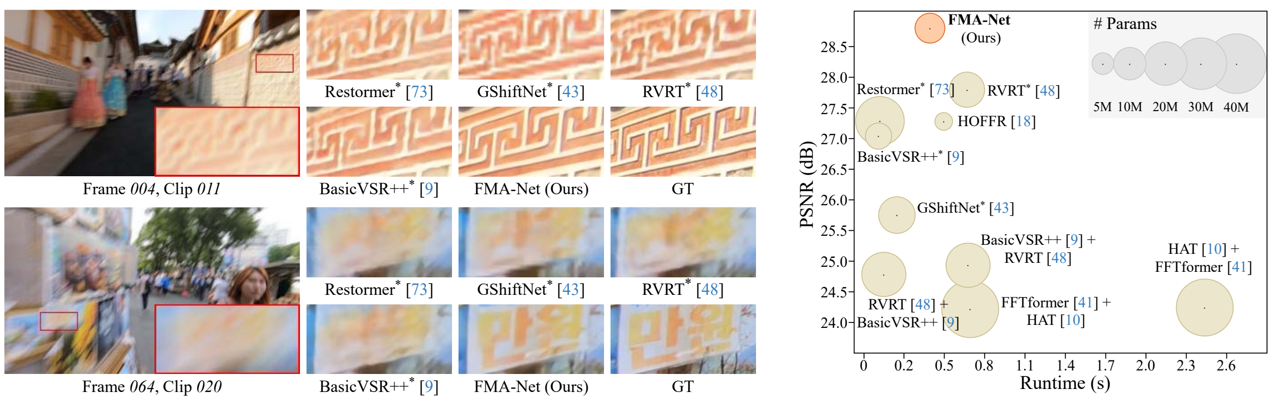
2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.00013

FMA-Net: Flow-Guided Dynamic Filtering and Iterative Feature Refinement with Multi-Attention for Joint Video Super-Resolution and Deblurring

Geunhyuk Youk
KAIST
rmsgurkjg@kaist.ac.kr

Jihyong Oh †
Chung-Ang University
jihyongoh@cau.ac.kr

Munchurl Kim †
KAIST
mkimee@kaist.ac.kr



(a) Visual comparison results of different methods on REDS4 [52] test set

(b) Performance Gain

Figure 1. Our FMA-Net outperforms state-of-the-art methods in both quantitative and qualitative results for $\times 4$ VSRDB.

Abstract

We present a joint learning scheme of video super-resolution and deblurring, called VSRDB, to restore clean high-resolution (HR) videos from blurry low-resolution (LR) ones. This joint restoration problem has drawn much less attention compared to single restoration problems. In this paper, we propose a novel flow-guided dynamic filtering (FGDF) and iterative feature refinement with multi-attention (FRMA), which constitutes our VSRDB framework, denoted as FMA-Net. Specifically, our proposed FGDF enables precise estimation of both spatio-temporally-variant degradation and restoration kernels that are aware of motion trajectories through sophisticated motion representation learning. Compared to conventional dynamic filtering, the FGDF enables the FMA-Net to effectively handle large motions into the VSRDB. Additionally, the stacked FRMA blocks trained with our novel temporal anchor (TA) loss, which temporally anchors and sharpens features, refine features in a coarse-to-fine manner through iterative updates. Extensive experiments demonstrate the superiority of the proposed FMA-Net over state-of-the-art methods in terms of both quantitative and qualitative quality. Codes and pre-trained models are available at: https://kaist-viclab.github.io/fmanet-site.

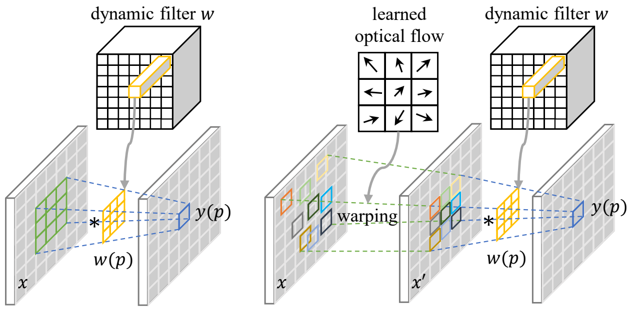
1. Introduction

Video super-resolution (VSR) aims to restore a high-resolution (HR) video from a given low-resolution (LR) counterpart. VSR can be beneficial for diverse real-world applications of high-quality video, such as surveillance [1, 78], video streaming [14, 82], medical imaging [2, 21], etc. However, in practical situations, acquired videos are often blurred due to camera or object motions [4, 75, 77], leading to a deterioration in perceptual quality. Therefore, joint restoration (VSRDB) of VSR and deblurring is needed, which is challenging to achieve the desired level of high-quality videos because two types of degradation in blurry LR videos should be handled simultaneously.

A straightforward approach to solving the joint problem of SR and deblurring is to perform the two tasks sequentially, *i.e.,* by performing SR first and then deblurring, or vice versa. However, this approach has a drawback with the propagation of estimation errors from the preceding operation (SR or deblurring) to the following one (deblurring or SR) [55]. To overcome this, several works pro-

†Co-corresponding authors.

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.00013



(a) Dynamic filtering (b) Our Flow-guided dynamic filtering

Figure 2. Comparison of 3×3 dynamic filtering. (a) conventional dynamic filtering at location p with fixed surroundings and (b) our flow-guided dynamic filtering (FGDF, Sec. 3.3) at position p with variable surroundings guided by learned optical flow.

posed joint learning methods of image SR and deblurring (ISRDB), and VSRDB methods [16, 18, 26, 72, 74, 79]. They showed that the two tasks are strongly inter-correlated. However, most of these methods are designed for ISRDB [16, 26, 72, 74, 79]. Since motion blur occurs due to camera shakes or object motions, efficient deblurring requires the use of temporal information over video sequences. Recently, Fang *et al.* [18] proposed the first deep learning-based VSRDB method, called HOFFR, which combines features from the SR and deblurring branches using a parallel fusion module. Although HOFFR exhibited promising performance compared to the ISRDB methods, it struggled to effectively deblur spatially-variant motion blur due to the nature of 2D convolutional neural networks (CNNs) with spatially-equivariant and input-independent filters.

Inspired by the Dynamic Filter Network [33] in video prediction, significant progress has been made with the dynamic filter mechanism in low-level vision tasks [35, 38, 53, 54, 83]. Specifically, SR [35, 38] and deblurring [19, 83] have shown remarkable performances with this mechanism in predicting spatially-variant degradation or restoration kernels. For example, Zhou *et al.* [83] proposed a video deblurring method using spatially adaptive alignment and deblurring filters. However, this method applies filtering only to the reference frame, which limits its ability to accurately exploit information from adjacent frames. To fully utilize motion information from adjacent frames, large-sized filters are required to capture large motions, resulting in high computational complexity. While the method [54] of using two separable large 1D kernels to approximate a large 2D kernel seems feasible, it loses the ability to capture fine detail, making it difficult to apply for video effectively.

