NTIRE 2020 Challenge on Video Quality Mapping: Methods and Results

Dario Fuoli Zhiwu Huang Martin Danelljan Radu Timofte Hua Wang
Longcun Jin Dewei Su Jing Liu Jaehoon Lee Michal Kudelski Lukasz Bala
Dmitry Hrybov Marcin Mozejko Muchen Li Siyao Li Bo Pang Cewu Lu
Chao Li Dongliang He Fu Li Shilei Wen

Abstract

This paper reviews the NTIRE 2020 challenge on video quality mapping (VQM), which addresses the issues of quality mapping from source video domain to target video domain. The challenge includes both a supervised track (track 1) and a weakly-supervised track (track 2) for two benchmark datasets. In particular, track 1 offers a new Internet video benchmark, requiring algorithms to learn the map from more compressed videos to less compressed videos in a supervised training manner. In track 2, algorithms are required to learn the quality mapping from one device to another when their quality varies substantially and weakly-aligned video pairs are available. For track 1, in total 7 teams competed in the final test phase, demonstrating novel and effective solutions to the problem. For track 2, some existing methods are evaluated, showing promising solutions to the weakly-supervised video quality mapping problem.

1. Introduction

Human captured and transmitted videos often suffer from various quality issues. For instance, despite the incredible development of current smartphone or depth cameras, compact sensors and lenses still make DSLR-quality unattainable for them. Due to bandwidth limit over internet, videos have to be compressed for easier transmission. The compressed videos inevitably suffer from compression artifacts. Therefore, quality enhancement over such videos are highly in demand.

The challenge aims at pushing competing methods into effective and efficient solutions to the newly emerging video quality mapping (VQM) tasks. Following [20], two tracks are studied in this challenge. Track 1 is configured to the task of fully-supervised video quality mapping between more compressed videos to less compressed videos collected from the Internet, while track 2 are designed for the weakly-supervised video quality mapping from a ZED camera to a Canon 5D Mark IV camera. Competing methods are evaluated with the most prominent metrics in the field, *i.e.*, Peak Signal-to-Noise Ratio (PSNR) and structural similarity index (SSIM).

Since PSNR and SSIM are not always well correlated with human perception of quality, we also consider to leverage perceptual measures, such as the Learned Perceptual Image Patch Similarity (LPIPS) [44] metric as well as mean opinion scores (MOS), which aim to evaluate the quality of the outputs according to human visual perception.

This challenge is one of the NTIRE 2020 associated challenges on: deblurring [26], nonhomogeneous dehazing [4], perceptual extreme super-resolution [43], video quality mapping (this paper), real image denoising [1], real-world super-resolution [24], spectral reconstruction from RGB image [5] and demoireing [42].

2. Related Work

Quality Enhancement on Compressed Videos aims to eliminate visual artifacts of compressed videos, which are transmitted over the bandwidth-limited Internet and often suffers from compression artifacts. There are emerging several algorithms like [9, 36, 40], which generally employ the original (uncompressed or less compressed) videos for full supervision on video quality map learning. For instance, [36] proposes an Auto-Decoder to learn the non-linear mapping from the decoded video to the original one, such that the artifacts can be removed and details can be enhanced on compressed videos. [9] suggests a post-processing algorithm for artifact reduction on compressed videos. Based on the observation that High in Efficiency Video Coding (HEVC) adopts variable block size transform, the suggested algorithm integrates variable filter size into convolutional networks for better reduction of the quantization error. To take advantage of the information available in the neighboring frames, [40] proposes a deep network to take both current frame and its adjacent high-quality frames into account

Dario Fuoli (dario.fuoli@vision.ee.ethz.ch), Zhiwu Huang, Martin Danelljan, and Radu Timofte at ETH Zürich are the NTIRE 2020 challenge organizers. The other authors participated in the challenge. Appendix A contains the authors' teams and affiliations.
https://data.vision.ee.ethz.ch/cvl/ntire20/

for better enhancement on compressed videos.

Video Super-Resolution (VSR) methods are used as well to enhance the texture quality of videos. The requirements for VSR and VQM are similar. It is important to enforce temporally consistent transitions between enhanced frames and to accumulate information over time, which is a fundamental difference to single image enhancement methods. Most deep learning based methods adopt the idea of concatenating adjacent frames with explicit motion compensation in order to leverage temporal information [19, 33, 7]. A more recent method [18] successfully explores the application of 3D convolutions as a natural extension for video data, without explicit motion compensation. In contrast to single image enhancers, many applications for video require real-time performance. Therefore, efficient algorithms for video processing are in high demand. Temporal information can be very efficiently aggregated with recurrent neural networks (RNN) which are developed in [31, 10]. For instance, [31] efficiently warps the previous high-resolution output towards the current frame according to optical flow. In [10], runtimes are further improved by propagating an additional hidden state, which handles implicit processing of temporal information without explicit motion compensation. Perceptual improvements over fully-supervised VSR methods are realized with generative adversarial networks (GAN) by [25] and [28].

Quality Enhancement on Device Captured Videos aims at enhancing the perceived quality of videos taken by devices, which includes enhancements like increasing color vividness, boosting contrast, sharpening up textures, etc. However, the major issue of enhancing such videos is the extreme challenge of collecting well-aligned training data, i.e., input and target videos that are aligned in both the spatial and the temporal domain. A few approaches address this problem using reinforcement learning based techniques like [13, 27, 21], which aims at creating pseudo input-retouched pairs by applying retouching operations sequentially.

Another direction is to develop Generative Adversarial Network (GAN) based methods for this task. For example, [8] proposes a method for image enhancement by learning from unpaired photographs. The method learns an enhancing map from a set of low-quality photos to a set of high-quality photographs using the GAN technique [11], which has proven to be good at learning real data distributions. Similarly, [17] leverages the GAN technique to learn the distribution of separate visual elements (i.e., color and texture) of images, such that the low-quality images can be mapped easier to the high-quality image domain which is encoded with more vivid colors and more sharpened textures. More recently, [15] suggests a divide-and-conquer adversarial learning method to further decompose the photo enhancement problem into multiple sub-problems. Such sub-problems are divided hierarchically: 1) a perception-based division for learning on additive and multiplicative components, 2) a frequency-based division in the GAN context for learning on the low- and high-frequency based distributions, and 3) a dimension-based division for factorization of high-dimensional distributions. To further smooth the temporal semantics during the enhancement, an efficient recurrent design of the GAN model is introduced. To the best of our knowledge, except for [15], there are very few works specially for weakly-supervised video enhancement.

