NTIRE 2021 Challenge on Quality Enhancement of Compressed Video: Dataset and Study

Ren Yang
Computer Vision Laboratory
ETH Zürich, Switzerland
ren.yang@vision.ee.ethz.ch

Radu Timofte
Computer Vision Laboratory
ETH Zürich, Switzerland
radu.timofte@vision.ee.ethz.ch

Abstract

This paper introduces a novel dataset for video enhancement and studies the state-of-the-art methods of the NTIRE 2021 challenge on quality enhancement of compressed video. The challenge is the first NTIRE challenge in this direction, with three competitions, hundreds of participants and tens of proposed solutions. Our newly collected Large-scale Diverse Video (LDV) dataset is employed in the challenge. In our study, we analyze the solutions of the challenges and several representative methods from previous literature on the proposed LDV dataset. We find that the NTIRE 2021 challenge advances the state-of-the-art of quality enhancement on compressed video. The proposed LDV dataset is publicly available at the homepage of the challenge: https://github.com/RenYang-home/NTIRE21_VEnh

Figure 1. The performance on the fidelity tracks.

1. Introduction

The recent years have witnessed the increasing popularity of video streaming over the Internet [6] and meanwhile the demands on high-quality and high-resolution videos are also increasing. To transmit larger number of high-resolution videos thought the bandwidth-limited Internet, video compression [28, 21] has to be applied to significantly reduce the bit-rate. However, compressed videos unavoidably suffer from compression artifacts, thus leading to the loss of both fidelity and perceptual quality and degrade the Quality of Experience (QoE). Therefore, it is necessary to study on enhancing the quality of compressed video, which aims at improving the compression quality at the decoder side without changing the bit-stream. Due to the rate-distortion trade-off in data compression, quality enhancement on compressed video is equivalent to reducing the bit-rate at the same quality, and hence it also can be seen as a way to improve the efficiency of video compression.

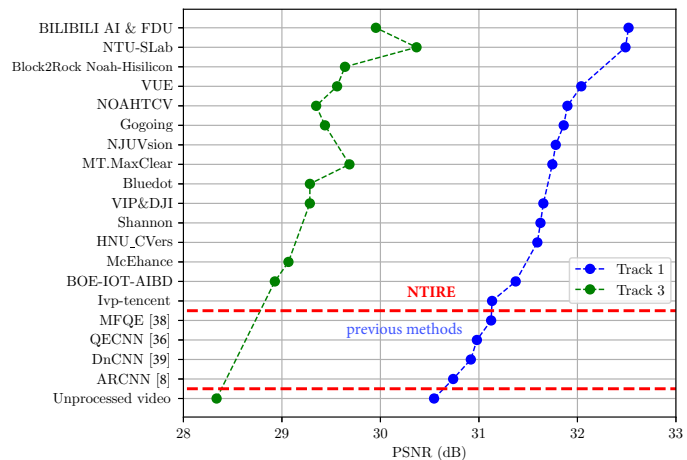In the past a few years, there has been plenty of works on

enhancing the quality of compressed video [37, 36, 25, 17, 38, 29, 34, 10, 30, 7, 33, 12, 24], among which [37, 36, 25] propose enhancing compression quality based on a single frame, while [17, 38, 34, 10, 30, 7, 33, 12, 24] are multi-frame quality enhancement methods. Besides, [24] aims at improving the perceptual quality of compressed video, which is evaluated by the Mean Opinion Score (MOS). Other works [37, 36, 25, 17, 38, 34, 10, 30, 7, 33, 12] focus on advancing the performance on Peak Signal-to-Noise Ratio (PSNR) to achieve higher fidelity to the uncompressed video.

