SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2023, Vol. 66 192301:1–192301:13
https://doi.org/10.1007/s11432-022-3634-x

A multi-frame sparse self-learning PWC-Net for motion estimation in satellite video scenes

Tengfei WANG[1], Yanfeng GU[1*] & Shengyang LI[2,3]

[1]*School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin* 150001*, China;*
[2]*Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing* 100094*, China;*
[3]*University of Chinese Academy of Sciences, Beijing* 100049*, China*

Received 19 June 2022/Revised 13 September 2022/Accepted 24 November 2022/Published online 21 August 2023

Abstract Motion estimation is an important approach to acquiring motion information of all targets in satellite video while it provides the ability to real-time monitor the Earth observation region. Compared with the case in computer vision, motion estimation in satellite video has to face two main difficulties: the large scale of observation and numerous weak targets of low signal-to-noise ratio. In this paper, a multi-frame sparse self-learning PWC-Net (MSSPWC-Net) is proposed to implement motion estimation of the weak targets in satellite video. To overcome the shortage that the existing PWC-Net fails to extract motion information from numerous weak targets, motion consistency and sparse self-learning are introduced to modify the pyramid, warping, and cost volume convolutional neural networks (CNN) network (PWC-Net). The motion consistency between neighboring frames as a multi-frame framework is mainly used to improve the accuracy of motion estimation of the weak targets, and sparse self-learning is adopted to deal with the case that labeled samples in satellite video are insufficient to train PWC-Net. Numerical experiments are conducted on 4 real satellite video datasets. Experimental results demonstrate that the proposed MSSPWC-Net achieves the excellent performance of motion estimation of the weak targets in satellite video and outperforms the state-of-the-art methods.

Keywords satellite video scenes, motion estimation, small blurry targets

Citation Wang T F, Gu Y F, Li S Y. A multi-frame sparse self-learning PWC-Net for motion estimation in satellite video scenes. Sci China Inf Sci, 2023, 66(9): 192301, https://doi.org/10.1007/s11432-022-3634-x

1 Introduction

The development of multi-temporal remote sensing data advances the requirement of real-time earth observation. The video satellites such as SkySat and Jilin series business satellites provide near real-time multi-temporal videos of a large-range ground with 1 m spatial resolution, which helps us observe the moving transportation on the ground. Benefiting from the large scale and real-time observation, massive numbers of events are contained in a satellite video. It brings the requirements and challenges of data processing. To analyze the motion state of the targets in the satellite video, we divide the large-scale satellite video into several small scenes. Referring to the definition of image scene, satellite video scene refers to the space-time data block represented by the spatial area that can independently describe an event in satellite video. Generally, we process the satellite video in the scenes but not a whole video because processing the large-scale images is a challenge to the computing device. The satellite video scene processing is different from the traditional videos while it is interested in the status of the targets but not the action of a target. The key problem of satellite video scene processing is extracting the dynamic information.

Motion estimation refers to estimating the 2-D image-plane moving speed and direction of the targets. It is the fundamental problem in analyzing the dynamic information of videos [1–3] in computer vision. Optic flow is often used to describe the motions of the images pixel by pixel. There are massive applications of motion estimation in satellite videos. Dynamic object detection [4], tracking [5–7], video

* Corresponding author (email: guyf@hit.edu.cn)

segmentation [8], video superpixels [9,10] and understanding [11–13] treat motion estimation as the basic features to find or track the targets. Unfortunately, those applications rely on accurate motion estimation but there is no pointed research on satellite video.

Motion estimation has a long history in computer vision. The methods of the motion estimation can be simply classified as traditional methods and deep learning methods. Traditional motion estimation methods require gradient information to calculate the motion; they failed to calculate the optical flow when the background is complex [14]. Researchers of motion estimation have made great progress based on deep learning in computer vision [15–18]. Thanks to the deep features and supervised method, the deep learning method achieved more accurate motion than the traditional methods. FlowNet [19, 20] brings the possibility to obtain optic flow via deep features. PWC-Net [21] improves FlowNet using the cost volume. Those studies extract precise features of targets based on massive amounts of training samples to solve the complex background question of traditional methods [22]. Deep learning methods become the main methods due to the advantages of more accurate motion estimation.

A basic deep-learning motion estimation method can be partitioned into three parts: feature extraction, correlation, and motion estimation. Feature extraction is used to match the same target between two frames, correlation is used to find the displacement relation of the point pairs from different frames, and motion estimation is used to estimate the motion from the correlation. We often define the loss function to train the network by the correlation but not the motion directly. The structure of a fully connected network (FCN) is always used as the end-to-end spatial feature extractor [23]. Instead of a traditional artificial feature extractor, FlowNet and PWC-Net use FCN or the other FCN-based method (such as Unet) to obtain the end-to-end features. Commonly, the features are similar to segmentation or pixel by pixel classification [24]. Training an FCN-based feature extractor requires massive amounts of data due to the large parameter scale. Transfer learning and fine-tuning methods are necessary when the network is used on a different dataset. The problem is deep feature extractors are not sensitive to small targets. The features may confuse the group targets with a single target. FlowNet opens the way to calculate the correlation features of two frames. Stacking and difference are commonly-used idea to combine the features of two frames, but they are ill-conditioned to calculate the motions. They calculate the features using a linear relationship but the optical flow field is not continuous on the time axis. The optical flow field is continuous along the direction of the motion in 3-D space, which means that the offset of the frames should be taken into account. Cost-volume [25–27] solves this problem by calculating the correlation of the patches. It is a commonly used method to calculate the offset of 3-D objects. PWC-Net successfully applies the cost volume and calculates the optical flow. Additionally, it is difficult to define the loss function and the backward function by the gradient information of frames. Calculating the correlation of different directions is more accurate but time-consuming. Commonly, training the networks needs a large number of training samples.

The primary aim of this paper is to extract the motions in satellite video scenes. The background of satellite video scenes is complex and noisy, which makes the traditional methods fail to extract the dense motions. Moreover, traditional methods are time-consuming in that they need to train the models when calculating the motions. It is also difficult to apply the deep learning method directly. Proper features can solve the problem of the complex background, but they are helpless for small targets and noise. The satellite video scenes, especially the city scenes, contain massive numbers of small blurry targets; labeling the ground truth of those targets' motion is a challenging job. In conclusion, there are two challenges to extracting the dynamic information in satellite video scenes. First, how to extract the motions of unlabeled small blurry targets. Second, how to extract accurate motions of the blurry targets from noisy backgrounds.

Self-learning [28–30] is an ideal method to solve the sample insufficient question. SelFlow [31] and UnFlow [32] train the optical flow on other datasets by defining a new loss function and retraining the networks to fit the dataset. Warping [33] the predicted optical flow to the frames is the common way to reconstruct the frames. The warping function is also used in video superpixels, video understanding, and video segmentation areas where the frames need to be fused by the optical flow. SelFlow and UnFlow conduct the self-training methods using the warping layers on other datasets. Although the self-training methods achieved good results benefiting from the adjusted deep features, they are sensitive to the noise of training samples because they are unsupervised methods [34–36]. It is difficult to apart the motions of blurry targets and the noise.

