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

Jiyang Yu[1], Jingen Liu[1], Liefeng Bo[1], Tao Mei[2]
[1]JD Explore Academy, Mountain View, USA,
[2]JD Explore Academy, Beijing, China
{jiyang173, jingenliu}@gmail.com, {liefeng.bo, tmei}@jd.com

Abstract

In this paper, we propose a simple yet effective 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 challenges 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 being more robust to large motions in the video. In addition, to acquire general video prior information beyond neighbor frames, and to compensate for the information loss caused by large motions, we design a novel memory-augmented attention module to memorize general video details during the super-resolution training. We have thoroughly evaluated our work on various challenging datasets. Compared to other recent video super-resolution approaches, our method not only achieves significant performance gains on large motion videos but also shows better generalization. Our source code and the new Parkour benchmark dataset is available at https://github.com/jiy173/MANA.

1. Introduction

Video super-resolution (VSR) task aims to generate high-resolution (HR) videos from low-resolution (LR) 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 VSR tasks. The first challenge comes from the dynamic nature of videos. To ensure temporal consistency and improve visual fidelity, previous methods generally seek to fuse information from multiple neighbor frames. Due to the motion across frames, neighbor frames need to be aligned before fusion. Recent works have proposed various ways for aligning neighbor frames to the current frame, either by explicit warping using optical flow [2, 17, 21, 28] or learning implicit alignment using deformable convolution [29, 32]. 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. As an example shown in Fig. 1 column (a), the method EDVR [32] and TOFlow [37] fail in the scenario of large motions due to fusing misaligned frames. This hinders the application of existing VSR methods in real-world videos such as sports videos (see our Parkour dataset in Sec. 4.1), and entertainment videos from animation, movies and vlogs.

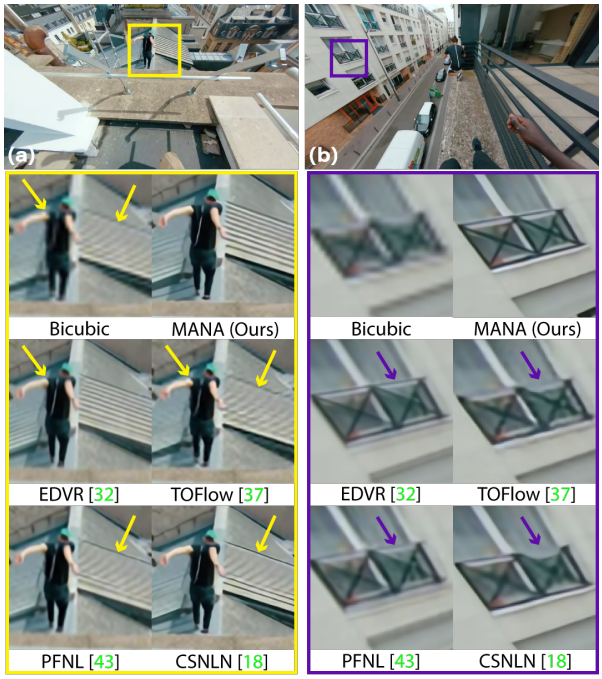

Figure 1. Our memory-augmented cross-frame non-local attention approach is robust to large motion videos (first row). Our method reconstructs visually pleasing details on repetitive patterns (left example) and thin structures (right example) while other video super-resolution methods fail in these cases. Best viewed in PDF.

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 (SISR) works [5, 12, 13, 16, 18, 26, 30, 34, 38, 49] have intensively studied the visual reconstruction from LR images by learning general image prior to help recover high-frequency details or transferring texture from an

HR reference image. One straightforward solution for video super-resolution (VSR) is to directly apply SISR methods to each frame, but it does not guarantee temporal consistency in the visual appearance. Instead, most VSR methods try to fuse information from neighbor frames for HR frame reconstruction, and thus generate results superior to SISR methods. However, we argue that the information acquired from neighbor frames is still limited, especially for videos with large motions. In such a scenario, the correlations among neighbor frames become smaller due to less similar neighbor frames, which makes it difficult to mine useful information from neighbor frames. As a result, the VSR essentially degrades to the single image super-resolution.

