# Leveraging Motion Priors in Videos for Improving Human Segmentation

Yu-Ting Chen[1], Wen-Yen Chang[1], Hai-Lun Lu[1], Tingfan Wu[2], Min Sun[1]

[1] National Tsing Hua University
{yuting2401,s0936100879,oscar.lu1007}@gmail.com, sunmin@ee.nthu.edu.tw
[2] Umbo Computer Vision
tingfan.wu@umbocv.com

**Abstract.** Despite many advances in deep-learning based semantic segmentation, performance drop due to distribution mismatch is often encountered in the real world. Recently, a few domain adaptation and active learning approaches have been proposed to mitigate the performance drop. However, very little attention has been made toward leveraging information in videos which are naturally captured in most camera systems. In this work, we propose to leverage "motion prior" in videos for improving human segmentation in a weakly-supervised active learning setting. By extracting motion information using optical flow in videos, we can extract candidate foreground motion segments (referred to as motion prior) potentially corresponding to human segments. We propose to learn a memory-network-based policy model to select *strong* candidate segments (referred to as *strong* motion prior) through reinforcement learning. The selected segments have high precision and are directly used to finetune the model. In a newly collected surveillance camera dataset and a publicly available UrbanStreet dataset, our proposed method improves the performance of human segmentation across multiple scenes and modalities (i.e., RGB to Infrared (IR)). Last but not least, our method is empirically complementary to existing domain adaptation approaches such that additional performance gain is achieved by combining our weakly-supervised active learning approach with domain adaptation approaches.

**Keywords:** Active Learning, Domain Adaptation, Human Segmentation

## 1 Introduction

Intelligent camera systems with the capability to recognize objects often encounter issues caused by data distribution mismatch in the real world. For instance, surveillance cameras encounter various weather conditions, view angles, lighting conditions, and sensor modalities (e.g., RGB, infrared or even thermal). A standard solution is to collect more labeled images from various distributions to train a more robust model. However, collecting high-quality labels is very expensive and time-consuming, especially for segmentation and detection tasks.

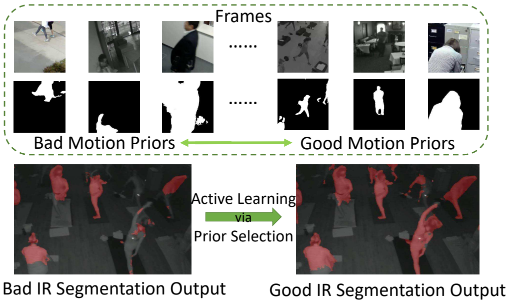

**Fig. 1.** (top): RGB patches and their corresponding patch-based motion priors extracted from videos. The priors can be classified into "good" and "bad" ones. (bottom): Our proposed active learning strategy can select good motion priors to improve performance in a cross-modality (RGB to IR) segmentation scenario.

These considerations raise two critical questions: (1) "how to select data points for training such that the accuracy improved as much as possible?" and (2) "how to obtain the label of the selected data points with cost as low as possible?"

Active learning is one of the common paradigms to address the "how to select" question since it is defined as learning to select data points to label, from a pool of unlabeled data points, in order to maximize the accuracy. There exist many heuristics [1] which have been proven to be effective when applied to classical machine learning models. However, Sener and Savarese [2] have shown that these heuristics are less effective when applied to CNN. To overcome the limitation, Sener and Savarese [2] propose a new active learning method specifically designed for Convolutional Neural Networks (CNNs). Despite recent advances, Most active learning methods require human to label the selected data points. For segmentation and detection tasks, the cost of labeling a small set of selected data points can still be relatively expensive and time-consuming.

On the other hand, instead of collecting independent images, it is generally easy to collect a sequence of images (i.e., a video) from always-on camera systems. Sequences of images have two main properties: (1) images close in time are similar/redundant, and (2) difference in two consecutive images reveals motion information potentially corresponding to moving objects. Very little attention, however, has been made toward exploiting these properties in a video to automatically provide supervision to boost recognition performance and mitigate the performance drop caused by distribution mismatch. This is related to the "how to obtain labels" question. If we can obtain labels automatically from videos, it will be immensely beneficial for intelligent camera systems. In fact, researchers have proposed to extract motion information from a sequence of images. For instance, given two consecutive frames, dense optical flow can be extracted for each pixel. Given a longer sequence of frames, sparse long-term trajectories of pixels can be extracted. In the rest of the paper, we refer to these motion information in a video as "motion prior".

In this work, we propose to leverage motion prior in videos for improving human segmentation accuracy. We first compute dense optical flow between two consecutive frames. Then, we treat pixels with flow higher than a threshold as candidates of foreground motion segments, which are referred to as "motion prior". Due to the nature of imperfect optical flow, a majority of the segments are quite noisy (see examples in Fig. 1). Considering that only some candidates are good and many candidates are noisy, we propose to learn a memory-network-based policy model to select good candidate segments through reinforcement learning. The selected good segments are then used as additional ground truth to finetune the human segmenter. In this way, we can achieve active learning without additional human annotation.

Our policy is trained on a hold-out dataset with unlabeled videos and a set of labeled images. The training of the policy is formulated as a reinforcement learning problem where the reward is the accuracy of the labeled images and the action is whether to select each motion segment. Once the policy is trained, we can apply the policy to select motion segments in challenging cross-modality (RGB to IR) or cross-scene settings. We refer our setting as weakly-supervised active learning since the policy needs to be trained on an additional hold-out dataset.

