Memory-Augmented Non-Local Attention for Video Super-Resolution

Jiyang Yu[1], Jingen Liu[1], Liefeng Bo[2], Tao Mei[3]
[1]JD AI Research, Mountain View, USA,
[2]JD Finance America Corporation, Mountain View, USA
[3]JD AI Research, Beijing, China
{jiyang.yu, jingen.liu, liefeng.bo, tmei}@jd.com

Abstract

In this paper, we propose a novel video super-resolution method that aims at generating high-fidelity high-resolution (HR) videos from low-resolution (LR) ones. Previous methods predominantly leverage temporal neighbor frames to assist the super-resolution of the current frame. Those methods achieve limited performance as they suffer from the challenge in spatial frame alignment and the lack of useful information from similar LR neighbor frames. In contrast, we devise a cross-frame non-local attention mechanism that allows video super-resolution without frame alignment, leading to be more robust to large motions in the video. In addition, to acquire the information beyond neighbor frames, we design a novel memory-augmented attention module to memorize general video details during the super-resolution training. Experimental results indicate that our method can achieve superior performance on large motion videos comparing to the state-of-the-art methods without aligning frames. Our source code will be released.

1. Introduction

Video super-resolution task aims to generate high-resolution videos from low-resolution input videos and recover high frequency details in the frames. It is attracting more attention due to its potential application in online video streaming services and the movie industry.

There are two major challenges in the video super-resolution tasks. The first challenge comes from the dynamic nature of videos. To ensure temporal consistency and improve visual fidelity, video super-resolution methods seek to fuse information from multiple neighbor frames. Due to the motion across the frames in the video, neighbor frames need to be aligned before fusion. Recent video super-resolution works have proposed various ways for aligning neighbor frames to the current frame, either by explicit warping using optical flow [2, 17, 20, 27] or learning implicit alignment using deformable convolution [28, 31]. However, the quality of these works highly depends on the accuracy of spatial alignment of neighbor frames, which is difficult to achieve in videos with large motions. This hinders the application of existing video super-resolution methods in real-world videos such as egocentric sports videos (see our Parkour dataset in Sec. 4.1), and some videos from animation, movies and vlogs (see additional examples in supplementary material).

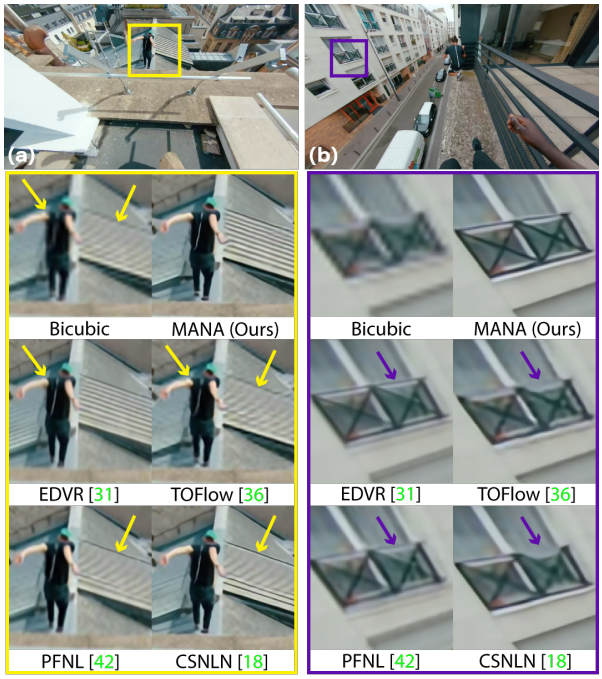

Figure 1. Our memory-augmented cross-frame non-local attention approach is robust to large motion videos (first row). Our method reconstructs visually pleasing details on repetitive patterns (left example) and thin structures (right example) while the state-of-the-art super-resolution methods requiring frame alignment (EDVR [31] and TOFlow [36]) and regular non-local attention (PFNL [42] and CSNLN [18]) fails in these cases.

The second challenge comes from the irreversible loss of high-frequency detail and the lack of useful information in the low-resolution video. Recent learning based single image super-resolution works [5, 12, 13, 16, 18, 25, 29, 33, 37, 48] have intensively studied the visual reconstruction from low-resolution images by learning general image

prior to help recover high-frequency details or transferring texture from a high-resolution reference image. Since these methods do not guarantee temporal consistency in the visual appearance, they usually generate results inferior to that of video super-resolution methods using neighbor frame information. However, in the video super-resolution task, the neighbor frames are largely similar to each other and the benefits from fusing them are limited. For large motion videos, the neighbor frames become less similar. In this scenario, the correlation among neighbor frames also becomes smaller, and the video super-resolution essentially degrades to the single image super-resolution since it cannot find any useful information by mining neighbor frames.

To address these issues, we propose a memory-augmented non-local attention framework for video super-resolution. Our method is a deep learning based method. Taking a set of consecutive low-resolution video frames as inputs, our network produces the high-resolution version of the temporal center frame by referring to the information from its neighbor frames. Since consecutive frames share a large portion of visual contents, this scheme implicitly ensures the temporal consistency in the result.

To solve the frame-alignment challenge, we design the *Cross-Frame Non-local Attention* module which allows us to fuse neighbor frames without aligning them towards the current frame. Although conventional non-local attention can capture temporally and spatially long-distance correspondences, it requires computing pair-wise correlation between each pixel in the query and key. This imposes a large burden on GPU memory since in the video super-resolution case down-sampling the video like Wang et al. [32] and losing more high-frequency detail is not desired. To make non-local attention practical in video super-resolution, in the cross-frame non-local attention module, we only query the current frame pixel within its 9×9 spatial neighborhood in the neighbor frames. Furthermore, instead of using softmax normalized correlation matrix to combine the value tensor like it is done in traditional non-local attention, we only sample the most correlated pixel in the value tensor, namely *one-hot attention*. Our one-hot non-local attention is effective especially for videos with large motions. In Fig. 1(a), while the state-of-the-art video super-resolution method EDVR [31] and TOFlow [36] fails due to fusing misaligned frames, our method reconstruct sharp details like the stripes on the roof and the waving arm. We provide complete verification of the effectiveness of our one-hot non-local attention framework in Sec. 4.