3. Challenge Setup

3.1. Track 1: Supervised VQM

For this track, we introduce the IntVid dataset [20]. It consists of videos downloaded from Internet websites. The collected videos cover 12 diverse scenarios: city, coffee, fashion, food, lifestyle, music/dance, narrative, nature, sports, talk, technique and transport. The resolution of the crawled videos is mostly 1920×1080. Their duration varies from 8 seconds to 360 seconds with frame rates in the range of 23.98-25.00 FPS.

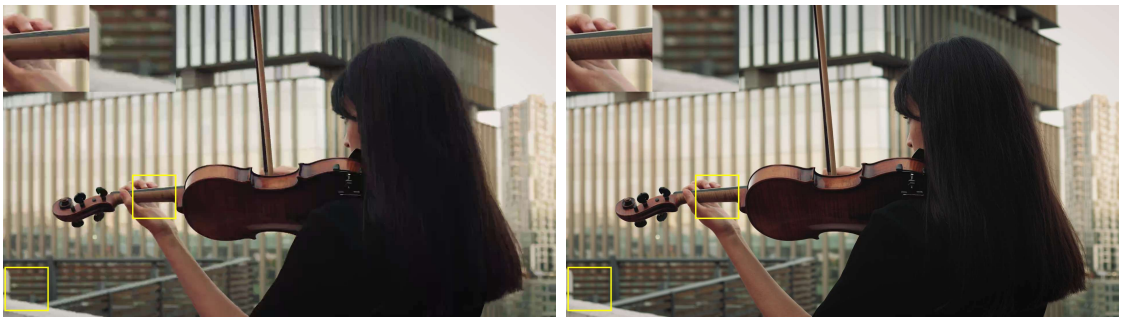
As most of the collected videos consist of changing scenes, a popular scene detection tool named PySceneDetect[1] is used to split the videos into three separate sets of clips for training, validation and test respectively. In particular, most of the resulting video clips are selected such that the majority of the original video content is employed for training. For the validation and test video clips, the video length is fixed to 4 seconds containing 120 frames, which are saved as PNG image files.

Due to the bandwidth limit of Internet, video compression techniques are often applied to reduce the coding bitrate. Inspired by this, [20] applied the standard video coding system H.264 to compress the collected videos. As a result, a total of 60 paired compressed and uncompressed videos are generated for training, and 32 paired compressed/uncompressed clips are produced for validation and testing. One example for track 1 is shown in Fig.1 (a)-(b).

3.2. Track 2: Weakly-Supervised VQM

For this track, we employ the Vid3oC dataset [20], which records videos with a rig containing three cameras. In particular, we use the Canon 5D Mark IV DSLR camera to serve as a high-quality reference, while utilizing the ZED camera, which additionally records depth information, to provide sequences of the same scene with a significantly lower video quality level. As the track focuses on the RGB-based visual quality mapping, we remove the depth information from the ZED camera. Using the two cameras, videos are recorded in the area in and around

[1] https://pyscenedetect.readthedocs.io



(a) More compressed (b) Less compressed

(c) ZED captured (d) Canon captured

Figure 1: Track 1 (a)-(b): quality mapping from more compressed (a) to less compressed (b) videos which are well aligned. Track 2 (c)-(d): quality mapping from low-quality videos captured by a ZED camera (c) to high-quality Canon DSLR videos (d), which are roughly aligned.

Zurich, Switzerland during the summer months. The locations and scenes are carefully chosen to ensure variety in content, appearance, and the dynamic nature. The length of each recording is between 30 and 60 seconds. Videos are captured in 30 FPS, using the highest resolution (i.e., 1920×1080) available at that frame rate.

In [20], the recorded videos are split into a training set of 50 weakly-paired videos, together with a validation and test set of 16 videos each. For all sets, a rough temporal alignment is performed based on the visual recording of a digital clock, which is captured by both cameras in the beginning of each video. The training videos are then trimmed down to 25-50 seconds by removing the first few seconds (which include the timer) and encoded with H.264. For each video in the validation and test set, a 4-second interval is selected. Each of such small video clips contains 120 frames, which are stored as individual PNG image files. Fig.1 (c)-(d) illustrates one example for track 2.

3.3. Evaluation Protocol

Validation phase: During the validation phase, the source domain videos for the validation set were provided on CodaLab. While the participants had no direct access to the validation ground truth, they could get feedback through the online server on CodaLab. Due to the storage limits on the servers, the participants could only submit a subset of frames for the online evaluation. PSNR and SSIM were reported for both tracks, even though track 2 only has weakly-aligned targets. The participants were allowed to make 10 submissions per day, and 20 submissions in total for the whole validation phase.

Test phase: In the test phase, participants are expected to submit their final results to the CodaLab test server. Compared to the validation phase, no feedback was given in terms of PSNR/SSIM to prevent comparisons with other teams and overfitting to the test data. By the deadline, the participants were required to provide the full set of frames, from which the final results were obtained.

4. Challenge Teams and Methods

In total 7 teams submitted their solutions to track 1. One team asked to anonymize their team name and references, since they found out to be using inappropriate extra-data for training after the test phase submission deadline. No submissions were made for track 2.

4.1. GTQ team

The team proposes a modified deformable convolution network to achieve high quality video mapping as shown in Fig. 2. The framework first down-samples the input frames

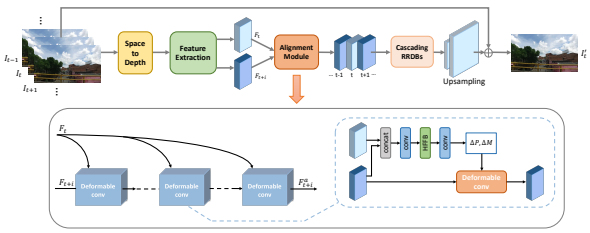


Figure 2: Illustration of the network design suggested by team GTQ.

