Split-Attention Multiframe Alignment Network for Image Restoration

YONGYI YU, MINGZHE LIU, HUAJUN FENG, ZHIHAI XU, AND QI LI

State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou 310027, China

Corresponding author: Qi Li (yuyongyi@zju.edu.cn)

This work was supported by the Civil Aerospace Pre-Research Project of China under Grant D040301.

ABSTRACT Image registration (or image alignment), the problem of aligning multiple images with relative displacement, is a crucial step in many multiframe image restoration algorithms. To solve the problem that most existing image registration approaches can only align two images in one inference, we propose a split-attention multiframe alignment network (SAMANet). Pixel-level displacements between multiple images are first estimated at low-resolution scales and then refined gradually with the increase in feature resolution. To better integrate the interframe information, we present a split-attention module (SAM) and a dot-product attention module (DPAM), which can adaptively rescale the cost volume features and optical flow features according to the similarity between features from different images. The experimental results demonstrate the superiority of our SAMANet over state-of-the-art image registration methods in terms of both accuracy and robustness. To solve the “ghosting effect” caused by pixelwise registration, we designed two “ghost” removal modules: warping repetition detection module (WRDM) and attention fusion module (AFM). WRDM detects “ghost” regions during the image warping process without increasing the time complexity of the registration algorithm. AFM uses an attention mechanism to rescale the aligned images and enables the registration network and the subsequent image restoration networks to be trained jointly. To validate the strengths of the proposed approaches, we apply SAMANet, WRDM and AFM to three image/video restoration tasks. Extensive evaluations demonstrate that the proposed methods can enhance the performance of image restoration algorithms and outperform the other compared registration algorithms.

INDEX TERMS Image registration, image restoration, multiframe, attention mechanism, neural networks, optical flow.

I. INTRODUCTION

Image restoration, the problem of restoring a clean image from its degraded version, which includes subtasks such as image denoising [1], image superresolution [2], image decompression [3] and image deblurring [4], is an ill-posed problem. Because multiple consecutive images can provide complementary information, especially in distorted areas, occluded areas and motion-blurred areas in images, utilizing multiple images provides a new way to improve the accuracy of image restoration algorithms. Multiframe image restoration algorithms generally follow the following pipeline: first, the input images are registered explicitly or implicitly, and then a restoration network is used to process the aligned images or features. Since image registration is an important step for multiframe image restoration algorithms, designing a high-performance image registration algorithm is crucial.

The current image registration techniques can be divided into four categories: global registration based on feature points and homography matrix [5]–[7], variational nonrigid registration [8], [9], optical-flow-based registration [10], [11] and implicit feature-domain registration [12], [13]. Among these techniques, global registration algorithms calculate a transformation matrix for the entire image. However, in the case where the depth of field of the image changes greatly or there are objects moving in the image, each part of the image follows different transformation matrices, and global alignment algorithms will fail in these situations. Traditional variational nonrigid alignment algorithms generally combine brightness and gradient constancy assumptions [14] and iteratively optimize displacements by minimizing the brightness difference between the reference image and the warped target image [8]. However, variational
nonrigid algorithms are generally computationally expensive, and state-of-the-art algorithms even take tens of minutes to calculate an image pair with 1.5 million pixels [15], [16]. The excessive computational expenditure limits the application of variational algorithms.

With the development of deep learning techniques, in the field of image and video processing, many algorithms use learning-based optical flow to align the input images of the multiframe network [10], [17]–[19]. However, there are still many problems with these methods. Most of the existing optical flow algorithms only handle two images. If we want to register multiple images, the network needs to be reasoned multiple times, resulting in a large number of repeated calculations. At present, most existing algorithms that utilize multiple images eventually output only one optical flow [20], [21], and they can only still register two images. Moreover, due to the lack of large-scale pretraining datasets containing consecutive optical flow labels, the multiframe algorithms are difficult to train in a supervised manner. Consequently, to the best of our knowledge, there is no supervised learning-based algorithm that can explicitly register multiple images at the pixel level.

The principles of pyramidal processing, cost volume processing and coarse-to-fine manner are well established in the field of optical flow estimation [18]. Various types of feature representations are input to the optical flow estimator to estimate the optical flow at each resolution scale. However, at present, almost all optical flow networks directly concatenate features and then input them into a network to calculate the optical flow directly in the flow estimator, neglecting to explore the relationship between features, which limits the network’s representational ability.

The human visual system can adaptively process visual information and focus on salient areas. Inspired by the human visual system, the attention mechanism [22]–[24] in computer vision can adaptively assign weights to feature representations according to their importance. The emergence and development of attention mechanisms provides us with a way to improve the optical flow network. For most computer vision tasks, such as image segmentation and image superresolution, the features transmitted in a network are abstract features extracted by the network, which have no clear physical meaning and are difficult to classify. Therefore, existing attention algorithms generally do not use independent attention modules for different types of features. This is not suitable for optical flow networks, where flow estimators generally address three types of features: optical flow from the previous scale, cost volume features and common features. The first two types have clear physical definitions, and the three types differ greatly from each other.

Another major problem with optical-flow-based image registration is that most algorithms do not consider “ghosts” caused by optical flow warping or are even unaware of the problem. Through the analysis of IV-A, we find that the “ghosting” effect is inevitable as long as there are occluded areas in the image sequence. Therefore, providing solutions to the “ghosting” effect is of great importance.

To address the above issues, we have established a large-scale multiframe optical flow training set and proposed a split-attention multiframe alignment network (SAMANet) and two “ghost” removal modules: the warning repetition detection module (WRDM) and the attention fusion module (AFM). Specifically, we construct the “FlyingObjects” pretraining dataset, which uses digital camera images from the Internet as backgrounds and uses labels in the image semantic segmentation dataset PASCAL VOC [25] as foregrounds. The dataset contains 22,000 image sequences and up to 1,000 categories of foreground objects, which is more abundant than all existing large-scale training sets. The pipeline of our model is shown in Fig. 1, and the overall model consists of three parts: image alignment, “ghost” removal and image restoration. Degraded images are first aligned by SAMANet, “ghosts” caused by pixel warping are then removed by WRDM or AFM, and the final image restoration result is obtained through a restoration network.

The design of SAMANet follows three key points: hierarchical feature extraction, cascaded displacement refinement and a split-attention module (SAM) to integrate multiframe information. In contrast to existing channel attention and spatial attention modules, our SAM uses two independent dot-product attention modules (DPAMs) to handle the cost volume features and optical flow features separately. The DPAM rescales feature representations from each frame depending on their similarity to the corresponding reference frame features. To address the “ghost” problem, WRDM starts from the definition of “ghost” and detects the pixels used multiple times in the optical flow warping process and their corresponding positions in the warping results. AFM uses DPAM to assign pixel-level weights to each image in the registration results based on the similarity between registered images and the reference image. We quantitatively and qualitatively compare SAMANet with
classical and state-of-the-art registration methods, and the results clearly demonstrate the superiority of our model in terms of both alignment performance and computational efficiency. In addition, SAMANet shows good robustness in special scenes, including non-texture areas, repeated texture areas, specular reflection areas, motion-blurred areas, dark-light areas, high-dynamic-range scenes and scenes containing white Gaussian noise.

Finally, we use the proposed SAMANet and the “ghost” removal modules in image denoising, image superresoultion and video superresolution tasks. For the image restoration task, we first classify the existing algorithms, select classical algorithms from each category and then transform them into multiframe versions. Next, we align the input images with different image registration algorithms and compare the image restoration results. For video superresolution, we also first classify the existing algorithms and then select the classical algorithms and embed the registration modules in them. Compared with the existing registration algorithms, the experimental results verify the superiority of our model as a premodule for image/video networks. We further demonstrate that our model can improve the performance of algorithms that contain implicit registration modules.

The main contributions of this paper can be summarized as follows:

1) We design SAMANet, which can align multiple images at the pixel level with excellent accuracy.

2) We propose a SAM and a DPAM to handle different types of feature representations and better integrate information from multiple images.