To solve the challenge of the lack of information, we seek to fuse useful information beyond the current video. This means that the network should *memorize* previous experiences in super-resolving other videos in the training set. Based on this principle, we introduce a *Memory-Augmented Attention* module to our network. In this module, we maintain a 2D memory bank which is completely learned during the video super-resolution training. The purpose of this module is to summarize the representative local details in the entire training set and use them as an external reference for super-resolving the current video frame. To our experience, by introducing the memory bank mechanism, our work is the first video super-resolution method that incorporates information beyond the current video. With the help of the memory-augmented attention module, our method can recover details that are missing in the low-resolution video like the balcony railings in Fig. 1(b).

In this paper, our contributions include the follows:

Cross-frame non-local attention. We introduce a novel cross-frame non-local attention that liberates the video super-resolution from the error-prone frame alignment process. This design makes our method robust to videos with large motions. (See Sec. 3.2)

Video super-resolution beyond current video. We proposed a novel memory-augmentation mechanism in video super-resolution, which memorizes previous experiences during the training process and uses the memory to assist current video super-resolution. (See Sec. 3.3 and Sec. 3.4).

2. Related Work

Single Image Super-Resolution Early image super-resolution works resort to image processing algorithms [24, 39, 40, 41]. Recent works in deep learning have been proved to obtain superior results in image super-resolution due to the ability to learn prior of high-resolution images. SRCNN proposed by Dong et al. [5] first introduces a convolutional neural network in image super-resolution. Kim et al. further develop VDSR [12] and DRCN [13] and explore deeper residual networks and recursive structures. ESPCN [22] encode the low-resolution image into multiple sub-pixel channels and upscale to a high-resolution image by shuffling the channels back in the spatial domain. This idea was widely used in recent super-resolution works. Other approaches using CNN includes pyramid strucutre (LapSRN [15]), recursive residual network (DRRN [26]), dense skip connections (SRDenseNet [30] and RDN [47]), and adversarial networks (SRGAN [16], EnhanceNet [21], ESRGAN [34] and GLEAN [3]).

Video Super-Resolution Video super-resolution typically generate better result than single image super-resolution thanks to the extra information from neighbor frames. The main focus of video super-resolution works is how to correctly fuse auxiliary frames in the presence of dynamic contents and camera motion. Some methods explicitly use optical flow (VESPCN [2], FRVSR [20], SPMC [27], TOFlow [36] and BasicVSR/IconVSR [4]) or homography (TGA [10]) to align neighbor frames. However, estimating accurate optical flow/transformation is challenging when the motion between the neighbor frame and current frame is large. Having observed this limitation, recent methods start to explore techniques to bypass alignment or implicitly align frames. Jo et al. proposed DUF [11] that learns dynamic upsampling filters that combine the entire spatial

neighborhood of a pixel in the auxiliary frames. TDAN [28] and EDVR [31] use deformable convolution layer to sample neighbor frames according to the estimated kernel offsets. However, these methods essentially still learn the spatial correspondence across frames. As we will show in Sec. 4, in large motion cases, the results from these methods are unsatisfactory. Unlike any previous video super-resolution methods, our method finds the pixel correspondence in an unstructured fashion by applying non-local attention.

Non-local Attention in Super-Resolution Attention mechanism has proven to be effective in various computer vision tasks[6, 9, 19, 43, 45, 46]. Non-local neural networks proposed by Wang et al.[32] capture pixel-wise correlations within a video segment, making temporally and spatially long distance attention possible. Recent image super-resolution methods using non-local attention includes CSNLN [18], RNAN [46] and TTSR [38]. Video super-resolution method PFNL [42] also utilize self-attention over a set of consecutive video frames. However, directly applying non-local attention requires storing pair-wise correlation between query and key. In the video super-resolution task, the size of the correlation matrix grows quadratically with the total number of pixels in the video segment and becomes intractable when the input frame size is large. Moreover, a larger number of pixels will potentially degrade the performance of non-local attention as we will discuss in Sec. 3.2. Our work performs one-hot non-local attention within the patch enclosing a query pixel and only selects the most correlated pixel in the neighbor frames. This approach greatly reduced the GPU memory usage and generate better results comparing to that of PFNL [42].

Memory models Neural networks with memory show their potential in natural language processing [1, 23], image classification [50] and video action recognition [8]. These works augment their model with an explicit memory bank that can be updated or read during the training. Inspired by these works, we design a memory-augmented attention module to incorporate previous knowledge gained from super-resolving other videos. In Sec. 4, we will show that the memory module provides a significant boost in the performance of video super-resolution.

3. Methodology

3.1. Overview

Fig. 2 demonstrates the structure of our video super-resolution network. The goal of our network is to super-resolve a single low-resolution frame $\mathbf{I}_t \in \mathbb{R}^{3\times H\times W}$, given the low-resolution temporal neighbor frames $\{\mathbf{I}_{t-\tau}, ..., \mathbf{I}_{t+\tau}\}$, where H and W are the video height and width respectively. To make the discussion more concise, we will use "current frame" to refer to $\mathbf{I}_t$ and "neighbor frames" to refer to $\{\mathbf{I}_{t-\tau}, ..., \mathbf{I}_{t+\tau}\}$. We will use $T = 2\tau + 1$ to represent the time span of neighbor frames. Note that neighbor frames include the current frame.

