

Progressive Training of A Two-Stage Framework for Video Restoration

Meisong Zheng^{1*}, Qunliang Xing^{1*}, Minglang Qiao^{1*}, Mai Xu[†], Lai Jiang, Huaida Liu¹ and Ying Chen^{1†}

¹Alibaba Group

{zhengmeisong.zms, xingqunliang.xql, qiaominglang.qml}@alibaba-inc.com

xumai@icloud.com, jianglai.china@gmail.com, {liuhuaida.lhd, yingchen}@alibaba-inc.com

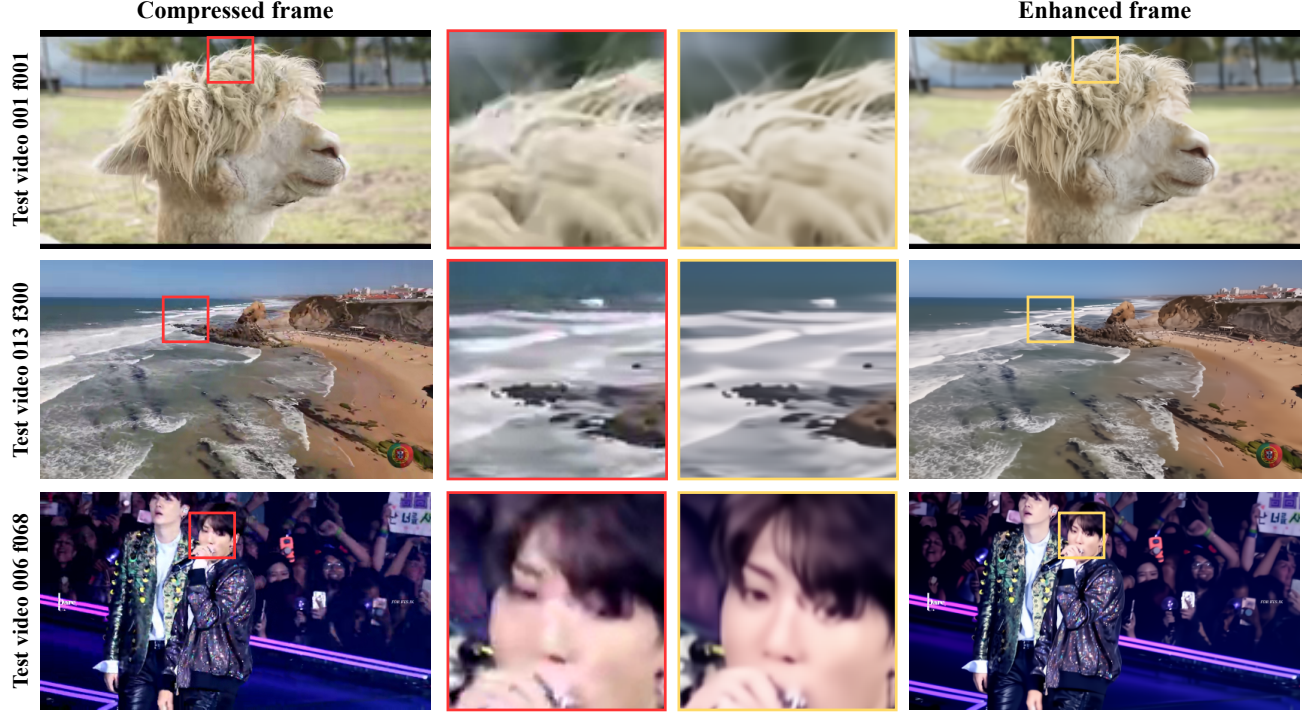


Figure 1. Subjective performance of our proposed method on the test set of the NTIRE 2022 challenge.

Abstract

As a widely studied task, video restoration aims to enhance the quality of the videos with multiple potential degradations, such as noises, blurs and compression artifacts. Among video restorations, compressed video quality enhancement and video super-resolution are two of the main tasks with significant values in practical scenarios. Recently, recurrent neural networks and transformers attract increasing research interests in this field, due to their impressive capability in sequence-to-sequence modeling. However, the training of these models is not only costly but also relatively hard to converge, with gradient exploding and vanishing problems. To cope with these problems, we proposed a two-stage framework including a multi-frame recurrent network and a single-frame transformer. Besides,

multiple training strategies, such as transfer learning and progressive training, are developed to shorten the training time and improve the model performance. Benefiting from the above technical contributions, our solution wins two champions and a runner-up in the NTIRE 2022 super-resolution and quality enhancement of compressed video challenges. Code is available at <https://github.com/ryanxingql/winner-ntire22-vqe>.