3) We construct a multiframe image registration training set that contains richer foreground and background images than all existing large-scale datasets. This dataset enables the multiframe network to be trained in a supervised manner.

4) Our model shows good robustness in various scenarios and under additive white Gaussian noise.

5) We analyze in detail the causes of “ghosts” appearing in the optical flow warping process and propose a WRDM to replace the misaligned areas without increasing the time complexity of the warping process. To enable the registration network and the restoration network to be trained jointly, we design an AFM, which adaptively adds pixel-level weights to the registered images using the attention mechanism.

6) We embed SAMANet and the ghost removal modules into image and video restoration networks and demonstrate their superiority over other registration algorithms.

II. RELATED WORK

A. FEATURE-POINT-BASED IMAGE ALIGNMENT

Most feature-point-based alignment algorithms assume that the transformation between two images satisfies a single transformation matrix. Initially, feature points such as SIFT [26], SURF [5], ORB [6] or CNN-based features [27] are extracted from images. Next, the matching feature point pairs are calculated. Then, the transformation matrix is calculated using random sample consensus (RANSAC) [7] or its variants [28], [29]. Finally, the target image is warped to the reference image according to the transformation matrix. Digital camera images often contain multiple depths of fields and moving objects. The image pairs do not conform to a single transformation matrix. Accordingly, global alignment algorithms have clear limitations in digital camera images. In contrast to the above methods, some studies, such as the homography flow [30], divide the image into several regions, perform independent feature point matching on the different regions, and use the image pyramid to refine the alignment results. This type of method also has obvious shortcomings; for example, the transformations calculated between image blocks are often discontinuous, and it is difficult to calculate the transformation matrices in regions that lack texture.

B. VARIATIONAL IMAGE ALIGNMENT

The variational methods generally use the brightness consistency loss as the main loss function of the displacement optimization [31]. Brox et al. [14] proposed a consistent gradient hypothesis to address brightness changes. Geraldo et al. [32] proposed a transformation model and a new optimization method for directly and robustly registering images. The idea of coarse-to-fine iteration is also widely used in variational methods [8], [9], [33]. In PatchMatch Filter [34] and PatchMatch [35], the researchers calculated the dense correspondence using patch-match algorithms. EpicFlow [9] uses an edge-preserving interpolation algorithm to improve the accuracy of optical flow. These variational algorithms are capable of applying dense transformations to each pixel of the image pairs. However, the large time consumption of the iterative computation limits the application of variational algorithms.

C. CNN-BASED OPTICAL FLOW

FlowNet [19] was the first algorithm to estimate optical flow using a deep neural network with limited accuracy. FlowNet2 [11] and MCN [36] gradually refine the optical flow estimated by stacking several FlowNets, but the overly large network structure limit their application prospects. With the help of the feature warping proposed by LiteFlowNet [37] and PWC-Net [18], optical flow can be refined in a coarse-to-fine manner in a more compact network. Additionally, Zhile Ren et al. [20] improved the accuracy of a single optical flow network by fusing multiple optical flow estimations through a fusion network. Rather than estimating optical flow directly, Zhichao Yin et al. proposed a hierarchical discrete distribution decomposition module [38], which enables the network to generate the optical flow and the uncertainty map simultaneously.

Recently, unsupervised learning has also been widely used in optical flow estimation. Among such studies, UnFlow [39] uses a forward-backward consistency check to estimate occlusion areas and specifically address them. SegFlow [40]
trains an optical flow network and an image segmentation network jointly to improve the accuracy of both networks. DDFlow [41] and SelFlow [42] employ two data distillation methods and train the networks in an “teacher-student” manner. Although unsupervised algorithms have the advantage of not requiring labeled data, they often have complex training steps and lower accuracy than supervised methods.

D. ATTENTION MECHANISM

Since the attention mechanism was proposed [43], it has been extensively studied [44]–[46]. Inspired by nonlocal denoising, Wang et al. [22] proposed a nonlocal operation to compute the long-range dependencies of features. Hu et al. proposed SENet [23] to exploit channelwise relationships and GENet [24] to exploit spatialwise relationships of features. Furthermore, the dual attention net [47] combined channelwise attention and spatialwise attention to achieve enhanced performance for scene segmentation. In low-level vision, SFT-GAN [48] integrates a high-level prior into the image superresolution network using the attention mechanism, EDVR [12] utilizes attention to fuse features from aligned frames, and RCAN [49] combines channel attention and skip connections to enhance the representational ability of deep CNNs. Although attention mechanisms have proven their effectiveness in many tasks, they have not been effectively applied in the field of optical flow and image registration.

E. IMAGE/VIDEO RESTORATION

Image restoration includes many subtasks, and we only introduce three related to this paper.

1) IMAGE DENOISING

Image/video denoising algorithms can be divided into two categories: model-based methods and learning-based methods. Model-based methods such as BM3D [50] and WNNM [51] depend heavily on handcrafted image priors and are generally time consuming, which makes these methods difficult to apply in practice. Recently, convolutional neural networks have achieved significant success in image denoising. DnCNN [52] uses residual learning and batch normalization in denoising networks. FFDNet [53] adds a tunable noise level map in the network to handle spatially variant noise and a wide range of noise levels. Ying et al. [54] introduced a very deep persistent memory network (MemNet) that incorporates a recursive unit and a gate unit to integrate representations with different receptive fields. CBDNet [55] utilizes a noise intensity estimation subnetwork for blind denoising of real-world noisy photographs. To address real raw image noise, Tim Brooks et al. [56] proposed a technique to invert each step of the in-camera image processing pipeline (ISP) to synthesize realistic raw sensor noisy photos and train networks on synthetic raw images.

2) IMAGE SUPERRESOLUTION

Since the three-layer end-to-end neural network SRCNN was proposed [57], many image superresolution methods have been presented. Among them, VDSR [58] and DRCN [59] train a deeper network through residual learning, and EDSR [60] enhances model performance by improving and stacking residual blocks. Combining a residual block and dense block, RDNet [61] can better transmit hierarchical information in the network. In addition to deepening and widening the networks, an attention module [49], [62] is also used in some works to improve the representational power of CNNs.

3) IMAGE SUPERRESOLUTION

Video restoration algorithms are generally based on image restoration methods and to also avoid the problem of spurious flickering artifacts. For video restoration, temporal alignment plays an important role. Some methods use motion compensation [10], [17], [63] networks to align images before restoring them, but they often neglect dealing with the ghosting effect, and the alignment accuracy is lower than the proposed alignment network in this paper. Recently, some methods based on deformable convolution [12], [64], [65] or dynamic filters [13] have been proposed to implicitly align feature maps. These methods generally have good time efficiency, but because they do not explicitly align images, their alignment effect is difficult to quantitatively analyze. Long short-term memory (LSTM) networks [66] and 3D convolutions [67], [68] are also widely used in the field of video reconstruction. Considering that they do not register images or features, we do not compare these algorithms in this paper.

III. SPLIT-ATTENTION MULTIFRAME ALIGNMENT NETWORK (SAMANET)

In this section, we introduce our SAMANet for aligning multiple images at the pixel level. Before providing a detailed description, we first define our notations.

A. NOTATION

SAMANet contains six resolution levels. Subscript \( l \) indicates the resolution scale of the feature map, and feature map from image \( i \) is marked as \( f^l_i \). Given three consecutive images \( I^{l-1}_i, I^l_i \), and \( I^{l+1}_i \), \( w^{l+1}_{l+1} \rightarrow l \) denotes the optical flow from \( I^l_i \) to \( I^{l+1}_i \) at the \( l \)th resolution scale, and \( w^{l-1}_{l+1} \) represents the flow from \( I^{l-1}_i \) to \( I^l_i \). With the estimated optical flow \( w^{l-1}_{l+1} \), we can backward warp \( f^l_i \) towards \( f^{l-1}_i \) and obtain the results of \( f^{l-1}_i, \tilde{I}^{l-1}_i \) as described in III-B. The \( f^l_i \) and the warped \( f^{l-1}_i \) are used to construct a cost volume \( C \), as introduced in III-B. The cost volume between \( f^l_i \) and \( f^{l+1-1}_{l+1} \) is denoted as \( \tilde{C} \). Similarly, \( \tilde{C} \) denotes the cost volume between \( f^l_i \) and \( f^{l-1-1}_i \). In III-C, we present our SAM, and feature representations rescaled by SAM are marked by subscript \( \sim \), for
example, $\tilde{w}_{l-1}^{l}$. In the following paper, we call $I'$ a reference image and $I_{t+1}$ and $I_{t-1}$ target images.