It has been noted that multi-frame information is helpful for accurate motion estimation of complex moving targets. The results from earlier studies demonstrate that three frames can help improve the

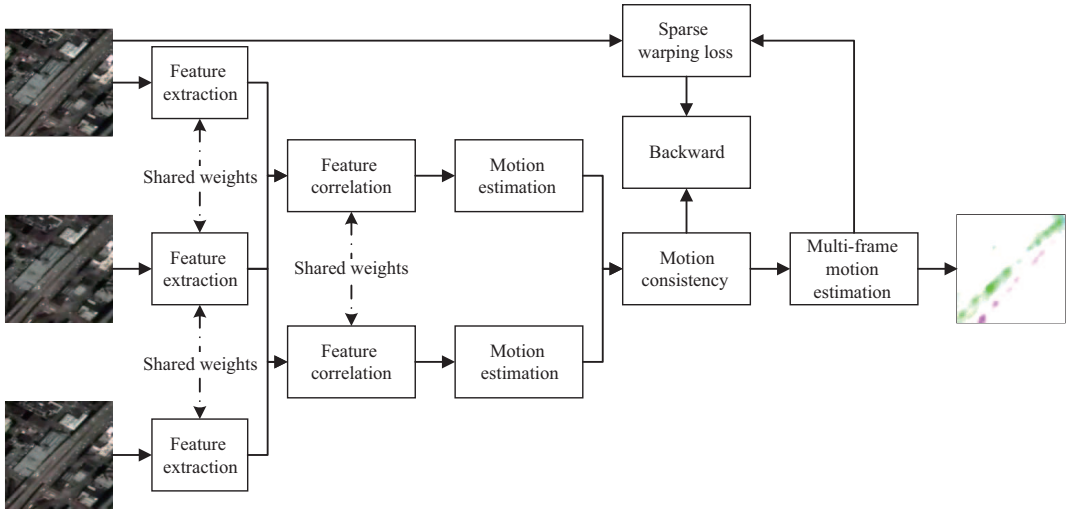


Figure 1 (Color online) The mainframe.

optical flow [37]. However, those methods are designed to solve the mismatch question but not the blurry moving target question. The potential information of multiple frames should be utilized to extract blurry targets' motion from the complex noisy background.

This study points to solve the questions of motion flow estimation in satellite video scenes. Self-learning methods can solve the sample insufficient question but those methods need improvements to enhance the features of the small targets. The self-learning method is an unsupervised method which means it needs additional information about the moving targets. We design a sparse self-learning method to solve the small target sample insufficient question. A sparse warping loss function is proposed to enhance the small targets' sensitivity to the proposed self-learning method. The targets of satellite videos are sparse on the images relative to the background, which is helpful in the detection methods. We solve the blurry target motion problem by using a motion consistency constraint. It is necessary to add multi-frame information to estimate the motion flow of blurry targets. To enhance the motion of the blurry target, we design a multi-frame framework according to the motion consistency of the moving targets. By the multi-frame framework, we succeed in fusing the motions of neighbor frames to estimate the accurate motions of the blurry targets.

There are two contributions of the proposed method.

(1) A sparse self-learning PWC-Net is proposed to extract the motion flows of small targets under satellite video scenes. We improve the warping loss function of the self-learning method using a sparsity constraint so that the feature extractors of PWC-Net are sensitive to the small targets without labeled samples.

(2) A motion consistency constraint that points to the characteristic of the satellite video targets is proposed. It is helpful to filter the noise and obtain the accurate motion flow of blurry targets combined with the multi-frame framework.

The remainder of the paper proceeds as follows: Section 2 presents the mainframe and the detailed method. Section 3 proposes the experimental results to prove the effectiveness of the proposed method on satellite videos. Section 4 concludes the paper.

2 The proposed method

2.1 The main framework

The multi-frame sparse self-learning PWC-Net (MSSPWC-Net) consists of spares self-learning PWC-Net and multi-frame framework. The section begins by laying out the framework of motion flow extraction and then explains the sparse self-learning PWC-Net (SSPWC-Net). Finally, we propose the multi-frame framework and obtain the motion flows by combining the methods. Figure 1 shows the mainframe of motion estimation.

A video V is composed of multiple frames. Let $\boldsymbol{I}(\boldsymbol{x}, t)$ denote the image of a frame in time t, where

$\boldsymbol{x} = (x, y)$ is the spatial location of a pixel in the image. When $\boldsymbol{x}$ and t are fixed, $\boldsymbol{I}$ stands for the band vector of the pixel. Let $\boldsymbol{s}(\boldsymbol{x}, t)$ denote the motion vector between two frames $\boldsymbol{I}(\boldsymbol{x}, t)$ and $\boldsymbol{I}(\boldsymbol{x}, t + \Delta t)$. According to the gray invariant hypothesis, we have

$$\boldsymbol{I}(\boldsymbol{x}, t) = \mathcal{F}(\boldsymbol{x}), \tag{1}$$

$$\boldsymbol{I}(\boldsymbol{x}, t + \Delta t) = \mathcal{F}(\boldsymbol{x} + \boldsymbol{s}(\boldsymbol{x}, t)), \tag{2}$$

where $\mathcal{F}(\boldsymbol{x})$ denotes the intensity function of the first frame $\boldsymbol{I}(\boldsymbol{x}, t)$. Consider the third frame:

$$\boldsymbol{I}(\boldsymbol{x}, t + 2\Delta t) = \boldsymbol{I}(\boldsymbol{x} + \boldsymbol{s}(\boldsymbol{x}, t + \Delta t), t + \Delta t). \tag{3}$$

The relation between $\boldsymbol{s}(\boldsymbol{x}, t)$ and $\boldsymbol{s}(\boldsymbol{x}, t + \Delta t)$ is not linear:

$$\boldsymbol{s}(\boldsymbol{x}, t + \Delta t) = \boldsymbol{s}(\boldsymbol{x} + \boldsymbol{s}(\boldsymbol{x}, t), t). \tag{4}$$

The purpose of motion estimation is to solve $\boldsymbol{s}(\boldsymbol{x}, t)$. Obviously, $\boldsymbol{s}(\boldsymbol{x}, t)$ is a hidden function and has no unique solution. Traditional motion estimation methods use the partial derivative to transform the hidden function into an explicit solution.

$$\frac{\partial \boldsymbol{I}}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial \boldsymbol{I}}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} + \frac{\partial \boldsymbol{I}}{\partial t}\frac{\mathrm{d}t}{\mathrm{d}t} = 0. \tag{5}$$

The solution of (4) relies on additional assumptions. The commonly used assumption is the smoothness assumption. The method cannot achieve ideal results under a complex noisy background. Obviously, from (2), we can assume that the existing function $\mathcal{G}(\cdot)$ satisfies

$$\mathcal{G}(I(\boldsymbol{x}, t + \Delta t)) = \mathcal{G}(\mathcal{F}(\boldsymbol{x} + \boldsymbol{s}(x, t))) = \boldsymbol{x} + \boldsymbol{s}(\boldsymbol{x}, t). \tag{6}$$

Thus

$$\boldsymbol{s}(x, t) = \mathcal{G}(\boldsymbol{I}(\boldsymbol{x}, t + \Delta t)) - \boldsymbol{x}, \quad \text{where } \mathcal{G}(\mathcal{F}(\boldsymbol{x})) = \boldsymbol{x}. \tag{7}$$

It is difficult to solve $\mathcal{G}(\boldsymbol{x})$ because the equation is discrete and one-to-many. We should first translate the equation to one-to-one, so we must obtain the position invariant features. It is similar to moving target detection. Let $\mathcal{P}(\cdot)$ denote the feature extraction function. Thus,

$$\boldsymbol{s}(x, t) = \hat{\mathcal{G}}(\mathcal{P}(\boldsymbol{I}(\boldsymbol{x}, t + \Delta t))) - \boldsymbol{x}, \tag{8}$$

where $\hat{\mathcal{G}}(\mathcal{P}(\boldsymbol{I})) = \boldsymbol{x}$. It can be regarded as a codec process.