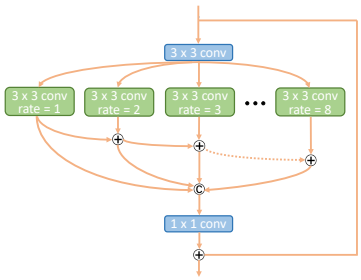


Figure 3: Illustration of the hierarchical feature fusion block (HFFB) suggested by team GTQ.

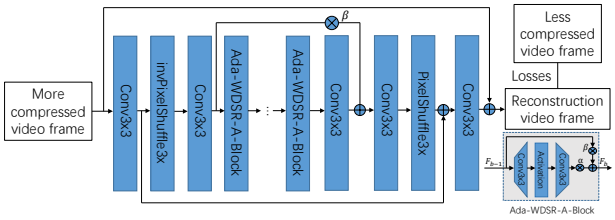


Figure 4: The network architecture of the proposed C2CNet by team ECNU.

with scale factor 4 through a space to depth shuffling operation. Then, the extracted features pass through an alignment module which applies a cascade of deformable convolutions [47] to perform implicit motion compensation. In the alignment module, the team takes advantage of hierarchical feature fusion blocks (HFFB) [16] to predict more precise offset and modulation scalars used in deformable convolutions. As shown in Fig. 3, HFFB introduces a spatial pyramid of dilated convolutions to effectively enlarge the receptive field with relatively low computational cost, which contributes to dealing with complicated and large motions between frames. After the alignment operation, the features are concatenated and fed into stacked residual in residual dense blocks (RRDB) [39] to reconstruct high quality frames.

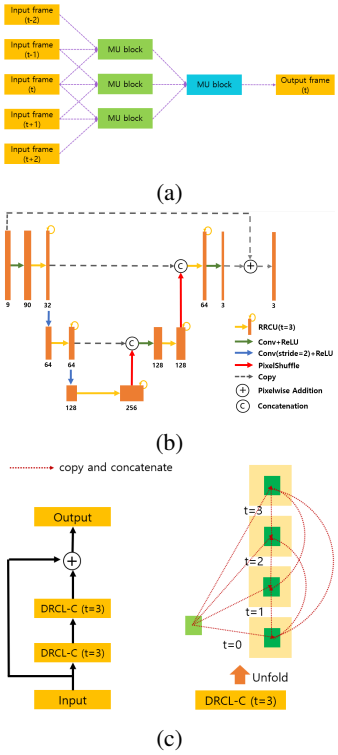


Figure 5: Architecture of team GIL's model. (a) Overall Network architecture. (b) MU block. (c) RRCU (t=3) at the left and Unfolded DRCL-C (t=3) at the right.

4.2. ECNU team

Team ECNU proposes a Compression to Compression Network (C2CNet). The input to C2CNet is a more compressed video frame and the ground truth is a less compressed video frame. As shown in Fig. 4, C2CNet is composed of a head 3×3 convolutional layer, a de-sub-pixel convolutional layer composed of an inverse pixel-shuffle layer and a 3×3 convolution, a non-linear feature mapping module composed of 64 Adaptive WDSR-A-Blocks, a 3×3 convolution and a short skip connection with residual scaling β=0.2, an upsampling skip connection, a sub-pixel convolutional layer composed of a 3×3 convolution and a pixel-shuffle layer, a global skip connection and a tail 3×3 convolution. The number of channels for C2CNet is 128. The Adaptive WDSR-A-Block is composed of 64, 256 and 64 channels. The Adaptive WDSR-A-Block is modified from a WDSR-A-Block [41], by adding learnable weight α (initialized with 1) for body scaling and learnable weight β (initialized with 0.2) for residual scaling. Each 3×3 convolution is followed by a weight normalization layer (omitted in Fig. 4).

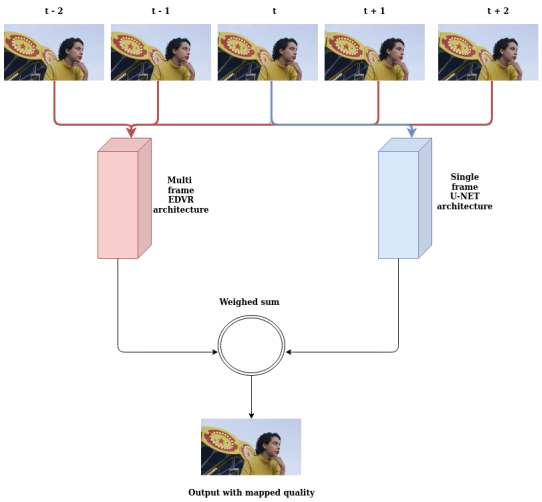


Figure 6: Schematic representation of team TCL's approach.

4.3. GIL team

The team employs a network with two-stage architecture proposed in FastDVDnet [34] and is shown in Fig. 5a. It takes five consecutive frames as an input and generates a restored central frame. Three MU blocks in the first stage (shown in green) share parameters. Each MU block is a modified U-Net [29] shown in Fig. 5b. It uses a convolutional layer with stride=2 for down-sampling and a pixel-shuffle [32] layer for up-sampling. It features a skip connection for global residual learning and contains several RRCUs (recurrent residual convolutional unit) inspired from R2U-Net [3]. Each RRCU consists of two DRCL-C (dense recurrent convolutional layer-concatenate) and a skip connection for residual learning. Figures of RRCU and DRCL-C are shown in Fig. 5c. States of the DRCL-C change over discrete time steps and the maximum time step is limited to 3. The DRCL-C is different from a standard RCL (recurrent convolutional layer) [22]. It reuses previous features by concatenating them [14]. A convolutional layer with 1x1 filters is used after every concatenation in DRCL-C to make the number of channels constant. The network has approximately 3.6 million parameters.