**B. THE ARCHITECTURE OF SAMANET**

As shown in Fig. 2, our SAMANet mainly consists of four parts: hierarchical feature extraction, optical flow estimation, cascaded displacement refinement and image warping. Three consecutive images are input to the network, and optical flows between the reference image and target images are estimated by flow estimators at each resolution level. Optical flows are refined stage by stage to generate the final $w_{l-1}^{l+1}$ and $w_{l-2}^{l-1}$. Using these two optical flows, $I_{t+1}$ and $I_{t-1}$ are warped towards $I'$ to obtain the initial alignment results. In optical flow estimators, we design a novel SAM to better integrate information from multiple frames. Note that the actual feature pyramid contains six resolution levels, but for brevity, we only draw three of them in Fig. 2.

1) **HIERARCHICAL FEATURE EXTRACTION**

The feature extractor constructs feature pyramids for three input images $I'$, $I_{t+1}$ and $I_{t-1}$ with six resolution levels, and each level contains three convolution layers with kernel size $3 \times 3$. Feature maps are downsampled by a convolution layer with $stride = 2$ between scales. The number of channels per level is 16, 32, 64, 96, 128, and 196, respectively.

2) **FEATURE WARPING AND THE CONSTRUCTION OF COST VOLUMES**

Flow estimation becomes more challenging if images are captured far away from each other; we use feature warping to tackle large misalignments between feature maps of different images, similar to PWC-Net [18] and LiteFlowNet [37]. Cost volume [11], [19] measures the response of feature maps from two images to different relative displacements and is widely used as a basic module for optical flow estimation.

Figure 3 shows the details of feature warping and the construction of cost volumes in SAMANet. $w_{l-1}^{l+1}$, which denotes the coarse flow estimated in level $l+1$, is first upsampled in resolution and magnitude to match the resolution of level $l$. Next, the upsampled $w_{l-1}^{l+1}$ is used to warp $f_{l}^{j}$ towards $f_{l}^{t}$ as the following equation:

$$f_{l}^{j \rightarrow t} = f_{l}^{j}(x + w_{l-1}^{l+1}(x)),$$

where $x$ is the coordinate value of a point, $i = t$, and $j \in [t-1, t+1]$. Subsequently, the aligned feature representations $f_{l}^{j \rightarrow t}$ and $f_{l}^{t}$ are used to construct the cost volume as follows:

$$C_{l}(x, d) = \frac{1}{N}(f_{l}^{j}(x) \cdot f_{l}^{j \rightarrow t}(x + d)),$$

where $C_{l}(x, d)$ is the matching cost between point $x$ in feature map $f_{l}^{j}$ and point $x + d$ in $f_{l}^{j \rightarrow t}$. $d$ is the displacement vector from $x$, and $N$ is the length of vector $f_{l}^{j}(x)$. The matching costs are then aggregated into a 3D cost volume.

3) **STRUCTURE OF THE FLOW ESTIMATOR**

The flow estimator integrates multiframe information and estimates optical flow. The dashed box in Fig. 2 shows the structure of the flow estimator. Take the flow estimator that estimates $w_{l-1}^{l+1}$ as an example; $w_{l-1}^{l+1}$, $-w_{l-1}^{l+1}$, $f_{l}^{t}$, $\overrightarrow{C_{l}}$ and $\overleftarrow{C_{l}}$ are concatenated and fed into the SAM to generate the corresponding optical flow. The structure of SAM is presented in the following III-C.

**C. SPLIT-ATTENTION MODULE (SAM)**

Although the effectiveness of convolutional networks in optical flow estimation has been demonstrated, almost all existing
We first classify features into reference features, which limits the representational power of these networks. Recently, attention mechanisms have been extensively studied and applied in the field of image processing and computer vision. However, most attention algorithms input all feature representations directly into a certain attention module without discriminating them according to their characteristics. The input of our attention algorithm is composed of $w_{i+1}^{t-1}$, $w_{i+1}^{t-1}$, $f_l$, $C_l$ and $C_l$. The elements in optical flow represent the pixelwise displacement between images, and the elements in cost volume $C$ represent the response intensity between the feature vectors with different relative displacements. Common features are some abstract representations extracted by the network. There are significant differences among these three types of features, and it is not reasonable to use a single attention module to address all of them. To address the above problems, we propose SAM.

1) STRUCTURE OF SAM

Fig. 4(b) shows the structure of SAM in the flow estimator at resolution level $l$. We use two independent DPAMs, $DPAM^w$ and $DPAM^c$, for optical flow features and cost volume features to obtain the enhanced feature representations $\tilde{w}$ and $\tilde{C}$. Subsequently, a dense block is used to integrate enhanced representations to generate the optical estimation $w_{i-1}^{t+1}$.

![Figure 4. Split-attention module: (a) dot-product attention module (DPAM), (b) the structure of SAM.](image)

2) DOT-PRODUCT ATTENTION MODULE (DPAM)

The structural details of DPAM are shown in Fig. 4(a), which explores the interframe temporal relationship of features. We first classify features into reference features ($feaa'$ in 4(a)) and supplemental features ($feas'$). Taking SAM that generates $w_{i-1}^{t+1}$ as an example, $w_{i+1}^{t-1}$ and $C_l$ are reference features, while $w_{i+1}^{t-1}$ and $\tilde{C}_l$ are supplemental features. Intuitively, a feature that is more similar to the reference feature should receive more attention. DPAM compares the similarity between the supplemental feature and reference feature in a nonlinear embedding space and assigns spatially aggregation weights on each feature accordingly. Specifically, we use two convolution layers to construct embeddings $E'$ and $E^{s}$ for $feaa'$ and $feas'$. Next, we calculate the attention maps according to the following equations:

$$M'(x) = \text{sigmoid}(E'(x) \cdot E'(x)), \quad (3)$$

$$M'(x) = \text{sigmoid}(E'(x) \cdot E'(x)), \quad (4)$$

where $x$ represents a coordinate position. $M'$ and $M^{s}$ are the attention maps for $feaa'$ and $feas'$, respectively. The sigmoid function restricts the output in $[0, 1]$. Following this step, we multiply $feaa'$ and $feas'$ with their corresponding attention maps $M'$ and $M^{s}$ in a pixelwise manner to obtain the enhanced feature representations $\tilde{feaa}'$ and $\tilde{feas}'$:

$$\tilde{feaa}' = feaa' \odot M',$$

$$\tilde{feas}' = feas' \odot M^{s}, \quad (5)$$

IV. GHOST REMOVAL MODULES

A. CAUSES OF "GHOSTING EFFECT"

The "ghosting effect" can be defined as regions where the same object appears more than once. When using optical flow to align images, it is necessary to detect the area where the alignment result may be incorrect and remove it. Therefore, it is necessary to first analyze the causes of the "ghosting effect". This section uses the same notations as in III-A.

1) "GHOST" IN OCCLUSION REGIONS

In the warping process, we warp target images to the reference image using the optical flow as follows:

$$\tilde{I}^{t\rightarrow i}(x) = I^{t}(x + w^{t\rightarrow i}(x)), \quad (7)$$

where $I^t$ denotes the target image, $I^i$ denotes the reference image, $j \in [t - 1, t + 1], i = t$, $w^{t\rightarrow i}$ denotes the optical flow from the reference image to the target image, and $x$ denotes the coordinate value of a pixel. In this way, the "ghosting effect" will appear in occluded regions.

