Title

A Dual Camera System for High Spatiotemporal Resolution Video Acquisition

Journal

IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10)

ISSN

0162-8828

Authors

Cheng, Ming

Ma, Zhan

Asif, M Salman

et al.

Publication Date

2021-10-01

DOI

10.1109/tpami.2020.2983371

Peer reviewed

A Dual Camera System for High Spatiotemporal Resolution Video Acquisition

Ming Cheng, Zhan Ma, M. Salman Asif, Yiling Xu, Haojie Liu, Wenbo Bao, and Jun Sun

Abstract—This paper presents a dual camera system for high spatiotemporal resolution (HSTR) video acquisition, where one camera shoots a video with high spatial resolution and low frame rate (HSR-LFR) and another one captures a low spatial resolution and high frame rate (LSR-HFR) video. Our main goal is to combine videos from LSR-HFR and HSR-LFR cameras to create an HSTR video. We propose an end-to-end learning framework, AWnet, mainly consisting of a FlowNet and a FusionNet that learn an adaptive weighting function in pixel domain to combine inputs in a frame recurrent fashion. To improve the reconstruction quality for cameras used in reality, we also introduce noise regularization under the same framework. Our method has demonstrated noticeable performance gains in terms of both objective PSNR measurement in simulation with different publicly available video and light-field datasets and subjective evaluation with real data captured by dual iPhone 7 and Grasshopper3 cameras. Ablation studies are further conducted to investigate and explore various aspects (such as reference structure, camera parallax, exposure time, etc) of our system to fully understand its capability for potential applications.

Index Terms—Dual camera system, high spatiotemporal resolution, super-resolution, optical flow, spatial information, end-to-end learning

1 Introduction

High-speed cameras play an important role in various modern imaging and photography tasks including sports photography, film special effects, scientific research, and industrial monitoring. They allow us to see very fast phenomena that are easily overlooked and can not be captured at ordinary speed, such as a droplet, full-speed fan rotation, or even a gun fire. These cameras can capture videos at high frame-rates that range from several hundred to several thousand frames per second (FPS), while an ordinary camera operates at 30 to 60 FPS. The high frame-rates often come at the expense of spatial resolution; especially, consumer-level cameras that sacrifice the spatial resolution to maintain the high frame rate acquisition. For example, popular iPhone 7 can capture 4K videos at 30 FPS, but can only offer 720p resolution at 240 FPS because of the limitation of the data I/O throughput. Some special-purpose and professional high-speed cameras can capture high spatiotemporal resolution (HSTR) videos, but they are typically very expensive (e.g., Phantom Flex4K[1] with price starting at $110K) and beyond the budget of a majority of consumers.

A naïve solution to obtain a HSTR video from a video with high spatial resolution and low frame rate (HSR-LFR) or low spatial resolution and high frame rate (LSR-HFR) is to upsample along temporal or spatial direction, respectively. Upsampling in temporal resolution or frame rate upconversion of an HSR-LFR

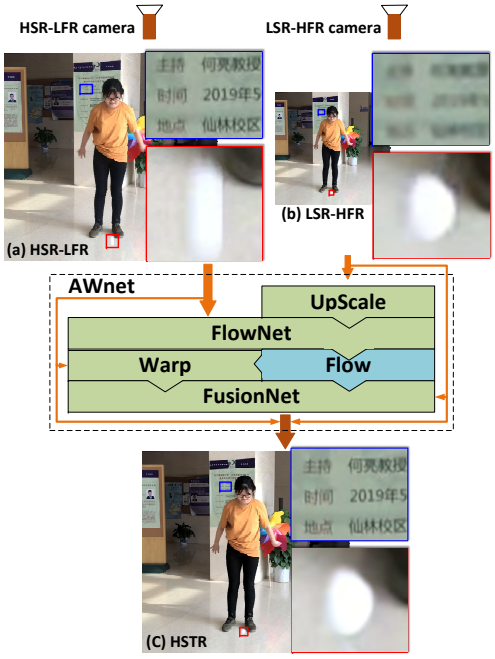


Fig. 1. Snapshots of high spatial resolution-low frame rate (HSR-LFR) and low spatial resolution-high frame rate (LSR-HFR) videos and synthesized high spatiotemporal resolution (HSTR) video. A woman is throwing a ping-pong ball in indoor space. (a) HSR-LFR video 4K@30FPS frame with zoomed-in region showing motion blur; (b) LSR-HFR video 720p@240FPS frame with zoomed-region showing spatial blur; (c) HSTR video 4K@240FPS frame.

M. Cheng, Z. Ma and H. Liu are with Nanjing University, Nanjing, Jiangsu, China. M. S. Asif is with the University of California at Riverside. M. Cheng is also with Shanghai JiaoTong University, Shanghai, China. Y. Xu, W. Bao and J. Sun are with Shanghai JiaoTong University, Shanghai, China. Z. Ma is the corresponding author of this paper. M. S. Asif and Y. Xu are co-corresponding authors of this paper.

This paper is supported in part by National Natural Science Foundation of China (61971282), National Key Research and Development Project of China Science and Technology Exchange Center (2018YFE0206700) and Scientific Research Plan of the Science and Technology Commission of Shanghai Municipality (18511105402).

1. https://www.phantomhighspeed.com/products/cameras/4kmedia/flex4k

video (e.g., 4K at 30FPS) involves imputing missing frames by interpolating motion between the observed frames, which is challenging because of the motion blur introduced by long exposure and inaccuracies in motion representation under the commonly-used uniform translational motion assumption [47]. On the other hand, upsampling spatial resolution of a LSR-HFR video can be performed using a variety of existing super-resolution methodologies [33], [51], but they often provide smoothed images in which high frequency spatial details of the captured scene are missing. Figure 1 highlights these effects, where HSR-LFR (4K at 30FPS) video contains motion blur in the regions of fast motion and LSR-HFR (720p at 240FPS) has a uniform spatial blur because of limited spatial resolution.

In this paper, we propose a dual camera system for HSTR video $\mathbf{Y}_{\texttt{HSTR}}(t)$ acquisition, as shown in Fig. 1, where one camera captures a HSR-LFR video $\mathbf{X}_{\texttt{HSR}-\texttt{LFR}}(t)$ with rich spatial information (i.e., sharp spatial details for textures and edges), and the other one records a HFR-LSR video $\mathbf{x}_{\texttt{LSR}-\texttt{HFR}}(t)$ with fine-grain temporal information (i.e., intricate motion flows). We then fuse these two videos via a learning-based approach to produce a final HSTR video with both appealing spatial details and accurate motion. In another words, we aim to transfer the rich spatial details from HSR-LFR frame to the associated LSR-HFR frames while retaining accurate motion in the entire sequence.

Our method performs spatiotemporal super-resolution in a frame recurrent manner within a synchronized GoP (group of pictures) to exploit the spatial-temporal priors in Fig. 2(a). A detailed block diagram of our proposed method is shown in Fig. 2(b). The synthesis process has two main parts: FlowNet and FusionNet, which are placed consecutively in Fig. 2(c). $I_{\texttt{LSR}}$ denotes a frame captured with LSR-HFR camera. The FlowNet accepts an upsampled LSR-HFR image ($I_{\texttt{LSR}\uparrow}$) and a reference image ($I_{\texttt{ref}}$) to provide the optical flow (denoted as $\mathscr{F}$) and a warped reference image ($I^{w}_{\texttt{ref}}$). The reference image can either be a frame from the HSR-LFR camera at a synchronized time instant or a synthesized frame Y. Our dual camera based system can be viewed as a method in the class of "super-resolution with reference" (RefSR) methods [7], [51]. The FusionNet accepts the optical flow, upsampled LSR-HFR image, and warped reference image, and learns dynamic filters and masking pattern that are used to adaptively weigh the contribution of $I_{\texttt{LSR}}$ and $I_{\texttt{ref}}$ for a high-quality reconstruction Y. We use PWC-Net [42] as our FlowNet and a U-net as our FusionNet [39]. More details about network architecture are provided in Section 3 and in Fig. 2(d). Our method learns adaptive weights to combine the hybrid inputs, therefore, we refer to it as an adaptive weighting network (AWnet).

Our proposed AWnet is trained using Vimeo90K dataset. We first evaluate the performance of our method using simulations on publicly available datasets. Then we measure the performance of our method on videos captured with our custom dual-camera prototype. In our experimental evaluations, we observe that the quality of HSTR video degrades when we directly use models trained with Vimeo90K training images. One reason for the performance degradation is the presence of large sensor noise in $I_{\texttt{LSR}}$ when LSR-HFR video is captured with short exposure time (especially under low light conditions). Vimeo90K training data is virtually free of noise and other nonidealities that a real data capture encounters. To make our system robust to noise, we introduce noise at various levels in original Vimeo90K training data when performing the end-to-end learning. Such noise regularization can intelligently shift weights between $I_{\texttt{LSR}}$ and $I_{\texttt{ref}}$, offering much better reconstruction quality, under the same framework.

Extensive simulations are conducted using both simulation data from publicly accessible videos, dual-camera captures, and light field datasets (such as Vimeo90K [47], KITTI [36], Flower [41], LFVideo [43] and Stanford Light Field [1] datasets[2]), and real data (captured with custom-built dual iPhone 7 or Grasshopper3 cameras). Our proposed AWnet demonstrates noticeable performance gains over the existing super-resolution and frame interpolation methods, in objective and subjective measures. In our tests with simulations using Vimeo90K testing samples, our proposed model offers ~ 0.7 dB PSNR gain compared to the state-of-the-art CrossNet [51] and ~ 6 dB PSNR compared to the popular single-image super-resolution (SISR) method EDSR [33]. Our proposed AWnet provides the best performance on other video and lightfield datasets. In our tests with real data, we observe perceptual enhancements for various scenarios with indoor and outdoor activities under different lighting conditions.

We also offer ablation studies to fully understand the capability of our dual camera AWnet system, by analyzing various aspects in practice, such as the impacts of upscaling filters, reference structure, camera parallaxes, exposure time, etc. All these tests demonstrate the efficiency of our dual camera system for super-resolution and frame interpolation, to maintain sharp spatial details and accurate temporal motions jointly, leading to the state-of-the-art performance.

Main contributions of this work are highlighted below.

- A practical system for high spatiotemporal video acquisition uses a dual off-the-shelf camera setup. Videos from two cameras, operating at different spatial and temporal resolution, are combined using an end-to-end learning-based adaptive weighting to preserve spatial and temporal information in both inputs for a high-quality reconstruction.
- Cascaded FlowNet and FusionNet are applied to learn embedded spatial and temporal features for adaptive weights derivation in a frame recurrent way. These weights can be regularized using added noise to efficiently handle noise and other nonidealities in real data captured with consumer cameras.
- Our dual camera AWnet system demonstrates the state-of-the-art performance for super-resolution and frame interpolation, using both simulation data from public and real data captured by cameras.
- We analyze the robustness and efficiency of our system through a series of ablation studies to explore the impacts of upscaling filters, reference structure, camera parallaxes, exposure time, etc, which promises generalization in a variety of practical scenarios.

The remainder of this paper is structured as follows. Section 2 provides a brief overview of related work in literature, including system prototypes and applications. Section 3 details our proposed system and associated learning algorithms, followed by training processing in Section 4. The experimental results on simulation data and real data captured by cameras are shown in Section 5. We further break down our system to analyze and study its various aspects, such as the camera parallax, scaling filters, etc, in Section 6. Finally, conclusion is drawn in Section 7. Table 1 contains a list of all the notions and acronyms used throughout this paper.

