IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 28, 2019, accepted August 12, 2019, date of publication August 21, 2019, date of current version September 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936549

FI-Net: A Lightweight Video Frame Interpolation Network Using Feature-Level Flow

HAOPENG LI, YUAN YUAN, (Senior Member, IEEE), AND QI WANG, (Senior Member, IEEE)
School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Yuan Yuan (y.yuan1.ieee@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant U1864204 and Grant 61773316, in part by the State Key Program of National Natural Science Foundation of China under Grant 61632018, in part by the Project of Special Zone for National Defense Science and Technology Innovation, and in part by the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University.

ABSTRACT Video frame interpolation is a classic computer vision task that aims to generate in-between frames given two consecutive frames. In this paper, a flow-based interpolation method (FI-Net) is proposed. FI-Net is a lightweight end-to-end neural network that takes two frames in arbitrary size as input and outputs the estimated intermediate frame. Novelly, it computes optical flow at feature level instead of image level. Such practice can increase the accuracy of estimated flow. Multi-scale technique is utilized to handle large motions. For training, a comprehensive loss function that contains a novel content loss (Sobolev loss) and a semantic loss is introduced. It forces the generated frame to be close to the ground truth one at both pixel level and semantic level. We compare FI-Net with previous methods and it achieves higher performance with less time consumption and much smaller model size.

INDEX TERMS Video frame interpolation, lightweight network, feature-level flow, Sobolev loss.

I. INTRODUCTION

Video frame interpolation technique synthesizes intermediate frames between any two consecutive frames, which is a classic computer vision task in image and video processing [1]–[6]. It generates smooth transitions from the former frame to the latter one, compensating the motion information and enriching the changing details. The generated frames, together with the original ones, form both spatially and temporally coherent video sequences.

Video frame interpolation is widely applied to various fields, such as video frame rate conversion between broadcast standards, temporal super-resolution of videos for better visual effect, virtual view synthesis, etc. Different broadcast standards may request videos of different frame rates. So only when its frame rate is up-conversed, can the video be broadcasted with a higher standards. Besides, due to video temporal compression or the limitation of camera hardware, videos could suffer from low frame rate, leading to poor visual effect. Increasing the rate by frame interpolation can substantially improve the video details and ease the above problem. Moreover, frame interpolation can synthesize virtual views of objects from two given adjacent views, enriching the depiction of objects. In virtue of its wide applications, video frame interpolation has been drawing increasing attention to many researchers.

Three reasons account for the fact that video frame interpolation is a challenging task. 1) Cameras capture various complicated scenes in the real world. Generating images of such scenes is considerably difficult even though two adjacent frames are given. 2) Large motions are commonly existed in videos. In that case, the two original frames have big differences in motion regions, greatly increasing the complexity of frame interpolation. 3) The objects in videos could easily be occluded, which demands strong robustness of interpolation algorithms.

Learning-based video frame interpolation methods achieve great performance in recent year due to the development of machine leaning techniques. General procedure of learning-based video frame interpolation is shown in FIGURE 1. Given two consecutive frames in time order, the intermediate frame is firstly estimated by a certain frame interpolation algorithm. Then, the difference between the synthesized frame and the real one is computed under a certain criterion. To minimize the difference, the interpolation algorithm is improved repeatedly. Noting that the core of video frame

The associate editor coordinating the review of this article and approving it for publication was Zahid Akhtar.

interpolation is the interpolation method and the distance metric between two images.

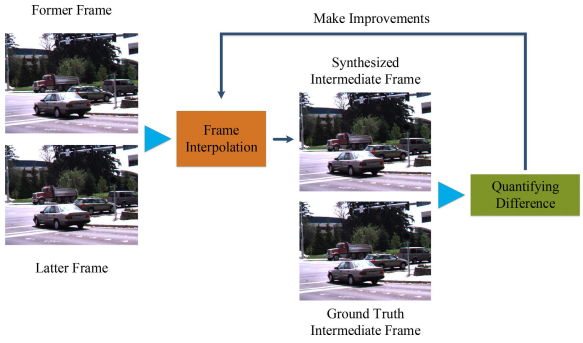


FIGURE 1. Learning-based video frame interpolation. The intermediate frame is estimated by a frame interpolation algorithm. Then the difference between the estimation and the ground truth is measured. Based on the interpolation error, the algorithm is improved or adjusted to generate images that is more similar to the real ones.

Standard approaches to synthesize interpolated frames in a video sequence involves two steps: motion estimation and pixel hallucination. Typical methods to model motions require accurate pixel correspondences between the two frames (e.g. using optical flow) [2]. The motions are considered as pixel spatial displacement. The quality of interpolated results is directly related to the accuracy of optical flow. That is to say, theoretically, perfect intermediate frames can be generated if the computed optical flow is flawless. However, flow-based methods have three drawbacks: 1) The estimated optical flow is not accurate when the motion regions suffer from blur, occlusion and abrupt brightness change. 2) The computation of global optical flow requires considerable time consumption and memory usage, which increases the difficulty of frame interpolation. 3) Some algorithms calculating optical flow are sensitive to parameter settings [7]. The effectiveness of those methods is not guaranteed under certain circumstances.

Recently, a novel thought to model motions has been brought up. It considers the motions as the change of color per pixel though time. Those methods are based on complex steerable pyramid and phase shift [2], [8]. While achieving favorable performance, they are limited by three aspects: 1) Large motions cannot be well dealt with by them. The blur effect is inevitable when the two original frames are far part. 2) Because the phase shift of each pixel is required, the computational complexity cannot be ignored. 3) Experiments prove that they cannot reach the performance of flow-based methods.