Fig. 5 shows the "ghost" caused by optical flow warping. We select two consecutive images (a)(b) and their optical flow ground truth (c) from the "FlyingThings" dataset [69]. The position of the objects in the ground truth is referred to as image (a). We then warp image (b) towards image (a) according to (c). In the warping result, an obvious "ghosting effect" occurs, and the objects indicated by the arrows appear twice. Fig. 6 illustrates the cause of the "ghosting effect". There is a blue egg at position 1 in the reference image and an egg at position 2 in the target image, 5 cm to the right of position 1. The ground truth of the optical flow is defined to be aligned with the reference image; thus, it is $u = 5, v = 0$ ($u$ is the horizontal displacement, and $v$ is the vertical displacement) at position 1 and $u = 0, v = 0$.
at position 2. We warp the target image toward the reference image using the ground truth. At position 1 of the warped image, the egg at position 2 of the target image is filled according to the ground truth at position 1 \((u = 5, v = 0)\). At position 2 of the warped image, the egg at position 2 of image 2 is filled according to the ground truth at position 2 \((u = 0, v = 0)\). Two blue eggs appeared in the warping result, which is the so-called “ghosting effect”. In an ideal alignment result, the warped image should be as similar as the reference image, which means that the egg at position 2 should be erased.

### 2) “GHOST” IN OPTICAL FLOW DISTORTION REGIONS

Optical flow will be irregularly distorted in some incorrectly estimated areas, which means that in these areas, pixels of the target image will be irregularly warped to the reference image. Because the target image and the reference image have the same number of pixels, there must be pixels that are used multiple times in these regions, resulting in “ghosts”.

From the above analysis, we know that in the process of pixel-level registration, as long as there are occluded areas between image pairs, “ghosts” will appear in the warping results. However, at present, most learning-based algorithms that use optical flow warping to align images, whether in high-level vision tasks or in low-level vision tasks, do not consider the influence of “ghosts”. This limits the rationality and effectiveness of these algorithms. Next, we design two modules to address the “ghost” problem.

### B. WARPING REPETITION DETECTION MODULE (WRDM)

After analyzing the causes of “ghosts”, we design a module to eliminate “ghosts” on the basis of the definition of “ghost” (the same object appears multiple times), which we call the warping repetition detection module (WRDM).

Figure 7 illustrates the pipeline of our WRDM. Fig. 7(a) shows the target image and the reference image. In this figure, the person’s arm swings downward, and the background also has a displacement. First, we warp the target image toward the reference image using the estimated optical flow and obtain the result shown in Fig. 7(c). Due to occlusion, there are two arms on the initial warping result, and the upper one should be erased. Next, we detect the repeated pixels in the warping result, except for the most similar pixels to the reference image. We add other pixels to the suspected “ghost” mask, as shown in Fig. 7(d). There is noise on the initial mask, and the mask on the upper arm is discrete. Therefore, we perform morphological operations on the initial mask to obtain a clean and continuous “ghost” mask, as shown in Fig. 7(d). In the final step, we replace the pixels under the “ghost” mask with the pixels in the reference image and obtain the final warping result, as shown in Fig. 7(f), which is a 50%-50% overlap of the warping result and the reference image. The two overlapping images are almost completely coincident, and there is no “ghost” on the warping result, indicating that our WRDM not only aligns the swinging arm and the background displacement but also eliminates the “ghosting effect” successfully.

We use bilinear interpolation in the initial warping, while in the “ghost” detection algorithm, we round pixel values to count the number of times the pixels are used. The time complexity of WRDM is \(O(N)\). It performs “ghost” detection during the warping process without increasing time complexity, and the time is mainly consumed in the morphological operation of the high-risk mask.

The WRDM still needs to be improved in the following two cases. First, when there is a moving object in the image pair.
but there is no such object in the corresponding optical flow, the error at such a moving object cannot be detected because no pixels are used multiple times in the warping process. Second, when the image is dramatically enlarged (enlarged more than twice) from the target frame to the reference frame, many normal pixels will be used multiple times when warping and thus be incorrectly detected as high-risk areas and cannot be eliminated by the morphological operation. Fortunately, in successive image sequences, we find that such a large expansion hardly occurs.

C. ATTENTION FUSION MODULE (AFM)
WRDM detects “ghosts” from the definition of “ghosting effect” with low time complexity. However, when SAMANet is used as the prenetwork of image restoration networks, if the “ghost” is eliminated by using WRDM, the gradient backpropagated from image restoration networks cannot be transmitted to SAMANet, resulting in SAMANet and image restoration networks not being trained jointly. To enable the entire registration and restoration network to be trained end-to-end, we designed the AFM.

The structure of AFM is the same as that of the DPAM proposed in III-C, where the reference image $I_t$ is the reference input $\text{feat}$ in DPAM, and the aligned target images $\hat{I}^{t+1\rightarrow t}$ and $\hat{I}^{t-1\rightarrow t}$ output by SAMANet are the supplemental $\text{feat}$ in DPAM. DPAM adaptively assign weights to $I_t$, $\hat{I}^{t+1\rightarrow t}$ and $\hat{I}^{t-1\rightarrow t}$. The regions that are similar to the reference image are multiplied by larger weights, while the regions that are not similar are multiplied by smaller weights. Similarity is calculated by vector dot-product in an embedding space. Please refer to Fig. 4(a) for details of the structure.

V. EXPERIMENTAL RESULTS OF IMAGE REGISTRATION
A. DATASETS
1) TRAINING DATASETS OF SAMANET
For image sequences captured in the real world, displacement labels are difficult to obtain. Existing large-scale optical flow datasets such as “FlyingChairs” [19], “FlyingThings” [69], and “MPI-Sintel” [70] are all synthetic datasets. Sun et al. [71] demonstrated that pretraining an optical flow network on a simpler dataset and then fine-tuning it on a harder one is more likely to achieve a stable training result. However, the most widely used pretraining dataset “FlyingChairs” only contains image pairs, and no image sequences are available for multiframe network training. Moreover, as shown in the first row of Fig. 8, the “FlyingChairs” dataset only has chairs as foregrounds, and the monotonicity of the training images in “FlyingChairs” limits its effectiveness as a training set for digital camera image alignment. Therefore, we construct a multiframe optical flow training set with reference to the foreground and background displacement patterns of “FlyingChairs”. We use the segmentation label of the semantic segmentation dataset PASCAL VOC [25] with more than 1000 classes as foregrounds and thousands of digital camera images from the Internet as backgrounds. We apply random and independent affine transformations to the foregrounds and backgrounds and then generate optical flow labels according to the pixel displacements in image sequences. Compared to “FlyingChairs”, our training set has more diverse foreground and background images. Our dataset contains 22,000 image sequences, and we call it “FlyingObjects”. The second and third rows of Fig. 8 are examples of the “FlyingObjects” dataset.
2) EVALUATION DATASETS OF IMAGE REGISTRATION

We use a Canon 80D camera to take 200 image sequences in the low-speed continuous shooting mode. The pictures are taken under various conditions in terms of weather, locations, and time, including sunny, rainy, cloudy days; indoor and outdoor; and day and night. We call the dataset “Own-real” in this paper. The second dataset that we use is the “Kitti2015” [72], [73] dataset built for autonomous driving scenarios.

B. TRAINING DETAILS OF SAMANET

Our SAMANet generates optical flows at five resolution levels, and the highest resolution of the predicted optical flow is one quarter of the original image size. The ground truth is downsampled and scaled to each corresponding level, and the L2 losses between the predicted results and the scaled ground truth are calculated at each resolution level and aggregated into the final loss as follow:

$$\zeta(\Theta) = \sum_{l=0}^{L} \alpha_l \sum_{j \in \{t-1, t, t+1\}} \sum_{x} \| w_{\Theta,j} \rightarrow_f (x) - w_{GT,j} \rightarrow_f (x) \|_2 + \lambda |\Theta|_2, \quad (8)$$

where \( w_{\Theta,j} \rightarrow_f \) denotes the optical flow estimated by SAMANet, \( w_{GT,j} \rightarrow_f \) denotes the optical flow ground truth, \( | \cdot |_2 \) computes the L2 norm of a vector and the second term regularizes the parameters of the model. The loss weights \( \alpha_l \) from the lowest to the highest resolution are set to 0.32, 0.08, 0.02, 0.01, and 0.005 as in FlowNet [19].