2. Note that these datasets are widely used in literature for performance benchmark [51].

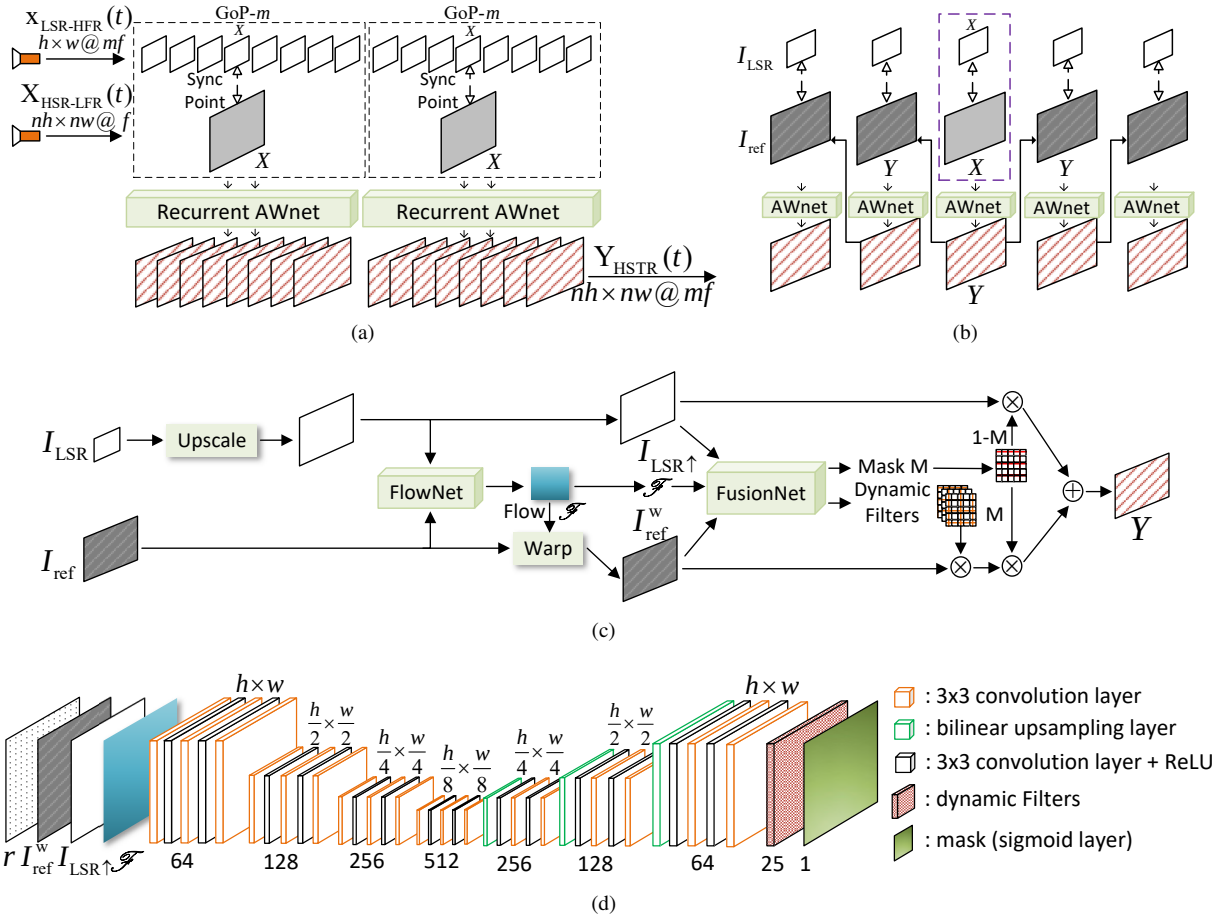


Fig. 2. **A Dual Camera System for High Spatiotemporal Acquisition:** (a) dual camera setup with one LSR-HFR video capture (e.g., $\mathbf{x}_{\text{LSR}-\text{HFR}}(t)$ with $h \times w$ at mf FPS), the other HSR-LFR video shooting (e.g., $\mathbf{X}_{\text{HSR}-\text{LFR}}(t)$ with $nh \times nw$ at f FPS) and synthesized HSTR video (e.g., $\mathbf{Y}_{\text{HSTR}}(t)$ with $nh \times nw$ at mf FPS); (b) Recurrent RefSR structure for Y synthesis using I_{LSR} and I_{ref} at each time instant; (c) Proposed AWnet for dual camera input with cascaded FlowNet and FusionNet to learn adaptive weights for final synthesis; (d) An U-net style [39] FusionNet structure for dynamic filter and mask generation. $\oplus$ and $\otimes$ are element-wise addition and multiplication.

TABLE 1
Notations and Abbreviations

Abbr.	Description
HSTR	High Spatiotemporal Resolution
HFR	High Frame Rate (or Temporal Resolution)
LFR	Low Frame Rate (or Temporal Resolution)
HSR	High Spatial Resolution (or Frame Size)
LSR	Low Spatial Resolution (or Frame Size)
$\mathbf{Y}_{\text{HSTR}}(t)$	Output HSTR Video
$Y_{t_i} = \mathbf{Y}_{\text{HSTR}}(t_i)$	A Frame of HSTR Video at time t_i
$\mathbf{x}_{\text{LSR}-\text{HFR}}(t)$	Input LSR-HFR Video
$x_{t_i} = \mathbf{x}_{\text{LSR}-\text{HFR}}(t_i)$	A Frame of LSR-HFR Video at time t_i
$\mathbf{X}_{\text{HSR}-\text{LFR}}(t)$	Input HSR-LFR Video
$X_{t_i} = \mathbf{X}_{\text{HSR}-\text{LFR}}(t_i)$	A Frame of HSR-LFR Video at time t_i
$\bar{\mathbf{X}}_{\text{LSR}-\text{HFR}}(t)$	Upscaled $\mathbf{x}_{\text{LSR}-\text{HFR}}(t)$
I_{ref}	A Frame from either $\mathbf{X}_{\text{HSR}-\text{LFR}}(t)$ or $\mathbf{Y}_{\text{HSTR}}(t)$
I_{LSR}	A Frame from $\mathbf{x}_{\text{LSR}-\text{HFR}}(t)$
$I_{\text{LSR}\uparrow}$	A Frame from $\bar{\mathbf{X}}_{\text{LSR}-\text{HFR}}(t)$
I^{w}_{ref}	warped I_{ref}
h, w	Height & width of x_{t_i}
f	frame rate of $\mathbf{X}_{\text{HSR}-\text{LFR}}(t)$
n, m	scaling factor of respective SR & FR
SISR	Single-Image Super Resolution
RefSR	Super Resolution with Reference
SNR	Signal-to-Noise Ratio
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity

2 Related Work

This work is closely related to the high-speed video acquisition, multi-camera (or array) system, super-resolution, and frame interpolation. we will briefly review these topics below.

High-speed Cameras. High-speed (or slow-motion) video capturing has been widely used these days. For example, iPhone Xs Max can capture a slow-motion video with 1080p at 240FPS, and Samsung Galaxy S10 offers 720p at 960FPS for a short time (0.2s). Consumer mobile devices often sacrifice spatial resolution and other image quality-related factors to ensure the high frame rate throughput. Professional high-speed cameras can support both high spatial and temporal resolution, typically coming along with a bulky body and an inconvenient price. Professional i-SPEED 726 can shoot 2K at 8512FPS and 1080p at 12742FPS, but its price range is above USD100,000. In recent years, we have noticed the increasing adoptions of multi-camera design from professional device to mobiles, such as dual camera iPhone or four camera Huawei P30. Thus, we have new opportunities to combine inputs from hybrid cameras operating at different speeds and synthesize high spatiotemporal resolution video.

Multi-Camera System. In pursuit of ultra-high spatial resolution, such as gigapixel, multi-camera systems and arrays have been developed [8], [9], [35]. In particular, cameras with different properties have been combined in a hybrid setup to sample and

synthesize images from a variety of light components, such as hyperspectral imaging [12], low light imaging [32], and light fields [45], [51]. Pelican imaging and Light are two recent companies that launched products with multiple cameras on board capturing a diverse set of images and synthesizing a desired image with high dynamic range, long-range depth, or light field [45] . The hybrid camera or multi-camera systems have also been used to combine different imaging sources to perform super-resolution and frame interpolation (for frame rate up-conversion) [51]. Our proposed work also belongs to the hybrid camera setup that captures two video streams, one at HSR-LFR and the other one at LSR-HFR, and combine them to synthesize a final HSTR video.

Super-Resolution. Single image super resolution (SISR) methods upscale individual images, which include traditional methods based on bilinear and bicubic filters and recently-introduced learning-based techniques [31], [33]. SISR methods can be easily extended to support video or multi-frame super-resolution [11], [27], [40]. In the case of multi-camera setup, super-resolution can be performed using some source as a reference [7], [17]. Low resolution images/videos can be upscaled with references (e.g., RefSR) from other viewpoints, leading to significant quality improvement [7], [49]–[51]. Such RefSR approaches have also been widely used in lightfield imaging [7], [51].

Frame Interpolation. Linear translation motion is a conventional assumption that has been extensively used in different interpolation-based methods to impute missing frames for frame rate up-conversion. Motion estimation can be performed using classical block-based or dense optical flow-based methods [2]–[5], [25], [37], [47]. Classical optimization-based and modern learning-based methods mainly try to retain smooth motion along the temporal trajectory while resolving occlusion-induced artifacts. Accurate motion flow estimation remains a challenging task because of the inconsistent object movements and motion-induced occlusion. As we discussed below, this issue can be significantly alleviated with the help of a high frame-rate video as a reference.

3 Dual Camera System for High Spatiotemporal Resolution Video

We propose a dual camera system for HSTR video acquisition, as illustrated in Fig. 2. One camera records a HSR-LFR video $\mathbf{X}_{\texttt{HSR}-\texttt{LFR}}(t)$ with $nh \times nw$ at f FPS, while the other one captures a LSR-HFR video $\mathbf{x}_{\texttt{LSR}-\texttt{HFR}}(t)$ with $h \times w$ at mf FPS. We learn an adaptive model to weigh contributions from the two input videos and synthesize a final HSTR output video $\mathbf{Y}_{\texttt{HSTR}}(t)$ with $nh \times nw$ at mf FPS. We use integer multipliers, m and n, for simplicity in this work, but different multipliers can be easily used. In subsequent sections, we first offer experimental observations that a single camera setup could not provide high-quality reconstruction of HSTR video via naïve spatial super-revolution or temporal frame interpolation. Then we discuss our dual camera system and algorithm development. Note that even though we particularly emphasize current work in a dual camera setup, this work can be generalized to other multi-camera configurations since the RefSR structure can be flexibly extended.

3.1 Single Camera System

Let us consider the following model for an image frame captured at time instance t of a camera as

$$I(t) = \int_{t-T}^{t} S(\tau) d\tau + n(t), \tag{1}$$

where $S(\tau)$ is the instantaneous photon density reflected from the physical scene, T denotes the exposure time, and $n(t)$ is the noise accumulated in the camera during a single exposure and the subsequent readout process. In other words, image is represented as the accumulated photons during the exposure time (according to the shutter speed). A typical consumer camera used in mobile devices[3] usually automatically adjusts the aperture size, ISO settings, and shutter speed according to a specific "shooting mode". Thus, exposure time duration T is uncertain and changes according to the scene content.

For example, normal video capture in iPhone 7 offers HSR (e.g., 4K) at f = 30FPS (i.e., HSR-LFR video), leading to rich spatial details but blurred motion. On the other hand, the slow-motion mode provides LSR (e.g., $<$ 720p) at mf = 240FPS (LSR-HFR video), resulting in accurate motion acquisition at the expense of spatial information, dynamic range, and SNR. Figure 1 shows two snapshots for respective HSR-LFR and LSR-HFR videos. As we can see, the spatial quality of the LSR-HFR frame is poor because some spatial information is missing in the slow-motion mode. Similarly, motion blur is clearly observed for the HSR-LFR frame because of insufficient temporal sampling.

Our experimental results in Tables 2, 3 and Figs. 6, 7, 8 show that frame-by-frame super-resolution or temporal interpolation of frames from a single camera could not provide the same quality as the dual camera setup, both objectively and subjectively.

These observations suggest that we are not capable of reconstructing high-quality HSTR video by applying spatial interpolation on the LSR-HFR video or temporal interpolation on HSR-LFR video. Therefore, we propose to use a dual camera system in which the main problem is to synthesize two input videos while preserving the sharp spatial details from the HSR-LFR video and real motions from the LSR-HFR video.

3.2 Dual Camera System

For a dual camera system setup, given a LSR-HFR video $\mathbf{x}_{\texttt{LSR}-\texttt{HFR}}(t)$ recorded at mf FPS and an additional HSR-LFR video $\mathbf{X}_{\texttt{HSR}-\texttt{LFR}}(t)$ recorded at f FPS from another view, we wish to generate a final HSTR video $\mathbf{Y}_{\texttt{HSTR}}(t)$ at mf FPS. To recover high-quality HSTR video, we seek to preserve the real motion field from the LSR-HFR camera input and the detailed spatial information from the HSR-LSR input. To do that, we need to design an effective mechanism to extract, transfer, and fuse appropriate information or features intelligently from the two inputs.

Dual cameras are configured and synchronized with a certain baseline distance to capture the instantaneous frames of the same scene. Because of the different frame rates for the respective HSR-LFR and LSR-HFR cameras, the exposure time of the HSR-LFR frame are often much longer than the LSR-HFR frames, especially in the low-light settings. Based on the imaging model

3. We use mobile phone camera as an example for its massive market adoption.

in (1), we can also consider the HSR-LFR frame as the high-spatial-resolution version of the summation of associated LSR-HFR frames, which will model the motion blur.

We divide the entire processing into two major steps: (1) optical flow estimation to compensate for motion and parallaxes to facilitate image fusion; (2) fusion processing to compute appropriate weighting functions (e.g., dynamic filter and mask in our work) through extensive feature learning. We propose to apply learned networks to perform aforementioned temporal and spatial feature extraction, transfer, and fusion. For convenience, we note them as "FlowNet" and "FusionNet" respectively, shown in Fig. 2(c).

We will synthesize the HSTR video in batches of group-of-pictures (GoP), as shown in Fig. 2(a). Because the same processing pattern is applied in each GoP, we will explain the specific steps for a single GoP. A GoP is a set of frames from the LSR-HFR video that are aligned to a specific time stamp of the HSR-LFR video. Since we assume that the LSR-HFR video is recorded at mf FPS and the HSR-LFR at f FPS, we find it convenient to use a GoP with m frames that we call GoP-m in the remainder of this paper. A GoP-m consists of m LSR-HFR frames that are aligned with one HSR-LFR frame. In particular, we assume that the HSR-LFR frame timestamp is in the middle of the timestamps for all the LSR-HFR frames in a GoP-m. For instance, a GoP-m that contains LSR-HFR frames at times $[t-k'\Delta t,\ldots,t+k\Delta t]$ are synchronized with HSR-LFR frame at t, where $k'=\lceil\frac{m}{2}\rceil-1$, $k=\lfloor\frac{m}{2}\rfloor$, and Δt is the time interval between two adjacent HSR-LFR frames. To maximize the use of spatial information from synchronized HSR-LFR frames, we use the synchronized HSR-LFR frame at t to first super-resolve its synchronized LSR-HFR frame at time t and then super-resolve its adjacent $m-1$ frames in a frame-recurrent manner.

