NTIRE 2021 Challenge on Quality Enhancement of Compressed Video: Methods and Results

Ren Yang   Radu Timofte   Jing Liu   Yi Xu   Xinjian Zhang   Minyi Zhao
Shuigeng Zhou   Kelvin C.K. Chan   Shangchen Zhou   Xiangyu Xu   Chen Change Loy
Xin Li   Fanglong Liu   He Zheng   Lielin Jiang   Qi Zhang   Dongliang He   Fu Li
Qingqing Dang   Yibin Huang   Matteo Maggioni   Zhongqian Fu   Shuai Xiao
Cheng li   Thomas Tanay   Fenglong Song   Wentao Chao   Qiang Guo   Yan Liu
Jiang Li   Xiaochao Qu   Dewang Hou   Jiayu Yang   Lyn Jiang   Di You
Zhenyu Zhang   Chong Mou   Iaroslav Koshelev   Pavel Ostyakov   Andrey Somov
Jia Hao   Xueyi Zou   Shijie Zhao   Xiaopeng Sun   Yiting Liao   Yuanzhi Zhang
Qing Wang   Gen Zhan   Mengxi Guo   Junlin Li   Ming Lu   Zhan Ma
Pablo Navarrete Michelini   Hai Wang   Yiyun Chen   Jingyu Guo   Liliang Zhang
Wenming Yang   Sijung Kim   Syehoon Oh   Yucong Wang   Minjie Cai   Wei Hao
Kangdi Shi   Liangyan Li   Jun Chen   Wei Gao   Wang Liu   Xiaoyu Zhang
Linjie Zhou   Sixin Lin   Ru Wang

Abstract

This paper reviews the first NTIRE challenge on quality enhancement of compressed video, with focus on proposed solutions and results. In this challenge, the new Large-scale Diverse Video (LDV) dataset is employed. The challenge has three tracks. Tracks 1 and 2 aim at enhancing the videos compressed by HEVC at a fixed QP, while Track 3 is designed for enhancing the videos compressed by x265 at a fixed bit-rate. Besides, the quality enhancement of Tracks 1 and 3 targets at improving the fidelity (PSNR), and Track 2 targets at enhancing the perceptual quality. The three tracks totally attract 482 registrations. In the test phase, 12 teams, 8 teams and 11 teams submitted the final results of Tracks 1, 2 and 3, respectively. The proposed methods and solutions gauge the state-of-the-art of video quality enhancement. The homepage of the challenge: `https://github.com/RenYang-home/NTIRE21_VEnh`

1. Introduction

During the recent years, there is increasing popularity

of video streaming over the Internet [12] and the demands on high-quality and high-resolution videos are also rapidly increasing. Due to the limited bandwidth of Internet, video compression [61, 49] plays an important role to significantly reduce the bit-rate and facilitate transmitting a large number of high-quality and high-resolution videos. However, video compression unavoidably leads to compression artifacts, thus resulting in the loss of both fidelity and perceptual quality and the degradation of 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. Due to the rate-distortion trade-off in data compression, enhancing 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 in this direction [71, 70, 55, 35, 72, 63, 68, 21, 64, 15, 67, 26, 53], among which [71, 70, 55] are single-frame quality enhancement methods, while [35, 72, 68, 21, 64, 15, 67, 26, 53] propose enhancing quality by taking advantage of temporal correlation. Besides, [53] aims at improving the perceptual quality of compressed video. Other works [71, 70, 55, 35, 72, 68, 21, 64, 15, 67, 26] focus on advancing the performance on Peak Signal-to-Noise Ratio (PSNR) to achieve higher fidelity to the uncompressed video. These works show the promising future of this research field. However, the scale of training sets used in

---

the existing methods are incremental and different methods are also tested on various test sets. For example, at the beginning, [55] utilizes the image database BSDS500 [4] for training, without any video. Then, [71] trains the model on a small video dataset including 26 video sequences, and [70] enlarges the training set to 81 videos. In the multi-frame methods, [35] is trained on the Vimeo-90K dataset [66], in which each clip only contains 7 frames, and thus it is insufficient for studying on enhancing long video sequences. Then, the Vid-70 dataset, which includes 70 video sequences, is used as the training set in [72, 68, 64, 26]. Meanwhile, [21] and [15] collected 142 and 106 uncompressed videos for training, respectively. Besides, the commonly used test sets in existing literature are the JCT-VC dataset [7] (18 videos), the test set of Vid-70 [68] (10 videos) and Vimeo-90K (only 7 frames in each clip). Standardizing a larger and more diverse dataset is not only beneficial for training video enhancement models but also meaningful for establishing a convincing benchmark in this area.

The NTIRE 2021 challenge on enhancing compressed video is a step forward in benchmarking the video quality enhancement algorithms. It uses the newly proposed Large-scale Diverse Video (LDV) [69] dataset, which contains 240 videos with the diversities of content, motion and frame-rate, *etc*. The LDV dataset is introduced in [69] along with the analyses of challenge results. In the following, we first describe the NTIRE 2021 challenge, and then introduce the proposed methods and their results.

2. NTIRE 2021 Challenge

The objectives of the NTIRE 2021 challenge on enhancing compressed video are: (i) to gauge and push the state-of-the-art in video quality enhancement; (ii) to compare different solutions; (iii) to promote the newly proposed LDV dataset; and (iv) to promote more challenging video quality enhancement settings.

This challenge is one of the NTIRE 2021 associated challenges: nonhomogeneous dehazing [3], defocus deblurring using dual-pixel [1], depth guided image relighting [17], image deblurring [40], multi-modal aerial view imagery classification [34], learning the super-resolution space [36], quality enhancement of compressed video (this report), video super-resolution [47], perceptual image quality assessment [19], burst super-resolution [5], high dynamic range [42].

2.1. LDV dataset

As introduced in [69], our LDV dataset contains 240 videos with 10 categories of scenes, *i.e.*, *animal*, *city*, *close-up*, *fashion*, *human*, *indoor*, *park*, *scenery*, *sports* and *vehicle*. Besides, among the 240 videos in LDV, there are 48 fast-motion videos, 68 high frame-rate (≥ 50) videos and 172 low frame-rate (≤ 30) videos. Additionally, the camera is slightly shaky (*e.g.*, captured by handheld camera) in 75 videos of LDV, and 20 videos in LDV are with the dark environments, *e.g.*, at night or in the rooms with insufficient light. In the challenge of NTIRE 2021, we divide the LDV dataset into training, validation and test sets with 200, 20 and 20 videos, respectively. The test set is further split into two sets with 10 videos in each for the tracks of fixed QP (Tracks 1 and 2) and fixed bit-rate (Track 3), respectively. The 20 validation videos consist of 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 out of the 20 validation videos and 4 among the 10 videos in each test set are with high frame-rates. There are five fast-motion videos in the validation set. In the test sets for fixed QP and fixed bit-rate tracks, there are three and two fast-motion videos, respectively.

2.2. Fidelity tracks

The first part of this challenge aims at improving the quality of compressed video towards fidelity. We evaluate the fidelity via PSNR. Additionally, we also calculate the Multi-Scale Structural SIMilarity index (MS-SSIM) [60] for the proposed methods.

Track 1: Fixed QP. In Track 1, the videos are compressed following the typical settings of the existing literature [71, 70, 55, 35, 72, 63, 68, 21, 64, 15, 26, 53], *i.e.*, using the official HEVC test model (HM) at fixed QPs. In this challenge, we compress videos by the default configuration of the Low-Delay P (LDP) mode (*encoder_lowdelay_P_main.cfg*) of HM 16.20[1] at QP = 37. In this setting, due to the regularly changed QPs at the frame-level, the compression quality normally fluctuates regularly among frames. Besides, it does not enable the rate control strategy, and therefore the frame-rate does not have impact on compression. This may make this track to be an easy task.

Track 3: Fixed bit-rate. Track 3 targets at a more practical scenario. In video streaming, rate control has been utilizing as a popular strategy to constraint the bit-rate into the limited bandwidth. In this track, we compress videos by the x265 library of FFmpeg[2] with rate control enabled and set the target bit-rate as 200 kbps, by 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
```

---

[1]https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.20

[2]https://johnvansickle.com/ffmpeg/releases/ffmpeg-release-amd64-static.tar.xz

Table 1. The results of Track 1 (fixed QP, fidelity)

| Team | PSNR (dB) | | | | | | | | | | | MS-SSIM |
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Average	Average
BILIBILI AI & FDU	**33.69**	31.80	**38.31**	34.44	28.00	32.13	29.68	**29.91**	35.61	31.62	**32.52**	**0.9562**
NTU-SLab	31.30	**32.46**	36.96	**35.29**	**28.30**	**33.00**	**30.42**	29.20	**35.70**	**32.24**	32.49	0.9552
VUE	31.10	32.00	36.36	34.86	28.08	32.26	30.06	28.54	35.31	31.82	32.04	0.9493
NOAHTCV	30.97	31.76	36.25	34.52	28.01	32.11	29.75	28.56	35.38	31.67	31.90	0.9480
Gogoing	30.91	31.68	36.16	34.53	27.99	32.16	29.77	28.45	35.31	31.66	31.86	0.9472
NJU-Vision	30.84	31.55	36.08	34.47	27.92	32.01	29.72	28.42	35.21	31.58	31.78	0.9470
MT.MaxClear	31.15	31.21	37.06	33.83	27.68	31.68	29.52	28.43	34.87	32.03	31.75	0.9473
VIP&DJI	30.75	31.36	36.07	34.35	27.79	31.89	29.48	28.35	35.05	31.47	31.65	0.9452
Shannon	30.81	31.41	35.83	34.17	27.81	31.71	29.53	28.43	35.05	31.49	31.62	0.9457
HNU_CVers	30.74	31.35	35.90	34.21	27.79	31.76	29.49	28.24	34.99	31.47	31.59	0.9443
BOE-IOT-AIBD	30.69	30.95	35.65	33.83	27.51	31.38	29.29	28.21	34.94	31.29	31.37	0.9431
Ivp-tencent	30.53	30.63	35.16	33.73	27.26	31.00	29.22	28.14	34.51	31.14	31.13	0.9405
MFQE [72]	30.56	30.67	34.99	33.59	27.38	31.02	29.21	28.03	34.63	31.17	31.12	0.9392
QECNN [70]	30.46	30.47	34.80	33.48	27.17	30.78	29.15	28.03	34.39	31.05	30.98	0.9381
DnCNN [74]	30.41	30.40	34.71	33.35	27.12	30.67	29.13	28.00	34.37	31.02	30.92	0.9373
ARCNN [16]	30.29	30.18	34.35	33.12	26.91	30.42	29.05	27.97	34.23	30.87	30.74	0.9345
Unprocessed video	30.04	29.95	34.16	32.89	26.79	30.07	28.90	27.84	34.09	30.71	30.54	0.9305

