Multi-motion and Appearance Self-Supervised Moving Object Detection

Fan Yang[1,2*] Srikrishna Karanam[1] Meng Zheng[1]
Terrence Chen[1] Haibin Ling[3] Ziyan Wu[1]
[1]United Imaging Intelligence, Cambridge MA, USA
[2]Temple University, Philadelphia PA, USA
[3]Stony Brook University, Stony Brook NY, USA
{fan.yang03,srikrishna.karanam,terrence.chen,ziyan.wu}@uii-ai.com
hling@cs.stonybrook.edu

Abstract

In this work, we consider the problem of self-supervised Moving Object Detection (MOD) in video, where no ground truth is involved in both training and inference phases. Recently, an adversarial learning framework is proposed [32] to leverage inherent temporal information for MOD. While showing great promising results, it uses single scale temporal information and may meet problems when dealing with a deformable object under multi-scale motion in different parts. Additional challenges can arise from the moving camera, which results in the failure of the motion independence hypothesis and locally independent background motion. To deal with these problems, we propose a Multi-motion and Appearance Self-supervised Network (MASNet) to introduce multi-scale motion information and appearance information of scene for MOD. In particular, a moving object, especially the deformable, usually consists of moving regions at various temporal scales. Introducing multi-scale motion can aggregate these regions to form a more complete detection. Appearance information can serve as another cue for MOD when the motion independence is not reliable and for removing false detection in background caused by locally independent background motion. To encode multi-scale motion and appearance, in MASNet we respectively design a multi-branch flow encoding module and an image inpainter module. The proposed modules and MASNet are extensively evaluated on the DAVIS dataset to demonstrate the effectiveness and superiority to state-of-the-art self-supervised methods.

1. Introduction

Moving object detection (MOD) is a fundamental ability of human visual system, which can be used in a wide range of real-world applications, such as autonomous vehicle, video surveillance, activity recognition, road condition monitoring, airport safety, monitoring of protection along marine border [7]. The works on MOD in video can be roughly divided into three categories in terms of the differences in training and inference protocol: supervised train/unsupervised [1] inference [27, 15, 26, 8], supervised train/ semi-supervised inference [6, 30], unsupervised train/ unsupervised inference [17, 21, 29, 32]. Note that the unsupervised here indicates no any ground truth label and supervised pretrained information (*e.g.,*, model weights trained on ImageNet[11]) are involved in both training and inference phases. Although supervised and semi-supervised methods have achieved excellent performance, they heavily rely on a large amount of fine labeled data, which is scarce and expensive. To alleviate these problems, more and more efforts are spent on how to effectively explore the inherent information of unlabeled data to supervise model training.

It is a considerable challenging task to perform MOD without constraints on the presence of annotation, supervised pretrained model, stationary camera, and clean background. Impressively, Yang *et al.* [32] creatively propose a Contextual Information Separation (CIS) principle to achieve a first fully unsupervised deep neural network based MOD. This framework is built on a hypothesis that moving regions are independent from contextual regions in motion space. In particular, the motion information (*i.e.,*, optical flow) of moving regions cannot be inferred from that of background regions and vice-versa.

While the work is novel and obtains promising results in public dataset, it fails to capture a complete object or to differentiate regions in background (light blue and red boxes in Figure 1). This phenomenon is caused by two reasons: (1) single scale temporal information, and (2) moving camera. Since a moving object can be composed of tem-

*This work was done during the internship of Fan Yang with United Imaging Intelligence.

[1]we use "unsupervised" and "self-supervised" alternatively.

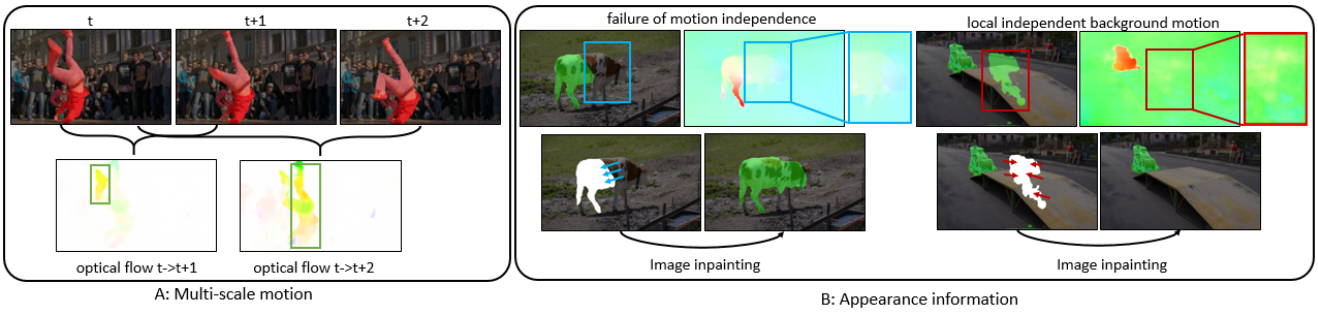

Figure 1. Illustration of multi-scale motion and appearance information. In subfigure **A**, the green boxes represent the different moving regions at different temporal scales. In subfigure **B**, the light blue boxes point to the failed regions of motion independence, the red boxes are the locally motion independent regions in background. The blue and red arrows indicate that in image space the masked regions can be recovered by means of the surrounding contextual information.

poral multi-scale of moving regions, one scale of temporal information can simply encode a fraction of object. For instance, in Figure 1A the optical flow from t to $t+1$ captures the motion of the dancer's left leg (smaller green box); the flow map from t to $t+2$ captures the motion of his right leg and body (larger green box). The optical flow maps of the dancer between various frames capture different moving regions of human body. Therefore, single scale temporal information is insufficient for comprehensive MOD. The joint movement of camera and object leads to the violation of motion independence hypothesis. For example, in the light blue box of Figure 1B the object and background motion are similar. Thus, it is impossible to differentiate object from background in the current framework. In addition, moving camera can yield locally independent moving regions in background. Such regions mislead the model to generate false detection in background (the red box of Figure 1B).

