Adaptive Spatial-Temporal Fusion of Multi-Objective Networks for Compressed Video Perceptual Enhancement

He Zheng, Xin Li, Fanglong Liu, Lielin Jiang, Qi Zhang,
Fu Li, Qingqing Dang, Dongliang He
Department of Computer Vision Technology (VIS), Baidu Inc.
Building #2, Baidu Science Park, Haidian Distinct, Beijing, China
{zhenghe01, lixin41, liufanglong, jianglielin, zhangqi44,
lifu, dangqingqing, hedongliang01}@baidu.com

Abstract

Perceptual quality enhancement of heavily compressed videos is a difficult, unsolved problem because there still not exists a suitable perceptual similarity loss function between two video pairs. Motivated by the fact that it is hard to design unified training objectives which are perceptual-friendly for enhancing regions with smooth content and regions with rich textures simultaneously, in this paper, we propose a simple yet effective novel solution dubbed "Adaptive Spatial-Temporal Fusion of Two-Stage Multi-Objective Networks" (ASTF) to adaptive fuse the enhancement results from networks trained with two different optimization objectives. Specifically, the proposed ASTF takes an enhancement frame along with its neighboring frames as input to jointly predict a mask to indicate regions with high-frequency textual details. Then we use the mask to fuse two enhancement results which can retain both smooth content and rich textures. Extensive experiments show that our method achieves a promising performance of compressed video perceptual quality enhancement.

1. Introduction

In recent years, we have witnessed the explosive growth of video data over the Internet. In order to transmit video with limited bandwidth, video compression is essential to significantly reduce the bit rate. However, existing compression algorithms often introduce artifacts, which severely degrade the Quality of Experience (QoE) [33, 9, 5, 1, 20]. Thus, it's crucial to study on compressed video quality enhancement (VQE).

Recently, there is only limited study on quality enhancement for compressed video [33, 9, 5, 25]. Multi-Frame Quality Enhancement (MFQE 1.0) [33] first leverage temporal information for VQE. MFQE 2.0 [9] was proposed to further to improve the performance which also adopts a temporal fusion scheme that incorporates dense optical flow for motion compensation. Spatio-Temporal Deformable Fusion (STDF) [5] aggregates temporal information while avoiding explicit optical flow estimation. However, all these methods use pixel-wise metrics, such as MSE, PSNR and SSIM, to compute the similarity between two images which fail to account for many nuances of human perception. MW-GAN [25] proposed a generative adversarial network (GAN) based on multi-level wavelet packet transform to recover the high-frequency details for enhancing the perceptual quality of compressed video.

While it is nearly effortless for humans to quickly assess the perceptual similarity between two images, the underlying processes are thought to be quite complex. lpips [35] has been proposed to assess the perceptual similarity between two images. However, there still not exists a suitable metric for VQE. Compared to single-image perceptual quality enhancement which focuses on the intrinsic properties of a single image in spatial space, video perceptual quality enhancement poses an extra challenge as it involves temporal flickering though each enhancement frame in video sequences seems to be enhanced well considering image perceptual quality individually. Specifically, VQE trained with PSNR and SSIM will generate smooth videos while VQE trained with lpips will generate temporal flickering videos with more textual details.

To address the aforementioned issues, we adopt multi-objective networks with adaptive spatial-temporal fusion module to enhance regions with smooth content and regions with rich textures simultaneously. Specifically, we conduct the enhancement using a two-stage strategy. The first stage aims at obtaining relatively good intermediate results with high fidelity. At the second stage, we train two BasicVSR [3] models for different refinement purposes. One for textual details and the other for temporal smooth

regions. To eliminate temporal flickering and retain textual details, we devise a novel adaptive spatial-temporal fusion scheme. Specifically, spatial-temporal mask generation module is proposed to produce spatial-temporal masks and it is used to fuse the two network outputs. Then we use image sharpen to further enhance the videos.

The main contributions are as follows:

(1) We observe that regions with smooth contents and rich textures are degraded non-identically due to compression loss, different optimization objectives are designed for better enhancement of these regions with a two-branch architecture.