```
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
```

Note that we utilize the two-pass scheme to ensure the accuracy of rate control. Due to the fixed bit-rate, the videos of different frame-rates, various motion speeds and diverse contents have to be compressed to a specific bit-rate per second. This makes the compression quality of different videos dramatically different, and therefore may make it a more challenging track than Track 1.

2.3. Perceptual track

We also organize a track aiming at enhancing the compressed videos towards perceptual quality. In this track, the performance is evaluated via the Mean Opinion Score (MOS). We also report the performance on other perceptual metrics as references, such as the Learned Perceptual Image Patch Similarity (LPIPS) [76], Fréchet Inception Distance (FID) [25], Kernel Inception Distance (KID) [6] and Video Multimethod Assessment Fusion (VMAF) [29].

Track 2: perceptual quality enhancement. In Track 2, we compress videos with the same settings as Track 1. The task of this track is to generate visually pleasing enhanced videos, and the scores are ranked according to MOS values from 15 subjects. The scores range from $s = 0$ (poorest quality) to $s = 100$ (best quality). The groundtruth videos are given to the subjects as the standard of $s = 100$, but the subjects are asked to rate videos in accordance with the visual quality, instead of the similarity to the groundtruth. We linearly normalize the scores (s) of

each subject to

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

in which s_{max} and s_{min} denote the highest and the lowest score of each subject, respectively. In the experiment, we insert five repeated videos to check the concentration of each subject to ensure the consistency of rating. Eventually, we omit the scores from the three least concentrated subjects, and the final MOS values are averaged among 12 subjects. Besides, we also calculate the LPIPS, FID, VID and VMAF values to evaluate the proposed methods.

3. Challenge results

3.1. Track 1: Fixed QP, Fidelity

The numerical results of Track 1 are shown in Table 1. On the top part, we show the results of the 12 methods proposed in this challenge. The unprocessed video indicates the compressed videos without enhancement. Additionally, we also train the models of the existing methods on the training set of the newly proposed LDV dataset, and report the results in Table 1. It can be seen from Table 1, the proposed methods in the challenge outperform the existing methods, and therefore advance the state-of-the-art of video quality enhancement.

The PSNR improvements of the 12 proposed methods range from 0.59 dB to 1.98 dB, and the improvements of MS-SSIM range between 0.0100 and 0.0257. The BILI-BILI AI & FDU Team achieves the best average PSNR and MS-SSIM performance in this track. They improve the average PSNR and MS-SSIM by 1.98 dB and 0.0257, respectively. The NTU-SLab and VUE Teams rank second

Table 2. The results of Track 2 (fixed QP, perceptual)

Team	MOS ↑											LPIPS ↓	FID ↓	VID ↓	VMAF ↑
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Average	Average	Average	Average	Average
BILIBILI AI & FDU	**85**	**75**	67	**71**	**83**	61	53	**89**	64	58	**71**	**0.0429**	**32.17**	**0.0137**	75.69
NTU-SLab	63	73	61	**71**	74	62	**62**	75	62	**82**	69	0.0483	34.64	0.0179	71.55
NOAHTCV	73	72	73	64	66	**84**	61	57	61	58	67	0.0561	46.39	0.0288	68.92
Shannon	67	66	74	67	64	65	56	61	57	54	63	0.0561	50.61	0.0332	69.06
VUE	62	67	72	62	73	56	61	36	47	60	60	0.1018	72.27	0.0561	**78.64**
BOE-IOT-AIBD	50	40	63	66	67	41	58	50	42	50	53	0.0674	62.05	0.0447	68.78
(*anonymous*)	45	38	70	58	60	33	**62**	22	**69**	48	50	0.0865	83.77	0.0699	69.70
MT.MaxClear	34	47	**80**	62	50	41	59	13	38	34	46	0.1314	92.42	0.0818	77.30
Unprocessed video	34	29	29	53	39	34	35	44	30	34	36	0.0752	48.94	0.0303	65.72

Table 3. The results of Track 3 (fixed bit-rate, fidelity)

Team	PSNR (dB)											MS-SSIM
	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20	Average	Average
NTU-SLab	**30.59**	**28.14**	35.37	**34.61**	**32.23**	**34.66**	28.17	20.38	27.39	**32.13**	**30.37**	**0.9484**
BILIBILI AI & FDU	29.85	27.01	34.17	34.25	31.62	34.34	**28.51**	**21.13**	**28.01**	30.65	29.95	0.9468
MT.MaxClear	29.47	27.89	**35.63**	34.16	30.93	34.29	26.25	20.47	27.38	30.38	29.69	0.9423
Block2Rock Noah-Hisilicon	30.20	27.31	34.50	33.55	31.94	34.14	26.62	20.43	26.74	30.96	29.64	0.9405
VUE	29.93	27.31	34.58	33.64	31.79	33.86	26.54	20.44	26.54	30.97	29.56	0.9403
Gogoing	29.77	27.23	34.36	33.47	31.61	33.71	26.68	20.40	26.38	30.77	29.44	0.9393
NOAHTCV	29.80	27.13	34.15	33.38	31.60	33.66	26.38	20.36	26.37	30.64	29.35	0.9379
BLUEDOT	29.74	27.09	34.08	33.29	31.53	33.33	26.50	20.36	26.35	30.57	29.28	0.9384
VIP&DJI	29.64	27.09	34.12	33.44	31.46	33.50	26.50	20.34	26.19	30.56	29.28	0.9380
McEhance	29.57	26.81	33.92	33.10	31.36	33.40	25.94	20.21	26.07	30.27	29.07	0.9353
BOE-IOT-AIBD	29.43	26.68	33.72	33.02	31.04	32.98	26.25	20.26	25.81	30.09	28.93	0.9350
Unprocessed video	29.17	26.02	32.52	32.22	30.69	32.54	25.48	20.03	25.28	29.41	28.34	0.9243

and third, respectively. The average PSNR performance of NTU-SLab is slightly lower (0.03 dB) than the BILIBILI AI & FDU Team, and the PSNR of VUE is 0.48 dB lower than the best method. We also report the detailed results on the 10 test videos (#1 to #10) in Table 1. The results indicate that the second-ranked team NTU-SLab outperforms BILIBILI AI & FDU on 7 videos. It shows the better generalization capability of NTU-SLab than BILIBILI AI & FDU. Considering the average PSNR and generalization, the BILIBILI AI & FDU and NTU-SLab Teams are both the winners of this track.

3.2. Track 2: Fixed QP, Perceptual

Table 2 shows the results of Track 2. In Track 2, the BILIBILI AI & FDU Team achieves the best MOS performance on 5 of the 10 test videos and has the best average MOS performance. The results of the NTU-SLab team are the best on 3 videos, and their average MOS performance ranks second. The NOAHTCV Team is the third in the ranking of average MOS. We also report the results of LIPIS, FID, VID and VMAF, which are the popular metrics for evaluating perceptual quality of image and video. It can be seen from Table 2 that BILIBILI AI & FDU, NTU-SLab and NOAHTCV still rank at the first, second and third places on LPIPS, FID and VID. It indicates that these per-

ceptual metrics are effective on measuring the subjective quality. However, the rank on VMAF is obviously different from MOS. Besides, some teams perform worse than the unprocessed videos on LPIPS, FID and VID, while their MOS values are all higher than the unprocessed videos. This may show that the perceptual metrics are not always reliable for evaluating the visual quality of video.

3.3. Track 3: Fixed bit-rate, Fidelity

Table 3 shows the results of Track 3. In this track, we use the different videos as the test set, denoted as #11 to #20. The top three teams in this track are NTU-SLab, BILIBILI AI & FDU and MT.MaxClear. The NTU-SLab Team achieves the best results on 6 videos and also ranks first on average PSNR and MS-SSIM. They improve the average PSNR by 2.03 dB. BILIBILI AI & FDU and MT.MaxClear enhance PSNR by 1.61 dB and 1.35 dB, respectively.

3.4. Efficiency

Table 4 reports the running time of the proposed methods. In the methods with top quality performance, NTU-SLab is most the most efficient method. The NTU-SLab and BILIBILI AI & FDU Teams achieve the best and the second best quality performance for all three tracks, but the method of NTU-SLab is several times faster than BILIBILI

Team	Running time (s) per frame			Platform	GPU	Ensemble / Fusion	Extra training data
	Track 1	Track 2	Track 3				
BILIBILI AI & FDU	9.00	9.45	9.00	PyTorch	Tesla V100/RTX 3090	Flip/Rotation x8	Bilibili [27], YouTube [28]
NTU-SLab	1.36	1.36	1.36	PyTorch	Tesla V100	Flip/Rotation x8	Pre-trained on REDS [39]
VUE	34	36	50	PyTorch	Tesla V100	Flip/Rotation x8	Vimeo90K [66]
NOAHTCV	12.8	12.8	12.8	TensorFlow	Tesla V100	Flip/Rotation x8	DIV8K [20] (Track 2)
MT.MaxClear	2.4	2.4	2.4	PyTorch	Tesla V100	Flip/Rotation/Multi-model x12	Private dataset
Shannon	12.0	1.5	-	PyTorch	Tesla T4	Flip/Rotation x8 (Track 1)	-
Block2Rock Noah-Hisilicon	-	-	300	PyTorch	Tesla V100	Flip/Rotation x8	YouTube [28]
Gogoing	8.5	-	6.8	PyTorch	Tesla V100	Flip/Rotation x4	REDS [39]
NJUVision	3.0	-	-	PyTorch	Titan RTX	Flip/Rotation x6	SJ4K [48]
BOE-IOT-AIBD	1.16	1.16	1.16	PyTorch	GTX 1080	Overlapping patches	-
(*anonymous*)	-	4.52	-	PyTorch	Tesla V100	-	Partly finetuned from [57]
