SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

July 2024, Vol. 67, Iss. 7, 172104:1–172104:14
https://doi.org/10.1007/s11432-023-4030-y

Towards imbalanced motion: part-decoupling network for video portrait segmentation

Tianshu YU[1], Changqun XIA[2*] & Jia LI[1,2*]

[1]*State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing* 100191*, China;*
[2]*Peng Cheng Laboratory, Shenzhen* 518055*, China*

Received 10 October 2023/Revised 24 January 2024/Accepted 12 April 2024/Published online 25 June 2024

Abstract Video portrait segmentation (VPS), aiming at segmenting prominent foreground portraits from video frames, has received much attention in recent years. However, the simplicity of existing VPS datasets leads to a limitation on extensive research of the task. In this work, we propose a new intricate large-scale multi-scene video portrait segmentation dataset MVPS consisting of 101 video clips in 7 scenario categories, in which 10843 sampled frames are finely annotated at the pixel level. The dataset has diverse scenes and complicated background environments, which is the most complex dataset in VPS to our best knowledge. Through the observation of a large number of videos with portraits during dataset construction, we find that due to the joint structure of the human body, the motion of portraits is part-associated, which leads to the different parts being relatively independent in motion. That is, the motion of different parts of the portraits is imbalanced. Towards this imbalance, an intuitive and reasonable idea is that different motion states in portraits can be better exploited by decoupling the portraits into parts. To achieve this, we propose a part-decoupling network (PDNet) for VPS. Specifically, an inter-frame part-discriminated attention (IPDA) module is proposed which unsupervisedly segments portrait into parts and utilizes different attentiveness on discriminative features specified to each different part. In this way, appropriate attention can be imposed on portrait parts with imbalanced motion to extract part-discriminated correlations, so that the portraits can be segmented more accurately. Experimental results demonstrate that our method achieves leading performance with the comparison to state-of-the-art methods.

Keywords video portrait segmentation, imbalanced motion, unsupervised part decoupling, motion correlation, inter-frame attention

1 Introduction

Video portrait segmentation (VPS) [1], which aims to discover and separate prominent foreground portraits from video frames, has drawn much interest due to a great number of different application scenarios in video creation such as background replacement [2] and portrait transformation [3].

Currently, several video segmentation datasets with portraits have been proposed, like DAVIS [4], PVSD2.5K [1], and PP-HumanSeg14K [5]. In DAVIS, video clips with humans as foreground objects are all captured in outdoor scenes. PVSD2.5K consists of only 2530 annotated frames, and there is only one prominent person in each video clip. PP-HumanSeg14K only contains videos in the remote conference, in which background environments are quite simple. Although recent unsupervised video object segmentation (VOS) methods [6–16] and a specialized VPS method [1] based on existing datasets achieve good performance, due to the uniformity of scenarios and the monotonousness of background environments, the simplicity of these datasets leads to weak robustness of models to handle complex situations in practical applications.

In this work, we propose a new intricate large-scale multi-scene video portrait segmentation dataset MVPS. The dataset consists of 101 video clips in 7 categories of different scenarios including entertainment, indoor handwork, interview, lecture, news, outdoor activity and online shopping which appear

* Corresponding author (email: xiachq@pcl.ac.cn, jiali@buaa.edu.cn)

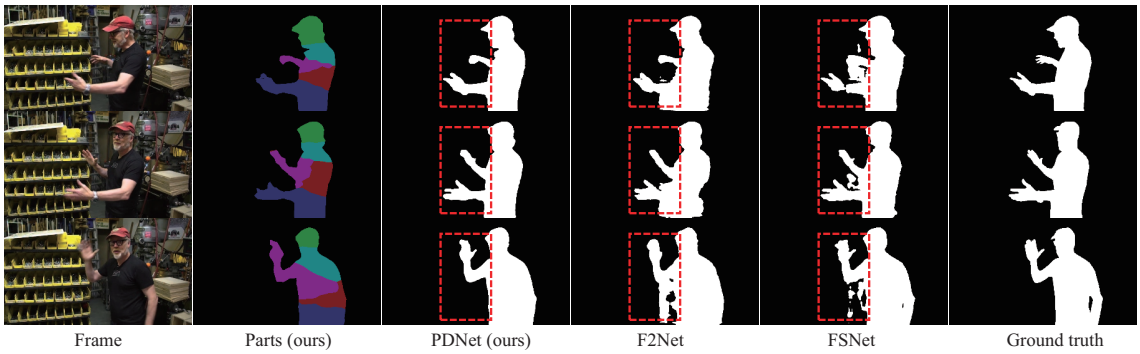

Figure 1 (Color online) Motion imbalance between arms and main body. Due to the joint structure of the human body, different parts in portraits have relatively independent motion states.

frequently on the Internet. Portraits in these scenarios are in different poses and gestures, and background scenes are also diverse. These complications are much closer to practical application scenarios on the Internet. We sample 53923 frames from these video clips in total, where 10843 of them are finely annotated. It is currently the most complex dataset for VPS to our best knowledge.

Based on the observation of a large number of videos with portraits during dataset construction, we find that different from other moving objects like vehicles and airplanes which motion state is consistent for the whole object, the motion of portraits is imbalanced due to the joint structure of the human body. As shown in Figure 1, the motion of the arms in the red box is independent from that of the main body, which leads to inaccurate prediction near the arm with a greater range of motion in existing methods such as F2Net [10] and FSNet [12]. Due to this part-associated imbalance, utilizing the same attentiveness on the motion of different parts may cause imprecise location and segmentation. Since part-discriminated features can be extracted through unsupervised part segmentation [17–22], an intuitive idea is to introduce this discriminated part cue to extract correlations of imbalanced motion.

To this end, we propose a novel part-decoupling network (PDNet) for VPS, which captures the motion correlations of different parts respectively. Specifically, we propose an inter-frame part-discriminated attention (IPDA) module, which decouples imbalanced integral portrait motion into independent part motion. This module unsupervisedly segments portrait parts of both target frame and reference frame, utilizes cross-attention operation between the same part in different frames to obtain part-discriminated motion features, and finally assembles them to generate global motion features based on masks of these parts predicted by the module. The effectiveness of the proposed method is demonstrated in our experiments.

Our major contributions can be summarized as follows:

• We propose a new intricate large-scale dataset MVPS with 7 categories of scenarios, which is the most complex VPS dataset currently.

• We explore the imbalance of portrait motion and propose a PDNet, which leverages the part-associated imbalance of portrait motion and separately captures this imbalanced motion from different parts explicitly.

• We propose an IPDA module to extract discriminative part-level correlations of imbalanced motion.

• Experimental results demonstrate that our method achieves superior performance compared to state-of-the-art methods.

2 Related work

In this paper, we mainly discuss a new VPS dataset and a newly designed VPS method based on the characteristics summarized from the observation during dataset construction. In addition, we use part information as a clue and perceive the imbalanced motion of portraits in videos. We provide an introduction to relevant studies.

2.1 Video portrait datasets

Existing segmentation studies based on images have achieved great success on salient object detection [23–30] and image portrait segmentation [3, 31–33]. With the increasing demand for video applications, VPS

has received much attention by researchers in recent years. Chu et al. [5] proposed a large-scale VPS dataset PP-HumanSeg14K for remote conference scene with 14117 annotated frames, in which portraits are always in half-body and there are usually no changes in shooting distance and angle. Wang et al. [1] introduced a dataset PVSD2.5K for VPS which includes 2530 annotated frames, where there is only one prominent person in each video clip. However, existing datasets only involve some simple scenarios, which leads to poor generalization ability when applied in practical application situations. The lack of datasets with more complex scenarios limits extensive research of the task.

2.2 VPS

VPS can be regarded as a special circumstance of VOS where the object category is fixed as human [1]. As a hot topic, VPS has been studied by many researchers [34–37] and is regarded as an independent task. A recent work [1] studied the VPS task under the settings of unsupervised VOS, which can provide a better assessment of the comprehensive performance of different video clips, so we follow these settings in our work.