Thanks to the rapid development of deep learning [16], all aforementioned methods are deep-learning-based and data-driven. In single-frame enhancement literature, [25] uses the image database BSDS500 [1] as the training set, without any video. [37] trains the model on a small video dataset including 26 video sequences, and then [36] enlarges the training set to 81 videos. Among the existing multi-frame works, the model of [17] is trained on the Vimeo-90K dataset [32], in which each clip only contains 7 frames, and thus it is insufficient for the research on en-
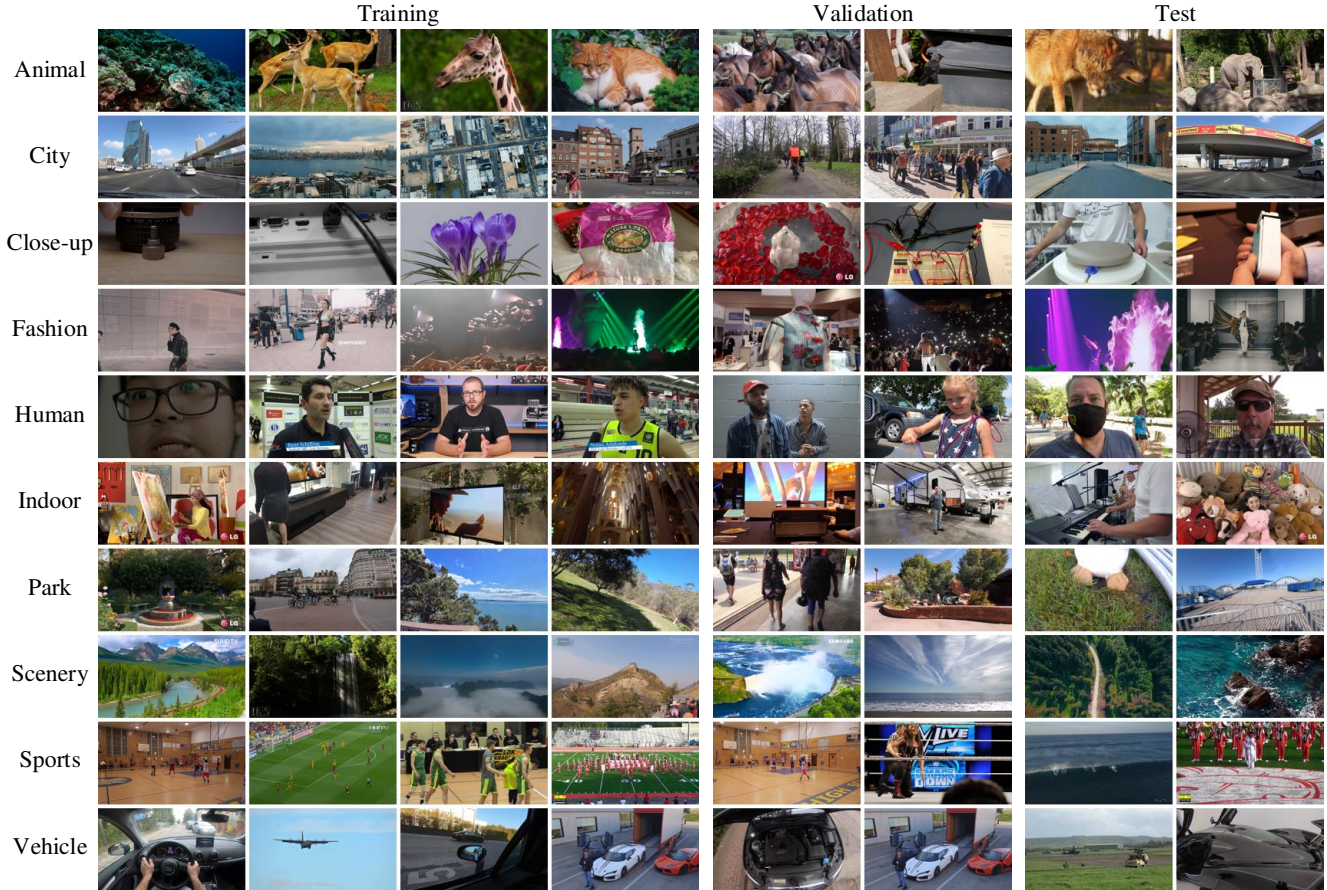
Figure 2. Example videos in the proposed LDV dataset, which contains 10 categories of scenes. The left four columns show a part of videos used for training in the NTRIE challenge. The two columns in the middle are the videos for validation. The right two columns are the test videos, in which the left column is the test set for Tracks 1 and 2, and the most right column is the test set for Track 3.

hancing long video sequences, especially not applicable for recurrent frameworks. Then, the Vid-70 dataset, which includes 70 video sequences, is used as the training set in [38, 34, 30, 12]. Meanwhile, [10] and [7] collected 142 and 106 uncompressed videos for training, respectively. [24] uses the same training set as [10]. In conclusion, all above works train the models on less than 150 video sequences, except Vimeo-90K which only has very short clips. Therefore, establishing a large scale training database with high diversity is essential to promote the future research on video enhancement. Besides, the commonly used test sets in existing literature are the JCT-VC dataset [3] (18 videos), the test set of Vid-70 [34] (10 videos) and Vimeo-90K (only 7 frames in each clip). Standardizing a larger and more diverse test set is also meaningful for the deeper studies and fairer comparisons for the proposed methods.

In this work, we collect a novel Large-scale Diverse Video (LDV) dataset, which contains 240 high quality videos with the resolution of 960×536. The proposed LDV dataset includes diverse categories of contents, various kinds of motion and different frame-rates. Moreover, we

organized the online challenge at NTIRE 2021[1] for enhancing the quality of compressed video using the LDV dataset. The challenge has three tracks. In Tracks 1 and 2, videos are compressed by HEVC [21] at a fixed QP, while in Track 3, we enable rate-control to compress videos by HEVC at a fixed bit-rate. Besides, Tracks 1 and 3 aim at improving the fidelity (PSNR) of compressed video, and Track 2 targets at enhancing the perceptual quality, which is evaluated by the MOS value.

Moreover, we also study the newly proposed LDV dataset via the performance of the solutions in the NTIRE 2021 video enhancement challenge and the popular methods from existing works. Figure 1 shows the PSNR performance on the fidelity tracks. It can be seen from Figure 1 that all NTIRE methods effectively improve the PSNR of compressed video, and the top methods obviously outperforms the the existing methods. The PSNR improvement of NTIRE methods on Track 1 ranges from 0.59 dB to 1.98 dB, and on Track 3 ranges from 0.59 dB to 2.03 dB. We also

---

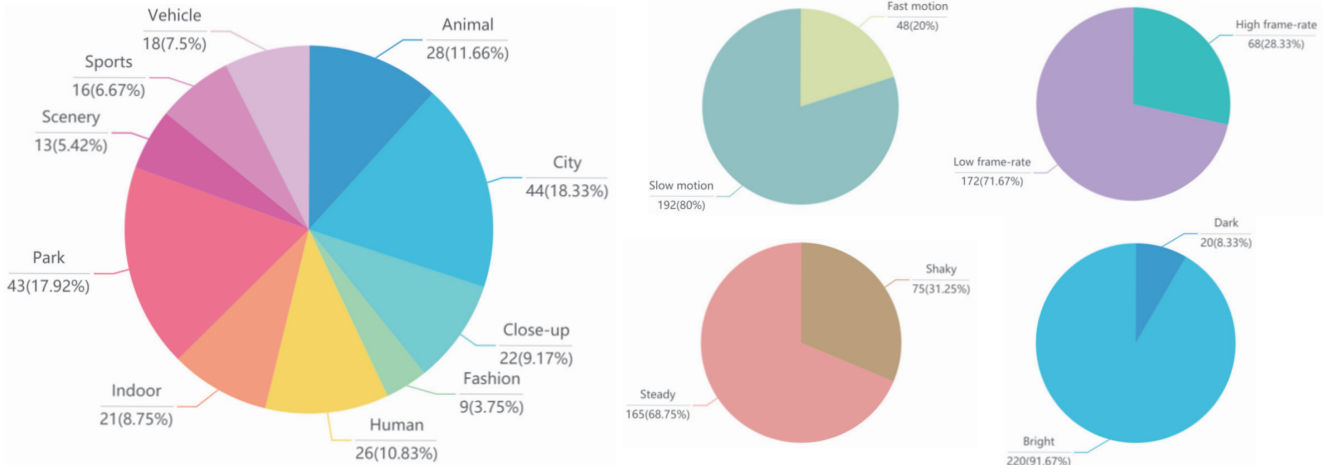[1]https://data.vision.ee.ethz.ch/cvl/ntire21/