(2) An adaptive spatial-temporal fusion module is proposed to combine advantages of both network branches, meanwhile, spatial-temporal consistency is achieved to avoid flickering.

(3) BasicVSR is leverage as VQE backbones for proof-of-concept purpose and experimental results validate the effectiveness of our solution.

2. Related Work

2.1. Quality Enhancement

In the past few years, extensive works have been proposed to enhance the objective quality of compressed images [19, 8, 14, 16, 7, 10, 28, 18, 34]. Specifically, non-deep learning methods use Shape-Adaptive DCT or sparse coding to reduce the blocking effects, ringing effects and JPEG artifacts [8, 14, 16]. Deep learning methods like D^3 [28] and deep dual-domain convolutional network (DDCN) [10] utilize the prior knowledge of JPEG compression to enhance the quality of JPEG compression image.

For the compressed videos, most methods use single-frame quality enhancement approaches to tackle video enhancement [4, 26, 32]. Motivated by multi-frame super-resolution, MFQE [33] was the first to take advantage of neighboring frames for compressed video enhancement. Then MFQE 2.0 [9] is proposed which is an extended version of MFQE. Both MFQE methods adopt a temporal fusion scheme that incorporates dense optical flow for motion compensation. Since compression artifacts could seriously distort video contents and break pixel-wise correspondances between frames, the estimated optical flow tends to be inaccurate and unreliable, thereby resulting in ineffective quality enhancement. Spatio-Temporal Deformable Fusion (STDF) [5] aggregates temporal information while avoiding explicit optical flow estimation. All the above methods try to minimize the pixel-wise loss, such as MSE, PSNR and SSIM, to obtain high objective quality which disagree with human judgments. Recently, MW-GAN [25] proposed a generative adversarial network based on multi-level wavelet packet transform to recover the high-frequency details for enhancing the perceptual quality of compressed video.

2.2. Video Super Resolution

The closest work to ours is the video super-resolution (VSR). The significant difference between VSR and VQR is VSR need the final upsample layer. Several VSR approaches [2, 23, 29] use optical flow to estimate motions between frames and use spatial warping for alignment. Other methods use a more sophisticated approach of implicit alignment [24, 27, 15, 12, 13, 3]. Specifically, TDAN [24] and EDVR [27] adopt deformable convolutions to align different frames. BasicVSR [3] proposes to untangle some most essential components for VSR such as Propagation, Alignment, Aggregation, and Upsampling and find that bidirectional propagation coupled with a simple optical flow-based feature alignment suffice to outperform many state-of-the-art methods. In this work, we adopt BasicVSR as our base model which we will remove the final upsample layer.

3. Proposed Method

Given a heavily compressed video, the goal of our method is to produce high quality results with the best perceptual quality to the reference ground truth. To be specific, we conduct the enhancement using a two-stage strategy. As shown in Figure 1, The first stage aims at obtaining relatively good intermediate results with high fidelity. In this stage, a BasicVSR [3] model is trained with Charbonnier loss [17]. At the second stage, we train two BasicVSR models for different refinement purposes. One refine BasicVSR model (we term it as *EnhanceNet2*) is trained with a trade-off loss function $Charbonnier\_loss + lpips\_loss$. The other refine BasicVSR model (termed as *EnhanceNet1*) is trained with merely lpips loss. Here lpips loss [35] is a learned objective video quality measurement, which is more consistent with human perception. In this way, EnhanceNet1 is more good at recovering textures to satisfying human perception requirement but it can result in temporal flickering for smooth regions of videos, meanwhile EnhanceNet2 will produce much more smooth results, especially, temporal flickering is well eliminated. To overcome this issue, we devise a novel adaptive spatial-temporal fusion scheme. Specifically, spatial-temporal mask generation module is proposed to produce spatial-temporal masks and it is used to fuse the two network outputs. Then we use image sharpen to further enhance the videos with a Gaussian kernel size of 3.