VIP&DJI	18.4	-	12.8	PyTorch	GTX 1080/2080 Ti	Flip/Rotation x8	SkyPixel [45].
BLUEDOT	-	-	2.85	PyTorch	RTX 3090	-	Dataset of MFQE 2.0 [21]
HNU_CVers	13.72	-	-	PyTorch	RTX 3090	Overlapping patches	-
McEhance	-	-	0.16	PyTorch	GTX 1080 Ti	-	-
Ivp-tencent	0.0078	-	-	PyTorch	GTX 2080 Ti	-	-
MFQE [72]	0.38	-	-	TensorFlow	TITAN Xp	-	-
QECNN [70]	0.20	-	-	TensorFlow	TITAN Xp	-	-
DnCNN [74]	0.08	-	-	TensorFlow	TITAN Xp	-	-
ARCNN [16]	0.02	-	-	TensorFlow	TITAN Xp	-	-

AI & FDU. Therefore, the NTU-SLab Team makes the best trade-off between quality performance and time efficiency. Moreover, among all proposed methods, the method from the Ivp-tencent Team is most time-efficient. It is able to enhance more than 120 frames per second, so it may be practical for the scenario of high frame-rates.

3.5. Training and test

It can be seen from Table 4 that the top teams utilized extra training data in additional to the 200 training videos pf LDV [69] provided in the challenge. It may indicate that the scale of training database has obvious effect on the test performance. Besides, the ensemble strategy [51] has been widely used in the top methods, and the participants observe the quality improvement of their methods when using ensemble, showing the effectiveness of the ensemble strategy for video enhancement.

4. Challenge methods and teams

4.1. BILIBILI AI & FDU Team

Track 1. In Track 1, they propose a Spatiotemporal Model with Gated Fusion (SMGF) [65] for enhancing compressed video, based on the framework of [15]. The pipeline of the proposed method is illustrated in the top of Figure 1-(a).

As the preliminary step, they first extract the metadata from the HEVC bit-stream with HM-Decoder. They set the frame whose QP score in metadata is lower than that of the two adjacent frames (PQF [72]) as candidate frame. 1) They fixedly select adjacent frame I_{t-1}/I_{t+1} of I_t as the first reference frame. 2) Taking the preceding part as example, they recursively take the next preceding candidate frame of the last selected reference frame as a new reference frame until
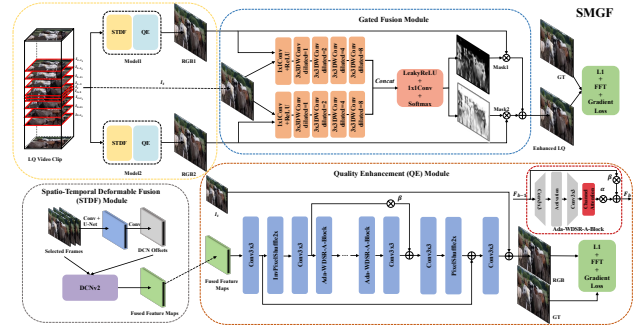
(a) Network architecture.
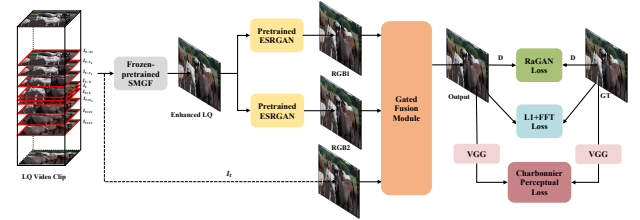
(b) The proposed pSMGF.

Figure 1. Network architectures of the BILIBILI AI & FDU Team.

there are 4 reference frames or no candidate frames are left. 3) If there is no more candidate frame and the number of selected reference frames is smaller than 4, then repeatedly pad it with the last selected frame until there are 4 frames.

They feed 9 frames (8 references and a target frame) into the Spaito-Temporal Deformable Fusion (STDF) [15] module to capture spatiotemporal information. The output of STDF module is then sent to the Quality Enchancement (QE) module. They employ a stack of adaptive WDSR-A-Block from C2CNet [18] as the QE module. As illustrated in Figure 1, a Channel Attention (CA)

(a) An Overview of BasicVSR++
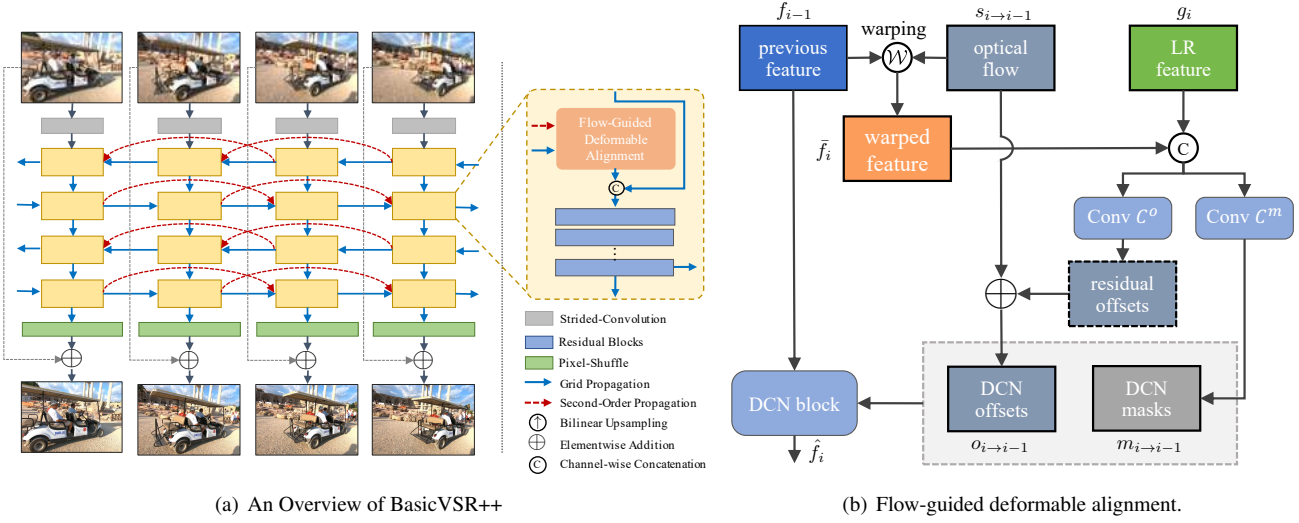
(b) Flow-guided deformable alignment.

Figure 2. The method proposed by the NTU-SLab Team.

layer [77] is additionally attached at the bottom of WDSR-A-Block [73]. Comparing with the CA layer in RCAN [77], there are two learnable parameters α and β initialized with 1 and 0.2 in Ada-WDSR-A-Block. Besides, the channels of the feature map and block in the QE module are 128 and 96, respectively. The channels of Ada-WDSR-A-Block are implemented as {64, 256, 64}.

Additionally, they propose a novel module to improve the performance of enhancement at the bottom of the pipeline. As shown in the middle-top of Figure 1-(a), though each model has the same architecture (STDF with QE) and training strategy (L1 + FFT + Gradient [56] loss), one is trained on the official training sets, and the other is on extra videos crawled from Bilibili [65] and YouTube [28], named as BiliTube4k. To combine the predictions of two models, they exploit a stack of layers to output the mask M and then aggregate predictions. The mask M in gated fusion module is with the same resolution of the target frame ranging from $[0, 1]$, the final enhanced low-quality frame is formulated as

$$\hat{I} = M \otimes \hat{I}_1 \oplus (1 - M) \otimes \hat{I}_2. \quad (2)$$

Track 2. In Track 2, they reuse and freeze the models from Track 1, attach ESRGAN [58] at the bottom of SMGF, and propose a perceptual SMGF (pSMGF). As shown in Figure 1-(b), they first take the enhanced low-quality frames from Track 1. Then they feed these enhanced frames into ESRGAN and train the Generator and Discriminator iteratively. Specifically, they use the ESRGAN pre-trained on DIV2K dataset [2], remove the pixel shuffle layer in ESR-GAN, and supervise the model with {L1 + FFT + RaGAN + Perceptual} loss. They also utilize the gated fusion module proposed after ESRGAN, which is proposed in SMGF.

Specifically, one of the ESRGANs is tuned on the official training sets, and the other is on extra videos collected from Bilibili [65] and YouTube [28], named as BiliTube4k. The predictions of two models are aggregated via (2).

Track 3. They utilize the model in Track 1 as the pre-trained model, and then fine-tune it on training data of Track 3 with early stopping. Another difference is that they take the neighboring preceding/following I/P frames as candidate frames, instead of PQFs.

4.2. NTU-SLab Team

Overview. The NTU-SLab Team proposes the BasicVSR++ method for this challenge. BasicVSR++ consists of two deliberate modifications for improving *propagation* and *alignment* designs of BasicVSR [9]. As shown in Figure 2-(a), given an input video, 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.

Second-Order Grid Propagation. Motivated by the effectiveness of the bidirectional propagation, they devise a grid propagation scheme to enable *repeated refinement through propagation*. More specifically, the intermediate features are propagated backward and forward in time in an alternating manner. Through propagation, the information from different frames can be "revisited" and adopted for feature refinement. Compared to existing works that propagate features only once, grid propagation repeatedly extracts information from the entire sequence, improving

feature expressiveness. To further enhance the robustness of propagation, they relax the assumption of first-order Markov property in BasicVSR and adopt a second-order connection, realizing a second-order Markov chain. With this relaxation, information can be aggregated from different spatiotemporal locations, improving robustness and effectiveness in occluded and fine regions.

Flow-Guided Deformable Alignment. Deformable alignment [52, 57] has demonstrated significant improvements over flow-based alignment [23, 66] thanks to the offset diversity [10] intrinsically introduced in deformable convolution (DCN) [13, 79]. However, deformable alignment module can be difficult to train [10]. The training instability often results in offset overflow, deteriorating the final performance. To take advantage of the offset diversity while overcoming the instability, they propose to employ optical flow to guide deformable alignment, motivated by the strong relation between deformable alignment and flow-based alignment [10]. The graphical illustration is shown in Figure 2-(b).

Training. The training consists of only one stage. For Tracks 1 and 3, only Charbonnier loss [11] is used as the loss function. For Track 2, perceptual and adversarial loss functions are also used. The training patch size is 256×256, randomly cropped from the original input images. They perform data augmentation, *i.e.*, rotation ($0°$, $90°$, $180°$, $270°$), horizontal flip, and vertical flip. For the Track 1, they initialize the model from a variant trained for video super-resolution to shorten the training time. The models for the other two tracks are initialized from the model of Track 1. During the test phase, they test the proposed models with ensemble ($\times 8$) testing, *i.e.*, rotation $90°$, flipping the input in four ways (none, horizontally, vertically, both horizontally and vertically) and averaging their outputs.

4.3. VUE Team

Tracks 1 and 3. In the fidelity tracks, the VUE Team proposes the methods based on BasicVSR [9], as shown in Figure 3. For Track 1, they propose a two-stage method. In stage-1, they train two BasicVSR models with different parameters followed by the self-ensemble strategy. Then, they fuse the two results using average sum. In stage-2, they train another BasicVSR model. For Track 3, they propose to tackle this problem by using VSR methods without the last upsampling layer. They train four BasicVSR models with different parameter settings followed by the self-ensemble strategy. Then, they average the four outputs as the final result.