There are three advantages of using the features. First, it redefines the method to calculate the motions instead of the partial derivative method and raises the accuracy theoretically. Second, it can filter the weakly changed background and the noise. Third, it can combine multichannel information. Relatively, it is time-consuming.

Therefore, the motion estimation problem is transferred to a feature extraction problem. We can design proper features of the targets as the target detection methods do, however, the information of a current image is not sufficient. We can only utilize the neighbor pixels and the band information. It may be complex to obtain one-to-one features, so most methods limit the areas to extract the features.

Artificial design of the features is difficult. Most researchers obtain the features using training models. Additionally, the model should fit the prior information. We should create the relationship between the ground truth and the predicted motions. The loss function is used to evaluate the relationship.

$$\mathcal{L}(\theta) = \|\hat{\boldsymbol{s}}(\theta) - \boldsymbol{s}\|, \tag{9}$$

where $\hat{\boldsymbol{s}}(\theta)$ is the predicted motion by the features θ. We use the feedback to adjust the parameters of the model by minimizing the $\mathcal{L}(\theta)$.

There are massive methods to obtain the motions from the images, but all of the methods should create the relationship between the gray image and the motion information. The main challenge of deep learning methods is also to solve (2).

Feature extraction is the core topic of the deep learning method. Spatial feature extraction is not important to most traditional motion estimation methods because of the clear dataset, but deep learning

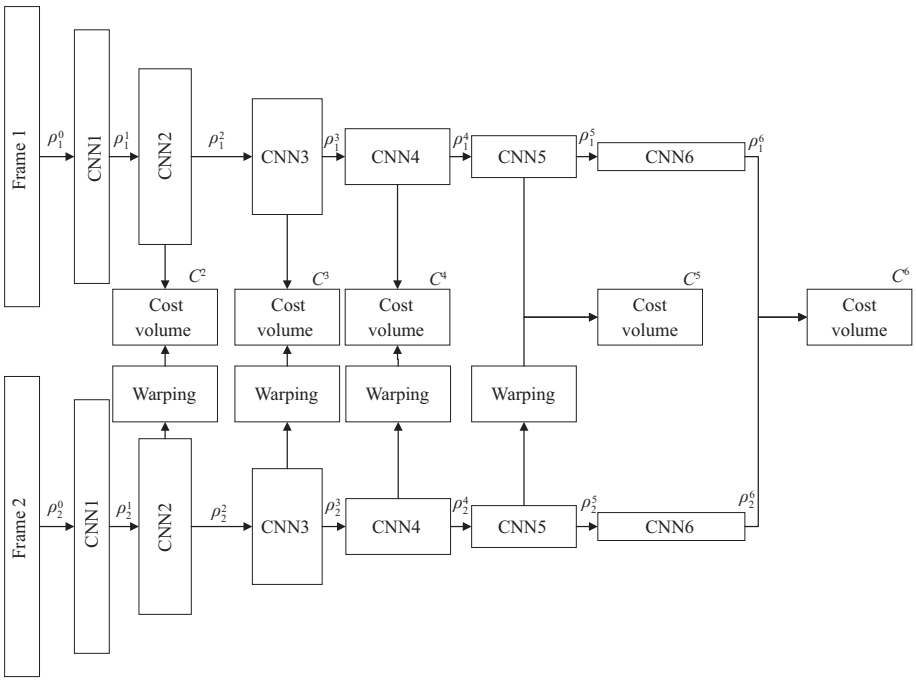


Figure 2 The feature extraction of PWC-Net.

methods rely on good features to simplify the following steps. For example, if the feature is good enough to detect the targets, it is unnecessary to use (5). How to obtain proper features from the massive data is the main topic.

It is impossible to obtain the motions from a single image, so finding the correlation features is an indispensable step. The commonly used methods are stacking, frame difference, and correlation. Stacking and frame difference treat the spatial locations of two frames as the same position, which is unable to calculate the motions. The gray value of moving objects is discontinuous on the time axis, so the stacking and frame differences are improper. Correlation is helpful to calculate the relationship, but it is difficult to calculate the subpixel displacement.

We use the feature correlation to calculate the relationship between two frames, such as (7), and train the network by regressing the features. The prevailing method of deep learning deconvolutes the correlation features to end-to-end results to estimate the motions [38].

The loss function is helpful to improve the features and must be continuous layer by layer. Most deep learning methods fail in the training step because of the discontinuous loss function or motion estimation method. Generally, we add the prior information by adding the constraint to the loss function to improve the method. In the proposed method, a sparse warping loss and a motion consistency loss are used to improve the network.

2.2 Sparse self-learning PWC-Net

2.2.1 *PWC-Net*

PWC-Net is the backbone network of the proposed method. Based on the mainframe, PWC-Net achieves motion information via a convolutional neural network (CNN) [39] feature extractor, cost volume correlation, deconvolution network, and mean-square error (MSE) loss. Figure 2 shows the structure of the feature extractor.

The deep learning method provides new insights into extracting spatial information and color information. According to (8), both FlowNet and PWC-Net proposed two CNNs to extract the features from two frames. The two CNNs shared their weights to obtain the spatial features. Each CNN has four layers and uses small size convolution modules of 3×3. The features of the l-th layer are defined as

$$\rho_i^l(\boldsymbol{x}) = \theta(\rho_i^{l-1}(\boldsymbol{x})), \tag{10}$$

where the subscript denotes the order of the frames, and $\theta(\cdot)$ denotes the convolutions.

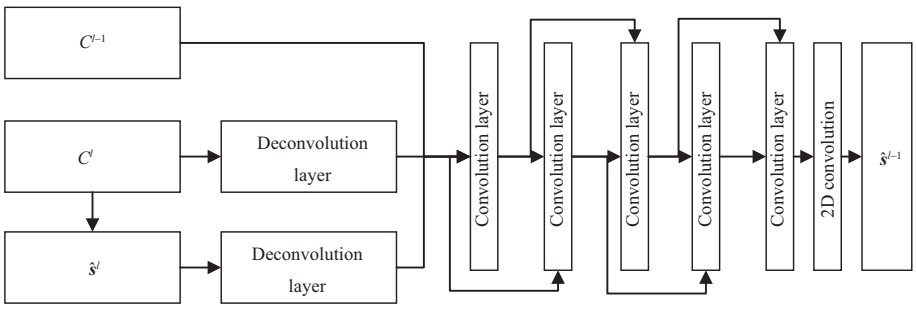


Figure 3 The network to estimate the motion using the correlation pyramid.