1. Introduction

The recent decades have witnessed an explosive growth of video data over the internet. Meanwhile, the resolution of the videos becomes higher and higher to satisfy the in-

*These authors contributed equally to this work.

†Corresponding authors.

creasing demand for the quality of experience (QoE). However, due to the limited bandwidth, the videos are commonly down-sampled and compressed, which causes inevitably degradation on video quality. Therefore, it draws a great attention in the computer vision community for video restoration tasks, such as video super-resolution, de-artifacts of compressed video.

Video restoration is challenging because it requires aggregating information from multiple highly related but misaligned low-quality frames in video sequences. Most existing methods of video restoration consider it as a spatial-temporal sequence prediction problem, and can be mainly divided into two categories: sliding window methods [9, 15, 32, 35, 42] and recurrent-based methods [4, 5, 19]. For instance, BasicVSR++ [5] proposes a second-order grid propagation network to better mining the spatial-temporal information. It demonstrates the great effectiveness of the recurrent framework and wins the NTIRE 2021 quality enhancement of heavily compressed video challenge. However, the recurrent framework processes the video frames sequentially, which limits the efficiency of the recurrent-based methods. Recent works [2, 23] try to enhance video frames in parallel, on the top of transformer architecture. However, both recurrent network and transformer have square computational complexity with respect to sequence length and image size, resulting in $O(n^4)$ computational complexity. Subject to the huge memory consumption, these networks can only be fed by the clipped sequence with no more than 16 frames, even on a NVIDIA A100 GPU. This degrades PSNR performance compared to BasicVSR++ [5] on the REDS dataset [27]. Besides the large consumption of GPU memory, the models with larger structures, such as Transformer, are also hard to be tuned. That is, we sometimes are unable to finely adjust the key hyper-parameters, like batch size and learning rate, which are essential on stabilizing the training process. Moreover, the “large” models also prone to suffer from the problems of over-fitting and performance fluctuation across the restored frames.

To address the above problems, we propose a two-stage framework combining a multi-frame recurrent-based network and single-frame transformer-based network. Specifically, the first stage is developed to coarsely restore the video frames and alleviate the quality fluctuation across the frames. Given the restored frames from the first stage, the second stage further effectively removes the severe artifacts frame by frame. Specifically, the first stage model is an improved BasicVSR++ [5], and in the second stage we adopt SwinIR [24] as the backbone model. We train these two models separately to save memory resources and further improve the accuracy. Besides, multiple strategies of transfer learning and progressive training are conducted in both two stages, to not only accelerate the convergence but also improve final restoration performance. In summary, the con-

tributions of this paper are as follows:

- We propose a two-stage framework to simultaneously remove compression artifacts and mitigate the quality fluctuation in compressed videos.
- We introduce a progressive training scheme to stabilize training and improve finally performance.
- We introduce a transfer learning strategy with pre-trained models to shorten training time.
- Our proposed method achieves a good trade-off between the enhancement performance and model complexity, and wins the NTIRE 2022 challenge of super-resolution and quality enhancement of compressed video [40].

2. Related Work

2.1. Video Restoration

As one of the main tracks of video restoration, compressed video quality enhancement on has been widely studied [11, 15, 35, 41, 42] in the past years. Among them, most of the existing methods are based on the single-frame quality enhancement [11, 35, 41]. Observing that the frame quality remarkably fluctuates after compression, MFQE [42] and its extended version MFQE 2.0 [15] take advantage of neighboring high-quality frames. They adopt a temporal fusion scheme that incorporates dense optical flow for motion compensation. Similarly, STDF [9] aggregates temporal information while avoiding explicit optical flow estimation.

Video Super Resolution. In addition to the compressed video quality enhancement, video super resolution (VSR) aims to restore the videos by improving their resolution. Different from the single image super resolution (SISR), VSR utilizes neighboring frames to reconstruct the high-resolution sequence. The existing VSR methods can be divided into two categories: window-based [22, 32, 36, 43] and recurrent methods [4, 5, 18, 19]. Specifically, EDVR [32] adopts deformable convolutions [8, 46] to align neighbouring frames. Similar to EDVR, D3DNet [43] uses deformable 3D convolution network to fully exploit the spatio-temporal information for video SR. Besides, BasicVSR [4] proposes to untangle the basic components for VSR such as propagation, alignment, aggregation and up-sampling. On the top of BasicVSR, BasicVSR++ [5] further improves performance with extensive bi-directional propagation strategy and flow-guided deformable alignment. In this work, we adopt BasicVSR++ as our backbone model in the first stage.