We first train the model on our “FlyingObjects” dataset using the \( S_{\text{long}} \) learning rate schedule proposed by FlowNet2 [11], starting from 0.0001 and reduced by half at 0.4M, 0.6M, 0.8M and 1.0M iterations. We crop the images into \( 448 \times 384 \) patches and use a batch size of 8. Subsequently, we fine-tune the model on the “FlyingThings” dataset using the \( S_{\text{fine}} \) learning rate schedule, starting from 0.00001 and reducing by half every 100000 iterations. The crop size in this stage is \( 768 \times 384 \), and the batch size is 4.

C. ALIGNMENT RESULTS AT DIFFERENT RESOLUTION LEVELS

Our SAMANet uses a coarse-to-fine refinement strategy to estimate and refine optical flows at five resolution scales. Fig. 9 shows a set of optical flow estimations of SAMANet and their corresponding registration results at five resolution levels. The resolution at level 2 is a quarter of the original image resolution; after that, for every increase in level number by 1, the resolution of the optical flow decreases by half. At relatively low-resolution levels (high in level number), flow estimators can roughly estimate the overall displacement of the image pair but fail to estimate the displacements of the foreground objects. At higher-resolution levels, flow estimators gradually refine the displacement estimations of the foregrounds, making the alignment results gradually accurate. Note that to present the result of image registration more clearly, in contrast to backward warping in 7, we use forward warping in Fig. 9, which warps the reference image to the target image and will not cause a “ghosting effect” but will leave black holes in the warping results. A comparison of forward warping and backward warping is provided in the supplementary material. SAMANet has two weight-sharing subnetworks for aligning the reference image and two target images. For brevity, only the results of one of the subnetworks are shown in Fig. 9.

D. EFFECTIVENESS OF PROPOSED SPLIT-ATTENTION MODULE

In III-B, we analyze the possibility of improving the accuracy of the image registration algorithm by utilizing information from adjacent multiple images and design the SAM to integrate the features from multiple images. Here, we verify the validity of integrating multiframe information and the SAM.

| TABLE 1. Quantitative assessment of attention mechanism, where PSNRs are calculated for five models in two datasets. ”Two” is the two-frame version network without the attention module, and the others are multiframe versions. ”No”: without attention module; ”SE”: with channel attention module as SE-net; ”GE”: with spatial attention module as GE-net; ”SAM”: with the proposed split-attention module. |
|-----------------|------------|----------|----------|--------|
|                | Two       | No       | SE       | GE     | SAM    |
| Own            | 28.75     | 29.13    | 29.06    | 28.72  | 29.51  |
| Kitti          | 22.33     | 22.54    | 22.35    | 22.27  | 22.82  |

Table 1 shows the results of the quantitative analysis, which compares the PSNR results of five structures on
two test sets. We first briefly describe the five structures. As shown in Fig. 2, our SAMANet contains two subnetworks to estimate $w_{t-1}^0$ and $w_{t-1}^0$, respectively. The optical flow estimator inputs five components $w_{t-1}^0$, $-w_{t-1}^0$, $f_t^0$, $\mathcal{C}_t$ and $\mathcal{C}_{t-1}$ into SAM to generate optical flow. Among the five structures in Table 1, “Two” retains only one subnetwork of SAMANet whose flow estimator accepts features from two images and has no attention module. “NO” retains two subnetworks, but it feeds the five input components of the estimator directly into a dense block without using an attention mechanism. The structures of “SE” [23] and “GE” [24] are consistent with that of SAMANet, except that the SAM in the flow estimators is replaced by the channel attention module and spatial attention module proposed by SENet and GENet. SAM in the table is SAMANet.

The comparison between the PSNR results of the “No” version and the “Two” version shows that using multiple images can improve the accuracy of image registration. Among the three versions of the structure that use the attention mechanism to integrate multi-image information, the PSNR of the “SE” version using the channel attention module is lower than that of the “No” version without the attention module, and the index of the “GE” version using the spatial attention module is even lower than that of the two-frame network. This result indicates that it is unreasonable to use a single attention module to address multiple feature representations with distinct physical meanings. In all the models, our SAM structure has the best PSNR results, which illustrates the effectiveness of SAM.

E. COMPARISON WITH OTHER ALGORITHMS

We compare our SAMANet with five algorithms. Among them, SURF + homography [5] is a traditional algorithm that combines feature point matching with global transformation. Homography flow [30] is a regional feature point matching and transformation algorithm that can apply different transformation matrices to different regions of the image. Coarse-to-fine flow [14], [74] is a variation-based optical flow approach. FlowNet-c [19] and PWC-Net [18] are algorithms based on neural networks. These methods are state-of-the-art or widely used for comparison. All deep learning methods are trained using the same training steps and tested with a Nvidia GTX-1080ti graphics card.

The image pair to be aligned in Fig. 10 follows multiple transformation matrices. The sky, arena, bird and front platform obey different transformations. We superimpose the reference image and alignment results to observe the accuracy of the alignment. Because only one global transformation can be applied to the image, the SURF+homography algorithm fails to align in the sky, the edge of the arena, the bird, and the front platform areas, as shown in Fig. 10(a). Although homography flow can apply multiple transformations to one image, its results often contain a blocking effect, and there are often mismatches and dislocations between blocks. The yellow boxes in Fig. 10(b) mark the seams between blocks, and the areas indicated by other red boxes are also not aligned. Coarse-to-fine flow aligns the image pair relatively well, but it is slightly flawed in the sky area. Furthermore, without the help of WRDM, the warping algorithm provided by coarse-to-fine flow cannot eliminate the misalignments in the bird’s beak and body regions.

Fig. 10(d)-(f) and (g)-(i) compare the alignment results of the state-of-the-art deep learning method PWC-Net [18] and our SAMANet. As shown in Fig. (d) and (g), the optical flow estimated by SAMANet has sharper edges than that estimated by PWC-Net, which means that although both algorithms can estimate the displacement of each pixel in the image pair, SAMANet is more accurate at the edges of objects.

The occluded regions in image pairs cannot be aligned because in these regions, pixels of one image cannot find the corresponding points in the other. The WRDM that we designed detects suspicious “ghost” areas and replaces them; it can be extended to any optical flow algorithms. Fig. (e) and (h) show the “ghost” masks estimated for PWC-Net and SAMANet, where 7.25% of the pixels in the PWC-Net result are determined to be “ghost”, while 5.37% are determined to be “ghost” in SAMANet. After being corrected by the WRDM, the final alignment results are as shown in (f) and (i), and SAMANet aligns almost all regions successfully and achieves the highest PSNR. However, PWC-Net still has errors at the top left of the arena and at the beak. This is because the accuracy of the optical flow estimated by
PWC-Net is lower than that estimated by our SAMANet, and the optical flow at the top of the arena and the edge of the beak is blurry but not distorted, making the WRDM module fail to detect these errors.

Table 2 lists the average PSNR [75] and structural similarity index (SSIM) [76] between the alignment results and the reference images in two test sets, the average running time of the algorithm to align three images and the average “ghost” rate determined by WRDM. SAMANet outperforms all competing algorithms in terms of PSNR and SSIM on both test sets, and the evaluation metrics are improved in all three algorithms using WRDM when only less than 8% of pixels in the “Own-real” test set and less than 17% of pixels in the “Kitti” test set are replaced, as shown in the fourth and eighth columns of the table. The “GR” in Table 2 represents the suspicious “ghost” area rate detected by WRDM, which may be affected by the characteristics of the test set and the robustness of the optical flow model. Details are described in the supplementary material. Moreover, the running time of our SAMANet is less than that of the state-of-the-art two-frame-based algorithm PWC-Net, and it is only slightly inferior to that of FlowNet-c. We also list the running times of the three traditional algorithms. Because we cannot guarantee that they use the GPU in the same way as other deep learning algorithms, their data are marked with ∗.