Track 2. In Track 2, they propose a novel solution dubbed "Adaptive Spatial-Temporal Fusion of Two-Stage Multi-Objective Networks" [78]. It is motivated by the fact that it is hard to design unified training objectives which are
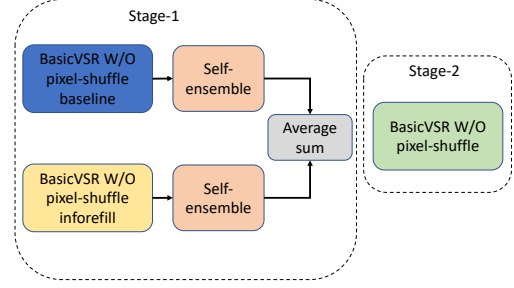
(a) The method for Track 1.
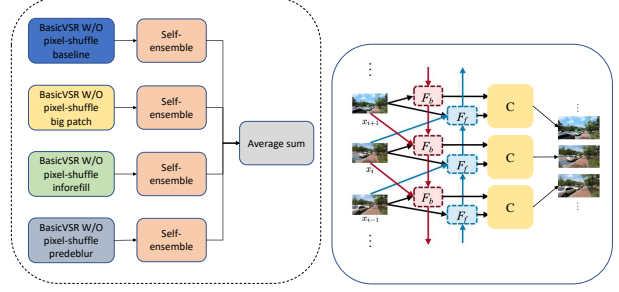
(b) The method for Track 3. (c) BasicVSR w/o pixel-shuffle.

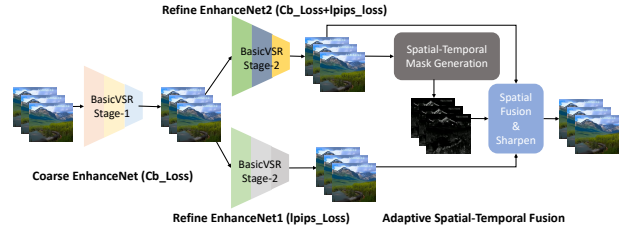Figure 3. The methods of the VUE Team for Tracks 1 and 3.

Figure 4. The proposed method of the VUE Team for Track 2.

perceptual-friendly for enhancing regions with smooth content and regions with rich textures simultaneously. To this end, they propose to adaptively fuse the enhancement results from the networks trained with two different optimization objectives. As shown in Figure 4, the framework is designed with two stages. 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 [11]. At the second stage, they train two BasicVSR models for different refinement purposes. One refined BasicVSR model (denoted as EnhanceNet2) is trained with

$$3 \cdot \text{Charbonnier loss} + \text{LPIPS loss.} \qquad (3)$$

Another refined BasicVSR model (denoted as EnhanceNet1) is trained with the mere LPIPS loss [76]. This way, EnhanceNet1 is good at recovering textures to satisfying human perception requirement but it can result in temporal flickering for smooth regions of videos, meanwhile EnhanceNet1 produces much more smooth results, espe-

cially, temporal flickering is well eliminated.

To overcome this issue, they devise a novel adaptive spatial-temporal fusion scheme. Specifically, the spatial-temporal mask generation module is proposed to produce spatial-temporal masks and it is used to fuse the two network outputs:

$$I_{out}^t = (1 - mask_t) \times I_{out,1}^t + mask_t \times I_{out,2}^t, \quad (4)$$

where $mask_t$ is the generated mask for the t-th frame, $I_{out,1}^t$ and $I_{out,2}^t$ are the t-th output frames of EnhanceNet1 and EnhanceNet2, respectively. The mask $mask_t = f(I_{out,2}^{t-1}, I_{out,2}^t, I_{out,2}^{t+1})$ is adaptively generated from $I_{out,2}^{t-1}$, $I_{out,2}^t$ and $I_{out,2}^{t+1}$ as follows. First, the variance map V^t is calculated from $I_{out,2}^t$ by:

$$
\begin{aligned}
V_{i,j}^t &= Var(Y_{out,2}^t[i-5:i+5, j-5:j+5]) \\
Y_{out,2}^t &= (I_{out,2}^t[:,:,0] + I_{out,2}^t[:,:,1] + I_{out,2}^t[:,:,2])/3,
\end{aligned}
\quad (5)
$$

where $Var(x)$ means the variance of x. Then, they normalize the variance map in a temporal sliding window to generate the mask $mask_t$:

$$
\begin{aligned}
mask_t &= (V^t - q)/(p - q) \\
p &= max([V^{t-1}, V^t, V^{t+1}]) \\
q &= min([V^{t-1}, V^t, V^{t+1}]).
\end{aligned}
\quad (6)
$$

Intuitively, when a region is smooth, its local variance is small, otherwise, its local variance is large. Therefore, smooth region more relies on the output of EnhanceNet2 while the rich-texture region gets more recovered details from EnhanceNet1. With temporal sliding window, the temporal flickering effect is also well eliminated.

4.4. NOAHTCV Team

As show in Figure 5, the input images includes three frames, *i.e.*, the current frame plus the previous and the next Peak Quality Frame (PQF). The first step consists in a shared feature extraction with a stack of residual blocks and subsequently an U-Net is used to jointly predict the individual offsets for each of the three inputs. Such offsets are then used to implicitly align and fuse the features. Note that, there is no loss used as supervision for this step. After the initial feature extraction and alignment, they use a multi-head U-Net with shared weights to process each input feature, and at each scale of the encoder and decoder, they fuse the U-Net features with scale-dependant deformable convolutions, which are denoted in black in Figure 5. The output features of the U-Net are fused for a final time, and the output fused features are finally processed by a stack of residual blocks to predict the final output. This output is in fact a residual compression information which is added to
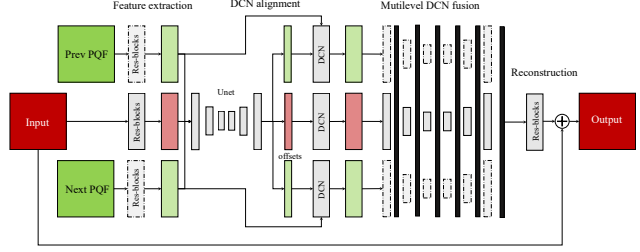

Figure 5. The proposed method of the NOAHTCV Team.

the input frame to produce the enhanced output frame. The models utilized for all three tracks are the same. The difference is the loss function, *i.e.*, they use the $L2$ loss for Tracks 1 and 3, and use GAN Loss + Perceptual loss + $L2$ loss for Track 2.

4.5. MT.MaxClear Team

The proposed model is based on EDVR [57], which uses the deformable convolution to align features between neighboring frames and the reference frame, and then combines all aligned frame features to reconstruct the reference frame. The deformable convolution module in EDVR is difficult to train due to the unstable of DCN offset. They propose to add two DCN offset losses to regularize the deformable convolution module which makes the training of DCN offset much more stable. They use Charbonnier penalty loss [11], DCN offsets Total Variation loss and DCN offsets Variation loss to train the model. The Charbonnier penalty loss is more robust than $L2$ loss. 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. The training of DCN is much more stable due to the aforementioned two offsets losses, but the EDVR model performs better if the loss weights of DCN offsets Total Variation loss and DCN offsets Variation loss gradually decays to zero during training. In Track 2, they add the sharpening operation on the enhanced frames for better visual perception.
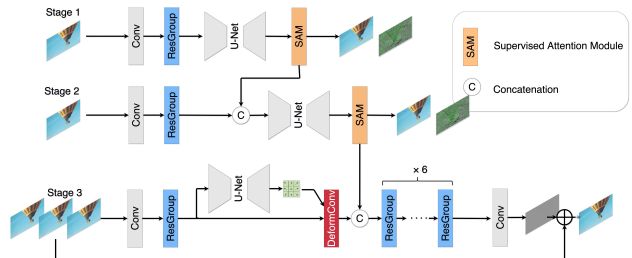

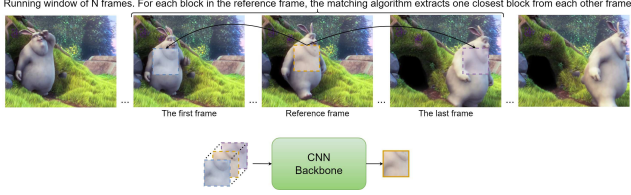Figure 6. The proposed generator of the Shannon Team.

Figure 7. The illustration of the proposed method of the Block2Rock Noah-Hisilicon Team.

4.6. Shannon Team

The Shannon Team introduces a disentangled attention for compression artifact analysis. Unlike previous works, they propose to address the problem of artifact reduction from a new perspective: disentangle complex artifacts by a disentangled attention mechanism. Specifically, they adopt a multi-stage architecture in which early stage also provides a disentangled attention. Their key insight is that there are various types of artifacts created by video compression, some of which result in significant blurring effect in the reconstructed signals, and some of which generate artifacts, such as blocking and ringing. Algorithms could be either too aggressive and amplify erroneous high-frequency components, or too conservative and tend to smooth over ambiguous components, both resulting in bad cases that seriously affect subjective visual impression. The proposed disentangled attention aims to reduce these bad cases.

In Track 2, they use the LPIPS loss and only feed the high-frequency components to the discriminator. Before training the model, they analyze the quality fluctuation among frames [72] and train the model from-easy-to-hard. To generate the attention map, they use the supervised attention module proposed proposed in [37]. The overall structure of the proposed generator is shown in Figure 6. The discriminator is simply composed of several convolutional layer-ReLU-strided convolutional layers blocks, and its final output is a 4×4 confidence map.

Let F_{lp} denote the low-pass filtering, which is implemented in a differentiable manner with *Kornia* [43]. They derive the supervised information for attention map:

$$\mathcal{R}_{lp} = F_{lp}(x) - x, \tag{7}$$

$$\mathcal{R}_y = y - x, \tag{8}$$

$$D(y, x) = sgn(\mathcal{R}_y \odot \mathcal{R}_{lp}), \tag{9}$$

where $sgn(\cdot)$ denotes signum function that extracts the sign of a given pixel value; $\odot$ is the element-wise product, and y refers to the output, and x refers to the compressed input.
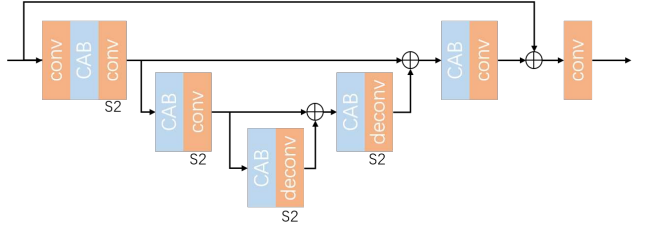
4.7. Block2Rock Noah-Hisilicon Team

This team makes a trade-off between spatial and temporal sizes in favor of the latter by performing collaborative CNN-based restoration of square patches extracted
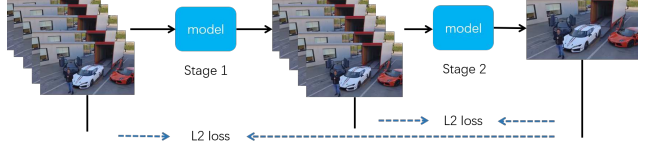
(a) Overall architecture
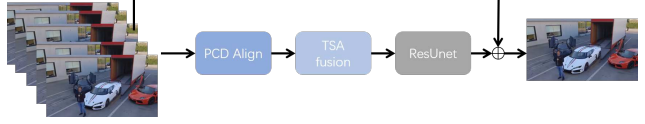
(b) Model architecture

(c) ResUNet architecture

Figure 8. The proposed method of Gogoing Team.

