

# Boosting Video Super Resolution with Patch-Based Temporal Redundancy Optimization

**Yuhao Huang<sup>1</sup> <sup>\*</sup>, Hang Dong<sup>2</sup> <sup>\*</sup> <sup>†</sup>, Jinshan Pan<sup>3</sup>, Chao Zhu<sup>1</sup>, Boyang Liang<sup>1</sup>, Yu Guo<sup>1</sup>, Ding Liu<sup>2</sup>, Lean Fu<sup>2</sup>, Fei Wang<sup>1</sup>**

<sup>1</sup>Xi'an Jiaotong University

<sup>2</sup>ByteDance Intelligent Creation Lab

<sup>3</sup>Nanjing University of Science and Technology

hyhsimon@gmail.com

## Abstract

The success of existing video super-resolution (VSR) algorithms stems mainly exploiting the temporal information from the neighboring frames. However, none of these methods have discussed the influence of the temporal redundancy in the patches with stationary objects and background and usually use all the information in the adjacent frames without any discrimination. In this paper, we observe that the temporal redundancy will bring adverse effect to the information propagation, which limits the performance of the most existing VSR methods and causes the severe generalization problem. Motivated by this observation, we aim to improve existing VSR algorithms by handling the temporal redundancy patches in an optimized manner. We develop two simple yet effective plug-and-play methods to improve the performance and the generalization ability of existing local and non-local propagation-based VSR algorithms on widely-used public videos. For more comprehensive evaluating the robustness and performance of existing VSR algorithms, we also collect a new dataset which contains a variety of public videos as testing set. Extensive evaluations show that the proposed methods can significantly improve the performance and the generalization ability of existing VSR methods on the collected videos from wild scenarios while maintain their performance on existing commonly used datasets. The code is available at <https://github.com/HYHsimon/Boosted-VSR>.

## 1 Introduction

Video Super-Resolution (VSR) aims to reconstruct a high-resolution visual-pleasing video from a low-resolution one. Recent years have witnessed significant advances due to the use of deep convolutional neural networks (CNNs). As more frames are used, VSR methods achieve better performance than the single image SR methods (Haris, Shakhnarovich, and Ukita 2018; Zhang et al. 2018; Dai et al. 2019; Zhou et al. 2020; Mei, Fan, and Zhou 2021; Chen et al. 2021) on existing VSR datasets (e.g., REDS (Nah et al. 2019), Vid4 (Liu and Sun 2013), Vimeo-90K (Xue et al. 2019)). However, the VSR task introduces another challenging problem, i.e., how to effectively exploit the temporal information for better results.

<sup>\*</sup>These authors contributed equally.

<sup>†</sup>Corresponding author.

To solve this problem, most existing deep learning-based methods usually employ optical flow, deformable convolution networks, and recurrent neural networks to explore useful information from adjacent frames for better high-resolution video restoration. Existing deep learning-based VSR methods can be roughly categorized into local propagation-based (e.g., EDVR) and non-local propagation-based (e.g., BasicVSR) methods according to the propagation scheme of the input frames. The success of existing VSR stems mainly exploiting the temporal information from the neighboring frames through propagation.

Meanwhile, we note that the neighboring frames also contains similar contents (i.e., temporal redundancy) in the patches with the stationary objects and background. If these temporal redundancy contents dominate the propagation process, they will not facilitate the VSR problem as no additional useful information is introduced from the temporal domain. However, most existing methods usually use all the information from adjacent frames without any discrimination. Therefore, the temporal redundancy are likely to be involved in the high-resolution frame reconstruction process.

In this paper, we find that the temporal redundancy in stationary objects and background interfere with the high-resolution frame reconstruction if they are not specially handled. As shown in Figure 1, we select one patch sequence with stationary objects and background  $s_{[t-2:t+2]}$  and one patch sequence with dynamic scene  $d_{[t-2:t+2]}$  from input frames  $I_{[t-2:t+2]}$  and super-resolve them with two typical VSR networks in the local (EDVR (Wang et al. 2019)) and non-local (BasicVSR (Chan et al. 2021a)) propagation-based methods. To evaluate the benefit of neighboring frames, we also super-resolve the reference patches ( $s_t$  and  $d_t$ ) with two single frame counterparts of these two methods for comparisons. The super-resolved results of these two patch sequences are shown in the right side of Figure 1. As expected, by exploiting the temporal information from the neighboring patches, both networks can achieve better results in the dynamic patch. In the meantime, due to the existence of temporal redundancy contents, the single frame counterparts outperform the VSR networks in the patch with stationary objects and background. The inconsistent performance of the VSR networks on two patch sequences demonstrates that the temporal redundancy may bring adverse ef-

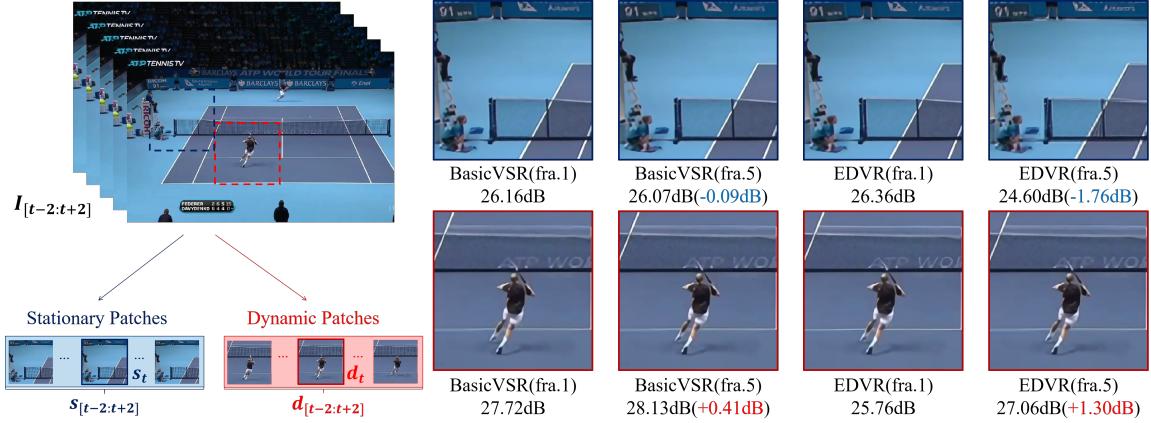


Figure 1: **Effect of the temporal redundancy in stationary objects and background.** Since the EDVR can only received five frames as inputs, we modify the original EDVR to adapt the single frame input (more details can be find in Sec. 4.1). Both the EDVR and BasicVSR are trained on the REDS dataset. fra.N means method takes N frames as input.

fect on the VSR problem and patches with stationary background and dynamic objects should be handled separately. VSR networks exploit the useful subpixel temporal information though alignment from the neighboring patches with dynamic objects. However, pixels may still change due to noisy and information loss during encoding and decoding in the patches with stationary background, which will be regarded as the useful temporal information by the alignment module and bring adverse effect on the VSR problem.

To overcome this problem, we try to handle the temporal redundancy patches in an optimized manner and develop two simple yet effective plug-and-play methods to improve the performance and the generalization ability of existing VSR algorithms. Our work is motivated by two observations: the temporal redundancy content is universal on different types of videos and the single frame super-resolution is more suitable for handling patches with temporal redundancy contents. This inspired us to propose a new VSR pipeline with temporal redundancy detection module for local propagation-based methods and deploy it to the original EDVR, namely **Boosted EDVR**. Specifically, the proposed pipeline first decomposes the input frames into overlapping patches and super-resolve the detected patches with a fine-tuned EDVR model for single frame (EDVR-1F). Since the EDVR-1F is more suitable for super-resolving patches with temporal redundancy and has lower computational cost than the original EDVR, the **Boosted EDVR** could improve the performance and accelerate the inference time simultaneously.

Moreover, we also optimize the non-local propagation-based VSR methods in a different way based on another observation: As for the non-local propagation-based VSR methods, one frame may strongly affect the next adjacent frame, but its influence is quickly lost after few time steps. Therefore, the temporal redundancy in the patch sequences will hinder the propagation of the hidden states since the temporal information in the distant frame may be gradually vanished by the temporal redundancy from neighboring patches. To improve the effectiveness of the hidden states

propagation in the presence of temporal redundancy contents, we propose a patch-based dynamic propagation (PDP) scheme to better accumulate and exploit the long-term information. Unlike existing propagation schemes, where the information is sequentially propagated frame-by-frame, the proposed patch-based dynamic propagation can directly propagate the long-term information to the current frame in a patch-wise way without accumulating useless redundancy. We deploy this propagation scheme to BasicVSR, namely **Boosted BasicVSR**, and largely improve the performance and the generalization ability without any training process.

In addition, we also collect a new testing dataset which contains a variety of public videos to comprehensively evaluate the robustness and performance of VSR algorithms. More specifically, the collected testing dataset contains videos from live streaming, TV program, sports live, movie and television, surveillance camera, advertisement, and some first person videos captured with irregular trajectories. We believe that the new dataset is suitable for evaluating the importance of temporal redundancy and can enrich the video types of the existing datasets.

The contributions of this work are summarized as follows:

- In this paper, we find that the temporal redundancy is universal in public videos and will limit the potential of the existing VSR methods. To the best of our knowledge, this is the first work to investigate the influence of the temporal redundancy in the VSR task.
- We develop two plug-and-play methods for both the local and non-local propagation-based VSR methods, which can optimize the super-resolving process for the temporal redundancy patches and save computational cost.
- We collect a dataset with a variety of public videos to enrich the existing datasets. Extensive evaluations demonstrate that the proposed methods can largely improve the performance and the generalization ability of existing VSR algorithms.