As the prominent foreground objects are not marked in VPS, researches have made efforts to locate prominent foreground objects in video frames based on appearance and motion cues. Some studies [38–41] applied recurrent structures to extract temporal-correlated spatial features, which may accumulate interference signals caused by complex backgrounds. Other researches [8,11–15,42–46] utilize optical flow to obtain motion-based temporal features. These methods usually extract spatial and temporal features in two different branches and fuse them, which explicitly exploit appearance and motion cues separately. Some other methods [1, 6, 7, 9, 10, 16, 47–51] establish correlations between features extracted from the target frame and reference frames, which is conducive to enhancing temporal global understanding. However, distinct from other common objects like cars, motion states of different parts of portraits are usually different. Many existing methods regard the prominent object as a whole and extract motion information uniformly, which results in incorrect handling of parts with different motion states from the main body. How to treat parts with different motion states separately is another focus of our work.

Note that portrait matting is another distinct task outputting alpha mattes usually guided by trimap or background image, which focuses on obtaining fine boundaries, while VPS focuses on locating and segmenting portraits with motion continuity [1].

2.3 Unsupervised part segmentation

Unsupervised part segmentation provides a paradigm to represent object semantics robust to environment and object gesture without any part-level mask annotation, which is first proposed in [17]. This work utilizes self-supervision losses based on saliency map to obtain pixel-wise co-part segmentation results among images of a specific object category, which is robust to instance and posture differences. In [18], a method is proposed to learn part-level invariant features between the origin image and spatial-transformed image through image reconstruction to generate parts that are stable to spatial displacement. Recent studies also utilize part semantics generated by unsupervised part segmentation on several different computer vision tasks related to regional property, such as fine-grained recognition [19, 20] and 3D reconstruction [21]. In our work, we utilize unsupervised part segmentation to extract correlations of part-discriminated imbalanced motion for VPS.

Note that the final output of our work is not for the segmentation of human parts but the whole portraits, while the part cue is introduced as a guidance to promote the effectiveness of VPS.

2.4 Video motion perception

By perceiving imbalanced human motion in videos, the accuracy of portrait localization and segmentation can be improved. Some approaches [52–54] suppress background frame signals to enhance the perceptual response to action frames. Some other methods [55, 56] capture representation responses in the temporal domain to detect potential motion in videos. In our work, we achieve the perception of imbalanced motion in videos by extracting part-discriminated motion correlation from decoupled parts.

3 MVPS dataset

In VPS, the richness of background scenes and the complexity of portrait motion in a dataset have a significant impact on the robustness of segmentation models. Existing video segmentation datasets with portraits [1,4,5,57–59] cannot meet the requirements of complexity in scenarios. In order to alleviate the large generalization gap of VPS methods, we construct a new intricate large-scale dataset MVPS with 7 categories of scenarios, which is the most complex VPS dataset available.

3.1 Dataset construction

The diversity of scenes is emphasized during data collection. We first collect initial candidate videos from the public Internet according to the following rules: (1) there is at least one foreground portrait in the video; (2) the number of foreground portraits in videos is different; (3) the background scene of these videos are varied, containing many common indoor and outdoor environments; (4) the videos are clear enough, of which the pixel number on the short side must not be less than 720; (5) videos captured both horizontally and vertically are included. From these candidate videos, we select available video clips that last at least 6 s and include at least one foreground portrait with obvious motion and no camera cut. Finally, we obtain 101 video clips for the MVPS dataset which are divided into 7 categories of scenarios: entertainment, indoor handwork, interview, lecture, news, outdoor activity and online shopping. Then we sample original frames from these video clips at 30 FPS and obtain 53923 frames in total, which are finely annotated with portrait masks at pixel-level every 5 frames by 8 experienced annotators. All the annotations are cross-checked to ensure accuracy. Totally there are 10843 frames annotated in MVPS. For convenience of use, we randomly divide the dataset into a training set and a test set. The training set contains 61 video clips and 6732 annotated frames, while the test set contains 40 video clips and 4111 annotated frames. All frames in MVPS have a high resolution with pixel number of the short side between 720 and 1080. Same as DAVIS [4], we also provide a 480p-resolution version of our MVPS dataset for easier training and evaluation. Note that all videos are available on the public Internet, and our annotations will be publicly available for academic research only.

3.2 Dataset analysis

Scenario distribution. Our MVPS dataset includes videos of 7 scenario categories, where the number of clips in each category distributes quite evenly as shown in Figure 2(a). Examples for each scenario category are shown in Figure 2(b). It can be seen that not only the scenario category has a complex distribution, but the background scenes are also significantly different inside each category. In the category "outdoor activity", for example, there are videos captured in the jungle, on the river boat, and in the snowfield. Meanwhile, both videos shot during the day and night are included. Some videos in the dataset contain subtitle occlusion, which is common in practical application scenarios.

Complexity. To evaluate the complexity of image scenes, we utilize gray level co-occurrence matrix (GLCM) entropy according to [60]. We calculate the average GLCM entropy of frames in each video clip on the typical PP-HumanSeg14K [5] dataset and our MVPS dataset, which frequency distributions are shown in Figure 3. From the comparison results it can be seen that our MVPS dataset has not only a larger mean but also a larger value range of GLCM entropy, while its values in PP-HumanSeg14K are significantly smaller and more concentrated. This indicates the complexity of scenes in our MVPS dataset.

Duration. Duration is also an important factor affecting the complexity of portrait motion. We show the distribution of the number of frames in each video clip in Figure 4. Most of the video clips in MVPS have a number of frames between 300 and 900, corresponding to a time length between 10 and 30 s. In long video clips, there is more complicated portrait motion. We have not selected clips less than 6 s, in which the actions of figures are not abundant enough.

In conclusion, our MVPS dataset contains a variety of intricate scenes and complex portrait motion, which is the most complex dataset in VPS to our best knowledge. It addresses the lack of intricate large-scale multi-scene VPS datasets, which may promote extensive researches on VPS. Our annotations will be publicly available under request for academic research only.

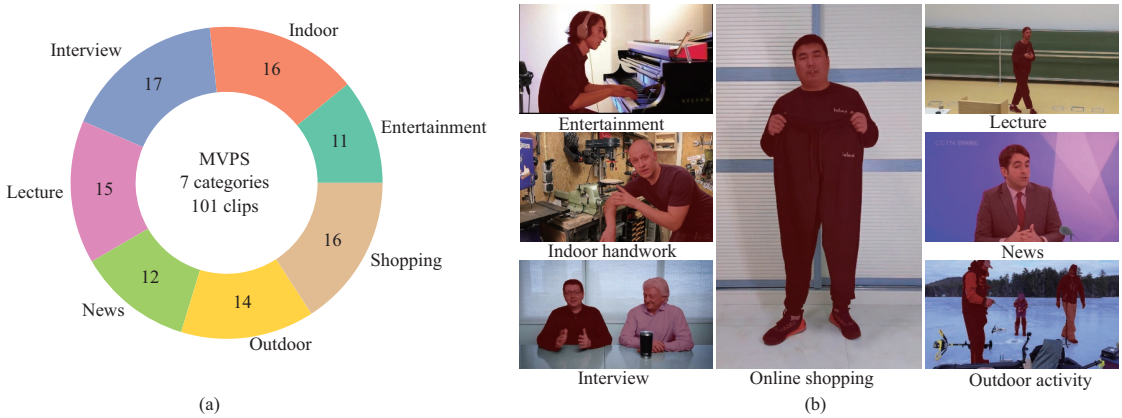

Figure 2 (Color online) Overview of our MVPS dataset. (a) Number of video clips in each of 7 scenario categories; (b) examples for each scenario category.