To address the aforementioned challenges, we propose a Memory-Augmented Non-local Attention (MANA) framework for video super-resolution (VSR). Our MANA takes a set of consecutive low-resolution video frames as inputs, and 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 temporal consistency in the result. But most importantly, MANA consists of two novel modules, which are specifically designed for solving the VSR challenges.

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. Conventional non-local attention [33] computes the pair-wise correlation between each pixel in the query and key. In the video super-resolution (VSR) case, however, it is improper to treat pixels in all spatial locations equally like conventional non-local attention. We observe that the pixels near the query are more likely to be good correspondences thanks to the nature of continuity. Therefore, unlike conventional non-local attention, we employ a trainable Gaussian map centered at the query pixel to weight the correlations. This is helpful for keeping a good balance between all information sources, and effectively reduces mistaken correspondences that will negatively affect the accuracy of super-resolution. The Gaussian weighted cross-frame non-local attention enables our work to circumvent the frame-alignment operation, which usually performs poorly in videos with large motions. As an example, Fig. 1(a) illustrates that our method can reconstruct sharp details like the stripes on the roof and the waving arm in fast-moving frames.

To solve the challenge of the lack of information from neighbor frames, we seek to fuse useful video prior 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 training. The purpose 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 knowledge, our work is the first VSR method that leverages the memory bank mechanism to incorporate information beyond the current video. Thanks to the general video prior captured by this module, our method can recover details that are missing in the LR video like the balcony railings shown in Fig. 1(b).

To verify the superiority of our MANA method on videos with large motions, we collect the Parkour benchmark dataset. Both qualitative and quantitative results on this dataset have demonstrated that our MANA significantly outperforms all previous approaches. In addition, we also evaluate MANA on other public datasets including Vimeo90K [37], SPMC [28], and Vid4 [21]. Our approach still achieves better or very competitive results. It is worth noting that MANA shows better generalization, because it is superior to other approaches on SPMC and Parkour datasets, which are very different from the training dataset Vimeo90K.

To summarize, our contributions include the follows:
Cross-frame non-local attention. We introduce a Gaussian weighted cross-frame non-local attention that liberates the video super-resolution from the error-prone frame alignment process, and effectively balances the local and non-local information sources. This design makes our method robust to videos with large motions (See Sec. 3.2).
Video super-resolution beyond current video. To the best of our knowledge, we are the first to leverage the memory-augmentation attention to incorporate general video prior to assist current video super-resolution. (See Sec. 3.3).
New benchmark for large motion video super-resolution. We introduce the Parkour dataset containing large motion videos. To our knowledge, this is the first benchmark for evaluating VSR methods in large motion cases (See Sec. 4.1).

2. Related Work

Single Image Super-Resolution. Early single image super-resolution (SISR) works resort to image processing algorithms [25, 40–42]. Recent works in deep learning have been proved to obtain superior results in SISR 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 SISR. Kim et al. further explore deeper residual networks (VDSR [12]) and recursive structures (DRCN [13]). ESPCN [23] encodes the low-resolution image into multiple sub-pixel channels and upscales to a high-resolution image by shuffling the channels back in the spatial domain, which was widely used in recent super-resolution works. Other approaches using CNN includes pyramid structure (LapSRN [15]), recursive residual network (DRRN [27]), dense skip connections (SRDenseNet [31] and RDN [48]), and adversarial networks [3, 16, 22, 35].
Video Super-Resolution. Video super-resolution (VSR) typically generates better results than SISR thanks to the extra information from neighbor frames. The main fo-

