017.

[9] Jianyi Wang, Xin Deng, Mai Xu, Congyong Chen, and Yuhang Song, "Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of compressed video," *European Conference on Computer Vision*, pp. 405–421, 2020.

[10] Ting Wang, Jingbo He, Shuhua Xiong, Pradeep Karn, and Xiaohai He, "Visual perception enhancement for hevc compressed video using a generative adversarial network," in *2020 International Conference on UK-China Emerging Technologies (UCET)*, 2020, pp. 1–4.

[11] Zhijie Huang, Jun Sun, Xiaopeng Guo, and Mingyu Shang, "One-for-all: An efficient variable convolution neural network for in-loop filter of vvc," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[12] Chao Liu, Heming Sun, Jiro Katto, Xiaoyang Zeng, and Yibo Fan, "A qp-adaptive mechanism for cnn-based filter in video coding," *arXiv preprint*, arXiv:2010.13059, Oct. 2020.

[13] Olaf Ronneberger, Fischer Philipp, and Brox Thomas, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[15] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2813–2821.

[16] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[17] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[18] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[19] Anurag Ranjan and Michael J. Black, "Optical flow estimation using a spatial pyramid network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2720–2729.