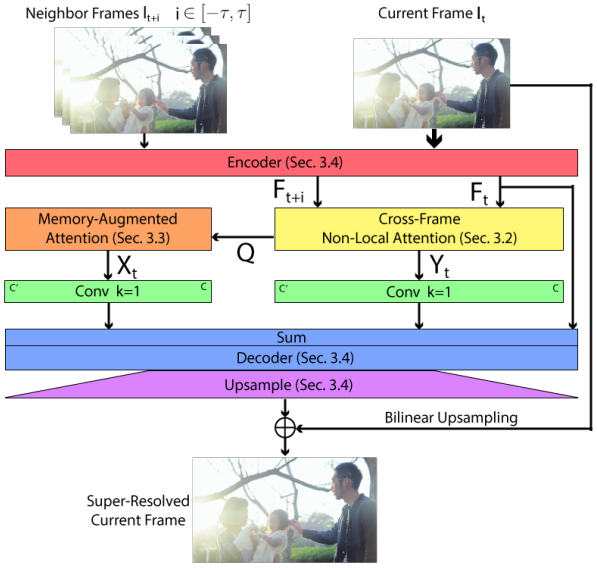

Figure 2. An overview of the structure of our video super-resolution network. The network super-resolves the current frame $\mathbf{I}_t$ using the neighbor frames $\mathbf{I}_{t-\tau}, ..., \mathbf{I}_{t+\tau}$ as the input. The cross-frame non-local attention aims at mining information from neighbor frames and the memory-augmented attention targets at memorizing experience in super-resolving other videos. The output of these modules are used as residual to enhance the details of a bilinearly upsampled low-resolution frame.

The first stage of our network embeds all the video frames into the same feature space by applying the same encoding network to each input frame. We denote the embedded features as $\{\mathbf{F}_{t-\tau}, ..., \mathbf{F}_{t+\tau}\} \in \mathbb{R}^{C\times H\times W}$, where C is the dimension of the feature space.

As we discussed in Sec. 1, our super-resolution process refers to both the current video and general videos. Based on this principle, we adapt the attention mechanism which allows us to query the pixels that need to be super-resolved in the keys consist of auxiliary pixels. Specifically, the second stage of our network includes two parts: *Cross-Frame Non-local Attention* and *Memory-Augmented Attention*.

Cross-Frame Non-local Attention aims to mine useful information from neighbor frame features. In this module, neighbor frame features are queried by the current frame feature and the most possible match will be selected as the output. We denote the output of the cross-frame non-local attention module as $\mathbf{X}_t \in \mathbb{R}^{C'\times H\times W}$, where $C' = C/2$ is the dimension of the embedding space of the cross-frame non-local attention module. The design of this module will be discussed in Sec. 3.2.

Memory-Augmented Attention maintains a global memory bank $\mathbf{M} \in \mathbb{R}^{C'\times N}$ to memorize useful information from general videos in the training set, where N represents an arbitrary number of entries in the memory bank. We use the current frame feature to query the memory bank directly. However, unlike the cross-frame non-local attention module in which the keys are embedded versions of neighbor frame features, the memory bank is completely learned.

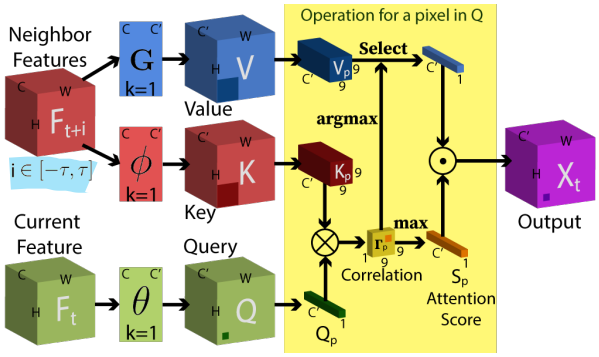

Figure 3. The cross-frame non-local attention module in our network. The size are marked on the edges of the tensors. The operation marked by the yellow box is done in parallel for each pixel $\mathbf{Q}_p$ in the query tensor $\mathbf{Q}$. Best viewed in PDF.

The output of this module is denoted as $\mathbf{Y}_t \in \mathbb{R}^{C' \times H \times W}$. This module will be discussed in Sec. 3.3.

Finally, the output of the cross-frame non-local attention module $\mathbf{X}_t$ and memory-augmented attention module $\mathbf{Y}_t$ are convolved by two different convolutional layers with kernel size 1 and added to the input current frame feature $\mathbf{F}_t$ as residuals. A decoder decodes the output of attention modules and an up-sampling module shuffles the pixels to generate a high-resolution residual. The residual adds details to the bilinearly up-sampled blurry low-resolution frame, resulting in a clear high-resolution frame.

3.2. Cross-Frame Non-local Attention

One of the major procedures in the conventional video super-resolution methods is to align the neighbor frames so that the corresponding pixels can be fused and improve the quality of the super-resolution of the current frame. To achieve the alignment, the typical approaches in video super-resolution works include optical flow [20, 36] and deformable convolution [28, 31]. However, aligning pixels according to color consistency is known to be a challenging task under large motion or illumination change. As a consequence, the inaccuracy in alignment will negatively impact the performance of video super-resolution. In our work, we seek to avoid this performance overhead. As we discussed in Sec. 2, the non-local attention [32] enables capturing temporally and spatially long distance correspondence. Therefore, the frame alignment can be omitted if non-local attention is used to query pixels of the current frame in neighbor frames.

The cross-frame non-local attention module is demonstrated in Fig. 3. We first normalize the input frame features using group normalization [35], resulting in the normalized neighbor frame features $\{\overline{\mathbf{F}}_{t-\tau}, ..., \overline{\mathbf{F}}_{t+\tau}\}$. In our non-local attention setup, the center feature $\overline{\mathbf{F}}_t$ is used as the query tensor, and neighbor frame features $\{\overline{\mathbf{F}}_{t-\tau}, ..., \overline{\mathbf{F}}_{t+\tau}\}$ serve as both the key and value tensors. The embedded version of query, key and value tensor are noted as $\mathbf{Q} \in \mathbb{R}^{C' \times H \times W}$, $\mathbf{K} \in \mathbb{R}^{C' \times T \times H \times W}$ and $\mathbf{V} \in \mathbb{R}^{C' \times T \times H \times W}$ in Fig. 3. In the traditional setup of non-local attention, the next step is to flatten the temporal and spatial dimension of $\mathbf{Q}$ and $\mathbf{K}$ to $\widehat{\mathbf{Q}} \in \mathbb{R}^{HW \times C'}$ and $\widehat{\mathbf{K}} \in \mathbb{R}^{C' \times HWT}$ and calculating the correlation matrix $\mathbf{\Gamma} = \widehat{\mathbf{Q}}\widehat{\mathbf{K}}$. However, the size of $\mathbf{\Gamma}$ is $HW \times HWT$. Even though the input to our network is low-resolution frames, the size of this matrix is large and grows quadratically when involving more neighbor frames. This makes the training difficult due to the limited GPU memory. Moreover, note that the number of columns in $\mathbf{\Gamma}$ depends on the frame size. With the same pre-trained network, normalizing HWT dimensions using softmax makes the columns in $\widehat{\mathbf{K}}$ that are most correlated with a row in $\widehat{\mathbf{Q}}$ less significant when the input frame size is larger. Intuitively, traditional non-local attention weakens the contribution of the most useful auxiliary pixels in the super-resolution case. In Sec. 4.3, we will show that traditional non-local attention did not benefit the video super-resolution method PFNL [42] which directly applies it to the entire group of neighbor frames.