by block matching algorithm, which finds correlated areas across consequent frames. Due to performance concerns, the proposed block matching realization is trivial: for each patch $\boldsymbol{p}$ at the reference frame, they search for and extract a single closest patch $\boldsymbol{p}_i$ from each other frame $\boldsymbol{f}_i$ in a sequence, based on squared $L2$ distance:

$$\boldsymbol{p}_i = \boldsymbol{C}(\hat{u}, \hat{v})\boldsymbol{f}_i, \text{ where } \hat{u}, \hat{v} = \arg\min_{u,v} ||\boldsymbol{C}(u, v)\boldsymbol{f}_i - \boldsymbol{p}||_2^2. \tag{10}$$

Here $\boldsymbol{C}(u, v)$ is a linear operator that crops patch which top-left corner is located at (u, v) pixel coordinates of the canvas. As it is shown for example in [24], the search for the closest patch in (10) requires a few pointwise operations and two convolutions (one of which is a box filter), which could be done efficiently in the frequency domain based on convolution theorem. The resulted patches are then stacked and passed to a CNN backbone, which outputs a single enhanced version of the reference patch. The overall process is presented in Figure 7.

Since the total number of pixels being processed by CNN depends quadratically on spatial size and linearly on the input's temporal size, they propose to use inference on small patches of size 48×48 pixels to decrease the spatial size of backbone inputs. For example, the two times decrease in height and width allows increasing temporal dimension by a factor of four. With existing well-performing CNNs, this fact allows the temporal dimension to increase up to 50 frames since such models are designed to be trained on patches with spatial sizes of more than 100 pixels (typically
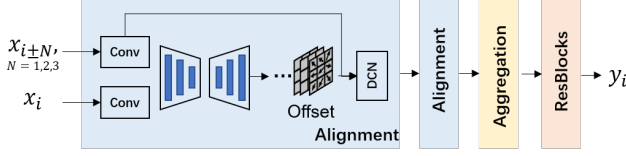
Figure 9. The proposed method of the NJU-Vision Team.

128 pixels).

In the solution, they use the EDVR network [57] as a backbone, and the RRDB network [59] acted as a baseline. For EDVR, they stacke patches in a separate dimension, while for RRDB, they stacke patches in channel dimension. Reference patch was always the first in a stack.

For training network weights, they use the $L1$ distance between output and target as an objective to minimize through back-propagation and stochastic gradient descent. They use the Adam optimizer [32] with learning rate increased from zero to $2e^{-4}$ during a warmup period, which then was gradually decreased by a factor of 0.98 after each epoch. The total number of epochs was 100 with 2000 unique batches passed to the network during each one. To stabilize the training and prevent divergence, they use the adaptive gradient clipping technique with weight $\lambda = 0.01$, as proposed in [8].

4.8. Gogoing Team

The overall structure adopts a two-stage model, as shown in the Figure 8-(a). As Figure 8-(b) shows, for the temporal part, they use the temporal module in EDVR [57], that is PCD module and TSA module. The number of input frames is 7. In the spatial part, they combine the UNet [41] and the residual attention module [77] to form ResUNet, as shown in Figure 8-(c). In the training phase, they use the 256×256 RGB patches from training set as input, and augment them with random horizontal flips and 90 rotations. The number of input frames is seven. All models are optimized by using the Adam [32] optimizer with mini-batches of size 12, with the learning rate being initialized to 4×10^{-4} using CosineAnnealingRestartLR strategy. Loss function is the $L2$ loss.

4.9. NJU-Vision Team

As shown in Figure 9, the NJU-Vision Team proposes a method utilizing a progressive deformable alignment module and a spatial-temporal attention based aggregation module, based on [57]. Data augmentation is also applied with training data augmentation by randomly flipping in horizontal and vertical orientations, and rotating at $90°$, $180°$, and $270°$, and evaluation ensemble by flipping in horizontal and vertical orientations, and rotating at $90°$, $180°$, and $270°$ to obtain the averaged results.
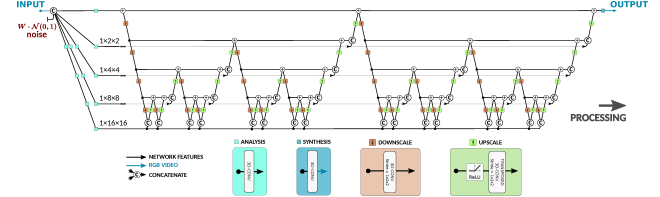
(a) 3D–MultiGrid BackProjection network (MGBP–3D).
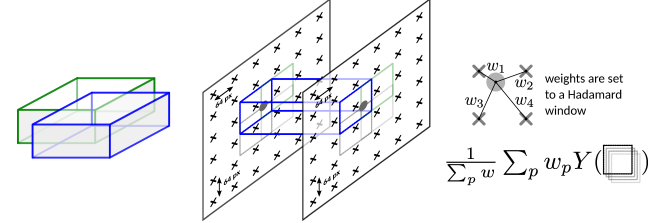
(b) Multiscale discriminator.
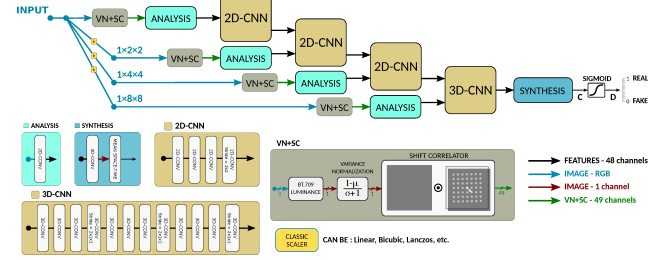
(c) Overlapping–patches inference strategy.

Figure 10. Network architectures of the BOE-IOT-AIBD Team.

4.10. BOE-IOT-AIBD Team

Tracks 1 and 3. Figure 10-(a) displays the diagram of the MGBP–3D network used in this challenge, which was proposed by the team members in [38]. The system uses two backprojection residual blocks that run recursively in five levels. Each level downsamples only space, not time, by the factor of 2. The Analysis and Synthesis modules convert an image into features space and vice–versa using single 3D–convolutional layers. The Upscaler and Downscaler modules are composed of single strided (transposed and conventional) 3D–convolutional layers. Every Upscaler and Downscaler module shares the same configuration in a given level but they do not share parameters. Small number of features are set at high resolution and they increase at lower resolutions to reduce the memory footprint in high resolution scales. In addition, they add a 1–channel noise video stream to the input that is only used for the Perceptual track. In the fidelity tracks, the proposed model is trained by the $L2$ loss.

To process long video sequences they use the patch based approach from [38], in which they average the output of overlapping video patches taken from the compressed degraded input. First, they divide input streams into overlap-

ping patches (of same size as training patches) as shown in Figure 10-(b); second, they multiply each output by weights set to a Hadamard window; and third, they average the results. In the experiments they use overlapping patches separated by 243 pixels in vertical and horizontal directions and one frames in time direction.

Track 2. Based on the model for Tracks 1 and 3, they add noise inputs to activate and deactivate the generation of artificial details for the Perceptual track. In MGBP–3D, they generate one channel of Gaussian noise concatenated to the bicubic upscaled input. The noise then moves to different scales in the Analysis blocks. This change allows using the overlapping patch solution with noise inputs, as it simply represent an additional channel in the input.

They further employ a discriminator shown in Figure 10-(c) to achieve adversarial training. The loss function used for Track 2 is a combination of the GAN loss, LPIPS loss and the $L1$ loss. $Y_{n=0}$ and $Y_{n=1}$ as the outputs of the generator using noise amplitudes $n = 0$ and $n = 1$, respectively. X indicates the groundtruth. The loss function can be expressed as as follows:

$$
\begin{aligned}
\mathcal{L}(Y, X; \theta) = {} & 0.001 \cdot \mathcal{L}_G^{RSGAN}(Y_{n=1}) \\
& + 0.1 \cdot \mathcal{L}^{perceptual}(Y_{n=1}, X) \\
& + 10 \cdot \mathcal{L}^{L1}(Y_{n=0}, X) \,,
\end{aligned} \tag{11}
$$

Here, $\mathcal{L}^{perceptual}(Y_{n=1}, X) = \text{LPIPS}(Y_{n=1}, X)$ and the Relativistic GAN loss [31], is given by:

$$
\begin{aligned}
& \mathcal{L}_D^{RSGAN} \\
& = -\mathbb{E}_{(real, fake) \sim (\mathbb{P}, \mathbb{Q})} \left[\log(\text{sigmoid}(C(real) - C(fake))) \right], \\
& \mathcal{L}_G^{RSGAN} \\
& = -\mathbb{E}_{(real, fake) \sim (\mathbb{P}, \mathbb{Q})} \left[\log(\text{sigmoid}(C(fake) - C(real))) \right],
\end{aligned} \tag{12}
$$

where C is the output of the discriminator just before the sigmoid function, as shown in Figure 10-(c). In (12), $real$ and $fake$ are the sets of inputs to the discriminator, which contains multiple inputs with multiple scales, *i.e.*,

$$
fake = \{Y_{n=1}\}, real = \{X\}. \tag{13}
$$

4.11. (*anonymous*)

Network. For enhancing the perceptual quality of heavily compressed videos, there are two main problems that need to be solved: spatial texture enhancement and temporal smoothing. Accordingly, in Track 2, they propose a multi-stage approach with specific designs for the above problems. Figure 11 depicts the overall framework of the proposed method, which consists of three processing stages. In stage I, they enhance the distorted textures of manifold objects in each compressed frame by the proposed Texture Imposition Module (TIM). In stage II, they
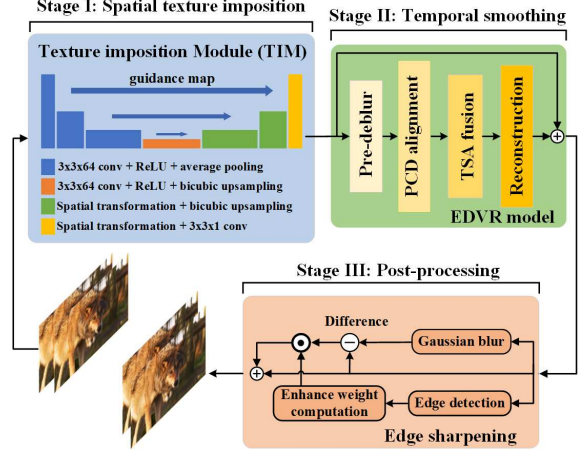

Figure 11. The proposed method of (*anonymous*).

suppress the flickering and discontinuity of the enhanced consecutive frames by a video alignment and enhancement network, *i.e.*, EDVR [57]. In stage III, they further enhance the sharpness of the enhanced videos by several classical techniques (opposite to learning-based network) .