Figure 3. The diversity of the proposed LDV dataset.

report the results in various quality metrics in Section 5 for analyses.

The remainder of the paper is as follows. Section II introduces the proposed LDV dataset, and Section III describes the video enhancement challenge in NTIRE 2021. The methods proposed in the challenge and some classical existing methods are overviewed in Section IV. Section V reports the challenge results and the analyses on the LDV dataset.

2. The proposed LDV dataset

We propose the LDV dataset with 240 high quality videos with diverse content, different kinds of motion and large range of frame-rates. The LDV dataset is intended to complement the existing video enhancement datasets to enlarge the scale and increase the diversity to establish a more solid benchmark. Some example videos in the proposed LDV dataset are shown in Figure 2.

Collecting. The videos in LDV are collected from YouTube[2]. To ensure the high quality, we only collect the videos with 4K resolution, and without obvious compression artifacts. All source videos used for our LDV dataset have the licence of *Creative Commons Attribution licence (reuse allowed)*[3]. Note that the LDV dataset is for academic and research proposes.

Diversity. We mainly consider the diversity of videos in our LDV dataset from three aspects: category of scenes, motion and frame-rate. As Figure 2 illustrates, our LDV dataset contains 10 categories of scenes, *i.e.*, *animal*, *city*, *close-up*, *fashion*, *human*, *indoor*, *park*, *scenery*, *sports* and *vehicle*. In the 240 videos of LDV, 48 videos are with fast motion. The frame-rates of the videos in LDV range from

24 to 60, in which 172 videos are with low frame-rates ($\leq$ 30) and 68 videos are with high frame-rates ($\geq$ 50). Additionally, the camera is slightly shaky (*e.g.*, captured by handheld camera) in 75 videos of LDV. The shakiness results in irregular local movement of pixels, which is a common phenomenon especially in the videos of social media (cameras hold by hands). Besides, 20 videos of LDV are in dark environments, *e.g.*, at night or in the rooms with insufficient light.

Downscaling. To remove the compression artifacts of the source videos, we downscale the videos by the factor of 4 using the Lanczos filter [23]. Then, the width and height of each video are cropped to the multiples of 8, due to the requirement of the HEVC test model (HM). We follow the standard datasets, *e.g.*, JCT-VC [3], to convert videos to the format of YUV 4:2:0.

Partition. In the challenge of NTIRE 2021, we divide the LDV dataset into three parts for training (200 videos), validation (20 videos) and test sets (20 videos), respectively. The 20 test videos are further split into two sets with 10 videos each. The two test sets are used for the track of fixed QP (Tracks 1 and 2) and the track of fixed bit-rate (Track 3), respectively. The validation set contains the videos from the 10 categories of scenes with two videos in each category. Each test set has one video from each category. Besides, 9 in the 20 validation videos and 4 of the 10 videos in each test set are with high frame-rates. There are five fast-motion videos in the validation set, and there are three and two fast-motion videos in the test sets for fixed QP and fixed bit-rate tracks, respectively.

3. NTIRE 2021 video enhancement challenge

The video enhancement challenge of NTIRE 2021 has three tracks. Tracks 1 and 2 target at enhancing the quality of video compressed by HM 16.20 at fixed QP (= 37),

---

[2]https://www.youtube.com/
[3]https://support.google.com/youtube/answer/2797468

and Track 3 is for the enhancement of video compressed by x265 at fixed bit-rate (= 200 kbps). Besides, Tracks 1 and 3 aim at improving the fidelity of compressed video, and Track 3 aims at enhancing compressed video towards perceptual quality. The websites of the three tracks are at:

Track 1: https://competitions.codalab.org/competitions/28033

Track 2: https://competitions.codalab.org/competitions/28034

Track 3: https://competitions.codalab.org/competitions/28035

3.1. Video Quality Assessment (VQA)

Fidelity. In the fidelity tracks, we evaluate the quality of enhanced video by PSNR and the Multi-Scale Structural SIMilarity index (MS-SSIM) [27]. In our challenge, PSNR and MS-SSIM are calculated in the RGB domain on each frame, and then they are averaged among all frames of each video. The rank of proposed methods in the fidelity tracks depends on PSNR, and we additionally report MS-SSIM for analyses.

User study. In the track of perceptual quality, we rank the proposed methods by the Mean Opinion Score (MOS), which ranges from 0 to 100. There are 15 subjects participating in the MOS experiment. The groundtruth videos are given to the subjects as a standard of the highest score ($s = 100$), but the subjects are asked to rate videos in accordance with the visual quality, instead of the similarity to the groundtruth. Then, we linearly normalize the scores (s) of each subject to

$$s' = 100 \cdot \frac{s - s_{min}}{s_{max} - s_{min}}, \qquad (1)$$

in which s_{max} and s_{min} denote the highest and the lowest score of each subject, respectively. To ensure the rating consistency of each subject, we insert five repeated videos to check the concentration of each subject, and the scores of the three least concentrated subjects are omitted. Therefore, the final MOS values are averaged among 12 subjects.