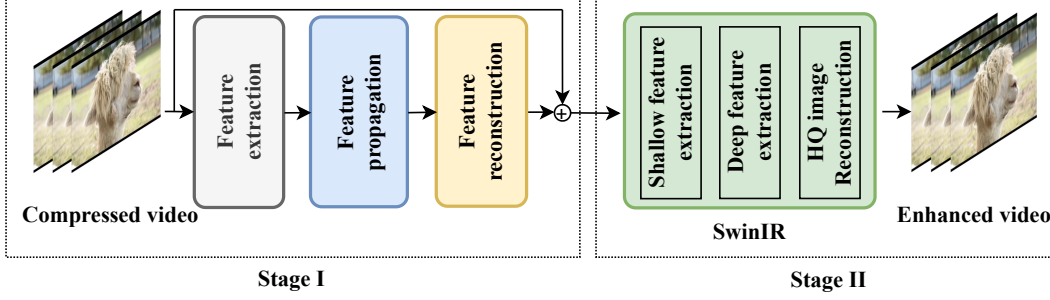


Figure 2. Two-stage framework of our method.

2.2. Vision Transformer

Recently, sourced from the area of natural language processing (NLP) [10, 21, 25], Transformers have shown the outstanding performance and outperforms the state-of-the-art models in many vision tasks, including image classification, object detection, semantic segmentation, human pose estimation and video classification [1, 3, 12, 16, 26, 26, 34, 34, 44]. Specifically, Swin transformer [26] proposes a hierarchical transformer structure with shifted windows mechanism, which integrates the advantages of build-in inductive biases of CNN and long-range self-attention of transformers.

There also exists some attempts to apply transformers in low-level vision tasks [6, 7, 20, 24, 33, 37, 45]. For instance, SwinIR [24] proposes an image restoration model based on Swin transformer, which not only handles local context but also efficiently captures long-range dependencies. Uformer [33] proposes a general U-shaped transformer-based structure, which shows strong performance on real de-noising tasks.

Transformers have also been introduced for video restoration [2, 13, 23]. VSRT [2] utilizes the parallel computing ability of transformer to align the features between neighboring frames in parallel. VRT [23] introduces a temporal mutual self attention module to better mining spatial-temporal information. Unfortunately, these approaches can not be trained with longer video clips as they require large memory of GPU. In this work, we adopt SwinIR as our backbone model in the second stage.

3. Method

3.1. Proposed Two-stage Framework

We first introduce our two-stage framework for video restoration, as shown in Fig. 2. In stage I, the network is developed on the top of BasicVSR++ [5]. Based on this, we replace the second-order flows in BasicVSR++ by PQF flows [15, 42]. Besides, we deepen the reconstruction module of BasicVSR++ from 5 residual blocks to 55 blocks.

In stage II, we further improve the quality of the enhanced consecutive frames by a state-of-the-art image restoration network, *i.e.*, SwinIR [24]. This stage helps remove severe artifacts and further improve the quality upon the previous stage. Finally, the networks of stage I and II are cascaded for producing the final results. In summary, we first feed the compressed video with N compressed frames $\{F_t\}_{t=1}^N$ into the stage I model. Then, we obtain the enhanced video frames $\{\tilde{F}_t\}_{t=1}^N$ by stage I. Next, we feed $\{\tilde{F}_t\}_{t=1}^N$ into the stage II model frame by frame. Finally, we get the enhanced video frames $\{\hat{F}_t\}_{t=1}^N$, which are sequentially combined into the final enhanced video.

3.2. First Stage and Progressive Training

Our stage I model consists of three developed modules: feature extraction, propagation and image reconstruction. Given an input video, two strided convolution and five residual blocks are first applied to extract spatial features from the input frames. At the same time, all input frames are down-sampled by the factor of 4 with an bicubic filter, and then applied to SpyNet [29] to calculate the forward and backward flows. Next, as shown in Fig. 3, for enhancing the t -th frame, the features of neighboring $(t-1)$ -th and $(t+1)$ -th frames as well as the features of previous and subsequent PQFs are propagated to the spatial feature of the t -th frame. For the propagation of each frame, the frame is warped by its estimated flow. Finally, we use 55 residual blocks to decode the propagated features, and reconstruct the video. To be more specific, in stage I model, we use pixel shuffling [30] to restore the resolution of decoded features. Besides, residual learning [17] is also conducted for generating the final enhanced image, by reducing the training complexity of the model.