## 2 Related Works

Most existing VSR algorithms (Chan et al. 2020; Isobe et al. 2020b; Huang, Wang, and Wang 2017, 2015; Caballero et al. 2017; Tao et al. 2017; Jo et al. 2018a; Yi et al. 2019; Tian et al. 2020; Wang et al. 2019; Isobe et al. 2020a,c; Lai et al. 2017; Kingma and Ba 2014; Kim et al. 2018; Jo et al. 2018b) focus on improving the motion compensation and frame aggregation modules to better exploit temporal information. In VESPCN (Caballero et al. 2017), a real-time deep motion compensation module is proposed for frames registration. SPMC (Tao et al. 2017) further improve the process by proposing a sub-pixel motion compensation (SPMC) strategy, which is validated by the physical imaging model. Since optical flow estimation is a challenging task in dynamic scenes, some recent works adopt implicit alignment without the optical flow estimation process. EDVR and TDAN (Tian et al. 2020) both adopt deformable convolutions (DCNs (Dai et al. 2017)) to align the features of the neighboring frames in a multi-scale architecture. In DUF (Jo et al. 2018a), a novel learned dynamic upsampling filter is proposed to exploiting the spatio-temporal of each pixel without explicit motion compensation. Although these sliding-window frames can achieve favorable results, none of them discuss the effect of the temporal redundancy, which leads to sub-optimal results and causes unnecessary consumption.

Since the RNN (non-local propagation-based) architecture has been validated to be effective in processing the time sequence signals, it is also applied in the some video super-resolution tasks. FRVSR (Sajjadi, Vemulapalli, and Brown 2018) first proposes a recurrent network to super-resolve the low resolution video by leveraging the HR output from last iteration. Since the propagation is one of the most influential components in non-local propagation-based VSR algorithms, subsequent methods propose new propagation schemes to improve the information-flow of the hidden states. RRN (Isobe et al. 2020c) proposes a new recurrent residual block to solve the gradient vanish problem and preserve the texture information over long periods. Recently, BasicVSR and BasicVSR++ (Chan et al. 2021b) achieves SotA performance on all the existing datasets by adopting a bidirectional propagation coupled with optical flow-based and deformable-based feature alignments. Despite the distinguished performance, the information in BasicVSR and BasicVSR++ are still sequentially propagated frame-by-frame which is not optimal when temporal redundancy patches exist. The most similar work to our paper is RSDN (Isobe et al. 2020a), where a spatially variant hidden state adaptation module is proposed to only propagate the similar information in previous frames to the current frame at each position. However, this strategy bring serious adverse effects when handling the video with temporal redundancy, since the useful information in the long-term frames will be totally replaced by the temporal redundancy contents.

Table 1: **Performance of EDVR-1f and two input types of EDVR-5f.** Type A and Type B sequences refer to the stationary and dynamic sequences.

Models	Type A sequences	Type B sequences
EDVR-1F	<b>39.20dB</b>	38.01dB
EDVR-5f(original)	37.81dB	<b>38.65dB</b>

## 3 Observations on Temporal Redundancy

### 3.1 The DTVIT Dataset

Currently, most VSR datasets are first-person videos, which contains only dynamic scenes due to consistent movement. However, there are a variety of videos with irregular movement in public videos. To better investigate temporal redundancy and its influence, We collected a Diverse Types Videos with Irregular Trajectories (DTVIT) Dataset. More specifically, we collect 96 videos with high-quality and high-resolution as ground-truth from the internet. To ensure the diversity of the datasets, the collected videos include live streaming, TV program, sports live, movie and television, surveillance camera, and advertisement. Besides, to further increase the quantity and diversity of the collected dataset, we also additionally capture 12 first-person videos with irregular trajectories (using iPhone 12 with DJI stabilizer). More details can be find in the supplementary. Then, we randomly select ten videos from DTVIT dataset as the validation set and try to investigate the influence of temporal redundancy based on it.

### 3.2 Temporal Redundancy in Videos

**Observation 1:** *The temporal redundancy contents is universal in widely-used public videos.*

As temporal redundancy occurs in the stationary objects and background, we conduct a statistical analysis on the sliced patches of the validation set to determine the ratio of the patch sequence with stationary objects and background. Here, based on the input length of most local propagation-based VSR algorithms, we define the five neighboring patches, where the PSNR of each neighboring patch is higher than 35, as a patch sequence with stationary objects and background. Based on the definition above, there are 69.92% patch sequences in the validation set can be discriminated as stationary. Even we extend the length of the patch sequence to 11 patches, there are still 64.79% patch sequences can be treated as stationary. These statistic results demonstrate that the patch sequence with stationary objects and background, as well as the temporal redundancy, is universal in widely-used public videos. For convenience, the patch sequences with stationary objects and background are denoted as the Type A sequences, while the dynamic patch sequences are denoted as the Type B sequences.

**Observation 2:** *Single frame super-resolution is more suitable for handling patches with temporal redundancy in stationary objects and background.*

Since the temporal redundancy contents is universal in widely-used public videos, we should also investigate whether it will interfere with existing local propagation-based VSR networks. Following the settings of the exper-

Table 2: **The performance of BasicVSR in the simulated Type A sequences.** For a fair comparison, the PSNR are calculated on the original dynamic frames.

Training dataset	DS	+10df	+20df	+30df	+40df	+50df
REDS	27.38 dB	27.31 dB	27.24 dB	27.12 dB	26.99 dB	26.84 dB
Vimeo	25.85 dB	25.78 dB	25.73 dB	25.69 dB	25.64 dB	25.58 dB

iment in Sec. 1, we super-resolve all the Type A and Type B sequences in the validation set with both the EDVR-1f and original EDVR (EDVR-5f). The EDVR-1f is modified upon the original EDVR for single frame input, which will be described in Sec. 4.1. As shown in Table 1, although the EDVR-5f achieves better results on the type B sequences, the single frame super-resolution method (EDVR-1f) can outperform EDVR-5f with lower computational cost on the type A sequences. Since the type A sequences refer to the sequences with temporal redundancy, we analyze that alignment and fusion module of original EDVR may regard these changed pixels due to noisy and information loss during encoding and decoding as the useful temporal information and bring adverse effect on the VSR problem in the sequences with temporal redundancy. Therefore, the single frame super-resolution is more suitable for handling patches with temporal redundancy.

**Observation 3:** *Patches with temporal redundancy in the video sequence will hinder the propagation of non-local propagation-based VSR networks.*

According to the **Observation 2**, the existence of temporal redundancy will bring negative effect to local propagation-based VSR algorithms, where only local information can be exploited. On the other hand, non-local propagation-based VSR algorithms can exploit the long-term temporal information by taking all the inference frames as inputs. To investigate the influence of the temporal redundancy on such longer input sequences, we conduct an experiment based on the BasicVSR model. Specifically, we selected 4 downsampling videos with dynamic scenes from the REDS and super-resolve them with the BasicVSR trained on the REDS and the Vimeo respectively. Then, to simulate the Type A sequence and introduce the temporal redundancy, we randomly choose 10 frames from each video and replicated them several times (range from 1 to 5), progressively. For each time, we super-resolve all the extended videos with BasicVSR and record its performance. As shown in Table 2, the two BasicVSR models both suffer from the performance decline as the length of frames with temporal redundancy increases, which demonstrates limitation of RNN due to the recurrent nature, where one frame may strongly affect the next adjacent frame but its influence is quickly lost after few time steps. Therefore, despite of long input sequences, the temporal redundancy will still bring negative effect to the RNN-based VSR network by hindering the information propagation. Similarly results in the realistic video can be found in the supplementary.

## 4 Methodology

From the observations in Sec. 3.1, the temporal redundancy is universal: almost 70% patch sequences in the validation

set are Type A sequences, which cannot provide any useful information for the VSR algorithms. Therefore, it's necessary to optimize the existing VSR algorithms to handle the patches with temporal redundancy. However, there are two categories in the existing VSR methods, which makes it difficult to propose a unified strategy to improve two frameworks simultaneously. In this section, based on the **Observation 2** and **Observation 3**, we introduce two effective plug-and-play methods for local and non-local propagation-based networks to optimize the super-resolving process for patches with temporal redundancy.

### 4.1 Boosting Local Propagation-Based Networks

The local propagation-based VSR methods (Wang et al. 2019; Tao et al. 2017; Jo et al. 2018a; Caballero et al. 2017) take LR images within a local window as inputs and employ the local information for restoration. However, based on the **Observation 2**, the patches with temporal redundancy should be specially handled. To achieve this, we try to introduce a temporal redundancy detection module to the existing methods and super-resolve each patch adaptively. In the following parts, we will use the EDVR as example to show how the proposed plug-and-play method optimize the local propagation-based VSR methods.

Inspired by the recent work, Class-SR(Kong et al. 2021), we extend the original EDVR to a new pipeline, namely **Boosted EDVR**, to perform temporal redundancy detection and super-resolution simultaneously. As shown in Figure 2, the proposed **Boosted EDVR** consists of two modules: Temporal Redundancy Detection Module (TRDM) and Adaptive Super-Resolution Module (ASRM). The input five LR neighboring frames  $X_{[t-2:t+2]}$  are first decomposed into  $N$  overlapping patch sequences  $\{x_{[t-2:t+2]}^i\}_{i=1}^N$ . Then, each decomposed patch sequence  $x_{[t-2:t+2]}^i$  is fed to the TRDM and assigned a movement label ( $L_j^i, j \in \{1, 3, 5\}$ ) according to its motion state among neighboring patches. After that, all the patch sets with the same label will be concatenated in the batch-size dimension and super-resolved by the optimal EDVR model in ASRM. Finally, we combine all the super-resolved patches  $\{y_t^i\}_{i=1}^N$  to get the final SR results  $Y_t$ .