To mitigate the GPU memory issue, we conduct non-local attention on each neighbor frames separately. Moreover, for each pixel in the query tensor, we conduct non-local attention only on its spatial neighbor region in the temporal neighbor frames. The insight is that video motion is limited during a short time period, and a pixel's correspondences in the neighbor frames lie in the neighborhood of its position. Specifically, we unfold each temporal slice of $\mathbf{K}$ with dimension $C' \times H \times W$ into 9×9 patches in the spatial domain. Therefore, for each pixel $\widehat{\mathbf{Q}}_p \in \mathbb{R}^{1 \times C'}$ in the query tensor $\widehat{\mathbf{Q}}$, the corresponding key tensor becomes $\mathbf{K}_p \in \mathbb{R}^{C' \times 81}$. As a result, the correlation matrix $\mathbf{\Gamma}_p = \widehat{\mathbf{Q}}_p\mathbf{K}_p$ has a dimension of 1×81. Note that $\mathbf{K}_p$ is different for each row of $\widehat{\mathbf{Q}}$; each $\mathbf{K}_p$ only represents the spatial neighborhood of the specific pixel in the current frame. By stacking the $\mathbf{\Gamma_p}$ for each pixel, we obtain the final correlation matrix $\mathbf{\Gamma}$.

To resolve the performance degradation issue in the traditional non-local attention, we reduce the non-local attention into an *one-hot* attention. Specifically, for each row in $\mathbf{\Gamma}$, we only select the largest entry as follows:

$$s_i = \max_j \mathbf{\Gamma}(i,j), \quad d_i = \operatorname*{argmax}_j \mathbf{\Gamma}(i,j) \tag{1}$$

where s_i is the attention score and d_i is the column coordinate in $\mathbf{K}_p$. Denote the *one-hot* attention results as $\mathbf{S} \in \mathbb{R}^{HW \times 1}$ and $\mathbf{D} \in \mathbb{R}^{HW \times 1}$ which consists of s_i and d_i respectively. We then select the value tensor $\mathbf{V}_p$ as follows:

$$\widehat{\mathbf{X}}_t = \mathbf{S} \cdot \mathbf{V}_p(\mathbf{D}) \tag{2}$$

In Eqn. 2, $\mathbf{V}_p \in \mathbb{R}^{C' \times 81}$ is the value tensor unfolded in the same way as $\mathbf{K}_p$. The operator $(\ )$ selects HW columns from $\mathbf{V}_p$ using the HW indices in $\mathbf{D}$. Then the selected HW columns are multiplied by HW attention scores in $\mathbf{S}$.

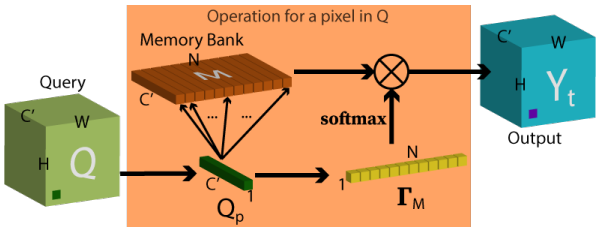

Figure 4. The memory-augmented attention module in our network. The operation marked by the orange box is done in parallel for each pixel $\mathbf{Q}_p$ in the query tensor $\mathbf{Q}$. Best viewed in PDF.

The $\widehat{\mathbf{X}}_t \in \mathbb{R}^{C' \times HW}$ is reshaped to $\mathbf{X}_t \in \mathbb{R}^{C' \times H \times W}$ as the output of the cross-frame non-local attention module.

3.3. Memory-Augmented Attention

Cross-frame non-local attention enables the fusion of the information from neighbor frames in the current video. However, the neighbor frames used in the attention are also low-resolution with similar content to the current frame. Therefore, the benefit from cross-frame non-local attention is limited. We seek to refer to more local detail information beyond the current video, which requires memorizing useful information from the entire training set. For this purpose, our network includes a memory-augmented attention module. The module maintains a global memory bank $\mathbf{M} \in \mathbb{R}^{C' \times N}$ which is learned as parameters of the network. We use regular non-local attention to query current frame features $\widehat{\mathbf{Q}}$ in the global memory bank $\mathbf{M}$, i.e. the correlation matrix is $\mathbf{\Gamma}_M = \widehat{\mathbf{Q}}\mathbf{M} \in \mathbb{R}^{HW \times N}$. Finally, we obtain the output

$$\widehat{\mathbf{Y}}_t = softmax(\mathbf{\Gamma}_M)\widehat{\mathbf{M}} \tag{3}$$

where $\widehat{\mathbf{M}} \in \mathbb{R}^{N \times C'}$ is the transposed version of the memory bank $\mathbf{M}$. Similar to the cross-frame non-local attention module, we reshape $\widehat{\mathbf{Y}}_t \in \mathbb{R}^{HW \times C'}$ to $\mathbf{Y}_t \in \mathbb{R}^{C' \times H \times W}$ as the output of the memory-augmented attention module.