F. ALIGNMENT RESULTS IN SPECIAL CIRCUMSTANCES

Optical flow and alignment algorithms are often unstable in the following regions or conditions: regions lacking texture, regions with mirror reflections, regions with repetitive textures, regions with motion blur, under dark lighting conditions or under high-dynamic-range circumstances.
TABLE 2. PSNR and SSIM between the alignment results and the reference images in the “Own-real” test set we made and the “Kitti” dataset. Our SAMANet achieves the highest PSNR and SSIM results in all algorithms being compared. The WRDM that we designed improves the alignment results of all three learning-based optical flow networks (FlowNet, PWC-Net and our SAMANet). “Ghost” rate (GR) and the running time to align three images are also listed.

| Methods         | Own-real PSNR | Own-real SSIM | Own-real GR | Kitti PSNR | Kitti SSIM | Kitti GR | time (s) |
|-----------------|---------------|---------------|-------------|------------|------------|----------|----------|
| Global homography | 26.18         | 0.7816        | -           | 1.406*     | 17.45      | 0.6250   | -        |
| Homography flow  | 23.33         | 0.7555        | -           | 2.123*     | 16.80      | 0.5560   | -        |
| Coarse2Fine flow| 28.26         | 0.8235        | -           | 38.79*     | 22.28      | 0.736    | -        |
| FlowNet+c+WRDM  | 27.03         | 0.7981        | -           | 0.363      | 19.91      | 0.6779   | -        |
| FlowNet+c+WRDM  | 27.59         | 0.8111        | 2.3%        | -          | 22.40      | 0.7375   | 10.2%    |
| PWC-net+warp    | 28.75         | 0.8577        | -           | 0.462      | 22.33      | 0.7680   | -        |
| PWC+WRDM        | 30.09         | 0.8688        | 6.3%        | -          | 25.34      | 0.8185   | 15.5%    |
| SAMANet+war     | 29.51         | 0.8668        | -           | 0.424      | 22.82      | 0.7869   | -        |
| SAMANet+WRDM    | 30.48         | 0.8730        | 7.6%        | -          | 26.32      | 0.8252   | 9.2%     |

TABLE 3. Comparison of PSNR between PWC-Net and SAMANet in regions lacking texture, regions with mirror reflections, regions with repetitive textures, regions with motion blur, under dark lighting conditions or under high-dynamic-range circumstances.

| Methods         | Lack | Mirror | Repetitive | Blur | Dark | HDR |
|-----------------|------|--------|------------|------|------|-----|
| PWC             | 29.15| 27.88  | 26.17      | 28.64| 26.43| 26.28|
| SAMANet         | 29.97| 29.12  | 28.58      | 29.58| 28.66| 28.57|

TABLE 4. Robustness test of PWC-Net and SAMANet for additive white Gaussian noise on the “Own-real” test set.

| σ    | 0    | 15   | 25   | 50  |
|------|------|------|------|-----|
| PWC  | 28.75| 28.53| 28.27| 27.64|
| SAMANet | 29.51| 29.50| 29.36| 29.01|

Under these six conditions, we take twenty test sequences each. Table 3 compares PWC-Net with our SAMANet under these sequences. In all six cases, our SAMANet is superior to PWC-Net.

VI. EXPERIMENTS ON IMAGE/VIDEO RESTORATION

This section applies image registration modules to image/video restoration tasks (image denoising, image super-resolution and video superresolution). First, we classify the existing image restoration algorithms and embed different image registration algorithms into them.

A. MODEL CLASSIFICATION AND MODIFICATION

1) TYPES OF IMAGE DENOISING NETWORKS

We classify image denoising networks according to the resolution of the image processed by the main body of the network. As shown in Fig. 12(a), most existing models use a nonlinear network to process the original resolution image [52], [54], [55]. By comparison, to accelerate the model reasoning, the main body of FFDNet [53] processes the downsampled subimages and then reshapes the denoised subimages to the original resolution, as shown in Fig. (b).

2) TYPES OF IMAGE SUPERRESOLUTION (SR) NETWORKS

The classification of single-image superresolution networks is shown in Fig. 13. In early works [57], [58], bicubic interpolation is generally used to upsample low-resolution images to the target resolution, and then nonlinear networks are used to calculate superresolution results. Because network reasoning is performed on high-resolution images, such methods have a large computational overhead. Another type of algorithm calculates the nonlinear mapping of the image on a low-resolution scale and then uses deconvolution, pixel shuffle or other techniques to upsample the results to a high-resolution level [60], [61], [77], as shown in Fig. (b).

3) MULTIFRAME VERSION OF NONLINEAR NETWORKS

To embed the registration module into the above image restoration models, we need to design multiframe versions for the nonlinear networks in Fig. 12 and Fig. 13. We try to minimize the modifications to the nonlinear networks and only add a few structures to their input and output ports. The nonlinear networks used in image restoration generally contain skip connections with different densities, possibly containing batch normalization, gate units and other elements. Fig. 14(a) is a schematic diagram of a single-image nonlinear network whose nonlinear mapping part can be any image.
restoration model. Fig. 14(b) is the corresponding multiframe version of Fig.(a). We change the input of the network to three images, reconstruct these three images using a nonlinear mapping network and skip connection, and fuse the three reconstructed results into one final output using a fusion network composed of three CNN layers. To ensure that the weight quantities of the two versions of the network are consistent, the nonlinear mapping part of the multiframe version is reduced by three layers compared with the single-image version.
4) TYPES OF VIDEO RESTORATION NETWORKS

Based on image restoration methods and to also grasp the temporal consistency, deep learning methods have been widely used in the video restoration field. Most video restoration algorithms restore a frame by integrating several adjacent frames into a restoration network and repeat this procedure until the entire video is restored. According to different approaches of aligning multiple input frames, we divide video restoration algorithms into two categories, as shown in Fig. 15. The first category uses a motion compensation module [10], [17], [63] to align multiple input frames and then inputs the aligned images into a restoration network. This framework is similar to ours but has the following main differences: 1) Our SAMANet is more accurate than the motion compensation modules in these algorithms, which are similar to PWC-Net or FlowNet, as demonstrated in Table 2. 2) These algorithms generally neglect to address the “ghosting” effect. The second category does not align images explicitly but uses a method such as deformable convolution to register the feature representations implicitly in the feature domain and then inputs the aligned features into the restoration network.

5) EMBEDDING IMAGE REGISTRATION MODULE INTO IMAGE/VIDEO RESTORATION NETWORKS

To compare the effect of different image registration algorithms on subsequent image restoration models, we first align three degraded images in Fig. 14(b) before feeding them to the restoration network using different image registration methods.

To verify the effectiveness of our SAMANet for video restoration models, we replace the motion compensation modules with the proposed SAMANet and ghost removal modules for the first category of video restoration models in Fig. 15, and then we compare the video restoration results with the original model. For the second category of algorithms in Fig. 15, we use SAMANet as a premodule to register input frames without changing the original model.

B. EXPERIMENTAL RESULTS OF IMAGE/VIDEO RESTORATION

This subsection shows the experimental results of applying image registration algorithms to image and video restoration tasks. First, we verify whether well-aligned multiple images can improve the performance of a single-image denoising network. Then, the effects of applying different image registration algorithms to image denoising and image superresolution models are compared using image sequences with relative displacement to verify the superiority of our SAMANet. Finally, we demonstrate the result of applying SAMANet to a video superresolution task.

1) DATASET

a: TRAINING DATASETS

In VI-B.3, input images need to be well aligned; we use the DIV2K dataset [78] as a training set, which contains...
900 HD images. We make three copies of an image, add Gaussian white noise with the same $\sigma$ value independently on the three images, and concatenate them as the input of the network. In VI-B.4, we use the slow continuous shooting mode of a Nikon D750 camera to take 200 image sequences as the training set. The image superresolution task uses bicubic interpolation to construct low-resolution images, and the image denoising task adds Gaussian white noise to image sequences to construct degraded images.

b: TESTING DATASETS
In VI-B.3, we use the widely used Set5 and BSD68 as test sets. In VI-B.4, the “Own-real” and “Kitti” test sets introduced in V-A are used. In the video superresolution task in VI-B.5, Vid4 [79] and the test set of the Vimeo-90K dataset [80] are used as test sets. Among them, the data of Vid4 have limited motion, while Vimeo-90K contains various motions.

