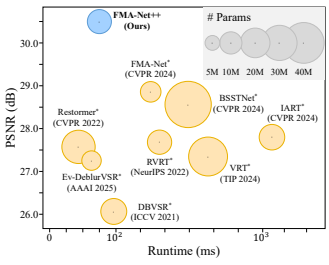
FMA-Net++: Motion- and Exposure-Aware Joint Video Super-Resolution and Deblurring

Geunhyuk Youk[1], Jihyong Oh[2†], and Munchurl Kim[1†]

[1] KAIST, Republic of Korea
{rmsgurkjg, mkimee}@kaist.ac.kr
[2] CMLab, Chung-Ang University, Republic of Korea
jihyongoh@cau.ac.kr

https://kaist-viclab.github.io/fmanetpp_site/

(a) Qualitative comparison on challenging real-world videos.



(b) Quantitative comparison on the GoPro [34] dataset.

Fig. 1: FMA-Net++ outperforms state-of-the-art methods in real-world qualitative results and quantitative benchmarks for joint video super-resolution and deblurring.

Abstract. Joint video super-resolution and deblurring (VSRDB) requires both efficient long-range temporal modeling and robustness to frame-wise exposure-duration variation, which changes the extent of motion blur across video frames. We propose FMA-Net++, a non-recurrent, sequence-level framework built from Hierarchical Refinement with Bidirectional Aggregation (HRBA) blocks. By stacking HRBA blocks, FMA-Net++ processes video frames in parallel while hierarchically expanding the temporal receptive field, avoiding the limited temporal receptive field of sliding-window designs and the sequential bottleneck of recurrent ones. To handle exposure-duration-dependent blur, we introduce an Exposure Time-aware Modulation (ETM) layer that conditions HRBA features on exposure embeddings from an Exposure Time-aware Feature Extractor (ETE). The conditioned features guide an exposure-aware flow-guided dynamic filtering module to predict motion- and exposure-aware degradation kernels. FMA-Net++ decouples degradation learning from restoration: the former predicts degradation priors and the latter exploits them for efficient high-resolution restoration. To evaluate VSRDB under controlled exposure-duration variation, we introduce the

† Co-corresponding authors.

REDS-ME (multi-exposure) and REDS-RE (random-exposure) benchmarks. Trained solely on synthetic data, FMA-Net++ achieves state-of-the-art accuracy and temporal consistency on these benchmarks. It further shows strong out-of-distribution performance on GoPro and challenging real-world videos, while outperforming recent methods in both restoration quality and inference speed.

Keywords: Joint Video Super-Resolution Deblurring · Temporal Modeling · Dynamic Exposure

1 Introduction

Joint video super-resolution and deblurring (VSRDB) [11,17,49] aims to restore sharp high-resolution (HR) videos from blurry low-resolution (LR) inputs. In practice, blurry LR videos are common, and treating SR or deblurring separately is inadequate: SR cannot remove motion blur, while deblurring cannot recover high-frequency details, motivating a joint VSRDB approach [35,49]. The physical degradation process underlying these blurry LR videos is driven by two deeply intertwined factors: the *motion field* determines the spatial patterns of blur, and the *exposure time* controls its temporal extent and intensity [33,34,46]. Compounding this, camera auto-exposure mechanisms vary the exposure dynamically across frames [20,46]. This continuous fluctuation causes the severity of motion blur to change drastically within a single video sequence, resulting in complex, spatio-temporally variant degradations that standard restoration methods struggle to model.

While significant progress has been made in various video restoration tasks [7, 27, 36, 49], most existing methods assume a *fixed exposure time*. This assumption severely limits their robustness, as they struggle to handle the dynamically changing blur severity arising from continuous exposure variations. For instance, VSR [5,7,16,24,29] and video deblurring [23,51–53,55] approaches may produce artifacts or temporally inconsistent results when faced with exposure shifts. Even methods designed for unknown degradations, such as Blind VSR [3,22,36], typically assume spatially-invariant kernels and fail to account for the coupled effect of motion and varying exposure. Furthermore, recent joint VSRDB approaches like FMA-Net [49], despite handling motion-dependent degradation, remain constrained by this fixed-exposure assumption. Thus, VSRDB methods that explicitly address frame-wise exposure-duration variation are critically needed.

Moreover, beyond the exposure issue, prevailing temporal modeling strategies face inherent limitations: Sliding-window architectures [16, 24, 40, 41] tend to suffer from limited temporal receptive fields, while recurrent architectures [5, 7, 12, 28, 29] lack parallelizability due to sequential bottlenecks, as conceptually compared in Fig. 2(a). Although recent transformer-based models like VRT [25] enable parallel processing, they incur heavy computational and memory burdens. To overcome these limits and address the aforementioned exposure variation, we introduce FMA-Net++, an efficient sequence-level framework that couples long-range temporal modeling with exposure-aware degradation estimation.

The core architectural unit of FMA-Net++ is the **Hierarchical Refinement with Bidirectional Aggregation (HRBA)** block. Instead of relying on restrictive sliding windows, such as those in FMA-Net [49], or inherently sequential recurrent structures [5, 7, 29], stacking HRBA blocks enables *sequence-level parallelization* and *hierarchically expands the temporal receptive field* to capture long-range dependencies. To handle exposure-duration-dependent blur, each HRBA block includes an **Exposure Time-aware Modulation (ETM)** layer that conditions features on per-frame exposure embeddings, producing representations rich in both temporal context and exposure information. Leveraging these representations, an exposure-aware Flow-Guided Dynamic Filtering (FGDF) module estimates *motion- and exposure-aware degradation kernels*. Architecturally, we decouple degradation learning from restoration: the former predicts these rich priors, and the latter utilizes them to restore sharp HR frames efficiently, as illustrated in Fig. 2(b).

To systematically evaluate VSRDB under controlled exposure-duration variation, we construct two new benchmarks, REDS-ME (multi-exposure) and REDS-RE (random-exposure). Trained solely on synthetic data, FMA-Net++ achieves state-of-the-art accuracy and temporal consistency on our new benchmarks and the GoPro [34] dataset. It outperforms recent methods in both restoration quality and inference speed, and we further validate robustness on real-world videos using qualitative results and no-reference metrics (see Fig. 1).

The main contributions of this work are as follows:

- We design an efficient, sequence-level architecture based on **Hierarchical Refinement with Bidirectional Aggregation (HRBA)** blocks. By hierarchically expanding temporal receptive fields without sequential bottlenecks, HRBA effectively captures long-range temporal dependencies while enabling sequence-level parallel processing.
- We formulate the physical coupling of motion blur and *frame-wise exposure-duration variation* in VSRDB. To address this, we propose an **Exposure Time-aware Modulation (ETM)** layer that conditions features on per-frame exposure embeddings, driving the estimation of motion- and exposure-aware degradation kernels.
- We introduce two new benchmarks, **REDS-ME** and **REDS-RE**, for systematic evaluation under dynamic exposure. Extensive experiments demonstrate that FMA-Net++ achieves state-of-the-art performance, high computational efficiency, and strong generalization to challenging real-world videos.