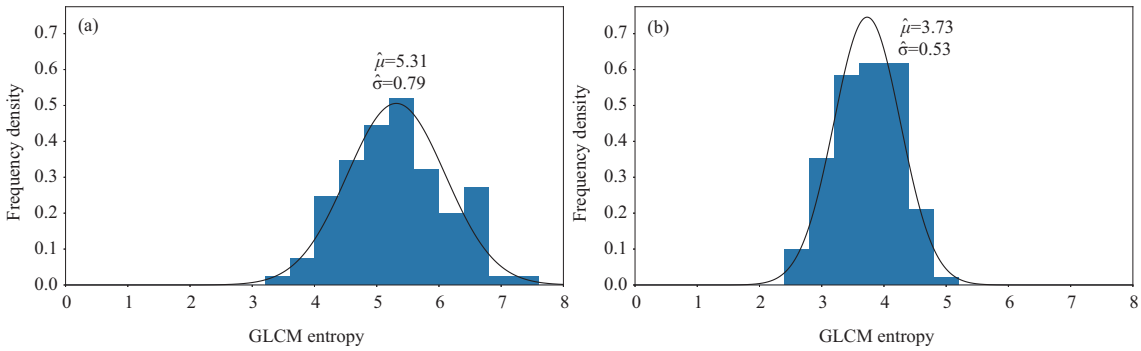

Figure 3 (Color online) Frequency distributions of GLCM entropy of video clips in MVPS and PP-HumanSeg14K [5] datasets. (a) Distribution of GLCM entropy in MVPS dataset; (b) distribution of GLCM entropy in PP-HumanSeg14K dataset.

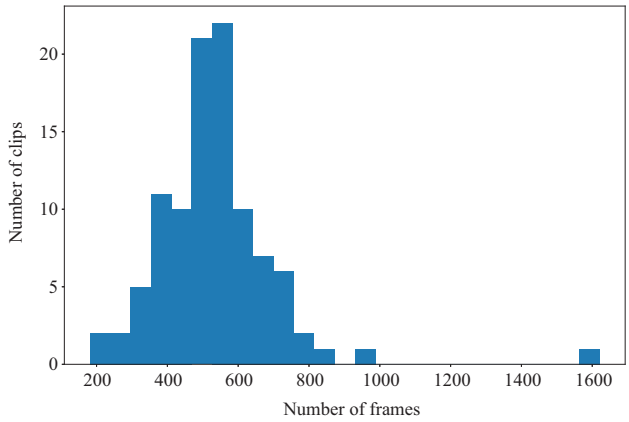

Figure 4 (Color online) Number of frames in each video clip.

4 Method

Through observation of a large number of videos during dataset construction, the motion of portraits is usually imbalanced due to the joint structure of the human body, where the motion of different parts of portraits is relatively independent. This motion imbalance leads to inaccurate prediction near the parts with a greater range of motion in existing methods, as shown in Figure 1. Towards this imbalance, we propose a PDNet for VPS, which leverages part cue to separately capture part-associated motion in VPS. The network architecture is introduced in Subsection 4.1. In our network, an IPDA module is proposed to extract motion information specified to each part, which is introduced in Subsection 4.2.

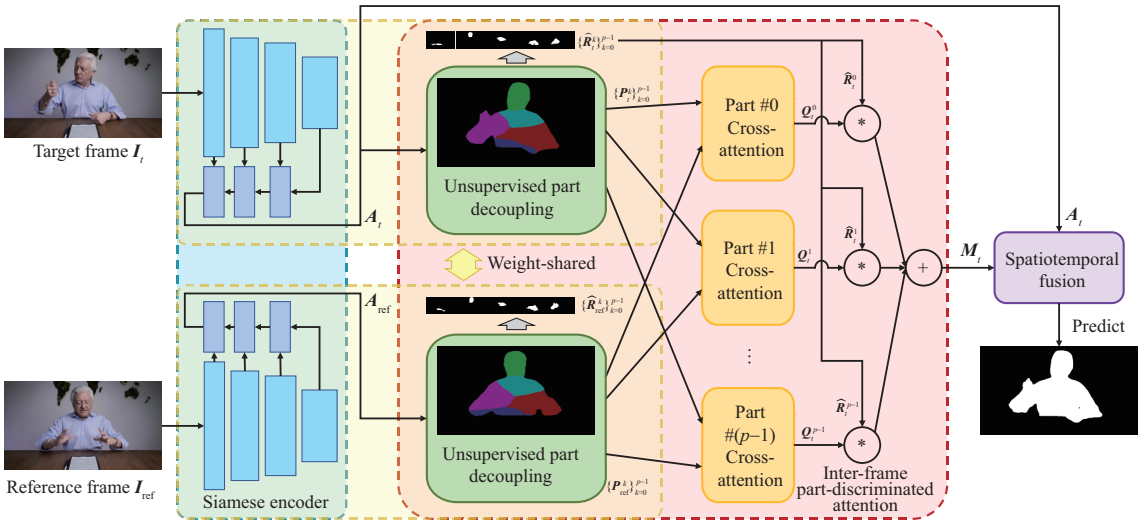

Figure 5 (Color online) Overview framework of our proposed PDNet.

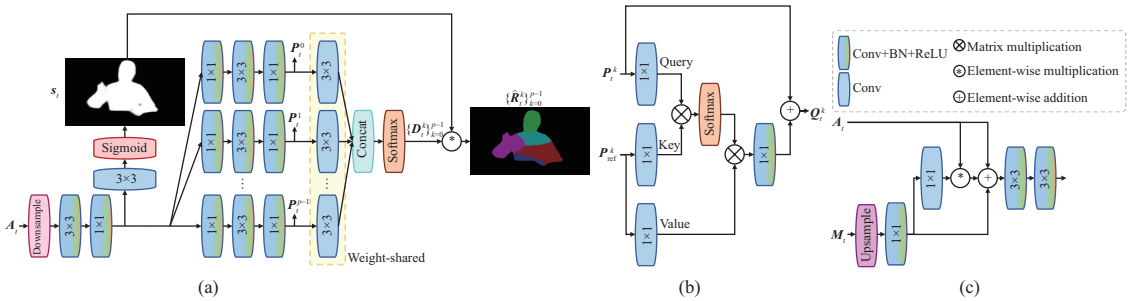

Figure 6 (Color online) Submodules in our proposed PDNet. (a) Pipeline of unsupervised part decoupling in IPDA; (b) cross-attention for correlation extraction in IPDA; (c) spatiotemporal fusion module in PDNet.

4.1 PDNet

The architecture of the proposed PDNet is shown in Figure 5. Similar to several unsupervised VOS methods [10,47], the first frame is used as a reference frame in our network to distill motion information. In the Siamese encoder, we use ResNet-50 [61] as the backbone to extract spatial features of both target frame $\boldsymbol{I}_t$ and reference frame $\boldsymbol{I}_{\text{ref}}$, and then take the advantages of feature pyramid network (FPN) [62] structure to fuse spatial semantic features and detailed features from deep to shallow. The whole encoder is weight-shared between the two frames, so that consistent appearance features $\boldsymbol{A}_t$ and $\boldsymbol{A}_{\text{ref}}$ are obtained for the same persons in different frames.

Towards the motion imbalance of portraits, we utilize our proposed IPDA module to unsupervisedly decouple $\boldsymbol{A}_t$ and $\boldsymbol{A}_{\text{ref}}$ into part-attached features, and utilize cross-attention operation on each part to extract part-discriminated motion correlation. These correlation features are then assembled to generate motion features $\boldsymbol{M}_t$ of the target frame.

After obtaining both appearance features $\boldsymbol{A}_t$ and motion features $\boldsymbol{M}_t$ of the target frame, we utilize a spatiotemporal fusion module to fuse these two features, which is introduced in Subsection 4.3. At last, the final result is predicted from the fused features.

4.2 IPDA

Due to the motion imbalance, our proposed IPDA module extracts motion correlation from the reference frame to the target frame separately for each part of the portraits, which consists of three stages: unsupervised part decoupling stage, correlation extraction stage and correlation assembly stage.