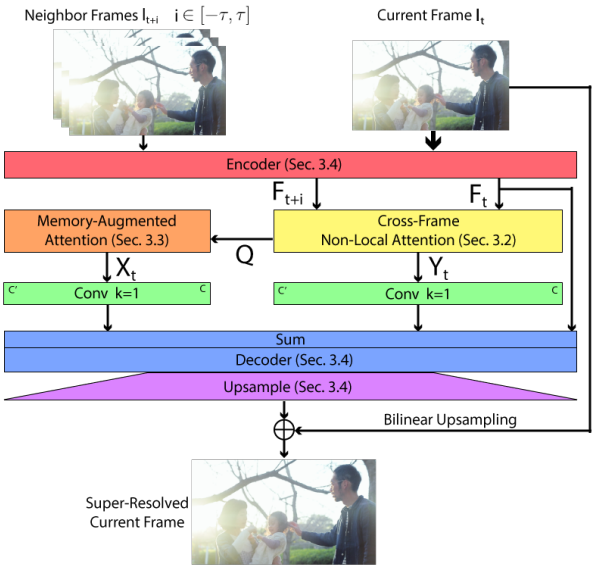

Figure 2. An overview of the structure of our video super-resolution network. The network super-resolves the current frame $\mathbf{I}_t$ using the neighbor frames $\mathbf{I}_{t-\tau}, ..., \mathbf{I}_{t+\tau}$ as the input. The cross-frame non-local attention aims at mining information from neighbor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

cus of VSR 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 [21], SPMC [28], TOFlow [37] 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 the current frame is large. Having observed this limitation, recent methods start to explore techniques to bypass alignment or implicitly align frames. Jo et al. propose DUF [11] that learns dynamic upsampling filters combining the entire spatial neighborhoods of a pixel in auxiliary frames. TDAN [29] and EDVR [32] use deformable convolution layers 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 VSR 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, 20, 44, 46, 47]. Some recent SISR works including CSNLN [18], RNAN [47] and TTSR [39] have designed various mechanisms of non-local attention for image super-resolution. Wang et al. [33] propose non-local neural networks to capture pixel-wise correlations within a video segment, making temporally and spatially long distance attention possible. Following this regular non-local attention, in the task of VSR, the authors of PFNL [43] also utilize self-attention as a feature preprocessing step for their progressive fusion of neighbor frames. This traditional non-local attention may find more matches to a query. But meanwhile, it can also introduce more mistaken correspondences, which will mess up the process of super-resolving the current frame. In contrast, following the nature of video continuity, we believe the matches near the query pixel carry more importance than the distant ones. Hence, we employ a trainable Gaussian map to weigh the non-local attention. The comparison experiments validate our approach indeed generates better results than PFNL [43].

Memory models Neural networks with memory show their potential in natural language processing [1, 24], image classification [51] 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 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 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 $\{\mathbf{I}_{t-\tau}, ..., \mathbf{I}_{t+\tau}\}$. $T = 2\tau + 1$ represents the time span of neighbor frames. Note that neighbor frames include the current frame.

The first stage of our network embeds all 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 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 consisting 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. 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 (See Sec. 3.2).

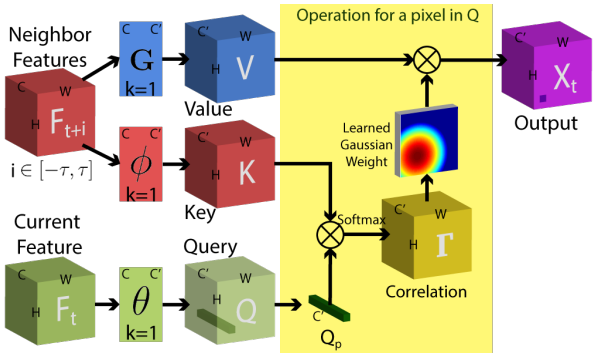

Figure 3. The cross-frame non-local attention module in our network. The size are marked on the edges of the tensors. The operation marked by the yellow box is done in parallel for each pixel $\mathbf{Q}_p$ in the query tensor $\mathbf{Q}$. Best viewed in PDF.