4.4. TCL team

The team uses a pyramidal architecture with deformable convolutions and spatio-temporal attention based on the work of [37] along with a single-frame U-Net [29]. The overview of the method is illustrated in Fig. 6. By combining these two methods, the local frame structure is preserved with the usage of U-Net and additional information from neighboring frames along with motion compensation, mostly by exploiting the PCD module from [37], is used to enhance output quality. Both networks are trained separately and the final result is obtained by a weighted sum with weight parameter β found by grid search, which is validated on a hold-out set from the training frames.

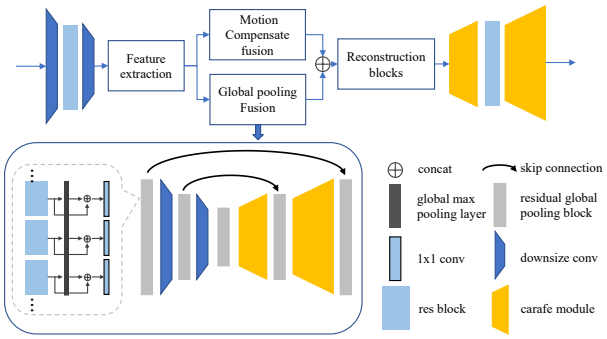


Figure 7: Network architectures used by team JOJO-MVIG.

4.5. JOJO-MVIG team

The team proposes a unified dual-path model to jointly utilize spatial and temporal information and map low-quality compressed frames to high-quality ones. As shown in Fig. 7, the model consists of a feature extraction stage, two spatio-temporal fusion paths, and a reconstruction module. The overall design of the pipeline follows [37].

In the feature extraction part, the multi-level features are calculated. The fusion stage explores spatial and temporal correlation across input frames and fuses useful information. Two fusion paths are designed for motion compensation and global pooling. The motion compensation fusion part measures and compensates the motion across frames by aligning them to the reference frame. The fusion is performed on aligned frames/features. The team adopts the alignment and fusion part from EDVR [37] for the motion compensation part.

Compared to the motion compensation path, the global pooling fusion path requires no alignment and adopts a U-net [30] like architecture in which global max-pooling layers are inserted into all residual blocks. Global pooling has been used in [2] to conduct permutation invariant deblurring. Here global pooling is used to exchange information between different frames, and since max-pooling is a selective process, different frames vote for the best information for restoration. Furthermore, the team adopts the CARAFE Module [35] to enable pixel-specific content-aware up-sampling. More specifically, the team uses 7 frames as input, with reconstruction blocks consisting of 40 residual blocks and feature extraction module consisting of 5 residual blocks. The channel number for each residual block is set to 128.

4.6. BossGao team

The BossGao team exploits cutting-edge deep neural architectures for the video quality mapping task. Specifically,

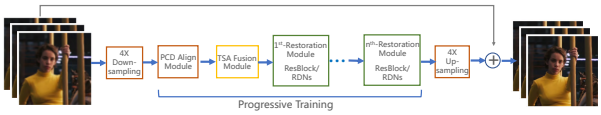


Figure 8: Illustration of the proposed framework by Boss-Gao. The team exploits cutting-edge deep neural architectures for the video quality mapping task, i.e PCD align module, TSA fusion module, residual blocks and RDN blocks. For progressive training, first, the PCD align module and the 1st Restoration module are trained together. Next, the TSA fusion module is plugged in and the existing parameters are used as initialization. Then, the new framework with TSA module is trained again. More restoration modules can be stacked to get a deeper framework, which can be trained to achieve better performance.

the team develops the following frameworks:

- Framework1: PCD+TSA+10ResBlocks+30ResBlocks
- Framework2: PCD+RDN1
- Framework3: PCD+TSA+RDN1
- Framework4: PCD+TSA+RDN2

where 10 ResBlock means 10 residual blocks [23], and there are two convolution layers in each ResBlock. RDN1 denotes 10 RDBs [45] with 8 convolution layers in each RDN. RDN2 denotes 8 RDBs with 6 convolution layers in each RDN. PCD and TSA are proposed in [38]. The framework is illustrated in Fig. 8.

Another contribution of the team is that, they propose to train the modules in these frameworks progressively. They train a framework by starting with fewer modules. More modules are added in progressively. When new modules are plugged in, the existing parameters are used as initialization, and the new modules and old modules are trained together. The modules in their frameworks are added in a carefully arranged order. Specifically, a framework with a PCD module and shallower restoration modules is trained first. Then, a TSA module is plugged in. Furthermore, more restoration modules can be stacked on to get a deeper frameworks. Frameworks trained by their method achieve better performance than the corresponding networks that are trained once-off.

In the final phase, the frameworks with the best performance are selected to produce the final test videos, i.e. Framework1, Framework3 and Framework4. Framework2 is only used for the last submission in the development phase.

4.7. DPE (baseline for track 2)

DPE [8] is originally developed for weakly-supervised photo enhancement. For track 2, we apply it to enhance videos frame by frame. In particular, DPE treats the problem with a two-way GAN whose structure is similar to CycleGAN [46]. To address the unstable training issue of GANs and obtain high-quality results, DPE proposes a few improvements along the way of constructing the two-way GAN. First, it suggests to augment the U-Net [29] with global features for the design of the generator. In addition, individual batch normalization layers are proposed for the same type of generators. For better GAN training, DPE proposes an adaptive weighting Wasserstein GAN scheme.

4.8. WESPE (baseline for track 2)

Similar to DPE [8], WESPE [17] is another baseline that exploits the GAN technique for weakly supervised per-frame enhancement. The WESPE model comprises a generator G paired with an inverse generator G_r. In addition, two adversarial discriminators D_c and D_t and total variation (TV) complete the model's objective definition. D_c aims at distinguishing between high-quality image y and enhanced image $\tilde{y} = G(x)$ based on image colors, and D_t distinguishes between y and $\tilde{y}$ based on image texture. More specially, the objective of WESPE consists of: i) content consistency loss to ensure G preserves x's content, ii) two adversarial losses ensuring generated images $\tilde{y}$ lie in the target domain Y: a color loss and a texture loss, and iii) TV loss to regularize towards smoother results.