As introduced, our reconstruction module contains 55 residual blocks, which is rather “heavy” for training. Thus, a progressive training [14, 28] strategy is conducted for our stage I model. Specially, we lighten the reconstruction module by using its first 5, 15, 25, 35, 45 and 55 residual blocks for reconstruction, respectively. Specifically, let R_1, R_2, \dots, R_5 and R_6 denote the 1-5, 6-15, 16-25, 26-35, 36-45, 46-55

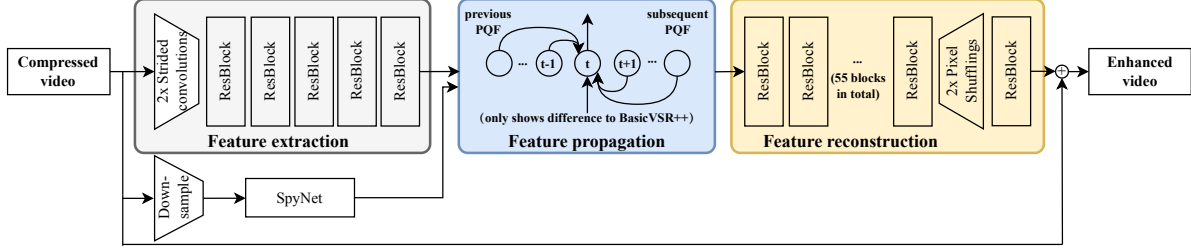


Figure 3. Structure of stage I model (for Track 1).

residual blocks; E and P refer to the modules of feature extraction and feature propagation; S and R are the two pixel shuffling layers and residual block at the end of our stage I model. Given the input frame I_{in} , the restored frame I_{out} can be obtained by the progressively training as follows

$$I_{out} = R(S(R_1(P(E(I_{in})))))) \quad (1)$$

$$I_{out} = R(S(R_2(R_1(P(E(I_{in})))))) \quad (2)$$

$$I_{out} = R(S(R_3(R_2(R_1(P(E(I_{in})))))) \quad (3)$$

$$I_{out} = R(S(R_4(R_3(R_2(R_1(P(E(I_{in})))))) \quad (4)$$

$$I_{out} = R(S(R_5(R_4(R_3(R_2(R_1(P(E(I_{in})))))) \quad (5)$$

$$I_{out} = R(S(R_6(R_5(R_4(R_3(R_2(R_1(P(E(I_{in})))))) \quad (6)$$

For the first training, we load the parameters of E , P , S and R from the open-sourced model of BasicVSR++. For the k -th training ($2 \leq k \leq 6$), we load the parameters of E , P , S , R and $\{R_i\}_{i=1}^{k-1}$ from the $(k-1)$ -th converged model. Note that the temporal information is embedded in the propagation module as illustrated in Fig. 2, which is simplified in the above equations.

3.3. Second Stage and Transfer Learning

Although a single BasicVSR++ could achieve state-of-the-art performance for the compressed videos restoration, the restored results are not satisfactory in the cases with severely distorted scenes. Thus, we develop a stage II model to further refine the enhanced video frames by stage I model, similar to the two-stage restoration strategy in [32]. However, different from [32], we empirically find that simply cascading a second BasicVSR++ on stage I can only bring slight improvement. Instead of cascading a video restoration model, we employ a single-image restoration model in stage II to further improve the quality of the enhanced frames.

Specifically, SwinIR [24] model is utilized in stage II to further enhance the outputs of stage I, which is proven to be still effective for enhancement of compressed video, in addition to the restoration of single image. Besides, due to the fact that transformer requires training in large-scale datasets, transfer learning is applied during the training of

SwinIR. More specifically, the SwinIR model is initialized by pre-trained parameters from [24], which is trained for RGB image denoising. The effectiveness of stage II is illustrated in Table 2.

4. Experiments

4.1. Datasets

We use two datasets for training our models in both two stages. First, we adopt the LDV dataset [39], which is released officially by the NTIRE 2022 challenge. It contains 240 qHD sequences belonging to 10 categories of scenes, including animal, city, closeup, fashion, human, indoor, park, scenery, sports and vehicle. Besides, we build a large-scale dataset with 870 4K sequences acquired from YouTube. Specially, for each above category, 87 sequences are collected. These sequences are with high-quality and without visible artifacts. Then, we follow the data processing procedure in NTIRE 2021 report [38], and convert our 4K sequences to qHD sequences. As a preprocessing, we further remove repeated frames in the compressed sequences and the corresponding frames in raw sequences.