Unsupervised part decoupling. In this stage, we unsupervisedly segment the portrait into several parts and extract the part-discriminated features simultaneously, which pipeline is shown in Figure 6(a). Different from usual part decoupling methods [17,18], in order to extract part-discriminated features that

can contribute to portrait segmentation, we enhance the guidance of saliency cue in part segmentation. Specifically, we decompose the process of unsupervised part segmentation into two steps: first supervisedly predict the portrait saliency map based on appearance features, and then self-supervisedly predict each part of the portrait.

We first downsample appearance features $\boldsymbol{A}_t, \boldsymbol{A}_{\text{ref}}$ extracted by the Siamese encoder, and use a 3×3 convolution operation with a 1×1 convolution operation which compresses dimensions of $\boldsymbol{A}_t$ and $\boldsymbol{A}_{\text{ref}}$ to reduce the computational cost, and then a saliency predictor is applied to predict the portrait saliency map $\boldsymbol{S}_t$ and $\boldsymbol{S}_{\text{ref}}$. Then, in order to extract discriminated features from each portrait part, we utilize p groups of a 1×1, a 3×3 and a 1×1 convolution operations to decouple compressed features into part-discriminated features $\{\boldsymbol{P}_t^k\}_{k=0}^{p-1}$ and $\{\boldsymbol{P}_{\text{ref}}^k\}_{k=0}^{p-1}$, where p denotes the number of parts. Stacked 1×1 and 3×3 convolution operations can enhance cross-channel information interaction and introduce more non-linearity, thus enhancing distinctiveness between part-discriminated features. Then, a weight-shared prediction head is applied on p part-discriminated features to generate decoupled part predictions $\{\boldsymbol{D}_t^k\}_{k=0}^{p-1}$ and $\{\boldsymbol{D}_{\text{ref}}^k\}_{k=0}^{p-1}$ under the constraint of self-supervision losses: geometric concentration loss $\mathcal{L}_{\text{geo}}$, semantic consistency loss $\mathcal{L}_{\text{sem}}$ and area variance loss $\mathcal{L}_{\text{area}}$, which are introduced in Subsection 4.4. This weight-sharing mechanism ensures semantic consistency among part-discriminated features. Predicted part masks $\{\hat{\boldsymbol{R}}_t^k\}_{k=0}^{p-1}$, $\{\hat{\boldsymbol{R}}_{\text{ref}}^k\}_{k=0}^{p-1}$ are obtained by multiplying portrait saliency map $\boldsymbol{S}_t, \boldsymbol{S}_{\text{ref}}$ and decoupled part predictions $\{\boldsymbol{D}_t^k\}_{k=0}^{p-1}$, $\{\boldsymbol{D}_{\text{ref}}^k\}_{k=0}^{p-1}$. Take the target frame for example, $\{\hat{\boldsymbol{R}}_t^k\}_{k=0}^{p-1}$ can be formulated as

$$\hat{\boldsymbol{R}}_t^k = \boldsymbol{S}_t \odot \boldsymbol{D}_t^k, k \in \{0, 1, \ldots, p-1\}, \tag{1}$$

where $\odot$ denotes element-wise production. Operations for the reference frame are the same. Note that all of the operations in the part decoupling stage are weight-shared between the features of the target and reference frame, which ensures the correspondence of parts and the consistency of features between the two frames.

Correlation extraction. After obtaining part-discriminated features $\{\boldsymbol{P}_t^k\}_{k=0}^{p-1}$ and $\{\boldsymbol{P}_{\text{ref}}^k\}_{k=0}^{p-1}$, we apply p cross-attention modules [63] on the part-discriminated features of p parts separately to obtain part-discriminated motion features $\{\boldsymbol{Q}_t^k\}_{k=0}^{p-1}$ of the target frame. Note that different from general non-local self-attention module [64], the cross-attention module we utilized is asymmetric, in which the query features are from discriminated features $\boldsymbol{P}_t^k$ of some part k in the target frame, while the key and value features are from discriminated features $\boldsymbol{P}_{\text{ref}}^k$ of the corresponding part k in the reference frame, as shown in Figure 6(b). Part-discriminated motion correlations from parts in the reference frame to the corresponding parts in the target frame are separately extracted in this way.

Correlation assembly. As shown in Figure 5, in this stage we assemble the k part-discriminated motion features $\{\boldsymbol{Q}_t^k\}_{k=0}^{p-1}$ of the target frame to acquire integrated motion features $\boldsymbol{M}_t$ by masking them with corresponding part masks $\{\hat{\boldsymbol{R}}_t^k\}_{k=0}^{p-1}$ and adding up them, which is formulated as

$$\boldsymbol{M}_t = \sum_{k=0}^{p-1} \boldsymbol{Q}_t^k \odot \hat{\boldsymbol{R}}_t^k. \tag{2}$$

4.3 Spatiotemporal fusion

As both appearance cue in the spatial domain and motion cue in the temporal domain are important for video segmentation, $\boldsymbol{A}_t$ and $\boldsymbol{M}_t$ are fused through our spatiotemporal fusion module. The pipeline is shown in Figure 6(c), where we first use a 1×1 convolution operation to expand the dimension of $\boldsymbol{M}_t$. After that, we use another 1×1 convolution operation on the expanded motion features and multiply it with $\boldsymbol{A}_t$ to obtain spatiotemporal features, in which regions that are significant in both appearance and motion are more responsive. At last, we sum up the three features and use two 3×3 convolution operations to generate fused features.

4.4 Loss functions

Losses for portrait segmentation. We utilize weighted binary cross-entropy (weighted BCE) loss and L1 loss for portrait segmentation, which are commonly used in VOS methods [6, 7]. We utilize these two losses on both the final prediction and other auxiliary outputs from the Siamese encoder, which can enhance the portrait feature extraction ability of the encoder.

Losses for part decoupling stage in IPDA module. In the saliency prediction step, we utilize weighted BCE loss and L1 loss as well on $\boldsymbol{S}_t$ and $\boldsymbol{S}_{\text{ref}}$. In the self-supervised part prediction step, we utilize geometric concentration loss $\mathcal{L}_{\text{geo}}(\{\boldsymbol{X}^k\})$ and semantic consistency loss $\mathcal{L}_{\text{sem}}(\{\boldsymbol{X}^k\}, \boldsymbol{Y})$ as introduced in [17], where $\boldsymbol{X}^k$ denotes a predicted mask of part k and $\boldsymbol{Y}$ denotes the mask of the portrait. Additionally, we introduce the area variance loss $\mathcal{L}_{\text{area}}(\{\boldsymbol{X}^k\})$ for self-supervision, which aims to prevent generating oversized or undersized parts which may result in reducing the ability of IPDA module to resolve motion, which is defined as

$$\mathcal{L}_{\text{area}}(\{\boldsymbol{X}^k\}) = D(\{\text{area}(\boldsymbol{X}^k)\}/\max(\|\{\text{area}(\boldsymbol{X}^k)\}\|_2, \varepsilon)), \tag{3}$$

where $D(\cdot)$ indicates the variance, ε denotes a small constant that prevents division by zero, and $\text{area}(\boldsymbol{X}^k)$ denotes the expected pixel number of the predicted part mask $\boldsymbol{X}^k$, which is computed as

$$\text{area}(\boldsymbol{X}^k) = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \boldsymbol{X}^k(i,j). \tag{4}$$

The constraint of $\mathcal{L}_{\text{geo}}$ ensures that pixels within each part tend to converge towards a specific region, enabling the model to generate parts with well-organized shape connectivity. Under the supervision of $\mathcal{L}_{\text{sem}}$, pixels with similar semantics tend to be assigned to the same part, while the discrepancies of semantic features between different parts are enlarged. This enables the model to distinguish portrait parts with different features more reasonably. $\mathcal{L}_{\text{area}}$ suppresses the generation of oversized or undersized parts, preventing the wasteful computation of these ineffective parts as the number of parts is limited. With the help of these three losses, reasonable parts can be generated to enhance the model's ability to extract part-discriminated motion correlations.