4.9. DACAL (baseline for track 2)

For track 2, we suggest the DACAL method [15] as the last baseline, which enhances videos directly. To further reduce the problem complexity, DACAL decomposes the photo enhancement process into multiple sub-problems. On the top level, a perception-based division is suggested to learn additive and multiplicative components, required to translate a low-quality image or video into its high-quality counterpart. On the intermediate level, a frequency-based division is exploited in the GAN context to learn the low- and high-frequency based distribution separately in a weakly-supervised manner. On the bottom level, a dimension-based division is suggested to factorize high-dimensional distributions into multiple one-dimensional marginal distributions for better training on the GAN model. To better deal with the temporal consistency of the enhancement, DACAL introduces an efficient recurrent design of the GAN model. The

	Method	↑PSNR	↑SSIM	↓LPIPS	TrainingReq	TrainingTime	TestReq	TestTime	Parameters	ExtraData
Participants	BossGao	**32.419**	**0.905**	0.177	8×V100	5-10d	1×V100	4s	n/a	No
	JOJO-MVIG	32.167	0.901	0.182	2×1080Ti	≈ 4d	1×1080Ti	2.07s	≈22.75M	No
	GTQ	32.126	0.900	0.187	2×2080Ti	≈ 5d	1×2080Ti	9.74s	19.76M	No
	ECNU	31.719	0.896	0.198	2×1080Ti	2-3d	1×1080Ti	1.1s	n/a	No
	TCL	31.701	0.897	0.193	2×1080Ti	≈ 3d	1×1080Ti	25s	≈8.92M	No
	GIL	31.579	0.894	0.195	1×970Ti	≈ 6d	1×970Ti	11.37s	3.60M	No
	7-th team	30.598	0.878	**0.176**	n/a	4d	n/a	0.5s	≈7.92M	Yes
	No processing	30.553	0.877	**0.176**						

Table 1: Quantitative results for Track 1. **Bold**: best, Underline: second and third best. TrainingTime: days, TestTime: seconds per frame.

5. Challenge Result Analysis

5.1. Track 1: Supervised VQM

This challenge track aims at restoring the discarded information, which has been lost due to compression, with the highest fidelity to the ground truth. Because of full supervision, the ranking among the participating teams can be computed objectively.

Metrics The most popular full reference metrics to evaluate the quality of images and videos are PSNR and SSIM. PSNR can be computed directly from the mean-squared-error (MSE). Therefore, L2-norm based objectives are commonly used to obtain high PSNR scores. SSIM is calculated from windows based statistics in images. In this challenge, both metrics are calculated per frame and averaged over all sequences. Table 1 reports the quantitative results of participating methods as well as the baseline, i.e. the input without any processing. With a PSNR value of 32.42dB and SSIM score of 0.91, team BossGao achieves the highest scores overall and is the winner of challenge track 1. Team JOJO-MVIG and GTQ follow closely with a PSNR difference of 0.25dB and 0.29dB to the winner respectively. The remaining teams also achieve respectable PSNR scores slightly below 32dB. The ranking in terms of SSIM is almost the same. In addition, as can be seen by the reported training times, capacity and test times, models with more parameters and teams with more processing power generally perform better. However, team ECNU manages to surpass more expensive methods with the fastest runtime. Team GIL targets for a compact network with the least parameters, which can be trained on a single lower-end GPU but still produces promising enhancement results.

Visual Comparison Selected samples from the test data are provided in Fig. 9 to compare the visual quality of the enhanced video frames among all teams. The visual comparison shows that team BossGao also performs the best for the quality enhancement on such sampled frames. It should be noted that due to the inherent loss of information after compression, fidelity based methods are not able reconstruct all high frequency details and tend to over-smooth the content. In order to assess continuity between frames, temporal profiles for all teams are provided in Fig. 10. A single vertical line of pixels is recorded over all frames in the sequence and stacked horizontally.

Additionally, we computed LPIPS [44] scores to compare perceptual quality among the teams. Optimizing for perceptual quality was not required by the participants in this challenge track, but the metric still provides interesting insights into quantitative quality assessment and its limitations. The scores among all teams is roughly consistent with PSNR and SSIM, which implies that the top teams also produce visually more pleasing results compared to their competitors. Interestingly, the input without processing along with team 7, which basically doesn't alter the input, achieves the best score. We assume that the distortions, due to smoothing of L2-norm based methods, cause worse scores for the top teams, despite much higher reconstruction quality. In contrast, compression algorithms are designed to optimize for perceptual quality, which could lead to the strong LPIPS score for the input.

5.2. Track 2: Weakly-Supervised VQM

In this challenge track, the goal of the task is to enhance the video characteristics from a low quality device (ZED camera) to the characteristics of a high-end device (Canon 5D Mark IV) with limited supervision. Weak supervision is provided by weakly-paired videos, which share approximately the same content and are roughly aligned in the spatial and temporal domain.

Metrics Since there is no pixel-aligned ground truth available, full reference metrics are no option for quality assessment. Usually, results for these types of problems are scored by a MOS study, conducted by humans visually comparing different methods. While there exist metrics to measure distances between probability distributions for high level content, e.g. Frchet Inception Distance (FID) [12], that are widely applied to generative models, finding reliable metrics for low-level characteristics remains an open problem. Popular perceptual metrics such as