Deep convolutional neural networks reveal great power in computer vision tasks such as image recognition [9]–[11], video understanding [12]–[15], pixel-wise image analyze [16], [17], etc. There is a big trend to use deep convolutional neural networks to synthesize intermediate frames. In general, there exist three ways to use DCNN for video frame interpolation. 1) DCNN is utilized only to compute global optical flow of high accuracy [18]–[20]. Then the given frames are warped to generate interpolated frames. These methods are better than those that use traditional algorithms to compute optical flow in most cases. 2) DCNN includes two steps, the computation of optical flow or phase shift, and frame synthesis [1], [4], [8], [21]. This kind of methods can synthesize better results than the first one does, because it models video frame interpolation as a whole task and optimizes all parameters simultaneously. 3) Instead of modeling motions in traditional ways, these methods use convolution operations to represent the motion of each pixel and directly synthesize frames [3], [22]. Though experiments show the effectiveness of these methods, their practicality is relatively low, because they need to compute the convolution kernel for each pixel.

Briefly speaking, previous works on video frame interpolation have limitations as follows. 1) The flow-based methods can generate incorrect frames due to inaccurate estimated optical flow. 2) The phase-based methods lead to blur effect when dealing with large motions. 3) The convolution-based methods require a great many computing resources, making them less practical when processing high resolution videos. 4) The DCNNs are commonly optimized by pixel-wise loss function (MSE loss or l_1 loss), generating images with high peak signal-to-noise ratio (PSNR) but low visual appearance.

Out of the deficiencies of previous video frame interpolation methods, in this paper we propose a novel neural network that uses robust feature-level optical flow for video frame interpolation, i.e., **FI-Net**. FI-Net is a lightweight end-to-end trainable neural network, given two consecutive frames in arbitrary size and directly outputting the estimated intermediate frame. It follows the typical procedure of video frame interpolation. First, the feature-level flow estimator computes the bidirectional optical flow between the two given frames. Then, the two frames are warped forward and backward through time to generate two estimated intermediate frames. The final output is obtained by three dimension (3D) convolutions [23] on the concatenation of the forward result and backward one. The training of FI-Net is end-to-end and directed by a comprehensive loss function that contains a new content loss and a semantic loss. The **main contributions** of this paper are condensed as follows:

- We propose a lightweight end-to-end trainable neural network (FI-Net) for video frame interpolation. Compared to previous methods, it achieves higher performance using less inference time and much smaller model size.
- We design a feature-level flow estimator to obtain robust optical flow for frame warping. The learnable estimator computes optical flow based on the original frames and their feature maps. Such practice increases the accuracy of interpolated results.
- A comprehensive loss that includes semantic loss and Sobolev loss is presented, where the semantic meanings and gradients of images are involved for optimization. It recovers high-level information and high frequency components of images.

II. RELATED WORK

As a classic and challenging computer vision task, video frame interpolation has become one of the most popular research areas. A great number of methods has been proposed.

Optical flow is the most common approach of video frame interpolation. The flow-based methods are generally named as Lagrangian methods. The accuracy of estimated optical flow would directly affect the quality of synthesized frame. In recent years, many algorithms to compute global optical flow come up and prove to be effective [18]–[20], [24]–[26]. However, directly using optical flow for video frame interpolation can cause blur and distorted results. There is a trend of merging optical flow computing into the frame interpolation procedure [1], [4], [21]. Liu *et al.* [1] combine optical flow and neural network to synthesize video frames by flowing pixel values from input frames rather than hallucinating them from scratch, which can be easily extended to multi-frame generation. Jiang *et al.* [4] also propose a frame interpolation method which is capable of synthesizing variable-length of intermediate frames. To overcome the artifacts of optical flow, a U-Net is employed to refine the optical flow in [4]. However, because the CNN to compute optical flow is self-supervised, the accuracy of the flow is relatively low. van Amersfoort *et al.* [21] design a generative network for frame interpolation, in which the prediction of optical flow and the synthesis of frames are constructed in a coarse-to-fine fashion. Owing to the sophisticated structure of the flow-based generator, the method in [21] achieves impressive performance.

Besides the Lagrangian methods, Eulerian methods are exploited to solve video frame interpolation problem. Eulerian methods are original used to enhance subtle motions and Meyer *et al.* [2], [8] extend the idea to model motions between frames. Meyer *et al.* model motions as phase shift of individual pixels and propose a bounded phase shift correction method to deal with large motions in [2]. Compared to flow-based methods, the proposed one is simple to implement and easy to parallelize, the computational cost of which is lower. However, it fails when there are significant appearance changes. To handel challenging scenarios and large motions, Meyer *et al.* extend the phase-based method by deep learning in [8]. This paper proposes PhaseNet that directly estimates the phase decomposition of the interpolated frame. Though Eulerian methods are promising for their lower computational cost and easy implementation, the performance can not reach the level of the methods that explicitly match and warp pixels [2].

Convolution-based methods have been drawing much attention to many researchers recently [3], [22], [27]. Niklaus *et al.* [3] propose a video frame interpolation method (AdaConv) which combines motion estimation and pixel synthesize into a single process. They model pixels synthesis as local convolution over two input frames. However, the method in [3] cannot handle any large motion beyond 41 pixels unless the frames are down-scaled. Besides, the memory demand increases quadratically with the kernel size. Niklaus *et al.* [22] improve AdaConv by formulating frame interpolation as local separable convolution over two input frames, which decreases the parameters of the model and makes it faster. But the method still cannot handle any large motion beyond 51 pixels. Chen *et al.* [27] propose a deep convolutional network to generate intermediate frames between two non-consecutive frames. The network can learn scene transformation through time and generate longer video sequences. The synthesized frames have randomness yet.

III. THE PROPOSED METHOD

In this section, we elaborate our method of video frame interpolation. First, we describe the task in mathematic and define a few notations. Then, we propose the distinctive FI-Net and explain its structure. At last, a comprehensive loss function is introduced to train FI-Net.