We propose FMA-Net, a novel VSRDB framework based on Flow-Guided Dynamic Filtering (FGDF) and an Iterative Feature Refinement with Multi-Attention (FRMA), to allow for small-to-large motion representation learning with good joint restoration performance. The key insight of the FGDF is to perform filtering that is aware of motion trajectories rather than sticking to fixed positions, enabling effective handling of large motions with small-sized kernels. Fig. 2 illustrates the concept of our FGDF. The FGDF looks similar to the deformable convolution (DCN) [13] but is different in that it learns position-wise $n \times n$ dynamic filter coefficients, while the DCN learns position-invariant $n \times n$ filter coefficients.

Our FMA-Net consists of (i) a degradation learning network that estimates motion-aware spatio-temporally-variant degradation kernels and (ii) a restoration network that utilizes these predicted degradation kernels to restore the blurry LR video. The newly proposed multi-attention, consisting of center-oriented attention and degradation-aware attention, enables the FMA-Net to focus on the target frame and utilize the degradation kernels in a globally adaptive manner for VSRDB. We empirically show that the proposed FMA-Net significantly outperforms the recent state-of-the-art (SOTA) methods for video SR and deblurring in objective and subjective qualities on the REDS4, GoPro, and YouTube test datasets under a fair comparison, demonstrating its good generalization ability.

2. Related Work

2.1. Video Super-Resolution

In contrast to image SR that focuses primarily on extracting essential features [15, 36, 38, 76, 80] and capturing spatial relationships [10, 46], VSR faces with an additional key challenge of efficiently utilizing highly correlated but misaligned frames. Based on the number of input frames, VSR is mainly categorized into two types: sliding window-based methods [5, 29, 31, 35, 43, 45, 64, 67] and recurrent-based methods [7, 9, 20, 24, 49, 50, 57].

Sliding window-based methods. Sliding window-based methods aim to recover HR frames by using neighboring frames within a sliding window. These methods mainly employ CNNs [31, 35, 37, 44], optical flow estimation [5, 62], deformable convolution (DCN) [13, 64, 70], or Transformer structures [6, 45, 47], with a focus on temporal alignment either explicitly or implicitly.

Recurrent-based methods. Recurrent-based methods sequentially propagate the latent features of one frame to the next frame. BasicVSR [7] and BasicVSR++ [9] introduced the VSR methods by combining bidirectional propagation of the past and future frames into the features of the current frame, achieving significant improvements. However, the recurrent mechanism is prone to gradient vanishing [11, 27, 50], thus causing information loss to some extent.

Although some progress has been made, all the above methods can handle not blurry but sharp LR videos.

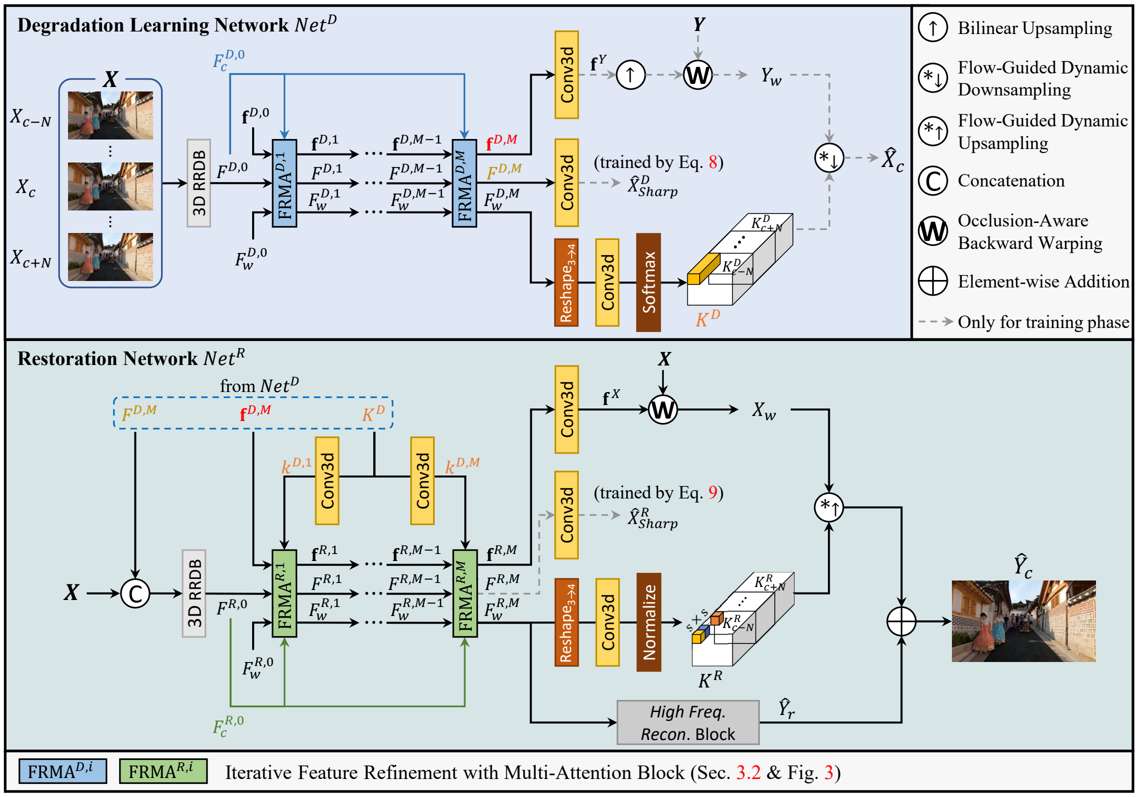


Figure 3. The architecture of FMA-Net for video super-resolution and deblurring (VSRDB).

2.2. Video Deblurring

Video deblurring aims to remove blur artifacts from blurry input videos. It can be categorized into single-frame deblurring [41, 42, 65, 73] and multi-frame deblurring [30, 34, 39, 47, 48]. Zhang *et al.* [75] proposed a 3D CNN-based deblurring method to handle spatio-temporal features, while Li *et al.* [43] introduced a deblurring method based on grouped spatial-temporal shifts. Recently, transformer-based deblurring methods such as Restormer [73], Stripformer [65], and RVRT [48] have been proposed and demonstrated significant performance improvements.

2.3. Dynamic Filtering-based Restoration

In contrast to conventional CNNs with spatially-equivariant filters, Jia *et al.* [33] proposed a dynamic filter network that predicts conditioned kernels for input images and filters the images in a locally adaptive manner. Subsequently, Jo *et al.* [35] introduced dynamic upsampling for VSR, while Niklaus *et al.* [53, 54] applied dynamic filtering for frame interpolation. Zhou *et al.* [83] proposed a spatially adaptive deblurring filter for recurrent video deblurring, and Kim *et al.* [38] proposed KOALAnet for blind SR, which predicts spatially-variant degradation and upsampling filters. However, all these methods operate naively on a target position and its fixed surrounding neighbors of images or features and cannot effectively handle spatio-temporally-variant motion.