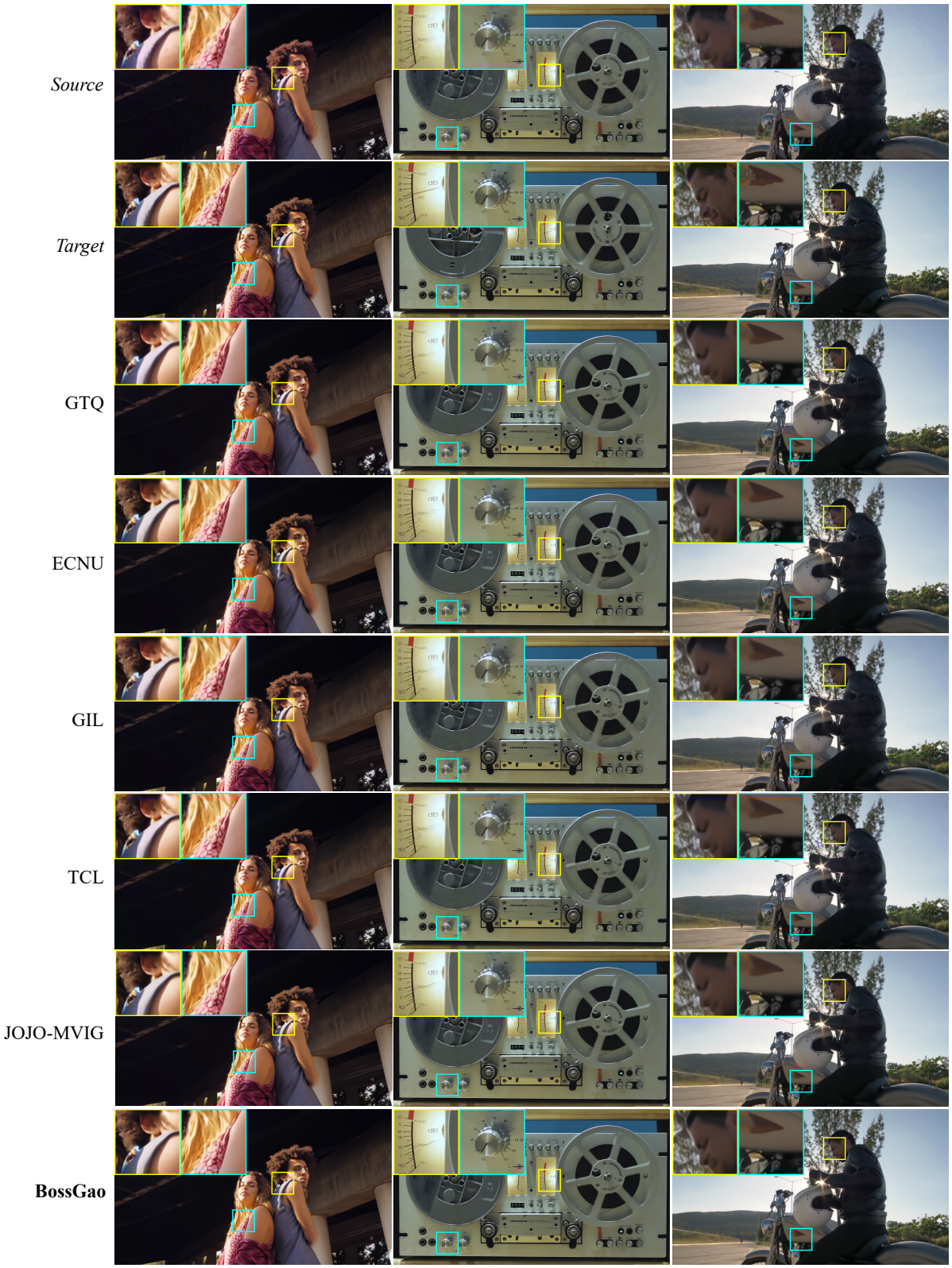


Figure 9: Visual Comparison for Track 1.

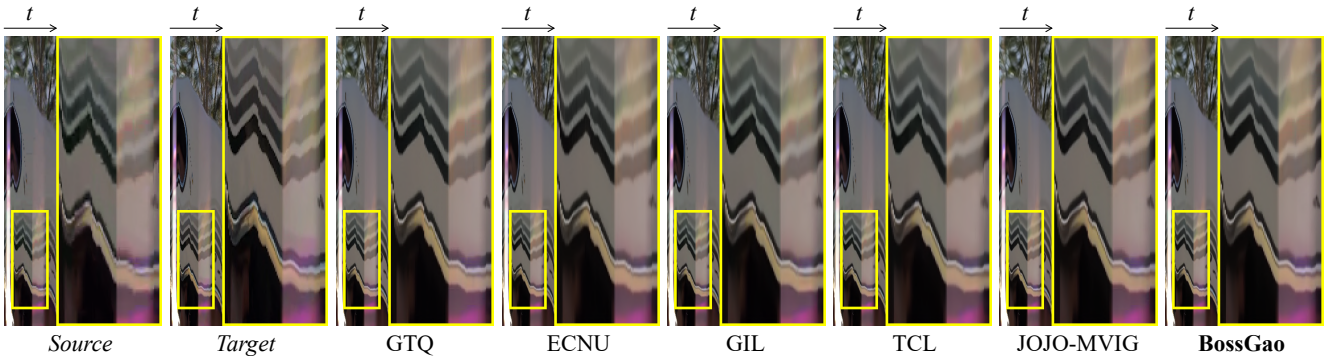


Figure 10: Temporal Profiles for Track 1.