Based on the analysis above, we note that current method fails to encode the multi-scale motion of object and to deal with problems caused by moving camera. To handle these problems, we propose a Multi-motion and Appearance Self-supervised Network (MASNet) to introduce multi-scale temporal information and spatial appearance information into CIS [32] for MOD. Specifically, the proposed MASNet contains two components: 1. a Multi-branch Flow Encoding (MFE) module (dashed box in Figure 2), 2. an IMage InPainter (IMIP) module (solid box in Figure 2). The MFE module takes in multiple optical flow maps simultaneously to encode multi-scale motion information and aggregates detections over different temporal scales to form a final detection. In this way, we can leverage multi-scale temporal information to generate a more comprehensive detection.

The IMIP is designed to tackle the problems caused by camera movement. In particular, when camera is moving, the motion information is not fully reliable. To differentiate object from background and remove false detection in background, the IMIP resorts to spatial appearance information. This is based on a hypothesis that object appearance is different from background. When the detection is in background, the IMIP can leverage the surrounding appearance information to recover it so as to force mask generator to yield better detection ((right Figure 1 B)). Similarly, for missing detection in moving object, IMIP can infer the masked region by the appearance in surrounding regions (left Figure 1 B).

The contributions of the work are summarized into following three-fold:

1. A MFE module is designed to introduce various scales of motion information to aggregate detections over different temporal scale for more complete moving object detection.

2. A IMIP is proposed to leverage spatial appearance information to strengthen the supervision for generator in a novel dimension.

3. Comprehensive experiments are conducted to validate the effectiveness of the proposed MFE and IMIP and to demonstrate the superior performance of our MASNet to state-of-the-art methods.

2. Related work

Classic methods. Due to the importance of moving object detection (MOD), the history of studying this topic can be tracked to long time ago. At early stage, numerous works [1, 28, 31, 10, 14, 5, 16, 25, 34] attempt to perform motion estimation and segmentation together. These methods heavily rely on a good initialization to obtain promising results. To deal with this dependency, Shi *et al.* [24] propose an alternative graph-based method to partition a spatiotemporal volume formed by a video into groups of moving pixel in space and time. In addition, variational infer-

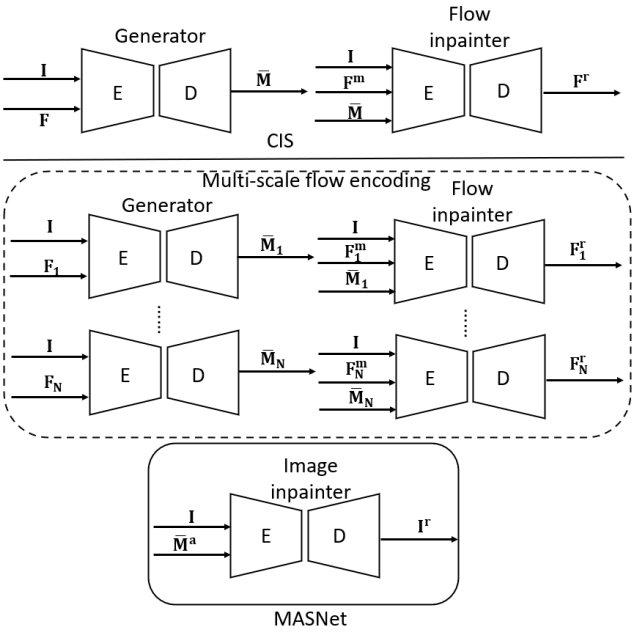

Figure 2. The comparison between CIS [32] (top) and our method (bottom). I, F, and $\bar{M}$ denote image, optical flow map and predicted mask, respectively. N is the index of frame. "r", "a", and "m" represent "recover", "average", and "masked", respectively.

ence based methods [9, 5, 25, 34] are proposed to explicitly model occlusions with piecewise affine regions. Instead of using classical optical flow as motion information, [20] proposes to leverage point trajectories to model long-term of temporal consistency and color constancy, which can span hundreds of frames and is less susceptible to short term variations. This idea has been widely applied in multiple fields: medical image analysis [12], traffic monitoring [3], and crowd motion detection [4].

Supervised methods. Recently, owe to the advance of deep neural network, supervised methods [27, 8, 15, 26] achieve significant performance improvement. In [15], Jain *et al.*propose an end-to-end framework to simultaneously learn appearance and motion information to yield pixel-level segmentation of moving object. Similarly, [8] designs a two-branch network to jointly predict object segmentation and optical flow in an unified framework. In addition to leveraging motion and appearance together, [26] attempts to simply use motion information for moving object detection via learning motion pattern from synthetic data with a fully convolutional network. [27] proposes a two-stream of network to encode spatial and temporal information respectively and a memory module to capture the evolution of object over time.