2.4. Joint Video Super-Resolution and Deblurring

Despite very active deep learning-based research on single restoration problems such as VSR [7, 35, 45, 50, 64] and deblurring [34, 39, 47, 48], the joint restoration (VSRDB) of these two tasks has drawn much less attention. Recently, Fang *et al.* [18] introduced HOFFR, the first deep learning-based VSRDB framework. Although they have demonstrated that the HOFFR outperforms ISRDB or sequential cascade approaches of SR and deblurring, the performance has not been significantly elevated, mainly due to the inherent characteristics of 2D CNNs with spatially-equivariant and input-independent filters. Therefore, there still remain many avenues for improvement, especially in effectively restoring spatio-temporally-variant degradations.

3. Proposed Method

3.1. Overview of FMA-Net

We aim to perform video super-resolution and deblurring (VSRDB) simultaneously. Let a blurry LR input sequence $\boldsymbol{X} = \{X_{c-N:c+N}\} \in \mathbb{R}^{T\times H\times W\times 3}$, where $T = 2N + 1$ and c denote the number of input frames and a center frame index, respectively. Our goal of VSRDB is set to predict a sharp HR center frame $\hat{Y}_c \in \mathbb{R}^{sH\times sW\times 3}$, where s represents the SR scale factor. Fig. 3 illustrates the architecture of our proposed VSRDB framework, FMA-Net. The FMA-Net consists of (i) a degradation learning network Net^D and (ii) a restoration network Net^R. Net^D predicts motion-aware spatio-temporally-variant degradation, while Net^R utilizes the predicted degradation from Net^D in a globally adaptive manner to restore the center frame X_c. Both Net^D and Net^R have a similar structure, consisting of the proposed iterative feature refinement with multi-attention (FRMA) blocks and a flow-guided dynamic filtering (FGDF) module. Therefore, in this section, we first describe the FRMA block and FGDF in Sec. 3.2 and Sec. 3.3, respectively. Then, we explain the overall structure of FMA-Net in Sec. 3.4. Finally, we present the loss functions and training strategy for the FMA-Net training in Sec. 3.5.

3.2. Iterative Feature Refinement with Multi-Attention (FRMA)

We use both types of image-based and feature-based optical flows to capture motion information in blurry videos and leverage them to align and enhance features. However, directly using a pre-trained optical flow network is unstable for blurry frames and computationally expensive [55]. To overcome this instability, we propose the FRMA block. The FRMA block is designed to learn self-induced optical flow and features in a residual learning manner, and we stack M FRMA blocks to iteratively refine features. Notably, inspired by [8, 28], the FRMA block learns multiple optical flows with their corresponding occlusion masks. This flow diversity enables the learning of one-to-many relations between pixels in a target frame and its neighbor frames, which is beneficial for blurry frames where pixel information is spread due to light accumulation [22, 25].

Fig. 4(a) illustrates the structure of the FRMA block at the (i+1)-*th* update-step. Note that FRMA block is incorporated into both Net^D and Net^R. To explain the operation of the FRMA block, we omit the superscript D and R for simplicity from its input and output notions in Fig. 3. The FRMA block aims to refine three tensors: temporally-anchored (unwarped) feature $F \in \mathbb{R}^{T\times H\times W\times C}$ at each frame index, warped feature $F_w \in \mathbb{R}^{H\times W\times C}$, and multi-flow-mask pairs $\mathbf{f} \equiv \{f^j_{c\to(c+t)}, o^j_{c\to(c+t)}\}^{t=-N:N}_{j=1:n} \in \mathbb{R}^{T\times H\times W\times(2+1)n}$, where n denotes the number of multi-flow-mask pairs from the center frame index c to each frame index, including learnable occlusion masks $o^j_{c\to(c+t)}$ which are sigmoid activations for stability [55].

(i+1)-*th* Feature Refinement. Given the features F^i, F^i_w, and $\mathbf{f}^i$ computed at the i-*th* update-step, we sequentially update each of these features. First, we refine F^i through a 3D RDB [81] to compute F^{i+1} as shown in Fig. 4(a), *i.e.,* $F^{i+1} = \text{RDB}(F^i)$. Then, we update $\mathbf{f}^i$ to $\mathbf{f}^{i+1}$, by warping F^{i+1} to the center frame index c based on $\mathbf{f}^i$ and concatenating the resultant with F^0_c and $\mathbf{f}^i$, which is given as:

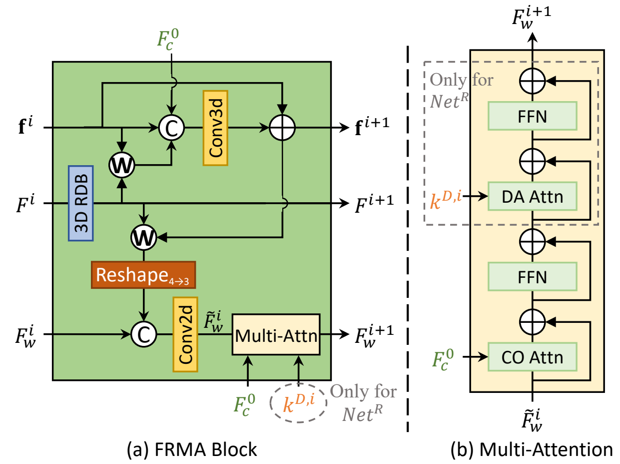


Figure 4. (a) Structure of i+1-*th* FRMA block (Sec. 3.2); (b) Structure of Multi-Attention. FFN refers to the feed-forward network of the transformer [17, 66].

$$\mathbf{f}^{i+1} = \mathbf{f}^i + \text{Conv}_{3d}(\text{concat}(\mathbf{f}^i, \mathcal{W}(F^{i+1}, \mathbf{f}^i), F^0_c)), \quad (1)$$

where $\mathcal{W}$ and concat denote the occlusion-aware backward warping [32, 55, 60] and concatenation along channel dimension, respectively. Note that $F^0_c \in \mathbb{R}^{H\times W\times C}$ represents the feature map at the center frame index c of the initial feature $F^0 \in \mathbb{R}^{T\times H\times W\times C}$. Finally, we update F^i_w by using warped F^{i+1} to the center frame index c by $\mathbf{f}^{i+1}$ as:

$$\tilde{F}^i_w = \text{Conv}_{2d}(\text{concat}(F^i_w, r_{4\to3}(\mathcal{W}(F^{i+1}, \mathbf{f}^{i+1})))), \quad (2)$$

where $r_{4\to3}$ denotes the reshape operation from $\mathbb{R}^{T\times H\times W\times C}$ to $\mathbb{R}^{H\times W\times TC}$ for feature aggregation.