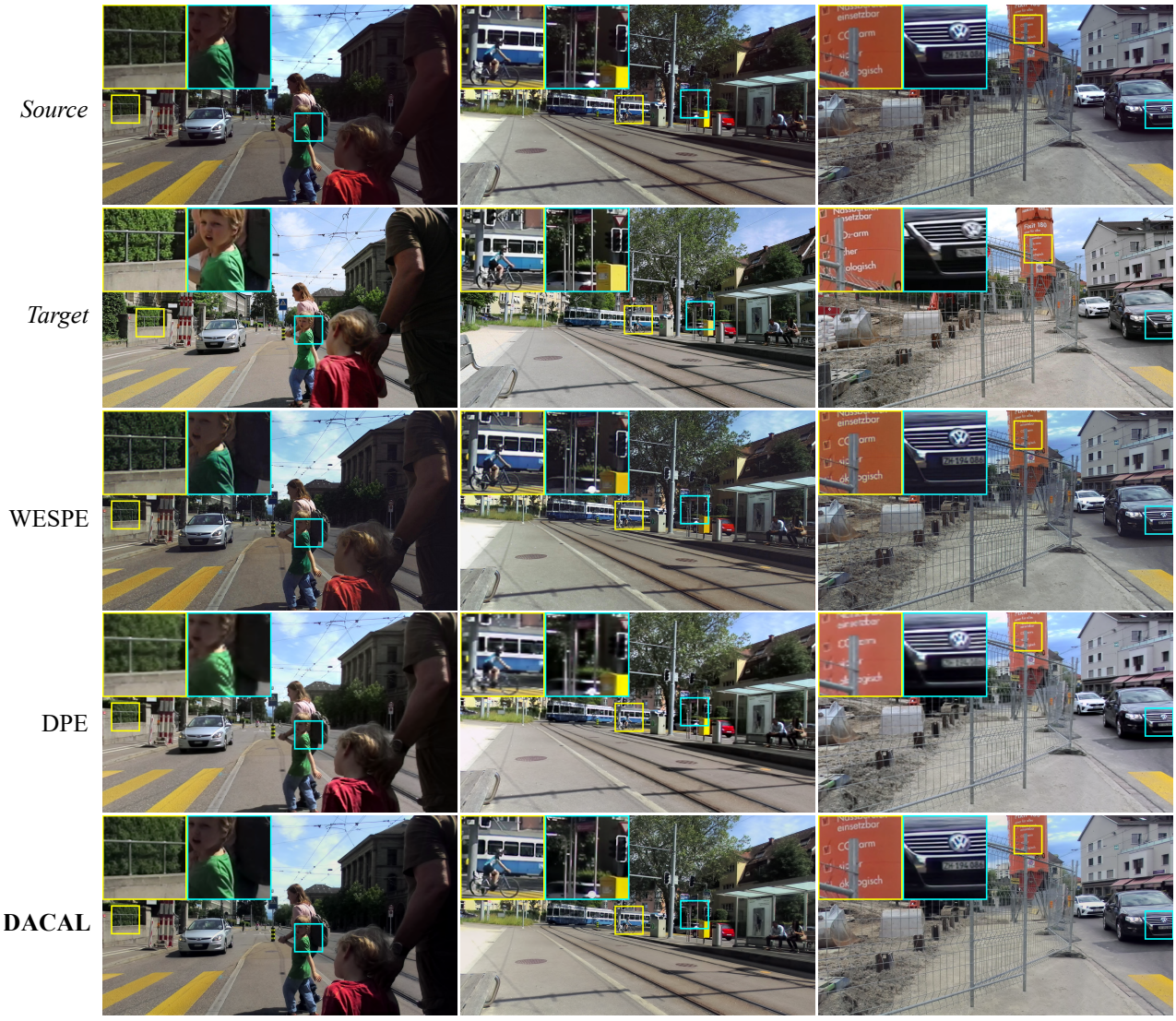


Figure 11: Visual Comparison for Track 2.

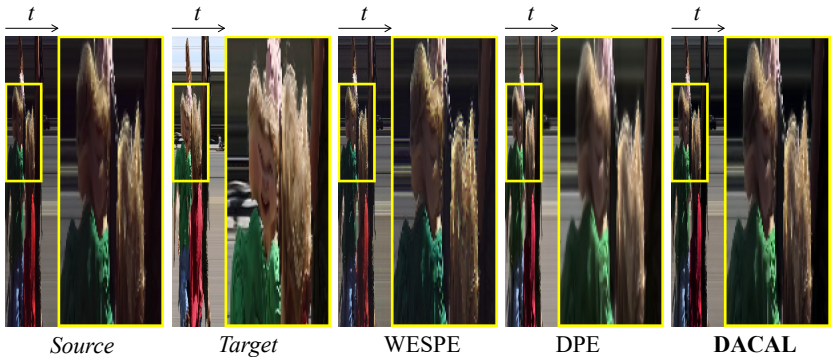


Figure 12: Temporal Profiles for Track 2.

	Source	WESPE	DPE	DACAL
LPIPS↓	**0.590**	0.755	0.793	<u>0.750</u>

Table 2: LPIPS scores for Track 2. **Bold**: best, <u>Underline</u>: second best.

Learned Perceptual Image Patch Similarity [44] metric and Perceptual Index [6] are used in the field too. However, we found the scores for these metrics are not suitable for the problem setting in this challenge and do not always correlate with human perception. Perceptual Index is not a relative score, it only measures general quality. However, we are interested in measuring the mapping quality from one domain to another. LPIPS requires aligned frames which is a problem since the frames are only roughly aligned. Nevertheless, we provide LPIPS scores for a selection of methods along with visual results, see Table 2, Fig. 11 and Fig. 12. Surprisingly, the source without processing achieves the best score by a large margin. While source and target frames are captured by a real camera, the methods alter the videos artificially. Since LPIPS relies on a feature extractor, which is trained on real images, this could lead to worse scores for the methods, due to low level distortions.

Visual Comparison Since there are no submissions for this track, visual results and temporal profiles for a selection of recent image and video quality mapping methods is provided as reference in Fig. 11 and Fig. 12 . WESPE [17] and DPE [8] are single image methods which are applied per frame, DACAL [15] is a true video enhancer. All the competing methods are trained on the Vid3oC dataset [20]. The visual results show that DACAL preserves more details and enhances contrast better, while WESPE introduces biased colorization and DPE produces blurry textures.

6. Conclusions

This paper presents the setup and results of the NTIRE 2020 challenge on video quality mapping. This challenge addresses two real world settings: track 1 concerns video quality mapping from more compressed videos to less compressed ones with available paired training data; track 2 focuses on video quality mapping from a lower-end device to a higher-end device, given a collected weakly-paired training set. 7 teams competed in Track 1 in total. The participating methods demonstrated interesting and innovative solutions to the supervised quality mapping on compressed videos. In contrast, we evaluated three existing methods for track 2, showing their performance is promising but much effort is still needed for better video enhancement. The evaluation with LPIPS on both challenge tracks reveals the limits of current quantitative perceptual quality metrics and shows the need for more research in that area, especially for track 2 where no pixel-aligned reference is available. Our goal is that this challenge stimulates future research in the area of video quality mapping in either supervised or weakly-supervised scenarios, by serving as a standard benchmark and by the evaluation of new baseline methods.

Acknowledgements

We thank the NTIRE 2020 sponsors: Huawei, Oppo, Voyage81, MediaTek, DisneyResearch|Studios, and Computer Vision Lab (CVL) ETH Zurich.