3.1. EnhanceNet

We use BasicVSR [3] without the pixel-shuffle [22] layer as our base model of EnhanceNet in both stage-1 and stage-2.

For Coarse EnhanceNet in stage-1, we use Charbonnier loss [17] to produce a coarse result with high fidelity

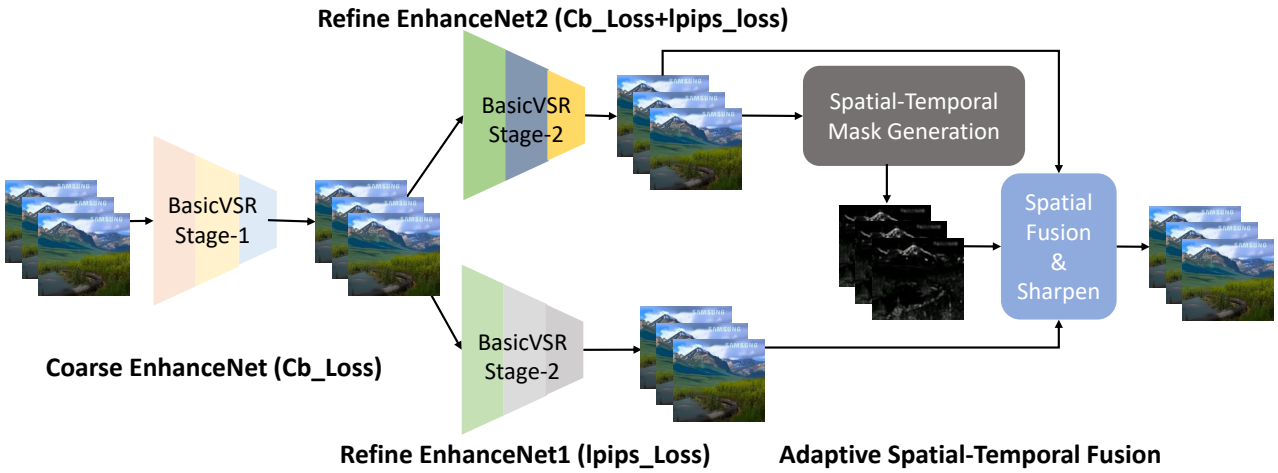

Figure 1. Workflow of our multi-objective networks with adaptive spatial-temporal fusion module. The whole framework consists of a coarse EnhanceNet as the first stage and two refine EnhanceNets with different training objectives. The outputs of two refine networks can make smooth regions or texture-rich regions perceptual-friendly to human eyes, respectively. At last, spatial-temporal masks are adaptively generated for combining their results such that advantages of the two refine networks are both leveraged.

such as PSNR and SSIM. However, the traditional metrics (L2/PSNR, SSIM) disagree with human judgments according to previous research [35]. Charbonnier loss is a differentiable variant of l_1 norm:

$$\rho(x) = \sqrt{x^2 + \epsilon^2} \tag{1}$$

We set ϵ to $1e-6$.

Based on our observation, it is hard to design unified training objectives which are perceptual-friendly for enhancing regions with smooth content and regions with rich textures simultaneously. Therefore, the trade-off between objective and perceptual quality is important to tackle this problem, which is similar to the Perception-distortion Trade-off [6].

In stage-2, we train two BasicVSR models focus on perceptual quality and the trade-off between the objective and perceptual quality, respectively. Specifically, we train Refine EnhanceNet1 with merely lpips loss [35] to produce a result focus on recovering texture details which is agree with human perception judgments. But it can also result in a drawback of temporal flickering at smooth regions of videos. To overcome this problem, we train Refine EnhanceNet2 with a trade-off loss function

$$L_{trade-off} = \alpha \times Charbonnier\_loss + \beta \times lpips\_loss \tag{2}$$

to produce a result focus on recovering smooth results, which is a trade-off between objective and perceptual quality. Note that we set $\alpha = 0.15$ and $\beta = 10000$ to make Charbonnier loss almost three times larger than lpips loss. We use VGG network in lpips loss for training and Alex network in lpips loss for validation.