2 Related Work

2.1 Joint Video Super-Resolution and Deblurring

VSRDB tackles the challenging task of jointly restoring sharp HR videos from blurry LR inputs, where blur degradations can be shaped by the coupled effects of motion and exposure. While single-task approaches for VSR [5, 7, 16, 24, 27, 42] or video deblurring [23, 27, 52, 53, 56] have advanced, applying them sequentially

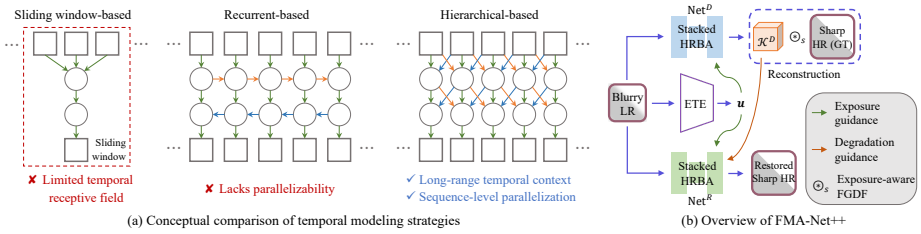


Fig. 2: Conceptual illustration and overview of the FMA-Net++ framework.

often amplifies artifacts [35, 49]. However, specific methods tackling this joint VSRDB challenge remain scarce. Early approaches like HOFFR [11] struggle with spatially variant blur due to standard CNN limitations. Although FMA-Net [49] handles motion-dependent degradation via Flow-Guided Dynamic Filtering, it is constrained by a sliding-window design and a fixed-exposure assumption. FMA-Net++ departs from FMA-Net in two key aspects: it replaces sliding-window restoration with a sequence-level HRBA backbone, and it generalizes degradation estimation to jointly model motion and exposure. Recently, Ev-DeblurVSR [17] utilized auxiliary event streams for VSRDB, but it requires non-standard event data and still assumes a fixed exposure time (a limitation explicitly discussed in [17]), and therefore does not address RGB-only VSRDB under frame-wise exposure variation. These gaps motivate our sequence-level, exposure-aware approach for robust VSRDB using only standard RGB inputs.

2.2 Temporal Modeling in Video Restoration

Effectively modeling long-range temporal dependencies is crucial for video restoration tasks like VSR. However, prevailing strategies face inherent architectural trade-offs. Sliding-window approaches [16, 24, 38, 42, 49] operate on fixed local neighborhoods, constraining input flexibility and limiting the capture of long-range context. Conversely, recurrent methods [5, 7, 27, 29] propagate information sequentially. While they enable longer temporal aggregation, they are inherently constrained by sequential processing, which restricts parallelization and makes them prone to vanishing gradients over long sequences [9, 29]. Recently, transformer-based models [4, 25] have been explored. Although VRT [25] achieves sequence-level parallelization via a shifted-window mechanism [30], its heavy architecture incurs massive computational and memory costs. Furthermore, most of these works target sharp inputs, lacking robustness to complex degradations. This landscape motivates the need for sequence-level backbones that hierarchically expand temporal receptive fields while enabling efficient parallel processing.

2.3 Exposure Time-Aware Restoration

In modern camera systems, auto-exposure mechanisms dynamically vary exposure across frames, yielding spatio-temporally variant blur that fixed-exposure models cannot faithfully capture [20, 37, 46]. While recent efforts in related tasks (*e.g.*, video deblurring [20, 37] and frame interpolation [37, 46, 54]) estimate exposure or exploit auxiliary sensing (events) to guide restoration, they do not

explicitly model the *joint* effects of motion and exposure within the VSRDB setting, and event-dependent designs limit practicality for standard RGB videos. To address these limitations, we propose an **Exposure Time-aware Modulation (ETM)** layer that injects per-frame exposure embeddings into temporal features. These conditioned features then drive our exposure-aware Flow-Guided Dynamic Filtering (FGDF), estimating *motion- and exposure-aware degradation kernels*. This design achieves robust VSRDB using only standard RGB inputs, integrating seamlessly with our sequence-level backbone.

3 Method

3.1 Problem Formulation

We address joint VSRDB under frame-wise varying exposure, where the per-frame exposure time $\Delta t_{e,i}$ is unknown at test time. Given a blurry LR video $\boldsymbol{X} = \{\boldsymbol{X}_i\}_{i=1}^{T} \in \mathbb{R}^{T\times H\times W\times 3}$, our goal is to restore the sharp HR video $\hat{\boldsymbol{Y}} = \{\hat{\boldsymbol{Y}}_i\}_{i=1}^{T} \in \mathbb{R}^{T\times sH\times sW\times 3}$ with an upscaling factor s.

To motivate an exposure-aware degradation model, we generalize the conventional fixed-exposure blur model [34]. The blurry LR frame $\boldsymbol{X}_i$ at a spatial position $\boldsymbol{p}$ is physically formed by spatially downsampling and temporally integrating the continuous latent sharp signal $\mathcal{S}$ over the exposure interval $\Delta t_{e,i}$ under the continuous motion field $\boldsymbol{M}$:

$$\boldsymbol{X}_i(\boldsymbol{p}) = \mathcal{D}_s\left(\frac{1}{\Delta t_{e,i}}\int_{i\cdot\Delta t}^{i\cdot\Delta t+\Delta t_{e,i}} \mathcal{S}\left(\boldsymbol{q}+\boldsymbol{M}(\boldsymbol{q},\tau),\tau\right)d\tau\right), \tag{1}$$

where $\mathcal{D}_s$ is the spatial downsampling operator, Δt is the frame interval, and $\boldsymbol{q}$ is the HR coordinate corresponding to $\boldsymbol{p}$. Since directly inverting this continuous physical process is intractable, we approximate it with a discrete, learnable formulation using a spatio-temporally variant degradation kernel $\mathcal{K}_i$:

$$\boldsymbol{X}_i \;\approx\; \mathcal{K}_i *_s \boldsymbol{Y}'_i, \tag{2}$$

where $*_s$ denotes a filtering operation with stride s, and $\boldsymbol{Y}'_i = \{\boldsymbol{Y}_{i-k},\ldots,\boldsymbol{Y}_{i+k}\}$ is a short temporal neighborhood of sharp HR frames for a small k. Conceptually, the ideal degradation kernel $\mathcal{K}_i$ at $\boldsymbol{p}$ depends jointly on the exposure time and the motion field through a complex mapping $\mathcal{F}$:

$$\mathcal{K}_i(\boldsymbol{p}) = \mathcal{F}\left(\Delta t_{e,i}, \{\boldsymbol{M}(\boldsymbol{q},\tau) \mid \boldsymbol{q}\in\Omega(\boldsymbol{p}); \tau\in[i\cdot\Delta t, i\cdot\Delta t+\Delta t_{e,i}]\}\right), \tag{3}$$

where $\Omega(\boldsymbol{p})$ denotes the spatial neighborhood of HR coordinates corresponding to $\boldsymbol{p}$. This formulation captures how dynamic exposure and motion couple to create complex, spatio-temporally variant blur.