In a newly collected surveillance camera dataset and a publicly available UrbanStreet dataset, our proposed method improves the performance of human segmentation across multiple scenes and modalities (i.e., RGB to Infrared (IR)). Last but not least, our method is empirically complementary to existing domain adaptation approaches such that additional performance gain is achieved by combining our weakly-supervised active learning approach with domain adaptation approaches.

In the following sections, we first describe the related works in Sec. 2. Then, we introduce our new surveillance cameras dataset in Sec. 3. Our main technical contribution—policy-based weakly-supervised active learning for strong motion prior selection—is introduced in Sec. 4. Finally, we report our experimental results in Sec. 5.

## 2 Related Works

We discuss the related work in the fields of motion segmentation, human segmentation, active learning and domain adaptation.

### 2.1 Motion Segmentation

Motion segmentation aims to decompose a video into foreground objects and background using motion information. Feature-based motion segmentation methods assume that segmentation of different motions is equivalent to segment the extracted feature trajectories into different clusters. These methods can be classified into two types: affinity-based methods [3,4] and subspace-based method [5,6]. Some of the works utilize properties of trajectory data. For example, Yan and Pollefeys [7] use geometric constraint and locality to solve the problem. Recently, [8,9] propose to jointly tackle the motion segmentation and optical flow tasks. Nirkin et al. [10] use motion as a prior and propose a man in the loop for producing segmentation labels. In our work, we simply obtain candidate moving object segments via high-quality optical flow. Most importantly, none of the work aforementioned leverage motion segmentation for weakly-supervised active learning.

### 2.2 Human Segmentation

Human segmentation has a wide range of applications. For instance, human segmentation in a high-density scene (crowded or occluded) acquired from a stationary camera has been discussed in early works [11,12]. Spina et al. [13] demonstrate applications in pose estimation and behavior study. On the other hand, in many applications, real-time performance is critical. Song et al. [14] achieve 1000 fps using a CNN-based architecture which outperforms traditional methods in both speed and accuracy. Some works use motion information for helping human segmentation, for instance, Guo et al. [15] base on local color distribution and shape priors through optical flow, and Lu et al. [16] describe a hierarchical MRF model to bridge low-level video fragments with high-level human motion and appearance.

In recent years, thermal and infrared systems have gained popularity for night vision. Hence, human segmentation on infrared images has become an important topic. For example, Tan et al. [17] propose a background subtraction based method for human segmentation on thermal infrared images. He et al. [18] further utilize predicted human segments on infrared images to guide robots search. To demonstrate severe domain shift, we evaluate our method mainly on cross-modality (RGB to IR) domain adaptation for human segmentation.

### 2.3 Active learning

An active learning algorithm can explore informative instances, querying desired output form users or other sources. Uncertainty-based approaches are widely used. These works consider uncertainty as the selection strategies. They find hard examples by dropout MC sampling [19], using heuristics like highest entropy [20], or geometric distance to decision boundaries [21,22]. Other approaches consider the diversity of selected samples, using k-means algorithms [2,23] or sparse representation for subset selection [24]. Still other important concepts also help the performance, such as selecting instances which will maximize the variance of output [25,26], or introducing the relationships between data points in structured data [27,28].

Recently, some works model the active learning process as a sequence of querying actions, using deep reinforcement learning. Fang et al. [29] demonstrates on cross-lingual setting and Bachman et al. [30] models the learning algorithm via meta-learning. Our approach is similar to these methods using learnable strategy rather than predefined heuristic. Above methods show their goal to reduce human label cost. However, we use active learning for unsupervised finetuning since our method selects automatically computed motion priors, requiring ZERO human label cost once the policy has learned.

### 2.4 Domain Adaptation

Domain adaptation leverages information from one or more source domains to improve the performance on target domain. Recent methods focus on learning deep representations to be robust to domain shift [31]. Several other works propose to align source and target domains in feature space based on Maximum Mean Discrepancy (MMD) [32] or Central Moment Discrepancy (CMD) [33].

On the other hand, adversarial training [34] has been applied for domain adaptation as well [35,36,37]. Liu et al. [35] propose Coupled GAN which generates a joint distribution of two domains for classification. Ganin et al. [36] applies adversarial training for achieving maximal confusion between the two domains. Other works such as Domain Separation Networks (DSN) [38] split the feature into shared representations and private ones, in order to improve the ability to extract domain-invariant features. Most of the works mentioned above focus on classification. Hoffman et al. [39], Chen et al. [40] and more recent works [41,42] extend to segmentation which is closer to our human segmentation task. In this work, we show that our proposed weakly-supervised active learning approach is complementary to state-of-the-art domain adaptation approaches.

## 3 Surveillance Datasets

In order to create challenging scenarios in videos, we have collected a new surveillance camera dataset consisting of large distribution mismatch due to cross-domains scenarios: cross-modalities (i.e., RGB to InfraRed (IR)) and across-scenes. It is surprisingly difficult to find existing segmentation annotated cross-domains video dataset. Due to the high cost of labeling, most public annotated video dataset are usually very small, not to mention about crossing multiple domains. In our dataset; we highlight cross-modalities for its high appearance mismatch and practical value. For surveillance application, good human segmentations across multiple sensor modality and scenes is essential. This dataset directly validates the proposed method in real-world surveillance scenarios.