Features of CNN are helpful to process the changing backgrounds and filter parts of noise. It is difficult for traditional methods to calculate the long-range motion. A feature pyramid is often used to solve this question. It notes that a CNN has a natural pyramid structure, and PWC-Net utilizes the structure to obtain the motions in multiple resolutions.

The best features should satisfy two conditions. First, the feature extractor should be a one-to-one function in the first frame. Second, the corresponding pixel can be easily found in the second frame. It is impossible to achieve under complex backgrounds. We utilize the neighbor areas and the color space to utilize the spatial information.

Cost volume correlation is a stereo matching idea and is used to create the relationship between two frames. Similar to the method to find the relationship between two frames, it has three ways to achieve the goal: stacking, difference, and correlation. As in the previous description, stacking and difference are improper for calculating the motion due to their discontinuous characteristics. PWC-Net uses the correlation method and achieves better results in the datasets. The cost volume calculates the correlation of each patch of frames by different displacements. There are two advantages to using cost volume: the first advantage is that it can be rewritten to convolutional form. As we know, the convolution layer is convenient to calculate for GPU. The second advantage is that the cost volume is continuous, which can give feedback to the features. The definition of the correlation cost volume is as follows:

$$C^l(\boldsymbol{d}, \boldsymbol{x}) = \frac{1}{N}\langle \rho_1^l(\boldsymbol{x}), \rho_2^l(\boldsymbol{x} - \boldsymbol{d})\rangle, \tag{11}$$

where $\boldsymbol{d}$ is the maximum search area. We arrange different $\boldsymbol{d}$ values to the feature channels to calculate the motion.

Estimating the motion from the cost volume can be difficult to understand. Unlike (5), PWC-Net uses deconvolution modules to restore the resolution of feature maps. The optical estimation can be regarded as a regression question. The other challenge is to combine the different resolutions to estimate long-range motion. PWC-Net up-samples the low-resolution features and adds them to the high-resolution features to restore the complete motion. The low-resolution features calculate the coarse motions; the high-resolution features calculate the residual motions by the warping images. The warping features first warp the features of the first frame to a predicted frame by (4) and calculate the cost volume with the features of the second image. The PWC-Net estimates the motion by the coarse-to-fine pyramid structure and the cost volume of different levels. Figure 3 shows the network used to estimate the motion using the correlation pyramid. First, two deconvolution layers are used to restore the resolution from the high-level motion and cost volume. Then they are stacked together with the low-level cost volume to calculate the low-level motion by five convolution layers. Finally, a 2-D convolution layer is used to reduce the band number to match the size of the motion.

It is easy to define the loss function if we have the ground truth. The loss function is defined as (9). Alternatively, we can define the loss function in different pyramid leaves.

$$\mathcal{L}(\theta) = \sum_l \Psi(\hat{\boldsymbol{s}}^l(\theta) - \boldsymbol{s}^l), \tag{12}$$

where $\Psi(\cdot)$ is the measure function. The commonly used method is the l2-norm function. PWC-Net achieved good results under complex background, however, this method fails to process small targets. Selecting the neighbor of a pixel to extract the features is difficult. The features cannot fit the first condition when the neighbor area is small; in contrary, it cannot fit the second condition when it is large.

It determines that designing a network to detect the targets and then calculate the motion is impossible on the satellite videos.

2.2.2 *Sparse self-learning*

Labeling the ground truth is a challenge for all datasets, especially satellite videos. The massive numbers of targets are hard to distinguish and the noise is strong. However, the deep learning method requires massive amounts of training samples to obtain high-quality features. Self-learning is the path to solve this challenge. SelFlow and UnFlow are the typical methods to calculate motions using self-learning.

The core question of the self-learning methods is how to train the networks without using the ground truth. In other words, we should redefine the loss function of (12). The Lukas-Kanade [40] method calculates the loss by

$$\mathcal{L}_u(\theta) = \|\boldsymbol{I}(\boldsymbol{x}, t+\Delta t) - \boldsymbol{I}(\boldsymbol{x} + \hat{\boldsymbol{s}}(\boldsymbol{x}, t; \theta), t)\|, \tag{13}$$

where $\hat{\boldsymbol{s}}(\boldsymbol{x}, t; \theta)$ is the predicted motion. Improvements such as the constraints should be added to the loss function. Similar to SelFlow, the proposed method uses a warping function to achieve $\boldsymbol{I}(\boldsymbol{x} + \hat{\boldsymbol{s}}(\boldsymbol{x}, t; \theta), t)$.

Traditional methods always restrain the motion to smooth using a l2-norm, but it will against the feature extraction of small targets. We observed that the small targets are sparsely distributed in a frame of satellite video scenes, which inspire us to redefine the warping loss by the sparsity constraint.

$$\mathcal{L}_s(\theta) = \|\boldsymbol{I}(\boldsymbol{x}, t+\Delta t) - \boldsymbol{I}(\boldsymbol{x} + \hat{\boldsymbol{s}}(\boldsymbol{x}, t; \theta), t)\| + \alpha|\hat{\boldsymbol{s}}(\boldsymbol{x}, t; \theta)|_1. \tag{14}$$

l1-Norm is used to constrain the motion to be sparse. The sparsity constraint limits the correlation features so that it enhances the salience of the small targets.

2.3 Multiframe framework

To solve the estimation of the blurry target motion question, the main topic of the proposed method is the use of multi-frame information. We need a long time to find and track the objects when we observe blurry moving objects. Similar to a human being, multiple frames help to obtain accurate motion information. Limited by the dataset, most methods use multi-frame information to solve occlusion questions. PWC-Net trains a linear model to fuse the motions of different frames. According to (4), SelFlow achieves an excellent result by using the forward cost volume and backward cost volume to restore and fuse the motions. Cycle consistency [41–43] is a better method to fuse the multi-frame motions. It calculates the backward and forward motions in different time scales and then fuses them. These methods can achieve better results in common videos, but they are not appropriate for obtaining the motions from satellite videos that contain massive numbers of blurry targets and noisy backgrounds.