Self-supervised methods. Although supervised methods can obtain excellent precision in public benchmark datasets, these methods require a large mount of labeled data, which is expensive and time-consuming to collect. Besides, the generalizability of the these methods are limited in unseen data. To deal with the problems, [17] reformulates the point trajectory based moving object segmentation as a minimum clustering problem and proposes a graph construction method to encode a long-term trajectory information. [29] incorporates saliency information of spatiotemporal maps as prior of object to compute the initial foreground region, then appearance model and dynamic model are used to indicate the object in subsequent frames. [21] achieves a fast moving object segmentation in video by yielding a rough estimation first then refining the estimation with a spatio-temporal extension of GrabCut [23]. More recently, [32] proposes a first deep-learning based unsupervised moving object detection method, where an adversarial learning framework is designed to utilize the independence property of object and background in motion field to generate moving regions.

Differences from previous work. To clarify the differences of our work with the closely related work [32], we dedicatedly compare the two methods in two aspects (Figure 2). 1. Instead of involving one scale of temporal information (Figure 2 top), we introduce multi-scale of motion information per iteration in training phase via the proposed multi-branch flow encoding module. 2. In addition to motion information, our method takes appearance information into account as well. Specifically, based on the different characteristics between object and background, an image inpainter module is designed to enhance the supervision signal for generator.

3. Method

The architecture of the proposed moving object detection method, *i.e.*, multi-motion and appearance self-supervised network (MASNet), is shown in Figure 3. The network consists of two components: a multi-branch flow encoding (MFE) module (red box in Figure 3), an image inpainter (IMIP) module (dashed orange box in Figure 3). The MFE module is composed of several pairs of generator and flow inpainter. The generator takes images, I, at time t and its corresponding optical flow maps with frames, $F_1, ..., F_N$, at $t_1, ..., t_N$ as input to encode multi-scale motion information and generate various motion segmentation masks, $\bar{M}_1, ..., \bar{M}_N$. The image I, segmentation mask $\bar{M}$, and masked flow maps F^m are forwarded into the flow inpainters to recover the masked flow maps, F^r. In addition to the flow inpainter for each scale of motion, we introduce an additional flow inpainter to recover average motion over the N frames. The image inpainter is adopted to reconstruct the masked image to provide supervision information from appearance dimension. In subsequent paragraphs, we will introduce these components in detail one by one.

3.1. Adversarial learning

The framework of the adversarial learning for MOD is originally proposed in [32], we here briefly review the es-

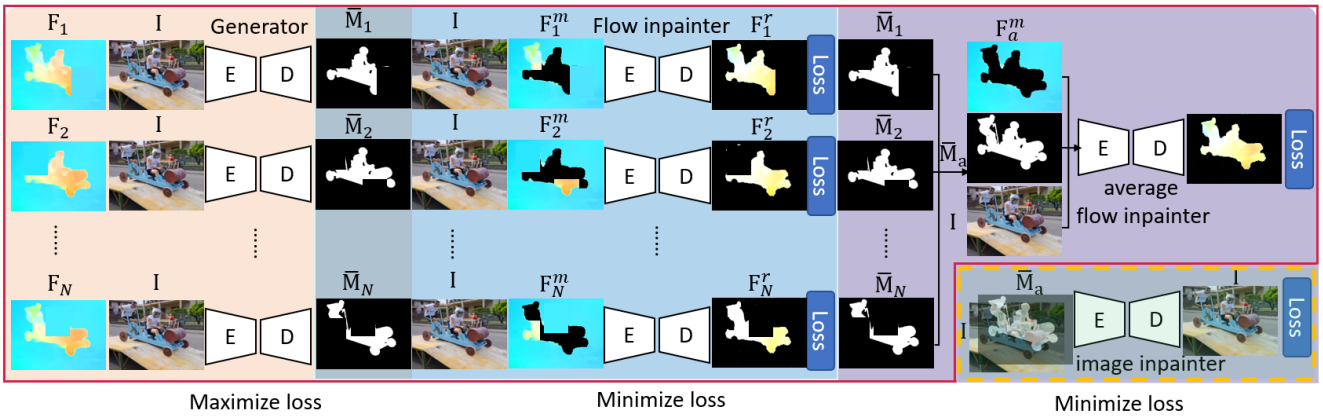

Figure 3. The framework of the proposed motion and appearance self-supervised network (MASNet). The multi-branch flow encoding (MFE) module is enclosed in red box, Image inpainter (IMIP) is in orange dashed box. I, F, and $\bar{M}$ denote image, optical flow map and predicted mask, respectively. N is the index of frame. "r", "a", and "m" represent "recover", "average", and "masked". E: encoder, D: decoder.

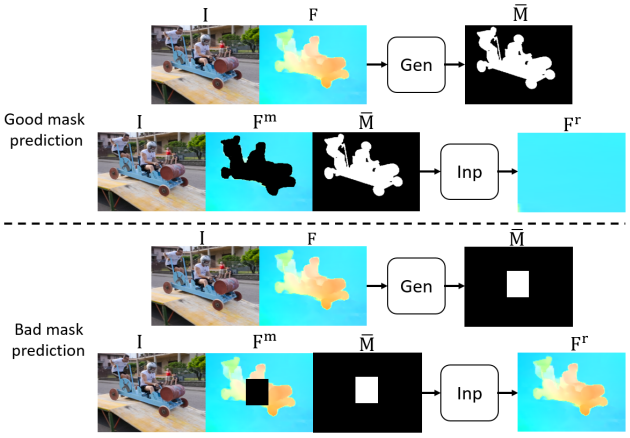

Figure 4. The adversarial learning framework of generator and inpainter. Gen: generator, Inp: inpainter. Top is successful to detect moving object. Bottom is failed to predict moving object,