2) TRAINING DETAILS
In the image denoising task, we use gray patches with a size of $96 \times 96$ as input, while in the image superresolution task, we use RGB patches with a size of $64 \times 64$. The batch size in both tasks is set to 16, and L2 loss is adopted as a loss function with a regularization term weighted by $1 \times 10^{-3}$. We use the Adam optimizer [81] and set $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the learning rate is initialized as $2 \times 10^{-4}$. In VI-B.4, we use the trained registration network to prealign input images of the image restoration networks and then only train the image restoration network. When using the AFM to connect the registration network and image restoration network, we change the mini-patch size to $128 \times 128$, initialize the registration network with the pretrained weights, and then train the registration network, AFM and image restoration network jointly. The weight update speed of the registration network is set to one tenth of the other two modules. In VI-B.5, we replace the motion compensation module in a trained video superresolution network with the trained SAMANet or use SAMANet as a premodule of a video superresolution network and then retrain the network.

3) EXPERIMENTS ON WELL-ALIGNED IMAGES
In this part, we verify whether the multiframe version model performs better in the image denoising task than the single-frame version model when the input images are fully aligned. In VI-A, we divide the single-image denoising model into two categories. We selected the classic network DnCNN and FFDNet from the two categories and transformed them into the three-frame version as shown in Fig.14. The experimental results are shown in Table 5. The three-frame version models improve the PSNR of the two-frame version models by 0.4-0.8 dB. Fig. 16 is a qualitative comparison of the four models on the BSD68 dataset. All four models can roughly remove the noise on the image. However, in the area marked by the red box, obvious artifacts appear in the result of the original DnCNN, and the result of FFDNet is blurred. In contrast, the two multiframe versions retain image details better while successfully removing image noise.

4) EXPERIMENTS ON MISALIGNED IMAGES
From the previous subsection, we know that multiple aligned images can enhance the performance of the image denoising model. However, in most cases, consecutive images taken by digital cameras or smartphones are not well aligned; there are relative displacements between images, and the images may contain complex object motions. Therefore, we need to explore the impact of the registration quality of the input images on the subsequent image restoration algorithms.

![FIGURE 16. Qualitative comparison between two-frame version models and multiframe version models on the BSD68 dataset. Suffix ‘-m’ denotes multiframe version.](image-url)
TABLE 6. PSNR results of multiframe version of DnCNN using different image registration algorithms as the premodule on the “Own-real” and “Kitti” datasets. “Single” represents single-image version of DnCNN, “No” represents inputting unregistered images directly into the network, “S” denotes SAMANet, “W” denotes WRDM and “A” represents AFM.

| Dataset | σ | Single | No | SURF | PWC | S | S+W | S+A |
|---------|----|--------|----|------|-----|---|------|-----|
| Own     | 15 | 32.44  | 30.17 | 32.53 | 32.59 | 32.76 | 32.85 | 32.78 |
|         | 25 | 29.97  | 27.46 | 30.25 | 30.38 | 30.41 | 30.46 | 30.37 |
|         | 50 | 27.24  | 25.37 | 27.41 | 27.62 | 27.83 | 27.94 | 27.75 |
| Kitti   | 15 | 31.58  | 29.14 | 31.66 | 31.71 | 31.84 | 32.91 | 31.87 |
|         | 25 | 29.06  | 26.23 | 29.31 | 29.29 | 29.47 | 29.54 | 29.38 |
|         | 50 | 26.29  | 23.95 | 26.62 | 26.75 | 26.84 | 26.89 | 26.72 |

First, we compare different image registration algorithms on the image denoising task. As shown in Table 6, we compare the denoising results of the multiframe version of DnCNN after using five image registration algorithms as its premodule. The registration algorithms involved in the comparison include SURF+homography, PWC-Net, our SAMANet and the combination of SAMANet and the proposed ghost removal modules WRDM and AFM. As a control group, single-frame DnCNN and the results of feeding unregistered images directly into the multiframe DnCNN are also listed. When multiple input images are not aligned, the PSNR results of the multiframe version of DnCNN are significantly worse than those of the single-frame version. When combined with SURF+homography or PWC-Net, the PSNR results of the multiframe version network are slightly better than those of the single-frame network. Our SAMANet has the highest accuracy among all registration algorithms, and WRDM can further improve the PSNR of the model by a small amount. In contrast, AFM does not perform well in this task and may even diminish the PSNR of the SAMANet version network. This result may be because AFM assigns weights to registered images based on their similarity with the reference image in an embedding space. When each image is added with noise independently, this similarity measure can be distorted by noise and is no longer accurate.

FIGURE 17. Qualitative comparison between the denoising results of DnCNN-m combined with three different image registration algorithms.

Figure 17 shows a qualitative comparison between DnCNN-m combined with three different image registration algorithms. The denoising results of surf+homography and PWC-net versions contain an obvious “ghosting” effect, as indicated by the red arrows, while the version using our SAMANet as a premodule has no “ghost” in the result due to the better registration quality.

Similarly, we compare different preregistration models on the image superresolution task. In VI-A, we classify SR networks into two categories according to the location of the upsampling process in SR networks. We selected the classic networks VDSR and RDN from the two categories as basic models and transformed them into the three-frame version as shown in Fig. 14 and use different image registration models as their premodules. The results are presented in Table 7, and the abbreviations of the registration models are the same as in Table 6. Our SAMANet version models outperform the other algorithms. In contrast to the image denoising task, the combination of SAMANet and AFM achieves the highest PSNR results in most cases in the image superresolution task, and the WRDM can still slightly improve the PSNR results.

TABLE 7. PSNR results of multiframe version of VDSR and RDN using different image registration algorithms as premodule on image superresolution task.

| Dataset | Multiframe VDSR | Multiframe RDN |
|---------|-----------------|----------------|
|         | SURF | PWC | S | S+W | S+A  | SURF | PWC | S | S+W | S+A  |
| Own     | X2   | 31.97 | 32.26 | 32.35 | 32.44 | 32.36 | X2   | 31.93 | 32.24 | 32.36 | 32.40 | 32.39 |
|         | X3   | 27.57 | 27.68 | 27.76 | 27.79 | 27.90 | X3   | 27.58 | 27.69 | 27.93 | 27.96 | 28.16 |
| Kitti   | X2   | 29.78 | 29.89 | 29.96 | 29.98 | 30.10 | X2   | 29.84 | 29.92 | 30.14 | 30.17 | 30.21 |
|         | X3   | 24.28 | 24.37 | 24.49 | 24.53 | 24.57 | X3   | 24.51 | 24.69 | 24.96 | 24.99 | 25.08 |

FIGURE 18. Qualitative comparison for image superresolution (X2).

Figure 18 presents a qualitative comparison of the PWC-Net version of multiframe RDN and the SAMANet plus AFM version one for image superresolution. In the results of bicubic upsampling and PWC-Net as the premodule, the direction of stripes in the roof is opposite to the actual one, while the result of our SAMANet plus AFM version model is basically correct. At present, both RDN-based models have the problem of over sharpening, which needs to be improved in the future.
5) RESULTS ON VIDEO SUPERRESOLUTION

In VI-A, we divide video restoration methods into explicit alignment and implicit alignment categories according to whether the restoration network first aligns the input frames explicitly. From these two categories, we select classical networks VESPCN [10] and EDVR [12] respectively to verify our algorithm. Among them, VESPCN uses two cascaded U-shaped networks as the motion compensation module to explicitly register the input frames. Since this module can only output one optical flow, it needs to be inferred twice to register the three input frames. In contrast, EDVR uses a CFD alignment module (alignment with pyramid, cascading and deformable convolution) to implicitly align the abstract features of the input frames extracted by a network. After obtaining the aligned images or features, VESPCN and EDVR feed them into a reconstruction network to compute the restoration results. For VESPCN, we directly replace its motion compensation module with our SAMANet, and for EDVR, we use SAMANet to prealign its input frames. EDVR accepts seven input frames. We use the middle frame as a reference frame, and we simply transform the three-frame structure of SAMANet into a seven-frame structure. Specifically, the optical flow estimators in SAMANet are copied three times, and each optical flow estimator only integrates features from the reference frame and two symmetrical target frames.

