DEFORMABLE CONVOLUTION DENSE NETWORK FOR COMPRESSED VIDEO QUALITY ENHANCEMENT

Jiahui Liu *Mingcai Zhou* *Meng Xiao*

Alibaba Group

ABSTRACT

Different from the traditional video quality enhancement, the goal of compressed video quality enhancement is to reduce the artifacts brought by the video compression. The existing multi-frame methods for compressed video quality enhancement heavily rely on optical flow, which is both inefficient and limited in performance. In this paper, a Multi-frame Residual Dense Network (MRDN) with deformable convolution is developed to improve the quality of the compressed video, by utilizing high-quality frame to compensate the low-quality frame. Specifically, the proposed network consists of the developed Motion Compensation (MC) module and Quality Enhancement (QE) module, aiming to compensate and enhance the quality of the input frame, respectively. Besides, a novel edge enhancement loss is conducted on the enhanced frame, in order to enhance edge structure during the training. Finally, the experimental results over a public benchmark show that our method outperforms the state-of-the-art methods for compressed video quality enhancement task.

Index Terms— compressed video quality enhancement, deformable convolution, compression artifact reduction

1. INTRODUCTION

Aiming to reduce the required bandwidth and storage space, video compression algorithms are widely used in many real application scenarios [1], but these algorithms also bring about the problem of video quality degradation. Therefore, how to enhance the quality of compressed videos is a common concern of research community and industry. As an important approach to reduce compression artifacts, compressed video quality enhancement ranges from techniques of removing blocking artifact, reducing edge/texture floating and reconstructing videos from mosquito noises and jerkiness [2]. However, since details are lost during compression, it is a challenge to reconstruct high-quality frames from distorted frames.

Recently, many methods have been proposed for compressed image/video quality enhancement, especially with the help of deep learning [3, 4, 5]. Early methods [5, 6, 7, 8] enhance each frame independently, these methods are simple but fail to exploit details from adjacent frames. In order to leverage temporal information, Yang et al. [9] first proposed a multi-frame strategy for compressed video quality enhancement. Since then, Guan et al. [10] have further developed this method by refining the key modules. However, since video contents are seriously distorted by compression artifacts, the optical flow method used in most of the existing multi-frame-based methods is not reliable enough. As a result, the enhanced videos are far from satisfactory.

Compared with predicting optical flow at the pixel level, extracting features in a receptive field can be more robust for compressed video quality enhancement. Motivated by this idea, we proposed a multi-frame network with deformable convolution [11, 12] to achieve motion compensation for multiple moving objects. Specifically, different from traditional deformable convolutional network used in video tasks[13, 14, 15], we develop a new pyramidal deformable structure to extract multi-scale alignment information, and add a new constrain to reduce the noise from the reference frames.

Furthermore, motivated by the application of residual dense network in image super-resolution task [16], we develop a new dense connected network with residual blocks, called MRDN, to further improve the ability of extracting more hierarchical features, and then achieve better compressed video quality enhancement. In addition, through the analysis on the compressed videos, we discover that the quality degradation of the compressed video usually occurs on the object edges in the video. Thus, an edge enhancement loss is designed to make the network focus more on the edge reconstruction.

The main contributions of this paper are summarized as follows: (1)We propose a novel method for compressed video quality enhancement. This method develops a new pyramidal deformable structure with an effective constrain for motion compensation, and adopts a residual dense network for quality enhancement; (2)Through the analysis on the causes of compressed videos quality degradation, we develop a novel loss to improve the performance of edge reconstruction; (3)We evaluate the proposed method on compressed video quality enhancement benchmark database, and achieve state-of-the-art performance.