To validate the performance of our proposed method, we select one sequence from each scenes to construct a offline validation set. These 10 sequences are 109, 030, 125, 056, 189, 124, 119, 102, 106 and 158 from LDV dataset. In general, we use 1100 sequences for training, and 10 sequences for validation.

4.2. Implementation Detail

For stage I, we first fine-tune the official pre-trained BasicVSR++ model for 300K iterations with Charbonnier loss. Adam optimizer is adopted with a initial learning rate of 2×10^{-5} . We also adopt the Cosine Restart scheduler with the period of 300K iterations. The learning rate is linearly increased for the first 10% iterations. Besides, we progressively train and converge our model by increasing the number of residual reconstruction blocks from 5 to 55. Then, we fine-tune our model with L2 loss for 100K iterations. All experiments are conducted with four NVIDIA V100 GPUs.

For stage II, we first fine-tune the image restoration model of SwinIR via the default Charbonnier loss, which is

Table 1. Quantitative results of average PSNR \uparrow on stage I model in Track 1. Note that LDV refers to the 230 official training data, and EX refers to the 870 training data collected from YouTube. MSE is the Mean Squared Error training loss, and RMD indicates that we removed duplicated frames at test time.

Model	Params	Settings	Our Offline	Official(10th frames)
BasicVSR++_c128n25 [5]	44.08M	LDV	32.5930	31.8269
StageI_c128n25	44.08M	LDV	32.6130	31.8611
StageI_c128n25	44.08M	LDV+EX	32.6957	32.0252
StageI_c128n25_rec2	47.51M	LDV+EX	32.7484	32.0587
StageI_c128n25_rec3	50.61M	LDV+EX	32.7850	32.0826
StageI_c128n25_rec4	53.71M	LDV+EX	32.7945	32.0933
StageI_c128n25_rec5	56.80M	LDV+EX	32.8054	32.1025
StageI_c128n25_rec6	59.90M	LDV+EX	32.8240	32.1067
StageI_c128n25_rec6	59.90M	LDV+cleaned_EX	32.8672	32.1934
StageI_c128n25_rec6	59.90M	LDV+cleaned_EX, MSE	32.8968	32.2224
StageI_c128n25_rec6	59.90M	LDV, MSE	32.9055	32.2323
StageI_c128n25_rec6	59.90M	LDV, MSE, RMD	32.9193	32.2395

Table 2. Quantitative results of PSNR \uparrow on our 10 offline validation videos in Track 1. Note that LDV refers to the 230 official training data, and EX refers to the 870 training data collected from YouTube. MSE is the Mean Squared Error training loss; RMD indicates that we removed duplicated frames at test time, and TTA indicates the employing of self-ensemble. TTA.I and TTA.II indicates applying self-ensemble in stage I and II, respectively.

Stage	Model	Settings	30	56	102	106	109	119	124	125	158	189	Avg. Offline	Avg. Official
Baseline	LQ Input	-	29.39	32.89	27.84	34.09	30.04	28.90	30.14	34.19	31.9	26.79	30.6170	30.1768
	BasicVSR++_c128n25 [5]	LDV	31.65	35.13	29.08	35.65	31.00	30.30	32.84	37.40	34.75	28.14	32.5930	31.8269
	BasicVSR++_c128n25 [5]	LDV, TTA	31.90	35.29	29.20	35.70	31.30	30.42	33.19	37.71	35.00	28.30	32.8019	32.1188
I	StageI_c128n25	LDV	31.64	35.01	29.07	25.64	31.17	30.29	32.88	37.47	34.82	28.13	32.6130	31.8611
	StageI_c128n25	LDV+EX	31.60	35.18	29.26	35.76	32.25	30.32	32.89	37.54	34.99	28.16	32.6957	32.0252
	StageI_c128n25_rec6	LDV+EX	31.74	35.31	29.3	35.79	31.36	30.39	33.20	37.81	35.07	28.27	32.8240	32.1067
	StageI_c128n25_rec6	LDV+cleaned_EX	31.81	35.29	29.29	35.74	31.4	30.42	33.34	37.94	35.11	28.33	32.8672	32.1934
	StageI_c128n25_rec6	LDV, MSE, RMD	31.84	35.35	29.39	35.78	31.52	30.45	33.37	37.97	35.14	28.36	32.9193	32.2395
	StageI_c128n25_rec6	LDV, MSE, RMD, TTA	32.09	35.49	29.48	35.84	31.62	30.54	33.60	38.19	35.38	28.49	33.0721	32.4334
II	SwinIR [24]	LDV+EX	32.05	35.43	29.40	35.80	31.57	30.51	33.61	38.08	35.23	28.47	33.0148	32.3687
	SwinIR [24]	LDV+cleaned_EX	32.06	35.42	29.39	35.80	31.57	30.51	33.63	38.13	35.25	28.46	33.0227	32.3757
	SwinIR [24]	LDV+cleaned_EX, MSE	32.07	35.43	29.40	35.81	31.58	30.53	33.65	38.13	35.26	28.48	33.0327	32.3873
	SwinIR [24]	LDV+cleaned_EX, MSE, TTA.I	32.25	35.54	29.46	35.86	31.65	30.59	33.81	38.28	35.42	28.58	33.1451	32.5425
	SwinIR [24]	LDV+cleaned_EX, MSE, TTA.II	32.27	35.55	29.47	35.86	31.66	30.60	33.83	38.32	35.45	28.59	33.1619	32.5525