3.2. ASTF

We devise a novel adaptive spatial-temporal fusion scheme (ASTF) motivated by the fact that Refine EnhanceNet1 is good at recovering texture details and Refine EnhanceNet2 is good at recovering smooth regions, which are two trade-off models we both need. Specifically, spatial-temporal mask generation module is proposed to produce spatial-temporal masks to indicate non smooth regions of videos with the results of Refine EnhanceNet2 as input. We adopt a spatial-temporal block with $3 \times 3 \times 3$ pixels.

It is used to fuse the outputs of Refine EnhanceNet1 and Refine EnhanceNet2:

$$I^t_{out} = mask_t \times I^t_{out,1} + (1 - mask_t) \times I^t_{out,2}, \tag{3}$$

where $mask_t$ is the generated mask for the t-th frame, $I^t_{out,1}$ and $I^t_{out,2}$ are the t-th output frame of EnhanceNet1 and EnhanceNet2, respectively. The mask $mask_t = f(I^{t-1}_{out,2}, I^t_{out,2}, I^{t+1}_{out,2})$ is adaptively generated from $I^{t-1}_{out,2}, I^t_{out,2}, I^{t+1}_{out,2}$ as follows:

1) convert the frame $I^t_{out,2}$ from BGR space to YUV space and choose the Y-channel (luminance component):

$$Y^t_{out,2} = BGR2YUV(I^t_{out,2})[0] \tag{4}$$

2) variance map V^t is calculated from $Y^t_{out,2}$ by:

$$V^t_{i,j} = Var(\mathcal{N}_k(Y^t_{out,2}[i,j])) \tag{5}$$

where $Var(x)$ means the variance of x. $\mathcal{N}_k$ denotes a neighbor pixel set of location (i,j) . Here $\forall V_{(p,q)} \in \mathcal{N}_k(i,j)$, we have $p = i + k, q = j + k, k^{\mathbb{R}} \in [-11, 11]$.

	Avg.	001	021	041	061	081	101	121	141	161	181
lq input	30.58/.85	26.65/.72	30.51/.89	33.30/.92	28.93/.89	37.58/.96	26.18/.79	31.65/.91	30.27/.86	30.40/.82	30.35/.79
stage-1	**31.95/.88**	**27.27/.75**	**32.11/.92**	**34.24/.94**	**30.82/.92**	**40.00/.98**	**27.03/.81**	**33.76/.94**	**32.31/.91**	**30.90/.83**	**31.09/.81**
cb+lpips	31.87/.88	27.20/.75	32.03/.91	34.15/.94	30.78/.92	39.82/.97	26.97/.81	33.68/.93	32.37/.90	30.78/.83	31.04/.81
lpips	30.80/.85	24.98/.68	31.14/.89	33.05/.92	29.98/.90	39.02/.97	25.38/.77	33.07/.92	31.66/.89	29.52/.79	30.21/.79
fuse	31.81/.88	27.13/.75	31.97/.91	34.04/.94	30.72/.92	39.77/.97	26.93/.81	33.62/.93	32.22/.90	30.74/.83	31.00/.81
fuse+sharp	26.83/.81	23.31/.67	25.78/.83	25.85/.86	23.25/.84	36.69/.95	21.94/.73	27.82/.88	26.90/.84	27.92/.77	28.80/.78

Table 1. Quantitative results of PSNR ↑ and SSIM ↑ on 10 validation videos. For both PSNR and SSIM, the higher value is better.