Multi-Attention. Our multi-attention structure is shown in the Fig. 4(b). To better align $\tilde{F}^i_w$ to the center frame index c and adapt to spatio-temporally variant degradation, we enhance $\tilde{F}^i_w$ using center-oriented (CO) attention and degradation-aware (DA) attention. In the case of 'CO attention', for the input $\tilde{F}^i_w$ and F^0_c, it generates *query* (Q), *key* (K), and *value* (V) as $Q = W_q F^0_c$, $K = W_k \tilde{F}^i_w$, and $V = W_v \tilde{F}^i_w$, respectively. Then, we calculate the attention map between Q and K, and use it to adjust V. While this process may resemble self-attention [17, 66] at first, our empirical findings indicate better performance when $\tilde{F}^i_w$ focuses on its relation with F^0_c rather than on itself. The CO attention process is expressed as:

$$\text{CO Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d})V, \quad (3)$$

where $\sqrt{d}$ denotes the scaling factor [17, 66]. The 'DA attention' is the same as the CO attention except that the *query* is derived from feature $k^{D,i} \in \mathbb{R}^{H\times W\times C}$, which is adjusted by convolution with the novel motion-aware degradation kernels K^D from Net^D, rather than from F^0_c. This process enables $\tilde{F}^i_w$ to be globally adaptive to degradation.

The motion-aware kernel K^D will be described in detail in Sec. 3.4. It should be noted that DA attention is only applied in Net^R since it utilizes the predicted K^D from Net^D as shown in Fig. 4. Specifically, we empirically found out that the adoption of the transposed-attention [3, 73] in Eq. 3 shows more efficient and better performances.

3.3. Flow-Guided Dynamic Filtering

We start with a brief overview of dynamic filtering [33]. Let p_k represent the *k-th* sampling offset in a standard convolution with a kernel size of $n \times n$. For instance, we have $p_k \in \{(-1,-1),(-1,0),\dots,(1,1)\}$ when $n = 3$. We denote the predicted $n \times n$ dynamic filter at position p as F^p. The dynamic filtering for images can be formulated as:

$$y(p) = \sum_{k=1}^{n^2} F^p(p_k) \cdot x(p + p_k), \tag{4}$$

where x and y are the input and output features. Its naive extension to video can be expressed as:

$$y(p) = \sum_{t=-N}^{+N} \sum_{k=1}^{n^2} F^p_{c+t}(p_k) \cdot x_{c+t}(p + p_k), \tag{5}$$

where c represents the center frame index of the input frames. However, such a naive extension of filtering at a pixel position with fixed surrounding neighbors requires a large-sized filter to capture large motions, resulting in an exponential increase in computation and memory usage. To overcome this problem, we propose flow-guided dynamic filtering (FGDF) inspired by DCN [13]. The kernels are dynamically generated to be pixel-wise motion-aware, guided by the optical flow. This allows effective handling of large motion with relatively small-sized kernels. Our FGDF can be formulated as:

$$y(p) = \sum_{t=-N}^{+N} \sum_{k=1}^{n^2} F^p_{c+t}(p_k) \cdot x'_{c+t}(p + p_k), \tag{6}$$

where $x'_{c+t} = \mathcal{W}(x_{c+t}, \mathbf{f}_{c+t})$ and $\mathbf{f}_{c+t}$ denotes optical flow with its occlusion mask from frame index c to $c + t$.

3.4. Overall Architecture

Degradation Learning Network. Net^D, shown in the upper part of Fig. 3, takes a blurry LR sequence $\boldsymbol{X}$ as input and aims to predict a motion-aware spatio-temporally variant degradation kernels that are assumed to be used to obtain center frame X_c from the sharp HR counterpart $\boldsymbol{Y}$. Specifically, we first compute the initial temporally-anchored feature $F^{D,0}$ from $\boldsymbol{X}$ through a 3D RRDB [69]. Then, we refine $F^{D,0}$, $F^{D,0}_w$, and $\mathbf{f}^{D,0}$ through M FRMA blocks (Eqs. 1 and 2). Meanwhile, $F^{D,i}_w$ is adaptively adjusted in the CO attention of each FRMA block based on its relation to $F^{D,0}_c$ (Eq. 3), the *center* feature map of $F^{D,0}$. It should be noted that we initially set $F^{D,0}_w = \mathbf{0}$ and $\mathbf{f}^{D,0} = \{f^j_{c\to(c+t)} = \mathbf{0}, o^j_{c\to(c+t)} = \mathbf{1}\}^{t=-N:N}_{j=1:n}$. Subsequently, using the final refined features $\mathbf{f}^{D,M}$ and $F^{D,M}_w$, we calculate an *image* flow-mask pair $\mathbf{f}^Y \in \mathbb{R}^{T\times H\times W\times(2+1)}$ for Y and its corresponding motion-aware degradation kernels $K^D \in \mathbb{R}^{T\times H\times W\times k_d^2}$, where k_d denotes the degradation kernel size. Here, we use a sigmoid function to normalize K^D, which mimics the blur generation process [51, 55, 58, 61] where all kernels have positive values. Finally, we synthesize $\hat{X}_c$ with K^D and $\mathbf{f}^Y$ as:

$$\hat{X}_c = (\mathcal{W}(\boldsymbol{Y}, s\cdot(\mathbf{f}^Y \uparrow_s)) \circledast K^D) \downarrow_s, \tag{7}$$

where $\circledast \downarrow_s$ represents novel $k_d \times k_d$ FGDF via Eq. 6 at each pixel location with stride s and $\uparrow_s$ denotes $\times s$ bilinear upsampling. Additionally, $F^{D,M}$ is mapped to the image domain via 3D convolution to generate $\hat{X}^D_{Sharp} \in \mathbb{R}^{T\times H\times W\times 3}$, which is only used to train the network.

Restoration Network. Net^R differs from Net^D which predicts flow and degradation in $\boldsymbol{Y}$. Instead, Net^R computes the flow in $\boldsymbol{X}$ and utilizes it along with the predicted K^D for VSRDB. Net^R takes $\boldsymbol{X}$, $F^{D,M}$, $\mathbf{f}^{D,M}$, and K^D as inputs. It first computes $F^{R,0}$ through a concatenation of $\boldsymbol{X}$ and $F^{D,M}$ using a RRDB and then refines three features, $F^{R,0}$, $F^{R,0}_w$, and $\mathbf{f}^{R,0}$ through the cascaded M FRMA blocks. Notably, we set $F^{R,0}_w = \mathbf{0}$ and $\mathbf{f}^{R,0} = \mathbf{f}^{D,M}$ in this case. During this FRMA process, each $F^{R,i}_w$ is globally adjusted based on both $F^{R,0}_c$ and the adjusted kernel $k^{D,i}$ through CO and DA attentions, where $k^{D,i}$ represents the degradation features adjusted by convolutions from K^D. Subsequently, $\mathbf{f}^{R,M}$ is used to generate an image flow-mask pair $\mathbf{f}^X \in \mathbb{R}^{T\times H\times W\times(2+1)}$ for $\boldsymbol{X}$, while $F^{R,M}_w$ is used to generate the high-frequency detail $\hat{Y}_r$ and the pixel-wise motion-aware $\times s$ upsampling and deblurring (*i.e.* restoration) kernels $K^R \in \mathbb{R}^{T\times H\times W\times s^2 k_r^2}$ for warped $\boldsymbol{X}$, where k_r denotes the restoration kernel size. $\hat{Y}_r$ is generated by stacked convolution and pixel shuffle [59] (*High Freq. Recon.* Block in Fig. 3). The pixel-wise kernels K^R are normalized with respect to all kernels at temporally co-located positions over $\boldsymbol{X}$, similar to [38]. Finally, $\hat{Y}_c$ can be obtained as $\hat{Y}_c = \hat{Y}_r + (\mathcal{W}(\boldsymbol{X}, \mathbf{f}^X) \circledast K^R) \uparrow_s$, where $\circledast \uparrow_s$ represents proposed flow-guided $\times s$ dynamic upsampling at each pixel location based on Eq. 6. Furthermore, $F^{D,R}$ is also mapped to the image domain through 3D convolution, similar to Net^D, to generate $\hat{X}^R_{Sharp} \in \mathbb{R}^{T\times H\times W\times 3}$, which is only used in FMA-Net training.