sential idea. As shown in Figure 4, the basic idea is based on the independence assumption of object and contextual motion. In particular, given the image I and flow map F, the mask generator can be written as $\bar{M} = \mathcal{G}(F, I)$, where $\mathcal{G}(\cdot)$ is the function of generator. The inpainter can be expressed as $F^r = \mathcal{I}(\bar{M}, F^m, I)$, where $\mathcal{I}$ is the function of inpainter, $F^m = F \times (1 - \bar{M})$ ($\times$ denotes element-wise product) is the masked flow map, F_r is recovered flow map. The loss for flow map recovery in predicted mask is represented as $\frac{\|\bar{M} \times (F - F^r)\|_2^2}{\|F_{in}^m\|_2^2}$, where F_{in}^m is the flow regions in mask. Accordingly, the loss for regions outside of predicted mask has similar form, $\frac{\|(1-\bar{M}) \times (F - F^r)\|_2^2}{\|F_{out}^m\|_2^2}$, where F_{out}^m is the flow map regions outside of mask. The variables in loss can be replaced by corresponding functions. Therefore, the final loss function is written as:

$$\mathcal{L}(\mathcal{G}, \mathcal{I}; I) = \frac{\|\mathcal{G}(F, I) \times (F - \mathcal{I}(\mathcal{G}(F, I), F_{out}^m, I))\|_2^2}{\|F_{in}^m\|_2^2} + \frac{\|(1 - \mathcal{G}(F, I)) \times (F - \mathcal{I}(1 - \mathcal{G}(F, I), F_{in}^m, I))\|_2^2}{\|F_{out}^m\|_2^2} \tag{1}$$

The generator is to produce a mask that inpainter cannot infer the motion in the mask region from the contextual motion. However, flow inpainter is to recover the motion map from a masked one. Naturally, this gives rise to a min-max problem:

$$\hat{F} = \arg \min_{\mathcal{I}} \max_{\mathcal{G}} \mathcal{L}(\mathcal{G}, \mathcal{I}; I) \tag{2}$$

3.2. Multi-branch flow encoding (MFE)

In [32], both the generator and inpainter simply leverage optical flow information between the current frame and a randomly chosen noncurrent frame as a primary cue for moving region detection. However, the motion information between two frames encode parts of moving object, especially deformable object. As shown in Figure 5, the moving regions from frame t to $t+1$ are different from that of frame t to $t+2$, this demonstrate the temporal multi-scale motion of moving object. This property is existed in oppose time direction as well.

To encode the motion information, a multi-scale module is designed accordingly. The basic architecture in shown in red box of Figure 3. Formally, given image, I, at frame t, optical flow maps $F_1, ..., F_N$ with respect to frames at $t_1, ..., t_N$, the object function for each branch of generator and flow inpainter can be expressed as:

$$\mathcal{L}^n(\mathcal{G}^n, \mathcal{I}^n; I_t) = \frac{\|\mathcal{G}^n \times (F^n - \mathcal{I}^n(\mathcal{G}^n, F_{out}^m, I_t))\|_2^2}{\|F_{in}^n\|_2^2} + \frac{\|(1 - \mathcal{G}^n) \times (F^n - \mathcal{I}^n(1 - \mathcal{G}^n, F_{in}^n, I_t))\|_2^2}{\|F_{out}\|_2^2} \tag{3}$$

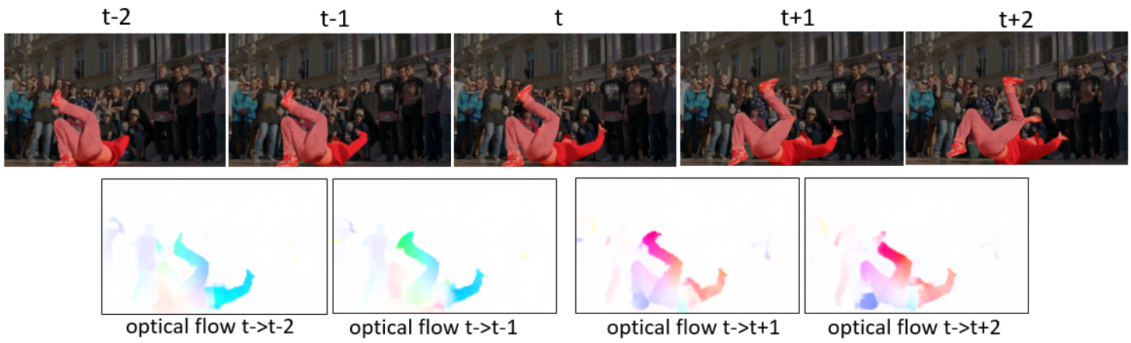

Figure 5. Illustration of temporal multi-scale moving regions in bi-direction. The bi-direction flow maps capture different body parts in different scale motion.

where $n = 1, ..., N$ is the index of branch. The loss function of the whole multi-scale generator and inpainter is written as:

$$\mathcal{L}_{ms}(\mathcal{G}, \mathcal{I}; I_t) = \sum_{n=1}^{N} \mathcal{L}(\mathcal{G}^n, \mathcal{I}^n; I) \tag{4}$$

where N is the total number of optical flow maps. The $\mathcal{G}^n$ and $\mathcal{I}^n$ share weights with other generator and inpainter over branches, respectively.