3.2.1 Updating Synchronization Frame

We start with the synchronization frame for every HSR-LFR video frame. Let us denote an HSR-LFR frame at time t_i as $X_{t_i}=\mathbf{X}_{\texttt{HSR}-\texttt{LFR}}(t_i)$. We upscale the LSR-HFR video to the same spatial size as HSR-LFR, i.e.,

$$\bar{\mathbf{X}}(t)=\mathcal{U}(\mathbf{x}_{\texttt{LSR}-\texttt{HFR}}(t)), \tag{2}$$

where $\mathcal{U}(\cdot)$ denotes a spatial upscaling operator that either performs bilinear/bicubic interpolation or some other type of SISR methods (e.g., EDSR [33]). Let us denote the upscaled LSR-HFR frames as $\bar{X}_{t_i}=\bar{\mathbf{X}}(t_i)$. At synchronization time t_i, we use the $\bar{X}_{t_i}$ and X_{t_i} to produce an HSTR frame $Y_{t_i}=\mathbf{Y}_{\texttt{HSTR}}(t_i)$, as shown in Fig. 2(b).

The remaining HSTR frames in the GoP-m, are then generated in a frame-recurrent manner to fully exploit and leverage temporal priors of reconstructions. In other words, we first create super-resolved version of the synchronization frame and then reconstruct one HSTR frame at a time using its immediate super-resolved neighbor. The GoP-m centered at timestamp t_i can be written as $[\bar{X}_{t_i-k'\Delta t},\ldots,\bar{X}_{t_i+k\Delta t}]$, where $k'=\lceil\frac{m}{2}\rceil-1$, $k=\lfloor\frac{m}{2}\rfloor$, and $\Delta t=(t_{i+1}-t_i)/m$. Thus, we start at the center and estimate $Y_{t_i+\Delta t}$ and $Y_{t_i-\Delta t}$ using Y_{t_i}; then we estimate $Y_{t_i+k\Delta t}$ using $Y_{t_i+(k-1)\Delta t}$ and $Y_{t_i-k'\Delta t}$ using $Y_{t_i-(k'-1)\Delta t}$ for all the frames in the GoP-m.

The entire process described above is analogous to a RefSR method, where the reference is a high spatial resolution frame either from a snapshot captured by a HSR-LFR camera or from a synthesized HSTR frame. We use a frame-recurrent method for super-resolution because the motion between two adjacent frames in LSR-HFR video is often small and the estimates are reliable, which provides a robust recovery. In comparison, motion estimates between the synchronization frame and all the other frames in a GoP-m can be large and unreliable, which seriously affects the performance of RefSR [51].

3.2.2 FlowNet

A popular approach to obtain the temporal motion fields or features is by using the *optical flow* [10]. Let us assume that we can compute optical flow between two frames $I_{\texttt{ref}}, I_{\texttt{LSR}\uparrow}$ as

$$\mathscr{F}=\mathsf{FlowNet}\left(I_{\texttt{ref}},I_{\texttt{LSR}\uparrow}\right), \tag{3}$$

where $I_{\texttt{ref}}$ refers to a reference frame and $I_{\texttt{LSR}\uparrow}$ denote the upscaled LSR-HFR frame at any specific time stamp. For the synchronization timestamp t_i, $I_{\texttt{LSR}\uparrow}=\bar{X}_{t_i}$ and $I_{\texttt{ref}}=X_{t_i}$. For the frames at timestamp $t_i+k\Delta t$, $I_{\texttt{LSR}\uparrow}=\bar{X}_{t=t_i+k\Delta t}$ and $I_{\texttt{ref}}=Y_{t_i+(k-1)\Delta t}$ with $1\le k\le\lfloor m/2\rfloor$. Similarly for the frames at timestamps $t_i-k'\Delta t$, $I_{\texttt{LSR}\uparrow}=\bar{\mathbf{X}}_{\texttt{LSR}-\texttt{HFR}}(t_i-k'\Delta t)$ and $I_{\texttt{ref}}=Y_{t_i-(k'-1)\Delta t}$ with $1\le k'\le\lceil m/2\rceil-1$.

A number of deep neural networks have been proposed to deal with the optical flow [22], [37], [42]. We adopt a pretrained optical flow model PWC-Net [42] as the FlowNet in our frame recurrent AWnet, and then use our data to fine-tune the model through retraining. Pretraining the optical flow network with mass labeled data greatly improves the convergence speed and accuracy of the flow calculation in our work. The size of the estimated optical flow by PWC-Net is $\frac{1}{4}$th of the input image along both spatial dimensions. We use a simple bilinear upsampling method to upscale the low-resolution optical flow field to the same size of the input images.

Spatial details of the reference frame can be transferred using extracted optical flow through the warping operation. Let us denote the warped reference image as $I^w_{\texttt{ref}}$. Since the optical flow is upsampled using a bilinear filter, it often leads to an over-smoothed output.

3.2.3 FusionNet

To preserve the fine motion details, we borrow the idea of dynamic filtering to refine our flow. Dynamic filtering for motion estimation and motion compensation has been recently used in [24]. It estimates an independent convolution kernel for each pixel, which can correctly describe the motion behaviors of each pixel individually. It especially shows accurate estimation and compensation of small motions, but it is not as effective as global optical flow for estimating large motion because of the limited size of convolutional filters. Thus we use dynamic filters to complement the optical flow devised in FlowNet for better performance. As far as we know, we are the first to combine flow network for flow estimation and dynamic filter network for motion refinement.

We observe in our experiments that even with the refinement of the optical flow using dynamic filters, the warped $I^w_{\texttt{ref}}$ fails in region with occlusions and suffers from motion and warping artifacts. In such regions, we need the information from $I_{\texttt{LSR}\uparrow}$. Therefore, we learn a mask to create a weighted combination of the warped reference image and $I_{\texttt{LSR}\uparrow}$ for every pixel. Our FusionNet is designed to perform motion refinement using dynamic filters and provide an adaptive weighting mask as an output. The structure of FusionNet is illustrated in Fig. 2(d).

To utilize all the information in the available frames, we explicitly feed warped reference frame $I^w_{\texttt{ref}}=\mathsf{Warp}(I_{\texttt{ref}},\mathscr{F})$,

upscaled LSR-HFR frame $I_{\mathtt{LSR}\uparrow}$, optical flow $\mathscr{F}$, residual between warped reference and upscaled LSR-HFR frame $r = I^{w}_{\mathtt{ref}} - I_{\mathtt{LSR}\uparrow}$, to the FusionNet as inputs. We can describe the FusionNet as the following function:

$$\mathbf{Fm} = \mathsf{FusionNet}(I^{w}_{\mathtt{ref}}, I_{\mathtt{LSR}\uparrow}, \mathscr{F}, r), \tag{4}$$

whose output has 26 channels that we use to calculate the dynamic filter and adaptive weighting mask.

We use the popular U-net architecture [39] for our FusionNet, as shown in Fig. 2(d). We downscale the feature maps by a factor of two in each of the three downscaling layers and then upscale the features using bilinear interpolation for computational efficiency. In contrast, existing approaches use transposed convolution layers for upscaling, which is computationally expensive and also causes some checkerboard artifacts. We do not use any skip connection in conventional U-net, which greatly reduces the memory requirements of our network. The output of FusionNet is a three-dimensional tensor of feature maps that has 26-channels and same spatial size as the input image frame. We use the first 25 channels to produce 5×5 dynamic filters (one filter per pixel), and the last one to produce the weighted mask for every pixel. Let us denote the dynamic filter for pixel (x, y) as a 5×5 $K_{x,y}$ matrix that can be written as

$$K_{x,y}(i, j) = \mathbf{Fm}(x, y, 5(i-1) + j - 1), \quad \text{for } i, j = 1, \ldots, 5. \tag{5}$$

Let us denote the weighted mask for the entire image as a matrix M with same size as input image whose value at pixel (x, y) can be written as

$$M(x, y) = \mathsf{sigmoid}[\mathbf{Fm}(x, y, 25)]. \tag{6}$$

To summarize, an output frame Y of the reconstructed HSTR video can be synthesized for pixel (x, y) as

$$\begin{aligned} Y(x, y) = M(x, y) I^{wk}_{\mathtt{ref}}(x, y) \\ + (1 - M(x, y)) I_{\mathtt{LSR}\uparrow}(x, y), \end{aligned} \tag{7}$$

where

$$I^{wk}_{\mathtt{ref}}(x, y) = \sum_{i,j=1}^{5} K_{x,y}(i, j) I^{w}_{\mathtt{ref}}(x - 3 + i, y - 3 + j). \tag{8}$$

The reconstructed Y in (7) (together with $I_{\mathtt{LSR}\uparrow}$ from next time instant) will be fed into our AWnet module (as a typical RefSR) in a recurrent manner to recover other HSTR frames reconstruction as exemplified in Fig. 2(b).

3.2.4 Reference Structure

The synthesis model discussed in the previous sections assumes an ultra-low latency application scenario. Thus, we use a single reference frame in AWnet and denote it as a *Single-Reference* AWnet. In practice, we can easily extend this method to include multiple references. One obvious example is to use two references such that one reference frame precedes current LSR-HFR frame, and the other one succeeds it (e.g., two consecutive HSR-LFR frames in Fig. 2(a) used to super-resolve $I_{\mathtt{LSR}}$s in between). We refer to it as the *Multi-Reference* AWnet.

To enable *Multi-Reference* AWnet with two references, we use two FlowNets to estimate the flows between two pairs: ($I_{\mathtt{ref0}}$, $I_{\mathtt{LSR}\uparrow}$) and ($I_{\mathtt{ref1}}$, $I_{\mathtt{LSR}\uparrow}$). Then we warp $I_{\mathtt{ref0}}$ and $I_{\mathtt{ref1}}$ to compute $I^{w}_{\mathtt{ref0}}$ and $I^{w}_{\mathtt{ref1}}$ accordingly. Similar to the *Single-Reference* AWnet, we input warped reference frames, upscaled LSR-HFR frame, optical flows, and residual frames between warped reference and upscaled LSR-HFR frame to the same FusionNet for dynamic filters and masks generation. In the case of Two-Reference AWnet, we increase the number of output channels in the FusionNet from 26 to 53, where the first 50 produce two sets of 5×5 dynamic filters $K^{\mathtt{ref0}}$ and $K^{\mathtt{ref1}}$, and the remaining 3 provide adaptive weighting masks $M_{\mathtt{ref0}}$, $M_{\mathtt{ref1}}$ and $M_{\mathtt{LSR}\uparrow}$. We replace the original sigmoid function in (6) with a softmax function to enable the multi-reference weighting, where the sum of these three masks at each pixel position is 1. Finally, reconstructed $Y(x, y)$ with two reference frames can be expressed as

$$\begin{aligned} Y(x, y) = & M_{\mathtt{ref0}}(x, y) I^{wk}_{\mathtt{ref0}}(x, y) + M_{\mathtt{ref1}}(x, y) I^{wk}_{\mathtt{ref1}}(x, y) \\ & + M_{\mathtt{LSR}\uparrow}(x, y) I_{\mathtt{LSR}\uparrow}(x, y). \end{aligned} \tag{9}$$

In this example, we can set original $I_{\mathtt{ref}}$ in Fig. 2(c) as $I_{\mathtt{ref0}}$, and duplicate a reference branch for incoming $I_{\mathtt{ref1}}$.

We can further extend two-reference AWnet to support more reference frames by duplicating reference branches in Fig. 2(c) and modifying the last layer of the FusionNet in Fig. 2(d) for dynamic filters and mask generation appropriately. Note that softmax activation can be utilized to support multiple reference weighting.

We will show in the ablation studies below that multiple-reference AWnet provides noticeable improvement over single-reference AWNet in the synthesized reconstruction with better spatial texture details and temporal continuity. Objective improvement measured by averaged PSNR over thousands of videos in Vimeo90K test dataset is given in Table 5, and subjective comparisons are supplemented with the demonstration videos at our website[4]. Applying either *Single-Reference* or *Multi-Reference* AWnet is dependent on the underlying application, where *Single-Reference* is preferred if ultra-low latency condition is demanded, otherwise, *Multi-Reference* AWnet is favored.

4 Training

4.1 Training Dataset

We use the Vimeo90K dataset [47] to train our model. The Vimeo90K dataset has 64,612 septuplets for training, where each septuplet contains 7 consecutive video frames at a size of 256×448 pixels. For each septuplet, we randomly select two consecutive frames as a pair for training. Specifically, one frame is used as the *reference* $I_{\mathtt{ref}}$ to mimic the input image from a HSR-LFR camera, and a downscaled version of the next frame is used as the *target frame*. We apply a native bicubic downsampling filter offered by the open source FFmpeg[5] to mimic the input image $I_{\mathtt{LSR}}$ from a LSR-HFR camera. And, we randomly crop each frame from its original resolution to a size of 256×384 on-the-fly for training data augmentation.