**Temporal redundancy detection module.** The goal of TRDM is to detect the temporal redundancy and assign a movement label to each patch sequence. Based on the **Observation 2**, the temporal redundancy exists in the stationary objects and background, which means we should find a way to represent the motion state between two patches. Since the optical flow is a widely-used metric to describe the motion information, we use the mean values of the optical flow to represent the motion state, which can be formulated as:

$$m_{-1 \rightarrow 0}^i = \text{mean}(|f(x_{t-1}^i, x_t^i)|), \quad (1)$$

where  $f$  denotes the optical flow estimator,  $|\cdot|$  denotes absolute value,  $\text{mean}$  is the mean value, and  $m_{-1 \rightarrow 0}^i$  denotes the motion state between the reference patch ( $x_t^i$ ) and its neighboring patch ( $x_{t-1}^i$ ) in the patch sequence  $i$ . We choose the traditional DIS (Kroeger et al. 2016) algorithm as the optical flow estimator since it only slightly increase the computational cost.

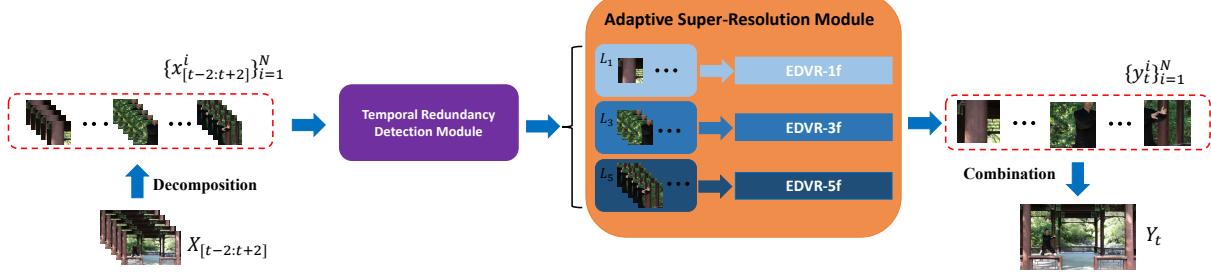


Figure 2: Overview of the proposed Boosted EDVR.

For  $i$ -th patch sequence, we successively calculate the motion states of all the neighboring patches, which denote as  $m_{-2 \rightarrow -1}^i, m_{-1 \rightarrow 0}^i, m_{1 \rightarrow 0}^i$ , and  $m_{2 \rightarrow 1}^i$ . Then, we assigned a movement label ( $L_j(x_{[t-2:t+2]}^i), j \in \{1, 3, 5\}$ ) according to these motion states:

$$L_j^i = \begin{cases} L_1^i & \text{if } m_{-1 \rightarrow 0}^i < \gamma \text{ and } m_{1 \rightarrow 0}^i < \gamma, \\ L_3^i & \text{elif } m_{-2 \rightarrow -1}^i < \gamma \text{ and } m_{2 \rightarrow 1}^i < \gamma, \\ L_5^i & \text{otherwise,} \end{cases} \quad (2)$$

where  $\gamma$  is the threshold to discriminate the patch with stationary objects and background and  $L_j^i$  denotes  $j$  dynamic patches involved in  $i$ -th patch sequence. With TRDM, we can determine which model in the following ASRM should be used to obtain better super-resolved results.

**Adaptive Super-Resolution Module.** The ASRM, which consists of the original EDVR (EDVR-5f) and two of its variants (EDVR-3f and EDVR-1f), is designed to super-resolve each patch sequence with the optimal model. Specifically, we adopt the EDVR-1f model, which is modified for single frame input based on EDVR, to super-resolve all the patch sets with the movement label  $L_1$ , since there is no useful temporal information in the neighboring patches. Similarly, the EDVR-3f model and EDVR-5f model will process the patch sequences with the movement labels  $L_3$  and  $L_5$ , respectively. Different from experiment in **Observation 2**, we introduce the EDVR-3f model by taking the situation that the temporal redundancy only occurs at the border frames of the patch sequence into consideration.

To acquire EDVR-1f and EDVR-3f with minimal modification, we only slightly changes the forward flow of the original EDVR (EDVR-5f) without any changes on the network architecture. For EDVR-1f and EDVR-3f, the PCD alignment module and the temporal attention layers in TSA module are only performed once and three times, respectively, and the features will be replicated to the same shape as EDVR-5f before sending to the fusion convolutional layer in the TSA module. Since we remove the unnecessary calculation in the PCD alignment and TSA modules of the EDVR-1f and EDVR-3f, the proposed pipeline will be more efficient than the original EDVR. More detailed of the EDVR-1f and EDVR-3f can be found in the supplementary. To ensure the EDVR-1f and EDVR-3f can achieve comparable super-resolving ability as EDVR, we also fine-tune them on the same training dataset (REDS) and with same hyper-parameter as EDVR. Experiments show that such a simple

pipeline can improve the performance of EDVR close to the upper bound with less FLOPs.

## 4.2 Boosting Non-Local Propagation-Based Networks

Unlike local propagation-based methods, the non-local propagation-based methods can exploit long-term information by taking all the inference frames as inputs and sequentially propagation. However, based on the **Observation 3**, the patches with temporal redundancy in the video sequence will hinder the propagation, which inevitably limits the potential of the existing non-local propagation-based VSR methods. To better exploit the long-term information, we propose a new plug-and-play method by introducing a Patch-based Dynamic Propagation (PDP) branch to dynamically propagate the long-term information in a patch-wise way. As shown in Figure 3(a), we deploy the proposed plug-and-play method to BasicVSR, namely **Boosted BasicVSR**, by replacing the original propagation branches with the proposed PDP branches. In the following parts, we will show how the PDP branch works in forward propagation ( $PDP_f$ ), and the PDP branch in the backward propagation ( $PDP_b$ ) can be derived accordingly.

Unlike the propagation branch in the BasicVSR, the proposed forward PDP branch adopts dynamical propagation, where each patch of the current frame can receive information from different frames. To achieve this, the proposed forward PDP branch maintains a patch pool  $P_{rgb}^f$  and its corresponding hidden state pool  $P_\phi^f$  to restore the useful information of patches from different frames. Then, the forward PDP branch takes the current LR frame  $X_t$ ,  $P_{rgb}^f$ , and  $P_\phi^f$  as inputs and generates the forward features  $h_t^f$  while updating  $P_{rgb}^f$  and  $P_\phi^f$  based on the temporal redundancy detection. The advantage of maintaining an independent patch-wise hidden states pool and propagating it to current frame instead of the neighboring hidden states is that the useful information in the long-term frame can directly connect to current frame without accumulating useless redundancy information. The detail of the PDP branch is shown in Figure 3(b), which consists of two stages: features aggregation and patch pools update.

**Features aggregation.** This stage is design to aggregate the information in the maintained pools ( $P_{rgb}^f$  and  $P_\phi^f$ ) with the

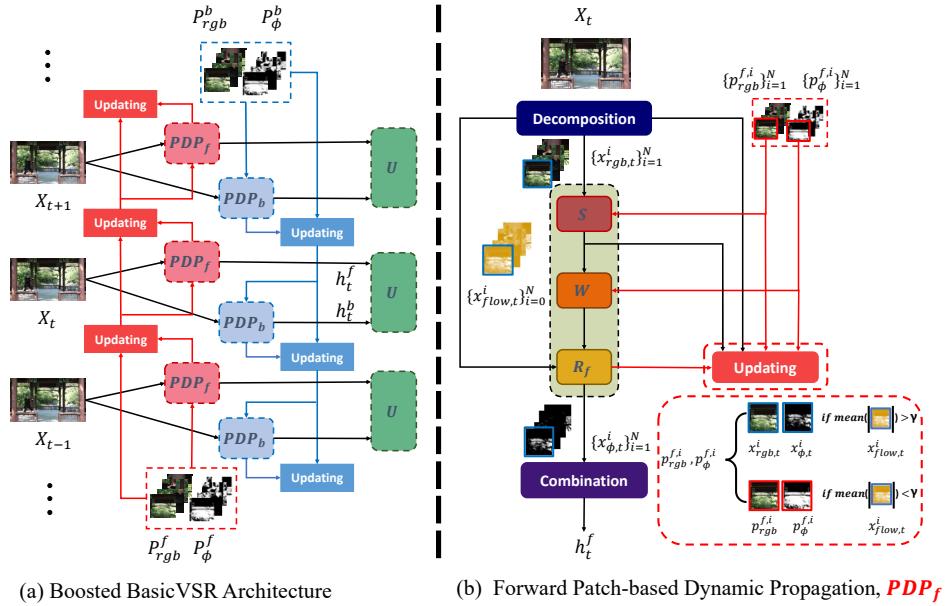


Figure 3: Overview of the proposed **Boosted BasicVSR**. (a) The upsampling module  $U$  contains multiple pixel-shuffle and convolutions. The red and blue colors represent the forward and backward propagations, respectively. (b)  $S$ ,  $W$ , and  $R_f$  refer to the flow estimator, spatial warping module, and residual blocks for forward branch, respectively.

current frame. The  $P_\phi$  and  $P_{rgb}$  contains  $N$  overlapping patches  $\{p_{rgb}^{f,i}\}_{i=1}^N$  and corresponding hidden state patches  $\{p_\phi^{f,i}\}_{i=1}^N$  which are sorted by their positions in the frame. Since the BasicVSR adopts the first-order propagation, we only maintain the information of one patch for each position.