As shown in Figure 3, the produced moving region segmentation from each branch of generator can be aggregated into a complete segmentation of moving object. Naturally, we can think of using this segmentation mask as a input of an additional flow inpainter, which is to recover the average optical flow map over the N frames. This inpainter can provide extra supervision information for the generators during the training process. The object function of the average flow inpainter and the generators can be written as:

$$\begin{aligned}\mathcal{L}_{avg}(\mathcal{G}^a, \mathcal{I}^a; I) = &\frac{\|\mathcal{G}^a \times (F^a - \mathcal{I}^a(\mathcal{G}^a, F^a_{out^i}, I))\|_2^2}{\|F^a_{in}\|_2^2} \\ &+ \frac{\|(1 - \mathcal{G}^a) \times (F^a - \mathcal{I}^a(1 - \mathcal{G}^a, F^a_{in}, I))\|_2^2}{\|F_{out}\|_2^2}\end{aligned} \tag{5}$$

where 'a' indicates the average over N frames. From the function, we note that based on the adversarial learning framework, the average flow inpainter guides the training of generator in each branch.

3.3. Image inpainter (IMIP)

In addition to the motion information, the appearance information can also serve as a cue to differentiate object from background in video. In particular, the appearance of object is different from its background so that we cannot infer the other when only one of them is accessible. Formally, given a predicted mask $\bar{M}$ and an image I, we can obtain a masked image I_m. Then the masked image is fed into an image inpainter (orange dashed box in Figure 3). The image inpainter and generators construct an adversarial learning relationship to compete each other. Specifically, when the masks produced by generators are not accurate (containing parts of background or foreground), the IMIP can infer the maksed region in image from the surrounding appearance information. The goal of generators is to make the IMIP cannot recover the masked region, such that it needs to yield accurate object mask. The object function of the inpainter is written as:

$$\begin{aligned}\mathcal{L}_{im}(\mathcal{G}^a, \mathcal{I}^{im}; I) = &\|\mathcal{G}^a \times (I - \mathcal{I}(I^m_{out}))\|_2^2 \\ &+ \|(1 - \mathcal{G}^a) \times (I - \mathcal{I}(I^{1-m}_{in}))\|_2^2\end{aligned} \tag{6}$$

where 'a' indicates the average.

3.4. Training of MASNet

The components mentioned above are integrated into the adversarial learning framework, it is well-known that a pair of adversarial network is not easy to train, not to mention training so many adversarial modules in an unified network. To mitigate the difficulty of training process, we propose an alternative training schedule. In specific, we first train the MFE module until it is stable. Then we train the IMIP with the mask from the output of frozen trained generators. After training the IMIP, we continue training MFE with fixed IMIP, so that the generator can be supervised by both motion and appearance information. The finally trained generator is applied in inference to detect moving object.

4. Experiment

4.1. Implementation details

Experiment setting. The proposed method is implemented on deep learning framework TensorFlow [2] and codebase [3] released by [32]. The dataset used for evaluation is Densely Annotated VIdeo Segmentation (DAVIS) [22], where a single moving object exists in most scenarios. The metrics used for fairly quantitative comparison are mean, recall, and standard deviation of Jaccard score $\mathcal{J}$ and F

[2]https://www.tensorflow.org/

[3]https://github.com/antonilo/unsupervised$_detection$

Table 1. The quantitative results of ablation study. MFE represents multi- flow encoding. IMIP is image inpainter, respectively. 'm', 'r', and 's' indicate 'mean', 'recall', and 'std', respectively.

Method	$\mathcal{J}_m \uparrow$	$\mathcal{J}_r \uparrow$	$\mathcal{J}_s \downarrow$	$\mathcal{F}_m \uparrow$	$\mathcal{F}_r \uparrow$	$\mathcal{F}_s \downarrow$
Baseline	59.3	68.0	25.4	60.2	70.9	21.5
+MFE	61.4	74.3	23.3	61.4	72.2	20.5
+MFE+IMIP	63.2	73.9	25.0	64.3	77.1	20.4

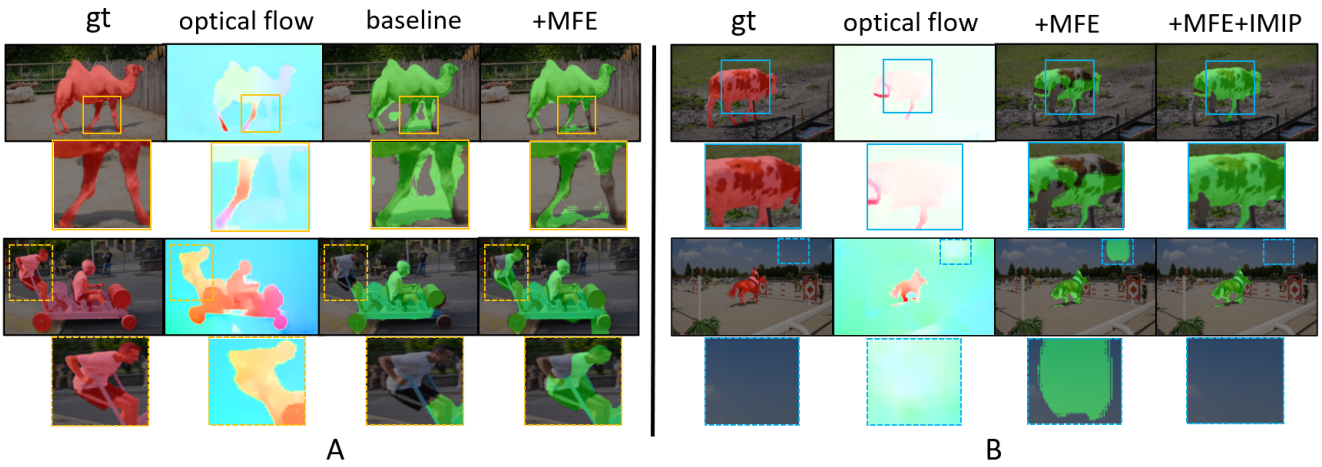

Figure 6. The visualization of ablation study results. A is the comparison between w/o and w/ MFE module. B is the comparison between w/ or w/o IMIP. Yellow solid and dashed boxes indicate MFE is better at capturing motion boundary and various scales of motion, respectively, Light blue solid and dashed boxes demonstrate IMIP can capture relatively static foreground parts and dynamic background parts, respectively.

score $\mathcal{J}$. These two are most commonly adopted in motion segmentation. To fair comparison, we use the sample post-processing protocol as CIS [32]. All experiments are conducted on a RTX 2080 GPU with 12G memory.