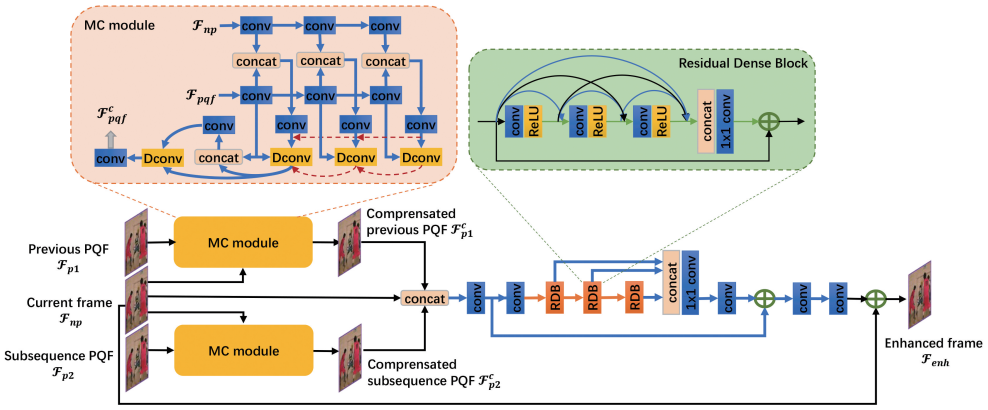

Fig. 1. The proposed network for video quality enhancement

2. THE PROPOSED SYSTEM

2.1. Overview

Different from the traditional video quality enhancement, the goal of compressed video enhancement is to reduce or remove the artifacts and blur brought by the video compression. To this end, inspired by [9], we propose a multi-frame network that is able to utilize Peak Quality Frames (PQFs)* to compensate the low-quality frame, as a result, improves the quality of the compressed video. As illustrated in Fig. 1, the network consists of the developed MC module and QE module. let $\mathcal{F}_{np}$ denotes the current frame, $\mathcal{F}_{p1}$ and $\mathcal{F}_{p2}$ are the closest previous and subsequent PQFs, respectively. Taking PQF ($\mathcal{F}_{p1}$ or $\mathcal{F}_{p2}$) as the reference frame, a deformable convolution based MC module is developed to predict the temporal motion and compensate the input frame $\mathcal{F}_{np}$ with more details. Subsequently, the compensated frames are concatenated as the input of the QE module, which is developed to further enhance the quality of the frame. Finally, a novel edge enhancement loss is conducted on the enhanced frame, in order to enhance edge structure during the training. The details of our new MC, QE and edge enhancement loss are introduced in the following sections.

2.2. MC module

For traditional deformable convolutional network in video-related tasks, most of them learn an offset δ on the reference frame, and then use deformable convolution to extract the aligned features on the current frame. The obtained aligned features F_a with N pixels are defined as:

$$F_a = DConv(F, \delta) = \sum_{i=1}^{N} \sum_{k=1}^{K^2} w_k \cdot F(p_i + p_k + \delta_k) \quad (1)$$

Where F is the features of the current frame, defined as F_{np} in this paper. p_i is the i-th location in F, K is the size of the convolution kernel, w_k is the weight of the k-th location, and p_k is the pre-specified offset of the k-th location. For example, $p_k \in (-1,-1),(-1,0),\ldots,(1,1)$ for $K=3$.

Considering that there are usually multiple moving objects on the frame at the same time, we construct multiple deformable convolution into a pyramidal structure to extract multi-scale alignment features, and enhance information interaction through the cascade. Specifically, the pyramidal deformable structure has 3 layers, each layer extracts the aligned features from the input with different resolution. The deeper the layer, the smaller the input resolution. At the same time, through cascading, the offset δ^l and aligned features F_a^l of the l-th layer are merged with δ^{l+1} and F_a^{l+1} of the next layer. The aligned features can be defined as:

$$\delta^l = f(g([F_{pqf}^l, F_{np}^l]), (\delta^{l+1})^{\uparrow 2}) \quad (2)$$

$$F_a^l = h(DConv(F_{np}^l, \delta^l), (F_a^{l+1})^{\uparrow 2}) \quad (3)$$

f, g, and h are all nonlinear transformation layers using ReLU activation, and $(\cdot)^{\uparrow s}$ is up-sampling by a factor s. s is 2 in this paper. Finally, using another convolutional layer to predict the compensated frame on the aligned features.

2.3. QE module

After obtaining compensated PQFs ($\mathcal{F}_{p1}^c$ and $\mathcal{F}_{p2}^c$), QE module is needed to fuse the information between compensated frames and current frame, then further improve the quality of current frame. In order to improve the long-term memory ability of the QE module, we adopt a residual dense network with parameters θ_{qe} to extract more hierarchical features. The proposed QE module concatenates the compensated frames and current frame as input, and then output a residual $R_{\theta_{qe}}$. A higher quality frame $\mathcal{F}_{enh}$ can be generated through adding this residual to the current frame:

$$\mathcal{F}_{enh} = \mathcal{F}_{np} + R_{\theta_{qe}}([\mathcal{F}_{p1}^c, \mathcal{F}_{np}, \mathcal{F}_{p2}^c]) \quad (4)$$

2.4. Loss functions

Edge enhancement loss. In compressed video quality enhancement, Mean Squared Error (MSE) is widely used. However, MSE loss cannot well guide the network to improve the

*PQF indicate the frame whose quality is higher than its previous and subsequent frames. In this paper, we detect the intra frame as PQF.