Perceptual metrics. In addition to the user study, we further evaluate the methods for perceptual quality by various perceptual metrics, including the Learned Perceptual Image Patch Similarity (LPIPS) [40], Fréchet Inception Distance (FID) [11], Kernel Inception Distance (KID) [2] and Video Multimethod Assessment Fusion (VMAF) [15]. LPIPS [40] measures the distance in the feature space of DNN. FID [11] and KID [2] calculates the similarity between the distributions of the groundtruth and distorted frames. VMAF [15] is developed by Netflix for assessing the perceptual quality of video. These metrics have been validated to be effective for evaluating perceptual quality.

3.2. Tracks 1 and 2: Fixed QP

In Tracks 1 and 2, videos are compressed by the official HEVC test model HM 16.20[4] with the default configuration of the Low-Delay P (LDP) mode (*encoder_lowdelay_P_main.cfg*) at QP = 37. In this configuration, the frame at each position in every Group of Pictures (GOPs) are compressed at a fixed QP, and there is no rate control during the compression. As such, the quality fluctuates regularly in each GOP and the framerate has no effect for compression. This is the widely used configuration in the existing video enhancement literature [37, 36, 38, 34, 10, 30].

3.3. Track 3: Fixed bit-rate

This track aims at a more practical scenario. Since HM 16.20 lacks practicability due to its high time complexity, we use the x265 library of FFmpeg[5] in this track. More importantly, rate control is a widely used strategy in video streaming to fit the compressed video to the limited bandwidth. Therefore, we compress the groundtruth videos by x265 with rate control at the fixed bit-rate of 200 kbps, using the following commands:

```
ffmpeg -pix_fmt yuv420p -s WxH -r FR
-i name.yuv -c:v libx265 -b:v 200k
-x265-params pass=1:log-level=error
-f null /dev/null

ffmpeg -pix_fmt yuv420p -s WxH -r FR
-i name.yuv -c:v libx265 -b:v 200k
-x265-params pass=2:log-level=error
name.mkv
```

In above commands, W and H are the width and height of the video, respectively. FR denotes the frame rate. Note that, we use the two-pass strategy to ensure the accuracy of rate control.

4. Video enhancement methods

In this section, we briefly introduce the top methods in the NTIRE 2021 challenge [35] and the existing methods in previous literature. The results of the challenge and the study on the proposed LDV dataset are to be detailed in Section 5.

4.1. Methods in the NTIRE 2021 challenge

BILIBILI AI & FDU Team [35, 31] ranks first in Tracks 1 and 2. They propose the Spatiotemporal Model

---

[4] https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.20
[5] https://johnvansickle.com/ffmpeg/releases/ffmpeg-release-amd64-static.tar.xz

Table 1. The platforms, test strategies and training data of the challenge methods.

Team	Platform	GPU	Ensemble / Fusion	Extra training data
BILIBILI AI & FDU	PyTorch	Tesla V100/RTX 3090	Flip/Rotation x8	Bilibili [13], YouTube [14]
NTU-SLab	PyTorch	Tesla V100	Flip/Rotation x8	Pre-trained on REDS [18]
VUE	PyTorch	Tesla V100	Flip/Rotation x8	Vimeo90K [32]
NOAHTCV	TensorFlow	Tesla V100	Flip/Rotation x8	DIV8K [9] (Track 2)
MT.MaxClear	PyTorch	Tesla V100	Flip/Rotation/Multi-model x12	Private dataset
Shannon	PyTorch	Tesla T4	Flip/Rotation x8 (Track 1)	-
Block2Rock Noah-Hisilicon	PyTorch	Tesla V100	Flip/Rotation x8	YouTube [14]
Gogoing	PyTorch	Tesla V100	Flip/Rotation x4	REDS [18]
NJUVision	PyTorch	Titan RTX	Flip/Rotation x6	SJ4K [20]
BOE-IOT-AIBD	PyTorch	GTX 1080	Overlapping patches	-
(anonymous)	PyTorch	Tesla V100	-	Partly finetuned from [26]
VIP&DJI	PyTorch	GTX 1080/2080 Ti	Flip/Rotation x8	SkyPixel [19].
BLUEDOT	PyTorch	RTX 3090	-	Dataset of MFQE 2.0 [10]
HNU_CVers	PyTorch	RTX 3090	Overlapping patches	-
McEhance	PyTorch	GTX 1080 Ti	-	-
Ivp-tencent	PyTorch	GTX 2080 Ti	-	-
MFQE [38]	TensorFlow	TITAN Xp	-	-
QECNN [36]	TensorFlow	TITAN Xp	-	-
DnCNN [39]	TensorFlow	TITAN Xp	-	-
ARCNN [8]	TensorFlow	TITAN Xp	-	-

with Gated Fusion (SMGF) for enhancing quality of compressed video. The proposed SMGF method explores the temporal correlation by selecting eight reference frames to benefit the enhancement of each target frame. The selection of reference frames considers temporal distance and QP values. The nine frames are fed into the Spaito-Temporal Deformable Fusion (STDF) [7] module to explore both spacial and temporal features, and then these features are fed to the Quality Enhancement (QE) module. Additionally, they propose training two models on two datasets, and both models consist of STDF and QE. Finally, they fuse the two enhanced frames to improve the quality performance.