The simple fusing of the motions is not proper for small targets. We note that the motion of the targets in satellite videos is approximately uniform motion, and there exists a linear relationship between the multi-frame motions in the time axis. We define a motion consistency instead of the time consistency to solve this problem.

$$\boldsymbol{I}(\boldsymbol{x}, t+\lambda\Delta t) = \mathcal{F}(\boldsymbol{x} + \lambda\boldsymbol{s}(\boldsymbol{x}, t)). \tag{15}$$

Considering the noise $\mathcal{N}$ and the motions $\boldsymbol{s}_b(\boldsymbol{x}, t)$ caused by a complex background, the predicted motion can be rewritten as

$$\lambda\hat{\boldsymbol{s}}(\boldsymbol{x}, t) = \lambda\boldsymbol{s}(\boldsymbol{x}, t) + \mathcal{N} + d_b(\boldsymbol{x}, t). \tag{16}$$

Obviously, $\lambda\hat{\boldsymbol{s}}(\boldsymbol{x}, t)$ is a linear relationship by different λ. $\lambda\hat{\boldsymbol{s}}(\boldsymbol{x}, t)$ is calculated by $\boldsymbol{I}(\boldsymbol{x}, t+\lambda\Delta t)$ and $\boldsymbol{I}(\boldsymbol{x}, t)$. n is the frame number of the multi-frame method ($n > 2$). Consider multiple frames and different λ:

$$\begin{cases} \hat{\boldsymbol{s}}_1(\boldsymbol{x}, t) = \boldsymbol{s}(\boldsymbol{x}, t) + \mathcal{N} + \boldsymbol{s}_b(\boldsymbol{x}, t), \\ \hat{\boldsymbol{s}}_2(\boldsymbol{x}, t) = \boldsymbol{s}(\boldsymbol{x}, t) + \dfrac{1}{2}\mathcal{N} + \dfrac{1}{2}\boldsymbol{s}_b(\boldsymbol{x}, t), \\ \cdots \\ \hat{\boldsymbol{s}}_{n-1}(\boldsymbol{x}, t) = \boldsymbol{s}(\boldsymbol{x}, t) + \dfrac{1}{n-1}\mathcal{N} + \dfrac{1}{n-1}\boldsymbol{s}_b(\boldsymbol{x}, t). \end{cases} \tag{17}$$

We define motion consistency loss $\mathcal{L}_M$ as

$$\mathcal{L}_i = \Psi(\hat{\boldsymbol{s}}_{i-1} - \hat{\boldsymbol{s}}_{i-2}), \tag{18}$$

$$\mathcal{L}_M = \sum_{i=2}^{n} \mathcal{L}_i, \tag{19}$$

where $\mathcal{L}_i$ is the residual noises. We minimize $\mathcal{L}_i$ to filter the noise and make the $\boldsymbol{s}(\boldsymbol{x}, t)$ significant.

2.4 MSSPWC-Net

Combining with the sparse self-learning PWC-Net and multi-frame framework, we finally calculate the motion flow by

$$\hat{\boldsymbol{s}}(\boldsymbol{x}, t) = \frac{2}{(n-1)(n-2)} \sum_{i=2}^{n} ((i-1)\hat{\boldsymbol{s}}_{i-1}(\boldsymbol{x}, t) - \hat{\boldsymbol{s}}_1(\boldsymbol{x}, t)). \tag{20}$$

Through the loss function, we can adjust the feature to minimize $\mathcal{N}$ and $\boldsymbol{s}_b(\boldsymbol{x}, t)$ and thus obtain better features of small blurry targets. We redefine the entire loss function to train the whole network by

$$\mathcal{L} = \mu\mathcal{L}_s + (1-\mu)\mathcal{L}_M, \tag{21}$$

where $\mu \in (0, 1)$ is a hyperparameter. The proposed MSSPWC-Net solved the problems of estimating small blurry targets' motions.

3 Experiments

We design experiments to prove the effectiveness of the proposed method, including (a) the accuracy of the sparse self-learning method, and (b) the effect of the multi-frame framework. In this section, we first introduce the setting of the experiments. Then, we propose the subjective and objective results compared with the baseline methods and illustrate the effectiveness of the proposed methods.

3.1 Experimental settings

Labeling the motion of satellite video targets is difficult. Because of the reason of small blurry target challenges, artificial segmentation labels are not sufficiently accurate to evaluate the effectiveness of the methods. We use the warping loss to approximate the error. Similar to the classification, we train the network on some of the frames and test the results on other frames. The consistent results of the training set and testing set illustrate the effectiveness of the features via the warping loss.

The traditional deep learning methods are tested in the experiments. FlowNet and PWC-Net are the baseline methods that estimate the motion using deep features. SelFlow and UnFlow are self-learning methods. Both PWC-Net and SelFlow use multi-frame fusion to obtain accurate motions, and we compare the multi-frame method with them.

All of the videos are taken by the Jilin-1 video satellite. The spatial resolution is approximately 1 m and the frame rate is 10. We choose the city scenes that contain massive numbers of targets because they are the regions of interest. We only label the motion direction of the satellite videos. The direction color can refer to Figure 4. We use flow field color coding [44] to visualize the motion. The direction of the vector whose center points to the color point indicates the moving direction of the target, and the distance from the color point to the center point indicates the moving speed of the target. The data examples are shown in Figure 5.

The proposed method, SelFlow, and UnFlow all use PWC-Net as the backbone network. We use the same structure to illustrate the effectiveness. The layer of the feature pyramid is 6, and lr = 0.01. The hyperparameter α is set to 0.3 and μ is set to 0.5.

3.2 Results

The warping loss of the compared methods is shown in Table 1. It shows that the proposed method using four frames obtained the best results. The results of PWC-Net are better than those of FlowNet illustrating that the features of PWC-Net are more effective than those of FlowNet.

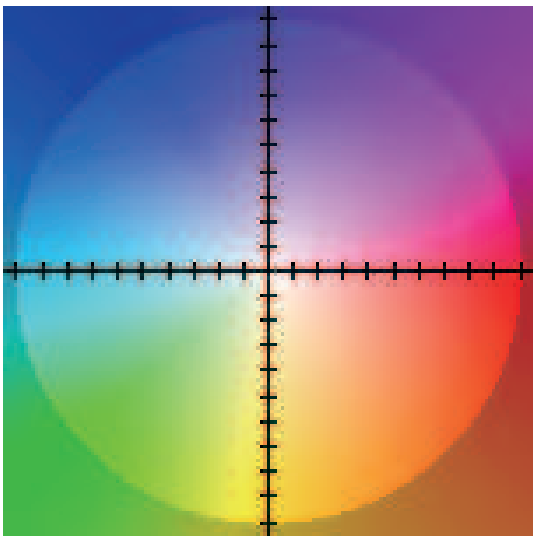


Figure 4 (Color online) The flow field color coding.