Mask generator. As depicted in Figure 3, the generator is constructed by a pair of encoder and decoder, which is a shrunk version of SegNet [2]. The encoder and decoder consists of 5 convolution layers and deconvolution layers, respectively. After each convolution layer, a batch normalization layer is attached. The deconvolution layers are followed by a softmax layer to generate the probability map for each pixel to be classified as moving object and background.

Flow inpainter. The flow inpainter is originally proposed in [33], the structure is shown in Figure 3. The input includes image I, predicted mask M, and masked optical flow map F_m using $\bar{M}$. The M and F_m are concatenated in channel-wise and forwarded into an encoder for feature extraction. The image I is fed into an another encoder to extract feature. The two features are concatenated and forwarded into a decoder for optical flow recovery.

Image inpainter. The architecture of image inpainter is similar with that of flow inpainter, but with a little modification in input layer. The input of image inpainter is image masked with M, I^m. The output is the recovered image, which is used to compute the error with image I.

4.2. Ablation study

To demonstrate the effectiveness of the proposed multi-branch flow encoding (MFE) and image inpainter (IMIP), we thoroughly conduct ablation experiment on DAVIS dataset. The baseline in ablation study is CIS [32]. Without extra notation, results are reported in single scale testing.

The effect of MFE. Compared to baseline, we design a multi-branch network to introduce multi-scale motion information, so that during training and inference phases the network can take advantage of sufficient temporal information to accurately capture motion from different moving regions of object. This can be demonstrated in Table 1 gray row, where we note that by adding MFE the baseline performance is improved by 2.1 and 1.2 points in $\mathcal{J}_m$ and $\mathcal{F}_m$, respectively. In terms of $\mathcal{J}_r$, MFE can boost the baseline by 6.3 points. From the metric $\mathcal{J}_s$ and $\mathcal{F}_s$, we notice that the MFE can also improve the stability of model. To further illustrate the MFE can benefit the accurate motion capture, we display the visualization of detection results and corresponding optotical flow map (Figure 6 A).

The effect of IMIP. In addition to motion information, we introduce appearance information as another supervision signal. The effectiveness of adding appearance infor-

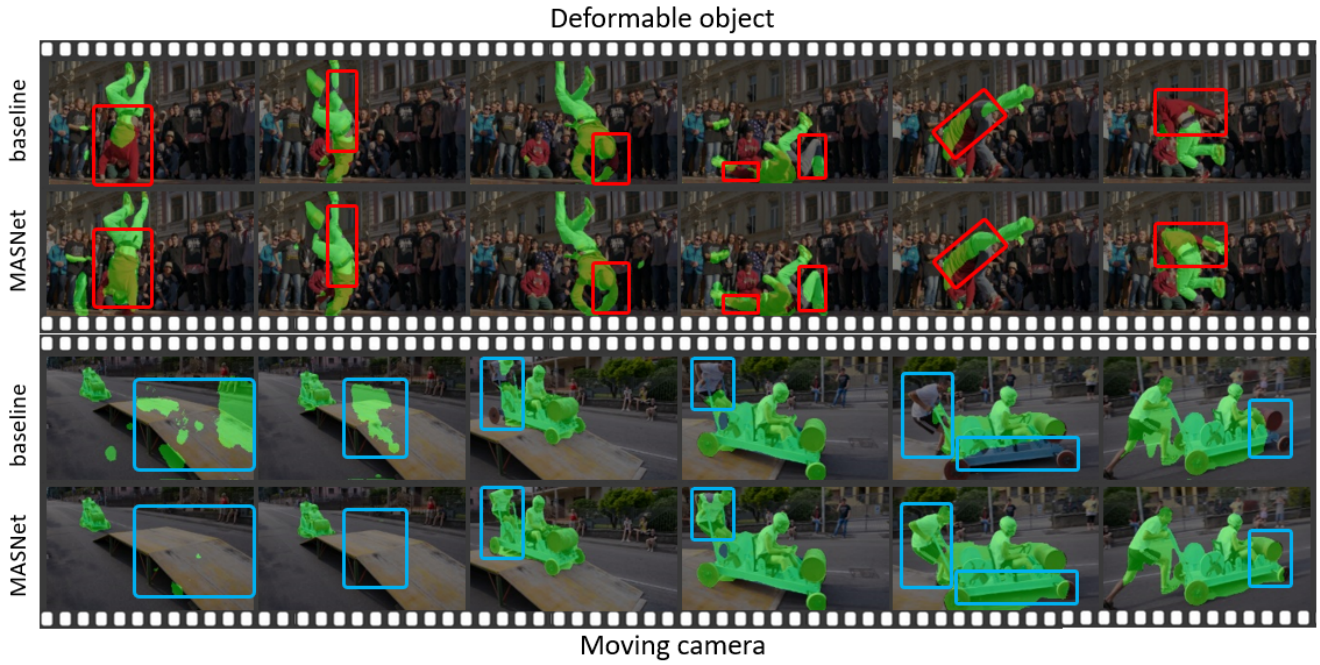

Figure 7. The visualization comparison between the proposed MASNet and baseline (CIS [32]) on representative sequences of deformable object and moving camera. Both are in single-scale testing setting.