As the predicted saliency maps $\boldsymbol{S}_t, \boldsymbol{S}_{\text{ref}}$ are not accurate and stable enough for self-supervision on part segmentation, we also introduce ground truth masks $\boldsymbol{G}_t, \boldsymbol{G}_{\text{ref}}$ in the training stage and produce part masks $\{\boldsymbol{R}_t^k\}_{k=0}^{p-1}, \{\boldsymbol{R}_{\text{ref}}^k\}_{k=0}^{p-1}$ masked by the ground truths, which can be formulated as

$$\boldsymbol{R}_t^k = \boldsymbol{G}_t \odot \boldsymbol{D}_t^k, k \in \{0, 1, \ldots, p-1\} \tag{5}$$

for the target frame and the same operation for the reference frame.

Meanwhile, as the ground truths are not available during inference time, to enhance consistency between the training and inference time while ensuring accuracy, we use both $\{\hat{\boldsymbol{R}}_t^k\}_{k=0}^{p-1}, \{\hat{\boldsymbol{R}}_{\text{ref}}^k\}_{k=0}^{p-1}$ and $\{\boldsymbol{R}_t^k\}_{k=0}^{p-1}, \{\boldsymbol{R}_{\text{ref}}^k\}_{k=0}^{p-1}$ for self-supervision. The summed-up part segmentation loss $\mathcal{L}_{\text{total\_part}}$ of the target frame can be formulated as

$$\mathcal{L}_{\text{total\_part}} = (\mathcal{L}_{\text{part}}(\{\boldsymbol{R}_t^k\}, \boldsymbol{G}_t) + \mathcal{L}_{\text{part}}(\{\hat{\boldsymbol{R}}_t^k\}, \boldsymbol{S}_t))/2, \tag{6}$$

where

$$\mathcal{L}_{\text{part}}(\{\boldsymbol{X}_t^k\}, \boldsymbol{Y}_t) = \mathcal{L}_{\text{geo}}(\{\boldsymbol{X}_t^k\}) + \mathcal{L}_{\text{sem}}(\{\boldsymbol{X}_t^k\}, \boldsymbol{Y}_t) + \mathcal{L}_{\text{area}}(\{\boldsymbol{X}_t^k\}). \tag{7}$$

$\mathcal{L}_{\text{total\_part}}$ for the reference frame is the same.

5 Experiments

5.1 Implementation details

We implement our network by PyTorch [65] and use two NVIDIA RTX 3090 GPUs for conducting our experiments. The number of parts p in the IPDA module is experimentally set to 5. ResNet-50 [61] pre-trained on ImageNet [66] is adopted as the backbone. As for semantic consistency loss $\mathcal{L}_{\text{sem}}$, we adopt a weight-fixed ResNet-18 [61] to generate semantic features for parts, which is not used during inference. Besides the MVPS dataset, we also use several image portrait segmentation datasets described in Subsection 5.2 to fine-tune the backbone as the same training strategy in [6, 7, 10]. That is, images and video frames are used alternately during training, where 1 iter of backbone fine-tuning by images is conducted before per 2 iters of training by video frames. All of the images, target frames and reference frames fed into the network are randomly cropped, randomly flipped in the horizontal direction and resized to a fixed size of 480×480 during training. Stochastic gradient descent (SGD) optimizer with a

Table 1 Quantitative comparison of our proposed PDNet with 7 state-of-the-art unsupervised VOS methods on MVPS dataset[a)]

Method	Origin	Backbone	$\mathcal{J}\&\mathcal{F}$ Mean↑	$\mathcal{J}$			$\mathcal{F}$			FPS
				Mean↑	Recall↑	Decay↓	Mean↑	Recall↑	Decay↓	
COSNet [6]	CVPR 2019	ResNet-101	85.7	87.8	97.7	2.4	83.5	98.2	1.8	2.5
AGNN [7]	ICCV 2019	ResNet-101	80.6	83.1	93.5	2.4	78.1	93.3	2.0	1.2
MATNet [8]	TIP 2020	ResNet-101	79.2	83.1	94.6	1.2	75.2	92.0	**0.9**	10.2
F2Net [10]	AAAI 2021	ResNet-101	77.5	82.2	93.1	2.1	72.9	86.4	2.9	11.1
FSNet [12]	ICCV 2021	ResNet-50	84.2	86.5	95.8	2.0	81.9	97.5	1.8	23.6
AMC-Net [13]	ICCV 2021	ResNet-101	85.9	88.1	98.1	**0.6**	83.6	97.4	1.1	22.8
HFAN-medium [14]	ECCV 2022	MiT-B2	86.9	88.6	96.9	2.7	85.1	95.5	1.2	10.2
IMCNet [16]	TCSVT 2022	ResNet-101	84.1	86.6	97.2	2.0	81.6	97.6	1.3	23.9
PDNet	Ours	ResNet-50	**88.1**	**90.0**	**98.3**	2.2	**86.1**	**98.3**	**0.9**	**26.6**

a) The best results in each column are shown in **bold**.

Table 2 Quantitative comparison on PP-HumanSeg14K [5] dataset[a)]

Method	$\mathcal{J}\&\mathcal{F}$ Mean	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
COSNet [6]	95.7	96.9	94.6
AGNN [7]	93.9	96.1	91.6
MATNet [8]	93.1	95.6	90.7
F2Net [10]	89.3	93.9	84.7
IMCNet [16]	94.5	96.4	92.7
PDNet	**96.0**	**97.1**	**94.9**

a) The best results in each column are shown in **bold**.

momentum of 0.9 and a weight decay of 5×10^{-4} is used for training. The size of the mini-batch is set to 4 and the number of epochs is set to 60. During fine-tuning by image, the maximum learning rate is set to 2.5×10^{-4}, while during training by video frames, it is set to 2.5×10^{-5} for backbone and 2.5×10^{-3} for the rest of our model. During inference, both target and reference frames are directly resized to 480×480 for prediction.

5.2 Datasets and evaluation metrics

We use our proposed MVPS dataset for training and evaluating our network and state-of-the-art unsupervised VOS methods. As for the image datasets used for fine-tuning the backbone, we use images in EG1800 [3] (1735 images available), SuperviselyPerson (5711 images available) and MSCOCO [67]. For the MSCOCO dataset, we select images including persons of which the pixel number is no less than 5000, and pick out 49479 images. For a fair comparison, the same image set is also used during training on VPS for those methods that utilize image datasets during training on VOS. For evaluation metrics, we adopt region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$ which are commonly used in VOS task [4].

5.3 Comparison results

Quantitative comparison on MVPS. We compare our proposed PDNet with 8 state-of-the-art unsupervised VOS methods including COSNet [6], AGNN [7], MATNet [8], F2Net [10], FSNet [12], AMC-Net [13], HFAN-medium [14], and IMCNet [16] which codes are publicly available for retraining and evaluation. For a fair comparison, a unified threshold of 0.5 is utilized on predictions from all methods to generate binary portrait masks, while conditional random field (CRF) is not used for post-processing.

The quantitative results of our PDNet and the other 7 methods mentioned above are shown in Table 1. From the results, we can see that our PDNet achieves significantly leading performance with the comparison to state-of-the-art methods on the test set of MVPS, which outperforms the second best method HFAN-medium by 1.4% and 1.0% on $\mathcal{J}$ Mean and $\mathcal{F}$ Mean, respectively. Note that our PDNet utilizes ResNet-50 as a backbone, while most state-of-the-art unsupervised VOS methods utilize ResNet-101 [61] with more parameters and stronger feature extraction capability.

We also compare the efficiency of our PDNet with state-of-the-art methods under the same environment, as shown in the last column of Table 1. Our method achieves a speed of 26.6 FPS, which is the most competitive among these methods.