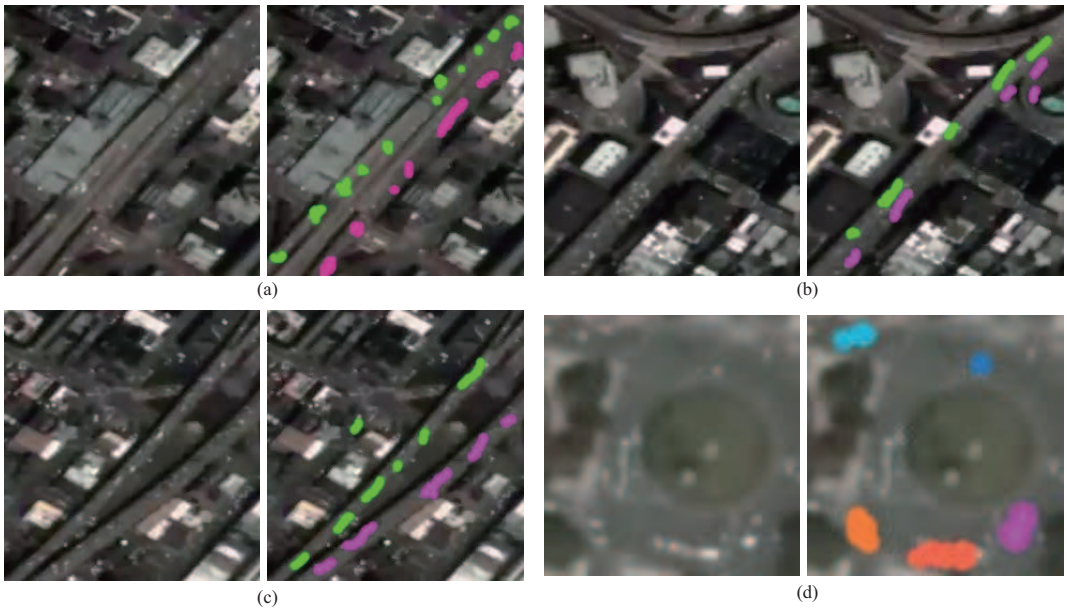


Figure 5 (Color online) The data samples. (a)–(d) Different videos (left) and the truth optical flow of the videos (right). (a) Video 1, (b) video 2, (c) video 3, (d) video 4.

Table 1 The warping loss of compared methods

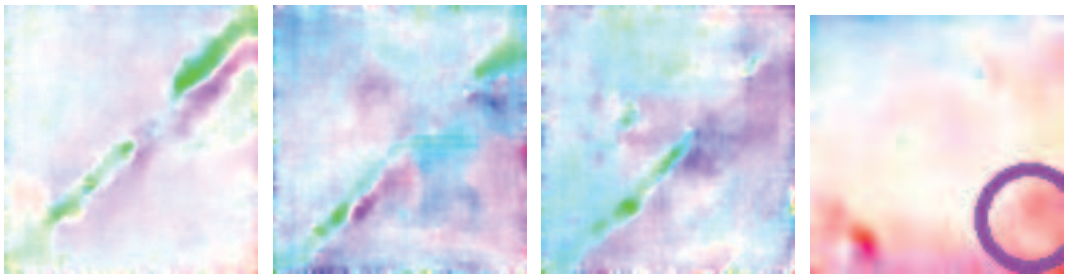
	(a) video dataset 1	(b) video dataset 2	(c) video dataset 3	(d) video dataset 4
FlowNet	6.56	7.36	7.63	12.78
PWC-Net	6.48	7.28	7.56	13.14
UnFlow	6.20	7.13	7.45	12.83
SelFlow	6.27	7.15	7.47	12.81
Multi-frame PWC-Net	6.47	7.33	7.53	12.91
SSPWC-Net	6.32	7.17	7.48	12.77
MSSPWC-Net	6.19	7.12	7.41	12.71

3.2.1 *Results of spares self-learning*

Figure 6 shows the subjective results of FlowNet. Figure 7 shows the subjective results of PWC-Net. FlowNet got smooth and low noise results but failed to extract the motions which are marked by red rectangles. PWC-Net successfully extracted the motions, but it regarded the motions of group targets as

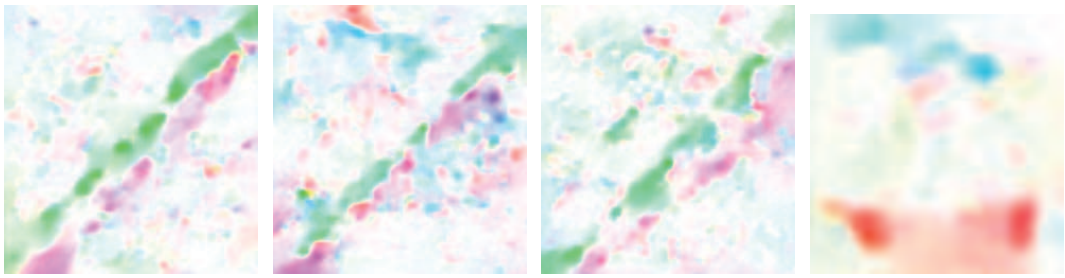
(a) (b) (c) (d)

Figure 6 (Color online) The results of FlowNet. (a) Video 1, (b) video 2, (c) video 3, (d) video 4.



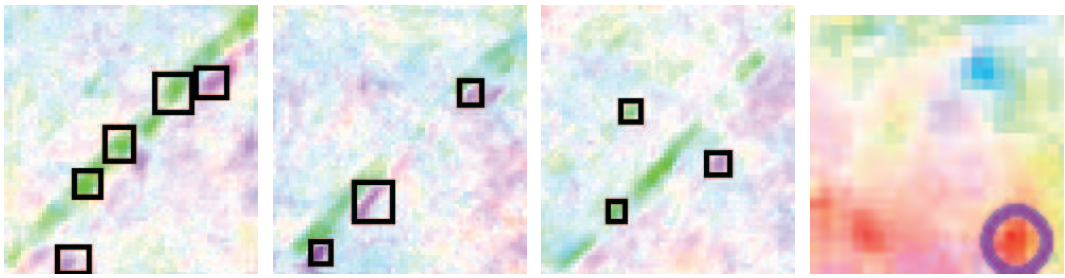
(a) (b) (c) (d)

Figure 7 (Color online) The results of PWC-Net. (a) Video 1, (b) video 2, (c) video 3, (d) video 4.



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

Figure 8 (Color online) The results of UnFlow. (a) Video 1, (b) video 2, (c) video 3, (d) video 4.

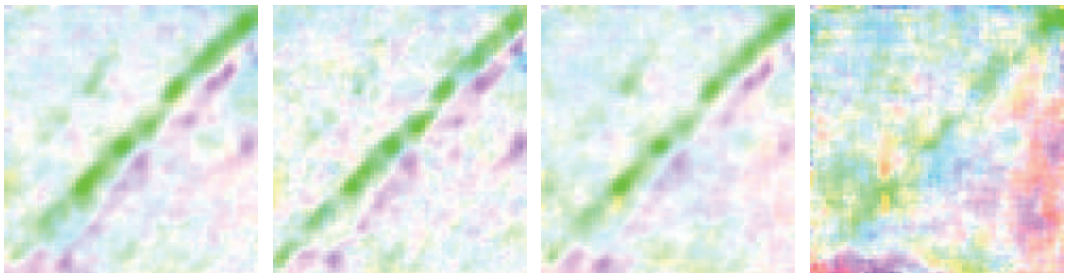


(a) (b) (c) (d)

Figure 9 (Color online) The results of SSPWC-Net. (a) Video 1, (b) video 2, (c) video 3, (d) video 4.