Table 2. The quantitative reults on deformable objects in single scale test setting.

Metric	Method	breakdance	camel	cows	dance-twirl	horsejump	soapbox
$\mathcal{J}_m \uparrow$	baseline	61.9	59.1	64.2	72.3	56.9	71.2
	ours	**72.0**	**76.1**	**83.0**	**76.3**	**76.9**	**81.3**
$\mathcal{J}_r \uparrow$	baseline	72.6	62.2	70.2	87.8	64.0	82.8
	ours	**100.0**	**97.8**	**100.0**	**93.3**	**100.0**	**98.0**
$\mathcal{J}_s \downarrow$	baseline	16.0	19.0	21.0	16.6	20.1	19.7
	ours	**10.0**	**9.2**	**3.2**	**11.8**	**8.6**	**10.3**
$\mathcal{F}_m \uparrow$	baseline	66.1	59.5	62.7	76.8	66.3	63.5
	ours	**73.4**	**73.0**	**75.7**	**79.1**	**80.7**	**72.1**
$\mathcal{F}_r \uparrow$	baseline	91.7	72.2	78.8	91.1	84.0	79.8
	ours	**100.0**	**96.7**	**100.0**	**96.7**	**100.0**	**97.0**
$\mathcal{F}_s \downarrow$	baseline	12.5	15.1	13.6	13.0	12.6	18.4
	ours	**9.3**	**9.7**	**6.7**	**12.0**	**6.9**	**11.7**

mation can be observed in Table 1 light blue row, where the model with IMIP is superior to the one without IMIP in almost all metrics. Especially, in $\mathcal{J}_m$, $\mathcal{F}_m$, and $\mathcal{F}_r$, the proposed IMIP increases the performance by 1.8, 2.9, and 4.9 points, respectively. To make it easy to understand the IMIP's functionality, we visually demonstrate the detection results and flow maps in Figure 6 B. In the dashed light blue box, we can note that the the proposed IMIP can successfully depress the false detection in background. This can be attributed to the leverage of appearance information. In particular, when generated mask is in the background, the IMIP can easily recover the image in the mask by using the appearance information from complementary regions (*i.e.,*surrounding background). As the adversary, the mask generator tries to yield mask only covering object. According to the same mechanism, the missing detection parts of object (blue box) can be complemented by generator with the help of IMIP.

The result on deformable object. To further demonstrate the superiority of the our proposed method to baseline, we thoroughly compare the results sequence by sequence. By doing this, we note that on deformable object (Table 2) our method is significantly better than baseline. For instance, in the categories of breakdance, camel, cows, horsejump, and soapbox, our method achieves more than or almost 10 points improvement in terms of $\mathcal{J}_m$ and $\mathcal{F}_m$.

Table 3. The quantitative reults on deformable objects in multi-scale test setting.

Metric	Method	breakdance	camel	cows	dance-twirl	horsejump	soapbox
$\mathcal{J}_m \uparrow$	baseline	61.1	62.2	65.0	70.7	62.9	67.8
	ours	**76.1**	**78.5**	**83.6**	**77.2**	**79.0**	**81.1**
$\mathcal{J}_r \uparrow$	baseline	82.1	73.3	82.7	90.0	88.0	89.9
	ours	**100.0**	**100.0**	**100.0**	**97.8**	**100.0**	**99.0**
$\mathcal{J}_s \downarrow$	baseline	11.7	15.1	15.2	13.2	11.7	22.6
	ours	**8.2**	**6.7**	**3.1**	**10.5**	**8.6**	**9.3**

Table 4. The comparison with state-of-the-art methods. "+" represents multi-scale test. "++" indicates post-processing.

Method	ARP [18]	FTS [21]	NLC [13]	SAGE [29]	CUT [17]	CIS [32]	CIS$_{++}$	MASNet	MASNet$_{+}$	MASNet$_{++}$
$\mathcal{J}_m$	76.2	55.8	55.1	42.6	55.2	59.3	69.4	63.2	64.7	71.0
$\mathcal{J}_r$	91.1	64.9	55.8	N/A	N/A	68.0	82.0	73.9	76.2	82.6
time (s)	74.5	0.5	11.0	0.88	103.0	0.053	4.7	0.053	0.053	4.7

This significant improvement indicates that our proposed MFE and IMIP indeed solve the two problems in baseline method: 1) insufficient utilization of motion, and 2) moving camera. In particular, in these sequences, due to the deformation of object, the different regions of object have different motion information. By introducing multi-scale information in training, model can access richer motion information in each batch which is especially critical for deformable object. In addition, the camera is moved to capture the moving object, this results in the moving background and its motion is not fully independent from object. Thus, the assumption of adversarial learning framework proposed in baseline is not always existed, which makes the motion-only method insufficient to deal with moving object detection problem. However, our method with help of IMIP complements the drawback of baseline and obtains promising improvement.

In Figure 7, we show the comparison of our method and baseline in two representative sequences, where each is dominated by multi-scale deformation or camera movement. From the figure, we note that our proposed method can more accurately detect deformable parts of human (red boxes) and can depress the detections on background (blue boxes). This visually justifies our method can handle the aforementioned problems.