	Avg.	001	021	041	061	081	101	121	141	161	181
lq input	0.21/77	0.29/44	0.12/66	0.16/94	0.18/54	0.14/103	0.21/97	0.13/26	0.20/76	0.24/99	0.37/108
stage-1	0.20/94	0.29/62	0.10/61	0.16/91	0.18/81	0.14/126	0.19/83	0.10/28	0.17/86	0.27/147	0.43/181
cb+lpips	0.17/74	0.24/43	0.09/50	0.14/71	0.15/68	0.12/106	0.15/68	0.09/23	0.15/76	0.22/96	0.37/144
lpips	**0.14/58**	**0.18/27**	**0.08/47**	**0.12/65**	**0.13/55**	**0.10/92**	**0.11/56**	**0.08/18**	**0.14/61**	**0.17/46**	**0.32/116**
fuse	0.17/72	0.24/41	0.08/50	0.14/71	0.14/67	0.12/105	0.15/68	0.09/23	0.15/74	0.21/95	0.36/137
fuse+sharp	0.21/82	0.27/42	0.15/56	0.18/133	0.21/69	0.14/106	0.22/87	0.14/25	0.18/73	0.23/90	0.34/139

Table 2. Quantitative results of lpips ↓ and FID ↓ on 10 validation videos. For both lpips and FID, the lower value is better.

3) normalize the variance map in a temporal sliding window to generate the mask $mask_t$:

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

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

4. Experiments

4.1. Datasets

We use the training videos and testing videos of Quality Enhancement of Heavily Compressed Videos Challenge (Track 2 Fixed QP, Perceptual) [30] for our experiments. The training data has a total of 200 paired compressed and uncompressed videos. The testing data has 10 compressed videos. Specifically, we split the training videos into training data (190 videos) and validation data (10 videos, '001', '021', '041', '061', '081', '101', '121', '141', '161', 181). Note that, we convert raw, compressed (and enhanced) videos to RGB domain by using the official code.

4.2. Implementation Details

We use BasicVSR without pixel-shuffle layer as our base model. For both stages of training, we randomly crop 64 × 64 clips from raw and the corresponding compressed videos as training samples. Data augmentation (i.e., rotation or flip) is further used to better exploit those training samples. Learning rate is initially set to 2×10^{-4} and learning scheme is set to CosineAnnealingLR_Restart throughout training. Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$. In stage-1, we train EnhanceNet from scratch with Charbonnier loss [17]. In stage-2, we use the results of stage-1 as input of stage-2. We train EnhanceNet1 and EnhanceNet2 with lpips loss [35] and a trade-off loss function, respectively. We adopt PSNR, SSIM, lpips [35] and FID [11, 21] to evaluate quality enhancement performance on our 10 validation videos.

4.3. Quantitative Results

In Table 1, we provide PSNR and SSIM for each video of validation. It can be seen that results of stage-1 have the highest values of PSNR and SSIM. ΔPSNR of EnhanceNet2 (cb + lpips) results is -0.08 dB compared with EnhanceNet (stage-1) results. ΔPSNR of EnhanceNet1 (lpips) results is -1.15 dB compared with EnhanceNet (stage-1) results. After our spatial-temporal fusion module, ΔPSNR of spatial-temporal fusion (fuse) results is -0.14 dB compared with EnhanceNet (stage-1) results. The adaptive spatial-temporal fusion (fuse) results are just trade-off of EnhanceNet1 and EnhanceNet2. We also provide PSNR and SSIM of results after image sharpen, which is the lowest values. Though it is the lowest PSNR of our all stages, it has a better human perceptual which we will discuss later.

In Table 2, we provide lpips [35] and FID [11, 21] for each video of validation. It can be seen that EnhanceNet1 (lpips) results have the lowest value which are the best results using image perceptual metric. However, EnhanceNet1 (lpips) results have temporal flickering problem which is very import for video human perceptual judgment. We use EnhanceNet2 (cb + lpips) to reduce temporal flickering and generate results with larger values of lpips and FID than EnhanceNet1 (lpips) results. To utilize spatial-temporal information, we use our proposed ASTF to fuse EnhanceNet1 (lpips) results and EnhanceNet2 (cb + lpips) results, which can generate results with values of

Compressed frames

Enhanced frames

Figure 2. Qualitative results. The first column are compressed frames. The second column are our enhanced frames. The first two rows are validation frames. The last two rows are test frames.

lpips and FID between values of EnhanceNet1 and EnhanceNet2 (cb + lpips). We also provide lpips and FID of results after image sharpen. Similar to PSNR and SSIM, results after image sharpen are not better than results before image sharpen, but if has a better human perceptual which we will discuss later.