Table 8 presents the PSNR results of video SR. Compared with the original VESPCN and EDVR, our SAMANet and WRDM can enhance the PSNR by 0.1-0.2 dB.

| input frames | V   | S+W+V | E   | S+W+E |
|--------------|-----|-------|-----|-------|
| Vid4         | 25.35† | 25.51 | 27.35† | 27.48 |
| Vimeo-90K    | 35.37† | 35.59 | 37.61† | 37.77 |

TABLE 8. PSNR results of embedding our SAMANet and WRDM into VESPCN and EDVR on Vid4 and Vimeo-90K test sets for 4 x video SR; "V" represents VESPCN, "S" denotes SAMANet, "W" means WRDM, "E" denotes EDVR and "+" means the values are taken from their publications. PSNR are evaluated on the Y (luminance) channel.

VII. CONCLUSION

In this paper, we proposed SAMANet to align multiple images with relative displacement at the pixel level. Taking advantage of the SAM, SAMANet can better integrate various types of feature representations from multiple images. Compared with the existing two-frame architectures, the multiframe architecture of SAMANet can avoid repeated computations caused by multiple inferences when aligning multiple images. To solve the “ghost” problem caused by optical flow warping, we designed two ghost removal modules. Among them, the WRDM can detect “ghost” at the same time as image warping with O(N) time complexity, and the AFM uses the attention mechanism to adaptively connect SAMANet and the subsequent image restoration network such that the two networks can be trained jointly. Furthermore, we applied SAMANet, WRDM and AFM to image denoising, image superresolution and video superresolution tasks. The experimental results demonstrate the superiority of SAMANet over existing registration algorithms and verify that our model can improve the performance of existing image restoration algorithms.

REFERENCES

[1] S. Li, Y. Chen, R. Jiang, and X. Tian, “Image denoising via multi-scale gated fusion network,” IEEE Access, vol. 7, pp. 49392–49402, 2019.
[2] P. Liu, Y. Hong, and Y. Liu, “Deep differential convolutional network for single image super-resolution,” IEEE Access, vol. 7, pp. 37555–37564, 2019.
[3] V. Itier, F. Kucharzak, O. Strauss, and W. Puech, “Interval-valued JPEG decompression for artifact suppression,” in Proc. 8th Int. Conf. Image Process. Theory, Tools Appl. (IPTA), Nov. 2018, pp. 1–6.
[4] Z. Dou, K. Gao, X. Zhang, and H. Wang, “Fast blind image deblurring using smoothing-enhancing regularizer,” IEEE Access, vol. 7, pp. 90904–90915, 2019.
[5] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2006, pp. 404–417.
[6] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, “Orb: An efficient alternative to sift or surf,” in Proc. ICCV, 2011.
[7] M. A. Fischer and R. C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” Commun. ACM, vol. 24, no. 6, pp. 381–395, 1981.
[8] P. Weinzaepfel, F. Revaud, Z. Harchaoui, and C. Schmid, “DeepFlow: Large displacement optical flow with deep matching,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 1385–1392.
[9] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, “EpisicFlow: Edge-preserving interpolation of correspondences for optical flow,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1164–1172.
[10] 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 Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4778–4787.
[11] 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), Jul. 2017, pp. 2462–2470.
[12] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, “Edvr: Video restoration with enhanced deformable convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2019.
[13] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, “Deep video super-resolution network using dynamic upsample filters without explicit motion compensation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3224–3232.
[14] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2004, pp. 25–36.
[15] A. Behl, O. H. Safari, S. K. Mustikovela, H. A. Alhaija, C. Rother, and A. Geiger, “Bounding boxes, segmentations and object coordinates: How important is recognition for 3D scene flow estimation in autonomous driving scenarios?” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017.
[16] C. Vogel, K. Schindler, and S. Roth, “3D scene flow estimation with a piecewise rigid scene model,” Int. J. Comput. Vis., vol. 115, no. 1, pp. 1–28, 2015.
[17] A. Kappeler, S. Yoo, Q. Dai, and A. K. Sattaggeles, “Video super-resolution with convolutional neural networks,” IEEE Trans. Comput. Imag., vol. 2, no. 2, pp. 109–122, Jun. 2016.
[18] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, “PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8934–8943.
[19] A. Dosovitskiy, P. Fischer, E. Ilg, P. Haussser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, “FlowNet: Learning optical flow with convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2758–2766. [Online]. Available: http://ml.informatik.uni-freiburg.de/Publications/2015/DFB15
[68] Y. Huang, W. Wang, and L. Wang, “Video super-resolution via bidirectional recurrent convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 1015–1028, Apr. 2018.

[69] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, “A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16 and https://arxiv.org/abs/1512.02134

[70] 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. (ECCV), in Lecture Notes in Computer Science, vol. 7577, A. Fitzgibbon, Ed. New York, NY, USA: Springer-Verlag, Oct. 2012, pp. 611–625.

[71] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, “Models matter, so does training: An empirical study of CNNs for optical flow estimation,” 2018, arXiv:1809.05571. [Online]. Available: https://arxiv.org/abs/1809.05571

[72] M. Menze, C. Heipke, and A. Geiger, “Object scene flow,” ISPRS J. Photogram. Remote Sens. (ISPRS), vol. 140, pp. 60–76, Jun. 2018.

[73] M. Menze, C. Heipke, and A. Geiger, “Joint 3d estimation of vehicles and scene flow,” in Proc. ISPRS Workshop Image Sequence Anal. (ISA), 2015.

[74] D. Pathak, R. Girshick, P. Dollar, T. Darrell, and B. Hariharan, “Learning features by watching objects move,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017.

[75] Q. Huynh-Thu and M. Ghanbari, “Scope of validity of PSNR in image/video quality assessment,” Electron. Lett., vol. 44, no. 13, p. 800, 2008.

[76] A. Hore and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in Proc. 20th Int. Conf. Pattern Recognit., Aug. 2010, pp. 2366–2369.

[77] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1874–1883.

[78] R. Timofte et al., “Ntire 2018 challenge on single image super-resolution: Methods and results,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, Jun. 2018.

[79] C. Liu and D. Sun, “On Bayesian adaptive video super resolution,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 2, pp. 346–360, Feb. 2014.

[80] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video enhancement with task-oriented flow,” Int. J. Comput. Vis., vol. 127, no. 8, pp. 1106–1125, Aug. 2019.

[81] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980

**YONGYI YU** received the B.E. degree from the College of Physics and Mechanical Engineering, Xiamen University, in 2016. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Modern Optical Instrumentation, Zhejiang University. His research interests include computer vision, deep learning, and image processing.

**MINGZHE LIU** received the B.S. degree from the School of Science, Beijing Jiaotong University, in 2017. He is currently pursuing the M.A.Sc. degree with the State Key Laboratory of Modern Optical Instrumentation, ZJU. His research interests include computer vision and image processing.

**HUAJUN FENG** received the B.S. degree from Zhejiang University, in 1983, and the master’s degree from Zhejiang University (ZJU), in 1986. He is currently a Professor with the State Key Laboratory of Modern Optical Instrumentation, ZJU. His research is in the field of imaging technique, imaging processing, precision testing technology, and optical system design.

**ZHIHAI XU** received the bachelor’s, master’s, and Ph.D. degrees from Zhejiang University (ZJU), in 1986, 1989, and 1996, respectively. He is currently a Professor with the State Key Laboratory of Modern Optical Instrumentation, ZJU. His research is in the field of optical remote sensing and imaging chain of cameras.

**QI LI** received the Ph.D. degree in optical engineering from Zhejiang University, in 2004. He is currently an Associate Professor with the State Key Laboratory of Modern Optical Instrumentation, ZJU. His research is in the field of optical system design and imaging technique.

---

**YONGYI YU**

**MINGZHE LIU**

**HUAJUN FENG**

**ZHIHAI XU**

**QI LI**