Our framework solves this inverse problem of restoring the sharp HR sequence by architecturally decoupling degradation learning from restoration. First, it uses an Exposure Time-aware Feature Extractor and learned optical flow to approximate the properties of $\Delta t_{e,i}$ and $\boldsymbol{M}$, thereby estimating a learnable approximation of $\mathcal{K}_i$. Second, these estimated priors guide the restoration network to reconstruct $\hat{\boldsymbol{Y}}$.

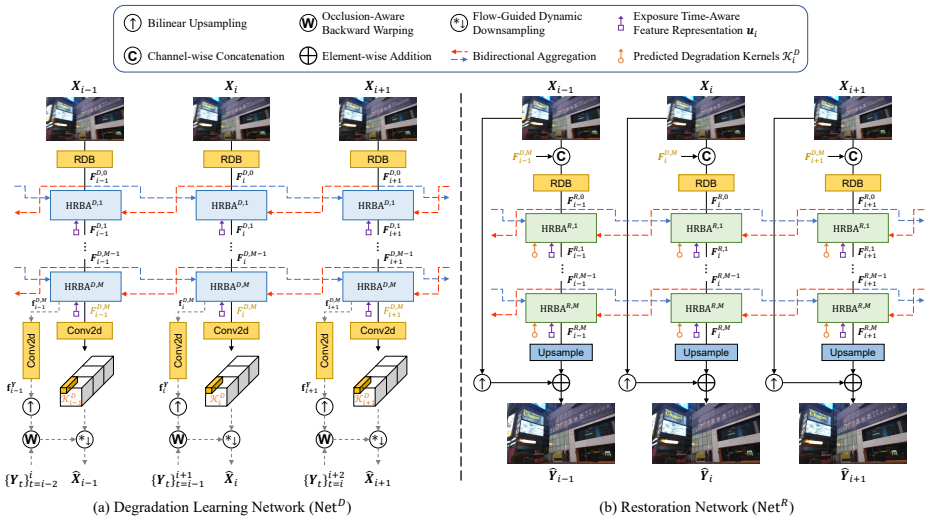


Fig. 3: Architecture of FMA-Net++ for joint VSRDB.

3.2 Overall Architecture of FMA-Net++

Fig. 3 illustrates the overall architecture of FMA-Net++, which consists of two main networks: a **Degradation Learning Network** (Net^D) and a **Restoration Network** (Net^R). Both networks are built upon stacks of **Hierarchical Refinement with Bidirectional Aggregation** (HRBA) blocks. This sequence-level backbone processes all frames in parallel, effectively avoiding sequential bottlenecks while hierarchically expanding the temporal receptive field.

Following the degradation formulation in Eq. 2, our architecture decouples degradation learning from restoration. Given an input blurry LR sequence $\boldsymbol{X}$ and exposure embeddings from a pretrained Exposure Time-aware Feature Extractor (ETE), Net^D first leverages HRBA and an **Exposure Time-aware Modulation** (ETM) layer to estimate degradation priors. These priors then drive the exposure-aware Flow-Guided Dynamic Filtering (FGDF) to model spatio-temporally variant degradations. Finally, Net^R restores the sharp HR video $\hat{\boldsymbol{Y}}$ guided by these priors, improving both accuracy and efficiency.

3.3 Hierarchical Refinement with Bidirectional Aggregation

As the core architectural unit shared by both Net^D and Net^R, the HRBA block overcomes the fundamental trade-offs faced by prior temporal modeling strategies (Sec. 2.2): namely, limited temporal receptive fields in sliding-window methods [42, 49] and the lack of parallelizability in sequential recurrent approaches [7, 27]. By stacking HRBA blocks, our architecture enables *sequence-level parallel processing*. At each refinement level, information from increasingly distant past and future frames is aggregated bidirectionally, thus *hierarchically expanding the temporal receptive field* to effectively capture long-range dependencies.

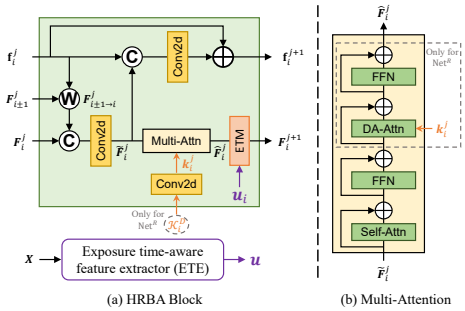


Fig. 4: Details of an HRBA block. (a) Structure of the HRBA block at $(j{+}1)$-*th* refinement step for *i-th* frame (Sec. 3.3). (b) Structure of Multi-Attention. FFN refers to the feed-forward network of the transformer [1, 10].

As shown in Fig. 4(a), each HRBA block iteratively refines the feature map $\boldsymbol{F}_i^j \in \mathbb{R}^{H\times W\times C}$ and a set of multi-flow-mask pairs $\mathbf{f}_i^j \in \mathbb{R}^{2\times H\times W\times(2+1)n}$ for a given frame i at refinement step $j+1$. Specifically, $\mathbf{f}_i^j$ is defined as:

$$\mathbf{f}_i^j \equiv \left\{(\boldsymbol{f}_{i\to i+1}^k,\, \boldsymbol{o}_{i\to i+1}^k), (\boldsymbol{f}_{i\to i-1}^k,\, \boldsymbol{o}_{i\to i-1}^k)\right\}_{k=1}^{n} \tag{4}$$

where n is the number of multi-flow-mask pairs, each containing an optical flow $\boldsymbol{f}$ and corresponding occlusion mask $\boldsymbol{o}$ representing motion towards neighbors $i\pm1$. Keeping multiple motion hypotheses ($n > 1$) enhances robustness under severe blur by providing one-to-many correspondences [6, 14]. The refinement process first computes intermediate features $\tilde{\boldsymbol{F}}_i^j$ via occlusion-aware warping [15, 35] of neighboring features $\boldsymbol{F}_{i\pm1}^j$ using $\mathbf{f}_i^j$, followed by fusion using concatenation and convolution. The multi-flow-mask pairs are then updated residually, $\mathbf{f}_i^{j+1} = \mathbf{f}_i^j + \Delta\mathbf{f}_i^j$, where the residual $\Delta\mathbf{f}_i^j$ is predicted based on $\tilde{\boldsymbol{F}}_i^j$ and $\mathbf{f}_i^j$. The intermediate feature $\tilde{\boldsymbol{F}}_i^j$ is further enhanced through two crucial modules before producing the final output $\boldsymbol{F}_i^{j+1}$.

Multi-Attention. As shown in Fig. 4(b), the multi-attention module employs self-attention [1] to capture spatial dependencies and integrate the aggregated hierarchical temporal context. Within Net^R, it subsequently applies Degradation-Aware (DA) attention. This cross-attention mechanism uses query $\boldsymbol{Q}$ derived from the estimated exposure- and motion-aware degradation kernel $\mathcal{K}_i^D$ (predicted by Net^D), while key $\boldsymbol{K}$ and value $\boldsymbol{V}$ are projected from the self-attention output. This allows Net^R features to adapt specifically to the estimated degradation characteristics of each frame.