A. PROBLEM DEFINITION

Video frame interpolation aims to generate smooth transitions between two consecutive frames in a video. Here, we use $I^1, I^2 \in \mathbf{R}^{H\times W\times C}$ to represent the two given time-ordered frames. H, W, C are the hight, width and channel number of the frame, respectively. I^R is the ground truth intermediate frame between I^1 and I^2, which is only available during training. The synthesized intermediate frame is denoted as I^S with the same size of I^R.

Our ultimate goal is to find a map G to synthesize I^S which is the estimation of I^R for the given I^1, I^2, namely,

$$G(I^1, I^2) = I^S \to I^R. \tag{1}$$

To achieve this, we train a neural network $G_{\boldsymbol{\theta}}$ with the parameters $\boldsymbol{\theta}$. $\boldsymbol{\theta}$ is obtained by minimizing a certain loss function L. For N training samples $\{(I_i^1, I_i^2), I_i^R\}_{i=1}^N$, we solve the following optimization problem,

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} L(G_{\boldsymbol{\theta}}(I_i^1, I_i^2), I_i^R). \tag{2}$$

In this work, we specifically design a neural network and a loss function for video frame interpolation, the details of which are described next.

B. FI-NET FOR VIDEO FRAME INTERPOLATION

We propose FI-Net for video frame interpolation. It takes two given frames as input and generates the intermediate frame directly. FI-Net is a flow-based method and follows typical procedure of frame interpolation: motion estimation and frame warping [28]. Overview of FI-Net is shown in FIGURE 2.

To handle large motions and obtain stable performance, we adopt the multi-scale estimation technique. For each scale, we estimate the intermediate frame in FI-Net. The final result

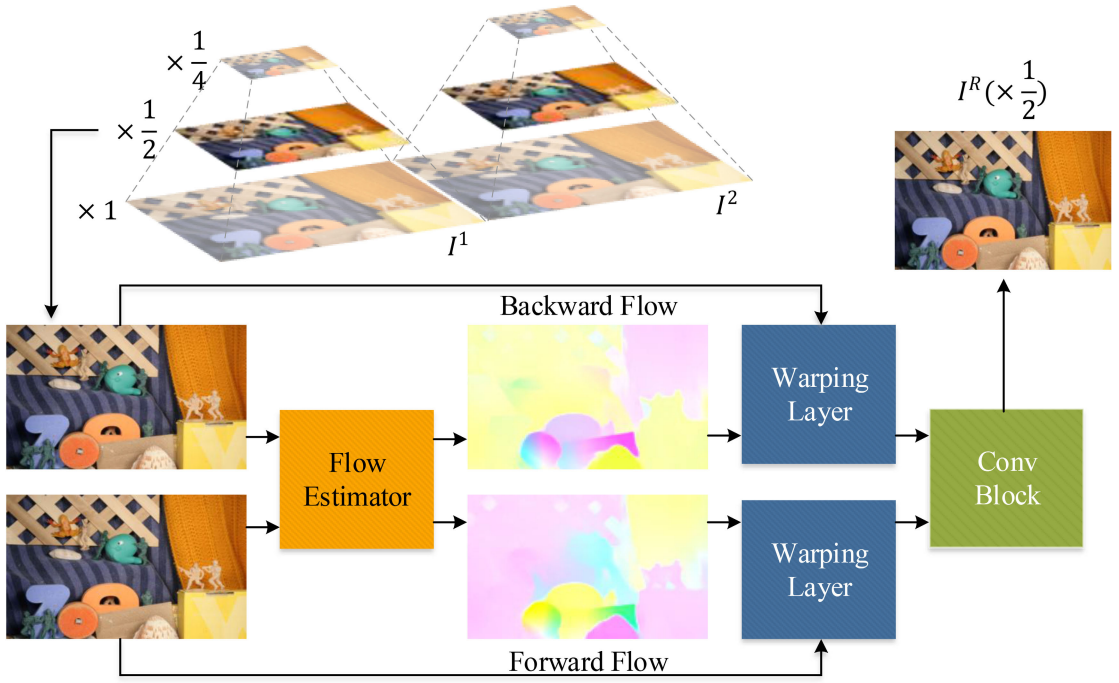


FIGURE 2. Overview of FI-Net. The estimated interpolated frame is computed in three scales ($\times 1$, $\times\frac{1}{2}$ and $\times\frac{1}{4}$). In each scale, the flow estimator outputs the forward and backward optical flow according to the frame order. Then the two given frames are warped based on the flow and two estimated transitions are obtained. The interpolated result in this scale is computed by convolutions. The final result is the weighted average of generated frames in all scales (firstly up-sampled to the same size of the original frame). Note that we only demonstrate how an intermediate frame is generated in $\times\frac{1}{2}$ scale, which means the process is the same in different scales.