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. 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 [21, 37] and deformable convolution [29, 32]. 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 [33] 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 [36], 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 $\mathbf{\Gamma} = \widehat{\mathbf{Q}}\widehat{\mathbf{K}}$. Since the size of $\mathbf{\Gamma}$ is $HW \times HWT$, this matrix imposes a large burden on the GPU memory. To make the network more memory efficient, we conduct non-local attention on each neighbor frames separately, i.e. the size of $\mathbf{\Gamma}$ is $HW \times HW$. The first dimension of $\mathbf{\Gamma}$ spans the spatial locations in $\mathbf{Q}$, and the second dimension spans the spatial locations in $\mathbf{K}$.

Unlike the high-level video classification task discussed in the original non-local attention [33], we aim to explore pixel-level information from neighbor frames in the video super-resolution. The goal of non-local attention is to find more matches to a query pixel. However, it can also introduce more inaccurate correspondences. Mistakenly matched pixels far away from the query pixel may have a negative effect on the video super-resolution performance. In Sec. 4.3, we will show that traditional non-local attention did not benefit the video super-resolution method PFNL [43] which directly applies it to the entire group of neighbor frames. Intuitively, most correspondences of a pixel should generally reside in its neighbor region in the neighbor frames thanks to the continuity characteristics of videos. To mitigate the effect of mistakenly matched pixels, we therefore multiply a Gaussian weight map $\mathbf{G} \in \mathbb{R}^{HW}$ on each slice in the second dimension of the correlation matrix $\mathbf{\Gamma}$. Note that the center of the Gaussian map is located at the location of the query pixel. In other words, the Gaussian map is different for each slice in the first dimension of $\mathbf{\Gamma}$. Instead of tuning an optimal standard deviation for the Gaussian map, we make it a trainable parameter and learn what value leads to the best overall performance. The final output of the cross-frame non-local attention module can be written as:

$$\mathbf{X}_t = (\mathbf{G} \bigotimes \mathbf{\Gamma}) \cdot \mathbf{V} \tag{1}$$

where $\bigotimes$ represents the slice-wise Hadamard product described above. The trainable Gaussian map maintains a good balance between the local and non-local sources for fusing information from neighbor frames in our VSR task.

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

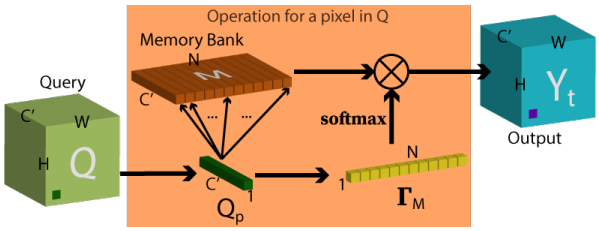

Figure 4. The memory-augmented attention module in our network. The operation marked by the orange box is done in parallel for each pixel $\mathbf{Q}_p$ in the query tensor $\mathbf{Q}$. Best viewed in PDF.

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

$$\widehat{\mathbf{Y}}_t = softmax(\mathbf{\Gamma}_M)\widehat{\mathbf{M}} \tag{2}$$

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. The Vimeo90K dataset is a large-scale video dataset proposed by Xue et al. [37]. Following recent super-resolution methods TOFlow [37], TDAN [29] and EDVR [32], 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. 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 [23]. 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

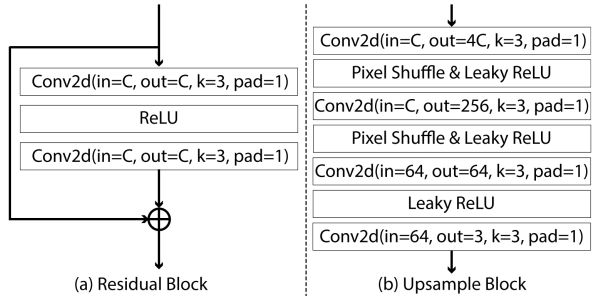

Figure 5. Basic building blocks in our network. (a) Residual blocks are used to build the encoder and decoder. (b) Upsample block shuffles pixels in different channels into a high-resolution frame.