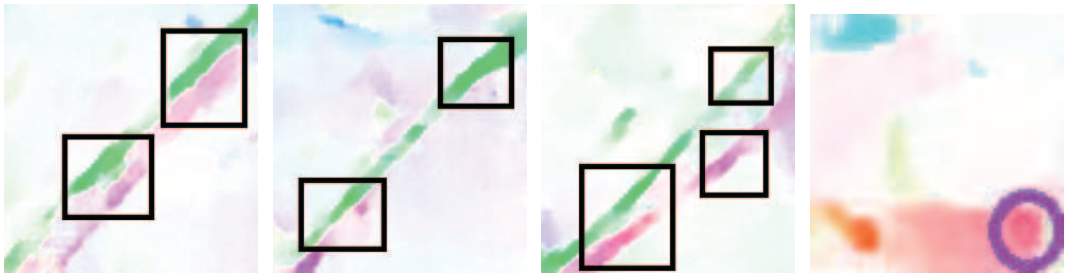
a single motion.

UnFlow designed a bidirectional census loss to adapt the dataset. The results of UnFlow are shown in Figure 8. UnFlow got better results for small targets but it is strongly influenced by the noise. Figure 9 shows the subjective results of the SSPWC-Net. The proposed SSPWC-Net got sparse results, it succeed in departing the small targets as the black rectangles show. The results illustrate the effectiveness of our sparse warping loss. The proposed SSPWC-Net achieved more accurate segmentation results. It shows more detail of the small objects than the other methods. The advantage of more accurate segmentation results is that it is helpful for target detection and tracking.



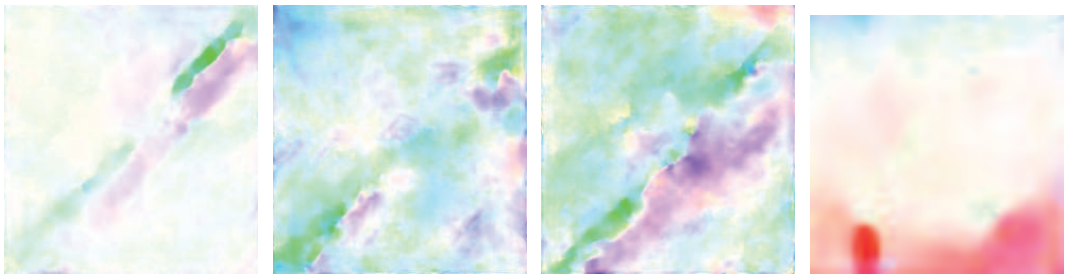
(a) (b) (c) (d)

Figure 10 (Color online) The results of different sparse coefficients. (a) $\alpha = 0.1$; (b) $\alpha = 0.5$; (c) $\alpha = 0.7$; (d) $\alpha = 0.9$.



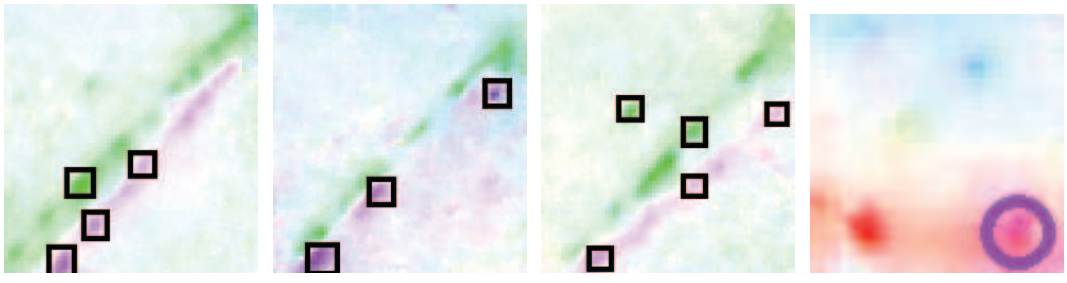
(a) (b) (c) (d)

Figure 11 (Color online) The results of SelFlow. (a) Video 1, (b) video 2, (c) video 3, (d) video 4.



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

Figure 12 (Color online) The results of the PWC-Net using three frames. (a) Video 1, (b) video 2, (c) video 3, (d) video 4.



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

Figure 13 (Color online) The results of MSSPWC-Net. (a) Video 1, (b) video 2, (c) video 3, (d) video 4.

Figure 10 shows the results of different sparse coefficients. The selection of sparsity coefficient α depends on the sparsity of small targets. The larger the target, the smaller the sparsity of the target distribution. The smaller the target, the larger the sparsity parameter to ensure the sparsity of the results. $\alpha = 0.3$ is the best choice.

3.2.2 *Results of multi-frame learning*

Figure 11 shows the subjective results of SelFlow. Figure 12 shows the fusion results of the multi-frame PWC-Net. Figure 13 shows the subjective results of the MSSPWC-Net. SelFlow, which improves the

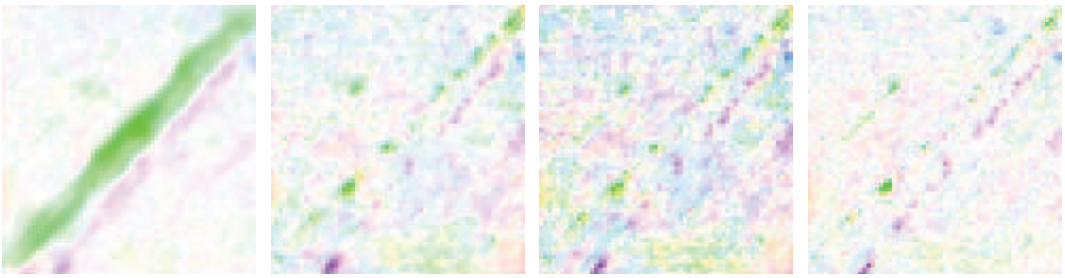


Figure 14 (Color online) The results of different number of frames. (a) $n = 3$; (b) $n = 5$; (c) $n = 6$; (d) $n = 7$.

UnFlow by PWC-Net and multi-frames using the time consistency got clear results, but SelFlow got the inaccurate speed of the motions. PWC-Net attempts to use linear fusion, but it failed to estimate the motions of blurry targets. As the purple circle marked, all the methods except the MSSPWC-Net predicted the wrong direction of this target. The results illustrate the effectiveness of the motion consistency. Accurate speed of motion can help us understand the status of the scenes better.

Figure 14 shows the results of different sparse coefficients. The selection of the sparsity coefficient depends on the sparsity of small targets. The larger the target, the smaller the sparsity of the target distribution. The smaller the target, the larger the sparsity parameter to ensure the sparsity of the results. $\alpha = 0.3$ is the best choice.

The computation of SSPWC-Net proposed in this paper is the same as that of PWC-Net. Depending on the number of frame n of the multi-frame learning, the efficiency of MSSPWC-Net is about $1/(n-1)$ of PWC-Net.

4 Conclusion

This study sets out to estimate the motion of unlabeled satellite video targets using the deep learning method. The proposed method combines the sparse self-learning method and multi-frame framework to estimate the motions of small blurry targets under complex satellite video scenes. The experimental results show that the MSSPWC-Net achieves the best results of small blurry targets under complex satellite video scenes. The results also prove the effectiveness of the proposed sparse self-learning method and motion consistency of the multi-frame framework. The present study will serve as a base feature for future scene understanding studies of satellite videos.