To estimate the optical flow for spatial alignment of the hidden state pool  $P_\phi^f$ , we first decompose current frame  $X_t$  into  $N$  overlapping patches ( $\{x_{rgb,t}^i\}_{i=1}^N$ ). Then, the optical flows of all the patches ( $\{x_{flow,t}^i\}_{i=1}^N$ ) are calculated by sending the correspond patches in the  $\{x_{rgb,t}^i\}_{i=1}^N$  and  $\{p_{rgb}^{f,i}\}_{i=1}^N$  to the optical estimator ( $S$ ). After that, we perform warping ( $W$ ) on the patches in the hidden state pool using the estimated flow for the further refinement in the residual blocks  $R_f$ . By feeding the warped hidden state pool and the overlapping patches of current frame into the residual blocks, the intermediate features patches of current frame ( $\{x_{\phi,t}^i\}_{i=1}^N$ ) can be obtained. Finally, the forward features  $h_t^f$  can be obtained by combining the  $\{x_{\phi,t}^i\}_{i=1}^N$ .

**Patch pools update.** In this stage, we try to update the patch pool  $P_{rgb}^f$  and hidden state pool  $P_\phi^f$  with the information of current frame. As shown in Figure 3(b), we use a similar temporal redundancy detection method in **Boosted EDVR** to decide which patches in  $P_{rgb}^f$  and  $P_\phi^f$  should be updated with current frame. Since we already obtain the optical flows ( $\{x_{flow,t}^i\}_{i=1}^N$ ) in the features aggregation stage, we direct use Equ. (1) to obtain the motion states  $\{m_t^i\}_{i=1}^N$  (the box with red dashed line) of all the corresponding patches between  $\{p_{rgb}^{f,i}\}_{i=1}^N$  and  $\{x_{rgb,t}^i\}_{i=1}^N$ . Then, to ensure the use-

ful information can be accumulated, each patch set ( $p_{rgb}^{f,i}$  and  $p_\phi^{f,i}$ ) in the two pools will be replaced by the information of corresponding patch of current frame ( $x_{rgb,t}^i$  and  $x_{\phi,t}^i$ ) when the motion state of this patch ( $m_t^i$ ) is larger than the threshold  $\gamma$ . Otherwise, which means temporal redundancy exists in these two patch, the information of current frame will be discarded to avoid vanishing the useful information. Finally, the updated pools will be propagated to the next frame. Experiments demonstrate that the proposed PDP scheme can significantly improve the performance of BasicVSR and solve the generalization ability without any training process.

## 5 Experiments

### 5.1 Datasets and Settings

Since we aim to boost existing VSR algorithms with minimal modifications, we only fine-tune the EDVR-1f and EDVR-3f on the training set of the REDS. Then, we use the REDS4, Vid4, and the DTVIT as the test set to compare the proposed models with existing VSR algorithms. For fair comparison, all the evaluated models are trained and tested on the dataset with  $4\times$  bicubic downsampling.

For the proposed two methods, each LR frame is decomposed into  $64\times 64$  patches with stride 56 (with 8 pixel overlaps), and the combination operation combines all the patches to an integrated frame by averaging overlapping areas. The threshold  $\gamma$  in **Boosted EDVR** and **Boosted BasicVSR** are set to 1 and 0.2, respectively. The DTVIT dataset and source code will be made available to the public.

Table 3: Analysis on each component of the proposed Boosted EDVR.

Methods	EDVR	TR-EDVR	Boosted EDVR-(15)	Boosted EDVR-(135)	Boosted EDVR-(UB)
EDVR-5f	✓	✓	✓	✓	✓
EDVR-3f				✓	✓
EDVR-1f	✓	✓	✓	✓	✓
TR detection	✓			✓	
DIS flow		✓	✓		
Flops	758M (100%)	661M (87%)	522M (69%)	<b>519M (68%)</b>	622M (82%)
PSNR	33.42	34.30	34.50	34.51	34.72

## 5.2 Experiments on Network Configurations

To investigate the effect of different network configurations and find the optimal one for the proposed **Boosted EDVR** and **Boosted BasicVSR**, we evaluate several models with alternative configurations. When evaluating **Boosted EDVR**, we also calculate the average FLOPs to evaluate the efficiency. For quick verification during the design stage, we still select the validation set of the DTVIT dataset for the ablation study.

**Study of the Boosted EDVR.** Starting from the original EDVR, we first use the mean square errors (MSE) of pixel values to detect temporal redundancy and use the fine-tuned EDVR-1f models to super-resolve the patch sequences with movement label  $L_1$ . We denote these configuration as TR-EDVR. As shown in Table 3, the TR-EDVR can achieve 0.88 dB performance gain over the original EDVR with less FLOPS. These results also demonstrate the effectiveness of the proposed pipeline with temporal redundancy detection module, which can adaptively super-resolve different patch sets with the optimal model. Since the optical flow is widely used to describe the motion information, we use the mean values of the DIS optical flow to represent the motion state and form the Boosted EDVR-(15). The Boosted EDVR-(15) outperforms TR-EDVR by a margin of 0.2 dB while the overall FLOPs drop from 661M to 522M, which demonstrates that the DIS optical flow is more suitable for redundancy detection than MSE. Furthermore, we also introduce a fine-tuned EDVR-3f model, namely Boosted EDVR-(135), to super-resolve the patch sets where the temporal redundancy only occurs at the border patches ( $m_{-2 \rightarrow -1}^i < \gamma$  and  $m_{2 \rightarrow 1}^i < \gamma$ ). Since both the performance and efficiency are slightly improved by introducing EDVR-3f, we choose the Boosted EDVR-(135) as the final configurations of **Boosted EDVR**. Compared with the EDVR, the proposed **Boosted EDVR** can achieve 1.09 dB performance gain with only 68% computational cost. Finally, we also obtain the upper bound of the **Boosted EDVR** by simultaneously feeding each patch sequence to three models respectively and choosing the best one (in terms of PSNR) as result. Since the performance gap between the **Boosted EDVR** and the upper bound model is relatively small (0.21 dB) and the proposed method can save more computational cost, we think our pipeline is acceptable by maintaining a good balance between the performance and efficiency.

**Study of the Boosted BasicVSR.** In this part, we will evaluate the importance of three key factors in the proposed Patch-based Dynamic Propagation (PDP): temporal redundancy detection, dynamic propagation, and patch-wise strategy. As shown in Table 4, the performance of BasicVSR trained on the REDS is much worse than the original

Table 4: Analysis on each key factor of the proposed PDP branch.

Methods	BasicVSR	TR-BasicVSR	DP-BasicVSR	Boosted BasicVSR
Frame type classification	✓	✓	✓	✓
Dynamic propagation			✓	✓
Patch-wise				✓
PSNR	27.96	32.57	33.22	<b>34.08</b>

EDVR (27.96 dB vs. 33.42 dB) on the validation set, which is contradictory to the results on the existing datasets. We owe this severe generalization problem of BasicVSR to the error accumulation of the optical flow: since the optical flow estimator in BasicVSR may regard these changed pixels due to noisy and information loss during encoding and decoding as the useful temporal information, it will produce inaccurate optical flow between the frames with stationary objects and background and the error will be accumulated through the propagation (**more analysis can be found in the supplementary**). To overcome this problem, we propose a new pipeline, namely TR-BasicVSR, to super-resolve stationary and dynamic frames separately. More specifically, we follow the sequence definitions in **Observation 1** and divide the types of each test video using the redundancy detection module in Sec. 4.1. Then, we combine all the Type B sequences into one sequence and super-resolve it with BasicVSR. For Type A sequences, where all the frames are similar in one sequence, we super-resolve each frame independently to avoid the error accumulation of the optical flow. As shown in Table 4, the TR-BasicVSR obtain significant performance gain over the original BasicVSR, which demonstrates that the temporal redundancy detection can solve the generalization problem effectively.

However, the TR-BasicVSR cannot exploit any temporal information from the Type B sequences when handling Type A sequences, which inevitably limits its performance. Therefore, we further introduce the dynamic propagation scheme to TR-BasicVSR (referred to as the DP-BasicVSR) and make sure each frame can exploit the useful temporal information. Specifically, the DP-BasicVSR maintains an anchor frame and its corresponding hidden states to restore the long-term information from the closest dynamic frame and propagate it to current frame. Since the dynamic propagation scheme can directly propagate the information from the long-term frame to current frame without accumulating useless redundancy information of the stationary objects and background, the DP-BasicVSR outperforms TR-BasicVSR by a margin of 0.65 dB.

Finally, due to the contents in different patches of a video may change independently, the final **Boosted BasicVSR** maintains a patch pool  $P_{rgb}^f$  and its corresponding hidden state pool  $P_\phi^f$  to restore long-term information in a patch-wise way. By adopting the patch-wise strategy, the **Boosted BasicVSR** achieves 0.86 dB performance gain over DP-BasicVSR. Overall, the proposed **Boosted BasicVSR** can solve the generalization problem of the pre-trained BasicVSR and boost its performance without any training process, which demonstrates the effectiveness of the proposed PDP scheme.

Table 5: **Quantitative comparison (PSNR/SSIM).** All results are calculated on RGB-channel.

Training dataset	Methods	REDS Val PSNR/SSIM	Vid4 PSNR/SSIM	DTVIT PSNR/SSIM
REDS	Bicubic	26.14/0.7292	23.78/0.6347	29.46/0.8870
	DUF	28.63/0.8251	18.45/0.5117	23.17/0.6517
	RBPN	30.09/0.8590	25.66/0.8024	32.74/0.9208
	MuCAN	30.88/0.8750	25.33/0.7994	30.58/0.9072
	EDVR	30.53/0.8699	25.34/0.7951	32.00/0.9205
	EDVR-L	31.09/0.8800	25.40/0.8008	32.39/0.9277
	BasicVSR	31.42/0.8909	25.75/0.8155	27.13/0.8165
	<b>Boosted EDVR</b>	30.53/0.8699	25.32/0.7950	32.91/0.9262
	<b>Boosted BasicVSR</b>	<b>31.42/0.8917</b>	<b>25.93/0.8202</b>	<b>33.21/0.9340</b>
	BasicVSR	30.32/0.8672	25.82/0.8085	33.31/0.9368
Vimeo	<b>Boosted BasicVSR</b>	<b>30.32/0.8673</b>	<b>25.84/0.8093</b>	<b>33.79/0.9503</b>

### 5.3 Comparisons with Existing VSR algorithms

To further evaluate the proposed methods, we conduct comprehensive experiments by comparing **Boosted EDVR** and **Boosted BasicVSR** with several state-of-the-art VSR algorithms: DUF, RBPN, MuCAN, EDVR, EDVR-L, and BasicVSR.