3.5. Training Strategy

We employ a two-stage training strategy to train the FMA-Net. Net^D is first pre-trained with the loss L_D as:

Methods	# Params (M)	Runtime (s)	REDS4 PSNR ↑ / SSIM ↑ / tOF ↓
Super-Resolution + Deblurring			
SwinIR [46] + Restormer [73]	11.9 + 26.1	0.320 + 1.121	24.33 / 0.7040 / 4.82
HAT [10] + FFTformer [41]	20.8 + 16.6	0.609 + 1.788	24.22 / 0.7091 / 4.40
BasicVSR++ [9] + RVRT [48]	7.3 + 13.6	0.072 + 0.623	24.92 / 0.7604 / 3.49
FTVSR [56] + GShiftNet [43]	45.8 + 13.0	0.527 + 2251	24.72 / 0.7415 / 3.69
Deblurring + Super-Resolution			
Restormer [73] + SwinIR [46]	26.1 + 11.9	0.078 + 0.320	24.30 / 0.7085 / 4.49
FFTformer [41] + HAT [10]	16.6 + 20.8	0.124 + 0.609	24.21 / 0.7111 / 4.38
RVRT [48] + BasicVSR++ [9]	13.6 + 7.3	0.028 + 0.072	24.79 / 0.7361 / 3.66
GShiftNet [43] + FTVSR [56]	13.0 + 45.8	0.102 + 0.527	23.47 / 0.7044 / 3.98
Joint Video Super-Resolution and Deblurring			
HOFFR [18]	3.5	0.500	27.24 / 0.7870 / -
Restormer* [73]	26.5	0.081	27.29 / 0.7850 / 2.71
GShiftNet* [43]	13.5	0.185	25.77 / 0.7275 / 2.96
BasicVSR++* [9]	7.3	0.072	27.06 / 0.7752 / 2.70
RVRT* [48]	12.9	0.680	27.80 / 0.8025 / 2.40
FMA-Net (Ours)	9.6	0.427	**28.83 / 0.8315 / 1.92**

Table 1. Quantitative comparison on REDS4 for $\times 4$ VSRDB. All results are calculated on the RGB channel. **Red** and blue colors indicate the best and second-best performance, respectively. Runtime is calculated on an LR frame sequence of size 180×320. The superscript * indicates that the model is retrained on the REDS [52] training dataset for VSRDB.

$$L_D = l_1(\hat{X}_c, X_c) + \lambda_1 \sum_{t=-N}^{+N} l_1(\mathcal{W}(Y_{t+c}, s \cdot (\mathbf{f}^Y_{t+c} \uparrow_s)), Y_c) + \lambda_2 l_1(f^Y, f^Y_{RAFT}) + \lambda_3 \underbrace{l_1(\hat{X}^D_{Sharp}, X_{Sharp})}_{\text{Temporal Anchor (TA) loss}}, \tag{8}$$

where f^Y represents the image optical flow contained in $\mathbf{f}^Y$, and f^Y_{RAFT} denotes the pseudo-GT optical flow generated by a pre-trained RAFT [63] model. X_{Sharp} is the sharp LR sequence obtained by applying bicubic downsampling to $\boldsymbol{Y}$. The first term on the right side in Eq. 8 is the reconstruction loss, the second term is the warping loss for optical flow learning in $\boldsymbol{Y}$ from center frame index c to $c+t$, and the third term is the loss using RAFT pseudo-GT for further refining the optical flow.

Temporal Anchor (TA) Loss. Finally, to boost performance, we propose a TA loss, the last term on the right side in Eq. 8. This loss sharpens F^D while keeping each feature temporally anchored for the corresponding frame index, thus constraining the solution space according to our intention to distinguish warped and unwarped features.

After pre-training, the FMA-Net in Fig. 3 is jointly trained as the second stage training with the total loss L_{total}:

$$L_{total} = l_1(\hat{Y}_c, Y_c) + \lambda_4 \sum_{t=-N}^{+N} l_1(\mathcal{W}(X_{t+c}, \mathbf{f}^X_{t+c}), X_c) + \lambda_5 \underbrace{l_1(\hat{X}^R_{Sharp}, X_{Sharp})}_{\text{Temporal Anchor (TA) loss}} + \lambda_6 L_D, \tag{9}$$

where the first term on the right side is the restoration loss, and the second and third terms are identical to the second and forth terms in Eq. 8, except for their applied domains.

4. Experiment Results

Implementation details. We train the FMA-Net using the Adam optimizer [40] with a mini-batch size of 8. The initial learning rate is set to 2×10^{-4}, and reduced by half at 70%, 85%, and 95% of total $300K$ iterations in each training stage. The training LR patch size is 64×64, the number of FRMA blocks is $M = 4$, the number of multi-flow-mask pairs is $n = 9$, and the kernel sizes k_d and k_r are 20 and 5, respectively. The coefficients $[\lambda_i]_{i=1}^6$ in Eqs. 8 and 9 are determined through grid searching, with λ_2 set to 10^{-4} and all other values set to 10^{-1}. We consider $T = 3$ (that is, $N = 1$) and $s = 4$ in our experiments. Additionally, we adopted the multi-Dconv head transposed attention (MDTA) and Gated-Dconv feed-forward network (GDFN) modules proposed in Restormer [73] for the attention and feed-forward network in our multi-attention block.

Datasets. We train FMA-Net using the REDS [52] dataset which consists of realistic and dynamic scenes. Following previous works [45, 50, 70], we use REDS4 [1] as the test set, while the remaining clips are for training. Also, to evaluate generalization performance, we employ the GoPro [51] and YouTube datasets as test sets alongside REDS4. For the GoPro dataset, we applied bicubic downsampling to its blurry version to evaluate VSRDB. As for the YouTube dataset, we selected 40 YouTube videos of different scenes with a resolution of $720 \times 1,280$ at 240fps, including extreme scenes from various devices. Subsequently, we temporally and spatially downsampled them, similar to previous works [23, 55, 58], resulting in blurry 30 fps of 180×320 size.

Evaluation metrics. We use PSNR and SSIM [71] to evaluate the quality of images generated by the networks, and tOF [12, 55] to evaluate temporal consistency. We also compare the model sizes and runtime.