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.

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 comparison methods based on their approaches to the super-resolution problem: VSR via explicit frame alignment (TOFlow [37], TGA [10] and DBVSR [19]), VSR via implicit frame alignment (EDVR [32]), VSR via regular non-local attention (PFNL [43]) 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

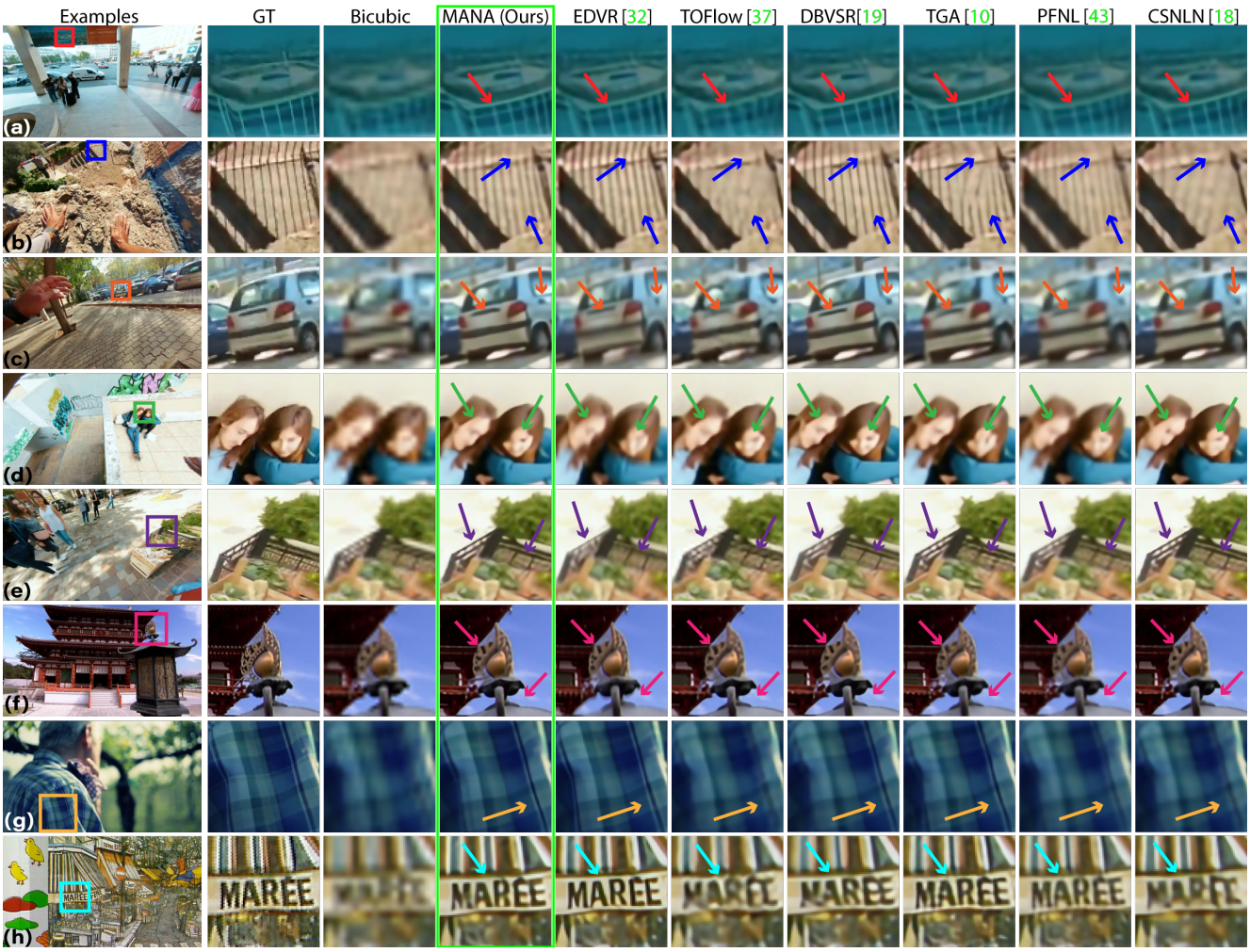

Figure 7. Visual comparison on the Parkour dataset, Vimeo90K [37] dataset and Vid4 [21] dataset. Example (a), (b), (c), (d) and (e) are selected from the large motion Parkour dataset. Example (f) is from SPMC [28] dataset. Example (g) is from the Vimeo90K [37] dataset. Example (h) is from Vid4 [21] dataset. 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.