The first and second columns in Table 5 show the quantitative results on the REDS and Vid4, where all the testing videos are first-person videos with consistent movement. As expected, the proposed **Boosted EDVR** and **Boosted BasicVSR** only achieve comparable performance with EDVR and BasicVSR on these two datasets, since they are optimized for videos with temporal redundancy. However, the stable performance on the first person videos demonstrates that the proposed methods are robustness and will not bring any adverse influence to existing datasets.

Meanwhile, for larger evaluation dataset, we select six clips in the REDS training clips and extend the REDS test set (i.e. REDS4) to ten clips, denoted by REDS10. The remaining training clips are used as new training dataset (a total of 260 clips). Based on the setting above, the proposed **Boosted BasicVSR** trained on the REDS outperforms BasicVSR by a margin of 0.15 dB on the REDS10 (30.86dB v.s. 31.01dB), which demonstrate that the proposed **Boosted BasicVSR** can still improve the performance on the REDS and temporal redundancy also exists in some of the REDS.

To comprehensively evaluate the performance of VSR algorithms on different types of public videos, we also evaluate these algorithms on the collected DTVIT dataset. As shown in the third column of Table 5, the BasicVSR trained on the REDS performs not well on the collected video dataset due to the generalization problem. Although the EDVR-M and EDVR-L achieve favorable performance than other methods, the proposed **Boosted EDVR** can further improve the performance by up to 0.91 dB over EDVR-M and outperform EDVR-L with much lower computational cost. Moreover, the proposed **Boosted BasicVSR** can solve the generalization problem and significantly improve the performance by a large margin of 6.28 dB over BasicVSR. In addition, the **Boosted BasicVSR** outperforms **Boosted EDVR** by a margin of 0.24 dB with comparable computational cost, which coincides with the results on the existing dataset. Overall, both **Boosted EDVR** and **Boosted BasicVSR** are able to achieve remarkable performance on the collected dataset, which demonstrates that the proposed plug-and-play methods can improve the performance and robustness of existing VSR algorithms.



Figure 4: **Qualitative comparison on the DTVIT dataset.**

To verify the proposed method can not only solve the generalization problem but also can enhance the effectiveness of the propagation branches, we also apply the proposed Patch-based Dynamic Propagation branch to the BasicVSR trained on the Vimeo to see whether the improvement can be obtained. As shown in Table 5, since the Vimeo contains more types of videos, the BasicVSR trained on it will not suffer of severe generalization problem and achieves favorable on the DTVIT dataset. In addition, the proposed **Boosted BasicVSR** can still outperform the BasicVSR without any training process, which demonstrates that the proposed method can effectively improve the performance of existing non-local propagation-based VSR algorithms. Moreover, our method largely boosts the BasicVSR trained on the REDS and achieves similar performance with the BasicVSR trained on the Vimeo (33.15 dB v.s. 33.31 dB) without extra training datasets and time-consuming training process, which makes our method practical in real-world applications and can be easily extended to other video restoration tasks.

Qualitative comparisons are shown in Figure 4. The **Boosted EDVR** and **Boosted BasicVSR** recover finer details and sharper texts in the videos from the DTVIT dataset. More examples are provided in the supplementary.

## 6 Conclusion

In this paper, we investigate the temporal redundancy in the video and note it as an important factor for VSR methods for three reasons: (1) it will bring unnecessary computational cost for local propagation-based networks (e.g., EDVR), (2) it will cause severe generalization problem for the models trained on the dynamic datasets (as BasicVSR trained on the REDS performs not well in the DTVIT), and (3) it will gradually vanish the useful temporal information in the distant frame and hinder the performance of the non-local propagation-based networks.

Therefore, we focus on optimizing the existing VSR algorithms by taking the adverse effect of the temporal redundancy into consideration. Through introducing a temporal redundancy detection and adaptive super-resolution module to the original EDVR, we propose **Boosted EDVR**, a simple yet effective method can improve the performance and accelerate the inference time simultaneously. We also pro-

pose **Boosted BasicVSR** by adopting a Patch-based Dynamic Propagation (PDP) scheme to solve the generalization problem of the original BasicVSR and boost its performance without any training process. Extensive evaluations show that the proposed modifications can largely improve the performance on the collected dataset without any adverse influence to existing datasets. We believe that these two plug-and-play methods can also be applied to others video restoration tasks since the temporal redundancy is universal in most public videos.

## References

- Caballero, J.; Ledig, C.; Aitken, A.; Acosta, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4778–4787.
- Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2020. Understanding deformable alignment in video super-resolution. *arXiv preprint arXiv:2009.07265*, 4: 3.
- Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021a. BasicVSR: The search for essential components in video super-resolution and beyond. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4947–4956.
- Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2021b. BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. *arXiv*.
- Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained image processing transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 12299–12310.
- Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.
- Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, 11065–11074.
- Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep Back-Projection Networks for Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- Huang, Y.; Wang, W.; and Wang, L. 2015. Bidirectional recurrent convolutional networks for multi-frame super-resolution. *Advances in neural information processing systems*, 28: 235–243.
- Huang, Y.; Wang, W.; and Wang, L. 2017. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 1015–1028.
- Isobe, T.; Jia, X.; Gu, S.; Li, S.; Wang, S.; and Tian, Q. 2020a. Video super-resolution with recurrent structure-detail network. In *European Conference on Computer Vision*, 645–660. Springer.
- Isobe, T.; Li, S.; Jia, X.; Yuan, S.; Slabaugh, G.; Xu, C.; Li, Y.-L.; Wang, S.; and Tian, Q. 2020b. Video super-resolution with temporal group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8008–8017.
- Isobe, T.; Zhu, F.; Jia, X.; and Wang, S. 2020c. Revisiting temporal modeling for video super-resolution. *arXiv preprint arXiv:2008.05765*.
- Jo, Y.; Oh, S. W.; Kang, J.; and Kim, S. J. 2018a. Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jo, Y.; Oh, S. W.; Kang, J.; and Kim, S. J. 2018b. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3224–3232.
- Kim, T. H.; Sajjadi, M. S.; Hirsch, M.; and Scholkopf, B. 2018. Spatio-temporal transformer network for video restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 106–122.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kong, X.; Zhao, H.; Qiao, Y.; and Dong, C. 2021. ClassSR: A General Framework to Accelerate Super-Resolution Networks by Data Characteristic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12016–12025.
- Kroeger, T.; Timofte, R.; Dai, D.; and Van Gool, L. 2016. Fast optical flow using dense inverse search. In *European Conference on Computer Vision*, 471–488. Springer.
- Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 624–632.
- Liu, C.; and Sun, D. 2013. On Bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2): 346–360.
- Mei, Y.; Fan, Y.; and Zhou, Y. 2021. Image Super-Resolution With Non-Local Sparse Attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3517–3526.
- Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; and Mu Lee, K. 2019. 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*, 0–0.
- Sajjadi, M. S. M.; Vemulapalli, R.; and Brown, M. 2018. Frame-Recurrent Video Super-Resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tao, X.; Gao, H.; Liao, R.; Wang, J.; and Jia, J. 2017. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 4472–4480.
- Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2020. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3360–3369.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8): 1106–1125.

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; and Ma, J. 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3106–3115.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, 286–301.

Zhou, S.; Zhang, J.; Zuo, W.; and Loy, C. C. 2020. Cross-scale internal graph neural network for image super-resolution. *Neural Information Processing Systems*.

# Boosting Video Super Resolution with Patch-Based Temporal Redundancy Optimization

## Supplementary Material

**Yuhao Huang<sup>1</sup> \*, Hang Dong<sup>2</sup> \* †, Jinshan Pan<sup>3</sup>, Chao Zhu<sup>1</sup>, Boyang Liang<sup>1</sup>, Yu Guo<sup>1</sup>, Ding Liu<sup>2</sup>, Lean Fu<sup>2</sup>, Fei Wang<sup>1</sup>**

<sup>1</sup>Xi'an Jiaotong University

<sup>2</sup>ByteDance Intelligent Creation Lab

<sup>3</sup>Nanjing University of Science and Technology

hyhsimon@gmail.com

### Overview

In this supplemental material, we provide more details about the proposed DTVIT dataset in Sec. 1. To further validate the **Observation 3**, we conduct an experiment on a realistic public video and find similarly observation in Sec. 2. The details of EDVR-1f and EDVR-3f and their differences with EDVR-5f are provided in Sec. 3. We also analysis the generalization problem of the pre-trained BasicVSR on the DTVIT dataset in Sec. 4, the inaccurate estimated optical flow cause accumulated errors which leads to performance decline. To better illustrate the efficiency of the proposed method, the evaluations of computational cost and inference time are provided in Sec. 5. Moreover, to verify the universality of our methods, we also deploy the proposed plug-and-play methods to the models trained on the DTVIT-Train (a newly collected training dataset according to the categories of the DTVIT dataset) in Sec. 6. Finally, more qualitative results on the DTVIT dataset and real-world videos are provided in Sec. 7.

### 1 More Details about DTVIT

As mentioned in the manuscript, we collect 96 videos with high-quality and high-resolution as ground-truth and obtain the corresponding low-resolution inputs with  $4 \times$  bicubic downsampling.