NTU-SLab Team is the winner of Track 3, and ranks at the the second place on average PSNR and MOS for Tracks 1 and 2, respectively. The method proposed by the NTU-SLab Team is called BasicVSR++ [35], which is based on BasicVSR [4] but additionally consists of deliberate improvements on propagation and alignment. Specifically, in BasicVSR++, residual blocks are first applied to extract features from each frame. The features are then propagated under the proposed second-order grid propagation scheme, where alignment is performed by the proposed flow-guided deformable alignment. After propagation, the aggregated features are used to generate the output image through convolution and pixel-shuffling.

VUE Team [35] proposes combining and stacking several BasicVSR [4] models for enhancing the fidelity of compressed videos. For the perpetual track, they propose the method called "Adaptive Spatial-Temporal Fusion of Two-Stage Multi-Objective Networks" [41]. Specifically, The first stage aims at obtaining the relatively good intermediate results with high fidelity. In this stage, a BasicVSR model is trained with Charbonnier loss [5]. At the second stage, they

train two BasicVSR models for different refinement purposes. One refined BasicVSR model is trained with a combination of the Charbonnier loss [4] and LIPIS loss [40]. Another refined BasicVSR model is only trained with the LPIPS loss [40]. Then, to fuse the results from the two networks, they design a novel adaptive spatial-temporal fusion scheme, which is able to eliminate flickering effects.

NOAHTCV Team [35] follows MFQE [38] to use Peak Quality Frames (PQFs) to facilitate the enhancement of other frames. The target frame and two neighboring PQFs are fed into the proposed network. They first use a deep network to estimate the offsets of the three input frames, and then the offsets are used to align and fuse the features. After the initial feature extraction and alignment, a multi-head U-Net is applied to process each feature, and they fuse the features with scale-dependant deformable convolutions at each scale. Finally, the output features are fused and fed into residual blocks to reconstruct the enhanced frame. They use the $L2$ loss for Tracks 1 and 3, and the GAN Loss + Perceptual loss + $L2$ loss for Track 2.

MT.MaxClear Team [35] proposes utilizing Charbonnier penalty loss, DCN offsets Total Variation loss and DCN offsets Variation loss to the EDVR [26] network to improve the stability of training. DCN offsets Total Variation loss encourages the predicted DCN offsets are smooth in spatial space. DCN offsets Variation loss encourages the predicted DCN offsets between different channels do not deviate too much from the offsets mean. In Track 2, they add the sharpening operation on the enhanced frames for better visual perception.

(a) PSNR on Track 1 (unprocessed video = 30.54 dB) (b) MS-SSIM on Track 1 (unprocessed video = 0.9305)
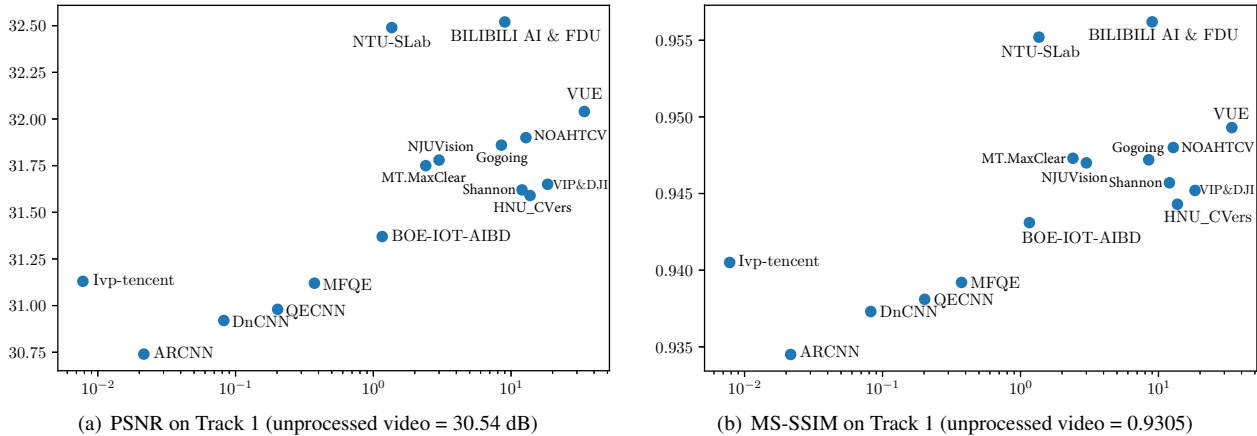
Figure 4. The results on Track 1. The running time is reported by the authors of the teams. The platforms used for test are listed in Table 1.

4.2. Methods in previous literature

ARCNN [8] and **DnCNN** [39] are the deep neural networks proposed for reducing the compression artifacts of JPEG images. ARCNN is the first method in this direction, which consists of four convolutional layers. DnCNN is a deeper network with 20 convolution layers and uses batch normalization. DnCNN learns to estimate the residual image, instead of directly reconstructing the enhanced image in ARCNN. These image enhancement methods are the fundamental for video enhancement. In this report, we re-train these networks by the training videos of our LDV dataset.

QE-CNN [36, 37] is a single-frame quality enhancement method proposed for enhancing HEVC compressed video. In QE-CNN, a five-layer convolutional neural network, called QE-CNN-I, is designed to enhance I-frames. Moreover, additional five CNN layers are proposed to extract and handle the features of inter-coding frames, which are then combined with the features of intra-frames extracted by QE-CNN-I to enhance the quality of B- and P-frames in HEVC. The network for enhancing B- and P-frames is named as QECNN-B. The original QE-CNN is designed for only enhancing the Y channel. In this report, we re-train QE-CNN on the LDV training videos in the RGB domain.