We collect four datasets: Gym-RGB, Gym-IR, Store-RGB , and Multi-Scene-IR. There are two different sensor modes on typical surveillance cameras, color and infrared, which we denote as "RGB" and "IR", respectively. To simulate real-world usage, we let the camera ambient light sensor to automatically switch between the two modes. Typically, when there is sufficient lighting, the cameras operate in RGB mode; on the other hand, when it gets dark, the IR mode is activated to improve sensitivity. All datasets are videos collected by stationary cameras, we label a subset of frame sparsely sampled from each video.

### 3.1 Cross-domains Settings

We divide our data into source $\mathcal{S}$ and target $\mathcal{T}$ domains. In this dataset, we treat all RGB data as source domain and all IR data as target domain in order to test challenging cross-modalities settings. In both domains, we further define training

$T$ and evaluation $E$ sets. All evaluation set contains labeled images. In the source domain, training $T$ consists of a few labeled images $\mathcal{I}_T^{\mathcal{S}}$ and unlabeled video frames $\mathcal{V}_T^{\mathcal{S}}$. The labeled training images $\mathcal{I}_T^{\mathcal{S}}$ are used to pre-train our segmenter. The unlabeled video frames $\mathcal{V}_T^{\mathcal{S}}$ are used to extract motion prior information (Sec. 4.1). Both the unlabeled video frames $\mathcal{V}_T^{\mathcal{S}}$ and the evaluation set $\mathcal{I}_E^{\mathcal{S}}$ in the source domain are used to train our motion prior selector using reinforcement learning (Sec. 4.2). In the target domain, training $T$ consists of only unlabeled video frames $\mathcal{V}_T^{\mathcal{T}}$ which are used to extract motion prior information. Finally, we report the cross-domains performance on the evaluation set $\mathcal{I}_E^{\mathcal{T}}$ in the target domain. The statistics about a number of videos and labeled images in each set of the source and target domain are shown in Table. 1 and 2, respectively.

| Gym-RGB | | | Store-RGB | | |
|---|---|---|---|---|---|
| Train | | Test | Train | | Test |
| Images | Videos | Images | Images | Videos | Images |
| 749 | 406 | 237 | 985 | 985 | 255 |

**Table 1.** Source domain datasets. “Images” refers to the number of images that are labeled. “Videos” refers to the number of videos that contain unlabeled frames.

| Gym-IR | | Multi-Scene-IR | |
|---|---|---|---|
| Train | Test | Train | Test |
| Videos | Images | Videos | Images |
| 929 | 492 | 253 | 89 |

**Table 2.** Target domain datasets. “Images” refers to the number of labeled images. “Videos” refers to the number of videos consist of unlabeled frames. Note that there are no labeled training images in the target domain.

### 3.2 Data Collection Details

For the Store-RGB dataset, we have only color (RGB) images since there is sufficient fluorescent lighting in the stores all day. On the other hand, we collect infrared data (Multi-Scene-IR) from multiple scenes, such as house, office, walkway, park, playground, etc. For Gym scene, the data comes in both RGB and IR modalities due to natural day-and-night lighting transitions. For all videos, there are about 6 to 15 frames in one video with 1080×1920 resolution.

## 4 Our Method

We describe how to obtain motion prior from optical flow (Sec. 4.1) and select a set of *strong* motion prior. Before that, we first define some common notations below.

**Notation.** We use $i$, $n$, and $k$ to index pixel, patch and the order of input data, respectively. $\mathbf{m}$ indicates motion prior, and $m_i$ denotes the motion prior of the $i^{\text{th}}$ pixel.

### 4.1 Motion Priors from Video Frames

Our goal is to obtain a set of motion prior $\mathbf{m}$ (i.e., candidate foreground mask) from video frames. Although many sophisticated motion segmentation methods can be used, we simply apply a state-of-the-art optical flow method [43]. Then, we obtain $\mathbf{m}$ as the binarized flow map such that $m_i = 1$ if its flow magnitude is larger than a threshold $\tau$. Since surveillance cameras in our dataset are typically stationary, we may assume that most background and foreground pixels corresponding to small and large flow magnitude, respectively. For non-stationary cameras, other motion segmentation methods (e.g., [44]) can be used to handle camera motion.

These automatically obtained motion priors inevitably will be noisy and contain outliers. Hence, we propose a memory-network-based policy model to select more accurate ones instead of directly finetuning the segmenter with all noisy labels. The usage of motion priors is illustrated in Fig. 3.

### 4.2 Motion Priors Selection

We train a policy model $\pi$ which learns to select a set of *strong* motion priors. Further, these *strong* motion priors are treated as ground truth to directly fine-tune our model using cross-entropy loss. Instead of manually labeling *strong* motion priors and training the policy in a supervised fashion, we train the policy using reinforcement learning, which rewards from directly improving the human segmentation accuracy on a hold-out evaluation set in source domain. The training procedure of our policy model is illustrated in Fig. 2.

**Policy model.** We define the policy $\pi$ as the following probability function:

$$\pi(a|I, \mathbf{m}(I); \phi) \ , \tag{1}$$

where $I$ is an image, $\mathbf{m}(I)$ is its corresponding motion prior, $a \in \{0, 1\}$ is the binary action to select $(a = 1)$ or not $(a = 0)$, and $\phi$ is the model parameters.