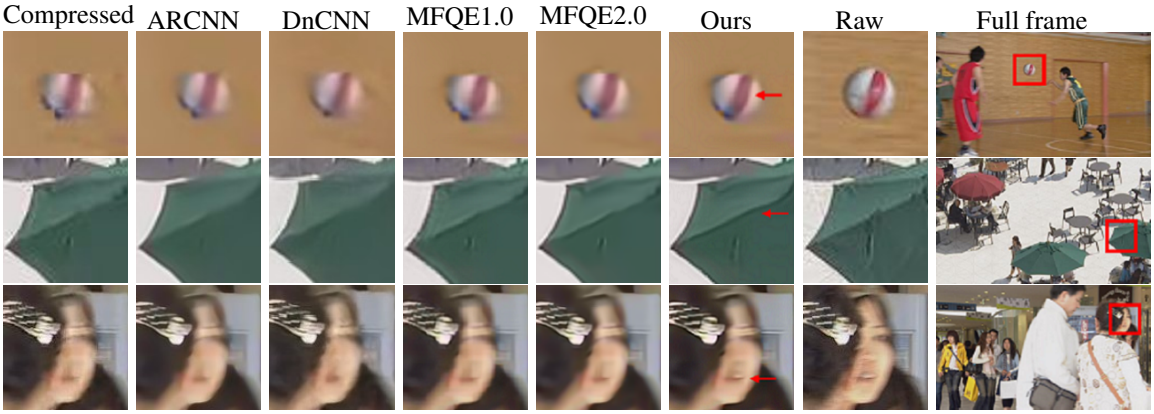

Fig. 2. Quality enhancement performance of different method at QP 37

quality of the object edges. In order to make the network pay more attention to edge reconstruction, we propose an edge enhancement loss. Given an enhanced frame $\mathcal{F}_{enh}$ which contains N pixels and its corresponding raw frame $\mathcal{F}_{raw}$, the edge enhancement loss between them is defined as:

$$L_e = \frac{1}{N}\sum_{i=1}^{N} W * (\mathcal{F}_{raw} - \mathcal{F}_{enh})^2 \tag{5}$$

Where $W = min(|dx| + |dy|, \alpha) + I$, $dx = \frac{\mathcal{F}^G_{raw}}{\partial x}$ is the derivation of the raw frame after Gaussian filtering $\mathcal{F}^G_{raw}$ in the x direction, and similarly $dy = \frac{F^G_{raw}}{\partial y}$ is the derivation of $\mathcal{F}^G_{raw}$ in the y direction. The edges information of the raw frame can be extracted by calculating the sum of dx and dy. In addition, in order to limit the influence of the edge with large response, we also limit the edge response value not to exceed α.In our work $\alpha = 0.45$ is used.

Total Loss. Different from the other video tasks, compressed video quality enhancement task is very sensitive to noise. Therefore, we not only optimizer the parameters θ_{qe} of the QE module, but also add constrains on the parameters θ_{mc} of the MC module. Specifically, for the MC module, the compensated frames are not only required to provide the aligned results, but also supposed to reserve the similar quality of the raw frame $\mathcal{F}_{raw}$. For the QE module, the enhanced frames $\mathcal{F}_{enh}$ are required to be as high quality as the raw frame. As such, the total loss is defined as:

$$L = a \cdot \sum_{i=1}^{2} L_e(\mathcal{F}^c_{pi}, \mathcal{F}_{raw}) + b \cdot L_e(\mathcal{F}_{enh}, \mathcal{F}_{raw}) \tag{6}$$

3. EXPERIMENT

3.1. Datasets

To train the proposed model, Guan et al.'s database [10] is utilized. This database consists of 160 uncompressed videos selected from datasets of Xiph.org, VQEG and Joint Collaborative Team on Video Coding (JCT-VC) [17], 106 videos of them are used for training. For testing, the proposed model is evaluated on 18 standard test videos [18], these videos are collected from JCT-VC and are widely used for video quality assessment. All the above videos are compressed by HM 16.5 in LDP mode, with 4 different QPs, i.e., 22, 27, 32, 37.