4.2 Training Strategy and Loss Function

Training process for our network has four main steps. We use the Adam [29] optimizer by setting its parameters β_1 and β_2 to 0.9 and 0.999, respectively. We use a batch size of 4. Details of every training step are as follows.

- **Step 0: FlowNet Initialization.** We use the pretrained PWC-Net [42] to initialize our FlowNet, which is trained with a

4. http://yun.nju.edu.cn/d/def5ea7074/?p=/MultiReferenceAWnet&mode=list

5. www.ffmpeg.org

TABLE 2
Objective Performance Comparison of Super-Resolution Methods on Vimeo90K Dataset [47].

Methods		4×		8×	
		PSNR	SSIM	PSNR	SSIM
SR	EDSR [33]	33.11	0.9413	28.20	0.8702
	ToFlow-SR [47]	33.08	0.9417	-	-
RefSR	PM [7]	35.06	0.9670	31.30	0.9380
	CrossNet [51]	39.17	0.9852	36.15	0.9766
	AWnet	**39.88**	**0.9862**	**36.63**	**0.9768**

TABLE 3
Objective Performance Comparison of Frame Interpolation using Vimeo90K [47] (downscaled 4th frame used as reference in AWnet).

	PSNR	SSIM
ToFlow-Intp. [47]	33.46	0.9615
AWnet with 1/64 reference	**36.63**	**0.9768**
AWnet with 1/16 reference	**39.88**	**0.9862**

large set of data with ground truth optical flow. The inputs of PWC-Net are two consecutive frames at the same resolution.

- **Step 1: FlowNet Fine-tuning.** The reference frame $I_{\texttt{ref}}$ and low-resolution target frame $I_{\texttt{LSR}}$ have a large gap in their sizes. Thus, we implement a fine-tuning step to improve the FlowNet. First, we upscale $I_{\texttt{LSR}}$ to $I_{\texttt{LSR}\uparrow}$ with the same size as $I_{\texttt{ref}}$. Then we compute optical flow between $I_{\texttt{LSR}\uparrow}$ and $I_{\texttt{ref}}$ using PWC-Net. The computed optical flow $\mathscr{F}$ is then used to warp $I_{\texttt{ref}}$ and produce $I^{w}_{\texttt{ref}}$. Then we apply an ℓ_1 norm-based warping loss to fine-tune the FlowNet, which is shown below:

$$\mathcal{L}_{warp} = ||I_{\texttt{gt}} - I^{w}_{\texttt{ref}}||_1, \tag{10}$$

  where $I_{\texttt{gt}}$ is the high-resolution ground truth of $I_{\texttt{LSR}}$. A small learning rate of $1e-6$ is used to fine-tune the FlowNet with 40k iterations. A similar loss function has been used in [23].
- **Step 2: FusionNet Pretraining.** A pretraining step is also used for FusionNet. To train the FusionNet, we fix the FlowNet and let the network select appropriate parameters for FusionNet during training. We use an ℓ_1 loss between the output Y and the ground truth $I_{\texttt{gt}}$, given as

$$\mathcal{L}_{rec} = ||I_{\texttt{gt}} - Y||_1. \tag{11}$$

  We set the learning rate to $1e-4$ and train the network for 100k iterations, according to our extensive simulation studies.
- **Step 3: End-to-End Joint Training.** Starting with our pretrained models, we jointly train FlowNet and FusionNet by minimizing the same end-to-end ℓ_1 loss in (11). In this step, we set learning rate to 10^{-5} for FusionNet and 3×10^{-6} for FlowNet over 100k iterations. With such pre- and joint-training, network model can converge faster with more robust and reliable behavior.

All networks are implemented and verified using PyTorch. In subsequent sections, we describe the experiments we performed to evaluate different aspects of proposed AWnet for our dual camera system.

5 Experiments

We conduct experiments on two types of videos. One type is the "simulation data" that has images/videos from the existing and public accessible datasets (e.g., Vimeo90K, KITTI, Flower, LFVideo and Stanford Light Field datasets); the other type is the "real data" captured by real cameras (e.g., iPhone 7 and Grasshopper3 cameras) under different settings.

5.1 Performance Comparison using Simulation Data

We first compare our method with the state-of-the-art SISR method EDSR [33], task-oriented video super-resolution method ToFlow-SR [47], conventional RefSR patchmatch (PM) [7], and the state-of-the-art learning-based RefSR CrossNet [51]. To be fair, we retrain CrossNet with our dataset following the training strategy suggested in [51].

Super-resolution: We first use the test set with 7,824 septuplets from Vimeo90K [47] for performance comparison. We select the fourth image in each septuplet for evaluation following the suggestion in [4], [47]. For video super-resolution method, the input is the downscaled septuplet sequence and the target is the super-resolved fourth frame. For a single frame or image RefSR, we downscale the fourth frame and use the fifth frame as the reference frame. The results are presented in Table 2. We use PSNR and Structural Similarity (SSIM) [44] as our performance metrics for evaluation. Results show that our method has superior performance in both PSNR and SSIM for $4\times$ super-resolution along both spatial dimensions. For PSNR, it yields $\approx$ 0.7 dB, 4.8 dB, 6.8 dB, and 6.7 dB gains against CrossNet, PM, ToFlow-SR, and EDSR, respectively. Similar gains are produced for $8\times$ super-resolution factor, demonstrating the generalization of our work to various application scenarios.

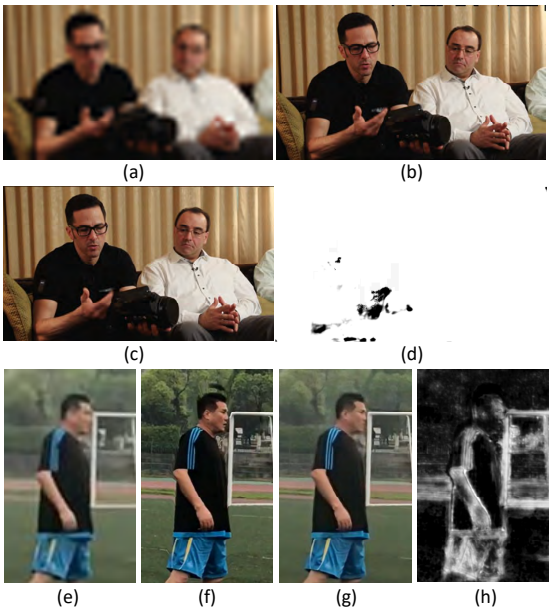
In addition to the experiments using Vimeo90K testing samples, we also tested other datasets such as KITTI, Flower, LFVideo and Stanford Light Field data to evaluate the performance of our proposed AWnet. We discuss those experiments in Section 6 where we analyze the impact of camera parallax.

Frame interpolation: Our AWnet can also be used to interpolate missing intermediate frames (usually at high spatial resolution) with the help from another LSR-HFR input. Such frame interpolation is also supported by optical flow based methods, such as ToFlow-Intp in [47]. We use the third and the fifth frames from the testing septuplets to interpolate missing fourth frame. But for our method, we downscale fourth frame (e.g., $8\times$ resolution downscaling at both spatial dimension) as another input. The results in Table 3 suggest that even a thumbnail-size image of its original source (e.g., $1/8 \times 1/8$ the size of the original image), can improve the quality of the interpolated intermediate frame significantly. A remarkable 6.4 dB PSNR gain is recorded compared to ToFlow-Intp [47] when scaling the fourth image to its $1/4 \times 1/4$ size (i.e., $16\times$ fewer pixels) and 3.2 dB PSNR gains for the case when scaling fourth image to its $1/8 \times 1/8$ size (i.e., $64\times$ fewer pixels).

Model efficiency: Our AWnet demands less system resource with less space and time complexity requirements. For example, AWnet model has 109.5 MB parameters, about 25% reduction when compared with the CrossNet model at a size of 140.8 MB parameters. When upscaling a snapshot at a factor of $8\times$ spatially to the size of 640×448, AWnet consumes about 0.12 second with 1499 MB running memory (e.g., about 60% reduction against the running memory consumption of CrossNet), while CrossNet is about 0.18 second with 4511 MB running memory. As an comparative anchor, traditional PM [7] uses 55.9 seconds due to iterative patch match.

(a) (b) (c) (d) (e)

Fig. 3. **Dual Camera Alignment.** (a) HSR-LFR frame $I_{\mathtt{ref}}$; (b) LSR-HFR frame $I_{\mathtt{LSR}}$; (c) HSR-LFR frame $I_{\mathtt{ref}}$ frame warped using optical flow only; (d) HSR-LFR frame $I_{\mathtt{ref}}$ warped using mesh-based homography; (e) HSR-LFR frame warped using both mesh-based homography and optical flow. (a) and (b) are captured using dual iPhone 7 with different views.



(a) (b) (c) (d) (e) (f) (g) (h)

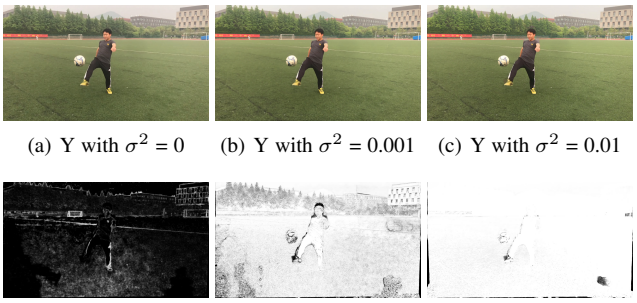
Fig. 4. **Synthesized Quality and Weighting Map** W**.** (a) to (d) are exemplified for Vimeo90K simulation data: (a) is the up-scaled $I_{\mathtt{LSR}\uparrow}$ using bicubic method from $I_{\mathtt{LSR}}$ by $8\times$ for both spatial dimensions; (b) is the warped reference frame $I^w_{\mathtt{ref}}$; (c) is the output synthesized frame Y; (d) is the adaptive weighting map W on (b); (e) to (h) are the visualization for camera captured real data with $3\times$ resolution scaling from $I_{\mathtt{LSR}}$ to $I_{\mathtt{LSR}\uparrow}$: (e) is the up-scaled image $I_{\mathtt{LSR}\uparrow}$ from the captured LSR-HFR frame; (f) is the warped reference frame $I^w_{\mathtt{ref}}$; (g) is the output synthesized frame Y; (h) is the adaptive weighting map W on (f).

5.2 Performance Studies using Real Data

We perform the real video data capture using dual iPhone 7 and Grasshopper3 cameras. One represents a consumer mobile camera used massively and the other one a camera commonly used for scientific or industrial imaging applications.

5.2.1 Camera Alignment

Dual camera setup requires careful calibration to map their relative coordinates and poses; especially, if we move the system around for shooting different scenes. As suggested by [34], we choose mesh-based homography for alignment. which greatly improves the accuracy of subsequent optical flow derivation as shown in Fig. 3. More specifically, we extract SURF features [6] for both HSR-LFR and LSR-HFR frames, and then use the matched feature points to derive the homography transformation matrix for alignment. Fig. 3(c) shows that optical flow-based alignment fails to produce a good image. We then show the HSR-LFR frame aligned using the mesh-based homography in Fig. 3(d). Fig. 3(e) reveals that homography-based alignment followed by optical flow-based refinement greatly improves the image quality.



(a) Y with $\sigma^2 = 0$ (b) Y with $\sigma^2 = 0.001$ (c) Y with $\sigma^2 = 0.01$

(d) W with $\sigma^2 = 0$ (e) W with $\sigma^2 = 0.001$ (f) W with $\sigma^2 = 0.01$

Fig. 5. **Noise Regularization.** (a)-(c) Reconstructions of synthesized Y with noise level σ^2 at 0, 0.001 and 0.01; (d)-(f) Weighting map W with noise level σ^2 at 0, 0.001 and 0.01. Video frames are captured using dual iPhone 7. Noise regularization shifts more weights to $I^w_{\mathtt{ref}}$ in general to improve the image quality, especially for those background stationary areas. But for those regions with occlusions (edge of the athletes), motion blurs (soccer ball) and warping artifacts (grassland), reconstruction still prefers pixels from $I_{\mathtt{LSR}\uparrow}$ to minimize the training loss. Ghosting artifact appear around the player's head in (b, c), which occur if the AWnet selects blurry regions from $I_{\mathtt{LSR}\uparrow}$ for the pixels that are not correctly aligned in $I^w_{\mathtt{ref}}$.

In temporal dimension, for dual iPhone 7 configuration, we use a millisecond timer to synchronize two cameras. Synchronization with industrial cameras is much easier where we can apply the hardware clock timer.

Our dual camera system is connected to a high performance computer for both HSR-LFR and LSR-HFR video caching. The computer is using an Intel Xeon E5-2620 running at 2.10 GHz with a GeForce GTX 1080Ti GPU. Without any platform and algorithmic optimization, it now takes about 7.4 seconds to synthesize a 3840×2160 frame from a pair of 1280×720 frames at 240FPS and 3840×2160 at 30FPS videos.

5.2.2 Noise Regularization

We first directly applied our model on data captured with real camera, but we found that the quality of synthesized images was not as good as we had in the case of simulation data, as shown in Fig. 4 (see (c) versus (g)). We observe that the reconstruction of data captured with iPhone 7 camera is blurry in Fig. 4(g), even though the reference image after warping retains sharp details, as shown in Fig. 4(f).