MFQE [38, 10] is proposed to enhance compressed video taking advantage of higher quality neighboring frames. In MFQE, the Peak Quality Frames (PQFs) are defined as the frames with higher quality then its previous and next frames, and these frames are used to benefit the enhancement of other frames. First, an SVM-based [38] or LSTM-based [10] PQF detector is used to estimate the PQFs and then the MF-CNN, which contains a motion compensation network and a quality enhancement network, is utilized to enhance the quality of non-PQFs. The same as QE-CNN, the MFQE model is re-trained for RGB images on the training set of LDV.

5. Challenge results and studies on LDV

5.1. Performance on Track 1

The quality performance vs. running time of Track 1 is illustrated in Figure 4. It can be seen that the methods proposed in NTIRE challenge obviously outperform the previous methods, and therefore advance the state-of-the-art of quality enhancement for compressed video.

PSNR vs. time efficiency. In the proposed methods, the BILIBILI AI & FDU and NTU-SLab Teams achieve the best quality with similar PSNR performance. The VUE Team ranks in the following at the third place. The NTU-SLab Team has the best time efficiency in the top methods (PSNR > 31.50 dB), and therefore it makes a good trade-off between quality performance and time efficiency. The most time-efficient method is Ivp-tencent, which is able to enhance video frames at 120 frames per second. The speed of Ivp-tencent significantly outperforms other methods, while its quality performance is comparable with and slightly higher than MFQE [38].

MS-SSIM performance. Figure 4-(b) shows the MS-SSIM performance on Track 1. We can observe from Figure 4-(b) that the MS-SSIM performance is generally consistent with the PSNR performance in Figure 4-(a). This may indicate that when optimizing the enhancement networks for PSNR, the MS-SSIM performance will be normally consistent with PSNR.

5.2. Performance on Track 2

The quality performance vs. running time of Track 2 is shown in Figure 5. Note that in the four perceptual metrics, the better MOS and VMAF values indicate better quality, while lower LPIPS and FID indicate better quality.

Normalized vs. non-normalized MOS. As mentioned in Section 3.1, we normalize the MOS values of each subject by (1) and rank the methods according to the aver-
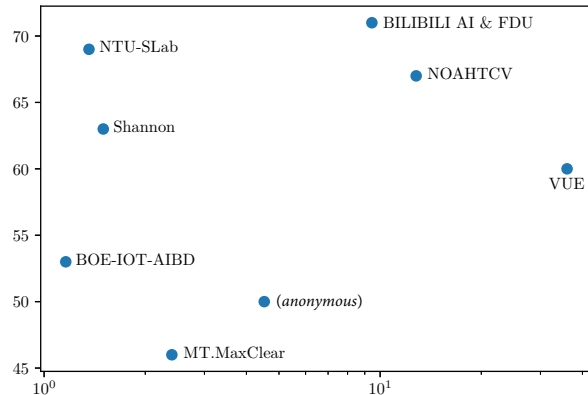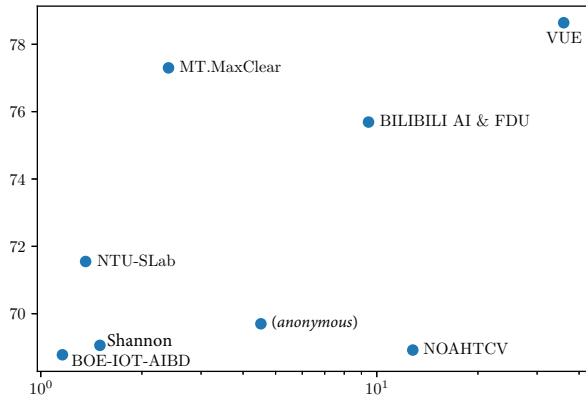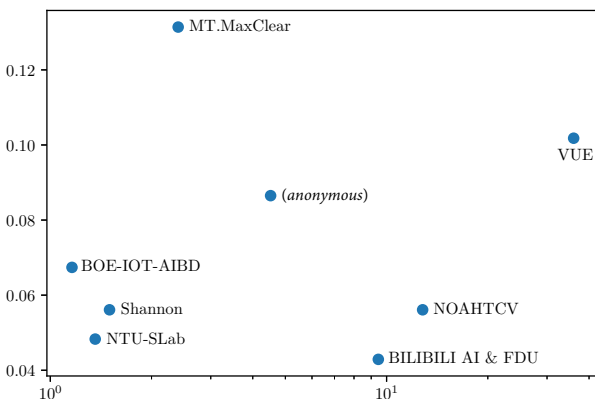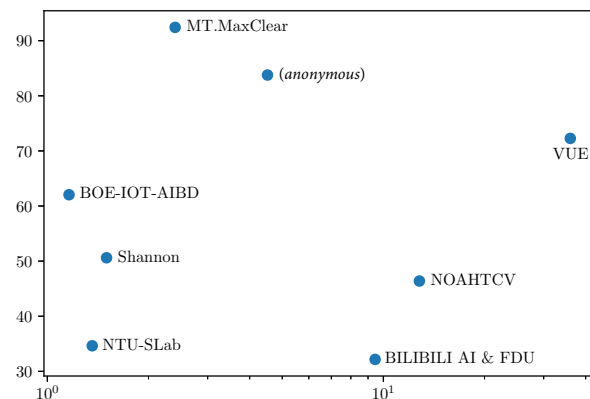
(a) MOS ↑ on Track 2

(b) VMAF ↑ on Track 2

(c) LPIPS ↓ on Track 2

(d) FID ↓ on Track 2

Figure 5. The results on Track 2. The running time is reported by the authors of the teams. The platforms used for test are listed in Table 1.