Appendix A: Teams and affiliations

NTIRE 2020 VQM organizers

Members: Dario Fuoli (dario.fuoli@vision.ee.ethz.ch), Zhiwu Huang (zhiwu.huang@vision.ee.ethz.ch), Martin Danelljan (martin.danelljan@vision.ee.ethz.ch), Radu Timofte (radu.timofte@vision.ee.ethz.ch)

Affiliations: Computer Vision Lab, ETH Zurich, Switzerland

GTQ team

Title: Modified Deformable Convolution Network for Video Quality Mapping

Members: Hua Wang, Longcun Jin (lcjin@scut.edu.cn), Dewei Su

Affiliations: School of Software Engineering, South China University of Technology, Guangdong, China

ECNU team

Title: Compression2Compression: Learning compression artifacts reduction without clean data

Members: Jing Liu (splinter02@163.com)

Affiliations: Multimedia and Computer Vision Lab, East China Normal University, Shanghai, China

GIL team

Title: Dense Recurrent Residual U-Net for Video Quality Mapping

Members: Jaehoon Lee (dlwogns0729@gmail.com)

Affiliations: Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea

TCL team

Title: Deformable convolution based multi-frame network with single-frame U-Net

Members: Michal Kudelski (michal.kudelski@tcl.com), Lukasz Bala, Dmitry Hrybov, Marcin Mozejko

Affiliations: TCL Research Europe, Warsaw, Poland

JOJO-MVIG team

Title: Video Quality Mapping with Dual Path Fusion Network

Members: : Muchen Li (muchenli1997@gmail.com), Siyao Li, Bo Pang, Cewu Lu

Affiliations: Machine Vision and Intelligence Group, Department of Computer Science, Shanghai Jiao Tong University, Shanghai, China

BossGao team

Title: Exploiting Deep Neural Architectures by Progressive Training for Video Quality Mapping

Members: Chao Li (uqcli1@gmail.com), Dongliang He, Fu Li, Shilei Wen

Affiliations: Department of Computer Vision Technology (VIS), Baidu Inc., Beijing, China

References

[1] Abdelrahman Abdelhamed, Mahmoud Afifi, Radu Timofte, Michael Brown, et al. Ntire 2020 challenge on real image denoising: Dataset, methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[2] Miika Aittala and Fredo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 5

[3] Md Zahangir Alom, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. Nuclei segmentation with recurrent residual convolutional neural networks based u-net (r2u-net). In *NAECON 2018-IEEE National Aerospace and Electronics Conference*, pages 228–233. IEEE, 2018. 5

[4] Codruta O. Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. Ntire 2020 challenge on nonhomogeneous dehazing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[5] Boaz Arad, Radu Timofte, Yi-Tun Lin, Graham Finlayson, Ohad Ben-Shahar, et al. Ntire 2020 challenge on spectral reconstruction from an rgb image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[6] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 7

[7] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[8] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6306–6314, 2018. 2, 6, 10

[9] 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*, pages 28–39. Springer, 2017. 1

[10] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *ICCV Workshops*, 2019. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[12] 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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6626–6637. 2017. 7

[13] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2):1–17, 2018. 2

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. 5

[15] Zhiwu Huang, Danda Pani Paudel, Guanju Li, Jiqing Wu, Radu Timofte, and Luc Van Gool. Divide-and-conquer adversarial learning for high-resolution image and video enhancement. *arXiv preprint arXiv:1910.10455*, 2019. 2, 6, 10

[16] Zheng Hui, Jie Li, Xinbo Gao, and Xiumei Wang. Progressive perception-oriented network for single image super-resolution. *arXiv preprint arXiv:1907.10399*, 2019. 4

[17] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Wespe: weakly supervised photo enhancer for digital cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 691–700, 2018. 2, 6, 10

[18] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[19] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. In *IEEE Transactions on Computational Imaging*, 2016. 2

[20] Sohyeong Kim, Guanju Li, Dario Fuoli, Martin Danelljan, Zhiwu Huang, Shuhang Gu, and Radu Timofte. The vid3oc and intvid datasets for video super resolution and quality mapping. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3609–3616. IEEE, 2019. 1, 2, 3, 10

[21] Satoshi Kosugi and Toshihiko Yamasaki. Unpaired image enhancement featuring reinforcement-learning-controlled image editing software. *arXiv preprint arXiv:1912.07833*, 2019. 2

[22] Ming Liang and Xiaolin Hu. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3367–3375, 2015. 5

[23] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 6

[24] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[25] Jonas Mayer Laura Leal-Taix Nils Thuerey Mengyu Chu, You Xie. Learning temporal coherence via self-supervision for gan-based video generation, 2018. 2

[26] Seungjun Nah, Sanghyun Son, Radu Timofte, Kyoung Mu Lee, et al. Ntire 2020 challenge on image and video deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[27] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5928–5936, 2018. 2

[28] E. Pérez-Pellitero, M. S. M. Sajjadi, M. Hirsch, and B. Schölkopf. Photorealistic video super resolution, 2018. 2

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5, 6

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 5

[31] Mehdi S. M. Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-Recurrent Video Super-Resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[32] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 5

[33] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[34] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time video denoising without explicit motion estimation. *arXiv preprint arXiv:1907.01361*, 2019. 4

[35] Jiaqi Wang, Kai Chen, Rui Xu, Ziwei Liu, Chen Change Loy, and Dahua Lin. CARAFE: content-aware reassembly of features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3007–3016. IEEE, 2019. 5

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

[37] 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*, pages 0–0, 2019. 5

[38] Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: video restoration with enhanced deformable convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. 6

[39] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 4

[40] 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*, pages 6664–6673, 2018. 1

[41] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas Huang. Wide activation for efficient and accurate image super-resolution. *arXiv preprint arXiv:1808.08718*, 2018. 4

[42] Shanxin Yuan, Radu Timofte, Ales Leonardis, Gregory Slabaugh, et al. Ntire 2020 challenge on image demoireing: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[43] Kai Zhang, Shuhang Gu, Radu Timofte, et al. Ntire 2020 challenge on perceptual extreme super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1, 7

[45] 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*, pages 2472–2481, 2018. 6

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 6

[47] 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*, pages 9308–9316, 2019. 4