The types of the videos in the DTVIT dataset are diverse and can be grouped into 7 categories: live streaming, TV program, sports live, movie and television play, surveillance video, advertisement, and first-person video with irregular trajectories. The examples of the collected DTVIT dataset are shown in Fig. 1 and the specific number of each type is shown in Table 1.

Moreover, we also collect a new training dataset according to the categories of the DTVIT dataset, referred as to DTVIT-Train, to verify the proposed method can not only solve the generalization problem but also can enhance the effectiveness of the propagation branches. The DTVIT-Train contains 100 video clips and each video clip is consist of 100 frames.

\*These authors contributed equally.

†Corresponding author.



Figure 1: **Example of the collected DTVIT dataset.** Best viewed on a high-resolution display.

Table 1: **The number of videos in the DTVIT dataset.** Each video contains 100 frames.

Video Style	Number of videos
Live streaming	19
TV program	20
Sports live	18
Movie and television play	9
Surveillance video	10
Advertisement	8
First person video	12
Total	96

### 2 More verification on Observation 3

In the manuscript, the **Observation 3** is validated in a simulated Type A sequence, where the temporal redundancy in a long sequence will still hinder the propagation of non-local propagation-based VSR networks. To further validate this, we apply the BasicVSR (Chan et al. 2021) on ten patch sequences from the realistic public video in the DTVIT dataset. As shown in Fig. 2-(a), all the selected patch sequences have a B-A-B type, which means a Type A sequence is inserted into a Type B sequence and separate the Type B sequence into two parts. In the meantime, to obtain the performance of VSR algorithms if the temporal information can be correctly propagated between the two Type B sequences, we remove the Type A sequence and combine the two Type B sequences as one (B-B type in Fig. 2-(b)). Then, we apply the BasicVSR on the B-A-B and B-B patch sequence and calculate the PSNR values only on the patches in the Type B sequences. According to the experiment, the BasicVSR perform much better on the B-B patch sequences (33.74 dB) than on the B-A-B patch sequences (32.40 dB), which demonstrates that the patches with temporal redundancy in the video sequence will hinder the propagation and

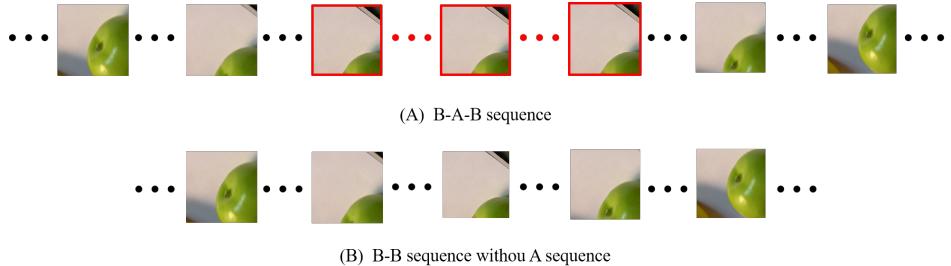


Figure 2: **Examples of the B-A-B and B-B sequences.**

cause generalization problem.

### 3 More Details of EDVR-1f and EDVR-3f

To obtain more suitable models for patch sequences with different motion states, We slightly modify the EDVR-5f (Wang et al. 2019) to acquire the EDVR-1f and EDVR-3f models for single-frame and three frames inputs, respectively. As shown in Fig. 3, unlike the original EDVR-5f, EDVR-1f and EDVR-3f only apply the PCD alignment module and temporal attention in the TSA module for 1 and 3 times, respectively. Then we replicate the feature in channel dimension to meet the requirement of the subsequent fusion convolution layer in TSA module. Specifically, the features in EDVR-1f and EDVR-3f,  $(C_0)$  and  $(C_{-1}, C_0, C_1)$ , will be replicated to the same shape as the corresponding features in EDVR-5f, i.e.,  $(C_0, C_0, C_0, C_0, C_0)$  and  $(C_{-1}, C_{-1}, C_0, C_1, C_1)$ , before feeding to the subsequent fusion convolution layer.

### 4 Analysis on the Generalization Problem of BasicVSR

To verify the inferior performance of BasicVSR on DTVIT dataset is caused by the error accumulation of the estimated optical flow (Ranjan and Black 2017), we present the intermediate optical flow during the inference time.

As shown in Fig. 4, where we visualize the estimated optical flow between neighboring frames, unexpected offsets occurs in stationary background. It is also noted that, no matter trained on the REDS datasets or the datasets of similar distribution (e.g., Vimeo and DTVIT-Train datasets), all the BasicVSR models suffer from the inaccurate flows due to regarding these changed pixels due to noisy and information loss during encoding and decoding as the useful temporal information. Then, the inaccurate flows will be used to perform back-warping on the hidden states and the errors will be pass to next frame via the propagation scheme. Therefore, the error of the optical flow will be accumulated progressively and cause unsatisfactory super-resolving results.

### 5 Efficiency Evaluations

To verify the efficiency of the proposed method, we evaluate the inference time of the proposed method on a TITAN RTX GPU. As shown in Fig. 5, due to more efficient and effective models, EDVR-1f and EDVR-3f, are chosen to super-resolve the patch sequence with stationary objects

Table 2: **The performance of BasicVSR and Boosted BasicVSR trained on the DTVIT-Train dataset.**

Method	Average PSNR on DTVIT
BasicVSR	33.36 dB
Boosted BasicVSR	33.73 dB

and background, the proposed **Boosted EDVR** can achieve better results with less inference time than EDVR. Although **Boosted BasicVSR** takes little more time than BasicVSR due to decomposed overlapping patches, it still faster than the proposed **Boosted EDVR** and EDVR.

### 6 More Quantitative Results

All the models in the manuscripts are trained on the REDS dataset (Nah et al. 2019), which is a first-person video dataset with consistent movement. To verify the proposed method can not only solve the generalization problem but also can enhance the effectiveness of the propagation branches, we apply the proposed Patch-based Dynamic Propagation (PDP) branch to the BasicVSR trained on the DTVIT-Train to see whether the improvement can be obtained. As shown in Table 2, the BasicVSR trained on the DTVIT-Train will not suffer of generalization problem and achieves favorable on the DTVIT dataset. In the meanwhile, the proposed **Boosted BasicVSR** can still outperform the BasicVSR without any training process, which demonstrates that the PDP can effectively improve the performance of existing non-local propagation-based VSR algorithms. It is noted that 0.37 dB performance gain based on a SotA method is remarkable for a plug-and-play method without taking more inference time.

### 7 More Qualitative Results

In this section, we provide additional qualitative results to clearly show the effectiveness of the proposed method. From Fig. 6 to Fig. 12, we select one scene from each type of the DTVIT dataset and show the super-resolving results of different methods. Besides, we also collect some real-world videos with low-quality and low-resolution to evaluate the performance of the state-of-the-art VSR algorithms on realistic scenes (Fig. 13). For the figures from the DTVIT dataset, we present the results from bicubic upsampling, EDVR and BasicVSR trained on the REDS dataset, the proposed **Boosted BasicVSR** and **Boosted EDVR**. For the fig-

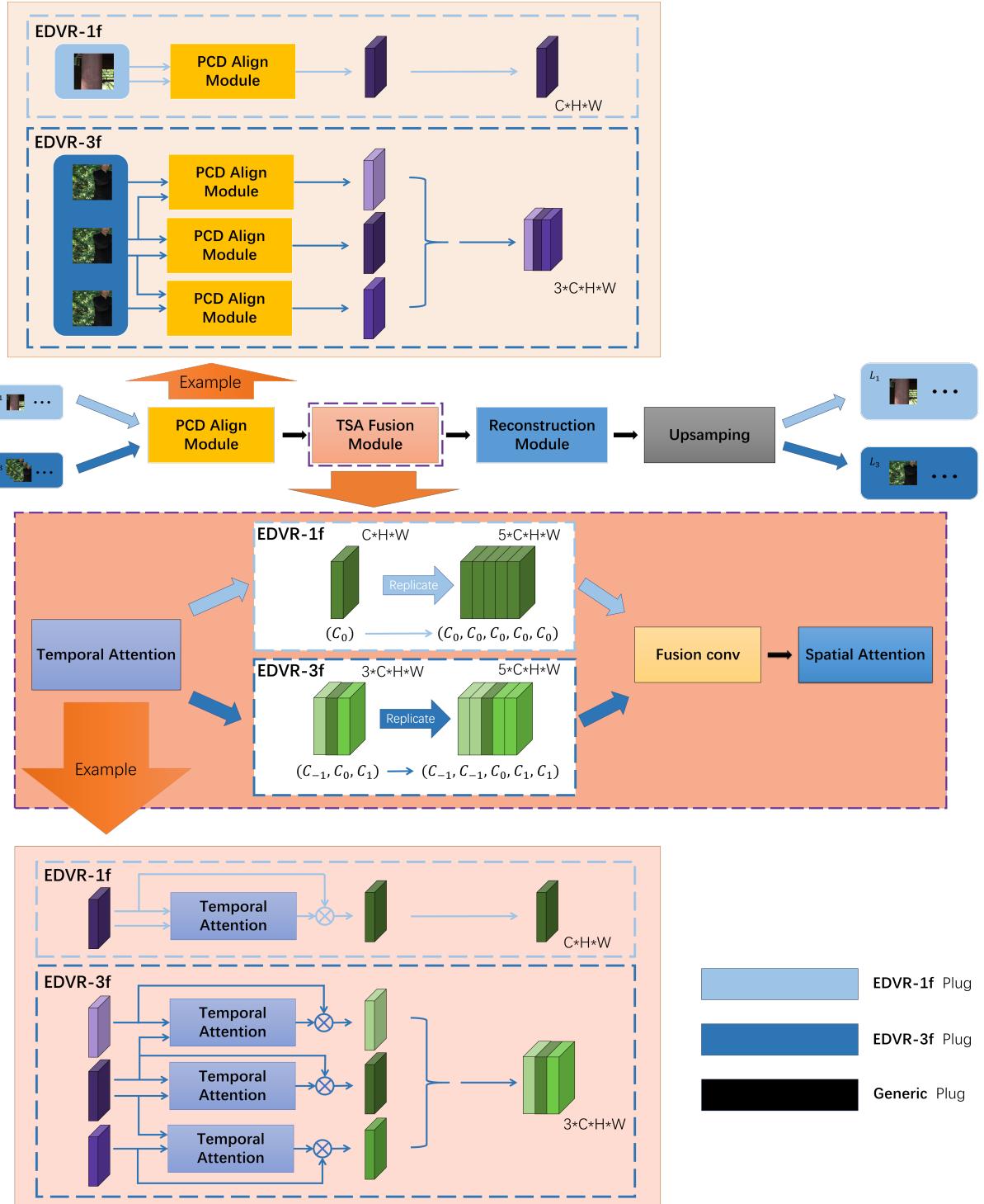


Figure 3: Details of the proposed EDVR-1f and EDVR-3f.

ures from the real-world videos, we only present the results from EDVR, BasicVSR, **Boosted BasicVSR**, and **Boosted EDVR**.