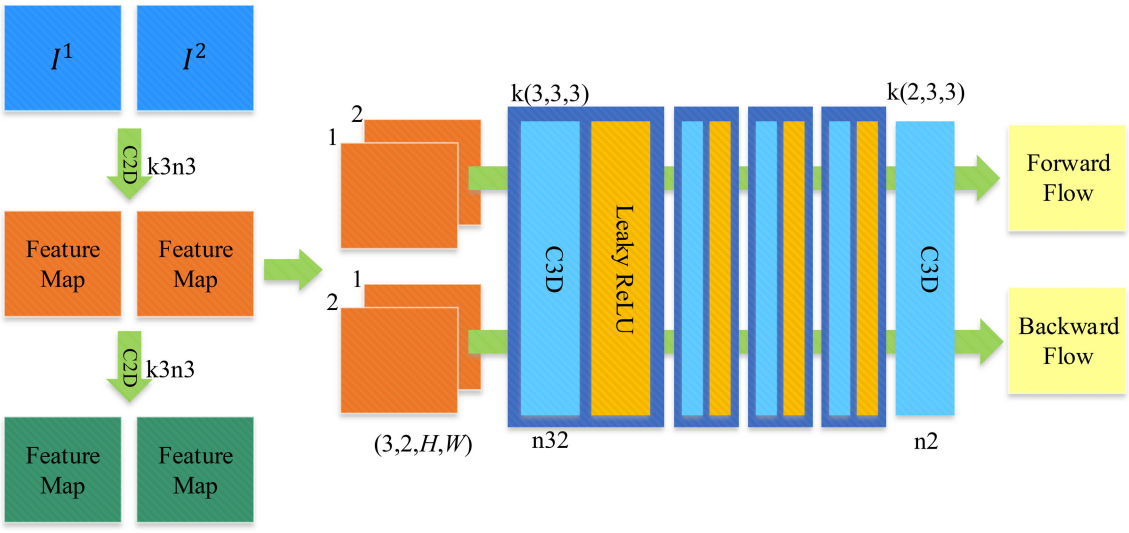


FIGURE 3. The structure of the flow estimator in FI-Net. We compute optical flow at feature level. Three levels of feature map are used to estimate the flow. The estimation is implemented by 3D convolution and Leaky ReLU activation.

is the weighted average of all scales. In consideration of model size, the network parameters are shared in all scales.

The core of FI-Net is the proposed feature-level flow estimator. The concrete structure of the flow estimator is illustrated in FIGURE 3. Different from other methods of optical flow which use CNN to extract different levels of feature maps and only use the deepest one to estimate optical flow, we construct different levels of feature maps and process them with the same networks but independently to obtain final optical flow. It computes the optical flow based on the two given frames and their feature maps. Given two frames I^1 and I^2, two convolutional layers (with kernel size 3 and channel number 3) are applied to extract features from them. So we get three levels of feature map

TABLE 1. The final convolutional block used in FI-Net.

Layer	Convolution Kernel	Padding	Activation
1	$3 \times 16 \times 3 \times 3 \times 3$	$1 \times 1 \times 1$	PReLU
2	$16 \times 16 \times 3 \times 3 \times 3$	$1 \times 1 \times 1$	PReLU
3	$16 \times 3 \times 2 \times 1 \times 1$	$0 \times 0 \times 0$	Identity

(the original frames are zeroth level). For each level, we use a 3D convolutional neural network to estimate the bidirectional optical flow. The details of the network are shown in FIGURE 3. Specifically, each 3D convolutional neural network resembles other convolutional networks in previous works. It is stacked with 4 blocks of "Conv-Activation" and one extra convolutional layer. The structure of the first "Conv-Activation" block is detailed in the figure: the number of filter is 32, the kernel size is $3 \times 3 \times 3$, and Leaky ReLU is used as activation. The following three blocks have the same structure as the first one does. The structure of the last convolutional layer is also shown in the figure: the number of filter is 2 and kernel size is $2 \times 3 \times 3$ (without activation). Thus, we obtain three maps of forward flow and three maps of backward one from three levels of features. In addition, the 3D convolutional networks at different feature levels do not share the parameters so as to increase the fitting ability of FI-Net. The final bidirectional flow is the weighted average of the corresponding flow.

We novelly compute optical flow at three levels for two motivations: 1) Optical flow aims to achieve global matches between pixels in two images. However, the matching process struggles to exploit the inner characteristic of pixels because the descriptor of pixels are fixed. However, in our work, a 3-level feature pyramid is constructed by convolution. Note that we use learnable convolutional layers to extract feature maps from images. So the feature maps we use to compute optical flow are adaptively changing during training, which means our model is capable of discovering better optima in extended solution space. Thus, such practice guarantees our method generate accurate intermediate frames. 2) The feature-level technique has the same functions as those of adaptive data augmentation (not in the strict sense): two more informative images (feature maps) are generated for the estimation of optical flow by trainable convolution layers (act like adding noise, smoothing, sharpening, etc). By integrating three results from three inputs, the robustness of optical flow estimator is improved.

After obtaining the bidirectional flow, we backward warp the original frames and get two estimated intermediate frames. Then we concatenate these two frame along the time dimension. The concatenated frames is processed by a 3D convolutional block (the Conv Block in FIGURE 2) whose structure is shown in TABLE 1. To increase the fitting ability of our model, parametric rectified linear unit (PReLU) [29] is added after each convolutional layer. Finally, we obtain the interpolated frame in this scale.

TABLE 2. The results of objective comparison. The best and the second best results are bold and underlined respectively.

Method	PSNR (dB)	SSIM	Runtime (ms)	Model Size (MB)
FlowNet2 [18]	29.774	0.874	**33.644**	444.68
PhaseS [2]	32.718	0.882	<u>75.048</u>	——
SepConv [22]	<u>33.304</u>	0.883	152.770	<u>86.73</u>
Super SloMo [4]	33.293	0.887	193.347	151.21
ContextFI [42]	33.283	<u>0.890</u>	188.452	108.95
FI-Net	**33.339**	**0.893**	123.221	**10.33**

C. LOSS FUNCTION

A comprehensive loss function is defined to training FI-Net. The loss is denoted as L_G that is a linear combination of content loss L_C and semantic loss L_S. That is to say,

$$L_G = L_C + \lambda_S L_S, \tag{3}$$

where λ_S is the weight to balance the two components. We elaborate each loss function as follows.

1) CONTENT LOSS

Content loss measures the distance between the generated frame and the ground truth one at pixel-wise level. Pixel-wise content loss such as MSE (l_2) and l_1 loss struggles to deal with the uncertainty nature of losing high-frequency details such as texture [30]. Experiments demonstrates that minimizing MSE tends to finding pixel-wise averages of plausible solutions that are often overly smooth, causing poor perceptual quality [31]–[33]. In consideration of the drawbacks of pixel-wise loss functions, gradient-based loss functions are exploited in image synthesis [3], [34]. Mathieu *et al.* prove that the blur effect can be alleviated by introducing image gradients into the loss functions [34]. The high-frequency compounds in images can be recovered and enhanced.