section. To obtain the low-resolution input, we use bicubic down-sampling on the ground truth high-resolution frames. According to our experiment, PFNL [43] and TGA [10] introduce serious aliasing artifacts to the results using bicubic down-sampled video. To make the comparison fair, for PFNL [43] 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 $N = 512$ 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 59ms per 448x256 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 collected 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 egocentric 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 videos using Vimeo90K [37] test set, SPMC dataset [28], and Vid4 [21] (it contains 4 videos).

For all test sets, we use the average PSNR and SSIM [50] scores on the RGB channel to quantitatively evaluate the performance. In addition, we apply LPIPS [45] 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-ran their codes and calculated

	(a) Parkour Dataset			(b) Vimeo90K-Motion [37]			(c)Vimeo90K Dataset [37]			(d)SPMC Dataset [28]		
	PSNR↑ in dB	SSIM↑	LPIPS↓	PSNR↑ in dB	SSIM↑	LPIPS↓	PSNR↑ in dB	SSIM↑	LPIPS↓	PSNR↑ in dB	SSIM↑	LPIPS↓
Bicubic	29.51	0.8712	0.3101	33.90	0.9194	0.2122	29.75	0.8476	0.2948	25.67	0.7241	0.4270
MANA(Ours)	**33.81**	**0.9397**	**0.1159**	**38.86**	**0.9630**	**0.0853**	**34.84**	**0.9404**	**0.1076**	**29.27**	**0.8449**	**0.2147**
EDVR [32]	31.61	0.9113	0.1900	38.33	0.9544	0.0813	35.68	0.9372	0.1019	27.98	0.8109	0.2715
TOFlow [37]	32.35	0.9197	0.1804	36.55	0.9471	0.1186	32.96	0.9041	0.1451	28.55	0.8327	0.2661
DBVSR [19]	32.09	0.9225	0.1534	37.77	0.9563	0.0943	33.47	0.9265	0.1240	28.00	0.8186	0.2247
TGA [10]	31.14	0.9033	0.2224	38.26	0.9588	0.0919	35.03	0.9310	0.1013	29.06	0.8449	0.2390
PFNL [43]	32.04	0.9189	0.2244	35.90	0.9449	0.1522	31.86	0.8959	0.2012	28.27	0.8270	0.3100
CSNLN [18]	32.93	0.9275	0.1357	37.79	0.9523	0.1062	33.55	0.9091	0.1338	28.79	0.8275	0.2343

Table 1. Quantitative comparison on (a) Parkour dataset, (b) Vimeo90K-Motion [37], (c) Vimeo90K [37] dataset and (d) SPMC dataset [28]. The metrics used are PSNR, SSIM and LPIPS. Larger numbers indicate better results for PSNR and SSIM, smaller numbers indicate better results for LPIPS.

the metrics in the same way on the same computer.

4.2. Visual Comparisons

The visual comparisons of various examples selected from the Parkour, SPMC, Vimeo90K and Vid4 dataset are shown in Fig. 7. To make the discussion concise, we label the ID at the bottom left of each video. We also add arrows pointing at the regions to be discussed.