Being limited to the deep features, the proposed method must train the network to fine-tune the features to obtain accurate results. It is recommended that further research be undertaken in the following areas: (a) improving the segmentation results by sparse prior constraints and (b) improving the cost volume to obtain more accurate small target information. It would be interesting to apply the method for the applications.

Acknowledgements This work was supported in part by National Natural Science Foundation of Key International Cooperation (Grant No. 61720106002) and National Natural Science Foundation for Outstanding Scholars (Grant No. 62025107).

References

1 Stiller C. Object-based estimation of dense motion fields. IEEE Trans Image Process, 1997, 6: 234–250
2 Oh T H, Jaroensri R, Kim C, et al. Learning-based video motion magnification. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2018
3 Zhu J S, Sun K, Jia S, et al. Urban traffic density estimation based on ultrahigh-resolution UAV video and deep neural network. IEEE J Sel Top Appl Earth Observations Remote Sens, 2018, 11: 4968–4981
4 Gupta M, Baireddy S, Comer M L, et al. Small target detection using optical flow. In: Proceedings of IEEE Aerospace Conference, 2021. 1–9
5 Du B, Cai S H, Wu C. Object tracking in satellite videos based on a multiframe optical flow tracker. IEEE J Sel Top Appl Earth Observations Remote Sens, 2019, 12: 3043–3055
6 Xuan S Y, Li S Y, Zhao Z F, et al. Rotation adaptive correlation filter for moving object tracking in satellite videos. Neurocomputing, 2021, 438: 94–106
7 Xuan S Y, Li S Y, Han M F, et al. Object tracking in satellite videos by improved correlation filters with motion estimations. IEEE Trans Geosci Remote Sens, 2020, 58: 1074–1086
8 Memin E, Perez P. Dense estimation and object-based segmentation of the optical flow with robust techniques. IEEE Trans Image Process, 1998, 7: 703–719
9 Liu H, Gu Y F, Wang T F, et al. Satellite video super-resolution based on adaptively spatiotemporal neighbors and nonlocal similarity regularization. IEEE Trans Geosci Remote Sens, 2020, 58: 8372–8383

10 Tanaka M, Yaguchi Y, Okutomi M. Robust and accurate estimation of multiple motions for whole-image super-resolution. In: Proceedings of IEEE International Conference on Image Processing, San Diego, 2008. 649–652
11 Dai W, Chen Y M, Huang C, et al. Two-stream convolution neural network with video-stream for action recognition. In: Proceedings of International Joint Conference on Neural Networks, Budapest, 2019. 1–8
12 Jin P, Mou L C, Hua Y S, et al. FuTH-Net: fusing temporal relations and holistic features for aerial video classification. IEEE Trans Geosci Remote Sens, 2022, 60: 1–13
13 Yin Z Y, Tang Y Q. Analysis of traffic flow in urban area for satellite video. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, 2020. 2898–2901
14 Bao L C, Yang Q L, Jin H L. Fast edge-preserving PatchMatch for large displacement optical flow. IEEE Trans Image Process, 2014, 23: 4996–5006
15 Sun D Q, Yang X D, Liu M Y, et al. Models matter, so does training: an empirical study of CNNs for optical flow estimation. IEEE Trans Pattern Anal Mach Intell, 2020, 42: 1408–1423
16 Revaud J, Weinzaepfel P, Harchaoui Z, et al. EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 1164–1172
17 Sun D, Roth S, Lewis J P, et al. Learning optical flow. In: Proceedings of European Conference on Computer Vision. Berlin: Springer, 2008. 83–97
18 Liu P P, King I, Lyu M R, et al. DDFlow: learning optical flow with unlabeled data distillation. In: Proceedings of American Association for Artificial Intelligence, Hawaii, 2019
19 Dosovitskiy A, Fischer P, Ilg E, et al. FlowNet: learning optical flow with convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 2758–2766
20 Ilg E, Mayer N, Saikia T, et al. FlowNet2.0: evolution of optical flow estimation with deep networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 1647–1655
21 Sun D Q, Yang X D, Liu M Y, et al. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 8934–8943
22 Lim J H, Choi H, Park J C, et al. Learning spatio-temporally invariant representations from video. In: Proceedings of International Joint Conference on Neural Networks, Brisbane, 2012. 1–6
23 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 640–651
24 Bai M, Luo W, Kundu K, et al. Exploiting semantic information and deep matching for optical flow. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2016
25 Jia X, Ranftl R, Koltun V. Accurate optical flow via direct cost volume processing. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 5807–5815
26 Žbontar J, LeCun Y. Computing the stereo matching cost with a convolutional neural network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 1592–1599
27 Hosni A, Rhemann C, Bleyer M, et al. Fast cost-volume filtering for visual correspondence and beyond. IEEE Trans Pattern Anal Mach Intell, 2013, 35: 504–511
28 Ren Z Z, Lee Y J. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 762–771
29 Wang X L, He K M, Gupta A. Transitive invariance for self-supervised visual representation learning. In: Proceedings of IEEE International Conference on Computer Vision, Salt Lake City, 2017. 1338–1347
30 Doersch C, Zisserman A. Multi-task self-supervised visual learning. In: Proceedings of IEEE International Conference on Computer Vision, Venice, 2017. 2070–2079
31 Liu P P, Lyu M, King I, et al. SelFlow: self-supervised learning of optical flow. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 4566–4575
32 Meister S, Hur J, Roth S. UnFlow: unsupervised learning of optical flow with a bidirectional census loss. In: Proceedings of American Association for Artificial, New Orleans, 2018
33 Brox T, Bruhn A, Papenberg N. High accuracy optical flow estimation based on a theory for warping. In: Proceedings of European Conference on Computer Vision, Berlin, 2004
34 Dosovitskiy A, Fischer P, Springenberg J T, et al. Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Trans Pattern Anal Mach Intell, 2016, 38: 1734–1747
35 Jia S, Jiang S G, Lin Z J, et al. A semisupervised siamese network for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2022, 60: 1–17
36 Liu L, Hong D F, Ni L, et al. Multilayer cascade screening strategy for semi-supervised change detection in hyperspectral images. IEEE J Sel Top Appl Earth Observations Remote Sens, 2022, 15: 1926–1940
37 Bergen J R, Burt P J, Hingorani R, et al. A three-frame algorithm for estimating two-component image motion. IEEE Trans Pattern Anal Machine Intell, 1992, 14: 886–896
38 Zeiler M D, Krishnan D, Taylor G W, et al. Deconvolutional networks. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, 2010. 2528–2535
39 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009. 248–255
40 Horn B K P, Schunck B G. Determining optical flow. Artif Intelligence, 1981, 17: 185–203
41 Wang X L, Jabri A, Efros A A. Learning correspondence from the cycle-consistency of time. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 2561–2571
42 Gao L R, Han Z, Hong D F, et al. CyCU-Net: cycle-consistency unmixing network by learning cascaded autoencoders. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
43 Dwibedi D, Aytar Y, Tompson J, et al. Temporal cycle-consistency learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 1801–1810
44 Baker S, Roth S, Scharstein D, et al. A database and evaluation methodology for optical flow. In: Proceedings of IEEE International Conference on Computer Vision, Rio de Janeiro, 2007. 1–8