[1] Clips 000, 011, 015, 020 of the REDS training set.

4.1. Comparisons with State-of-the-Art Methods

To achieve VSRDB, we compare our FMA-Net with the very recent SOTA methods: two single-image SR models (SwinIR [46] and HAT [10]), two single-image deblurring models (Restormer [73] and FFTformer [41]), two VSR models (BasicVSR++ [7] and FTVSR [56]), two video deblurring models (RVRT [48] and GShiftNet [43]), and one VSRDB model (HOFFR [18]). Also, we retrain one single-image model (Restormer* [73]) and three video models (BasicVSR++* [9], GShiftNet* [43], and RVRT* [48]) using our training dataset to perform VSRDB for a fair comparison. It should be noted that Restormer* [73] is modified to receive concatenated T frames in the channel dimension for video processing instead of a single frame, and we added a pixel-shuffle [59] block at the end to enable SR.

Table 1 shows the quantitative comparisons for the test set, REDS4. It can be observed in Table 1 that: (i) the sequential approaches of cascading SR and deblurring result in error propagation from previous models, leading to a significant performance drop, and the use of two models also increase memory and runtime costs; (ii) the VSRDB methods consistently demonstrate superior overall performance compared to the sequential cascade approaches, indicating that the two tasks are highly inter-correlated; and (iii) our FMA-Net *significantly* outperforms all SOTA methods including five joint VSRDB methods in terms of PSNR, SSIM, and tOF. Specifically, our FMA-Net achieves improvements of 1.03 dB and 1.77 dB over the SOTA algorithms, RVRT* [48] and BasicVSR++* [9], respectively. The clip-by-clip analyses for REDS4 and the results of all possible combinations of the sequential cascade approaches can be found in the *Supplemental*, including demo videos.

Methods	GoPro PSNR ↑ / SSIM ↑ / tOF ↓	YouTube PSNR ↑ / SSIM ↑ / tOF ↓
Restormer* [73]	26.29 / 0.8278 / 3.66	23.94 / 0.7682 / 2.87
GShiftNet* [43]	25.37 / 0.7922 / 3.95	24.44 / 0.7683 / 2.96
BasicVSR++* [9]	25.19 / 0.7968 / 4.04	23.84 / 0.7660 / 2.97
RVRT* [48]	25.99 / 0.8267 / 3.55	23.53 / 0.7588 / 2.78
FMA-Net (Ours)	**27.65 / 0.8542 / 3.31**	**26.02 / 0.8067 / 2.63**

Table 2. Quantitative comparison on GoPro [51] and YouTube test sets for ×4 VSRDB.

Table 2 shows the quantitative comparisons on GoPro [51] and YouTube test sets for *joint* models trained on REDS [52]. When averaged across both test sets, our FMA-Net achieves a performance boost of 2.08 dB and 1.93 dB over RVRT* [48] and GShiftNet* [43], respectively. This demonstrates that our FMA-Net has good generalization in addressing spatio-temporal degradation generated from various scenes across diverse devices. Figs. 1(a) and 5 show the visual results on three test sets, showing that the images generated by our FMA-Net are visually sharper than those by other methods.

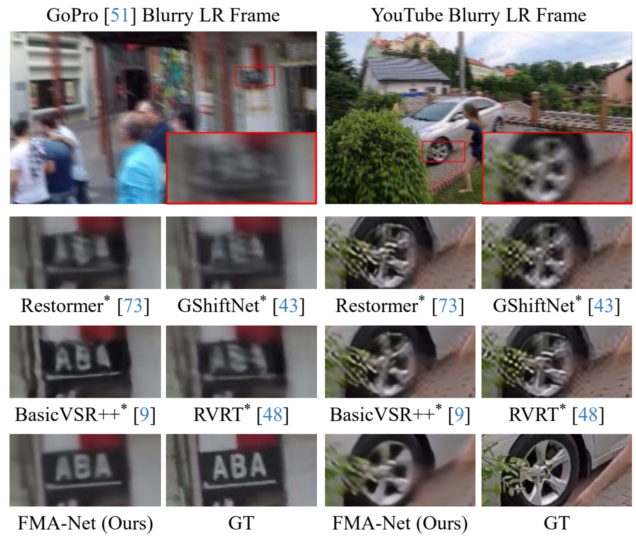


Figure 5. Visual results of different methods on REDS4 [52], GoPro [51], and YouTube test sets. *Best viewed in zoom.*

4.2. Ablation Studies

We analyze the effectiveness of the components in our FMA-Net through ablation studies for which we train the models on REDS [52] and test them on REDS4.

Effect of flow-guided dynamic filtering (FGDF). Table 3 shows the performance of Net^D and Net^R based on the degradation kernel size k_d and two dynamic filtering methods: the conventional dynamic filtering in Eq. 5 and the FGDF in Eq. 6. Average motion magnitude refers to the average absolute optical flow [63] magnitude between the two consecutive frames. Table 3 reveals the following observations: (i) conventional dynamic filtering [33, 35, 38] is not effective in handling large motion, resulting in a significant performance drop as the degree of motion magnitude increases; (ii) our proposed FGDF demonstrates better reconstruction and restoration performance than the conventional dynamic filtering for all ranges of motion magnitudes. This performance difference becomes more pronounced as the degree of motion magnitude increases. For $k_d = 20$, when the average motion magnitude is above 40, the proposed FGDF achieves a restoration performance improvement of

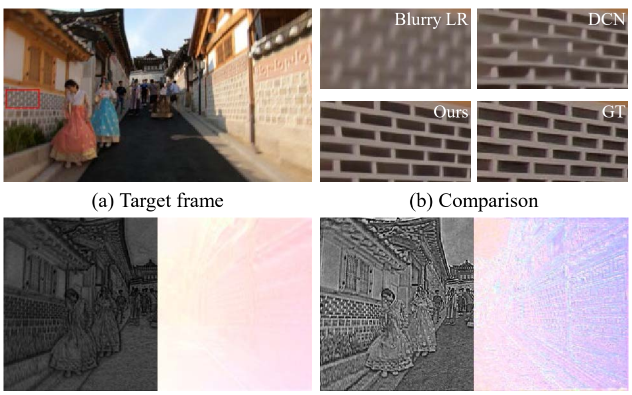
k_d	f	Network	Average Motion Magnitude [0, 20)	[20, 40)	≥ 40	Total
10	✗	Net^D	44.97 / 0.055	39.81 / 0.245	32.04 / 0.871	43.14 / 0.128
		Net^R	27.85 / 1.713	27.51 / 3.922	24.69 / 6.857	27.69 / 2.489
	✓	Net^D	45.38 / 0.049	42.18 / 0.165	37.72 / 0.474	44.25 / 0.092
		Net^R	28.64 / 1.436	28.46 / 3.469	25.54 / 6.558	28.52 / 2.157
20	✗	Net^D	45.94 / 0.047	42.02 / 0.193	35.50 / 0.689	44.53 / 0.104
		Net^R	28.10 / 1.566	27.54 / 3.835	24.24 / 6.989	27.86 / 2.365
	✓	Net^D	46.57 / 0.041	43.49 / 0.151	38.23 / 0.430	45.46 / 0.082
		Net^R	28.91 / 1.289	28.91 / 3.057	26.17 / 5.841	28.83 / 1.918
30	✗	Net^D	46.25 / 0.042	42.95 / 0.161	37.53 / 0.464	45.07 / 0.087
		Net^R	28.30 / 1.589	28.10 / 3.589	25.58 / 6.258	28.19 / 2.292
	✓	Net^D	**46.89 / 0.037**	**44.12 / 0.133**	**39.30 / 0.349**	**45.90 / 0.072**
		Net^R	**28.91 / 1.283**	**28.98 / 3.013**	**26.37 / 5.666**	**28.89 / 1.897**

Table 3. Ablation study on the FGDF (PSNR↑ / tOF↓).