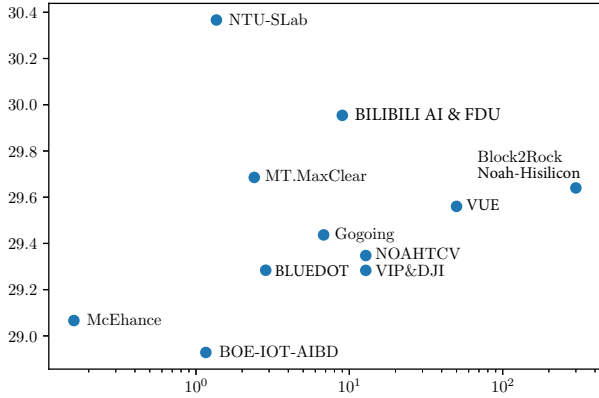Table 2. The normalized and non-normalized MOS values.

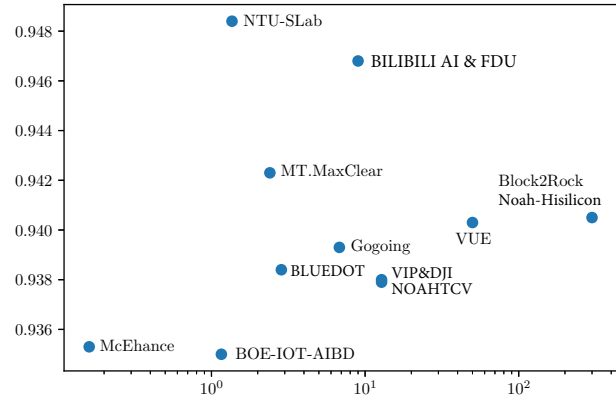| Team	MOS (norm)	MOS (w/o norm)
BILIBILI AI & FDU	70.6	69.6
NTU-SLab	68.5	68.0
NOAHTCV	67.0	65.1
Shannon	63.0	64.2
VUE	59.5	61.3
BOE-IOT-AIBD	52.7	55.7
(*anonymous*)	50.5	55.6
MT.MaxClear	46.0	51.3
Unprocessed video	36.3	43.7

in Figure 5.

MOS and time efficiency. As shown in Figure 5-(a), BILIBILI AI & FDU, NTU-SLab and NOAHTCV Teams achieve the top perceptual quality in terms of MOS. BILIBILI AI & FDU ranks at the first place in this track. The running time of NTU-SLab is several times lower than other top teams, *i.e.*, BILIBILI AI & FDU and NOAHTCV. Therefore, NTU-SLab achieves a good balance between perceptual quality and time efficiency. The BOE-IOT-AIBD Team achieves the best time efficiency. Moreover, it can be seen from Table 1 that all top methods use the ensemble strategies [22] during inference and enlarge the training set by adding extra videos for training. These may indicate the effectiveness of the ensemble strategy for perceptual enhancement and show the benefit of large-scale training data.

Other metrics. As Figures 5-(c) and -(d) show, the rankings of the top four methods in terms of LPIPS and FID are the same as MOS. This show that the perceptual metrics are effective for evaluating subjective quality. However, the LPIPS and FID values of unprocessed video are 0.0752 and 48.94, respectively. Therefore, some methods

age normalized MOS. We compare the results of normalized MOS and non-normalized MOS in Table 2. It can be seen from Table 2 that the normalization on MOS does not change the ranking of the proposed methods, but only enlarges the interval between the lowest and highest scores and therefore makes the difference between the methods more clear. Therefore, in the NTIRE 2021 challenge, we use the normalized MOS for Track 2. The results are shown

(a) PSNR on Track 3 (unprocessed video = 28.34 dB)

(b) MS-SSIM on Track 3 (unprocessed video = 0.9243)

Figure 6. The results on Track 3. The running time is reported by the authors of the teams. The platforms used for test are listed in Table 1.

even have worse LPIPS and FID performance than the unprocessed videos, but their MOS values are all obviously better the unprocessed videos (MOS = 36). Also, the ranking on VMAF is different from that on MOS. This may also indicate the limited reliability of the perceptual metrics.

5.3. Performance on Track 3

The PSNR and MS-SSIM performances on Track 3 are illustrated in Figure 6. The NTU-SLab, BILIBILI AI & FDU and MT.MaxClear Teams rank at the first, second and third places in terms of both PSNR and MS-SSIM. They improve the PSNR of unprocessed video (28.34 dB) by 2.03 dB, 1.62 dB and 1.35 dB, respectively. In these teams, NTU-SLab Team not only has the best quality performance, but achieves the best time efficiency as well. In all teams, the McEnhance Team proposes the most time-efficient method. It can be seen from Figure 6 that the ranks of PSNR and MS-SSIM are consistent. The same as Track 1, this may show that the MS-SSIM performance will be normally consistent with PSNR, in the case that the quality enhancement networks are optimized towards PSNR.

6. Analysis on training data

Finally, we analyze the scale of training data and the impact on the test performance. Figure 7 illustrates the PSNR results on the test set for different scales of training videos. The red line indicates the performance when the models are trained on all 200 training videos of the proposed LDV dataset. The blue and green lines show the results when training on 100 and 50 randomly selected videos from the training set of LDV, respectively. It can be seen from Figure 7 that enlarging the scale of training data from 50 to 100 obviously improves the test results, and further enlarging the number of training video to 200 also further slightly advances the performance.