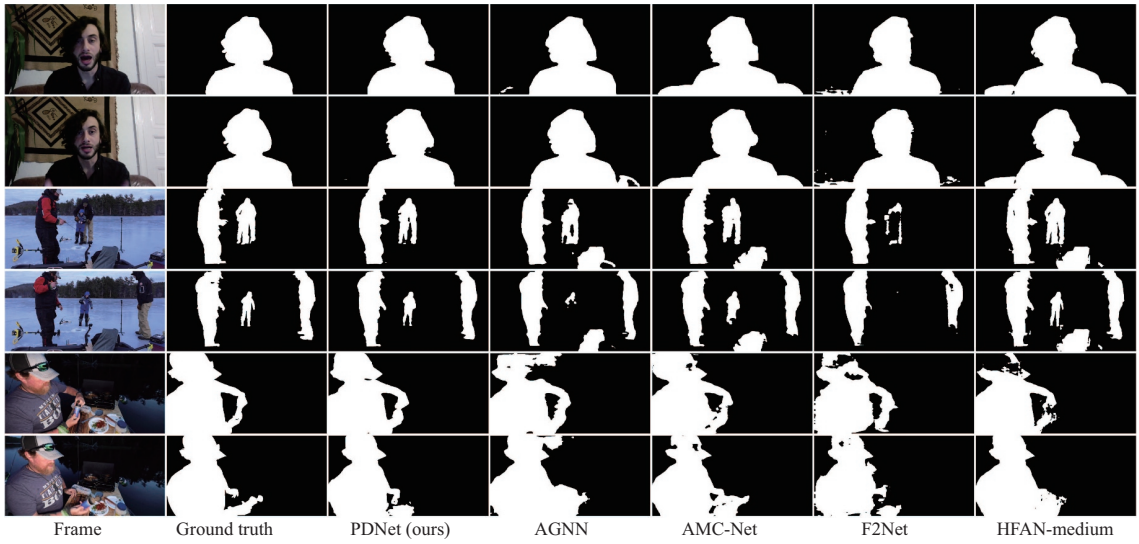

Figure 7 (Color online) Visual comparison between our method and state-of-the-art methods on MVPS dataset.

Quantitative comparison on PP-HumanSeg14K. To further demonstrate the robustness of our method and the complexity of our MVPS dataset, we also train and evaluate our PDNet and several state-of-the-art unsupervised VOS methods on PP-HumanSeg14K [5] dataset. The comparison results are shown in Table 2. Our approach also achieves the best performance. Note that the overall performance on PP-HumanSeg14K dataset is significantly higher than that on the MVPS dataset, which shows our proposed MVPS dataset is more challenging due to the complexity of our dataset.

Qualitative comparison. To intuitively demonstrate the excellent performance of our method, we compare the visualization results of our method with state-of-the-art VOS methods on the test set of our MVPS dataset. As shown in Figure 7, our method segments foreground portraits accurately in a variety of portrait numbers, portrait motion, portrait gestures and background scenes, especially in regions with imbalanced motion against the main bodies, while other methods incorrectly predict objects in backgrounds such as sofa and parcel, as foreground portrait area, or miss some regions of portraits even the whole portrait of someone. Although our method has achieved good performance, there is still room for improvement. For instance, the segmentation results of boundary areas or regions near occluded objects are not yet satisfactory. We will focus on further enhancing the model's ability to handle details in our future work.

5.4 Visual analysis of decoupled parts

As an important cue, part plays a crucial role in our method. Here we show some visual examples of decoupled parts as intermediate results in Figure 8. Under the effect of self-supervision, our method decouples portraits into parts which are usually with different motion states. Portrait parts generated by our IPDA module are robust to the variation of background scenes and portrait motion, even when there are multiple persons in one video clip. Based on this robustness, our method can extract the discriminated motion correlation of different portrait parts separately to capture the imbalanced motion of portraits. Note that although the current segmentation result is not perfect, it can already provide important auxiliary information for imbalanced motion extraction, which improves the accuracy of final segmentation results.

5.5 Ablation study

Impact of the IPDA module. To further verify the effectiveness of our IPDA module and self-supervision strategies in part decoupling for VPS, we conduct experiments as shown in Table 3. In the first row, we remove the whole IPDA and spatiotemporal fusion modules, while in the second row, we replace part segmentation losses with only $\mathcal{L}_{\text{geo}}$. It can be seen that introducing rough part cue can also improve the performance even without area and semantic supervision. In the third row, we reserve $\mathcal{L}_{\text{geo}}, \mathcal{L}_{\text{area}}$ and remove $\mathcal{L}_{\text{sem}}$, which demonstrates the effectiveness of our proposed area variance loss. It

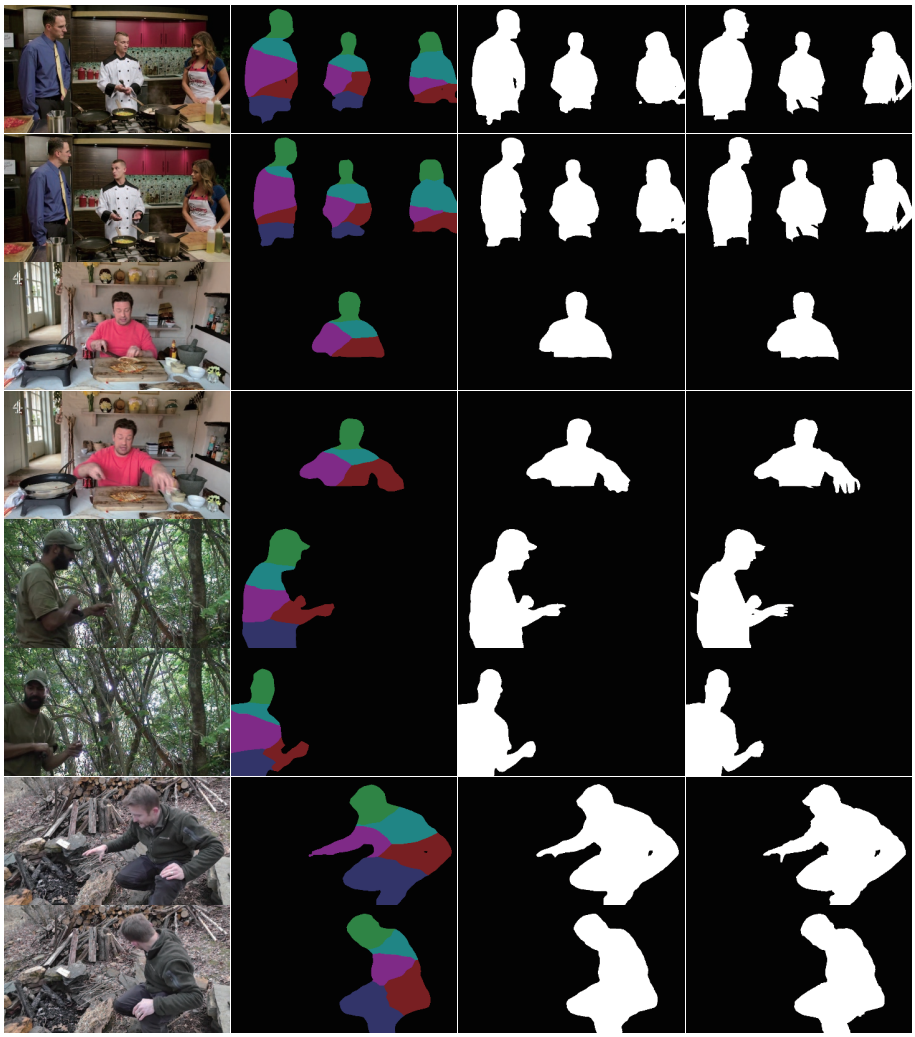

Figure 8 (Color online) Visual examples of decoupled parts and segmentation results of our PDNet on the test set of MVPS dataset.