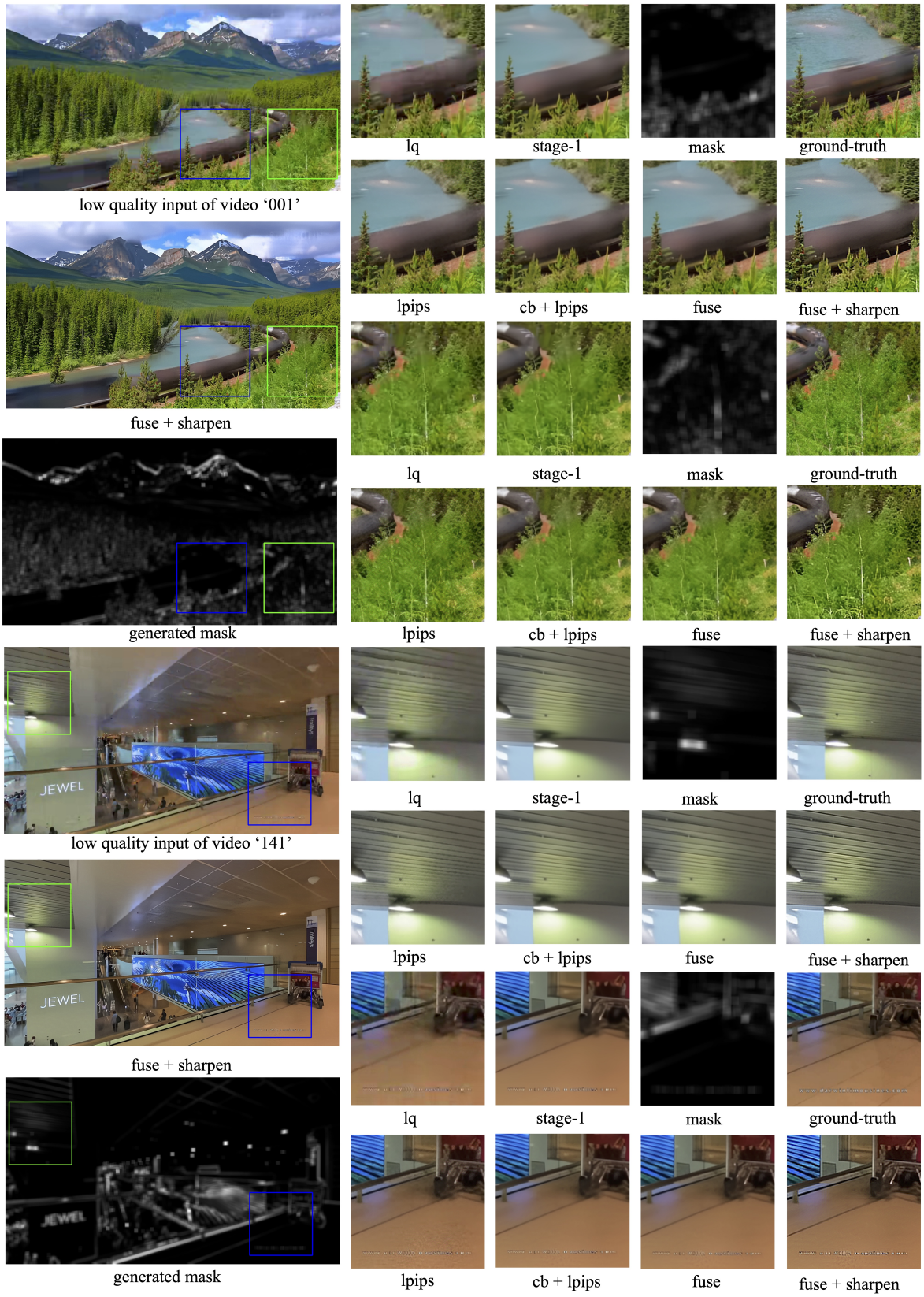

Figure 3. Qualitative results of validation video (’001’ and ’141’).