All the qualitative comparisons demonstrate that the proposed **Boosted BasicVSR** and **Boosted EDVR** are able to

reconstruct images with more details by effectively exploiting the temporal information.

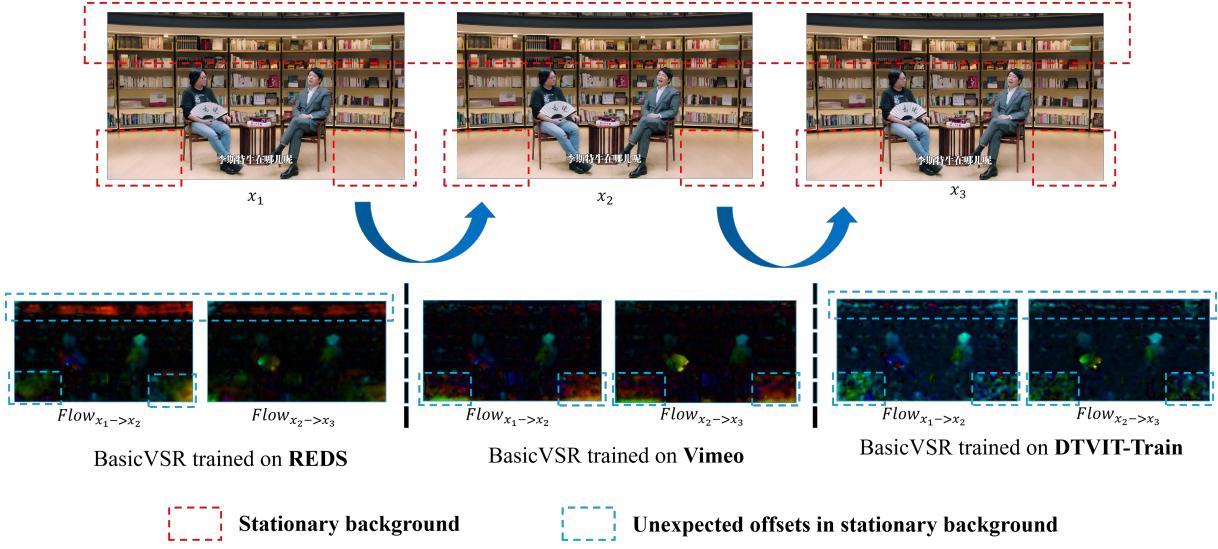


Figure 4: Visualization of the estimated optical flow in BasicVSR.

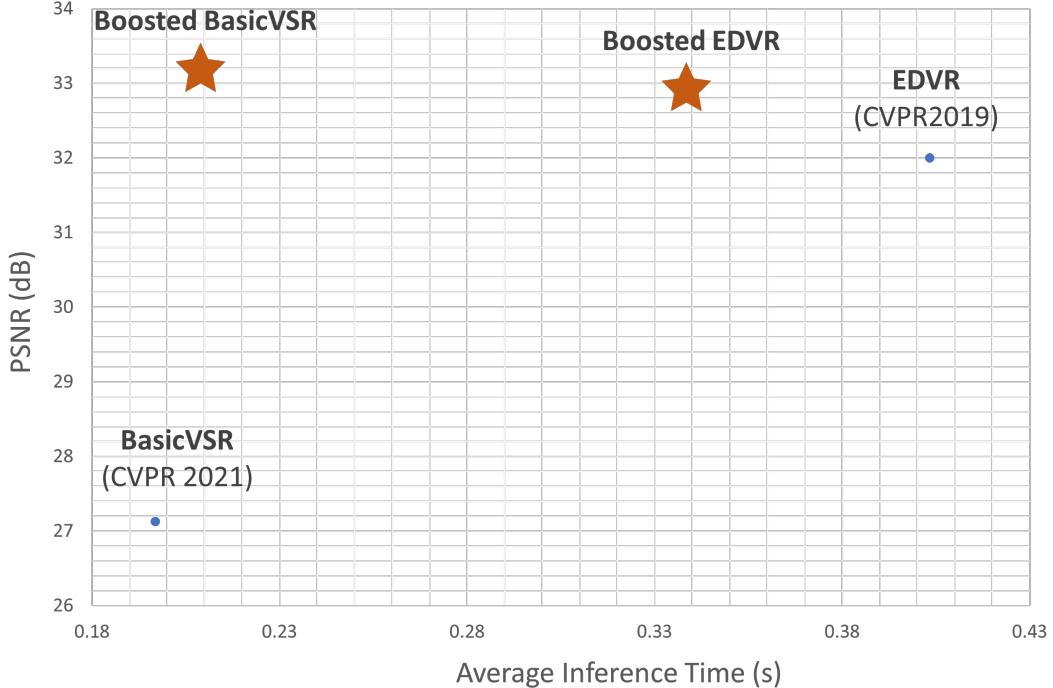


Figure 5: Speed and performance comparison. All the models are trained on the REDS (Nah et al. 2019) dataset and tested on the DTVIT dataset.

## References

- Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. BasicVSR: The search for essential components in video super-resolution and beyond. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4947–4956.
- Nah, S.; Baik, S.; Hong, S.; Moon, G.; Son, S.; Timofte, R.; and Mu Lee, K. 2019. 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*, 0–0.
- Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4161–4170.
- Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable

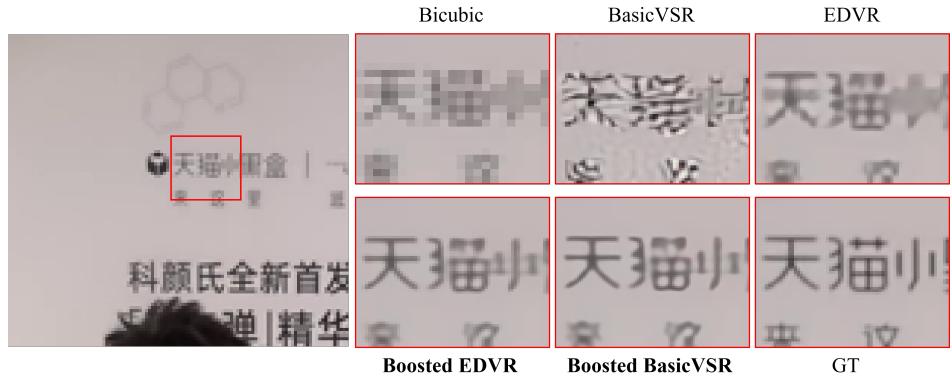


Figure 6: Qualitative results of the live streaming scene on the DTVIT dataset. Best viewed on a high-resolution display.

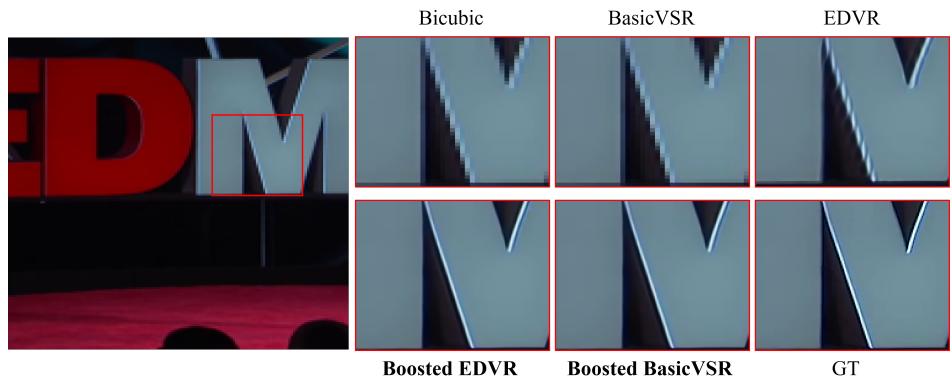


Figure 7: Qualitative results of the TV program scene on the DTVIT dataset. Best viewed on a high-resolution display.