Table 3 Ablation study on the IPDA module and self-supervision strategies in part decoupling on MVPS dataset

Base	Self-supervision losses			$\mathcal{J}\&\mathcal{F}$ Mean	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
	$\mathcal{L}_{\text{geo}}$	$\mathcal{L}_{\text{area}}$	$\mathcal{L}_{\text{sem}}$			
✓	w/o IPDA and fusion			85.0	88.2	81.9
✓	✓			86.5	89.0	84.0
✓	✓	✓		87.3	89.4	85.2
✓	✓	✓	✓	**88.1**	**90.0**	**86.1**

prevents oversized or undersized parts from being generated, which guarantees the stability of the ability of the IPDA module to decompose portrait motion. From the third and the last row, we can see that semantic cue is also important to enhance the motion-resolving ability, as owing to the joint structure of the human body, parts with different appearance semantics are often in different motion states. From Figure 9 we can see that under the guidance of more appropriate part maps generated by adding $\mathcal{L}_{\text{area}}$ and $\mathcal{L}_{\text{sem}}$ during training, the model can segment part boundary regions and regions with obvious motion more accurately.

Impact of the number of parts p. To verify the impact of the number of parts p in our IPDA module, we conduct experiments as shown in Table 4. Note that when the number of parts is set to 1, losses for self-supervised part segmentation are removed as they are not needed. From the table, we can see that within a certain range, the segmentation performance improves as the number of parts increases.

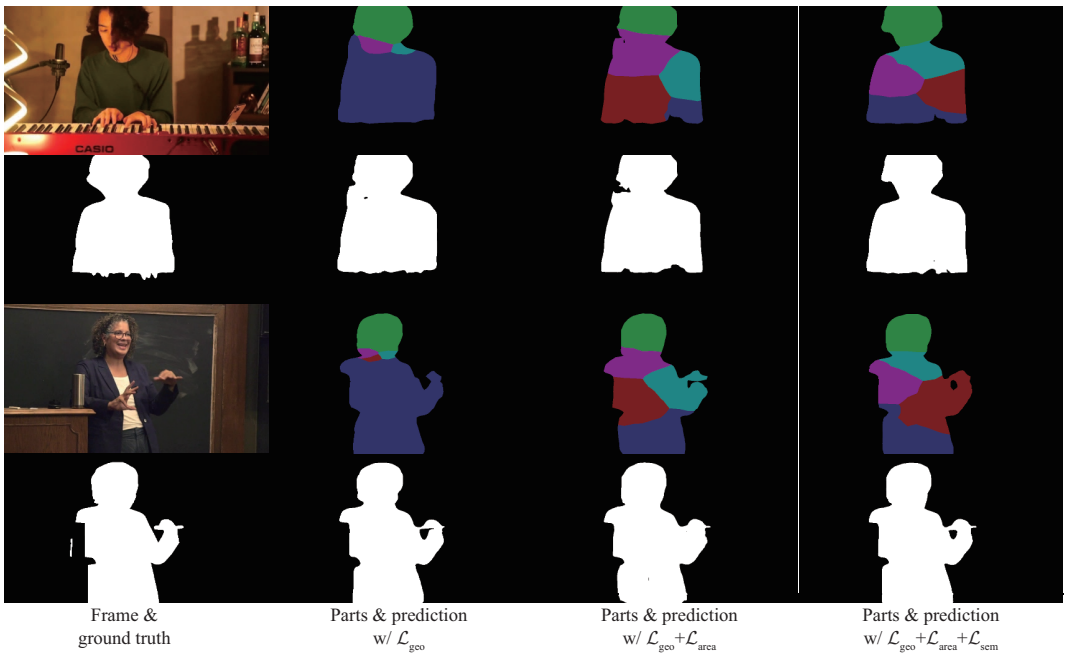

Figure 9 (Color online) Visual comparison for ablation study on self-supervision losses in part decoupling. Frames and decoupled parts are shown in odd lines, while ground truths and segmentation results are shown in even lines.

Table 4 Ablation study on the number of parts p in our IPDA module on MVPS dataset

Number of parts p	$\mathcal{J}\&\mathcal{F}$ Mean	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
1	86.2	88.8	83.6
3	86.7	88.7	84.7
5	**88.1**	**90.0**	**86.1**

Table 5 Ablation study on the number of reference frames of our PDNet on MVPS dataset

Number of reference frames	$\mathcal{J}\&\mathcal{F}$ Mean	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean	FPS
1	88.1	90.0	86.1	26.6
2	88.7	90.7	86.8	9.7

This is because the increase in the number of parts enables portrait regions with different motion states to be separated while extracting motion correlation, which enhances the decoupling ability of the model to complex motion. As we use NVIDIA RTX 3090 GPU for experiments, the memory is almost full during training when the number of parts is set to 5, so we use 5 as the final setting of number of parts in our method.

Impact of the number of reference frames. To verify the impact of the number of reference frames, we add the fifth annotated frame before the current frame as the second reference frame, and the experimental results are shown in Table 5. From Table 5, we can see that although there is an improvement in performance, there is a significant decline in the efficiency of the model, with FPS dropping from 26.6 to 9.7, a decrease of 63.5%. The involvement of more convolution and attention operations significantly increases computational overhead, resulting in a decrease in speed. Since the balance between performance and efficiency is important, and using 1 reference frame is a typical setting in video segmentation methods [1, 10, 47], we use 1 reference frame in our method.

6 Conclusion

In this paper, we propose a new intricate large-scale dataset MVPS and a PDNet for VPS. The proposed dataset MVPS addresses the lack of intricate large-scale multi-scene VPS datasets, which contains 101 video clips and 10843 annotated frames in 7 categories of scenarios. It is currently the most complex dataset for VPS. Towards the imbalance of portrait motion, our proposed PDNet decouples portrait

motion into part-level. In our network, the IPDA module is proposed to extract discriminative part-level motion features from the target and reference frames. Experimental results demonstrate that our method achieves superior performance in comparison to state-of-the-art unsupervised VOS methods.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62132002, 62102206) and Major Key Project of PCL (Grant No. PCL2023A10-1).