Multi-scale testing. In baseline, during training the input motion map is sampled at various temporal scale, thus theoretically it can also capture multi-scale motion. Thus, if the model is tested by using multi-scale motion, it can also deal with deformable object. In addition, multi-scale motion in test may solve the false detection problem by using the inconsistency of background motion. To demonstrate our method cannot be surpassed by trivial multi-scale test. We compare the proposed method and baseline by averaging detection over five frames in deformable categories (Table 3). We note that our method also outperforms the baseline in all metrics. These results are a compelling evidence to the efficacy of our proposed method.

4.3. Comparison with state-of-the-art

By comparing with other unsupervised methods, our method is considerably better than most of them in terms of speed and precision, except for ARP and CIS. The ARP achieves highest precision (5.2 points higher than ours in $\mathcal{J}_m$), but the speed is almost 16$\times$ slower than ours (74.5s per frame vs. 4.7s per frame). Thus, our method is better than ARP in speed and precision trade-off. Compared with CIS, our method can obtain superior precision in terms of $\mathcal{J}_m$ (63.2 vs. 59.3 and 71.0 vs. 69.4) at the same speed in both single and post-processing. In addition, we note that post-processing significantly increases the precision but meanwhile decreases the detection speed. It is because spatially multi-scale of CRF [19] is adopted to improve detections.

5. Conclusion

This work proposes a MASNet to perform self-supervised moving object detection by introducing multi-scale motion and appearance information via a multi-branch flow encoding (MFE) module and a image inpainter (IMIP) module, respectively. The two modules are designed dedicatedly to deal with single scale of motion and moving camera problems. The MFE aggregates temporally various scales of information to capture the different scales of regions of object in motion space, where single scale motion can simply capture a fraction of object. The moving camera problem leads to unindependent foreground and background motion and locally independent background motion. Thus, pure motion information is not sufficient to differentiate object and background. To handle this problem, the IMIP provides a novel dimension of information, appearance, by encoding spatial difference between object and background. The efficacy of the proposed modules and the superiority of the MASNet are evaluated extensively on DAVIS dataset.

References

[1] Serge Ayer and Harpreet S Sawhney. Layered representation of motion video using robust maximum-likelihood estimation of mixture models and mdl encoding. In *Proceedings of IEEE International Conference on Computer Vision*, pages 777–784. IEEE, 1995.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[3] David Beymer, Philip McLauchlan, Benjamin Coifman, and Jitendra Malik. A real-time computer vision system for measuring traffic parameters. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pages 495–501. IEEE, 1997.

[4] Gabriel J Brostow and Roberto Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 594–601. IEEE, 2006.

[5] Thomas Brox, Andrés Bruhn, and Joachim Weickert. Variational motion segmentation with level sets. In *European Conference on Computer Vision*, pages 471–483. Springer, 2006.

[6] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 221–230, 2017.

[7] Jose M Chaquet, Enrique J Carmona, and Antonio Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–s659, 2013.

[8] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, pages 686–695, 2017.

[9] Daniel Cremers and Stefano Soatto. Motion competition: A variational approach to piecewise parametric motion segmentation. *International Journal of Computer Vision*, 62(3):249–265, 2005.

[10] Trevor Darrell and Alexander Pentland. Robust estimation of a multi-layered motion representation. In *Proceedings of the IEEE Workshop on Visual Motion*, pages 173–174. IEEE Computer Society, 1991.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[12] Ahmed Elnakib, Georgy Gimel'farb, Jasjit S Suri, and Ayman El-Baz. Medical image segmentation: a brief survey. In *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*, pages 1–39. Springer, 2011.

[13] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, volume 2, page 8, 2014.

[14] Steve Hsu, P Anandan, and Shmuel Peleg. Accurate computation of optical flow by using layered motion representations. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 743–746. IEEE, 1994.

[15] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2126. IEEE, 2017.

[16] Allan Jepson and Michael J Black. Mixture models for optical flow computation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 760–761. IEEE, 1993.

[17] Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost multicuts. In *Proceedings of the IEEE international conference on computer vision*, pages 3271–3279, 2015.

[18] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7417–7425. IEEE, 2017.

[19] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24:109–117, 2011.

[20] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200, 2013.

[21] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE international conference on computer vision*, pages 1777–1784, 2013.

[22] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[23] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[24] Jianbo Shi and Jitendra Malik. Motion segmentation and tracking using normalized cuts. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 1154–1160. IEEE, 1998.

[25] Deqing Sun, Jonas Wulff, Erik B Sudderth, Hanspeter Pfister, and Michael J Black. A fully-connected layered model of foreground and background flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2451–2458, 2013.

[26] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3386–3394, 2017.

[27] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In

Proceedings of the IEEE International Conference on Computer Vision, pages 4481–4490, 2017.

[28] John YA Wang and Edward H Adelson. Representing moving images with layers. *IEEE transactions on image processing*, 3(5):625–638, 1994.

[29] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3395–3402, 2015.

[30] Wenguan Wang, Jianbing Shen, Fatih Porikli, and Ruigang Yang. Semi-supervised video object segmentation with super-trajectories. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):985–998, 2018.

[31] Yair Weiss and Edward H Adelson. A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 321–326. IEEE, 1996.

[32] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 879–888, 2019.

[33] Yanchao Yang and Stefano Soatto. Conditional prior networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 271–287, 2018.

[34] Yanchao Yang, Ganesh Sundaramoorthi, and Stefano Soatto. Self-occlusions and disocclusions in causal video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4408–4416, 2015.