Example (a), (b), (c), (d) and (e) are selected from the Parkour dataset. These examples contain large motions and are challenging to existing VSR methods. Our method can reconstruct repetitive patterns like Example (a) and (b), while explicit frame alignment methods TOFlow [37] and TGA [10] fail due to the inaccurate frame alignment. Recent method DBVSR [19] improves frame alignment by learning to deblur, but still cannot handle repetitive patterns in (b). EDVR [32] result is more blurry than our result in example (a) and (b), and *the blurry issue is more visible when viewed in dynamics as shown in the supplementary video*. This indicates that the deformable convolution alignment cannot handle the alignment with large frame displacements. Both PFNL [43] and CSNLN [18] using non-local attentions 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 the general details of an object. The frame-aligning methods introduce either ghosting artifacts (EDVR) or deformation (TOFlow and TGA) due to the inaccurate alignment. The results of PFNL and CSNLN have less details than ours, indicating that our Gaussian weighted 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 either blur (EDVR, TOFlow and PFNL) or reconstruct shapes that do not look like a human (TGA and CSNLN).

Example (e) and (f) contain thin structures. Similar to examples (a) and (b), the failures in frame alignment have negatively affected VSR methods. In these examples, the performance of EDVR, TOFlow and TGA are even worse than the SISR method CSNLN. The result of DBVSR is superior to that of TOFlow, but is more blurry than our result.

As being discussed in Sec. 4.3, the overall average quantitative score of our method is slightly inferior to that of EDVR and TGA in the Vimeo90K [37] and Vid4 dataset [21] which are relatively easy for frame aligning VSR methods. However, a larger deviation to the ground truth does not always indicate worse performance. As shown in example (g) selected from Vimeo90K, our method tends to produce visually sharper results than EDVR and TGA, which is often more preferred in the VSR task. Example (h) is a widely used example in Vid4. Our result is comparable to that of EDVR and TGA.

To further evaluate the robustness of our method in real-world scenarios, we arbitrarily selected videos of different types including animation, movies, and vlogs for video super resolution. The results further prove that our approach is superior to others (Due to limited space, results are included in the supplementary material).

4.3. Quantitative Comparisons

Table 1 displays the quantitative comparisons of our MANA to the state-of-the-art VSR approaches in terms of PSNR, SSIM, and LPIPS score, where 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. In this table, we illustrate the quantitative results of VSR on 4 datasets: Parkour dataset, Vimeo90K-Motion, Vimeo90K, and SPMC. The results on Vid4 can be found in our supplementary material due to the lack of space.

Table 1 Column (a) illustrates the result comparisons on the Parkour dataset. Videos in this dataset have extremely large motions, making the accurate alignment of the frames difficult. Among the comparison methods, TOFlow [37] and DBVSR [19] explicitly estimate the optical flow for warping neighbor frames; TGA [10] uses homography to align neighbor frames; EDVR [32] implicitly align frames using learned kernel offset for deformable convolution. Hence, the traditional VSR methods rely on the explicit or implicit alignment of neighbor frames, which generally can be affected by large motions in videos. The results also prove this point. As we can see, our MANA approach, which does not require frame alignment, has outperformed all VSR methods by a large margin. This observation indicates MANA is able to cope with large motions in videos. It is also interesting to notice that the performances of the frame-alignment VSR methods are even inferior to that of

	Parkour Dataset			Vimeo90K Dataset [37]			Vid4 Dataset [21]			SPMC Dataset [28]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
No_Mem	33.57	0.9367	0.1208	34.53	0.9377	0.1113	25.02	0.7739	0.2942	29.01	0.8384	0.2210
$N = 128$	33.72	0.9384	0.1214	34.73	0.9396	0.1089	25.17	0.7791	0.2931	29.17	0.8427	0.2189
$N = 256$	33.79	0.9395	0.1181	34.83	0.9407	0.1073	25.19	0.7815	0.2867	29.30	0.8452	0.2166
$\boldsymbol{N = 512}$	**33.81**	**0.9397**	**0.1159**	**34.84**	**0.9404**	**0.1076**	**25.21**	**0.7816**	**0.2842**	**29.27**	**0.8449**	**0.2147**
$N = 1024$	33.75	0.9390	0.1213	34.76	0.9398	0.1117	25.19	0.7802	0.2962	29.25	0.8447	0.2234

Table 2. Evaluation on memory size in the memory-augmented attention module. The $N = 512$ is selected for the experiments shown in Sec. 4.2 and Sec. 4.3

the SISR method CSNLN [18]. It is because fusing misaligned frames often cause ghosting artifacts in the result.