In particular, TIM is based on the U-Net [44] architecture to leverage the multi-level semantic guidance for texture imposition. In TIM, natural textures of different objects in compressed videos are assumed as different translation styles, which need to be learned and imposed; thus, they apply the affine transformations [30] in the decoder path of the U-Net, to impose the various styles in a spatial way. The parameters of the affine transformations are learned from several convolutions with the input of guidance map from the encoder path of the U-Net. Stage II is based on the video deblurring model of EDVR, which consists of four modules, the PreDeblure, the PCD Align, the TSA fusion and the reconstruction module. In stage III, a combination of classical unsharp masking [14] and edge detection [46] methods are adopted to further enhance the sharpness of the video frames. In particular, they first obtain the difference map between the original and its Gaussian blurred images. Then, they utilize the Sobel operator [46] to detect the edges of original image to weight the difference map, and add the difference map to the original image.

Training. For stage I, TIM is supervised by three losses: Charbonnier loss (epsilon is set to be $1e{-}6$) $\mathcal{L}_{\text{pixel}}$, VGG loss $\mathcal{L}_{\text{vgg}}$ and Relativistic GAN loss $\mathcal{L}_{\text{gan}}$. The overall loss function is defined as: $\mathcal{L} = 0.01 \times \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{gan}} + 0.005 \times \mathcal{L}_{\text{vgg}}$. For stage II, they fine-tune the video deblurring model of EDVR with the training set of NTIRE, supervised by the default Charbonnier loss. Instead of the default training setting in [57], they first fine-tune the PreDeblur module with 80,000 iterations. Then they fine-tune the overall model with a small learning rate of 1×10^{-5} for another 155,000 iterations.
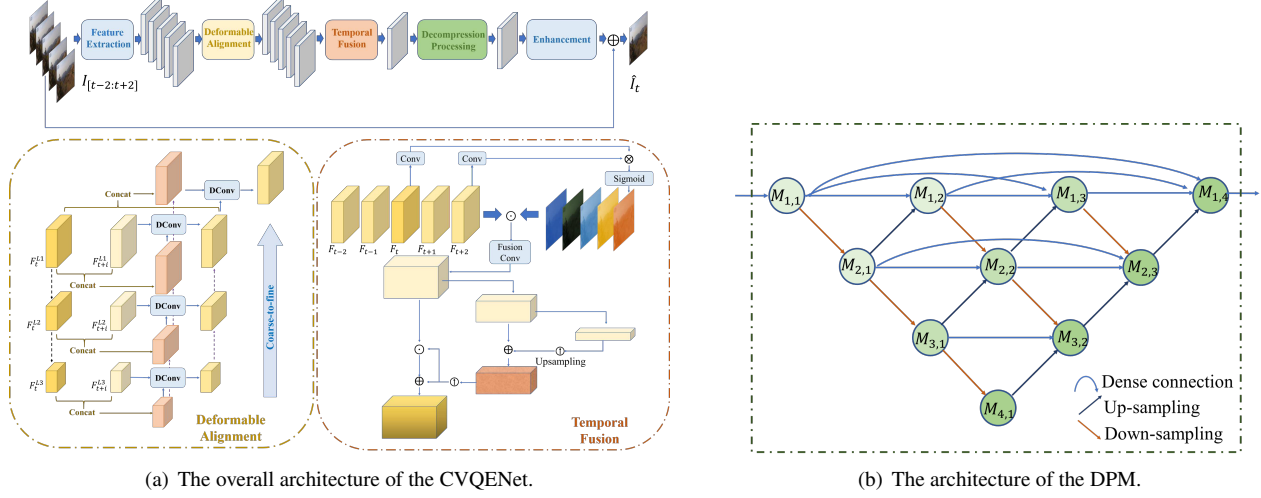
(a) The overall architecture of the CVQENet.　　(b) The architecture of the DPM.

Figure 12. The proposed method of the VIP&DJI Team for Track 1.

(a) The whole framework of the DUVE.　　(b) The architecture of U-Net.
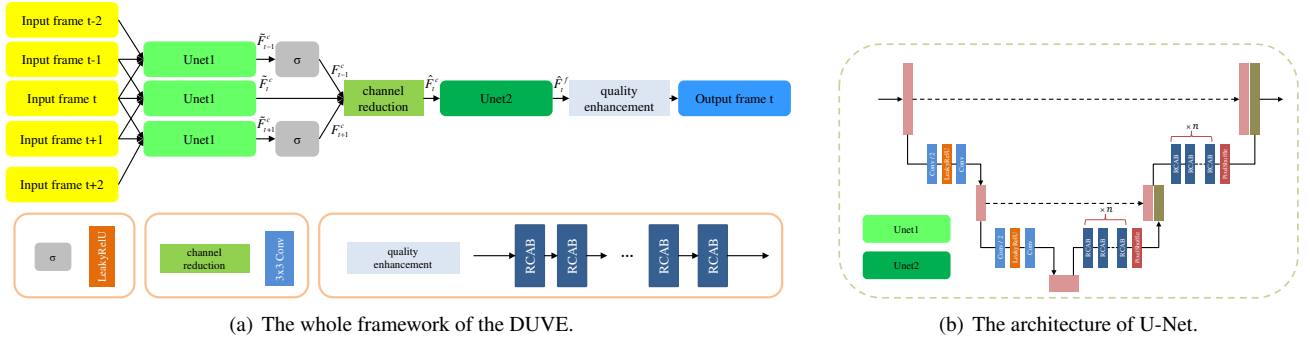
Figure 13. The proposed method of the VIP&DJI Team for Track 3.

4.12. VIP&DJI Team

4.12.1　Track 1

As shown in Figure 12-(a), the architecture of the proposed CVQENet consists of five parts, which are feature extraction module, inter-frame feature deformable alignment module, inter-frame feature temporal fusion module, decompression processing module and feature enhancement module. The input of CVQENet includes five consecutive compressed video frames $I_{[t-2:t+2]}$, and the output is a restored middle frame $\hat{I}_t$ that is as close as possible to the uncompressed middle frame O_t:

$$\hat{I}_t = CVQENet(I_{[t-2:t+2]}, \theta) \tag{14}$$

where θ represents the set of all parameters of the CVQENet. Given a training dataset, the loss function L is defined below to be minimized:

$$L = ||\hat{I}_t - O_t||_1 \tag{15}$$

where $||.||_1$ denotes the $L1$ loss. The following will introduce each module of CVQENet in detail.

Feature extraction module (FEM). The Feature extraction module contains one convolutional layer (Conv) to extract the shallow feature maps $F^0_{[t-2:t+2]}$ from the compressed video frames $I_{[t-2:t+2]}$ and 10 stacked residual blocks without Batch Normalization (BN) layer to process the feature maps further.

Inter-frame feature deformable alignment module (FDAM). Next, for the feature maps extracted by the FEM, FDAM aligns the feature map corresponding to each frame to the middle frame that needs to be restored. It can be aligned based on optical flow, deformable convolution, 3D convolution, and other methods, and CVQENet uses the method based on deformable convolution. For simplicity, CNet uses the Pyramid, Cascading and Deformable convolutions (PCD) module proposed by EDVR [57] to align feature maps. The detailed module structure is shown in Figure 12-(a). The PCD module aligns the feature map of each frame F^{L1}_{t+i} to the feature map of the middle frame F^{L1}_t. In the alignment process, $F^{L1}_{[t-2:t+2]}$ is progressively convolved and down-sampled to obtain a small-scale feature maps $F^{L2}_{[t-2:t+2]}, F^{L3}_{[t-2:t+2]}$, and then the align is processed from F^{L3} to F^{L1} in a coarse-to-fine manner. Align

$F_{[t-2:t+2]}^{L1}$ with F_t^{L1} respectively to obtain the aligned feature map $F_{[t-2:t+2]}$.

Inter-frame feature temporal fusion module (FTFM). The FTFM is used to fuse the feature maps from each frame to a compact and informative feature map for further process. CVQENet directly uses the TSA module proposed by EDVR [57] for fusion, and the detailed structure is shown in Figure12-(a). The TSA module generates temporal attention maps through the correlation between frames and then performs temporal feature fusion through the convolutional layer. Then, the TSA module uses spatial attention to further enhance the feature map.

Decompression processing module (DPM). For the fused feature map, CVQENet uses the DPM module to remove artifacts caused by compression. Inspired by RBQE [62], DPM consists of a simple densely connected UNet, as shown in Figure12-(b). The $M_{i,j}$ cell contains Efficient Channel Attention (ECA) [54] block, convolutional layer and residual blocks. The ECA block performs adaptive feature amplitude adjustment through the channel attention mechanism.

Feature quality enhancement module (FQEM). CVQENet add the output of DPM with F_t^0, then enters them into the feature quality enhancement module. The shallow feature map F_t^0 contains a wealth of detailed information of the middle frame, which can help restore the middle frame. The FQEM contains 20 stacked residual blocks without Batch Normalization (BN) layer to enhance the feature map further and one convolutional layer (Conv) to generate the output frame image $\hat{I}_t$.

4.12.2 Track 3

Motivated by FastDVDnet [50] and DRN [22], they propose the DUVE network for compressed video enhancement for Track 3, and the whole framework is shown in Figure 13-(a). It can be seen that given five consecutive compressed frames $I_{[t-2:t+2]}$, the goal of DUVE is to restore an uncompressed frame $\hat{O}_t$. Specifically, for five continuous input frames, each of the three consecutive images forms a group, so that the five images are overlapped into three groups $I_{[t-2:t]}$, $I_{[t-1:t+1]}$ and $I_{[t:t+1]}$. Then, the three groups are fed into Unet1 to get coarse restored feature maps $\widetilde{F}_{t-1}^c$, $\widetilde{F}_t^c$ and $\widetilde{F}_{t+1}^c$, respectively. Considering the correlation between different frames and the current reconstruction frame I_t, the two groups of coarse feature maps $\widetilde{F}_{t-1}^c$ and $\widetilde{F}_{t+1}^c$ are filtered by nonlinear activation function σ to get F_{t-1}^c and F_{t+1}^c. Next, F_{t-1}^c, $\widetilde{F}_t^c$ and F_{t+1}^c are concatenated along the channel dimension, and then pass through channel reduction module to obtain fused coarse feature map $\hat{F}_t^c$. To further reduce compression artifacts, they apply UNet2 on $\hat{F}_t^c$ to acquire more fine feature map $\hat{F}_t^f$. Finally, a quality enhancement module takes the fine feature map to achieve the
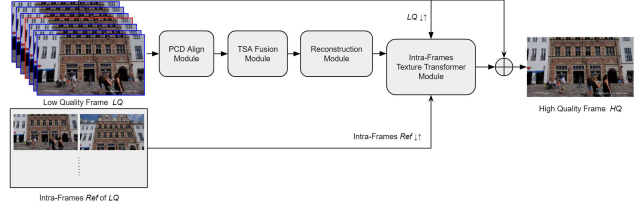
(a) The proposed framework.
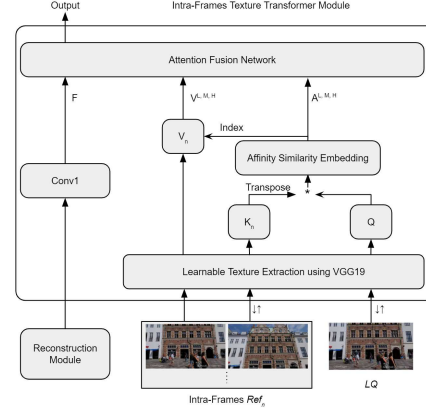
(b) Intra-Frames Texture Transformer (IFTT).
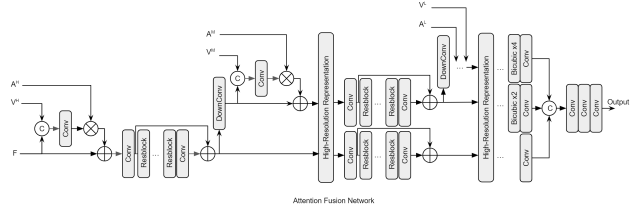
(c) Attention fusion network in IFTT.

Figure 14. The proposed method of the BLUEDOT Team.

restored frame $\hat{O}_t$. The detailed architecture of Unet1 and Unet2 is shown in Figure 13-(b). In the proposed method, the mere difference between Unet1 and Unet2 is the number n of Residual Channel Attention Blocks (RCAB) [77]. The L_2 loss is utilized as the loss function.

4.13. BLUEDOT Team

The proposed method is shown in Figure 14, which is motivated by the idea that intra-frames usually have better video quality than inter-frames. It means that more information about the texture of the videos from intra-frames can be extracted. They built and trained a neural network built on EDVR [57] and TTSR [67]. Relevances of all intra-frames in the video are measured and one frame of the high- est relevance with the current frame is embedded in the network. They carry out the two-stage training in the same network to obtain a video enhancement result with the restored intra-frame. In the first stage, the model is learned by low-quality intra-frames. Then, in the second stage, the model is trained