**4.2.1 Network Architecture.** Inspired by the ideal using Memory Network [45] in Deep Q-Network (DQN) proposed by Oh at el. [46], we use an memory-network-based policy model which consists of three components: (1) a feature encoder for extracting features from images and motion priors, (2) a memory retaining a recent history of observations, and (3) an action decision layers taking both content features and retrieved memory state to decide the action.

**Feature encoder.** We propose a two-stream CNN to firstly encode image appearance $I$ and motion prior $\mathbf{m}(I)$ separately. To fuse them, we concatenate the

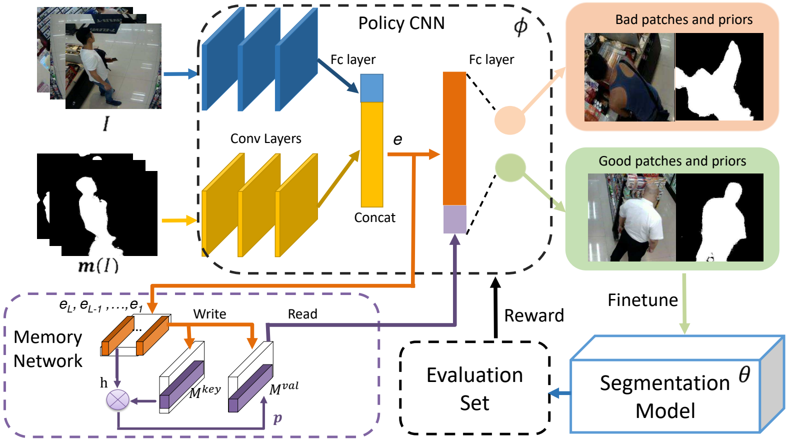

**Fig. 2.** Training Procedure of Policy Model via reinforcement learning. The policy model $\phi$ (consist of policy CNN and memory network) takes both the image $I$ and the motion prior $\mathbf{m}(I)$ as inputs and predicts an action, selecting $\mathbf{m}(I)$ as a good prior or not. The selected priors are further used to improve segmenter $\theta$, and then the improvement shown on a hold-out evaluation set will become a reward to update the policy model $\phi$.

embedded features from two streams. Then, we apply a linear transformation on the concatenated feature to mix the features. Not that it is essential to make our policy network robust to domain shift since it is trained only in source domain but applied in the target domain. We found motion priors are more invariant (relative to RGB images) across domains. Hence, we propose late-fusion and increase the number of features for motion priors.

**Memory network.** There are two operations, "write" and "read", in memory network, which is similar to the architecture proposed in [46].

- Write.
  The encoded features of last $L$ observations are stored into the memory by linear transformation. Two types of memories are represented as *key* and *value*, which are defined as follows,

$$M_k^{key} = W^{key} E_k \tag{2}$$

$$M_k^{val} = W^{val} E_k, \tag{3}$$

  where $M_k^{key}$, $M_k^{val} \in \mathbb{R}^{d \times L}$ are stored memories with embedding dimension $d$, and $k$ is the index of input data order. $W^{key}$ and $W^{val}$ are parameters of writing module. $E_k = \{e_{k-i}\}_{i=1,2,\ldots,L} \in \mathbb{R}^{e \times L}$ is concatenation of last $L$ features of observations which are selected as good priors.
- Read.
  Based on soft attention mechanism, the reading output will be the inner product between the content embedding $h$ and key memories $M_k^{key}$.

$$p_{k,\ell} = \frac{\exp(h_k^\intercal M_k^{key}[\ell])}{\sum_{j=1}^{L} \exp(h_k^\intercal M_k^{key}[j])}, \tag{4}$$

where $h_k = W^h e_k$, and $W^h$ are model parameters for content embedding. $p_{k,\ell}$ is the soft attention for $\ell^{th}$ memory block. Take the attention weights on *value* memories $M_k^{val}$ as the retrieved output, which can be represented as below,

$$o_k = M_k^{val} p_k, \tag{5}$$

where $o_k \in \mathbb{R}^d$ is retrieved memory output.

The memory network is expected to handle the problem of data redundancy, or the policy may tend to select very similar candidates. We concatenate the memory output $o_k$ with current content feature $e_k$ as last features for taking action (select or not).

**4.2.2 Reward.** We use the improved segmentation accuracy on a hold-out set in the source domain as the reward $r$ as follows,

$$r = \mathrm{IoU}(\mathcal{I}_E^{\mathcal{S}}; \theta) - \mathrm{IoU}(\mathcal{I}_E^{\mathcal{S}}; \theta^0) \ , \tag{6}$$

where IoU is the Intersection over Union (IoU) metric which is standard for semantic segmentation, $\theta^0$ is the initial parameters of the human segmentor, $\theta$ is the current parameters of the human segmentor, and $\mathcal{I}_E^{\mathcal{S}}$ is the set of images in the hold-out set in the source domain.

After few earlier episodes, $\mathrm{IoU}(\mathcal{I}_E^{\mathcal{S}}; \theta^0)$ is replaced with other estimated baseline value such as averaged reward in near episodes, in order to maintain learning efficiency.