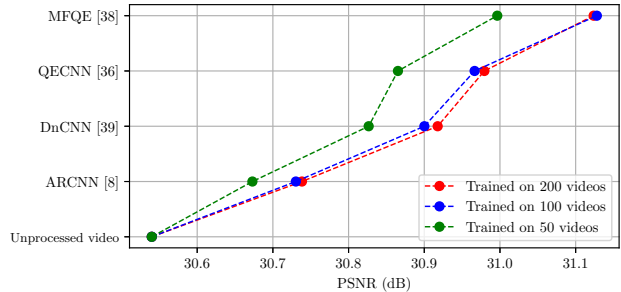As mentioned in Section 1, the training datasets used in

Figure 7. The PSNR performance on the test set of the models trained by different scales of data.

previous methods contain less than 150 videos. Therefore, the proposed LDV dataset may have the potential to promote the existing methods. However, in Figure 7, we also notice that the performance of MFQE [38] does not improve when increasing the number of training videos from 100 to 200. This may be because of the bias in the data, and it indicates that the scale of the proposed LDV dataset is still incremental, especially the test set. It is an interesting future work to further extend the LDV dataset and establish benchmarks on larger scales of data.

7. Conclusion

This paper has introduced the LDV database, which is a new large-scale dataset with diverse video. The proposed LDV dataset contains 240 videos with diverse categories of scenes, motions and frame-rates. The LDV dataset has been used in the video enhancement challenge of NTIRE 2021. We analyzed the proposed methods in the challenge and the methods in previous literature on the LDV dataset. These methods and results established a benchmark for quality enhancement of compressed video. We hope that the proposed LDV dataset and the benchmark of this challenge are able to provide solid baselines for the quality enhancement of

compressed video and further benefit the future research in this field.

Acknowledgments

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 1

[2] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 4

[3] Frank Bossen. Common test conditions and software reference configurations. *JCTVC-L1100*, 12, 2013. 2, 3

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[5] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing (ICIP)*, volume 2, pages 168–172. IEEE, 1994. 5

[6] Cisco. Cisco Annual Internet Report (2018–2023) White Paper. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html. 1

[7] Jianing Deng, Li Wang, Shiliang Pu, and Cheng Zhuo. Spatio-temporal deformable convolution for compressed video quality enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10696–10703, 2020. 1, 2, 5

[8] 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 (ICCV)*, pages 576–584, 2015. 5, 6

[9] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3512–3516. IEEE, 2019. 5

[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*, 43(3):949–963, 2021. 1, 2, 4, 5, 6

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 4

[12] Yongkai Huo, Qiyan Lian, Shaoshi Yang, and Jianmin Jiang. A recurrent video quality enhancement framework with multi-granularity frame-fusion and frame difference based attention. *Neurocomputing*, 431:34–46, 2021. 1, 2

[13] Bilibili Inc. Bilibili. https://www.bilibili.com/. 5

[14] Google Inc. YouTube. https://www.youtube.com. 5

[15] Netflix Inc. VMAF - video multi-method assessment fusion. https://github.com/Netflix/vmaf. 4

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1

[17] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Zhiyong Gao, and Ming-Ting Sun. Deep Kalman filtering network for video compression artifact reduction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 568–584, 2018. 1

[18] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 0–0, 2019. 5

[19] SkyPixel. https://www.skypixel.com. 5

[20] Li Song, Xun Tang, Wei Zhang, Xiaokang Yang, and Pingjian Xia. The sjtu 4k video sequence dataset. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 34–35. IEEE, 2013. 5

[21] 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*, 22(12):1649–1668, 2012. 1, 2

[22] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1865–1873, 2016. 7

[23] Ken Turkowski. Filters for common resampling tasks. *Graphics Gems I*, pages 147–165, 1990. 3

[24] 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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 1, 2

[25] Tingting Wang, Mingjin Chen, and Hongyang Chao. A novel deep learning-based method of improving coding efficiency from the decoder-end for HEVC. In *Proceedings of the Data Compression Conference (DCC)*, pages 410–419. IEEE, 2017. 1

[26] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 0–0, 2019. 5

[27] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multi-scale structural similarity for image quality assessment. In *Proceedings of The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003. 4

[28] 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*, 13(7):560–576, 2003. 1

[29] Mai Xu, Ren Yang, Tie Liu, Tianyi Li, and Zhaoji Fang. Multi-frame quality enhancement method and device for lossy compressed video, Dec. 24 2020. US Patent App. 16/968,409. 1

[30] Yi Xu, Longwen Gao, Kai Tian, Shuigeng Zhou, and Huyang Sun. Non-local ConvLSTM for video compression artifact reduction. In *Proceedings of The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 4

[31] Yi Xu, Minyi Zhao, Jing Liu, Xinjian Zhang, Longwen Gao, Shuigeng Zhou, and Huyang Sun. Boosting the performance of video compression artifact reduction with reference frame proposals and frequency domain information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 4

[32] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 1, 5

[33] Ren Yang, Fabian Mentzer, Luc Van Gool, and Radu Timofte. Learning for video compression with hierarchical quality and recurrent enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6628–6637, 2020. 1

[34] Ren Yang, Xiaoyan Sun, Mai Xu, and Wenjun Zeng. Quality-gated convolutional LSTM for enhancing compressed video. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 532–537. IEEE, 2019. 1, 2, 4

[35] Ren Yang, Radu Timofte, et al. NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 4, 5

[36] Ren Yang, Mai Xu, Tie Liu, Zulin Wang, and Zhenyu Guan. Enhancing quality for HEVC compressed videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018. 1, 4, 5, 6

[37] Ren Yang, Mai Xu, and Zulin Wang. Decoder-side HEVC quality enhancement with scalable convolutional neural network. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 817–822. IEEE, 2017. 1, 4, 6

[38] 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 (CVPR)*, pages 6664–6673, 2018. 1, 2, 4, 5, 6, 8

[39] 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*, 26(7):3142–3155, 2017. 5, 6

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 4, 5

[41] He Zheng, Xin Li, Fanglong Liu, Lielin Jiang, Qi Zhang, Fu Li, Qingqing Dang, and Dongliang He. Adaptive spatial-temporal fusion of multi-objective networks for compressed video perceptual enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 5