3.4. Implementation Details

Training Set Vimeo90K is a large-scale video dataset proposed by Xue et al. [36]. Following recent super-resolution methods TOFlow [36], TDAN [28] and EDVR [31], we use the training set of Vimeo90K to train our network. Each video clip in Vimeo90K consists of 7 consecutive frames. We use the center frame as the current frame to be super-resolved. All 7 frames are used as the neighbor frames.

Network Structure Besides the structures of cross-frame non-local attention and memory-augmented attention module shown in Fig. 2, we demonstrate the structure of other basic building blocks in Fig. 5. The residual blocks (Fig. 5(a)) are used to build the frame encoder and decoder. The frame encoder and decoder are the concatenation of 5 residual blocks and 40 residual blocks respectively. The structure of the up-sampling block is shown in Fig. 5(b). In this paper, we focus on 4x video super-resolution task.

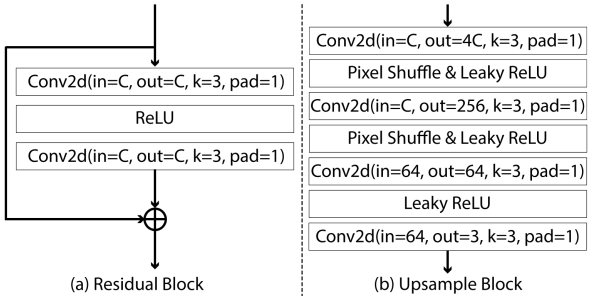

Figure 5. Basic building blocks in our network. (a) Residual blocks are used to build the encoder and decoder. (b) Upsample block shuffles pixels in different channels into a high-resolution frame.

Figure 6. Video stills from the *Parkour dataset*. Due to the large camera motion in this dataset, it is challenging for exisitng video super-resolution methods.

The up-sampling block is built by 2 pixel shuffle blocks, each up-sample the feature map by 2 using the pixel shuffle operation defined in ESPCN [22]. We use $C = 128$ for all experiments in this paper.

Training Procedure We implement our network in PyTorch [7] and use Adam optimizer [14] with $\beta_1 = 0.5$ and $\beta_2 = 0.99$ for training. The weight of the last convolutional layers of the cross-frame non-local attention module and the memory-augmented attention module is initialized to zero. The training of our network consists of three stages.

In the first stage, we fix the memory-augmented attention module and train the rest part of the network for 90,000 iterations at the learning rate of 10^{-4}. The loss function used is $L_1 = \|\mathbf{O}_t - \mathbf{G}_t\|_1$, where $\mathbf{O}_t$ stands for the output super-resolved current frame and $\mathbf{G}_t$ is the ground truth high-resolution frame.

In the second stage, we fix the network weights except for the memory-augmented attention module. The loss function $L_2 = \|\mathbf{Y}_t - \mathbf{Q}\|_1$ focus on training the memory bank. Note that the training process optimizes the memory bank $\mathbf{M}$ so that a query $\mathbf{Q}$ can be represented by the combination of the columns in $\mathbf{M}$ as accurate as possible. This is essentially clustering and summarizing the most representative general pixel features in the encoded space. We train this stage for 30,000 iterations at the learning rate of 10^{-4}.

In the final stage, we fine-tune the entire network using L_1 for 30,000 iterations at the learning rate of 10^{-5}.

4. Experiments

In this section, we compare our work with recent state-of-the-art video super-resolution (VSR) and single image super-resolution (SISR) methods. We select com-

Figure 7. Visual comparison on the Parkour dataset, Vimeo90K [36] dataset and Vid4 [20] dataset. Example (a), (b), (c), (d) and (e) are selected from the large motion Parkour dataset. Example (f) and (g) are selected in the Vimeo90K [36] dataset. Example (h) is from Vid4 [20] dataset. We mark the inset locations on the video stills on the left. To make our discussion clearer, we add arrows pointing to the region that we will be discussing in Sec. 4.2. Best viewed in PDF.

parison methods based on their approaches to the super-resolution problem: VSR via explicit frame alignment (TOFlow [36] and TGA [10]), VSR via implicit frame alignment (EDVR [31]), VSR via regular non-local attention (PFNL [42]) and SISR via regular non-local attention (CSNLN [18]) applied to each video frame individually. Similar to other VSR works, in this paper, we focus on the 4x scaling case for all the comparisons shown in this section. To obtain the low-resolution input, we use bicubic down-sampling on the ground truth high-resolution frames. According to our experiment, PFNL [42] and TGA [10] introduce serious aliasing artifacts to the results using bicubic down-sampled video. To make the comparison fair, for PFNL [42] and TGA [10], we apply Gaussian blur to the ground truth frames before down-sampling following the procedure in their papers. Unless otherwise stated, our results shown in this section are generated with the memory size of $N = 256$ in the memory-augmented attention module. We conduct the experiment on a desktop computer with an NVIDIA 2080Ti GPU. The average processing speed of our network is 251ms per 960x540 HR frame.

4.1. Datasets and Metrics

As discussed in Sec. 1, the cross-frame non-local attention in our method enables VSR without frame alignment. To validate the robustness of our method to large motion videos, we randomly collect 14 parkour video clips from the Internet. Parkour is a form of extreme sport focusing on passing obstacles in a complex environment by running, climbing, and jumping. Usually taken using egocen-