To deal with the problem of pixel-wise loss, we introduce a novel gradient-based content loss function named Sobolev loss which is based on Sobolev space [35]. Sobolev space is a subspace of L_2 function space. The functions in Sobolev space are required that their partial derivatives are also in L_2 space. Rigorously, Sobolev space is defined as

$$H(\Omega) = \left\{ u \in L_2(\Omega) \middle| \frac{\partial u}{\partial x_k} \in L_2(\Omega), k = 1, 2, \cdots, d \right\}, \tag{4}$$

where $\Omega \subset \mathbf{R}^d$ is the domain of functions. In addition, Sobolev space is equipped with a norm that is a combination of L_2-norms of the function itself and its derivatives, namely,

$$\|u\|_{H(\Omega)} = \left(\|u\|_{L_2(\Omega)} + \sum_{k=1}^{d} \left\| \frac{\partial u}{\partial x_k} \right\|_{L_2(\Omega)} \right)^{\frac{1}{2}} \tag{5}$$

$$= \left[\int_\Omega u^2 \mathrm{d}V + \sum_{k=1}^{d} \int_\Omega \left(\frac{\partial u}{\partial x_k} \right)^2 \mathrm{d}V \right]^{\frac{1}{2}} \tag{6}$$

where $\| \cdot \|_{L_2(\Omega)}$ denotes L_2-norm of functions.

In this work, we propose a novel content loss named Sobolev loss based on Sobolev space, denoted as follows,

$$Sob(I^S, I^R) = \|I^D\|_{\hat{l}_1} + \lambda_D\left(\left\|\frac{\partial I^D}{\partial x}\right\|_{l_2}^2 + \left\|\frac{\partial I^D}{\partial y}\right\|_{l_2}^2\right), \quad (7)$$

where $I^D = I^S - I^R$, $\frac{\partial I^D}{\partial x}$ and $\frac{\partial I^D}{\partial y}$ denote the gradient images of I^D along height dimension and along width dimension, respectively. $\|I\|_{l_2} = \frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W} I_{xy}^2$ is l_2-norm of images, and $\|I\|_{\hat{l}_1} = \frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W}\sqrt{I_{xy}^2 + \varepsilon^2}$ is Charbonnier norm (ε is positive and close to 0) [36]. λ_D is the weight of gradient terms. We use Charbonnier loss on the original images because it is less sensitive to outlier and has good global smoothness. We use MSE loss for the gradients to force the generated images to maintain high-frequency components such as edges and texture.

2) SEMANTIC LOSS

The generated frame has the similar semantic meanings to the ground truth one. Motivated by this idea, we propose the semantic loss that quantifies the difference between two images at semantic level. We define semantic loss as follows,

$$L_S(I^S, I^R) = \left\|\Phi(I^S) - \Phi(I^R)\right\|_{l_2}^2, \quad (8)$$

where $\Phi(I)$ is the semantic map of image I. We use MSE loss to force the generated image to has almost the same semantic meanings as the target does.

In this paper, we utilize a pre-trained convolutional neural network as the semantic map. Specifically, we pre-train Resnet50 on ImageNet and use the responses of the last average pooling layer as the semantic map $\Phi(\cdot)$.

IV. EXPERIMENTS

In this section, we conduct a series of experiments to prove the effectiveness of our method. First, we explain the experiment setups, including the datasets, implementation details and evaluation metrics. Then, we conduct ablation study to demonstrate the contribution of our method. We compare our method with several previous video frame interpolation methods to prove its superiority. At last, we further compare the optical flow compute by FI-Net and that by FlowNet2 to show the significance of our feature-level flow estimator.

A. EXPERIMENT SETUPS

1) DATASETS

The training of FI-Net requires no annotated sample like other computer vision tasks such as image classification [37], action recognition [38], etc. So we can make use of any video as the training data for FI-Net. We use the first episode of Planet Earth II (1080p, 25fps) that is available online to train our model. This video covers complicated real scenes in nature, including people, various of animals and plants, close-up shots and long shots. Besides, large motions as well as tiny movements are common in the video. Thus, this video is quite applicable to training FI-Net for better generalization.

To generate training samples, all the frames are grouped into triplets, each of which contains three consecutive frames. To eliminate the shot boundaries, the l_2 histogram distance between the first and third frames is computed in each triplet. Triplets with distances bigger than 0.95 upper-quantile are considered as shot boundaries and removed from the training set. After eliminating the boundaries, 83,479 triplets are remained. Finally, all the remaining 1920×1080 triplets are resized to the size of 320×256.

As for testing, we use videos from UCF-101 [38]. We choose 10 videos (320×240, 25fps) and group the frames into 1,715 triplets to construct the testing set. Note that the UCF-101 testing set in our paper is different from that in [1] and [4] which only contains 379 samples, and so our experiment results are more general. Besides, the training set and testing set are different in content, motion scale, etc. So the experiments require high generalization ability of models.

2) IMPLEMENTATION DETAILS

We empirically set $\lambda_S = 10^{-6}$ and $\lambda_D = \varepsilon = 10^{-3}$. We train FI-Net for 10 epochs by Adam algorithm [39] with initial learning rate 10^{-4} and $\boldsymbol{\beta} = (0.9, 0.999)$. The learning rate is halved every 2 epochs. The batch size is 16, which is enough for stable convergence. Our method is implemented by PyTorch on NVIDIA GeForce GTX 1080 Ti.

3) EVALUATION METRICS

Two widely-used metrics are applied to objective evaluation: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [40]. We compute PSNR between two RGB images by averaging the PSNR of three channels. As for SSIM, we convert the RGB images into YUV color space and compute the SSIM of the Y component.