Recall that the synthesized video frame Y in (7) is a weighted combination of the warped reference $I^w_{\mathtt{ref}}$ and upscaled LSR-HFR input $I_{\mathtt{LSR}\uparrow}$. The relative weighting factor describing the weight contribution from the warped reference for a pixel at (x, y)-th position can be written as

$$W(x,y) = \frac{M(x,y)\sum_{i,j=1}^{5} K_{x,y}(i,j)}{(1-M(x,y)) + M(x,y)\sum_{i,j=1}^{5} K_{x,y}(i,j)}. \quad (12)$$

Note that the value of $W(x, y)$ lies within [0, 1]. The larger the $W(x, y)$ is, the more the $I^w_{\mathtt{ref}}$ contributes and vice versa. As revealed in Fig. 4(e), and Fig. 4(f), our model shifts more weights to $I_{\mathtt{LSR}\uparrow}$, rather $I^w_{\mathtt{ref}}$ for captured real image synthesis, resulting in over smoothed reconstruction of Y.

We also observe that $I^w_{\texttt{ref}}$ in Fig. 4(f) not only preserves the sharp details but also the rich colors compared to Y in Fig. 4(g). One potential cause for this is that the camera parameters adjust automatically during recording to improve the image quality (e.g., ISO setting, aperture size, etc.). Therefore, the resulting LSR-HFR video has a narrow dynamic range and low SNR, as shown in Fig. 4(e) versus Fig. 4(f) from an associated HSR-LFR camera. In another words, the overall quality of $I_{\texttt{LSR}}$ (or $I_{\texttt{LSR}\uparrow}$) is much worse compared to the corresponding $I_{\texttt{ref}}$ (or $I^w_{\texttt{ref}}$).

In the case of simulation data, we do not face this problem because both $I_{\texttt{LSR}}$ and $I_{\texttt{ref}}$ are generated from the same ground truth. Such phenomena are also observed when using other RefSR methods (e.g., PM, CrossNet) to super-resolve real data. The differences between the simulated data and the real sensor data affect the accuracy of optical flow and the performance of FusionNet. The FusionNet relies on the similarity between $I_{\texttt{LSR}\uparrow}$ and $I^w_{\texttt{ref}}$ to combine them. The higher resolution of real sensor data and the difference between camera parameters cause errors in optical flow estimation and mislead the FusionNet to pay less attention to $I^w_{\texttt{ref}}$ as shown in Fig. 4(h). Thus, it seems that models trained using "clean" simulation data can not be directly extended to camera captured real data.

To apply our model on real data, we formulate a "regularization" problem that searches for a better weighting factor W between $I_{\texttt{LSR}\uparrow}$ and $I^w_{\texttt{ref}}$ in (7). We propose to add noise n in $I_{\texttt{LSR}\uparrow}$ for regularization during the training progress. Thus, end-to-end learning optimization in (11) can be updated using regularized Y and $I^w_{\texttt{ref}}$. The only difference here is that instead of using noiseless $I_{\texttt{LSR}\uparrow}$, we inject noise and use $I_{\texttt{LSR}\uparrow} + n$ in all the computations. With such noise regularization on $I_{\texttt{LSR}\uparrow}$, network learns to shift more weights to $I^w_{\texttt{ref}}$ to improve the quality of synthesized reconstruction. Adding synthetic noise at the time of training provides robustness and acts as data augmentation [15], [16], [19], [26], [38], [46], [48], [52].

We train our network with Gaussian noise at different variances σ^2s, e.g., 0, 0.001, and 0.01. Snapshots of reconstructed Y and weighting factor W with various σ^2s are shown in Fig. 5. More comparisons can be seen in the supplementary material. We can see that reconstruction shifts more weights to $I^w_{\texttt{ref}}$ (e.g., elements in W gets closer to 1 in Fig. 5(d)–(f) as the noise increases (i.e., σ^2 increases), yielding better image quality with higher dynamic range, better color and sharper details (see Fig. 5(a)–(c)). Misalignment due to camera parallax and occlusions can introduce artifacts in the recovered images as can be seen around the player's head. If we increase σ^2 further, the quality of images deteriorates. Therefore, to demonstrate the effect of noise regularization, we apply the noise regularization with σ^2 = 0.01 for all evaluations and comparisons in subsequent sections.

5.2.3 Subjective Evaluation

For data captured with the real cameras, we compare the performance of our method with the EDSR [33], PM [7][6], and CrossNet [51].

Super-Resolution. Dual iPhone 7 cameras are used in this study. One camera captures a 4K video at 30FPS as the HSR-LFR input, and the other synchronized camera records 720p video at 240FPS as the LSR-HFR input.

6. Because the size of camera captured video frame is larger than the simulation content in Vimeo90K, we enlarge the patch size from 8 to 16 and search range from 16 to 64 for patch matching.

TABLE 4
Performance Impact of Different Upscaling Filter

SISR	4×		8×	
	PSNR	SSIM	PSNR	SSIM
EDSR [33]	39.88	0.9862	36.63	0.9768
bicubic	39.75	0.9862	36.47	0.9766

We shoot videos for different scenes to validate the efficiency and generalization of our system. These scenes include indoor and outdoor activities with different illumination conditions, as illustrated in Figs. 6, and 7. We can observe the quality improvements of our proposed method when compared with the CrossNet [51], PM [7], and EDSR [33]. With appropriate noise regularization (e.g., σ^2 = 0.01 as exemplified), we could clearly observe that both the spatial details of $I_{\texttt{ref}}$ and accurate motions from $I_{\texttt{LSR}}$ are well retained and synthesized in the final reconstruction. We also notice that the subjective quality improvement is perceivable in our method with low and medium light illumination. With strong light illumination, the state-of-the-art CrossNet also provides good reconstruction, but our method still provides the best results, as shown in supplemental material[7].

Frame Interpolation. We extend our evaluations to frame interpolation. We present the entire GoP reconstructions in Fig. 8 for subjective comparison. $I_{\texttt{ref}}$-0 and $I_{\texttt{ref}}$-1 are the original frames from our HSR-LFR camera, while the frames in-between interpolated using ToFlow-Intp [47] are presented in the upper rows (highlighted with the green box). For comparison, reconstructed Y frames using our model are placed in the bottom rows. As we can see, ToFlow-Intp shows ghosting and motion blurring (e.g., almost invisible fast-dropping ping-pong ball in the upper part of Fig. 8). Our proposed method recovers the high-fidelity spatial details (see the woman's face and the texts on the wall) and accurate motions (see the fast-moving ping-pong ball and the woman's hands) at the same time.

6 Ablation Studies

In this section we investigate different parameters of our system individually to understand the system capability and the source of efficiency.

Upscale Filter. We upscale the $I_{\texttt{LSR}}$ in Fig. 2(c) to the same resolution as the $I_{\texttt{ref}}$ for subsequent processing. Previous explorations assume the state-of-the-art SISR method EDSR [33]. Here, we replace it with a straightforward bicubic filter. Models are re-trained with this new upscaling filter, and performance comparison is evaluated on the efficiency of respective $4\times$ and $8\times$ super-resolution application using the Vimeo90K testing dataset. Averaged PSNR and SSIM are listed in Table 6, showing that different upscaling method does not affect the overall performance noticeably, e.g., $\approx$ 0.1 dB PSNR and $<=$ 0.002 SSIM index variations reported. This observation suggests that we can use simple upsampling filters to scale up $I_{\texttt{LSR}}$ instead of complex super-resolution methods. In principle, this is mainly due to the fact that our AWnet-based dual camera system could learn and embed high frequency spatial information from its HSR-LFR input for final reconstruction synthesis. Thus, complex super-resolution method used to estimate high frequency component is not an inevitable step any more.

7. The reconstructed videos are available at http://yun.nju.edu.cn/d/def5ea7074/?p=/Illumination&mode=lis.

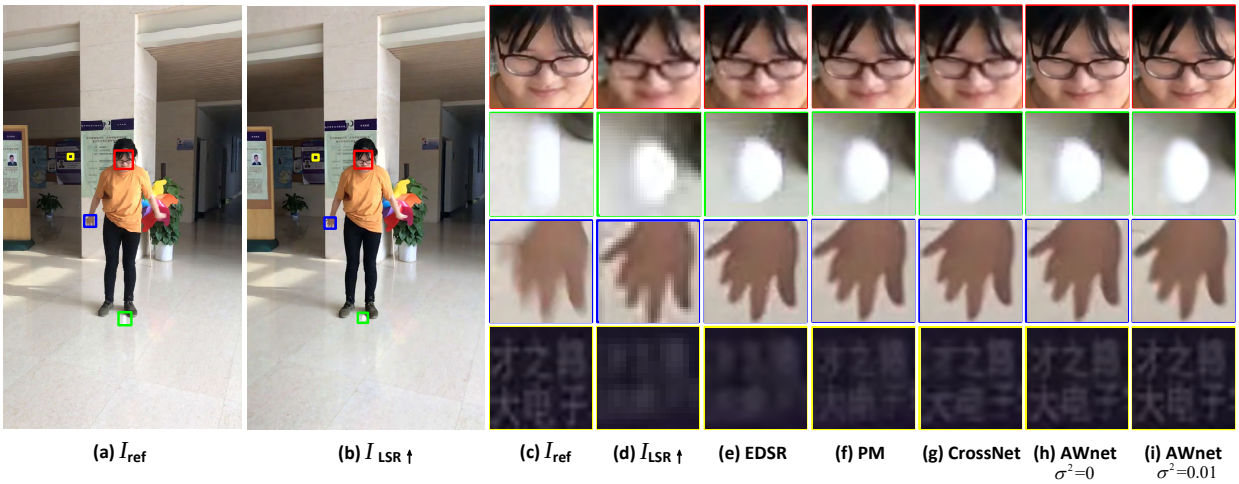


Fig. 6. **Super-Resolution:** Indoor activity with medium light illumination. $I_{\texttt{ref}}$ is the captured 4k frame from the HSR-LFR camera, and synchronized $I_{\text{LSR}\uparrow}$ is the up-scaled frame from the captured 720p frame of the LSR-HFR camera. Zoomed-regions are visualized for (c) $I_{\texttt{ref}}$, (d) $I_{\text{LSR}\uparrow}$, and super-resolved reconstructions using (e) EDSR, (f) PM, (g) CrossNet, (h) AWnet with $\sigma^2 = 0$ (no noise regularization), and (i) AWnet with $\sigma^2 = 0.01$.

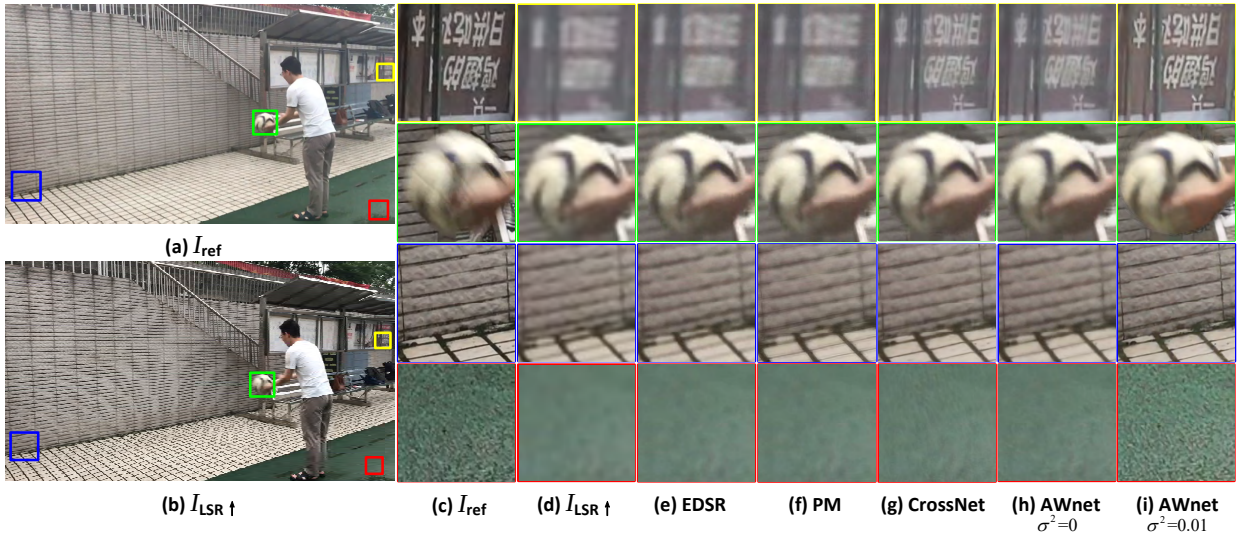


Fig. 7. **Super-Resolution:** Outdoor activity with low light illumination. $I_{\texttt{ref}}$ is the captured 4k frame from the HSR-LFR camera, and synchronized $I_{\text{LSR}\uparrow}$ is the up-scaled frame from the captured 720p frame of the LSR-HFR camera. Zoomed-regions are visualized for (c) $I_{\texttt{ref}}$, (d) $I_{\text{LSR}\uparrow}$, and super-resolved reconstructions using (e) EDSR, (f) PM, (g) CrossNet, (h) AWnet with $\sigma^2 = 0$ (no noise regularization), and (i) AWnet with $\sigma^2 = 0.01$.

TABLE 5
Objective PSNR of Reconstructed Images for *Single-Reference* and *Multi-Reference* AWnet.