Exposure Time-aware Modulation (ETM). To handle frame-wise exposure variation, every HRBA block applies ETM via a lightweight SFT layer [43]. The exposure embedding $\boldsymbol{u}_i$ is provided by the ETE, a ResNet-18 [13] backbone pretrained via supervised contrastive learning [19] to separate exposure anchors in a latent space. Conditioned on $\boldsymbol{u}_i$, a shallow network $\mathcal{M}$ predicts affine parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathcal{M}(\boldsymbol{u}_i)$ and modulates the attention output $\hat{\boldsymbol{F}}_i^j$ as $\boldsymbol{F}_i^{j+1} = (1+\boldsymbol{\alpha}) \odot \hat{\boldsymbol{F}}_i^j + \boldsymbol{\beta}$. This injects exposure information into all refinement stages with negligible overhead, enabling dynamic exposure adaptation.

In summary, compared to sliding windows [16, 24, 38, 42, 49], HRBA accesses long-range context via hierarchical aggregation. Compared to recurrent schemes [5, 7, 27, 29], it avoids sequential dependencies, enabling efficient parallelization and stable training on long sequences.

3.4 Degradation Learning and Restoration Networks

As outlined in Sec. 3.2, our framework comprises two main networks leveraging the HRBA backbone to solve the inverse problem defined in Sec. 3.1.

Degradation Learning Network (Net^D). As shown in the left part of Fig. 3, Net^D aims to estimate degradation priors from the input blurry LR sequence $\boldsymbol{X}$. It processes $\boldsymbol{X}$ through a stack of HRBA blocks with integrated ETM layers, producing refined features $\boldsymbol{F}^{D,M}$ and multi-flow-mask pairs $\mathbf{f}^{D,M}$. From these outputs, Net^D predicts two key priors for each frame $\boldsymbol{X}_i$: (i) image flow-mask pairs $\mathbf{f}_i^{\boldsymbol{Y}} = \left\{\boldsymbol{f}_{i\to i\pm1}^{\boldsymbol{Y}}, \boldsymbol{o}_{i\to i\pm1}^{\boldsymbol{Y}}\right\}$, representing motion between the sharp HR frame $\boldsymbol{Y}_i$ and its neighbors, and (ii) degradation kernels $\mathcal{K}_i^D \in \mathbb{R}^{3\times H\times W\times k_d^2}$.

To apply these predicted kernels, we utilize the Flow-Guided Dynamic Filtering (FGDF) [49] module. Crucially, because $\mathcal{K}_i^D$ is predicted from features already infused with exposure information via the ETM layer, this operation naturally becomes an exposure-aware FGDF. This kernel formulation, representing the degradation from three consecutive sharp HR frames $\{\boldsymbol{Y}_{i-1}, \boldsymbol{Y}_i, \boldsymbol{Y}_{i+1}\}$ to the blurry LR frame $\boldsymbol{X}_i$, follows the design principle of [49] as it offers a robust trade-off between performance and computational cost. The shape of $\mathcal{K}_i^D$ reflects its spatio-temporally variant and position-dependent nature, providing a structured learnable parameterization for modeling complex varying degradations.

To ensure accurate prior estimation, Net^D is trained with a reconstruction objective: the predicted priors must reconstruct the blurry LR frame $\hat{\boldsymbol{X}}_i$ from the ground-truth (GT) sharp HR frames $\boldsymbol{Y}$ as:

$$\hat{\boldsymbol{X}}_i = \left(\mathcal{K}_i^D \circledast_s \{\boldsymbol{Y}_{t\to i}\}_{t=i-1}^{i+1}\right), \tag{5}$$

where $\circledast_s$ denotes the FGDF operation [49] with stride s, defined as:

$$\left(\mathcal{K}_i^D \circledast_s \{\boldsymbol{Y}_{t\to i}\}_{t=i-1}^{i+1}\right)(\boldsymbol{p}) = \sum_{t=i-1}^{i+1}\sum_{k=1}^{k_d^2} \mathcal{K}_{i,t}^D(\boldsymbol{p}, \boldsymbol{p}_k)\, \boldsymbol{Y}_{t\to i}(s\boldsymbol{p} + \boldsymbol{p}_k), \tag{6}$$

where $\boldsymbol{p}_k$ denotes the k-th spatial offset within the $k_d \times k_d$ grid. Unlike conventional dynamic filtering that uses fixed spatial neighborhoods, FGDF explicitly performs filtering *along motion trajectories* by dynamically sampling features guided by the estimated optical flow. By utilizing our exposure-conditioned kernel $\mathcal{K}_i^D$, this operation naturally achieves jointly motion- and exposure-aware filtering, making Eq. 5 a *practical, learnable approximation* of the conceptual degradation model in Eq. 2. The warped HR frame $\boldsymbol{Y}_{t\to i}$ is defined as:

$$\boldsymbol{Y}_{t\to i} = \begin{cases} \boldsymbol{Y}_i, & \text{if } t = i \\ \mathcal{W}(\boldsymbol{Y}_t, \boldsymbol{f}_{i\to t}^{\boldsymbol{Y}}, \boldsymbol{o}_{i\to t}^{\boldsymbol{Y}}), & \text{if } t = i \pm 1 \end{cases} \tag{7}$$

Fig. 5: Example frames from our REDS-ME dataset across five exposure levels and two scenes. Each column corresponds to an exposure ratio from 5 : 1 (shortest exposure) to 5 : 5 (longest exposure). Longer exposures lead to increasingly severe motion blur, illustrating the controlled effect of exposure variation on blur extent.

where $\mathcal{W}$ denotes the occlusion-aware backward warping [47].

Restoration Network (Net^R). As shown in the right part of Fig. 3, Net^R performs the final restoration, taking the blurry LR sequence $\boldsymbol{X}$ along with the rich priors predicted by Net^D ($\boldsymbol{F}^{D,M}, \mathbf{f}^{D,M}$, and $\mathcal{K}^D$) as input. It first generates initial features by combining $\boldsymbol{X}$ and the context feature $\boldsymbol{F}^{D,M}$ using concatenation and an RDB [44]. These features are then refined through another stack of HRBA blocks, initializing the multi-flow-mask pairs with $\mathbf{f}^{D,M}$ from Net^D to leverage the motion prior. Crucially, within each HRBA block in Net^R, the DA attention utilizes the estimated kernel $\mathcal{K}_i^D$ as its query, after which the ETM layer continues to provide exposure conditioning, enabling degradation- and exposure-adaptive restoration. Finally, the refined features $\boldsymbol{F}^{R,M}$ pass through an upsampling block to predict a high-frequency residual $\hat{\boldsymbol{Y}}_i^{\text{res}}$. The final sharp HR frame is obtained by adding this residual to the bilinearly upsampled blurry LR input:

$$\hat{\boldsymbol{Y}}_i = \hat{\boldsymbol{Y}}_i^{\text{res}} + \boldsymbol{X}_i \uparrow_s, \tag{8}$$

where $\uparrow_s$ denotes the $\times s$ bilinear upsampling.