initialized by the pre-trained parameters on the task of image denoising. Then we jointly fine-tune the overall model with a small learning rate of 1×10^{-6} using L2 loss function, over our established dataset and NTIRE training dataset. It is noteworthy that we only sample one of every eight frames from each video for training, instead of sampling all video frames.

4.3. Quantitative Results

In the experiments, we adopt the peak signal-to-noise ratio (PSNR) to evaluate the video restoration performance. We report our performance of Track 1 on two parts: (1) 10 sequences of our offline validation set and (2) 15 sequences of the official online validation set.

As shown in Table 1, our stage I model achieves 0.326 dB PSNR improvement on the offline validation set. Specially, by training with extra data, we improve our performance by 0.083 dB. Besides, the performance can be further improved by 0.043 dB by removing some poor-quality sequences. By progressively training our model, 0.128 dB

PSNR improvement is achieved with 15.82M more parameters. Fine-tuning with MSE Loss and removing duplicated frames bring us 0.028 dB and 0.014 dB improvements, respectively.

We also provide the results on our offline validation set in Table 2. As can be seen, the employing of stage II model brings 0.11 dB performance gain in terms of PSNR upon the results of stage I. Furthermore, after applying self-ensemble in stage II, the performance (*i.e.*, PSNR) boost by 0.13 dB and achieves 33.16 dB in the offline validation set, and achieves a total improvement of 0.36 dB compared with the baseline BasicVSR++ model. This indicates that the utilizing of stage II helps achieve superior performance, and verifies the effectiveness of our proposed two-stage strategy in restoration of compressed videos.

4.4. Qualitative Results

We present our results on the official test set of NTIRE 2022 challenge in Fig. 1 and those on our validation set in Fig. 4. It is observed that our proposed method restore rich