	(a) Parkour Dataset			(b) Vimeo90K Dataset [36]			(c)Vid4 Dataset [20]			(d)SPMC Dataset [27]		
	PSNR in dB	SSIM	LPIPS	PSNR in dB	SSIM	LPIPS	PSNR in dB	SSIM	LPIPS	PSNR in dB	SSIM	LPIPS
Bicubic	29.51 (+3.97)	0.8712	0.3101	29.75 (+4.96)	0.8476	0.2948	22.34 (+2.81)	0.6131	0.5186	25.67 (+3.55)	0.7241	0.4270
MANA (Ours)	**33.48**	**0.9356**	**0.1241**	**34.71**	**0.9261**	**0.1101**	**25.15**	**0.7796**	**0.2744**	**29.22**	**0.8458**	**0.2119**
EDVR [31]	31.61 (+1.87)	0.9113	0.1900	35.68 (-0.97)	0.9372	0.1019	25.79 (-0.64)	0.8063	0.2489	27.98 (+1.24)	0.8109	0.2715
TOFlow [36]	32.35 (+1.13)	0.9197	0.1804	32.96 (+1.75)	0.9041	0.1451	24.41 (+0.74)	0.7435	0.3340	28.55 (+0.67)	0.8327	0.2661
TGA [10]	31.14 (+2.34)	0.9033	0.2224	35.03 (-0.32)	0.9310	0.1013	25.36 (-0.21)	0.7949	0.2834	29.06 (+0.16)	0.8449	0.2390
PFNL [42]	32.04 (+2.44)	0.9189	0.2244	31.86 (+2.85)	0.8959	0.2012	25.01 (+0.14)	0.7788	0.3204	28.27 (+0.95)	0.8270	0.3100
CSNLN [18]	32.93 (+0.55)	0.9275	0.1357	33.55 (+1.16)	0.9091	0.1338	24.09 (+1.06)	0.7202	0.3425	28.79 (+0.43)	0.8275	0.2343

Table 1. Quantitative comparison on (a) Parkour dataset, (b) Vimeo90K [36] dataset and (c) Vid4 dataset. The metrics used are PSNR, SSIM and LPIPS. Larger numbers indicate better results for PSNR and SSIM, smaller numbers indicate better results for LPIPS. We also note the PSNR gain of our method comparing to other methods; a positive gain means that our method performs better than the corresponding method.

	PSNR	SSIM	LPIPS		PSNR	SSIM	LPIPS
MANA	**38.62**	**0.9606**	**0.1586**	TGA	38.26	0.9588	0.1570
EDVR	38.33	0.9544	0.1641	PFNL	35.90	0.9449	0.1985
TOFlow	36.55	0.9471	0.1902	CSNLN	37.79	0.9523	0.1780

Table 2. Quantitative comparison on top 6% large motion videos in Vimeo90K [36] dataset.

tric wearable cameras, parkour videos are typical examples in the real-world where large camera motions are everywhere. Example video stills from the *Parkour dataset* are shown in Fig. 6. We further evaluate our method on regular small motion videos using Vimeo90K [36] test set and Vid4 [20]. For all the test sets, we use the average PSNR and SSIM [49] on the RGB channels to quantitatively evaluate the performance of the methods. In addition, we apply LPIPS [44] to evaluate the perceptual similarity between the super-resolved frames and the ground truth high-resolution frame. Since the performance can be different across computation platforms and the quantitative metric calculation might be different in these works, we re-run their code and calculate the metrics in the same way on the same computer.

4.2. Visual Comparisons

The visual comparison of the examples from the Parkour dataset, Vimeo90K [36] dataset and Vid4 [20] is shown in Fig. 7. To make the discussion concise, we label the ID at the bottom left of each video. We also added arrows pointing at the regions we will be discussing.

Example (a), (b), (c), (d) and (e) are selected from the Parkour dataset. These examples contain large motion and are challenging to existing VSR methods. Our method can reconstruct repetitive patterns like Example (a) and (b), while explicit frame alignment methods TOFlow [36] and TGA [10] fail due to the inaccurate frame alignment. EDVR [31] result is more blurry than our result in example (a) and (b), and *the blurry issue is more visible when viewed in dynamics as we will show in the supplementary video*. This indicates that the deformable convolution alignment cannot handle the alignment with large frame displacement. The VSR and SISR methods PFNL [42] and CSNLN [18] using non-local attention also suffer from the blurry issue, potentially due to the non-local attention performance degradation problem discussed in Sec. 3.2.

Example (c) focuses on general details of objects. The frame-aligning VSR methods introduce ghosting artifacts (EDVR [31]) and deformation (TOFlow [36] and TGA [10]) due to the inaccurate alignment. PFNL [42] and CSNLN [18] results have less detail than ours, indicating that our one-hot non-local attention improves the quality of regular non-local attention. Example (d) focuses on human face shape and details. As shown in the bicubic result, the original facial details are completely lost due to the downsampling. Our method reconstructs visually pleasing details of human faces thanks to the memory-augmented module, while the comparison methods introduce blur (EDVR [31], TOFlow [36] and PFNL [42]) or reconstruct shapes that do not look like a human (TGA [10] and CSNLN [18]).

Example (e) contains thin structures. Similar to examples (a) and (b), failure in frame alignment has negatively affected the VSR methods. In this case, the performance of VSR methods EDVR [31], TOFlow [36] and TGA [10] are even worse than the SISR method CSNLN [18]. Our method with one-hot non-local attention can achieve a comparable result to CSNLN [18] in this example since our network does not require frame alignment.

As we will discuss in Sec. 4.3, the overall average quantitative metric score of our method is slightly inferior to that of EDVR [31] and TGA [10] in the Vimeo90K [36] and Vid4 dataset [20] which are relatively easy for frame aligning VSR methods. However, a larger deviation to the ground truth does not always indicate worse performance. Example (f) and (g) are selected from the Vimeo90K [36] test set. Our method tends to produce visually sharper results than EDVR [31] and TGA [10], which is often more preferred in the video super-resolution task. Example (h) is a widely used example in Vid4 [20]. In this example, EDVR [31] and TGA [10] generate the best results, but our

result is comparable to their results.

To further evaluate the robustness of our method in real-world scenarios, we provide additional results in the supplementary material. These videos are arbitrarily selected from different types of videos, covering animation, movies, and vlogs. We encourage the readers to read the supplementary PDF for a more complete visual comparison.

4.3. Quantitative Comparisons

In Table 1(a), we compare the average PSNR and SSIM [49] of our method with the comparison methods on the Parkour dataset. In this table, larger PSNR and SSIM and smaller LPIPS loss indicate better results. We mark the best result in red and the second best result in blue.