Methods	# Params (M)	Runtime (s)	Net^R (sharp HR $\hat{Y}_c$) PSNR ↑ / SSIM ↑ / tOF ↓
The number of multi-flow-mask pairs n			
(a) $n = 1$	9.15	0.424	28.24 / 0.8151 / 2.224
(b) $n = 5$	9.29	0.429	28.60 / 0.8258 / 2.054
Deformable Convolution [13]			
(c) w/ DCN (#offset = 9)	10.13	0.426	28.52 / 0.8225 / 2.058
Loss Function and Training Strategy			
(d) w/o RAFT & TA Loss	9.61	0.434	28.68 / 0.8274 / 2.003
(e) w/o TA Loss	9.61	0.434	28.73 / 0.8288 / 1.956
(f) End-to-End Learning	9.61	0.434	28.39 / 0.8190 / 2.152
Multi-Attention			
(g) self-attn [73] + SFT [68]	9.20	0.415	28.50 / 0.8244 / 2.039
(h) CO attn + SFT [68]	9.20	0.416	28.58 / 0.8262 / 1.938
(i) self-attn [73] + DA attn	9.61	0.434	28.80 / 0.8298 / 1.956
(j) Ours	9.61	0.434	**28.83** / **0.8315** / **1.918**

Table 4. Ablation study on the components in FMA-Net.

1.93 dB compared to the conventional method. Additional analysis for Table 3 can be found in the *Supplemental*.



(a) Target frame (b) Comparison

(c) Offset & mask (DCN) (d) Optical flow & mask (Ours)

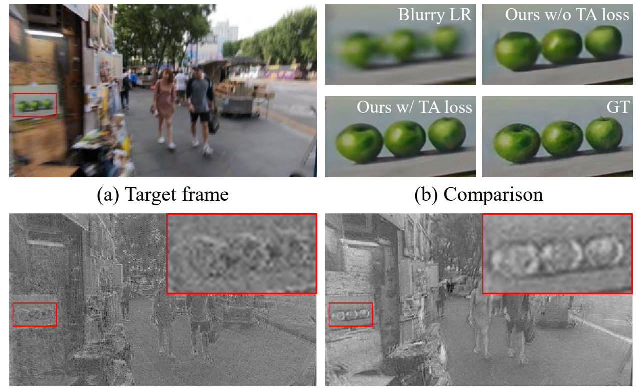
Figure 6. Offsets and mask (DCN [13]) vs. Multi-flow-mask pairs **f** (FMA-Net). Analysis of the multi-flow-mask pairs **f** compared to the DCN [13]. The offset and optical flow maps with their largest deviations are visualized with their corresponding masks.

Design choices for FMA-Net. Table 4 shows the ablation experiment results for the components of our FMA-Net:

(i) Table 4(a-b, j) shows the performance change in the number of multi-flow-mask pairs n. As n increases, there is a significant performance improvement in Net^R, accompanied by a slight increase in memory cost. The best results are observed in Table 4(j) with $n = 9$;

(ii) Table 4(c) shows the result of implicitly utilizing motion information through DCN [13] instead of using our multi-flow-mask pairs **f**. With the same number of offsets and n, our method achieves 0.31 dB higher performance compared to using DCN. This is due to the utilization of the self-induced sharper optical flows and occlusion masks, as shown in Fig. 6;

(iii) Table 4(d-f, j) shows the performance change depending on the used loss functions and training strategies. The 'RAFT' in Table 4(d) refers to the use of $l_1(f^Y, f^Y_{RAFT})$ in Eq. 8 for Net^D. The effectiveness of our loss functions in Eqs. 8 and 9 can be observed from Table 4(d-e, j), especially with our new TA loss which anchors and sharpens each feature with respect to the corresponding frame index as shown in Fig. 7, leading to 0.1 dB PSNR improvement (Table 4(e) and (j)). Also, our two-stage training strategy achieves 0.44 dB improvement (Table 4(f));



(a) Target frame (b) Comparison

(c) Average $F_c^{R,M}$ w/o TA loss (d) Average $F_c^{R,M}$ w/ TA loss

Figure 7. Analysis of the TA loss.

(iv) For the ablation study on our multi-attention (CO + DA attentions), we replaced them with self-attention [73] and spatial feature transform (SFT) [68] layer that is a SOTA feature modulation module, respectively. The results in Table 4(g-j) clearly demonstrate that our multi-attention approach outperforms the SOTA self-attention and modulation methods with a 0.33 dB improvement.

Similarly, Net^D exhibits the same tendencies as Net^R. See the results and analysis in the *Supplemental*.

5. Conclusion

We propose a novel VSRDB framework, called FMA-Net, based on our novel FGDF and FRMA. We iteratively update features including self-induced optical flow through stacked FRMA blocks, and predict a flow-mask pair with flow-guided dynamic filters, which enables the network to capture and utilize small-to-large motion information. The FGDF leads to a dramatic performance improvement compared to conventional dynamic filtering. Additionally, the newly proposed temporal anchor (TA) loss facilitates model training by temporally anchoring and sharpening unwarped features. Extensive experiments demonstrate that our FMA-Net achieves best performances for diverse datasets with significant margins compared to the recent SOTA methods.

Acknowledgement. This work was supported by the IITP grant funded by the Korea government (MSIT): No. 2021-0-00087, Development of high-quality conversion technology for SD/HD low-quality media and No. RS2022-00144444, Deep Learning Based Visual Representational Learning and Rendering of Static and Dynamic Scenes.