3.2. Implementation Details

During training, we use compressed frame and its previous and subsequent PQFs as input frames, then randomly crop these frames into 64x64 patches. Then, the model is optimized by Adam optimizer [19] with initial learning rate of 1e-4, and batch size is set to 16. Besides, the loss weight of the motion compensation module and quality enhancement module is set to 1 and 0.001 during training, respectively. After 100,000 iterations, the loss weight of these two modules will be changed to 0.001 and 1. Note that we train four models for QP 22, 27, 32 and 37. In addition, in order to compare the experimental results more clearly, our and compared methods are evaluated in terms of incremental Peak Signal-to-Noise Ratio (△PSNR) and Structural Similarity [20] (△SSIM), which measure the PSNR and SSIM difference between the enhanced and the raw frames.

3.3. Comparison with state of the art

Quantitative comparison. We compare the proposed method with five state-of-the-art methods, the △PSNR and △SSIM results are reported in Table 1. Among the compared methods, ARCNN [5], DnCNN [6] and RNAN [7] are methods for compressed image quality enhancement, they enhance each frame independently and have limited performance. MFQE 1.0 [9] proposed a novel strategy that looking for PQFs near the current frame, and extracting more information from multiple frames. On the basis of MFQE 1.0, MFQE 2.0 [10] further improves the performance by working with better PQF detector and QE module. In our work, an effective pyramidal deformable structure and a residual dense network are developed for multi-frame strategy. As one can see that the proposed method achieves better △PSNR and △SSIM than another five methods. Even more, for QP 37, our improvement relative to MFQE 2.0 is twice that of MFQE 2.0 relative

Table 1. Comparisons of six methods on 18 test sequences with 4 different QPs in terms of △PSNR △SSIM ($\times 10^{-2}$).

QP	Sequences		AR-CNN[5]	DnCNN[6]	RNAN[7]	MFQE 1.0[9]	MFQE 2.0[10]	Ours
37	Class A	Traffic	0.27/0.50	0.35/0.64	0.40/0.86	0.50/0.90	0.59/1.02	**0.72/1.16**
		PeopleOnStreet	0.37/0.76	0.54/0.94	0.74/1.3	0.80/1.37	0.92/1.57	**1.23/1.99**
	Class B	Kimono	0.20/0.59	0.27/0.73	0.33/0.98	0.50/1.13	0.55/1.18	**0.82/1.65**
		ParkScene	0.14/0.44	0.17/0.52	0.20/0.77	0.39/1.03	0.46/1.23	**0.60/1.54**
		Cactus	0.20/0.41	0.28/0.53	0.35/0.76	0.44/0.88	0.50/1.00	**0.67/1.30**
		BQTerrace	0.23/0.43	0.33/0.53	0.42/0.84	0.27/0.48	0.40/0.67	**0.55/0.97**
		BasketballDrive	0.23/0.51	0.33/0.63	0.43/0.92	0.41/0.80	0.47/0.83	**0.71/1.25**
	Class C	RaceHorses	0.23/0.49	0.31/0.70	0.39/0.99	0.34/0.55	0.39/0.80	**0.60/1.48**
		BQMall	0.28/0.69	0.38/0.87	0.45/1.15	0.51/1.03	0.62/1.20	**0.90/1.73**
		PartyScene	0.14/0.52	0.22/0.69	0.30/0.98	0.22/0.73	0.36/1.18	**0.55/1.66**
		BasketballDrill	0.23/0.48	0.42/0.89	0.50/1.07	0.48/0.90	0.58/1.20	**0.73/1.54**
	Class D	RaceHorses	0.26/0.59	0.34/0.80	0.42/1.02	0.51/1.13	0.59/1.43	**0.83/2.09**
		BQSquare	0.21/0.30	0.30/0.46	0.32/0.63	-0.01/0.15	0.34/0.65	**0.75/1.07**
		BlowingBubbles	0.16/0.46	0.25/0.76	0.31/1.08	0.39/1.20	0.53/1.70	**0.66/2.08**
		BasketballPass	0.26/0.63	0.38/0.83	0.46/1.08	0.63/1.38	0.73/1.55	**0.98/2.03**
	Class E	FourPeople	0.40/0.56	0.54/0.73	0.70/0.97	0.66/0.85	0.73/0.95	**0.94/1.18**
		Johnny	0.24/0.21	0.47/0.54	0.56/0.88	0.55/0.55	0.60/0.68	**0.75/0.78**
		KristenAndSara	0.41/0.47	0.59/0.62	0.63/0.8	0.66/0.75	0.75/0.85	**0.95/1.01**
	Average		0.25/0.50	0.36/0.69	0.44/0.95	0.46/0.88	0.56/1.09	**0.78/1.47**
32	Average		0.19/0.17	0.33/0.41	0.41/0.62	0.43/0.58	0.52/0.68	**0.81/1.02**
27	Average		0.16/0.09	0.33/0.26	-/-	0.40/0.34	0.49/0.42	**0.83/0.72**
22	Average		0.13/0.04	0.27/0.14	-/-	0.31/0.19	0.46/0.27	**0.70/0.40**

to MFQE 1.0.