In the test phase of the challenge our proposed method (Team VUE) achieves a score of 60 ranked the fifth place. The scores are ranked according to Mean Opinion Score (MOS) values from 15 subjects. The scores range from from $s = 0$ (poorest quality) to $s = 100$ (best quality). The ground-truth videos are given to the subjects as the standard of $s = 100$, but the subjects are told to rate videos in accordance with the visual quality, instead of the similarity to the ground-truth. More details can be found in the report paper [31].

4.4. Qualitative Results

Figure 2 provides the qualitative results of validation and test frames. The first column are compressed frames. The second column are our enhanced frames. The first two rows are validation frames. The last two rows are test frames. It can be seen that compressed frames are distorted, blurred and lack of textual details. While our enhanced frames can reduce these artifacts and have more textual details.

Further more, Figure 3 provides each step results of validation frames including compressed low quality input frames, EnhanceNet results of stage-1, EnhanceNet1 results trained with lpips loss, EnhanceNet2 results trained with a trade-off loss function, spatial-temporal fusion results, sharpen results of fusion, spatial-temporal mask and the ground-truth. It can be seen that EnhanceNet of stage-1 can reduce most distortion but still exists over-blurred problem. Results of EnhanceNet1 (lpips) have more textual details but it will cause some artifacts. Results of EnhanceNet2 (cb + lpips) are trade-off smooth and textual details compared with EnhanceNet and EnhanceNet1, respectively. The spatial-temporal mask generated by using results of EnhanceNet2 (cb + lpips) indicate high-frequency regions with high pixel values. The spatial-temporal fusion (fuse) results have both advantage of EnhanceNet1 (lpips) and EnhanceNet2 (cb + lpips), which are smooth in low-frequency regions and have more textual details in high-frequency regions. The results (fuse + sharpen) is processed by image sharpen to further enhance the human perception. Note that results (fuse) seems more similar to the ground-truth than results (fuse + sharpen). However, not consider the ground-truth results (fuse + sharpen) have better human perception effects. Therefore we choose results (fuse + sharpen) as our final submit results to track-2 of the NTIRE 2021 Quality enhancement of heavily compressed videos Challenge.

4.5. Analysis and Discussions

Compared with quantitative results and qualitative results we can see that results with highest values of PSNR and SSIM are not the best human perceptual results. We also conduct user study to evaluate the temporal performance. For both validation and test videos, our results (fuse) and results (fuse + sharpen) have a better performance to reduce temporal flickering in smooth regions than results of EnhanceNet1 (lpips). Charbonnier loss [17] is helpful for generating smooth regions and lpips loss [35] is helpful for generating textual details. We can see that either model trained with only one loss function (Charbonnier loss or lpips loss) can not achieve a well enhanced video judged by human perception. Temporal flickering and lacking of textual details need to be consider simultaneously. Our proposed adaptive spatial-temporal fusion (ASTF) utilize both advantages of EnhanceNet1 (lpips) and EnhanceNet2 (cb + lpips) and have a better human perceptual video enhancement performance.

5. Conclusion

We have introduced our approach of the Track 2 Fixed QP perceptual in the NTIRE 2021 Quality enhancement of heavily compressed videos Challenge. To handle the challenging we propose ASTF to adaptive fuse the enhancement results from networks trained with two different optimization objectives. BasicVSR enhancement models with different loss functions are trained to recover smooth and textual details, respectively. The fusion operation can combine advantages of both networks. Our method is a general approach and can be used for any other video enhancement backbones. Experiments show that our spatial-temporal fusion module can retain smooth and high-frequency details simultaneously, which results in a better human perceptual effect.