As we can see, although PFNL [43] works better than the frame-alignment method EDVR and TGA, its performance is even worse than the single-frame approach CSNLN. We conjecture that the performance gap between PFNL and CSNLN may be caused by the design of non-local attention in PFNL, which employs pair-wise non-local attention on all pixels in the entire spatiotemporal segment. This regular non-local attention can help find more correspondences globally, but meanwhile it can introduce more mistaken matches which may cause negative effects to the results of VSR. In contrast, our Gaussian weighted non-local attention is able to balance the fusion of local and non-local information. Hence, it significantly improves the performance of non-local attention as shown in Table 1 (a).

Table 1 column (b) exhibits additional experimental results on Vimeo90K-Motion, which consists of regular videos with relatively large motions. We computed the optical flow for videos in the Vimeo90K test set and ranked them based on the average flow magnitude. The top 6% videos are selected to form Vimeo90K-Motion. The results further confirm that our MANA works better on videos with some motions.

In addition, Table 1 column (c) and (d) illustrate more quantitative result comparisons on the dataset Vimeo90K and SPMC, respectively. As we can see, on these regular videos, MANA also achieves better performance than the explicit optical flow alignment methods TOFlow, DBVSR and the other non-local attention super-resolution methods PFNL and CSNLN. The PSNR score values of our method are slightly inferior to that of EDVR , and TGA in the Vimeo90K dataset. However, for the large motion videos in the Parkour dataset, our method has much larger PSNR gains (2.2dB and 2.67dB) in performance compared to EDVR and TGA.

It is worth noting that our approach MANA has better generalization than others. Although our MANA only obtains comparable results on Vimeo90K, it noticeably outperforms other VSR methods on both SPMC and Parkour. As all methods are trained on the Vimeo90K training set, the test results on both SPMC and Parkour datasets are more convincing. Please note the SPMC and Parkour dataset are very different from Vimeo90K. In contrast, EDVR could be biased towards Vimeo90K, given the significant performance drop in the SPMC dataset. Therefore, MANA is more robust and generalized than other approaches. This observation is further confirmed by our additional quantitative evaluations on more real-world videos shown in the supplementary material.

4.4. Evaluation on Memory Size

In Table 2, 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 2). Among these configurations, $N = 512$ achieves the best result and is selected in the comparisons in Sec. 4.2 and Sec. 4.3. Using smaller memory ($N = 128$ and $N = 256$) results in slight performance degradation. The benefits saturate when using a larger memory ($N = 1024$), 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 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 results in 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 that are relatively easy for frame aligning video super-resolution methods. The limitation is that our method cannot directly handle videos with compression artifacts. In real applications, separate pre-processing will be required to remove compression artifacts. We believe our method can be further improved by introducing a pyramid structure into the cross-frame non-local attention to increase the perception field or extend the memory bank from 2D to the 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, 3

[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. 3

[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. 3, 5, 6, 7

[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. 3

[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, 5, 7, 8

[19] Jinshan Pan, Haoran Bai, Jiangxin Dong, Jiawei Zhang, and Jinhui Tang. Deep blind video super-resolution. In *ICCV*, 2021. 5, 7

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

[21] Mehdi Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018. 1, 2, 3, 4, 6, 7, 8

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

[23] 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

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

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

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

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

[28] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017. 1, 2, 3, 6, 7, 8

[29] 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

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

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

[32] 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, 3, 4, 5, 7

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

[34] 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

[35] 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

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

[37] 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, 3, 4, 5, 6, 7, 8

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

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

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

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

[42] 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

[43] 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. 3, 4, 5, 6, 7, 8

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

[45] 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. 6

[46] 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

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

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

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

[50] 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. 6

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