3.5 Training Strategy

We adopt a three-stage training strategy to effectively optimize FMA-Net++. First, the ETE is pretrained using a supervised contrastive loss [19] on exposure labels derived from our data synthesis process (Sec. 4.1; see *Suppl.* for detailed contrastive loss formulations). Crucially, we freeze the ETE afterward to provide a stable exposure reference space, preventing representation drift during subsequent training. Second, guided by the frozen ETE, Net^D is trained to predict reliable degradation priors, supervised by a composite loss $\mathcal{L}_D$:

$$\mathcal{L}_D = \sum_{i=1}^{T} l_1(\hat{\boldsymbol{X}}_i, \boldsymbol{X}_i) + \lambda_1 \sum_{i=1}^{T} l_1(\boldsymbol{Y}_{i\pm1\to i}, \boldsymbol{Y}_i) + \lambda_2 l_1(\boldsymbol{f}^{\boldsymbol{Y}}, \boldsymbol{f}^{\boldsymbol{Y}}_{\text{RAFT}}), \tag{9}$$

where the first term is the reconstruction loss of the blurry LR input via Eq. 5, the second term is a warping loss, and the third term explicitly supervises the optical flow guided by a pretrained RAFT [39]. RAFT is used only for training-time pseudo-supervision and is not required during inference. Finally, the entire

framework is jointly trained end-to-end, fine-tuning Net^D alongside Net^R with the total loss:

$$\mathcal{L}_{\text{total}} = l_1(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) + \lambda_3 \mathcal{L}_D, \tag{10}$$

where the first term is the primary restoration loss on the final sharp HR output.

4 Experiments

4.1 Experimental Setup

Datasets and Benchmarks. To systematically evaluate VSRDB under controlled exposure-duration variation, we construct two new benchmarks derived from the REDS dataset [33]: **REDS-ME** (Multi-Exposure) and **REDS-RE** (Random-Exposure).

For REDS-ME, we synthesize five distinct exposure levels corresponding to duty cycles from 5 : 1 (shortest exposure) to 5 : 5 (longest exposure), as illustrated in Fig. 5. Crucially, these discrete levels serve as pseudo-labels for ETE pretraining. We train FMA-Net++ using all five levels from the REDS-ME training set and evaluate on the most challenging levels (5 : 4, 5 : 5) of the REDS4-ME test set (derived from the standard REDS4 partition). To explicitly assess robustness to dynamic exposure variations, we construct REDS-RE by temporally mixing frames across the five exposure levels within the REDS4-ME test scenes, yielding controlled exposure trajectories with varying blur extents. Furthermore, we use the standard GoPro dataset [34] to evaluate generalization to different scene and motion statistics, alongside qualitative assessments on real-world blurry videos. Details of the data synthesis pipeline are provided in the *Suppl.*

Evaluation Metrics. We evaluate restoration quality using PSNR and SSIM [45]. Temporal consistency is measured by tOF [2]. For real-world videos where GT is unavailable, we report no-reference metrics such as NIQE [32] and MUSIQ [18]. We also compare model efficiency in terms of parameter count and runtime.

Implementation Details. All implementation details, including network configurations and hyperparameter settings, are provided in the *Suppl.*

4.2 Comparisons with State-of-the-Art Methods

We compare FMA-Net++ against SOTA methods across relevant categories: single-image SR (SwinIR [26], HAT [8]), single-image deblurring (Restormer [50], FFTformer [21]), VSR (BasicVSR++ [7], IART [48]), video deblurring (VRT [25], RVRT [27], BSSTNet [52]), Blind VSR (DBVSR [36]), and joint VSRDB (FMA-Net [49], Ev-DeblurVSR [17]). For a fair comparison in the joint VSRDB setting under varying exposure conditions, relevant SOTA methods were adapted and retrained on our REDS-ME training set, denoted by * in Tables 1 and 2.

Quantitative Results. Table 1 presents the performance on REDS4-ME across two challenging exposure levels (5 : 4 and 5 : 5), representing severe motion blur. FMA-Net++ consistently outperforms all baselines across PSNR, SSIM, and tOF. For instance, FMA-Net++ achieves significant gains of 0.62 dB and 0.73

Table 1: Quantitative comparison of ×4 VSRDB on REDS4-ME for two challenging exposure levels (5:4 and 5:5). All metrics are computed on the RGB channels. **Red** and blue indicate the best and second-best performance, respectively. Runtime is measured per LR frame of resolution 180 × 320. The superscript * denotes models retrained on our proposed REDS-ME training set.

Methods	# Params (M)	Runtime (s)	REDS4-ME-5:4			REDS4-ME-5:5		
			PSNR↑	SSIM↑	tOF↓	PSNR↑	SSIM↑	tOF↓
Super-Resolution + Deblurring								
SwinIR [26] + Restormer [50]	11.9 + 26.1	0.221 + 0.753	26.23	0.7464	3.775	25.53	0.7229	4.558
HAT [8] + FFTformer [21]	20.8 + 16.6	0.352 + 1.414	26.66	0.7634	3.207	25.92	0.7400	3.995
BasicVSR++ [7] + RVRT [27]	7.3 + 13.6	0.048 + 0.349	27.28	0.7901	2.887	26.98	0.7621	3.164
IART [48] + BSSTNet [52]	13.4 + 52.0	1.041 + 0.482	27.50	0.8006	2.578	27.26	0.7888	2.721
Deblurring + Super-Resolution								
Restormer [50] + SwinIR [26]	26.1 + 11.9	0.043 + 0.221	26.36	0.7499	3.464	25.84	0.7316	3.948
FFTformer [21] + HAT [8]	16.6 + 20.8	0.066 + 0.352	26.36	0.7534	3.256	25.87	0.7356	3.739
RVRT [27] + BasicVSR++ [7]	13.6 + 7.3	0.019 + 0.048	26.35	0.7492	3.314	25.95	0.7424	3.610
BSSTNet [52] + IART [48]	52.0 + 13.4	0.025 + 1.041	26.51	0.7711	3.103	26.33	0.7564	3.313
Blind Video Super-Resolution								
DBVSR [36]	14.1	0.096	24.50	0.7208	3.449	22.19	0.6122	4.554
Joint Video Super-Resolution and Deblurring								
Restormer* [50]	26.5	0.045	27.45	0.7851	2.161	27.12	0.7750	2.516
DBVSR* [36]	14.1	0.096	26.77	0.7629	3.021	26.07	0.7405	3.765
BasicVSR++* [7]	7.3	0.048	27.70	0.7922	2.302	27.14	0.7770	2.746
IART* [48]	13.4	1.041	28.23	0.8153	2.143	27.64	0.7972	2.590
VRT* [25]	35.6	0.684	27.93	0.8045	2.366	27.41	0.7887	2.839
RVRT* [27]	12.9	0.385	28.11	0.8093	2.136	27.58	0.7944	2.558
BSSTNet* [52]	52.0	0.548	28.75	0.8342	1.893	28.11	0.8119	2.298
Ev-DeblurVSR [17]	8.3	0.062	24.51	0.7154	3.602	24.38	0.7047	4.094
Ev-DeblurVSR* [17]	8.3	0.062	27.40	0.7839	2.521	26.82	0.7672	3.059
FMA-Net [49]	9.6	0.318	26.42	0.7958	2.503	26.67	0.8005	2.443
FMA-Net* [49]	9.6	0.318	29.04	0.8275	1.891	28.51	0.8136	2.269
FMA-Net++ (Ours)	12.8	0.074	**29.66**	**0.8546**	**1.688**	**29.24**	**0.8453**	**1.956**

dB over the second-best model, FMA-Net*, on levels 5 : 4 and 5 : 5, respectively. Furthermore, FMA-Net++ demonstrates superior efficiency compared to methods with similar complexity like RVRT* [27]. It achieves remarkably higher performance while being significantly faster (over 5.2× speedup), an efficiency that primarily arises from our parallelizable HRBA architecture. While VRT [25] also enables sequence-level parallelization, it suffers from massive memory consumption and a heavy computational burden, requiring 20.5 GB for a 10-frame sequence at 180 × 320 resolution. In contrast, our FMA-Net++ requires only 6.2 GB, demonstrating that HRBA achieves parallel temporal modeling far more efficiently while maintaining state-of-the-art accuracy.