with predicted intra-frames by fisrt-stage model.

4.14. HNU_CVers Team

The HNU_CVers Team proposes the patch-based heavy compression recovery models for Track 1.
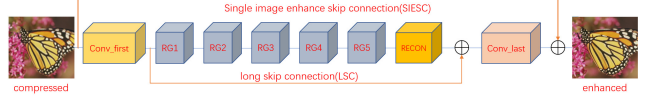
Single-frame residual channel-attention network. First, they delete the upsampling module from [77] and add a skip connection to build a new model. Different from [77], the RG (Residual Group) number is set as 5 and there are 16 residual blocks [33] in each RG. Each residual block is composed of 3×3 convolutional layers and ReLU with 64 feature maps. The single-frame model architecture is shown in Figure 15-(a), called RCAN_A. The model is trained with the $L1$ loss.

Multi-frame residual channel-attention video network. They further make the model compact by five consecutive frames of images which are enhanced by RCAN_A. The multi-frame model architecture is shown in Figure 15-(b). Five consecutive frames are stained with different colors in Figure 15-(b) after enhanced by RCAN_A. In order to mine consecutive frames information, they combine the central frame with each frame. For aligning, they designe the [conv (64 features) + ReLU + Resbolcks (64 features) $\times$ 5] network as the align network with the shared parameters to each combination. Immediately after it, temporal and spatial attention are fused [57]. After getting a single frame feature map colored yellow (shown in Figure 15-(b), they adopt another model called RCAN_B, which has the same structure as RCAN_A. Finally, the restored RGB image is obtained through a convolution layer. The model is also trained with the $L1$ loss.

Patch Integration Method (PIM). They further propose a patch-based fusion model to strengthen the reconstruction ability of the multi-frame model. The motivation for designing PIM is to mine the reconstruction ability of the model from part to the whole. For a small patch, the reconstruction ability of the model at the center will exceed the reconstruction ability at the edge. Therefore, they propose feed the overlapping patches to the proposed network, In the reconstructed patches, they remove the edges that overlap with the neighboring patches and only keep the high-confidence part in the center.

4.15. McEnhance Team

The McEnhance Team combines video super-resolution technology [15, 57] with multi-frame enhancement [72, 21] to create a new end-to-end network, as illustrated in Figure 16. First, they choose the current frame and it's neighbor peak Frames as the input data. Then, they feed them to the deformable convolution network to align. As a result, complementary information from both target and reference frames can be fused with in the operation. In the following, they feed them separately to the QE module [15] and the

(a) Single-frame Residual Channel-Attention Network (RCAN_A)
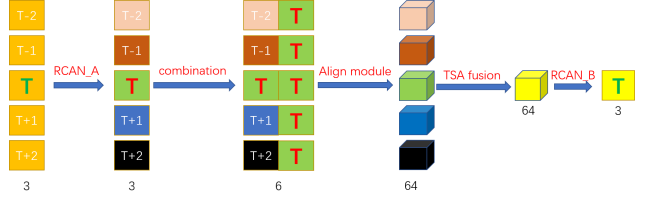
(b) Multi-frame Residual Channel-Attention Video Network (RCVN)

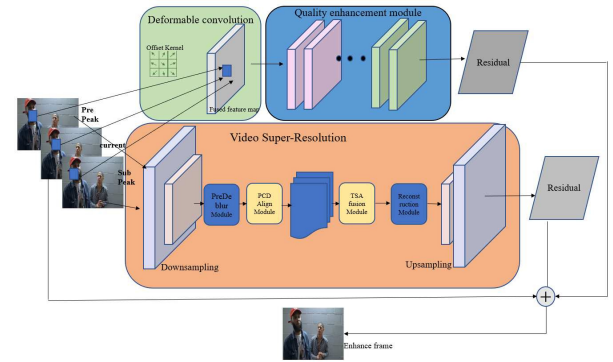Figure 15. The proposed method of the HNU_CVers Team.

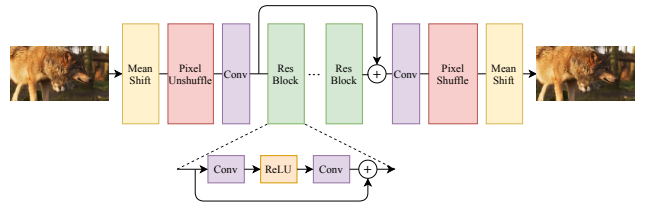Figure 16. The proposed method of the McEnhance Team.

Figure 17. The proposed BRNet of the Ivp-tencent Team.

Temporal and Spatial Attention (TSA) [57] network. Finally, they put the two residual frames on the raw target frame. There are two steps in training stage. First, they calculate the PSNR of each frame in training set and make labels of peak PSNR frames. Secondly, they send the current frame and two neighbor peak PSNR frames to the net.

4.16. Ivp-tencent

As Figure 17 shows, the Ivp-tencent Team proposes a Block Removal Network (BRNet) to reduce the block artifacts in compressed video for quality enhancement. Inspired by EDSR [33] and FFDNet [75], the proposed BRNet first uses a mean shift module (Mean Shift) to normalize the input frame, and then adopts a reversible downsampling operation (Pixel Unshuffle) to process the frame, which splits the compressed frame into four down-sampled

sub-frames. Then, the sub-frames are fed into a convolutional network shown in Figure 17, in which they use eight residual blocks. Finally, they use an up-sampling operation (Pixel Shuffle) and a mean shift module to reconstruct the enhanced frame. Note that, the up-sampling operation (Pixel Shuffle) is the inverse operation of the down-sampling operation (Pixel Unshuffle). During the training phase, they crop the given compressed images to 64×64 and feed to the network by batch size of 64. The Adam [32] algorithm is adopted to optimize $L1$ loss, and learning rate is set to 10^{-4}. The model is trained for 100,000 epochs.

The proposed BRNet achieves higher efficiency compared with EDSR and FFDNet. The reason is two-fold: First, the input frame is sub-sampled into several sub-frames as inputs to the network. While maintaining the quality performance, the network parameters are effectively reduced and the receiving field of the network is increased. Second, by removing the batch normalization layer of the residual blocks, about 40% of the memory usage can be saved during training.

Acknowledgments

Appendix: Teams and affiliations

NTIRE 2021 Team

Challenge:
NTIRE 2021 Challenge on Quality Enhancement of Compressed Video
Organizer(s):
Ren Yang (`ren.yang@vision.ee.ethz.ch`),
Radu Timofte (`radu.timofte@vision.ee.ethz.ch`)
Affiliation(s):
Computer Vision Lab, ETH Zurich, Switzerland

Bilibili AI & FDU Team

Title(s):
Tracks 1 and 3: Spatiotemporal Model with Gated Fusion for Compressed Video Artifact Reduction
Track 2: Perceptual Spatiotemporal Model with Gated Fusion for Compressed Video Artifact Reduction
Member(s):
Jing Liu (`liujing04@bilibili.com`), Yi Xu (`yxu17@fudan.edu.cn`), Xinjian Zhang, Minyi Zhao, Shuigeng Zhou
Affiliation(s):
Bilibili Inc.; Fudan University, Shanghai, China

NTU-SLab Team

Title(s):
BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment
Member(s):
Kelvin C.K. Chan (`chan0899@e.ntu.edu.sg`), Shangchen Zhou, Xiangyu Xu, Chen Change Loy
Affiliation(s):
S-Lab, Nanyang Technological University, Singapore

VUE Team

Title(s):
Tracks 1 and 3: Leveraging off-the-shelf BasicVSR for Video Enhancement
Track 2: Adaptive Spatial-Temporal Fusion of Two-Stage Multi-Objective Networks
Member(s):
Xin Li (`lixin41@baidu.com`), He Zheng (`zhenghe01@baidu.com`), Fanglong Liu, Lielin Jiang, Qi Zhang, Dongliang He, Fu Li, Qingqing Dang
Affiliation(s):
Department of Computer Vision Technology (VIS), Baidu Inc., Beijing, China

NOAHTCV Team

Title(s):
Multi-scale network with deformable temporal fusion for compressed video restoration
Member(s):
Fenglong Song (`songfenglong@huawei.com`), Yibin Huang, Matteo Maggioni, Zhongqian Fu, Shuai Xiao, Cheng li, Thomas Tanay
Affiliation(s):
Huawei Noah's Ark Lab, Huawei Technologies Co., Ltd.

MT.MaxClear Team

Title(s):
Enhanced EDVRNet for Quality Enhanced of Heavily Compressed Video
Member(s):
Wentao Chao (`cwt1@meitu.com`), Qiang Guo, Yan Liu, Jiang Li, Xiaochao Qu
Affiliation(s):
MTLab, Meitu Inc., Beijing, China

Shannon Team

Title(s):
Disentangled Attention for Enhancement of Compressed Videos
Member(s):
Dewang Hou (`dewh@pku.edu.cn`), Jiayu Yang, Lyn Jiang, Di You, Zhenyu Zhang, Chong Mou
Affiliation(s):
Peking university, Shenzhen, China; Tencent, Shenzhen, China

Block2Rock Noah-Hisilicon Team

Title(s):
Long Temporal Block Matching for Enhancement of Uncompressed Videos
Member(s):
Iaroslav Koshelev (`Iaroslav.Koshelev@skoltech.ru`), Pavel Ostyakov, Andrey Somov, Jia Hao, Xueyi Zou
Affiliation(s):
Skolkovo Institute of Science and Technology, Moscow, Russia; Huawei Noah's Ark Lab; HiSilicon (Shanghai) Technologies CO., LIMITED, Shanghai, China

Gogoing Team

Title(s):
Two-stage Video Enhancement Network for Different QP Frames
Member(s):
Shijie Zhao (`zhaoshijie.0526@bytedance.com`), Xiaopeng Sun, Yiting Liao, Yuanzhi Zhang, Qing Wang, Gen Zhan, Mengxi Guo, Junlin Li
Affiliation(s):
ByteDance Ltd., Shenzhen, China

NJU-Vision Team

Title(s):
Video Enhancement with Progressive Alignment and Data Augmentation
Member(s):
Ming Lu (`luming@smail.nju.edu.cn`), Zhan Ma
Affiliation(s):
School of Electronic Science and Engineering, Nanjing University, China

BOE-IOT-AIBD Team

Title(s):
Fully 3D–Convolutional MultiGrid–BackProjection Network

Member(s):
Pablo Navarrete Michelini (`pnavarre@boe.com.cn`)
Affiliation(s):
BOE Technology Group Co., Ltd., Beijing, China

VIP&DJI Team

Title(s):
Track 1: CVQENet: Deformable Convolution-based Compressed Video Quality Enhancement Network
Track 3: DUVE: Compressed Videos Enhancement with Double U-Net
Member(s):