**4.2.3 Policy Gradient.** According to above reward function, we compute the policy gradient to update the model parameters $\phi$, represented as below,

$$\nabla_\phi \frac{1}{K} \sum_{k=1}^{K} r \cdot \log \pi(a_k \mid I_k, \mathbf{m}(I_k); \phi) \ ; I_k \in \mathcal{V}_T^{\mathcal{S}} \ , \tag{7}$$

where $k$ is the image index, $K = \mid \mathcal{V}_T^{\mathcal{S}} \mid$, and $\mathcal{V}_T^{\mathcal{S}}$ is the set of unlabelled training video frames in source domain.

**4.2.4 Training Procedure.** We conduct the following steps iteratively until the reward and policy loss converge.

- Given $\phi$, we use the policy network to select a set of image (i.e., $\mathcal{K} = \{k; a_k = 1\}$) with motion priors.
- Given $\mathcal{K}$, we use $(I_k, \mathbf{m}(I_k))_{k \in \mathcal{K}}$ as additional pairs of image and ground truth segmentation to finetune the human segmentation parameters $\theta$.

- Given the new $\theta$, we compute the reward $r$ in Eq. 6.
- Given $r$, we compute policy gradient in Eq. 7 and update the policy parameters $\phi$ using Gradient Decent (GD).
- A budget of used data for training the segmenter $\theta$ is defined as $b$, i.e. an episode early stops at step $s$ as $\sum_{k=1}^{s} a_k = b$. Last, we reset the parameters of the segmentor $\theta = \theta^0$ when an episode finishes.

We further extend the procedure above from image-based to patch-based selection. We propose to select motion priors at patch-level since there are very few motion priors which are accurate throughout the entire image. In contrast, there are many patch-based motion priors which are almost completely accurate throughout the entire patch. Next, we define the patch-based selection process.

#### 4.2.5 Patch-based Selection.

Define the $n^{\text{th}}$ patch in an image corresponding to a set of pixels $R_n$, we can write patch-based motion prior as,

$$\mathbf{m}_n = \{m_i; i \in R_n\} \ . \tag{8}$$

The image-based policy gradient in Eq. 7 is modified to,

$$\nabla_\phi \frac{1}{KN} \sum_{k=1}^{K} \sum_{n=1}^{N} r \cdot \log \pi(a_{k,n} \mid I_{k,n}, \mathbf{m}(I_k)_n; \phi) \ , \tag{9}$$

where $I_{k,n}$ denotes the appearance of the $n^{\text{th}}$ patch on the $k^{\text{th}}$ image, $N$ is the number of patches in an image. In order to focus on foreground patches and reduce search space, we also automatically filter out patches with all background motion prior (i.e., $m_i = 0$ for all $i \in R(n)$).

#### 4.2.6 Inference on Target Domain.

We apply the trained policy $\pi$ to select a set of image patches $\mathcal{K}_{\mathcal{T}}$ along with strong motion prior from the unlabeled training frames in the target domain $\mathcal{V}_T^{\mathcal{T}}$. They are referred to as patch-wise *strong* motion prior as below,

$$\mathcal{K}_{\mathcal{T}} = \{(k, n); a_{k,n} = 1\} \ . \tag{10}$$

Given $\mathcal{K}_{\mathcal{T}}$, we use $(I_{k,n}, \mathbf{m}(I_k)_n)_{k \in \mathcal{K}_{\mathcal{T}}}$ as additional pairs of image and ground truth human segmentation and introduce cross-entropy loss for fine-tuning in the target domain. See Fig. 3.

## 5 Experiments

We conduct experiments to validate the proposed weakly-supervised active learning method in cross-modalities and cross-scenes settings. Firstly, the result shows that the proposed policy-based active learning method can select informative

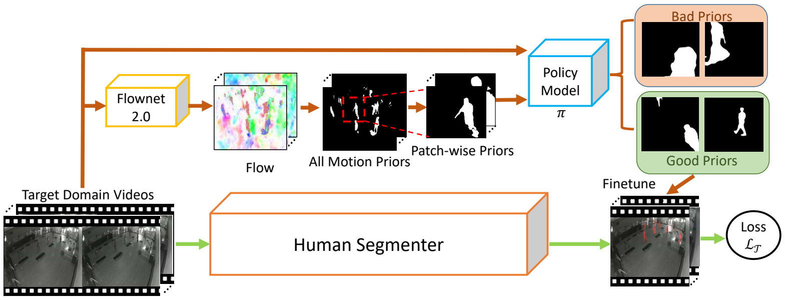

**Fig. 3.** The figure illustrates the extraction and usage of motion prior. Top-half shows the path to generate motion priors from videos, followed by policy model-based selection. Bottom-half shows selected priors for fine-tuning segmenter on target domain.

samples on a new target domain in Sec. 5.2. Moreover, we show the proposed active learning method is complementary to recent adversarial-based domain adaptation frameworks [38,40]. The performance gains of our method integrated with domain adaptation methods are shown in Sec. 5.3.

We demonstrate the weakly-supervised active learning with the cross-domains setting via our collected source datasets *Gym* and *Store* in camera modality-RGB, along with multiple target datasets, including our remaining datasets in camera modality-IR, and one public available pedestrian dataset, *UrbanStreet* [47], which contains 18 stereo sequences of pedestrians taken from a stereo rig mounted on a car driving in the streets of Philadelphia.