Table 2 evaluates robustness to dynamic exposure (REDS-RE) and generalization ability to an unseen dataset (GoPro [34]). On REDS-RE, featuring dynamic exposure transitions within sequences, the performance advantage of FMA-Net++ over other methods widens considerably compared to REDS-ME. This result validates the effectiveness of our explicit exposure-aware modeling (ETM) in adapting to temporally varying exposure conditions where fixed-exposure assumptions struggle. On the unseen GoPro dataset, which has different scene and motion statistics from REDS-ME, FMA-Net++ again achieves the best performance across all metrics, indicating strong generalization beyond the training domain.

Qualitative Results. Fig. 6 presents visual comparisons on synthetic benchmarks (REDS4-ME-5 : 5 and GoPro) that contain severe motion blur, while

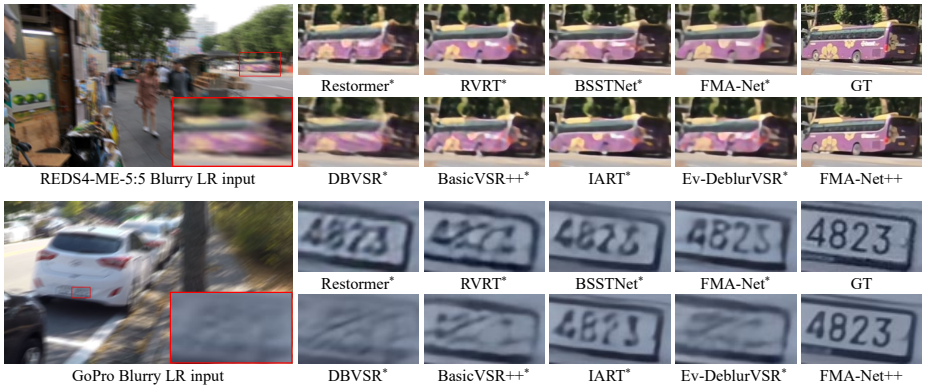


Fig. 6: Qualitative comparisons of ×4 VSRDB on REDS4-ME-5 : 5 and GoPro [34]. Each scene contains severe motion blur with different characteristics.

Fig. 1(a) shows results on challenging real-world videos captured with a smartphone. On both synthetic and real-world data, FMA-Net++ consistently restores sharper details, cleaner edges, and more legible text with fewer artifacts, achieving the best perceptual quality (NIQE/MUSIQ). We omit multi-modal methods such as Ev-DeblurVSR [17] from the real-world comparison, as they are fundamentally not applicable to standard RGB videos that lack the required event data. This highlights the strong practicality and generalization of our approach, which achieves these results using only conventional RGB inputs despite being trained solely on synthetic data. Further qualitative results are provided in the *Suppl.*

5 Ablation Study

We present ablation studies validating our key design choices. Further ablation studies and detailed analyses can be found in the *Suppl.*

5.1 Effectiveness of Hierarchical Architecture

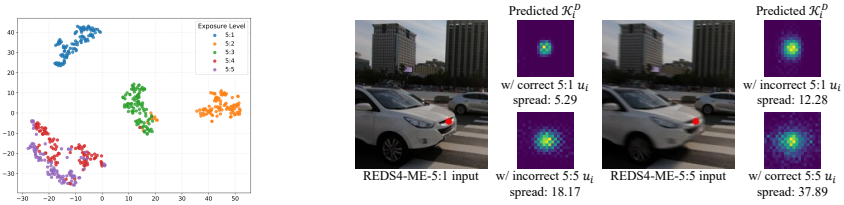
To validate the advantages of our proposed hierarchical temporal architecture (conceptually compared with other strategies in Fig. 2(a)), we compare the full FMA-Net++ against two variants built upon its core components but employing

Table 2: Quantitative comparison of ×4 VSRDB on REDS-RE and GoPro [34].

Methods	REDS-RE			GoPro		
	PSNR ↑	SSIM ↑	tOF ↓	PSNR ↑	SSIM ↑	tOF ↓
Restormer* [50]	27.79	0.7953	1.775	27.54	0.8350	3.302
DBVSR* [36]	27.30	0.7742	2.398	26.05	0.7815	4.730
BasicVSR++* [7]	28.14	0.8044	1.904	27.40	0.8282	3.285
IART* [48]	28.68	0.8248	1.852	27.76	0.8394	3.302
VRT* [25]	28.24	0.8124	2.071	27.39	0.8304	3.616
RVRT* [27]	28.56	0.8208	1.926	27.64	0.8364	3.223
BSSTNet* [52]	29.33	0.8427	1.602	28.57	0.8650	2.753
Ev-DeblurVSR* [17]	27.94	0.7987	2.039	27.25	0.8247	3.536
FMA-Net* [49]	29.29	0.8413	1.614	28.83	0.8655	2.727
FMA-Net++ (Ours)	**30.13**	**0.8643**	**1.360**	**30.49**	**0.9018**	**2.091**

different temporal modeling strategies: (i) a sliding-window variant processing three frames at a time, similar to [49], and (ii) a recurrent variant where the hierarchical refinement is adapted for sequential propagation. All variants maintain the same number of HRBA blocks and utilize the same ETM and multi-attention mechanisms, forming an identical backbone for a fair comparison.

Table 3 presents the comparison results on REDS4-ME-5:5. Our hierarchical FMA-Net++ demonstrates substantial improvements over both variants. Compared to the sliding-window variant, it yields markedly better results across all metrics, effectively overcoming the limitations imposed by a fixed temporal receptive field. Compared to the recurrent variant, it achieves superior performance in both PSNR and tOF. This noticeable gain in temporal consistency likely stems from its non-recurrent hierarchical structure, which mitigates gradient-vanishing issues that can affect sequential propagation over long sequences. We also empirically observe that this design achieves the most stable training dynamics among the compared variants. Furthermore, in terms of efficiency, the hierarchical design demonstrates a modest speed advantage over the recurrent approach. Overall, these results validate that our hierarchical strategy serves as a highly effective backbone for high-quality and temporally consistent video restoration.