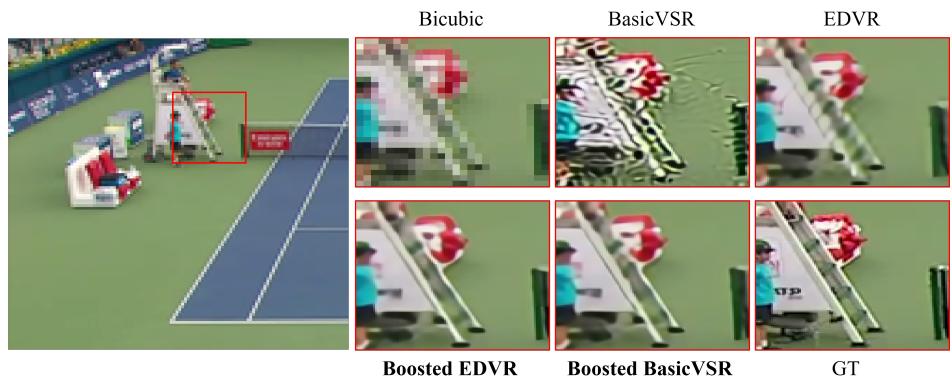


Figure 8: Qualitative results of the sports live scene on the DTVIT dataset. Best viewed on a high-resolution display.

convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

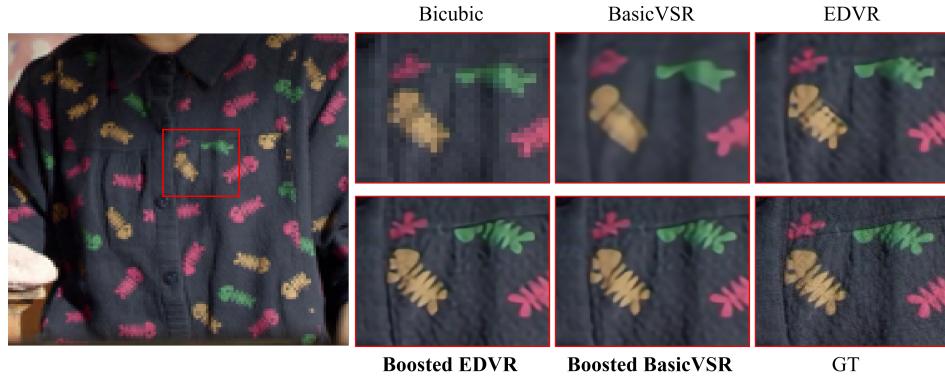


Figure 9: **Qualitative results of the movie and television play scene on the DTVIT dataset.** Best viewed on a high-resolution display.

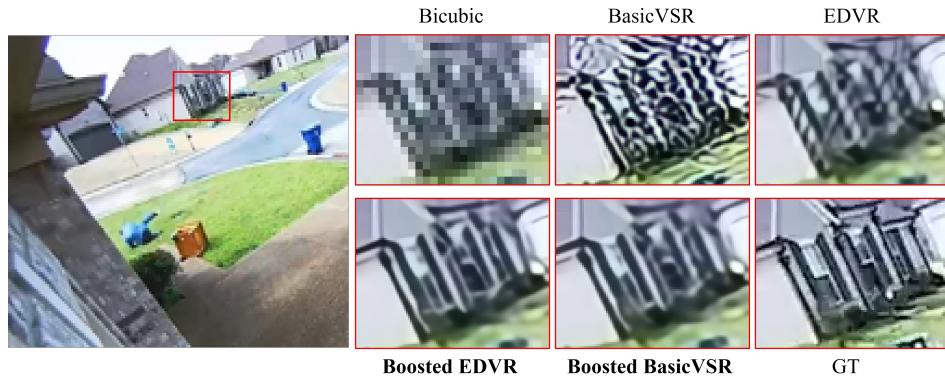


Figure 10: **Qualitative results of the surveillance video scene on the DTVIT dataset.** Best viewed on a high-resolution display.

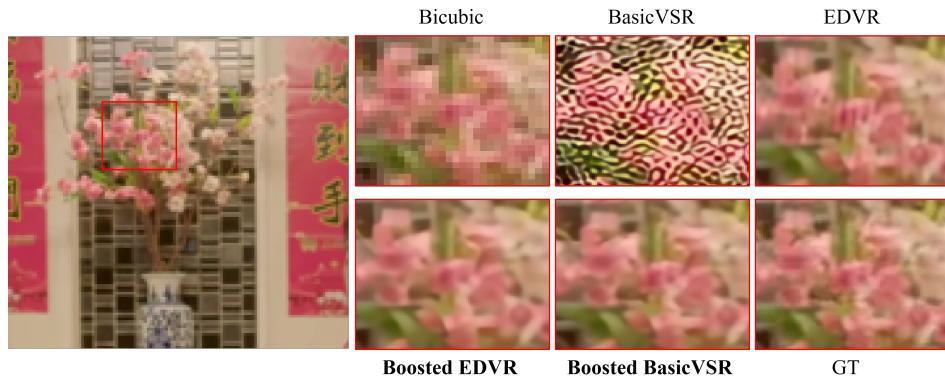


Figure 11: **Qualitative results of the advertisement scene on the DTVIT dataset.** Best viewed on a high-resolution display.

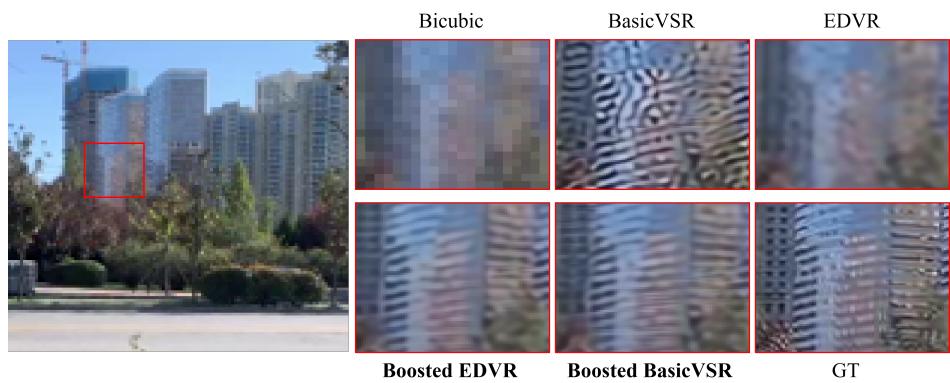


Figure 12: **Qualitative results of the first-person video scene on the DTVIT dataset.** Best viewed on a high-resolution display.

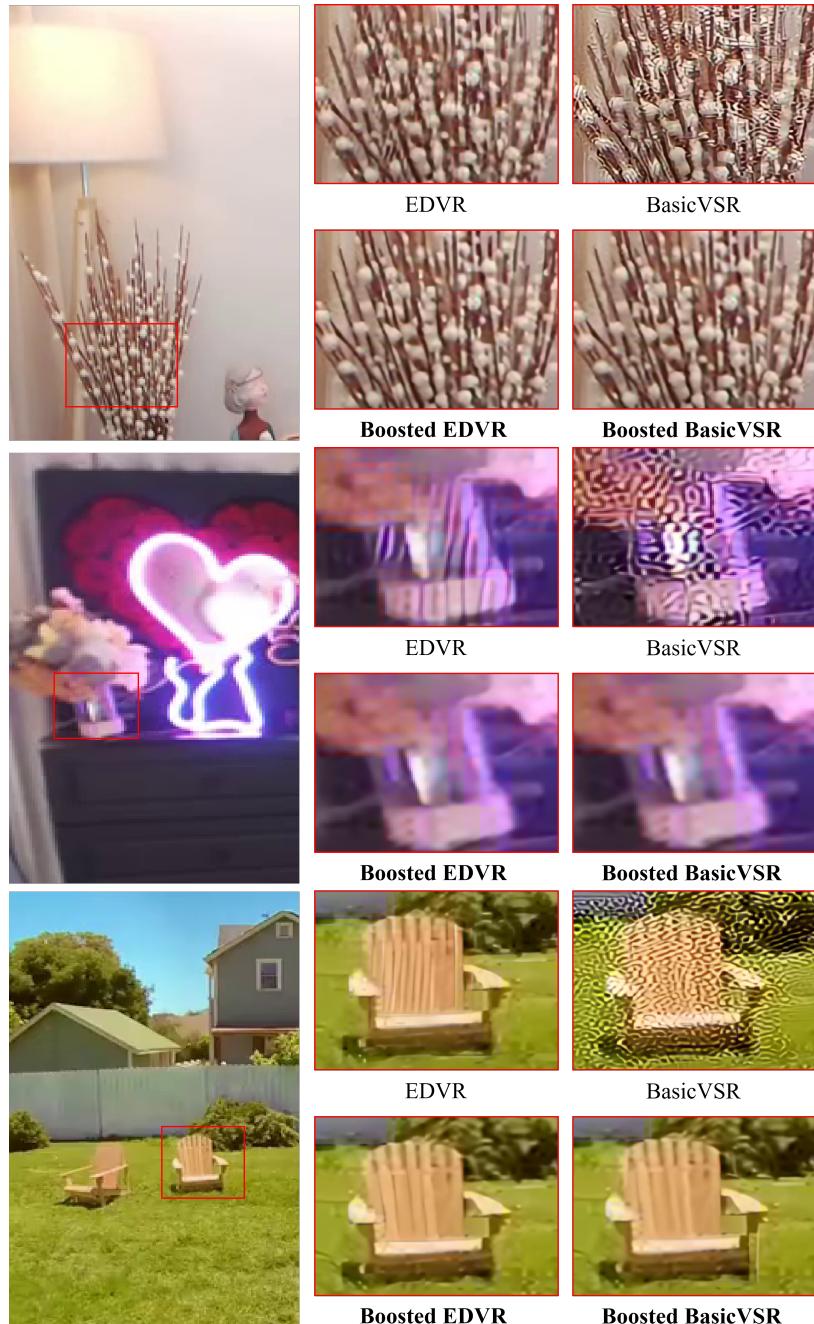


Figure 13: **Qualitative comparison on the real-world videos.** Best viewed on a high-resolution display.