### 5.1 Implementation Details

In all experiments, we use U-Net structure [48] as our baseline segmentation model for comparison. The code and models are evaluated in the Pytorch framework. For fair comparisons, we use the Intersection over Union (IoU) [49] as evaluation metrics for all experiments, where IoU $= \frac{TP}{TP+TF+FP}$. The quantitative results in Tables. 3 and 4 show the IoU scores of foreground class. For training our policy model, we use initial learning rate of $1 \times 10^{-4}$ with Adam optimizer [50]. The discount factor for policy gradient is set to 1. We train about 5000 episodes. In the training procedure, an initialized segmenter pre-trained on MSCOCO [51] is further fine-tuned with the policy model.

### 5.2 Weakly-supervised Active Learning with Cross-Domain Setting

We compare our Policy-based Active Learning method (referred to as *PAL*) with two methods: *Random* and *Human Selection* in Table. 3. The number of used motion-prior patches is pre-defined in all settings as a budget $b = 60$. Note that all methods share the same motion prior candidates (cropped patches).

**Table 3.** Cross-domain human segmentation performance (IoU) comparison of the proposed weakly-supervised active learning method "PAL" with other strategies. U- and Seg- denote the model architectures: U-Net and SegNet, respectively. First row "Source Only" is direct application of pre-trained model on target domain data. To best of our knowledge, none of the existing active learning algorithm use only prior instead of true label for fine-tuning on target domain.

| Source | | Gym-RGB | Gym-RGB | Gym-RGB | Store-RGB | Store-RGB | Store-RGB |
|---|---|---|---|---|---|---|---|
| Target | | Gym-IR | Multi-Scene-IR | UrbanStreet(-RGB) | Gym-IR | UrbanStreet(-RGB) | Multi-Scene-IR |
| **Source Only** | (U-) | 48.6% | 16.8% | 48.5% | 26.7% | 61.7% | 29.2% |
| | (Seg-) | 51.1% | 23.6% | 52.3% | 23.6% | 63.5% | 35.8% |
| **PAL** | (U-) | 55.6% | 30.5% | 51.2% | **32.3%** | 64.8% | 34.3% |
| | (Seg-) | **57.0%** | **38.4%** | **56.6%** | 26.9% | **65.3%** | **39.0%** |
| **Random** | (U-) | 52.5% | 26.5% | 49.3% | 29.3% | 62.4% | 30.2% |
| | (Seg-) | 56.7% | 37.2% | 55.3% | 24.8% | 63.4% | 33.2% |
| **Human-** | (U-) | 57.5% | 34.6% | 55.8% | 32.5% | 68.5% | 41.0% |
| **Selection** | (Seg-) | 57.5% | 42.3% | 59.7% | 32.7% | 65.9% | 46.5% |

**Random.** Randomly select a set of motion priors from a data pool. And we report the average results over ten selected sets.

**Human Selection.** We manually select a set of motion priors whose motion priors are closer to true annotations while also considering data divergence. The results can be viewed as an upper bound for our method.

We conduct three kinds of cross-domains applications: (1) cross-modalities, (2) cross-scenes, and (3) cross-modalities & -scenes. The experimental results are summarized in Table. 3. We choose two baseline segmentation models, U-Net and SegNet, to demonstrate generalization of the method. We also provide qualitative results in Fig. 5.

**Cross-modalities in same scene.** In our experiment, we change data in Gym from RGB images to infrared images. In Table. 3, the first column (Gym-RGB to Gym-IR) shows our method "PAL" has +3.1% IoU performance related to random selection and improves +7% IoU from "Source Only" (not using information on target domain).

**Cross-scenes in same modality.** We also validate our proposed method on public available datasets. However, it's hard to find a public dataset with IR videos with segmentation annotations. We replace with a public dataset *UrbanStreet* as the target domain whose appearance is very different from our surveillance camera dataset but captured in same modality (RGB). Our method still works under the condition of great appearance change. We conduct two experiments: Gym-RGB → UrbanStreet and Store-RGB → UrbanStreet showed in Table. 3. The results show +2.7% and +3.1% relative IoU form source model, respectively. Note that UrbanStreet contains many moving vehicles. Our method still can distinguish human motion segments form another moving segments, which may come from cars or slight camera motions. This result demonstrates the robustness of our weakly-supervised active learning approach.

**Cross-scenes and -modalities.** This is the most general situation to deal with for applications of surveillance cameras. We show the results of Gym →

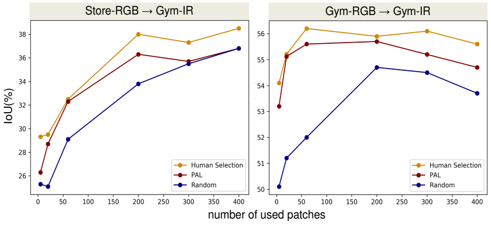

**Fig. 4.** The performance of human segmentations on target domain using our *PAL* method, where the policy-based active learning is trained on Gym-RGB and Store-RGB (Source), respectively, and is applied to Gym-IR (Target). Note that only motion prior (ZERO label) is used for target domain.