	Single-Reference	Multi-Reference
averaged	35.33 dB	36.21 dB

Reference Structure. We also compare the performance of *Single-Reference* and *Multi-Reference* AWnet. We evaluate the *Single-Reference* and *Multi-Reference* AWnet with Vimeo90K test dataset (e.g. 7824 video sequences in total). For *Multi-Reference* AWnet, we use 7 frames in each video sequence; treat the first and the last frame as the references; and recover the remaining five frames that are down-scaled $8\times$ along both horizontal and vertical directions. In comparison, *Single-Reference* AWnet is configured same as described in Section 5.1. The averaged PSNRs for all test videos are shown in Table 5. We observe that *Multi-reference* AWnet offers 0.9 dB gain compared to *Single-reference* AWnet. Subjective quality is also improved with better temporal continuity and spatial texture details by introducing the multiple reference in AWnet[8]. Since CrossNet uses only one reference frame, we only present results for *Single-Reference* structure in other comparative studies.

FusionNet. In our proposed method, FusionNet uses adaptive weighting fusion (AWFusion) on upscaled and warped images to produce output pixels. Alternatively, we can also apply convolution-based direct prediction. For such convolution-based direct prediction (ConvDP), we replace the U-net style architecture for dynamic filter and mask generation in Fig. 2(d) with stacked convolutions. Here, we utilize 36 convolutional layers with the same 3×3 kernel, and nonlinear ReLU activation. The remaining parts of the AWnet framework are kept without change. We then

8. The videos for multi-reference AWnet results are available at http://yun.nju.edu.cn/d/def5ea7074/?p=/MultiReferenceAWnet&mode=list

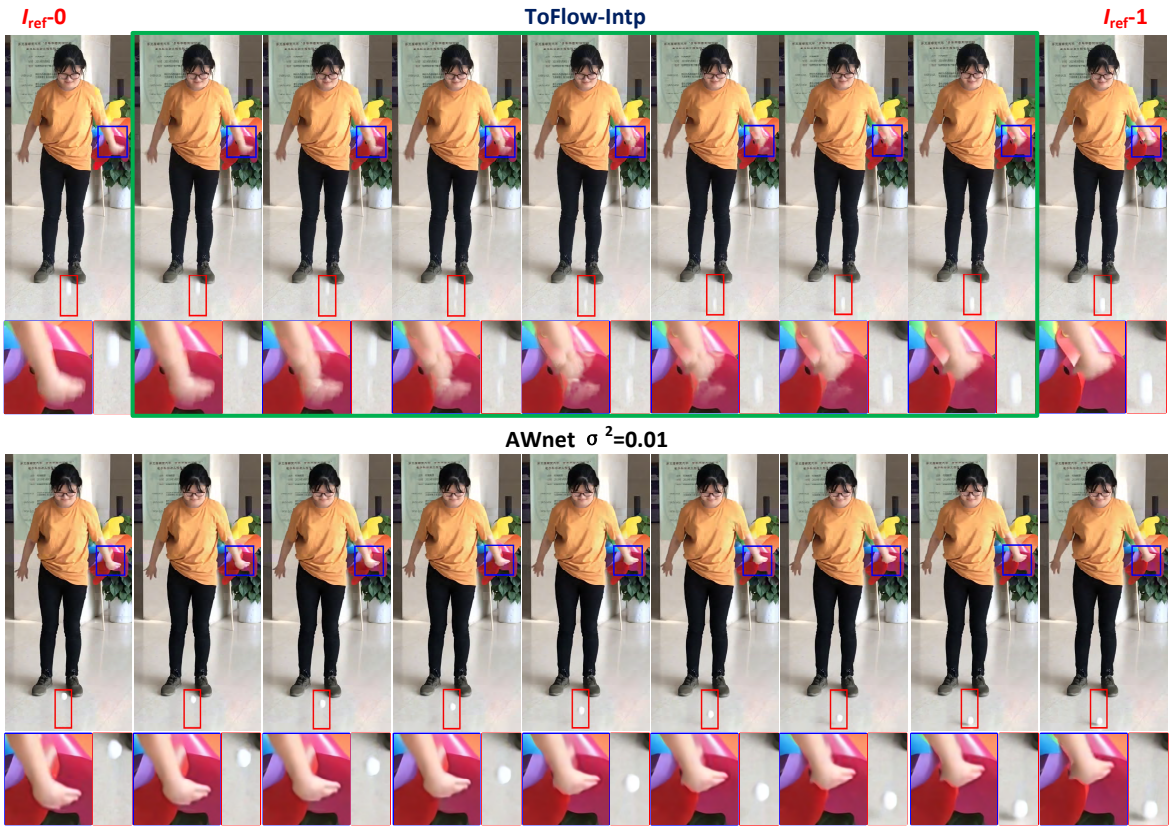


Fig. 8. **Frame interpolation.** Indoor activity with medium light illumination. The most left and the most right of first rows are the captured HSR-LFR frames. Seven frames in-between are interpolated using ToFlow-Intp [47]; The second rows are the synthesized HSTR frames using our AWnet, which is trained with noise regularization with the variance of 0.01. **Zoom in the pictures, and you will see more image details.** (More in supplemental material.)

TABLE 6
Objective Comparison between Adaptive Weighting Fusion and Convolution-based Direct Prediction for Output Pixel Reconstruction.

	4×		8×	
	PSNR	SSIM	PSNR	SSIM
AWFusion	39.88	0.9862	36.63	0.9768
ConvDP	39.67	0.9856	36.45	0.9756

evaluate the default AWFusion and ConvDP using the Vimeo90K test dataset (e.g., 7824 video sequences in total). Table 6 provides the averaged PSNR and SSIM results of the reconstructions using respective methods. Although the quantitative gains in terms of PSNR and SSIM are not very large, Fig. 9 shows noticeable quality improvement of synthesized frame with less noise and sharp/clear reconstruction for the AWFusion compared to ConvDP (e.g., Fig. 9(c) vs (d)).

Camera Parallax. Dual camera setup is used in our system. Thus camera parallax could be an issue that affects the system performance. We show in our studies below by using simulation data from available KITTI [36], Flower [41], LFVideo [43], Stanford light field [1] datasets, and real data captured by our dual camera system with various parallax settings to demonstrate the robustness of our method.

Comparison Using Simulation Data: We test our AWnet on Flower [41], LFVideo [43] and Stanford light field (Lego Gantry) [1] datasets following the same configuration in [51]. The Flower and LFVideo datasets are light field images captured using Lytro ILLUM camera. Each light field image has 376×541 spatial samples and 14×14 angular samples (grid). The same as the methods applied in [41], [51], we extract the central 8×8 grid of angular samples to avoid invalid images. Parallax is offered by setting the reference image $I_{\mathtt{ref}}$ at (0, 0), and associated low-resolution correspondence at (i, i), with $0 < i \leq 7$ by shifting position to another different angular sample. For example, Flower(1,1) and LFVideo(7,7) in Table 7, represent low-resolutions at (1,1) and (7,7) with respect to the respective references at (0,0). Images in both Flower and LFVideo datasets exhibit small parallax settings [41], [51]. On the other hand, Stanford light field dataset contains the light field images shot using a Canon Digital Rebel XTi with a canon 10-22 mm lens. It is placed using a movable Mindstorms motor on the Lego gantry, where the parallax is introduced by the baseline distances along with the camera movement. Under such equipment settings, the captured light-field images have much larger parallax than those captured by Lytro ILLUM camera. Both Table 7 and Table 8 also shows the leading performance of our proposed AWnet at a variety of parallax between testing and reference images, further demonstrating the generalization of our network in different application scenarios. Especially, on average, up to 1.3 dB PSNR improvement is obtained of our AWnet against the CrossNet in Table 8 for large parallax setting. This is mainly due to the reason that CrossNet was not initially designed for RefSR with larger parallax. Thus, additional parallax augmentation procedure was suggested in [51] for re-training.

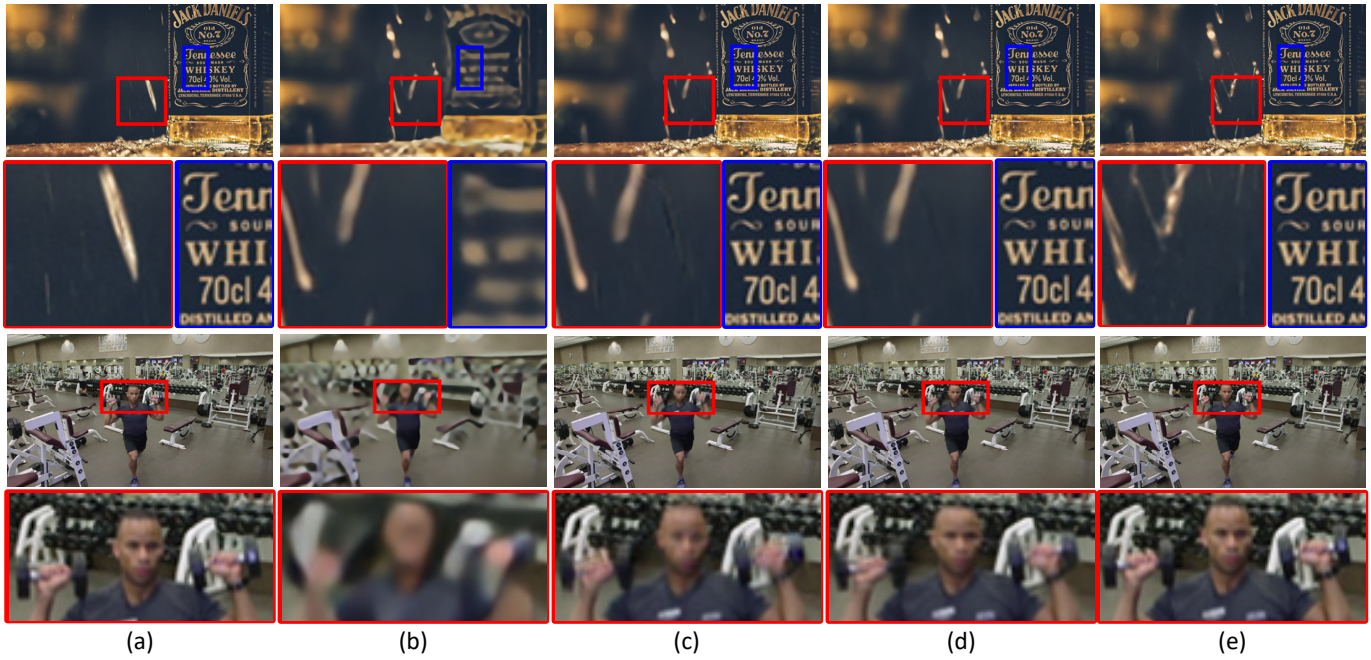


Fig. 9. **Subjective Evaluation of Adaptive Weighting Fusion and Convolution-based Direction Prediction Using Synthetic Vimeo90K Test Dataset.** (a) $I_{\mathtt{ref}}$; (b) $I_{\mathtt{LSR}}$; (c) Convolution-based Direct Prediction; (d) Adaptive Weighting Fusion; (e) Ground truth. The upper part is a raining scene with zoomed rain drop and label. The bottom part is a gym scene with zoomed face of a trainee.

TABLE 7
Objective Performance Comparison of 4× and 8× Super-Resolution Methods on Flower and LFVideo Datasets

Methods	Scale	Flower(1,1)		Flower(7,7)		LFVideo(1,1)		LFVideo(7,7)	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SRCNN [14]	4×	32.76	0.89	32.96	0.90	32.98	0.86	33.27	0.86
VDSR [28]	4×	33.34	0.90	33.58	0.91	33.58	0.87	33.87	0.88
MDSR [33]	4×	34.40	0.92	34.65	0.92	34.62	0.89	34.91	0.90
PM [7]	4×	38.03	**0.97**	35.23	0.94	38.22	0.95	37.08	0.94
CrossNet [51]	4×	41.23	0.9625	38.09	0.9475	41.57	**0.9758**	39.17	0.9627
AWnet	4×	**41.33**	0.9631	**38.31**	**0.9492**	**41.63**	0.9757	**39.36**	**0.9635**
SRCNN [14]	8×	28.17	0.77	28.25	0.77	29.43	0.75	29.63	0.76
VDSR [28]	8×	28.58	0.78	28.68	0.78	29.83	0.77	30.04	0.77
MDSR [33]	8×	29.15	0.79	29.26	0.80	30.43	0.78	30.65	0.79
PM [7]	8×	35.26	0.95	30.41	0.85	36.72	0.94	34.48	0.91
CrossNet [51]	8×	**39.35**	**0.9571**	34.11	0.9149	**40.63**	**0.9727**	36.97	0.9465
AWnet	8×	39.29	**0.9571**	**34.53**	**0.9199**	40.48	0.9725	**37.25**	**0.9487**

TABLE 8
Objective Performance (PSNR) Comparison of 8× Super-Resolution Methods on Stanford Light Field Dataset

Methods	parallax = (1,0)	(3,0)	(5,0)
MDSR [33]	29.66	29.66	29.67
PM [7]	34.61	32.55	30.42
CrossNet [51]	39.33	36.77	35.15
AWnet	**39.53**	**37.65**	**36.47**

TABLE 9
Performance Evaluation Using KITTI dataset for Super-Resolution

	4×		8×	
	PSNR	SSIM	PSNR	SSIM
EDSR [33]	27.03	0.8519	23.47	0.7377
CrossNet [51]	27.43	0.8631	24.92	0.7981
AWnet	**28.19**	**0.8882**	**26.01**	**0.8356**

KITTI dataset has 54cm baseline distance for two cameras. We apply our method and CrossNet with pretrained models using Vimeo90K dataset directly to KITTI test data with 400 stereo image pairs. We use the high-resolution left-view images as the reference for the low-resolution right-view images. Since both CrossNet and our method expect the image resolution to be divisible by 64, thus, we crop images to 1216×320 for testing. Table 9 gives the PSNR and SSIM for super-resolution evaluation. Our method still offers better PSNR (e.g., $>$ 1 dB gain for $8\times$ resolution scaling factor) and SSIM compared to the CrossNet. EDSR results are offered as a reference point, revealing that RefSR still exhibits superior performance, even with a large camera parallax (i.e., 54cm baseline in KITTI data). Results in Table 9 and Table 2 suggest that both PSNR and SSIM are dropped significantly when evaluating models on KITTI compared to the Vimeo90K test data. This is because the introduction of the (large) camera parallax leads to the inaccurate flow estimation for later processing. Similar observations are reported in CrossNet [51]

where PSNR and SSIM drop as the camera parallax increases.