B. ABLATION STUDY

We conduct ablation study to prove the effectiveness of our model and the proposed loss functions, including Sobolev loss and Semantic loss. A baseline model, simplification of FI-Net, is constructed for comparison: 1) We eliminate multi-scale technique and only maintain the original scale. 2) We abandon feature-level flow estimation and compute optical flow directly from original images. To eliminate the influence of model complexity, we compute the flow using three branches of 3D convolutional network in FIGURE 2

TABLE 3. The results of ablation study.

Model	PSNR (dB)	SSIM
Baseline	31.291	0.882
Baseline+Multi-scale	32.075	0.888
Baseline+Feature-level Flow	32.728	0.887
Baseline+Sobolev Loss	32.115	0.886
Baseline+Semantic Loss	32.015	0.890
FI-Net	33.339	0.893

TABLE 4. The results of subjective comparison.

Method	Mequon	Schefflera	Urban	Teddy	Backyard	Basketball	Dumptruck	Evergreen	Average
SepConv [22]	0.254	0.233	0.260	0.228	0.246	0.240	0.214	0.253	0.241
Super SloMo [4]	0.285	0.261	0.240	0.264	0.275	0.264	0.254	0.246	0.261
ContextFI [42]	0.228	0.261	0.274	0.259	0.229	0.279	0.269	0.254	0.256
FI-Net	0.234	0.246	0.226	0.250	0.250	0.218	0.264	0.248	0.242

(the three branches do not share parameters). 3) To demonstrate the effectiveness of our loss function, we use pure MSE loss to train the baseline model instead of the proposed loss. The results of ablation study are shown in TABLE 3.

The baseline model achieves considerable PSNR and SSIM. It proves the effectiveness of our basic structure. The multi-scale technique, feature-level flow, Sobolev Loss and semantic loss improve PSNR to varying degrees. The statistics show that feature-level flow has the strongest effect on PSNR, and hence feature-level flow is the core of our model. Besides, those four components also increase SSIM. Semantic loss has the greatest impact on SSIM, which means high-level similarity is crucial to visual appearance. Moreover, we find that the semantic loss can accelerate the optimization process. We believe the reason is that the network is oriented to a more abstract objective that is mathematically better than pixel-wise objectives.

C. OBJECTIVE COMPARISON

We compare our method with several previous video frame interpolation methods, including FlowNet2 (FlowNet2-CSS-ft-sd specifically) [18], PhaseS [2], SepConv [22], Super SloMo [4] and ContextFI [41]. These five methods are representative of various video frame interpolation techniques. FlowNet2 is a deep neural network to purely compute global optical flow. We use the optical flow computed by it to warp the input frames for frame interpolation. PhaseS novelly considers motions as the phase shifts of pixels, which is simple to implement and parallelize. SepConv directly synthesizes pixel values by local convolution without modelling the motions in physical ways. Super SloMo and ContextFI are two state-of-the-art flow-based methods that model motions as pixel spatial displacements. We use official implementations and pre-trained models for comparison. Hence, the comparison results are persuasive. The results of PSNR and SSIM, as well as average runtime and model size, are shown in TABLE 4.

FI-Net outperforms FlowNet2 to a large extent on PSNR and SSIM. It proves that directly using optical flow for video frame interpolation can not achieve considerable performance. Besides, FI-Net also achieves higher PSNR and SSIM than PhaseS does. PhaseS is a phase-based interpolation method that proves to be less effective than flow-based methods in other works. Our experiment also reveals this phenomenon. Compared to SepConv, Super SloMo and ContextFI, FI-Net achieves higher PSNR and SSIM but consumes less inference time with much smaller model size (approximately one tenth of ContextFI), which is applicable in practice. The comparisons prove the superiority of our method.

D. SUBJECTIVE COMPARISON

We follow Niklaus *et al.* in [22], [41] and conduct user studies to compare our method with previous ones. Eight examples (680×480) from the Middlebury benchmark [42] are used as the test samples for the user studies. 15 participants (graduate students or PhDs in computer science) are recruited and involved into our studies. For each example in Middlebury benchmark, they are given the synthesized images by different methods and asked to select the image with better visual appearance. The preference index of an image is calculated as the ratio of selections over views.

TABLE 4 demonstrates the preference indexes of generated images in each example. As shown in TABLE 4, a method performs differently in different examples. The expected average preference of all methods is 0.25, provided that their performance are the same. The results reveal that the four compared methods perform similarly, which means that the generated results from different methods have comparable visual appearance. Although our method achieves the best performance in the objective comparison, it does not perform the best in the subjective comparison. We believe it is because the objective criteria do not totally agree with the visual cognition of human. More work will be carried out to improve the visual appearance of generated images by our method.

FIGURE 4 shows some examples from the Middlebury benchmark, including the ground truth and images synthesized from different methods. As shown in FIGURE 4, video frame interpolation methods still suffer from blur and artifacts due to large and complicated motions, which are obvious in "Backyard" and "Basketball". Those phenomena reduce the realness of synthesized images to a large extent.

E. OPTICAL FLOW COMPARISON

As the ablation study shows, the core of FI-Net is the feature-level flow estimator. To further demonstrate the significance of our flow estimator, we compare the optical flow computed from it and that from FlowNet2-CSS-ft-sd (abbreviated as FlowNet2) [18]. The MPI Sintel dataset (clean and final) [43] and Middlebury dataset [42] are used for test. The average endpoint errors are shown in TABLE 5.