**Table 4.** Cross-domain human segmentation performance (IoU) comparison of the proposed method (**bold**) with other baselines in 6 diverse source-target domain pairs. The last two rows show the combined methods outperform each of sub-method, implying the active learning approach is complementary to original domain adaptation framework.

| Source<br>Target | Gym-RGB<br>Gym-IR | Gym-RGB<br>Multi-Scene-IR | Gym-RGB<br>UrbanStreet(-RGB) | Store-RGB<br>Gym-IR | Store-RGB<br>UrbanStreet(-RGB) | Store-RGB<br>Multi-Scene-IR |
|---|---|---|---|---|---|---|
| Source Only | 48.6% | 16.8% | 48.5% | 26.7% | 61.7% | 29.2% |
| **PAL** | 55.6% | 30.5% | 51.2% | 32.3% | 64.8% | 34.3% |
| DSN [38] | 54.3% | 25.9% | 52.6% | 31.8% | 62.3% | 34.4% |
| NMD [40] | 52.1% | 26.1% | 52.1% | 31.7% | 63.1% | 34.5% |
| **PAL+DSN** | **55.8**% | 35.8% | **54.5**% | **36.4**% | **66.2**% | **39.0**% |
| **PAL+NMD** | 55.6% | **36.7**% | **54.5**% | 34.0% | 64.6% | 36.3% |

Multi-scene, Store → Gym and Store → Multi-Scene in Table 3. Note that all settings are from RGB to IR. In all settings, the result shows that PAL offers significant improvement from "Source Only" and better than "Random". In the case of Store-RGB → Gym-IR, the result of our method is very close to the upper bound "Human Selection" with only a 0.2% gap.

The performance curves by exploring incrementally more amounts of priors are shown in Fig. 5.2. We show the effectiveness of PAL comparing with Random and Human selection results. Interestingly, the curve in Store-RGB → Gym-IR implies that the mIoU can increase by adding more strong priors. Since we can obtain motion priors from unlabeled videos with ZERO label cost, our method can be efficient practical to improve performance by simply collecting more unlabeled videos.

### 5.3 Combined with adversarial Domain Adaptation

In this part, we integrate the proposed weakly-supervised active learning with other existing unsupervised domain adaptation (DA) methods for two reasons. Firstly, unsupervised DA shares the same goal of ZERO label cost on target domain. Secondly, intuitively our method should be complementary to unsupervised DA. Most of the unsupervised DA methods only have fine-tuning loss on source domain, since the label is not available on target domain. However, our weakly-supervised active learning policy enables fine-tuning on target domain using the policy-selected strong motion priors.

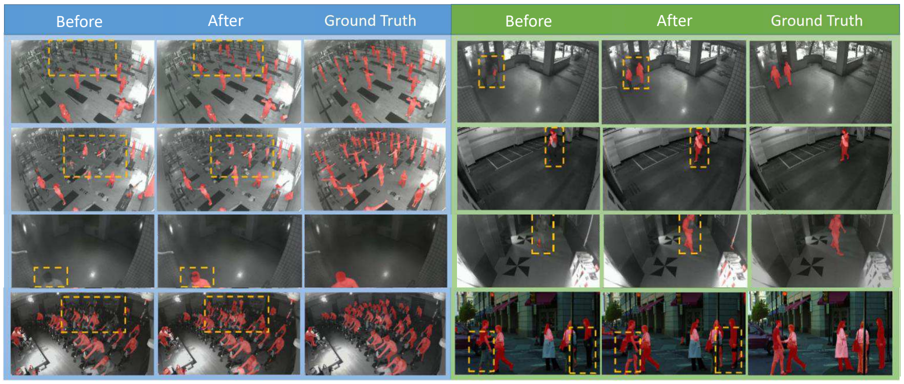

**Fig. 5.** Qualitative results of improving human segmentation on target domain of the following five source-target settings: Store-RGB→Gym-IR (top-left 6 images), Gym-RGB→Multi-Scene-IR (top-right 6 images), and Store-RGB→Multi-Scene-IR (the third row). The last row shows Gym-RGB→Gym-IR and Gym-RGB→UrbanStreet, respectively. The columns "After" denotes improved segmentations by **PAL+NMD**. Bounding-boxes in dash-line highlight the significant change.

On the concern of performance and complexity, we combine proposed PAL with two of existing methods, DSN [38] and NMD [40]. Demonstrating in same cross-domains settings as the previous section, we do the comparison between proposed PAL with these unsupervised domain adaptation baselines, and show these two types approaches (PAL vs. UDA) are complementary with each other since the combined method reach the greatest improvement on target domain. See results in Table. 4. For instance, in the setting Gym-RGB → Multi-Scene IR (second column), the combined method "PAL+NMD" achieve about 6.2% IoU improvements from each sub-approach.

## 6 Conclusion

We propose to leverage "motion prior" in videos to improve human segmentation with cross-domain setting. We propose a memory-network-based policy model to select "strong" motion prior through reinforcement learning. The selected segments have high precision and are used to fine-tune the model on target domain. Moreover, the active learning strategy is shown to be complementary to adversarial-based domain adaptation methods. In a newly collected surveillance camera datasets, we show that our proposed method significantly improves the performance of human segmentation across multiple scenes and modalities.

## 7 Acknowledgment

We thank Umbo CV, MediaTek, MOST 107-2634-F-007-007 for their support.