Hai Wang (`wanghai19@mails.tsinghua.edu.cn`), Yiyun Chen, Jingyu Guo, Liliang Zhang, Wenming Yang
Affiliation(s):
Tsinghua University, Shenzhen, China; SZ Da-Jiang Innovations Science and Technology Co., Ltd., Shenzhen, China

BLUEDOT Team

Title(s):
Intra-Frame texture transformer Network for compressed video enhancement
Member(s):
Sijung Kim (`jun.kim@blue-dot.io`), Syehoon Oh
Affiliation(s):
Bluedot, Seoul, Republic of Korea

HNU_CVers Team

Title(s):
Patch-Based Multi-Frame Residual Channel-Attention Networks For Video Enhancement
Member(s):
Yucong Wang (`1401121556@qq.com`), Minjie Cai
Affiliation(s):
College of Computer Science and Electronic Engineering, Hunan University, China

McEnhance Team

Title(s):
Parallel Enhancement Net
Member(s):
Wei Hao (`haow6@mcmaster.ca`), Kangdi Shi, Liangyan Li, Jun Chen
Affiliation(s):
McMaster University, Ontario, Canada

Ivp-tencent Team

Title(s):
BRNet: Block Removal Network
Member(s):
Wei Gao (gaowei262@pku.edu.cn), Wang Liu, Xiaoyu Zhang, Linjie Zhou, Sixin Lin, Ru Wang
Affiliation(s):
School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China; Peng Cheng Laboratory, Shenzhen, China; Tencent, Shenzhen, China

References

[1] Abdullah Abuolaim, Radu Timofte, Michael S Brown, et al. NTIRE 2021 challenge for defocus deblurring using dual-pixel images: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 126–135, 2017. 6

[3] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2021 nonhomogeneous dehazing challenge report. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[4] 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. 2

[5] Goutam Bhat, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 challenge on burst super-resolution: Methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[6] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 3

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

[8] Andrew Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. *arXiv preprint arXiv:2102.06171*, 2021. 10

[9] 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. 6, 7

[10] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 7

[11] 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. 7, 8

[12] 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

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 764–773, 2017. 7

[14] Guang Deng. A generalized unsharp masking algorithm. *IEEE transactions on Image Processing*, 20(5):1249–1261, 2010. 11

[15] 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, 14

[16] 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. 3, 5

[17] Majed El Helou, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, et al. NTIRE 2021 depth guided image relighting challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[18] Dario Fuoli, Zhiwu Huang, Martin Danelljan, and Radu Timofte. Ntire 2020 challenge on video quality mapping: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 476–477, 2020. 5

[19] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Yu Qiao, Shuhang Gu, Radu Timofte, et al. NTIRE 2021 challenge on perceptual image quality assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[20] 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

[21] 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. 1, 2, 5, 14

[22] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5407–5416, 2020. 13

[23] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3897–3906, 2019. 7

[24] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Trans. Graph.*, 35(6), Nov. 2016. 9

[25] 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. 3

[26] 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

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

[28] Google Inc. YouTube. https://www.youtube.com. 5, 6

[29] Netflix Inc. VMAF - video multi-method assessment fusion. https://github.com/Netflix/vmaf. 3

[30] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015. 11

[31] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 11

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 10, 15

[33] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 136–144, 2017. 14

[34] Jerrick Liu, Oliver Nina, Radu Timofte, et al. NTIRE 2021 multi-modal aerial view object classification challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[35] 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, 2

[36] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. NTIRE 2021 learning the super-resolution space challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[37] Armin Mehri, Parichehr B Ardakani, and Angel D Sappa. MPRNet: Multi-path residual network for lightweight image super resolution. In *Proceedings of the IEEE/CVF Win-ter Conference on Applications of Computer Vision (WACV)*, pages 2704–2713, 2021. 9

[38] Pablo Navarrete Michelini, Wenbin Chen, Hanwen Liu, Dan Zhu, and Xingqun Jiang. Multi–grid back–projection networks. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):279–294, 2021. 10

[39] 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

[40] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on image deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[41] Seungjun Nah, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2020 challenge on image and video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 416–417, 2020. 10

[42] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Aleš Leonardis, Radu Timofte, et al. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[43] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3674–3683, 2020. 9

[44] 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. 11

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

[46] Irwin Sobel. Camera models and machine perception. Technical report, Computer Science Department, Technion, 1972. 11

[47] Sanghyun Son, Suyoung Lee, Seungjun Nah, Radu Timofte, Kyoung Mu Lee, et al. NTIRE 2021 challenge on video super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2

[48] 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

[49] 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

[50] Matias Tassano, Julie Delon, and Thomas Veit. FastDVDnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1354–1363, 2020. 13

[51] 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. 5

[52] Hua Wang, Dewei Su, Chuangchuang Liu, Longcun Jin, Xianfang Sun, and Xinyi Peng. Deformable non-local network for video super-resolution. *IEEE Access*, 7:177734–177744, 2019. 7

[53] 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

[54] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 13

[55] 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, 2

[56] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 650–666. Springer, 2020. 6

[57] 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, 7, 8, 10, 11, 12, 13, 14

[58] 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 Workshops (ECCVW)*, pages 0–0, 2018. 6

[59] 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 *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018. 10

[60] 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. 2

[61] 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

[62] Qunliang Xing, Mai Xu, Tianyi Li, and Zhenyu Guan. Early exit or not: resource-efficient blind quality enhancement for compressed images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 275–292. Springer, 2020. 13

[63] 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, 2

[64] 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

[65] 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. 5, 6

[66] 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. 2, 5, 7

[67] 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, 13

[68] 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

[69] Ren Yang and Radu Timofte. NTIRE 2021 challenge on quality enhancement of compressed video: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 2, 5

[70] 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, 2, 3, 5

[71] 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, 2

[72] 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, 3, 5, 9, 14

[73] 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. 6

[74] 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. 3, 5

[75] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 14

[76] 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. 3, 7

[77] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 6, 10, 13, 14

[78] 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. 7

[79] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9308–9316, 2019. 7