(a) t-SNE visualization of ETE features (b) Effect of guidance on kernels

Fig. 7: Visual analysis of exposure-aware modeling. (a) t-SNE [31] visualization of ETE features ($\boldsymbol{u}$) shows clear clustering across exposure levels. (b) Bidirectional guidance test: for both a 5:5 and a 5:1 input, the predicted degradation kernel adapts its spatial spread to the supplied exposure guidance.

5.2 Effectiveness of Exposure-Aware Modeling

Table 4 evaluates our exposure-aware modeling by comparing FMA-Net++ with variants without ETE, including a capacity-matched baseline (13.1M), slightly larger than the full model (12.8M). Increasing the capacity of the w/o-ETE model absorbs most of the in-distribution PSNR gain on REDS4-ME, indicating that part of the improvement under fixed exposure comes from backbone capacity. However, the matched baseline does not close the gap on the exposure-varying REDS-RE split, where FMA-Net++ w/ ETE still improves PSNR from 29.88 to 30.13 and reduces tOF from 1.399 to 1.360. This indicates that the ETE/ETM pipeline contributes to exposure-conditioned robustness beyond pa-

Table 3: Comparison of temporal modeling variants within FMA-Net++ on REDS4-ME-5:5. Our HRBA design achieves the best accuracy and temporal consistency.

Temporal Modeling Strategy	Runtime (s)	PSNR ↑	tOF ↓
Sliding-window variant	0.314	28.57	2.231
Recurrent propagation variant	0.086	29.11	1.989
HRBA (Ours)	**0.074**	**29.24**	**1.956**

Table 4: Ablation study on the Exposure Time-aware Feature Extractor (ETE) for multiple datasets.

Methods	# Params (M)	Runtime (s)	In-distribution				Out-of-distribution			
			REDS4-ME-5:4		REDS4-ME-5:5		REDS-RE		GoPro	
			PSNR↑	tOF↓	PSNR↑	tOF↓	PSNR↑	tOF↓	PSNR↑	tOF↓
FMA-Net++ w/o ETE	9.8	0.071	29.55	1.764	29.12	2.054	29.72	1.436	29.78	2.267
FMA-Net++ w/o ETE (matched)	13.1	0.088	**29.67**	1.713	29.20	2.011	29.88	1.399	29.85	2.251
FMA-Net++ w/ ETE	12.8	0.074	29.66	**1.688**	**29.24**	**1.956**	**30.13**	**1.360**	**30.49**	**2.091**

rameter count. The consistent gain on the unseen GoPro dataset (+0.64 dB) further reflects improved generalization beyond the training domain.

Choice of ETE Design. To justify the contrastive ETE design, we compare it with three deployable exposure-conditioning alternatives that compute guidance from the input: frame-difference features, a classification-pretrained ETE, and an ordinal-regression ETE. As shown in Table 5, all alternatives underperform our contrastive ETE on REDS-RE, indicating that the gain does not come from merely adding an arbitrary exposure-related signal. The contrastive objective provides a more effective exposure embedding for the ETM/FGDF pipeline.

Visual Analysis of Learned Priors. To further validate this explicit conditioning, we visualize the learned representations in Fig. 7. The t-SNE [31] visualization of ETE features ($\boldsymbol{u}$) in Fig. 7(a) shows clear clustering across exposure levels, confirming that ETE successfully extracts exposure-dependent characteristics. Furthermore, Fig. 7(b) demonstrates the impact of this guidance through a bidirectional test. For a severely blurred 5:5 input, correct 5:5 guidance yields a spatially diffuse kernel, whereas incorrect 5:1 guidance contracts it into a concentrated one. Conversely, for a mildly blurred 5:1 input, 5:5 guidance drives the kernel to become more diffuse. These bidirectional responses show that the predicted kernel $\mathcal{K}_i^D$ adapts its spatial shape to the supplied exposure guidance.

Frame-wise Guidance Analysis. We further analyze the role of ETE guidance on REDS-RE by keeping the input sequence fixed and corrupting only the frame-wise guidance $\boldsymbol{u}$, as shown in Table 6. Correct guidance performs best, while fixed, sequence-shuffled, and random wrong-level guidance degrade both PSNR and tOF, with larger drops on transition frames where the exposure level changes. Notably, random wrong-level guidance performs worse than removing guidance entirely, showing that $\boldsymbol{u}$ is actively used rather than ignored.

Fixed-Input Sensitivity. Complementary to the REDS-RE corruption test, Table 7 reports the effect of replacing $\boldsymbol{u}$ with exposure embeddings from different levels on fixed REDS4-ME-5:5 inputs. Performance decreases gradually as the guidance deviates from the correct level, yet remains stable even under the farthest 5:1 guidance. This shows that, when the input exposure is fixed and

Table 5: Comparison of exposure-conditioning designs on REDS-RE.

Conditioning design	PSNR↑ / tOF↓
Frame-difference	29.75 / 1.429
Classification-pretrained ETE	29.92 / 1.398
Ordinal-regression ETE	29.97 / 1.382
Contrastive ETE (Ours)	**30.13** / **1.360**

Table 6: Effect of corrupting frame-wise guidance $\boldsymbol{u}$ on REDS-RE.

Guidance $\boldsymbol{u}$	Average	Transition Frames
	PSNR↑ / tOF↓	PSNR↑ / tOF↓
Correct	**30.13** / **1.360**	**29.98** / **1.395**
Same-frame fixed 5:5	29.88 / 1.387	29.40 / 1.631
Sequence-shuffled	29.59 / 1.525	29.18 / 1.713
Random wrong-level	29.47 / 1.605	29.05 / 1.739
w/o ETE (no guidance)	29.72 / 1.436	29.48 / 1.633

Table 7: Sensitivity to ETE guidance on fixed REDS4-ME-5:5 inputs.

Input Frame	Guidance $\boldsymbol{u}$	PSNR ↑ / tOF ↓
5:5	from 5:5 (Correct)	**29.24** / **1.956**
	from 5:4	29.20 / 1.972
	from 5:3	29.13 / 2.012
	from 5:2	29.11 / 2.027
	from 5:1	29.07 / 2.041
Baseline w/o ETE		29.12 / 2.054

Table 8: Ablation on key components of FMA-Net++.

Methods	# Params	Time (s)	PSNR↑	SSIM↑	tOF↓
The number of HRBA blocks M					
(a) $M = 1$	7.7M	0.035	28.29	0.8174	2.461
(b) $M = 2$	9.4M	0.048	28.74	0.8339	2.151
Multi-Attention					
(c) self-attn + SFT	13.2M	0.066	28.86	0.8378	2.132
Proposed					
(d) FMA-Net++	12.8M	0.074	**29.24**	**0.8453**	**1.956**

strong spatio-temporal evidence is available, FMA-Net++ leverages its HRBA backbone rather than relying exclusively on ETE.

5.3 Effectiveness of HRBA and Core Components

We conduct ablation studies to validate the key components and design choices of FMA-Net++, summarizing the main results in Table 8.

First, we investigate the impact of our hierarchical refinement strategy by varying the number of stacked HRBA blocks (M). As shown in Table 8(a, b, d), increasing M from 1 to 2 and finally to our full configuration ($M = 4$, row d) progressively improves both PSNR and tOF. This demonstrates the effectiveness of hierarchically expanding the temporal receptive field. As visualized in the *Suppl.*, features become progressively sharper and more structurally aligned through the stacked blocks, further validating our hierarchical design.