Comparison Using Real Data: We choose a pair of Grasshopper3 cameras to perform more parallax studies due to its easy setup using industrial cameras. We use two Grasshopper3 GS3-U3-51S5C cameras with respective 20mm and 6mm lens installed. There is nearly 4× resolution gap between these two cameras, e.g., one is at 2304×2048, and the other one is at 576×512. We fix the frame rate of the HSR-LFR camera at 240FPS and the frame rate of the LSR-HFR camera at 30FPS. Viewing distance from the cameras to the scene is about 2 meters, and the baseline distance between these two cameras are adjusted at 5cm, 10cm, 15cm, 20cm and 25cm for a variety of parallax configurations. Figure 10 plots the reconstructed images at different baseline distances. As we can see, our system has reliable performance at a variety of parallax settings. Image quality can be enhanced noticeably with noise regularization, as shown in enlarged thumbnails in Figure 10. And in the region with repeated patterns, the checkerboard, there are some ghosts on the results of CrossNet, but our network does not have this problem. Timer digits are over-smoothed by CrossNet, especially for the scenarios with larger baseline distance (e.g., 25cm), but ours still retain the sharp presentation.

Exposure time. The exposure time affects the number of arrival photons, thus having impacts on the image quality for each snapshot and subsequent synthesis performance. Identical dual camera setup is used as in parallax study, but with the baseline distance fixed at 4.4cm. We fix the aperture sizes of the two cameras and set the ISO gain to automatic mode.

The exposure time of the HSR-LFR camera is fixed as default to record 30FPS reference video with 2304×1920 resolution at 30FPS. For comparative studies, We set the exposure time of the LSR-HFR camera to 10ms, 5ms, 2.5ms , 1ms and 0.5ms respectively, to record video with 576×480 resolution with 100FPS. Both captured and reconstructed images are shown in Fig. 11. From the captured LSR-HFR frames, we can see that the signal-to-noise ratio (SNR) of the images decrease greatly (Fig. 11(d)) as the decrease of the exposure time.

Experiments reveal that CrossNet can remove some noise but the capacity is limited. This may be due to the smoothing effect of global convolutions applied, leading to blurred timing digits and jewelry contour (see Fig. 11(e)). Our AWnet can effectively alleviate the noise (even with strong level) and maintain the sharpness, yielding much high-quality HSTR frames. The noise induced by the lower SNR (with shorter exposure time), is greatly removed by our method (as illustrated in Fig. 11(d)-(e)), providing visually pleasant reconstruction with appealing spatial and temporal details. From these snapshots (and zoomed thumbnails), we can see that our system is robust and reliable to various exposure settings as well.

Motion blur. Camera acquisition with insufficient temporal resolution would introduce the motion blur as the zoomed regions of HSR-LFR frames in Fig. 12(a). Our work is trying to leverage the LSR-HFR video with accurate motion details to resolve it. Towards this goal, dedicated FlowNet and FusionNet are devised to extract and aggregate information (see Fig. 2(c)) for adaptive weighted synthesis. Optical flow extracted by the FlowNet is used to warp the $I_{\texttt{ref}}$ followed by the dynamic filter and mask generated by the FusionNet to weigh the respective information from $I_{\texttt{ref}}$ and $I_{\texttt{LSR}}$ appropriately.

As shown in Fig. 12, motion blurs are clearly observed in Fig. 12(a) around fast moving objects. Such effect is slightly reduced in Fig. 12(c) when flow information from LSR-HFR is utilized to warp the $I_{\texttt{ref}}$. Further improvement is achieved in Fig. 12(d) by utilizing the dynamic filter to restore the textures on blurred objects using information from the LSR-HFR input (see Fig. 12(b)(c)(d)). Sometimes, artifacts are induced due to over filtering, but subsequent adaptive mask in Fig. 12(e) will then intelligently combine pixels from $I_{\texttt{LSR}\uparrow}$ and $I^{wk}_{\texttt{ref}}$ for even better reconstruction shown in Fig. 12(f).

Resolution Gap. An interesting observation is that our model trained with image pairs (see Section 4) having $8\times$ resolution gap (noted as $8\times$-Model) provides much better reconstruction quality subjectively (Fig. 13(f)) compared to the model trained using image pairs with $4\times$ resolution gap (noted as $4\times$-Model) (Fig. 13(e)). For both models, noise level is set with $\sigma^2 = 0.01$ for regularization. For illustrative comparison, we downscale the $I_{\texttt{ref}}$ to its $\frac{1}{16}$th (e.g., $4\times$ downscaling at each spatial dimension) and $\frac{1}{64}$th (e.g., $8\times$ downscaling at each spatial dimension) sizes. Perceptually, a snapshot from the LSR-HFR camera in Fig. 13(d) is close to $8\times$ downscaled $I_{\texttt{ref}}$ in Fig. 13(c), but worse than $4\times$ downscaled $I_{\texttt{ref}}$ in Fig. 13(b). Thus, when we use $4\times$-Model, our network will evenly weigh information from $I_{\texttt{ref}}$ and $I_{\texttt{LSR}}$ yielding a smooth reconstruction with moderate quality in Fig. 13(e); but for $8\times$-Model, since the $8\times$ downscaled version in training is with pretty bad quality, more weights will be given to $I_{\texttt{ref}}$ in stationary areas and to $I_{\texttt{LSR}}$ in motion areas for adaptive fusion synthesis, leading to much sharp details in reconstruction as shown in Fig. 13(f). In other words, this is another example of adaptive weighting between the HSR-LFR and LSR-HFR camera inputs for final reconstruction quality improvement, where those weights can be regularized during training using sample pairs with different resolution gaps.

7 Conclusion

A dual camera system is developed in this work for high spatiotemporal resolution video acquisition where one camera captures the HSR-LFR video, and the other one records the LSR-HFR video. An end-to-end learning framework, AWnet, is then proposed to learn the spatial and temporal information from both camera inputs, and drive the final appealing reconstruction by intelligently synthesizing the content from either HSR-LFR or LSR-HFR frame. Towards this goal, separable FlowNet and FusionNet are devised in our framework, to explicitly exploit the information from two cameras so as to derive the adaptive weighting functions for reconstruction synthesis.

Our system demonstrates the superior performance, in comparison to the existing works, such as the state-of-the-art CrossNet, PM, EDSR, and ToFlow-SR for super-resolution, and ToFlow-Intp for frame interpolation, showing noticeable gains both subjectively and objectively, using simulation data and camera captured real data. ew We also analyze various aspects of our system by breaking down its modular components, such as upscaling filter, reference structure, camera parallax, exposure time, etc. These studies pave the way for the application of our model to different scenarios.

In general, our system belongs to a hybrid camera or multi-camera category, even though our current emphasis is the production of video at both high spatial resolution and high frame rate. But this approach can be easily extended to view synthesis since the different viewpoints can be also generalized using flow representation. Another interesting avenue is to extend current RefSR mechanism in AWnet to include more cameras (e.g., > 2)

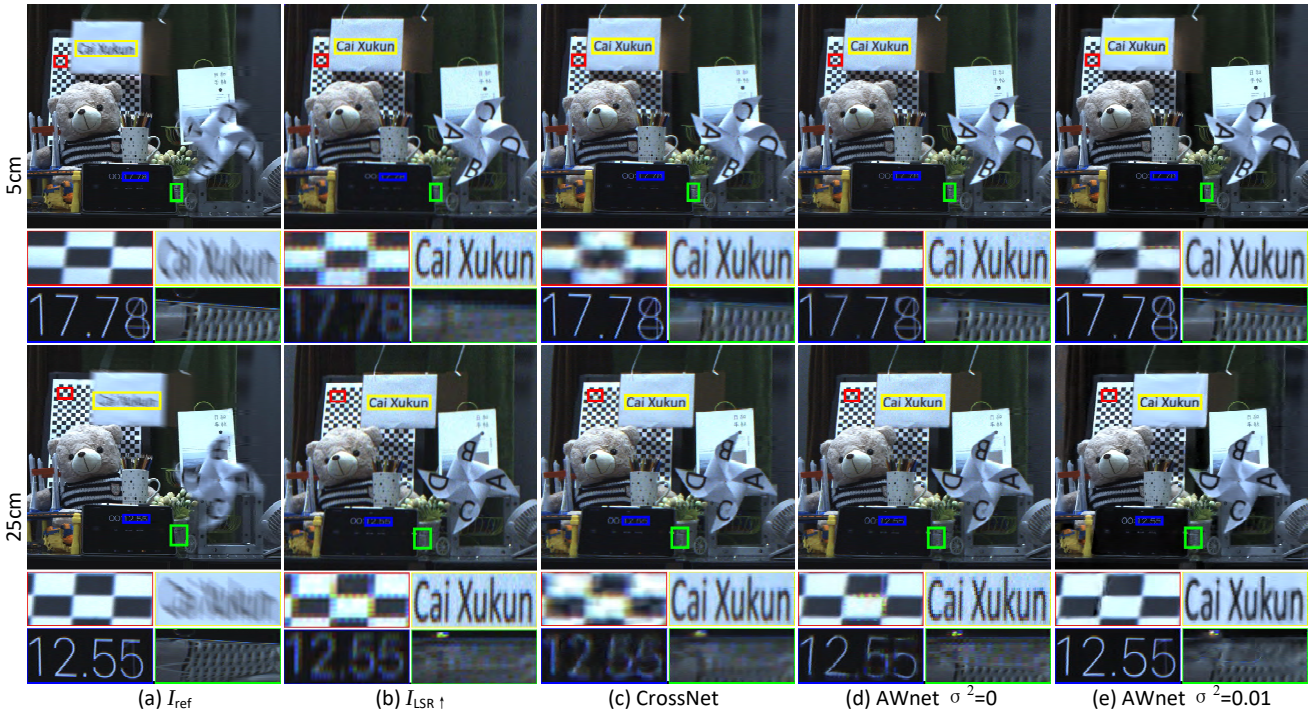


Fig. 10. **Camera Parallax.** Image reconstruction for our dual camera system when placing cameras with baseline distance at 5cm and 25cm. Our LSR-HFR camera operates at 240FPS. These images are captured using dual Grasshopper3 GS3-U3-51S5C cameras. The frames in the first column are the captured HSR-LFR frames. The frames in the second column are the captured LSR-HFR frames. Look at the repeated patterns on the checkerboard snapshots, there are some ghosts on the results of CrossNet because its multi-scale warping in feature domain, but our method does not have this problem. And our method has strong robustness when parallax is large. **Zoom in the pictures, you will see more details in the larger images.** More parallax settings in supplemental material.