The videos in the Parkour dataset have extremely large motions, making the accurate alignment of the frames difficult. Among the comparison methods, TOFlow [36] explicitly estimates the optical flow for warping neighbor frames; TGA [10] uses homography to align neighbor frames; EDVR [31] implicitly align frames using learned kernel offset for deformable convolution. The performances of these methods are even inferior to that of the SISR method CSNLN [18] in the Parkour dataset, since fusing misaligned frames often cause blurry or ghosting artifacts in the result. This indicates that the large motions have a negative effect on the performance of traditional VSR methods.

The comparison method PFNL [42] uses pair-wise non-local attention on all the pixels in the entire segment. As discussed in Sec. 3.2, the traditional non-local attention is difficult to train due to the large GPU memory consumption and the performance degradation with a large number of pixels. Although the performance of their method is better than frame alignment methods EDVR [31] and TGA [10], using traditional non-local attention on all video segments is worse than applying it only on a single frame like CSNLN [18]. Our cross-frame non-local attention mechanism significantly improves the performance of non-local attention by introducing the one-hot attention in the video super-resolution task.

In addition, we provide the quantitative comparison on the Vimeo90K dataset [36], the Vid4 dataset [20] and the SPMC dataset [27] in Table 1(b), (c) and (d) respectively. The metrics and color labels are the same as Table 1(a). For these general small motion videos datasets, we also achieve better results than the explicit optical flow alignment VSR method TOFlow [36] and the other non-local attention super-resolution methods PFNL [42] and CSNLN [18]. Note that the single image non-local attention method CSNLN [18] also outperforms video non-local attention method PFNL [42] in the Vimeo90K dataset. The PSNR and SSIM value of our method is slightly inferior to that of EDVR [31] and TGA [10] in the Vimeo90K and Vid4 datasets. EDVR [31] has a 0.97dB and 0.64dB PSNR gain to our method in Vimeo90K and Vid4 respectively. TGA [10] has a 0.32dB and 0.21dB PSNR gain to our method in Vimeo90K and Vid4 respectively. The perceptual quality measured by the LPIPS [44] are similar among EDVR [31], TGA [10] and our method, with around 0.02 differences. However, for the large motion examples in the Parkour dataset, our method has a larger PSNR gain (1.87dB and 2.34dB) and LPIPS gain (0.0659 and 0.0983) in the performance comparing to EDVR [31] and TGA [10]. Moreover, in Table. 2, we computed the optical flow for videos in the Vimeo90K [36] test set and ranked them based on the average flow magnitude. It can be observed that even though EDVR [31] and TGA [10] is slightly superior on average, our method is actually better on the large motion videos in Vimeo90K [36].

We also note that EDVR [31] is biased towards Vimeo90K [36], given the significant performance drop in the other small motion video dataset SPMC [27]. This implies that our method is more robust than the comparison methods. We also provide the quantitative evaluation on the real-world videos in the supplementary material, which also supports the superior of our method in robustness.

4.4. Ablation Study

In Table 3, we quantitatively compare the performance of different configurations in our network. Specifically, we set the memory size N of the memory-augmented attention module to 128, 256, 512 and 1024. To verify the effectiveness of the memory-augmented attention module, we also experimented with the network with cross-frame non-local attention module only (labeled as *No_Mem* in Table 3). Among these configurations, $N = 256$ achieves the best result and is selected in the comparisons in Sec. 4.2 and Sec. 4.3. Using smaller memory ($N = 128$) results in slight performance degradation. The benefits saturate when using a larger memory ($N = 512$ and $N = 1024$), implying that the local details of low-resolution frames can be well represented in low-dimensional space. The performance of our network degrades without the memory-augmented attention module. However, solely using the cross-frame non-local attention module, our network outperforms comparison methods in the Parkour dataset and achieves comparable performance in the Vimeo90K dataset.

5. Conclusion

We present a network for video super-resolution that is robust to large motion videos. Unlike typical video super-resolution works, our network is able to super-resolve videos without aligning neighbor frames through a novel one-hot cross-frame non-local attention mechanism. Thanks to the memory-augmented attention module, our method can also utilize information beyond the video that is being super-resolved by memorizing details of other videos during the training phase. Our method achieves significantly better result in the large motion videos compared to the state-of-the-art video super-resolution methods. The performance of our method is slightly inferior in the videos

	Parkour Dataset			Vimeo90K Dataset [36]			Vid4 Dataset [20]			SPMC Dataset [27]		
	PSNR (dB)	SSIM	LPIPS	PSNR (dB)	SSIM	LPIPS	PSNR (dB)	SSIM	LPIPS	PSNR (dB)	SSIM	LPIPS
No_Mem	33.31	0.9343	0.1196	34.47	0.9232	0.1696	25.09	0.7742	0.3018	29.03	0.8389	0.2256
$N = 128$	33.44	0.9350	0.1254	34.65	0.9254	0.1886	25.13	0.7787	0.2859	29.24	0.8449	0.2148
$\boldsymbol{N = 256}$	**33.48**	**0.9356**	**0.1241**	**34.71**	**0.9261**	**0.1644**	**25.15**	**0.7796**	**0.2744**	**29.22**	**0.8458**	**0.2119**
$N = 512$	33.47	0.9354	0.1245	34.71	0.9261	0.1647	25.14	0.7797	0.2840	29.25	0.8449	0.2129
$N = 1024$	33.47	0.9354	0.1238	34.68	0.9257	0.1649	25.15	0.7797	0.2852	29.26	0.8452	0.2142

Table 3. Ablation study on the memory size in the memory-augmented attention module. The $N = 256$ is selected for the experiments shown in Sec. 4.2 and Sec. 4.3

that are relatively easy for frame aligning video super-resolution methods. We believe our method can be further improved by introducing pyramid structure into the cross-frame non-local attention to increase the perception field or extend the memory bank from 2D to higher dimension, but these ideas are left for future work.