Next, we validate the effectiveness of the Degradation-Aware (DA) attention within Net^R's multi-attention module. Replacing DA attention with a standard SFT layer [43] for modulation significantly degrades performance, confirming that explicitly leveraging the estimated degradation priors via DA attention is crucial for targeted restoration.

6 Conclusion

In this paper, we addressed the challenging problem of joint VSRDB under unknown and dynamically varying exposure conditions. To tackle this challenge, we introduced FMA-Net++, a novel framework built upon HRBA blocks that enables effective sequence-level temporal modeling with efficient parallel processing. Crucially, our proposed ETM layer injects per-frame exposure conditioning into the features. This allows our exposure-aware FGDF module to predict degradation kernels that capture the coupled effects of motion and exposure. Extensive experiments on the proposed REDS-ME and REDS-RE benchmarks, as well as GoPro and real-world videos, demonstrate that FMA-Net++ achieves state-of-the-art results, showcasing superior performance, efficiency, and robustness while generalizing effectively despite being trained on synthetic data.

Acknowledgements. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government [Ministry of Science and ICT (Information and Communications Technology)] (Project Number: RS-2022-00144444, Project Title: Deep Learning Based Visual Representational Learning and Rendering of Static and Dynamic Scenes).

References

1. Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
2. Learning temporal coherence via self-supervision for gan-based video generation. ACM Transactions on Graphics **39**(4), 75–1 (2020)
3. Bai, H., Pan, J.: Self-supervised deep blind video super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(7), 4641–4653 (2024)
4. Cao, J., Li, Y., Zhang, K., Van Gool, L.: Video super-resolution transformer. arXiv preprint arXiv:2106.06847 (2021)
5. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: 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. pp. 4947–4956 (2021)
6. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Understanding deformable alignment in video super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 973–981 (2021)
7. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5972–5981 (2022)
8. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22367–22377 (2023)
9. Chiche, B.N., Woiselle, A., Frontera-Pons, J., Starck, J.L.: Stable long-term recurrent video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 837–846 (2022)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Fang, N., Zhan, Z.: High-resolution optical flow and frame-recurrent network for video super-resolution and deblurring. Neurocomputing **489**, 128–138 (2022)
12. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3897–3906 (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. Hu, P., Niklaus, S., Sclaroff, S., Saenko, K.: Many-to-many splatting for efficient video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3553–3562 (2022)
15. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in Neural Information Processing Systems **28** (2015)
16. Jo, Y., Oh, S.W., Kang, J., Kim, S.J.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3224–3232 (2018)
17. Kai, D., Zhang, Y., Wang, J., Xiao, Z., Xiong, Z., Sun, X.: Event-enhanced blurry video super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4175–4183 (2025)

18. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
19. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020)
20. Kim, T., Lee, J., Wang, L., Yoon, K.J.: Event-guided deblurring of unknown exposure time videos. In: European Conference on Computer Vision. pp. 519–538. Springer (2022)
21. Kong, L., Dong, J., Ge, J., Li, M., Pan, J.: Efficient frequency domain-based transformers for high-quality image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5886–5895 (2023)
22. Lee, S., Choi, M., Lee, K.M.: Dynavsr: Dynamic adaptive blind video super-resolution. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2093–2102 (2021)
23. Li, D., Shi, X., Zhang, Y., Cheung, K.C., See, S., Wang, X., Qin, H., Li, H.: A simple baseline for video restoration with grouped spatial-temporal shift. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9822–9832 (2023)
24. Li, W., Tao, X., Guo, T., Qi, L., Lu, J., Jia, J.: Mucan: Multi-correspondence aggregation network for video super-resolution. In: European Conference on Computer Vision. pp. 335–351. Springer (2020)
25. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. IEEE Transactions on Image Processing **33**, 2171–2182 (2024)
26. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1833–1844 (2021)
27. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Gool, L.V.: Recurrent video restoration transformer with guided deformable attention. Advances in Neural Information Processing Systems **35**, 378–393 (2022)
28. Lin, J., Huang, Y., Wang, L.: Fdan: Flow-guided deformable alignment network for video super-resolution. arXiv preprint arXiv:2105.05640 (2021)
29. Liu, C., Yang, H., Fu, J., Qian, X.: Learning trajectory-aware transformer for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5687–5696 (2022)
30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
31. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11) (2008)
32. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2012)
33. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: 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. pp. 1996–2005 (2019)

34. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3883–3891 (2017)
35. Oh, J., Kim, M.: Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In: European Conference on Computer Vision. pp. 198–215. Springer (2022)
36. Pan, J., Bai, H., Dong, J., Zhang, J., Tang, J.: Deep blind video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4811–4820 (2021)
37. Shang, W., Ren, D., Yang, Y., Zhang, H., Ma, K., Zuo, W.: Joint video multi-frame interpolation and deblurring under unknown exposure time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13935–13944 (2023)
38. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4472–4480 (2017)
39. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European Conference on Computer Vision. pp. 402–419. Springer (2020)
40. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3360–3369 (2020)
41. Wang, L., Guo, Y., Liu, L., Lin, Z., Deng, X., An, W.: Deep video super-resolution using hr optical flow estimation. IEEE Transactions on Image Processing **29**, 4323–4336 (2020)
42. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
43. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 606–615 (2018)
44. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: European Conference on Computer Vision Workshops (September 2018)
45. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
46. Weng, W., Zhang, Y., Xiong, Z.: Event-based blurry frame interpolation under blind exposure. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1588–1598 (2023)
47. Wolberg, G.: Digital image warping, vol. 10662. IEEE computer society press Los Alamitos, CA (1990)
48. Xu, K., Yu, Z., Wang, X., Mi, M.B., Yao, A.: Enhancing video super-resolution via implicit resampling-based alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2546–2555 (2024)
49. Youk, G., Oh, J., Kim, M.: Fma-net: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 44–55 (June 2024)

50. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
51. Zhang, H., Xie, H., Yao, H.: Spatio-temporal deformable attention network for video deblurring. In: European Conference on Computer Vision. pp. 581–596. Springer (2022)
52. Zhang, H., Xie, H., Yao, H.: Blur-aware spatio-temporal sparse transformer for video deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2673–2681 (2024)
53. Zhang, K., Luo, W., Zhong, Y., Ma, L., Liu, W., Li, H.: Adversarial spatio-temporal learning for video deblurring. IEEE Transactions on Image Processing **28**(1), 291–301 (2018)
54. Zhang, Y., Wang, C., Tao, D.: Video frame interpolation without temporal priors. Advances in Neural Information Processing Systems **33**, 13308–13318 (2020)
55. Zhong, Z., Gao, Y., Zheng, Y., Zheng, B.: Efficient spatio-temporal recurrent neural network for video deblurring. In: European Conference on Computer Vision. pp. 191–207. Springer (2020)
56. Zhu, C., Dong, H., Pan, J., Liang, B., Huang, Y., Fu, L., Wang, F.: Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3598–3607 (2022)