Qualitative comparison. Fig.2 shows the qualitative comparison of 5 methods, it's obvious that the proposed method can provide enhanced frames with higher quality. Taking the ball, the umbrella bone and the mouth in Fig.2 as examples, our method restores clearer object edges and more detail. It shows that for fast-moving objects (such as balls) in videos, our pyramidal deformable structure can more accurately compensate for motion, and with the help of an effective QE module and the guidance of edge enhancement loss, the proposed model have better performance on objects edge reconstruction and detail supplementation.

Table 2. Comparisons of different module in terms of △PSNR and △SSIM ($\times 10^{-2}$).

motion compensate module	results
optical flow estimation	0.46/0.88
pyramidal deformable convolution	0.72/1.41

quality enhancement module	results
general CNN	0.75/1.46
MRDN	0.78/1.47

3.4. Effects of the proposed module

Effects of MC module. MC module is a critical part of multi-frame strategy, to better understand the difference between the optical-flow-estimation-based MC module and the proposed deformable-convolution-based MC module, we compare the effects of these two module. We used optical flow method and the proposed pyramidal deformable convolution to train two models respectively, and results are shown in Table 2. Under the same training strategy and with the same QE module, the results of the model with our MC module have much higher quality, shows that the pyramidal deformable convolution is more reliable.

Effects of QE module. QE module extracts information from its input to further enrich the details of the compressed frame, its output determines the performance of the entire model. We used general CNN (used at MFQE 1.0) and the proposed MRDN to train two models respectively, these models use the same settings and the same MC module. Table 2 provides the evaluation results. As one can see, MRDN obtains higher results on △PSNR and △SSIM than general CNN. It means that the residual dense connection used in MRDN exploits feature information effectively, and presents a better performance than the super-large receptive field used in general CNN.

4. CONCLUSION

In this paper, we propose a novel multi-frame network for compressed video quality enhancement, which uses a pyramidal deformable structure to compensate for motion, and enhances the quality of compressed frames through a multi-frame residual dense network. In addition, an edge enhancement loss is designed for powerful edge reconstruction. The proposed model achieves state-of-the-art performance on benchmark database, and the model size of it is 1.32M, which is smaller than most of the compared methods. The future work will focus on further reducing computational complexity.

5. REFERENCES

[1] 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.

[2] Kai Zeng, Tiesong Zhao, Abdul Rehman, and Zhou Wang, "Characterizing perceptual artifacts in compressed video streams," in *Human Vision and Electronic Imaging XIX*. International Society for Optics and Photonics, 2014, vol. 9014, p. 90140Q.

[3] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8183–8192.

[4] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3883–3891.

[5] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang, "Compression artifacts reduction by a deep convolutional network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 576–584.

[6] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[7] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu, "Residual non-local attention networks for image restoration," *arXiv preprint arXiv:1903.10082*, 2019.

[8] Ren Yang, Mai Xu, and Zulin Wang, "Decoder-side hevc quality enhancement with scalable convolutional neural network," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 817–822.

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

[10] 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*, 2019.

[11] Yuanying Dai, Dong Liu, and Feng Wu, "A convolutional neural network approach for post-processing in hevc intra coding," in *International Conference on Multimedia Modeling*. Springer, 2017, pp. 28–39.

[12] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316.

[13] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu, "Tdan: Temporally-deformable alignment network for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3360–3369.

[14] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[15] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu, "Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3370–3379.

[16] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[17] Frank Bossen et al., "Common test conditions and software reference configurations," *JCTVC-L1100*, vol. 12, pp. 7, 2013.

[18] Jens-Rainer Ohm, Gary J Sullivan, Heiko Schwarz, Thiow Keng Tan, and Thomas Wiegand, "Comparison of the coding efficiency of video coding standards—including high efficiency video coding (hevc)," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1669–1684, 2012.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.