References

1 Wang Y, Zhang W, Wang L, et al. Temporal consistent portrait video segmentation. Pattern Recogn, 2021, 120: 108143
2 Pandey R, Escolano S O, Legendre C, et al. Total relighting: learning to relight portraits for background replacement. ACM Trans Graph, 2021, 40: 1–21
3 Shen X, Hertzmann A, Jia J, et al. Automatic portrait segmentation for image stylization. Comput Graph Forum, 2016, 35: 93–102
4 Perazzi F, Pont-Tuset J, McWilliams B, et al. A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 724–732
5 Chu L, Liu Y, Wu Z, et al. PP-HumanSeg: connectivity-aware portrait segmentation with a large-scale teleconferencing video dataset. In: Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, 2022. 202–209
6 Lu X, Wang W, Ma C, et al. See more, know more: unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 3618–3627
7 Wang W, Lu X, Shen J, et al. Zero-shot video object segmentation via attentive graph neural networks. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2019. 9235–9244
8 Zhou T, Li J, Wang S, et al. MATNet: motion-attentive transition network for zero-shot video object segmentation. IEEE Trans Image Process, 2020, 29: 8326–8338
9 Lu X, Wang W, Danelljan M, et al. Video object segmentation with episodic graph memory networks. In: Proceedings of European Conference on Computer Vision, 2020. 661–679
10 Liu D, Yu D, Wang C, et al. F2Net: learning to focus on the foreground for unsupervised video object segmentation. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021. 2109–2117
11 Ren S, Liu W, Liu Y, et al. Reciprocal transformations for unsupervised video object segmentation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 15430–15439
12 Ji G P, Fu K, Wu Z, et al. Full-duplex strategy for video object segmentation. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2021. 4902–4913
13 Yang S, Zhang L, Qi J, et al. Learning motion-appearance co-attention for zero-shot video object segmentation. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2021. 1544–1553
14 Pei G, Shen F, Yao Y, et al. Hierarchical feature alignment network for unsupervised video object segmentation. In: Proceedings of European Conference on Computer Vision, 2022. 596–613
15 Zhou Y, Xu X, Shen F, et al. Flow-edge guided unsupervised video object segmentation. IEEE Trans Circ Syst Video Technol, 2022, 32: 8116–8127
16 Xi L, Chen W, Wu X, et al. Implicit motion-compensated network for unsupervised video object segmentation. IEEE Trans Circ Syst Video Technol, 2022, 32: 6279–6292
17 Hung W C, Jampani V, Liu S, et al. SCOPS: self-supervised co-part segmentation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 869–878
18 Liu S, Zhang L, Yang X, et al. Unsupervised part segmentation through disentangling appearance and shape. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 8351–8360
19 Huang Z, Li Y. Interpretable and accurate fine-grained recognition via region grouping. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 8659–8669
20 Yu X, Wang J, Zhao Y, et al. Mix-ViT: mixing attentive vision transformer for ultra-fine-grained visual categorization. Pattern Recogn, 2023, 135: 109131
21 Li X, Liu S, Kim K, et al. Self-supervised single-view 3D reconstruction via semantic consistency. In: Proceedings of European Conference on Computer Vision, 2020. 677–693
22 Zhao Y, Li J, Zhang Y, et al. From pose to part: weakly-supervised pose evolution for human part segmentation. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 3107–3120
23 Xie C, Xia C, Ma M, et al. Pyramid grafting network for one-stage high resolution saliency detection. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 11707–11716
24 Zhao Z, Xia C, Xie C, et al. Complementary trilateral decoder for fast and accurate salient object detection. In: Proceedings of ACM International Conference on Multimedia, 2021. 4967–4975
25 Ma M, Xia C, Li J. Pyramidal feature shrinking for salient object detection. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021. 2311–2318
26 Zhuge M, Fan D P, Liu N, et al. Salient object detection via integrity learning. IEEE Trans Pattern Anal Mach Intell, 2022, : 1
27 Cong R, Qin Q, Zhang C, et al. A weakly supervised learning framework for salient object detection via hybrid labels. IEEE Trans Circ Syst Video Technol, 2023, 33: 534–548
28 Fang C W, Tian H B, Zhang D W, et al. Densely nested top-down flows for salient object detection. Sci China Inf Sci, 2022, 65: 182103
29 Zhou W J, Liu C, Lei J S, et al. RLLNet: a lightweight remaking learning network for saliency redetection on RGB-D images. Sci China Inf Sci, 2022, 65: 160107
30 Yue Y H, Zou Q, Yu H K, et al. An end-to-end network for co-saliency detection in one single image. Sci China Inf Sci, 2023, 66: 210101
31 Zhang S H, Dong X, Li H, et al. PortraitNet: real-time portrait segmentation network for mobile device. Comput Graphic, 2019, 80: 104–113
32 Park H, Sjösund L L, Yoo Y, et al. SINet: extreme lightweight portrait segmentation networks with spatial squeeze modules and information blocking decoder. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2020. 2055–2063

33 Zhang X Y, Wang L J, Xie J, et al. Human-in-the-loop image segmentation and annotation. Sci China Inf Sci, 2020, 63: 219101
34 Vineet V, Warrell J, Ladicky L, et al. Human instance segmentation from video using detector-based conditional random fields. In: Proceedings of British Machine Vision Conference, 2011
35 Bhole C, Pal C. Automated person segmentation in videos. In: Proceedings of International Conference on Pattern Recognition, 2012. 3672–3675
36 Xu M, Fan C, Wang Y, et al. Joint person segmentation and identification in synchronized first- and third-person videos. In: Proceedings of European Conference on Computer Vision, 2018. 656–672
37 Gruosso M, Capece N, Erra U. Human segmentation in surveillance video with deep learning. Multimed Tools Appl, 2021, 80: 1175–1199
38 Song H, Wang W, Zhao S, et al. Pyramid dilated deeper convLSTM for video salient object detection. In: Proceedings of European Conference on Computer Vision, 2018. 744–760
39 Ventura C, Bellver M, Girbau A, et al. RVOS: end-to-end recurrent network for video object segmentation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 5272–5281
40 Wang W, Shen J, Lu X, et al. Paying attention to video object pattern understanding. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 2413–2428
41 Fan J, Su T, Zhang K, et al. Bidirectionally learning dense spatio-temporal feature propagation network for unsupervised video object segmentation. In: Proceedings of ACM International Conference on Multimedia, 2022. 3646–3655
42 Tokmakov P, Schmid C, Alahari K. Learning to segment moving objects. Int J Comput Vis, 2019, 127: 282–301
43 Faisal M, Akhter I, Ali M, et al. EpO-Net: exploiting geometric constraints on dense trajectories for motion saliency. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2020. 1873–1882
44 Zhao X, Pang Y, Yang J, et al. Multi-source fusion and automatic predictor selection for zero-shot video object segmentation. In: Proceedings of ACM International Conference on Multimedia, 2021. 2645–2653
45 Zhang K, Zhao Z, Liu D, et al. Deep transport network for unsupervised video object segmentation. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2021. 8761–8770
46 Cong R, Song W, Lei J, et al. PSNet: parallel symmetric network for video salient object detection. IEEE Trans Emerg Top Comput Intell, 2023, 7: 402–414
47 Yang Z, Wang Q, Bertinetto L, et al. Anchor diffusion for unsupervised video object segmentation. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2019. 931–940
48 Zhang L, Zhang J, Lin Z, et al. Unsupervised video object segmentation with joint hotspot tracking. In: Proceedings of European Conference on Computer Vision, 2020. 490–506
49 Lee Y, Seong H, Kim E. Iteratively selecting an easy reference frame makes unsupervised video object segmentation easier. In: Proceedings of AAAI Conference on Artificial Intelligence, 2022. 1245–1253
50 Chen Y D, Hao C Y, Yang Z X, et al. Fast target-aware learning for few-shot video object segmentation. Sci China Inf Sci, 2022, 65: 182104
51 Wen P, Yang R, Xu Q, et al. DMVOS: discriminative matching for real-time video object segmentation. In: Proceedings of ACM International Conference on Multimedia, 2020. 2048–2056
52 Yang L, Han J, Zhao T, et al. Background-click supervision for temporal action localization. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 9814–9829
53 Zhao T, Han J, Yang L, et al. SODA: weakly supervised temporal action localization based on astute background response and self-distillation learning. Int J Comput Vis, 2021, 129: 2474–2498
54 Lee P, Uh Y, Byun H. Background suppression network for weakly-supervised temporal action localization. In: Proceedings of AAAI Conference on Artificial Intelligence, 2020. 11320–11327
55 Zhao T, Han J, Yang L, et al. Equivalent classification mapping for weakly supervised temporal action localization. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 3019–3031
56 Shi D, Zhong Y, Cao Q, et al. TriDet: temporal action detection with relative boundary modeling. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 18857–18866
57 Ochs P, Malik J, Brox T. Segmentation of moving objects by long term video analysis. IEEE Trans Pattern Anal Mach Intell, 2014, 36: 1187–1200
58 Fan D P, Wang W, Cheng M M, et al. Shifting more attention to video salient object detection. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 8546–8556
59 Xu N, Yang L, Fan Y, et al. YouTube-VOS: a large-scale video object segmentation benchmark. 2018. ArXiv:1809.03327
60 Rahane A A, Subramanian A. Measures of complexity for large scale image datasets. In: Proceedings of International Conference on Artificial Intelligence in Information and Communication, 2020. 282–287
61 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
62 Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 936–944
63 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017
64 Wang X, Girshick R, Gupta A, et al. Non-local neural networks. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 7794–7803
65 Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of Advances in Neural Information Processing Systems, 2019
66 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, 2009. 248–255
67 Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Proceedings of European Conference on Computer Vision, 2014. 740–755