References

[1] Christos George Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik. Study of temporal effects on subjective video quality of experience. *TIP*, 26(11):5217–5231, 2017. 1

[2] 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 *CVPR*, pages 4778–4787, 2017. 2

[3] 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. *arXiv preprint arXiv:2012.02181*, 2020. 1, 2

[4] 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. 2

[5] Jianing Deng, Li Wang, Shiliang Pu, and Cheng Zhuo. Spatio-temporal deformable convolution for compressed video quality enhancement. In *AAAI*, volume 34, pages 10696–10703, 2020. 1, 2

[6] Xin Deng, Ren Yang, Mai Xu, and Pier Luigi Dragotti. Wavelet domain style transfer for an effective perception-

distortion tradeoff in single image super-resolution. In *ICCV*, pages 3076–3085, 2019. 3

[7] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *ICCV*, pages 576–584, 2015. 2

[8] Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. *TIP*, 16(5):1395–1411, 2007. 2

[9] 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. *PAMI*, 2019. 1, 2

[10] Jun Guo and Hongyang Chao. Building dual-domain representations for compression artifacts reduction. In *ECCV*, pages 628–644. Springer, 2016. 2

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017. 4

[12] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In *ECCV*, pages 645–660. Springer, 2020. 2

[13] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *CVPR*, pages 8008–8017, 2020. 2

[14] Jeremy Jancsary, Sebastian Nowozin, and Carsten Rother. Loss-specific training of non-parametric image restoration models: A new state of the art. In *ECCV*, pages 112–125. Springer, 2012. 2

[15] 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 *CVPR*, pages 3224–3232, 2018. 2

[16] Cheolkon Jung, Licheng Jiao, Hongtao Qi, and Tian Sun. Image deblocking via sparse representation. *Signal Processing: Image Communication*, 27(6):663–677, 2012. 2

[17] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, pages 624–632, 2017. 2, 4, 7

[18] Ke Li, Bahetiyaer Bare, and Bo Yan. An efficient deep convolutional neural networks model for compressed image deblocking. In *ICME*, pages 1320–1325. IEEE, 2017. 2

[19] AW-C Liew and Hong Yan. Blocking artifacts suppression in block-coded images using overcomplete wavelet representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(4):450–461, 2004. 2

[20] Jens-Rainer Ohm, Gary J Sullivan, Heiko Schwarz, Thiow Keng Tan, and Thomas Wiegand. Comparison of the coding efficiency of video coding standards—including high efficiency video coding (hevc). *IEEE Transactions on circuits and systems for video technology*, 22(12):1669–1684, 2012. 1

[21] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.1.1. 4

[22] Wenzhe Shi, Jose Caballero, Ferenc Huszár, 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 *CVPR*, pages 1874–1883, 2016. 2

[23] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, pages 4472–4480, 2017. 2

[24] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, pages 3360–3369, 2020. 2

[25] 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 *ECCV*, pages 405–421. Springer, 2020. 1, 2

[26] 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. 2

[27] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, pages 0–0, 2019. 2

[28] Zhangyang Wang, Ding Liu, Shiyu Chang, Qing Ling, Yingzhen Yang, and Thomas S Huang. D3: Deep dual-domain based fast restoration of jpeg-compressed images. In *CVPR*, pages 2764–2772, 2016. 2

[29] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 127(8):1106–1125, 2019. 2

[30] Ren Yang and Radu Timofte. NTIRE 2021 challenge on quality enhancement of compressed video: Dataset and study. In *CVPRW*, 2021. 4

[31] Ren Yang, Radu Timofte, et al. NTIRE 2021 challenge on quality enhancement of compressed video: Methods and results. In *CVPRW*, 2021. 7

[32] Ren Yang, Mai Xu, and Zulin Wang. Decoder-side hevc quality enhancement with scalable convolutional neural network. In *ICME*, pages 817–822. IEEE, 2017. 2

[33] Ren Yang, Mai Xu, Zulin Wang, and Tianyi Li. Multi-frame quality enhancement for compressed video. In *CVPR*, pages 6664–6673, 2018. 1, 2

[34] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 26(7):3142–3155, 2017. 2

[35] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 1, 2, 3, 4, 7