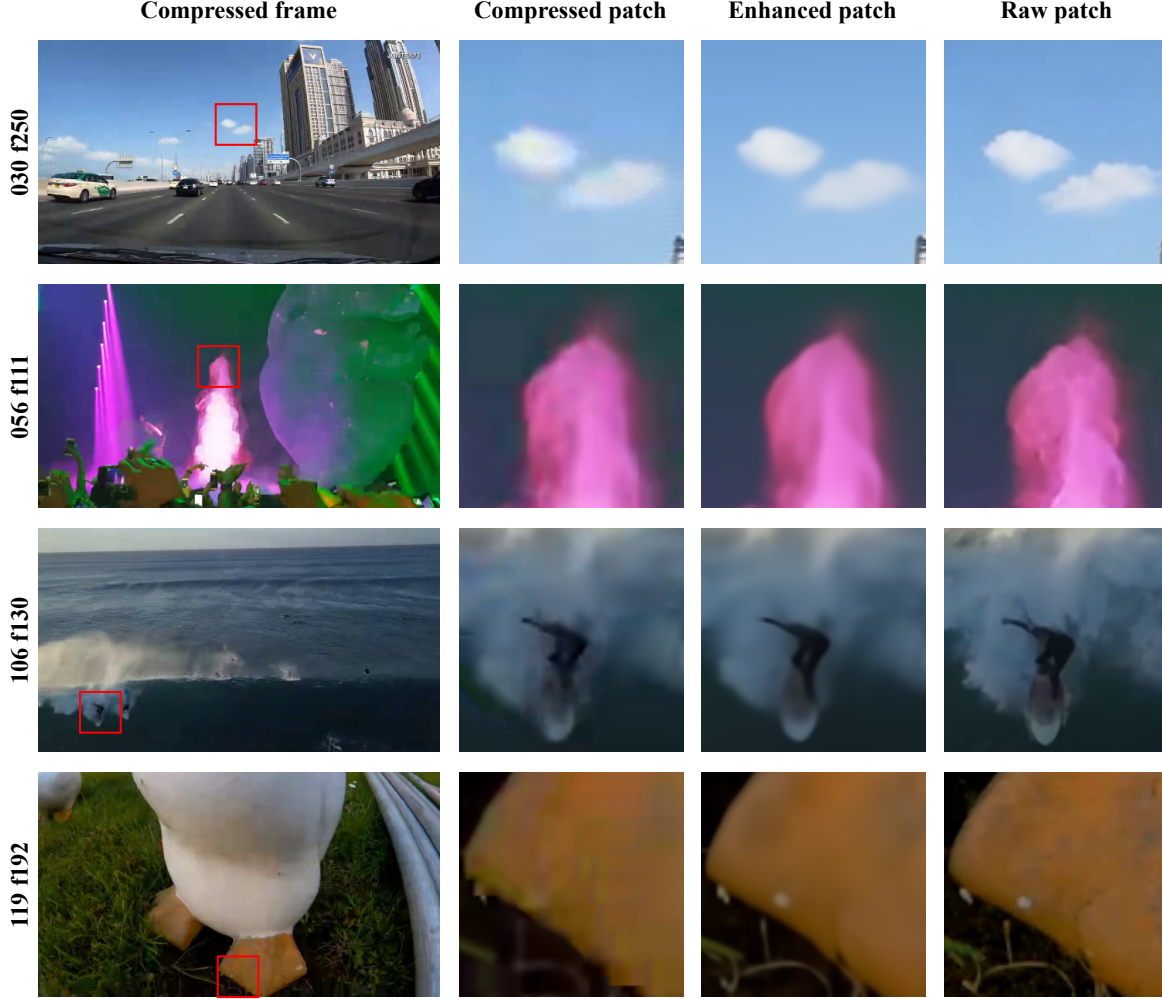


Figure 4. Subjective performance of our proposed method on our validation set.

Table 3. Ablation results on the effectiveness of transfer learning for stage II in terms of PSNR \uparrow and training time, respectively. The results are evaluated on our 10 offline validation videos.

Model	30	56	102	106	109	119	124	125	158	189	Avg.	Training time
SwinIR wo transfer	31.98	35.39	29.37	35.79	31.56	30.46	33.57	38.09	35.21	28.43	32.9822	66h
SwinIR wt transfer	32.06	35.42	29.39	35.80	31.57	30.51	33.63	38.13	35.25	28.46	33.0227	29h

details in the blurred regions of video frames. Besides, the output of our solution contains less motion blur, compared with the compressed video. The edge of objects are also much clearer.

4.5. Ablation Study

Table 3 shows the ablation results on transfer learning of stage II. As can be observed, with the application of transfer learning, the PSNR of stage II achieves an improvement of 0.04 dB compared with the model training from scratch. This indicates the effectiveness of transferring the

knowledge of image denoising to the compressed video enhancement. Besides, the training time of SwinIR is significantly reduced after employing transfer learning, which drops from 66 hours to 29 hours. This verifies the advantages of transfer learning on stage II.

5. NTIRE 2022 Challenge

We participate in all three tracks in the NTIRE 2022 super-resolution and quality enhancement of compressed video challenge. Quantitative results are presented in Ta-

Table 4. Our results of averaged PSNR \uparrow of all three tracks in the challenge. Note that for the evaluation on the validation set, we provide results of stage I/II.

Track	validation (10th frames)	test set (10th frames)	test set (all frames)
1 (Winner)	32.43/32.55	31.92	32.07
2 (Winner)	28.15/28.17	27.48	27.55
3 (Runner-up)	25.08/25.09	24.19	24.22

ble 4. In the competition, the self-ensemble [31, 32] is used in all the three tracks, while the model-ensemble is used only in Track 3. Specifically, for Track 1&2, we flip and rotate the input image to generate eight augmented inputs for each sample, and then merge the eight predicts as the input of the stage II model. For Track 3, in addition to the 8 augmentation in Track 1&2, we further conduct the model ensemble in the first stage. As a result, 16 predicts (two models with eight rotations of each) is used as the input of the stage II model.

6. Conclusion

In this paper, we proposed a two-stage framework to simultaneously remove compression artifacts and mitigate the quality fluctuation in compressed videos. Specifically, we introduced the progressive training and transfer learning strategies to stabilize the training process, shorten the training time, and improve final performance of video enhancement. Our method achieved a good trade-off between the enhancement performance and model complexity, and won two champions and one runner-up in the super-resolution and quality enhancement of compressed video challenge of NTIRE 2022.