References

[1] Nabiha Asghar, Lili Mou, Kira A. Selby, Kevin D. Pantasdo, Pascal Poupart, and Xin Jiang. Progressive memory banks for incremental domain adaptation. In *ICLR*, 2020. 3

[2] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, 2017. 1, 2

[3] Kelvin C.K. Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, 2021. 2

[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. In *CVPR*, 2021. 2

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 1, 2

[6] 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. In *ICLR*, 2021. 3

[7] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[8] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *ECCV*, 2020. 3

[9] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[10] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *CVPR*, 2020. 2, 6, 7, 8

[11] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, 2018. 2

[12] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 1, 2

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016. 1, 2

[14] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[15] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 2

[16] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1, 2

[17] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *ICCV*, 2017. 1

[18] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Humphrey Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, 2020. 1, 3, 6, 7, 8

[19] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *NeurIPS*, 2019. 3

[20] Mehdi Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018. 1, 2, 4, 6, 7, 8, 9

[21] Mehdi S. M. Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017. 2

[22] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 2, 5

[23] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *NeurIPS*, 2015. 3

[24] Jian Sun, Zongben Xu, and Heung-Yeung Shum. Image super-resolution using gradient profile prior. In *CVPR*, 2008. 2

[25] Libin Sun and James Hays. Super-resolution from internet-scale scene matching. In *ICCP*, 2012. 1

[26] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, 2017. 2

[27] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017. 1, 2, 7, 8, 9

[28] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally deformable alignment network for video super-resolution. In *CVPR*, 2020. 1, 3, 4, 5

[29] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, 2013. 1

[30] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *ICCV*, 2017. 2

[31] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPR*, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[32] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2, 3, 4

[33] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 1

[34] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018. 2

[35] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 4

[36] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019. 1, 2, 4, 5, 6, 7, 8, 9

[37] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 1

[38] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 3

[39] Jianchao Yang, Zhe Lin, and Scott Cohen. Fast image super-resolution based on in-place example regression. In *CVPR*, 2013. 2

[40] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution as sparse representation of raw image patches. In *CVPR*, 2008. 2

[41] Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing (TIP)*, 19(11):2861–2873, 2010. 2

[42] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 2019. 1, 3, 4, 6, 7, 8

[43] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, 2020. 3

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7, 8

[45] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 3

[46] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 3

[47] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 2

[48] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *CVPR*, 2019. 1

[49] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7, 8

[50] Linchao Zhu and Yi Yang. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *CVPR*, 2020. 3

Supplementary Material

In the main paper, we show that our method can generate superior super-resolution results for parkour videos captured by egocentric sports cameras. In addition to the parkour videos, we collect real-world video clips (See Fig. 1) from commonly seen categories: animation, movie, music video (MV) and vlog. The goal of this experiment is to test the robustness of our method in various scenarios. Note that the appearances of these videos are significantly different from the videos in our training set (Vimeo90K [4]). In this supplementary material, the results show that our method generalizes better than comparison methods EDVR [3], TOFlow [4] and TGA [1] which are also trained on Vimeo90K [4]. We now discuss the examples shown in Fig. 1.

1) Animation. Animation is challenging to frame alignment for its lack of textures and low frame rate. We show a 2D animation example (a) and 3D animation examples (b) and (c) in Fig. 1. In example (a), it is difficult to recover the facial details by simply fusing neighbor frames since the information is completely corrupted in the low-resolution frames (see bicubic and comparison results). Our method can recover the facial details thanks to the memory-augmented attention. In example (b), aligning the strings on a moving guitar is difficult for EDVR [3], TOFlow [4] and TGA [1]. Our one-hot cross-frame non-local attention can effectively avoid the artifacts caused by the misaligned frames. This mechanism also works better for still repetitive pattern regions like example (c), which are often recognized as moving patterns and shifted (EDVR [3], TOFlow [4] and TGA [1]). Image super-resolution method CSNLN [2] also fails due to the erroneous non-local attention.

2) Movie. Super-resolving movies are of interest in online video streaming services. Movies are also challenging for their large motion and low illumination. Our method can generate very small scale sharp details like the star on the shield (examples (d)), wrinkles on the face (example (e)) and eagle eye/feathers (example (f)), which are difficult to reconstruct using either frame alignment (EDVR [3], TOFlow [4] and TGA [1]) and regular non-local attention (PFNL [5] and CSNLN [2]).

3) MV. Music video (MV) is one of the most-watched video categories online. Music video usually focuses on close-up shots of dancing humans, making the reconstruction of details on clothes important. In examples (g) and (h), our cross-frame non-local attention module enables recovering the fine details under the presence of motion blur. The comparison non-lcoal attention based methods PFNL [5] and CSNLN [2] cannot achieve results comparable to ours since it is much less efficient to query a pixel in the entire video segment/frame than in the pixel's neighborhood like our method.

4) Vlog. Vlog is another type of daily video captured by hand-held devices. This category also covers video chat, which is also common in daily life. Due to the instability of hand-held devices, these videos are extremely shaky and difficult to super-resolve. Existing video super-resolution methods TOFlow [4], TGA [1] and PFNL [5] fail in example (j) and (k). EDVR [3] can recover some details in examples (i), (j) and (k), but their results are blurry in general due to the inaccurate frame alignment.

	PSNR	SSIM	LPIPS		PSNR	SSIM	LPIPS
MANA	**37.31**	**0.9614**	**0.0681**	TGA	37.12	0.9601	0.0707
EDVR	34.48	0.9456	0.1150	PFNL	37.04	0.9673	0.0819
TOFlow	35.58	0.9531	0.0968	CSNLN	36.09	0.9545	0.0844

Table 1: Quantitative comparison on the videos shown in Fig. 1. Larger numbers indicate better results for PSNR and SSIM, smaller numbers indicate better results for LPIPS.

We also show the average quantitative values for the example videos in Fig. 1. In summary, our method works consistently better than the existing state-of-the-art video super-resolution methods in various categories of real-world videos. This proves the robustness of our method, which is important for real applications.

References

[1] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In *CVPR*, 2020. 1

[2] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Humphrey Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, 2020. 1

[3] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPR*, 2019. 1

[4] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019. 1

[5] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 2019. 1

Figure 1: Additional visual comparison on examples from daily videos including animations (examples (a), (b) and (c)), movies (examples (d), (e) and (f)), MVs (examples (g) and (h)), and vlogs (examples (i), (j) and (k)). Our method works consistently better in common types of real-world video, indicating the robustness of our method.