As the results show, the feature-level flow estimator achieves almost the same accuracy as FlowNet2 does. Hence, our method is also capable of computing accurate global

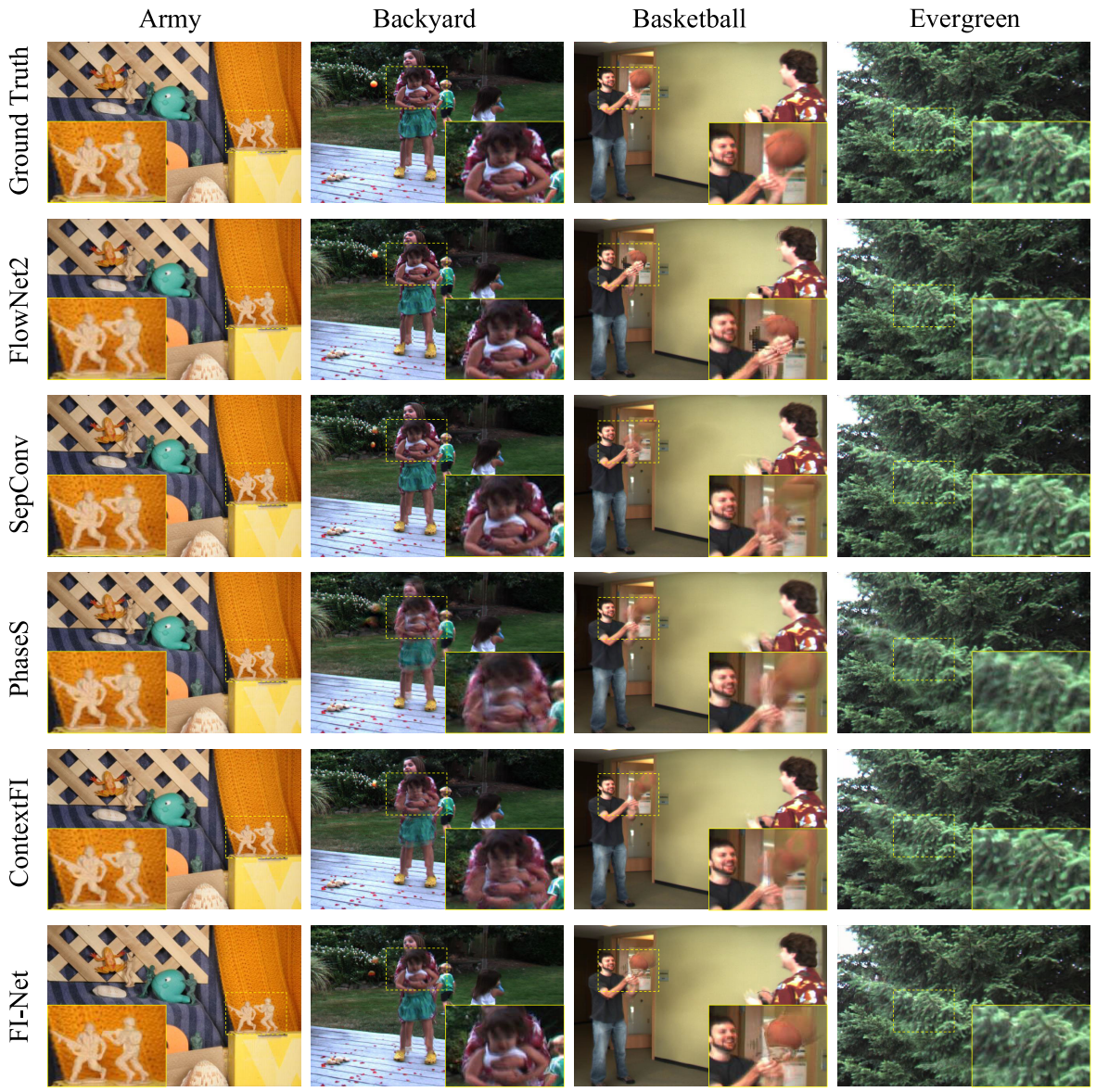


FIGURE 4. Generated images from the Middlebury benchmark by different methods as well as the ground truth images.

TABLE 5. Comparison of flow accuracy (average endpoint error).

Method	Sintel Clean	Sintel Final	Middlebury
FlowNet2	2.08	3.17	0.38
FI-Net	1.98	3.28	0.39

optical flow between two images. Compare with FlowNet2, our method owns two advantages: 1) FI-Net has much smaller model size than FlowNet2 does, which is more practical. 2) The training of FlowNet2 requires a large mount of samples labeled with ground truth optical flow, while the training of FI-Net only requires frame triplets of any video. So our method is superior to FlowNet2 for optical flow estimation.

V. CONCLUSION

In this work, we propose a novel lightweight neural network (FI-Net) for video frame interpolation. FI-Net leverages multi-scale technique and feature-level flow estimation to generate accurate intermediate frames. Moreover, a comprehensive loss that contains Sobolev loss and semantic loss is introduced to train FI-Net. The ablation study proves the effectiveness of the proposed network structure and the proposed loss functions. The objective comparison shows that our method surpasses the previous video frame interpolation methods. We also compare the optical flow estimated by our method with that by FlowNet2. the results show our method is capable of computing competitive optical flow with much smaller model size.

REFERENCES

[1] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4473–4481. doi: 10.1109/ICCV.2017.478.

[2] S. Meyer, O. Wang, H. Zimmer, M. Grosse, and A. Sorkine-Hornung, "Phase-based frame interpolation for video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1410–1418. doi: 10.1109/CVPR.2015.7298747.

[3] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2270–2279. doi: 10.1109/CVPR.2017.244.

[4] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. G. Learned-Miller, and J. Kautz, "Super SloMo: High quality estimation of multiple intermediate frames for video interpolation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 9000–9008.

[5] M. Ogaki, T. Matsumura, K. Nii, M. Miyama, K. Imamura, and Y. Matsuda, "Frame rate up-conversion using hoe (hierarchical optical flow estimation) based bidirectional optical flow estimation," *Int. J. Comput. Sci. Netw. Secur.*, vol. 12, no. 6, p. 52, 2012.