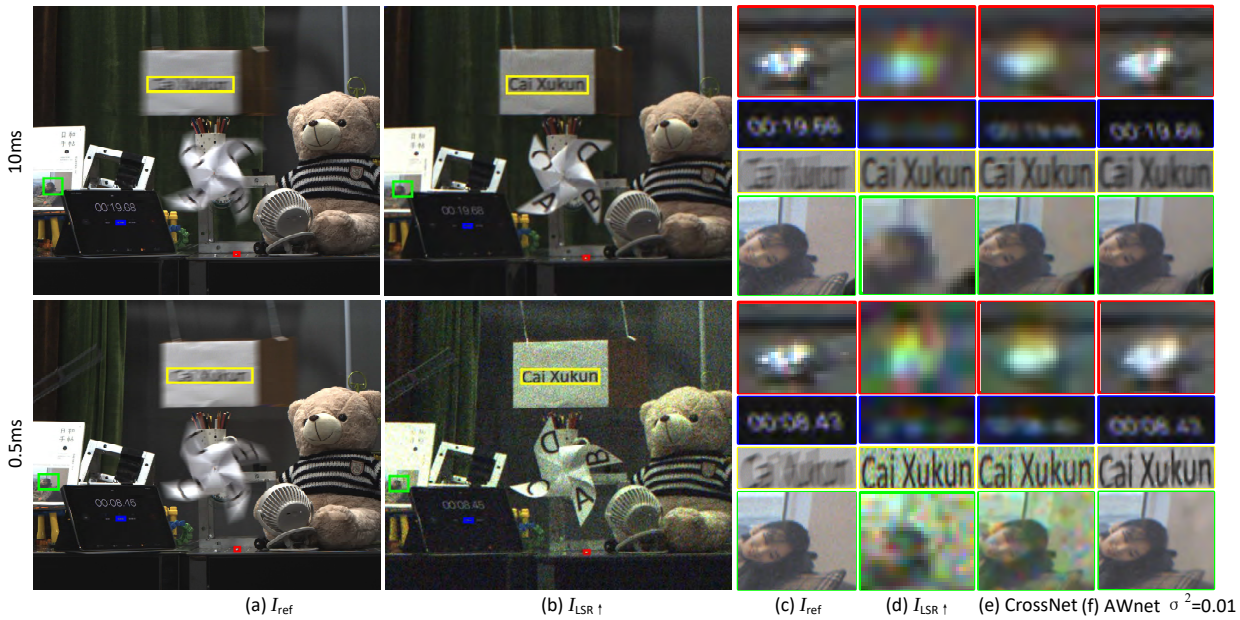


Fig. 11. **Exposure Time.** Various exposure time exemplified using our dual camera system. These images are captured using dual Grasshopper3 GS3-U3-51S5C cameras. The frames in the first column are the captured HSR-LFR frames using default exposure, the frames in the second column are the captured LSR-HFR frames with various exposure adjustments. Noise increases as exposure time decreases. CrossNet could remove some noise but generally lead to blurred artifacts induced by the global convolutions. Our AWnet can effective remove noise and improve the picture quality greatly. More exposure time settings in supplemental material.

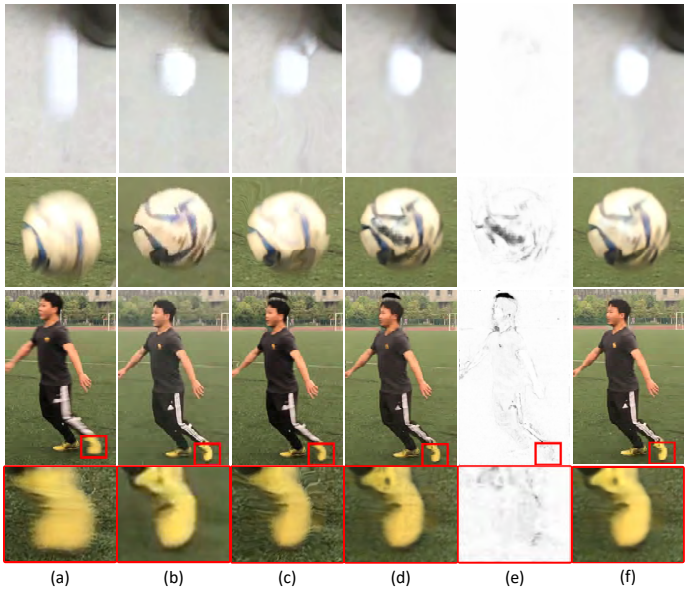


Fig. 12. **Motion blur in real data.** The images are captured with iPhone 7. (a) $I_{\texttt{HSR}}$; (b) $I_{\texttt{LSR}\uparrow}$; (c) $I^{w}_{\texttt{ref}}$; (d) $I^{wk}_{\texttt{ref}}$; (e) M; (f) Y with $\sigma^2 = 0.01$. The first row shows a fast ping-pong. The second row shows a fast soccer ball. The third and fourth rows are a fast-moving player and his enlarged shoe. **Zoom in to see more details.**

to enable the output video with more dimensional features, such as dynamic range (for low light imaging) [13], multi-spectra (beyond RGB), and depth (for 3D imaging) [18], etc.

Our AWnet could be further optimized towards the resource constrained embedded platform for broader applications using multi-camera equipped mobile phones, such as model compression [20], simple yet effective network structure [21], architecture-driven software optimization [30], etc.

REFERENCES

[1] The (new) stanford light field archive. http://http://lightfield.stanford.edu/lfs.html.

[2] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.

[3] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang. Depth-aware video frame interpolation. In *IEEE Conferene on Computer Vision and Pattern Recognition*, 2019.

[4] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *arXiv preprint arXiv:1810.08768*, 2018.

[5] W. Bao, X. Zhang, L. Chen, L. Ding, and Z. Gao. High-order model and dynamic filtering for frame rate up-conversion. *IEEE Transactions on Image Processing*, 27(8):3813–3826, 2018.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[7] V. Boominathan, K. Mitra, and A. Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2014.

[8] D. J. Brady, M. E. Gehm, R. A. Stack, D. L. Marks, D. S. Kittle, D. R. Golish, E. Vera, and S. D. Feller. Multiscale gigapixel photography. *Nature*, 486(7403):386, 2012.

[9] D. J. Brady, W. Pang, H. Li, Z. Ma, Y. Tao, and X. Cao. Parallel cameras. *Optica*, 5(2):127–137, Feb 2018.

[10] A. Burton and J. Radford. *Thinking in Perspective: Critical Essays in the Study of Thought Processes*. Psychology in progress. Methuen, 1978.

[11] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. 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.

[12] X. Cao, T. Yue, X. Lin, S. Lin, X. Yuan, Q. Dai, L. Carin, and D. J. Brady. Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world. *IEEE Signal Processing Magazine*, 33(5):95–108, 2016.

[13] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018.

[14] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[15] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, 36(6):1–15, 2017.

[16] G. Eilertsen, R. K. Mantiuk, and J. Unger. Single-frame regularization for temporally stable cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11176–11185, 2019.

[17] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, (2):56–65, 2002.

[18] T. Fujii. 3d image processing–from capture to display–. *Electronic Imaging*, 2019(3):625–1, 2019.

[19] J. Gu, H. Lu, W. Zuo, and C. Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1604–1613, 2019.

[20] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[21] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. *arXiv preprint arXiv:1905.02244*, 2019.

[22] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[23] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[24] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, pages 667–675, 2016.

[25] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018.

[26] M. Jin, Z. Hu, and P. Favaro. Learning to extract flawless slow motion from blurry videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8112–8121, 2019.

[27] Y. Jo, S. Wug Oh, J. Kang, and S. 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.

[28] J. Kim, J. Kwon Lee, and K. 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.

[29] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] N. D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, L. Jiao, L. Qendro, and F. Kawsar. Deepx: A software accelerator for low-power deep learning inference on mobile devices. In *Proceedings of the 15th International Conference on Information Processing in Sensor Networks*, page 23. IEEE Press, 2016.

[31] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[32] F. Li. *A HYBRID CAMERA SYSTEM FOR LOW-LIGHT IMAGING*. PhD thesis, University of Delaware, 2011.

[33] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017.

[34] S. Liu, L. Yuan, P. Tan, and J. Sun. Bundled camera paths for video stabilization. *ACM Transactions on Graphics (TOG)*, 32(4):78, 2013.

[35] S. Lu. High-speed video from asynchronous camera array. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages

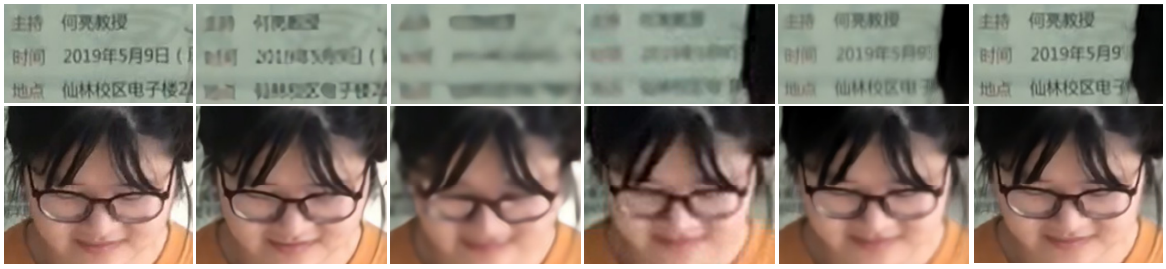


Fig. 13. **Resolution Gap Impact.** (a) $I_{\texttt{ref}}$, (b) $I_{\texttt{ref}}$ with $4\times$ resolution downscaling and upscaling to original size by EDSR, (c) $I_{\texttt{ref}}$ with $8\times$ resolution downscaling and upscaling to original size by EDSR, (d) $I_{\texttt{LSR}}$, (e) Y with $4\times$-Model, (f) Y with $8\times$-Model.

2196–2205. IEEE, 2019.

[36] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015.

[37] S. Niklaus and F. Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1710, 2018.

[38] H. Noh, T. You, J. Mun, and B. Han. Regularizing deep neural networks by noise: Its interpretation and optimization. In *Advances in Neural Information Processing Systems*, pages 5109–5118, 2017.

[39] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[40] M. S. Sajjadi, R. Vemulapalli, and M. Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6626–6634, 2018.

[41] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng. Learning to synthesize a 4d rgbd light field from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2243–2251, 2017.

[42] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[43] T.-C. Wang, J.-Y. Zhu, N. K. Kalantari, A. A. Efros, and R. Ramamoorthi. Light field video capture using a learning-based hybrid imaging system. *ACM Transactions on Graphics (TOG)*, 36(4):133, 2017.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[45] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu. Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954, 2017.

[46] Q. Xie, E. Hovy, M.-T. Luong, and Q. V. Le. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019.

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

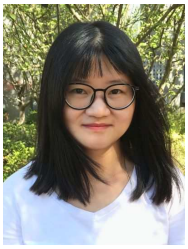
[48] K. Zhang, W. Zuo, and L. Zhang. Deep plug-and-play super-resolution for arbitrary blur kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1671–1681, 2019.

[49] H. Zheng, M. Guo, H. Wang, Y. Liu, and L. Fang. Combining exemplar-based approach and learning-based approach for light field super-resolution using a hybrid imaging system. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2481–2486, 2017.

[50] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang. Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution. In *BMVC*, 2017.

[51] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 88–104, 2018.

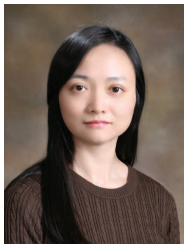
[52] S. Zheng, Y. Song, T. Leung, and I. Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4480–4488, 2016.



Ming Cheng is a Ph.D. candidate of Electrical Engineering with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. She received the B.S. degree in the University of Electronic Science and Technology of China, China, in 2016. Her research interests include computer vision, video processing and multi-cameras.

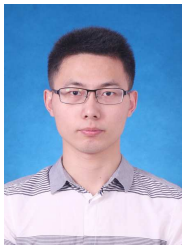
Zhan Ma (SM'19) is now on the faculty of Electronic Science and Engineering School, Nanjing University, Jiangsu, 210093, China. He received the B.S. and M.S. from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2004 and 2006 respectively, and the Ph.D. degree from the New York University, New York, in 2011. From 2011 to 2014, he has been with Samsung Research America, Dallas TX, and Futurewei Technologies, Inc., Santa Clara, CA, respectively. His current research focuses on the next-generation video coding, energy-efficient communication, gigapixel streaming and deep learning. He is a co-recipient of 2018 ACM SIGCOMM Student Research Competition Finalist, 2018 PCM Best Paper Finalist, and 2019 IEEE Broadcast Technology Society Best Paper Award.

M. Salman Asif is currently an Assistant Professor in the Department of Electrical and Computer Engineering at the University of California, Riverside. Prior to joining UC Riverside, he was a postdoctoral research associate in the DSP group at Rice University. Before that he briefly worked as a Research Engineer at Samsung Research America, Dallas. He received the Ph.D. at the Georgia Institute of Technology under the supervision of Justin Romberg. His research interests broadly lie in the areas of information processing and computational sensing with applications in signal processing, machine learning, and computational imaging.



Yiling Xu is a full researcher of School of Electronic Information and Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200145, China. She received the B.S., M.S. and Ph.D. from the University of Electronic Science and Technology of China, China, in 1999, 2001 and 2004 respectively. From 2004 to 2013, she was with Multimedia Communication Research Institute of Samsung Electronics Inc, Korea. Her main research interests include architecture design for next generation multimedia systems, dynamic data encapsulation, adaptive cross layer design, dynamic adaption for heterogenous networks and N-screen content presentation.

Haojie Liu is a Ph.D candidate in School of Electronic Science and Engineering, Nanjing University, Nanjing, China. He received the B.S degree in Nanjing University in 2016. His research interests includes video communication and processing, machine learning and computer vision.



Wenbo Bao is a Ph.D. candidate of Electrical Engineering with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. He received the B.S. degree in Electronic Information Engineering from Huazhong University of Science and Technology, Hubei, China, in 2014. His research interests include computer vision, machine learning, and video processing.

Jun Sun is currently a professor and Ph.D. advisor of Shanghai Jiao Tong University. He received his B.S. in 1989 from University of Electronic Sciences and technology of China, Chengdu, China, and a Ph.D. degree in 1995 from Shanghai Jiao Tong University, all in electrical engineering. In 1996, he was elected as the member of HDTV Technical Executive Experts Group (TEEG) of China. Since then, he has been acting as one of the main technical experts for the Chinese government in the field of digital television and multimedia communications. In the past five years, he has been responsible for several national projects in DTV and IPTV fields. He has published over 50 technical papers in the area of digital television and multimedia communications and received 2nd Prize of National Sci. & Tech. Development Award in 2003, 2008. His research interests include digital television, image communication, and video encoding.