## References

1. Settles, B.: Active learning literature survey. (2010) 2
2. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: ICLR. (2018) 2, 4
3. Dragon, R., Rosenhahn, B., Ostermann, J.: Multi-scale clustering of frame-to-frame correspondences for motion segmentation. In: ECCV, Springer (2012) 3
4. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE transactions on pattern analysis and machine intelligence **36**(6) (2014) 1187–1200 3
5. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: CVPR, IEEE (2009) 3
6. Yang, M.Y., Ackermann, H., Lin, W., Feng, S., Rosenhahn, B.: Motion segmentation via global and local sparse subspace optimization. arXiv preprint arXiv:1701.06944 (2017) 3
7. Yan, J., Pollefeys, M.: A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In: ECCV, Springer (2006) 3
8. Tsai, Y.H., Yang, M.H., Black, M.J.: Video segmentation via object flow. In: CVPR. (2016) 3
9. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: ICCV. (2017) 3
10. Nirkin, Y., Masi, I., Tuan, A.T., Hassner, T., Medioni, G.: On face segmentation, face swapping, and face perception. In: Automatic Face & Gesture Recognition IEEE International Conference. (2018) 3
11. Zhao, T., Nevatia, R.: Stochastic human segmentation from a static camera. In: Motion and Video Computing, Workshop. (2002) 4
12. Zhao, T., Nevatia, R.: Bayesian human segmentation in crowded situations. In: CVPR. (2003) 4
13. Spina, T.V., Tepper, M., Esler, A., Morellas, V., Papanikolopoulos, N., Falcão, A.X., Sapiro, G.: Video human segmentation using fuzzy object models and its application to body pose estimation of toddlers for behavior studies. arXiv preprint arXiv:1305.6918 (2013) 4
14. Song, C., Huang, Y., Wang, Z., Wang, L.: 1000fps human segmentation with deep convolutional neural networks. In: ACPR, IEEE (2015) 4
15. Guo, L.J., Cheng, T.T., Xiao, B., Zhang, R., Zhao, J.Y.: Video human segmentation based on multiple-cue integration. Signal Processing: Image Communication **30** (2015) 166–177 4
16. Lu, J., Corso, J.J., et al.: Human action segmentation with hierarchical supervoxel consistency. In: CVPR. (2015) 4
17. Tan, Y., Guo, Y., Gao, C.: Background subtraction based level sets for human segmentation in thermal infrared surveillance systems. Infrared Physics & Technology **61** (2013) 230–240 4
18. He, F., Guo, Y., Gao, C.: Human segmentation of infrared image for mobile robot search. Multimedia Tools and Applications (2017) 1–14 4
19. Yarin Gal, R.I., Ghahramani, Z.: Deep bayesian active learning with image data. In: ICML. (2017) 4
20. Colwell, S.R., Joshi, A.W.: Multi-item scale development for measuring institutional pressures in the context of corporate environmental action. In: IABS. (2009) 4

21. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: ICML. (2003) 4
22. Ducoffe, M., Precioso, F.: Adversarial active learning for deep networks: a margin based approach. arXiv preprint arXiv:1802.09841 (2018) 4
23. Xianglin Li, R.G., Cheng, J.: Incorporating incremental and active learning for scene classification. In: ICMLA. (2012) 4
24. Ehsan Elhamifar, Guillermo Sapiro, A.Y., Sastry, S.S.: A convex optimization framework for active learning. In: ICCV. (2013) 4
25. Yang, Y., Loog, M.: A variance maximization criterion for active learning. arXiv preprint arXiv:1706.07642 (2017) 4
26. Christoph Kading, Alexander Freytag, E.R.A.P., Denzler, J.: Large-scale active learning with approximations of expected model output changes. In: GCPR. (2016) 4
27. Kuwadekar, A., Neville, J.: Relational active learning for joint collective classification models. In: ICML. (2011) 4
28. Sujoy Paul, J.H.B., Roy-Chowdhury, A.: Non-uniform subset selection for active learning in structured data. In: CVPR. (2017) 4
29. Fang, M., Li, Y., Cohn, T.: Learning how to active learn: A deep reinforcement learning approach. In: EMNLP. (2017) 4
30. Philip Bachman, A.S., Trischler, A.: Learning algorithms for active learning. In: ICML. (2017) 4
31. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: ICCV. (2015) 5
32. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML. (2015) 5
33. Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., Saminger-Platz, S.: Central moment discrepancy (cmd) for domain-invariant representation learning. In: ICLR. (2017) 5
34. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014) 5
35. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: NIPS. (2016) 5
36. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML. (2015) 5
37. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. arXiv preprint arXiv:1702.05464 (2017) 5
38. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: NIPS. (2016) 5, 11, 13, 14
39. Hoffman, J., Wang, D., Yu, F., Darrell, T.: Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649 (2016) 5
40. Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Wang, Y.C.F., Sun, M.: No more discrimination: Cross city adaptation of road scene segmenters. In: ICCV. (2017) 5, 11, 13, 14
41. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV. (2017) 5
42. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Unsupervised domain adaptation for semantic segmentation with gans. arXiv preprint arXiv:1711.06969 (2017) 5
43. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. arXiv preprint arXiv:1612.01925 (2016) 7

44. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. TPAMI **33**(3) (2011) 500–513 7
45. J. Weston, S.C., networks., A.B.M.: Bordes. memory networks. In: ICLR. (2015) 7
46. Oh, J., Chockalingam, V., Singh, S., Lee, H.: Control of memory, active perception, and action in minecraft. In: ICML. (2016) 7, 8
47. Fragkiadaki, K., Zhang, W., Zhang, G., Shi, J.: Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions. In: ECCV. (2012) 11
48. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, Springer (2015) 11
49. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV **111**(1) (2015) 98–136 11
50. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR. (2015) 11
51. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. (2014) 11