[6] T. Thaipanich, P.-H. Wu, and C.-C. J. Kuo, "Low complexity algorithm for robust video frame rate up-conversion (FRUC) technique," *IEEE Trans. Consum. Electron.*, vol. 55, no. 1, pp. 220–228, Feb. 2009. doi: 10.1109/TCE.2009.4814438.

[7] M. W. Tao, J. Bai, P. Kohli, and S. Paris, "SimpleFlow: A non-iterative, sublinear optical flow algorithm," *Comput. Graph. Forum*, vol. 31, no. 2, pp. 345–353, 2012. doi: 10.1111/j.1467-8659.2012.03013.x.

[8] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. H. Gross, and C. Schroers, "Phasenet for video frame interpolation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 498–507.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

[10] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019. doi: 10.1109/TGRS.2018.2864987.

[11] Q. Wang, Z. Qin, F. Nie, and X. Li, "Spectral embedded adaptive neighbors clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1265–1271, Apr. 2018. doi: 10.1109/TNNLS.2018.2861209.

[12] Y. Yuan, H. Li, and Q. Wang, "Spatiotemporal modeling for video summarization using convolutional recurrent neural network," *IEEE Access*, vol. 7, pp. 64676–64685, 2019.

[13] Y. Yuan, D. Wang, and Q. Wang, "Anomaly detection in traffic scenes via spatial-aware motion reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1198–1209, May 2017. doi: 10.1109/TITS.2016.2601655.

[14] B. Zhao, X. Li, and X. Lu, "HSA-RNN: hierarchical structure-adaptive RNN for video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7405–7414.

[15] Q. Wang, J. Wan, F. Nie, B. Liu, C. Yan, and X. Li, "Hierarchical feature selection for random projection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1581–1586, 2018. doi: 10.1109/TNNLS.2018.2868836.

[16] C. Wang, Y. Yuan, and Q. Wang, "Learning by inertia: Self-supervised monocular visual odometry for road vehicles," 2019, *arXiv:1905.01634*. [Online]. Available: https://arxiv.org/abs/1905.01634

[17] Z. Jiang, Q. Wang, and Y. Yuan, "Modeling with prejudice: Small-sample learning via adversary for semantic segmentation," *IEEE Access*, vol. 6, pp. 77965–77974, 2018.

[18] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1647–1655. doi: 10.1109/CVPR.2017.179.

[19] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-preserving interpolation of correspondences for optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1164–1172. doi: 10.1109/CVPR.2015.7298720.

[20] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large displacement optical flow with deep matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sydney, NSW, Australia, Dec. 2013, pp. 1385–1392. doi: 10.1109/ICCV.2013.175.

[21] J. R. van Amersfoort, W. Shi, A. Acosta, F. Massa, J. Totz, Z. Wang, and J. Caballero, "Frame interpolation with multi-scale deep loss functions and generative adversarial networks," 2017, *arXiv:1711.06045*. [Online]. Available: https://arxiv.org/abs/1711.06045

[22] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 261–270. doi: 10.1109/ICCV.2017.37.

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4489–4497. doi: 10.1109/ICCV.2015.510.

[24] J. Wulff and M. J. Black, "Efficient sparse-to-dense optical flow estimation using a learned basis and layers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 120–130. doi: 10.1109/CVPR.2015.7298607.

[25] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1744–1757, Sep. 2012. doi: 10.1109/TPAMI.2011.236.

[26] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 2758–2766. doi: 10.1109/ICCV.2015.316.

[27] X. Chen, W. Wang, and J. Wang, "Long-term video interpolation with bidirectional predictive network," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, St. Petersburg, FL, USA, Dec. 2017, pp. 1–4. doi: 10.1109/VCIP.2017.8305029.

[28] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 2017–2025. [Online]. Available: http://papers.nips.cc/paper/5854-spatial-transformer-networks

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1026–1034. doi: 10.1109/ICCV.2015.123.

[30] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 105–114. doi: 10.1109/CVPR.2017.19.

[31] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, May 2016, pp. 1–17. [Online]. Available: http://arxiv.org/abs/1511.05666

[32] C. Li and M. Wand, "Combining Markov random fields and convolutional neural networks for image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2479–2486. doi: 10.1109/CVPR.2016.272.

[33] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, "Video (language) modeling: A baseline for generative models of natural videos," 2014, *arXiv:1412.6604*. [Online]. Available: https://arxiv.org/abs/1412.6604

[34] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, May 2016, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1511.05440

[35] S. L. Sobolev, "On a theorem of functional analysis," *Mat. Sbornik*, vol. 4, pp. 471–497, 1938.

[36] W. Lai, J. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

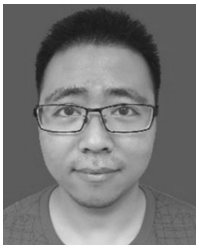
[37] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[38] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: https://arxiv.org/abs/1212.0402

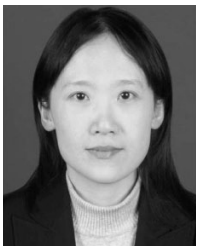
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–15. [Online]. Available: http://arxiv.org/abs/1412.6980

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004. doi: 10.1109/TIP.2003.819861.

[41] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 1701–1710. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Niklaus_Context-Aware_Synthesis_for_CVPR_2018_paper.html
[42] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, 2011. doi: 10.1007/s11263-010-0390-2.
[43] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 611–625. doi: 10.1007/978-3-642-33783-3_44.



HAOPENG LI received the B.E. degree in mathematics and applied mathematics from Northwestern Polytechnical University, Xi'an, China, in 2017, where he is currently pursuing the master's degree with the Center for OPTical IMagery Analysis and Learning (OPTIMAL). His research interests include computer vision and pattern recognition.



YUAN YUAN (M'05–SM'09) is currently a Full Professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. She has authored or coauthored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS and *Pattern Recognition*, as well as conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

QI WANG (M'15–SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.