References

[1] Andreas Aakerberg, Kamal Nasrollahi, and Thomas B Moeslund. Real-world super-resolution of face-images from surveillance cameras. *IET Image Processing*, 16(2):442–452, 2022. 1

[2] Waqar Ahmad, Hazrat Ali, Zubair Shah, and Shoaib Azmat. A new generative adversarial network for medical images super resolution. *Scientific Reports*, 12(1):9533, 2022. 1

[3] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021. 5

[4] Yuval Bahat, Netalee Efrat, and Michal Irani. Non-uniform blind deblurring by reblurring. In *Proceedings of the IEEE international conference on computer vision*, pages 3286–3294, 2017. 1

[5] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4778–4787, 2017. 2, 1

[6] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv*, 2021. 2

[7] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4947–4956, 2021. 2, 3, 7

[8] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, pages 973–981, 2021. 4

[9] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022. 2, 6, 7, 1, 5

[10] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 2, 6, 7

[11] Benjamin Naoto Chiche, Arnaud Woiselle, Joana Frontera-Pons, and Jean-Luc Starck. Stable long-term recurrent video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 837–846, 2022. 2

[12] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Transactions on Graphics (TOG)*, 39(4):75–1, 2020. 6

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2, 5, 8, 3

[14] Mallesham Dasari, Arani Bhattacharya, Santiago Vargas, Pranjal Sahu, Aruna Balasubramanian, and Samir R Das. Streaming 360-degree videos using super-resolution. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1977–1986. IEEE, 2020. 1

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 2

[16] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on image processing*, 20(7):1838–1857, 2011. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[18] Ning Fang and Zongqian Zhan. High-resolution optical flow and frame-recurrent network for video super-resolution and deblurring. *Neurocomputing*, 489:128–138, 2022. 2, 3, 6, 7, 5

[19] Zhenxuan Fang, Fangfang Wu, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Self-supervised non-uniform kernel estimation with flow-based motion prior for blind image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18105–18114, 2023. 2

[20] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019. 2

[21] Hayit Greenspan. Super-resolution in medical imaging. *The computer journal*, 52(1):43–63, 2009. 1

[22] Ankit Gupta, Neel Joshi, C Lawrence Zitnick, Michael Cohen, and Brian Curless. Single image deblurring using motion density functions. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, pages 171–184. Springer, 2010. 4

[23] Akash Gupta, Abhishek Aich, and Amit K Roy-Chowdhury. Alanet: Adaptive latent attention network for joint video deblurring and interpolation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 256–264, 2020. 6

[24] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3897–3906, 2019. 2

[25] Stefan Harmeling, Hirsch Michael, and Bernhard Schölkopf. Space-variant single-image blind deconvolution for removing camera shake. *Advances in Neural Information Processing Systems*, 23, 2010. 4

[26] Linyang He, Gang Li, Jinghong Liu, et al. Joint motion deblurring and superresolution from single blurry image. *Mathematical Problems in Engineering*, 2015, 2015. 2

[27] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998. 2

[28] Ping Hu, Simon Niklaus, Stan Sclaroff, and Kate Saenko. Many-to-many splatting for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3553–3562, 2022. 4

[29] Yan Huang, Wei Wang, and Liang Wang. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):1015–1028, 2017. 2

[30] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5426–5434, 2015. 3

[31] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8008–8017, 2020. 2

[32] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 4

[33] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 2, 3, 5, 7, 1

[34] Meiguang Jin, Givi Meishvili, and Paolo Favaro. Learning to extract a video sequence from a single motion-blurred image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6334–6342, 2018. 3

[35] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018. 2, 3, 7, 1

[36] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2

[37] Soo Ye Kim, Jeongyeon Lim, Taeyoung Na, and Munchurl Kim. 3dsrnet: Video super-resolution using 3d convolutional neural networks. *arXiv preprint arXiv:1812.09079*, 2018. 2

[38] Soo Ye Kim, Hyeonjun Sim, and Munchurl Kim. Koalanet: Blind super-resolution using kernel-oriented adaptive local adjustment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10611–10620, 2021. 2, 3, 5, 7, 1

[39] Tae Hyun Kim, Seungjun Nah, and Kyoung Mu Lee. Dynamic video deblurring using a locally adaptive blur model. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2374–2387, 2017. 3

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[41] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 3, 6, 7

[42] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018. 3

[43] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9822–9832, 2023. 2, 3, 6, 7, 5

[44] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10522–10531, 2019. 2

[45] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 335–351. Springer, 2020. 2, 3, 6

[46] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2, 6, 7

[47] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 2, 3

[48] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022. 3, 6, 7, 2, 5

[49] Jiayi Lin, Yan Huang, and Liang Wang. Fdan: Flow-guided deformable alignment network for video super-resolution. *arXiv preprint arXiv:2105.05640*, 2021. 2

[50] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5687–5696, 2022. 2, 3, 6, 1

[51] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 5, 6, 7, 2

[52] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 6, 7, 2, 5, 8, 9, 10

[53] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 670–679, 2017. 2, 3

[54] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE international conference on computer vision*, pages 261–270, 2017. 2, 3

[55] Jihyong Oh and Munchurl Kim. Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In *European Conference on Computer Vision*, pages 198–215. Springer, 2022. 1, 4, 5, 6

[56] Zhongwei Qiu, Huan Yang, Jianlong Fu, and Dongmei Fu. Learning spatiotemporal frequency-transformer for compressed video super-resolution. In *ECCV*, 2022. 6, 7

[57] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6626–6634, 2018. 2

[58] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5114–5123, 2020. 5, 6

[59] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5, 7

[60] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: extreme video frame interpolation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 4

[61] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1279–1288, 2017. 5

[62] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 4472–4480, 2017. 2, 1

[63] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 6, 7

[64] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3360–3369, 2020. 2, 3

[65] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 3

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[67] Longguang Wang, Yulan Guo, Li Liu, Zaiping Lin, Xinpu Deng, and Wei An. Deep video super-resolution using hr optical flow estimation. *IEEE Transactions on Image Processing*, 29:4323–4336, 2020. 2

[68] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8, 1, 3

[69] 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 *The European Conference on Computer Vision Workshops (ECCVW)*, 2018. 5

[70] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2, 6

[71] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[72] Si Xi, Jia Wei, and Weidong Zhang. Pixel-guided dual-branch attention network for joint image deblurring and super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 532–540, 2021. 2

[73] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 3, 5, 6, 7, 8, 2

[74] Dongyang Zhang, Zhenwen Liang, and Jie Shao. Joint image deblurring and super-resolution with attention dual supervised network. *Neurocomputing*, 412:187–196, 2020. 2

[75] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Wei Liu, and Hongdong Li. Adversarial spatio-temporal learning for video deblurring. *IEEE Transactions on Image Processing*, 28(1):291–301, 2018. 1, 3

[76] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3262–3271, 2018. 2

[77] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020. 1

[78] Liangpei Zhang, Hongyan Zhang, Huanfeng Shen, and Pingxiang Li. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3):848–859, 2010. 1

[79] Xinyi Zhang, Hang Dong, Zhe Hu, Wei-Sheng Lai, Fei Wang, and Ming-Hsuan Yang. Gated fusion network for joint image deblurring and super-resolution. *arXiv preprint arXiv:1807.10806*, 2018. 2

[80] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2

[81] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 4

[82] Yinjie Zhang, Yuanxing Zhang, Yi Wu, Yu Tao, Kaigui Bian, Pan Zhou, Lingyang Song, and Hu Tuo. Improving quality of experience by adaptive video streaming with super-resolution. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1957–1966. IEEE, 2020. 1

[83] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2482–2491, 2019. 2, 3