7. Acknowledgement

This work was supported by Alibaba Group through Alibaba Research Intern Program.

References

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *International Conference on Computer Vision (ICCV)*, 2021. 3
- [2] Jiezhong Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv*, 2021. 2, 3
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020. 3
- [4] Kelvin C. K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. *CoRR*, abs/2012.02181, 2020. 2
- [5] Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. *CoRR*, abs/2104.13371, 2021. 2, 3, 5
- [6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer, 2021. 3
- [7] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. *CoRR*, abs/2104.14730, 2021. 3
- [8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *CoRR*, abs/1703.06211, 2017. 2
- [9] 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, Apr. 2020. 2
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 3
- [11] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. *CoRR*, abs/1504.06993, 2015. 2
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 3
- [13] Dario Fuoli, Martin Danelljan, Radu Timofte, and Luc Van Gool. Fast online video super-resolution with deformable attention pyramid, 2022. 3
- [14] Dario Fuoli, Zhiwu Huang, Martin Danelljan, Radu Timofte, Hua Wang, Longcun Jin, Dewei Su, Jing Liu, Jaehoon Lee, Michal Kudelski, Lukasz Bala, Dmitry Hrybov, Marcin Mozejko, Muchen Li, Siyao Li, Bo Pang, Cewu Lu, Chao Li, Dongliang He, Fu Li, and Shilei Wen. Ntire 2020 challenge on video quality mapping: Methods and results, 2020. 3
- [15] Zhenyu Guan, Qunliang Xing, Mai Xu, Ren Yang, Tie Liu, and Zulin Wang. Mfqc 2.0: A new approach for multi-frame quality enhancement on compressed video. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):949–963, 2019. 2, 3
- [16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. 3
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [18] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. *CoRR*, abs/2008.00455, 2020. 2

- [19] Lielin Jiang, Na Wang, Qingqing Dang, Rui Liu, and Baohua Lai. PP-MSVSR: multi-stage video super-resolution. *CoRR*, abs/2112.02828, 2021. 2
- [20] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34, 2021. 3
- [21] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019. 3
- [22] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. *CoRR*, abs/2007.11803, 2020. 2
- [23] Jingyun Liang, Jiezhong Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 2, 3
- [24] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE International Conference on Computer Vision Workshops*, 2021. 2, 3, 4, 5
- [25] Yinhan Liu, Mylène Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 3
- [26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. 3
- [27] 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 *CVPR Workshops*, June 2019. 2
- [28] Yali Peng, Yue Cao, Shigang Liu, Jian Yang, and Wangmeng Zuo. Progressive training of multi-level wavelet residual networks for image denoising, 2020. 3
- [29] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 3
- [30] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3
- [31] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. *CoRR*, abs/1511.02228, 2015. 7
- [32] Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: video restoration with enhanced deformable convolutional networks. *CoRR*, abs/1905.02716, 2019. 2, 4, 7
- [33] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *CoRR*, abs/2106.03106, 2021. 3
- [34] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention, 2022. 3
- [35] Qunliang Xing, Mai Xu, Tianyi Li, and Zhenyu Guan. Early exit or not: Resource-efficient blind quality enhancement for compressed images. In *European Conference on Computer Vision*, pages 275–292. Springer, 2020. 2
- [36] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T. Freeman. Video enhancement with task-oriented flow. *CoRR*, abs/1711.09078, 2017. 2
- [37] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, June 2020. 3
- [38] Ren Yang. Ntire 2021 challenge on quality enhancement of compressed video: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2021. 4
- [39] Ren Yang and Radu Timofte. Ntire 2021 challenge on quality enhancement of compressed video: Dataset and study, 2021. 4
- [40] Ren Yang, Radu Timofte, et al. NTIRE 2022 challenge on super-resolution and quality enhancement of compressed video: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 2
- [41] Ren Yang, Mai Xu, and Zulin Wang. Decoder-side hevc quality enhancement with scalable convolutional neural network. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 817–822, 2017. 2
- [42] Ren Yang, Mai Xu, Zulin Wang, and Tianyi Li. Multi-frame quality enhancement for compressed video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6664–6673, 2018. 2, 3
- [43] Xinyi Ying, Longguang Wang, Yingqian Wang, Weidong Sheng, Wei An, and Yulan Guo. Deformable 3d convolution for video super-resolution. *IEEE Signal Processing Letters*, 27:1500–1504, 2020. 2
- [44] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. *CoRR*, abs/2110.09408, 2021. 3
- [45] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *CoRR*, abs/2111.09881, 2021. 3
- [46] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. *CoRR*, abs/1811.11168, 2018. 2