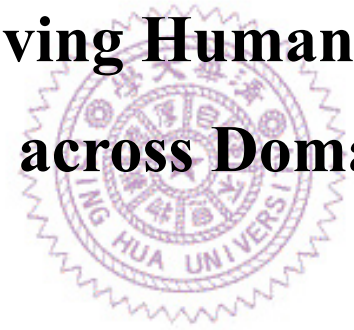
國 立 清 華 大 學

碩 士 論 文

基於影片中移動資訊改善跨域人體目標分割

Leveraging Motion Priors in Videos for Improving Human Segmentation across Domains



系所別：電機工程學系碩士班

學號姓名：105061528 陳玉亭 (Yu-Ting Chen)

指導教授：孫民 博士 (Prof. Min Sun)

中 華 民 國 107 年 6 月

摘要

儘管深度學習近期在語意分割 (sementic segmentation) 已發展成熟，但當一個訓練好的模型遇到真實情況下的測資，因實際資料的特徵分布與訓練集有偏差將造成測試效果不如預期。近年來一些域適應學習 (Domain Adaptation) 與主動學習 (Active Learning) 被提出用來解決上述問題。然而極少的研究強調利用影片的資訊來輔助解決模型跨域表現不佳的情況。

在這篇論文中，我們提出一個一個弱監督式主動學習的方法來改善人體目標分割 (human segmentation)，並利用影片中容易取得的移動資訊 (motion prior)。在我們固定攝影機的情況下，使用光流 (Optical Flow) 得到影片中像素移動的資訊，可將之轉換為前景與背景的分割區塊，前景相當於人體目標的分割區。我們提出以強化學習訓練得到一個基於記憶網路的決策模型，去挑選較佳的前景分割塊。被挑選出的分割塊通常代表有較正確的分割邊界，將被當成訓練目標並直接用來微調模型參數。在評估模型方面，我們蒐集了一個監視攝影機畫面的資料庫，以及在現有公開的資料庫-UrbanStreet 做測試。我們提出的方法改善模型在跨域 (含多場景與多攝影光模態) 的表現。最後，我們的方法可與現有的域適應學習算法結合，協同訓練後達到更好的跨域表現。

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.

誌謝

回顧兩年的碩士生涯，從入門的實驗室讀書會到國際論文的發表，這一路走來收穫甚多受益匪淺，無論是在學術研究或是能力培養都有舉足輕重的影響。首先，我要感謝我的指導教授—孫民老師，在學生遇到瓶頸時，總是可以透過 Facebook 或是 E-mail 和老師即時地討論。在每個禮拜的進度會議學生或許有那麼一點的天馬行空，老師也會以一個 open-minded 的角度適時地給予分析與意見。而在未來生涯規劃上，老師也透過與學生定期的討論，幫助學生及早確立目標並釐清未來方向，我從老師身上學到的不僅是學術知識，更重要的是清楚目標而努力的積極人生態度。

求學過程中，最難忘的總是與同儕一同打拼、在期末報告考試的苦海中互相扶持鼓勵的時光。感謝富翔、奕欣、博丞學長在我碩一時的教導與帶領，讓我能夠快速地進入狀況，也有發表論文的機會。感謝文彥、海侖在做 ECCV 計畫時密切的幫忙與討論，也讓我不至於在趕 deadline 時兵荒馬亂。同時也要感謝 UmboCV 的吳亭範博士一年來的合作與指導，提出許多不同的觀點且適時的對實驗方向做引導，使論文的完成更加順利且有充足的方法與實驗。

本論文最後的完成要感謝口試委員陳煥宗老師及陳祝嵩老師提供寶貴的建議與意見，使得本篇論文更完整且嚴謹。另外，感謝廷安、姿瑩、菀庭、柏瑜、琦雯、家瑀，有了你們的碩士兩年，實驗室的生活點滴顯得更加多彩多姿。希望大家在未來都有個美好的前程，畢業後也保持聯絡。最後我要感謝我的家人一路上的支持與陪伴，提供了我良好的家庭環境，使我能夠無憂無慮地完成學業。這篇論文的完成代表人生中一個階段的結束，也代表著自己已朝下一個階段邁進，在此期許自己未來無論在人際或工作上都能夠更進步，並且把握每一個學習的機會。

Contents

摘要 ii

Abstract iii

誌謝 iv

1 Introduction 1
- 1.1 Motivation and Problem Description . . . 1
- 1.2 Main Contribution . . . 4
- 1.3 Thesis Structure . . . 4

2 Related Work 5
- 2.1 Human Segmentation . . . 5
- 2.2 Motion Segmentation . . . 6
- 2.3 Active learning . . . 6
- 2.4 Domain Adaptation . . . 7

3 Preliminaries 8
- 3.1 U-Net for Semantic Segmentation . . . 8
- 3.2 Adversarial Domain Adaptation . . . 9
  - 3.2.1 Global and Class-wise Domain Shift . . . 10
  - 3.2.2 Global Adversarial Domain Adaptation . . . 10
  - 3.2.3 Class-wise Adversarial Domain Adaptation . . . 11
- 3.3 Policy Gradient . . . 12

4 Surveillance Datasets 14
- 4.1 Cross-domains Settings . . . 15
- 4.2 Data Collection Details . . . 16

5 Policy-based Active Learning in Cross-domain Setting 17
- 5.1 Motion Priors from Video Frames . . . 17
- 5.2 Motion Priors Selection . . . 18
  - 5.2.1 Network Architecture . . . 19
  - 5.2.2 Reinforcement Learning . . . 20
  - 5.2.3 Patch-based Selection. . . . 21
  - 5.2.4 Inference on Target Domain. . . . 22
- 5.3 Combined with Adversarial Domain Adaptation . . . 23
  - 5.3.1 Fine-tuning in Both Domains . . . 23
  - 5.3.2 Full Optimization Problem . . . 24

6 Experiments **25**
6.1 Introduction 25
6.1.1 Additional Dataset 25
6.1.2 Motion Analysis 26
6.2 Implementation Details 26
6.3 Weakly-supervised Active Learning with Cross-Domain Setting 27
6.4 Combined with adversarial Domain Adaptation 31

7 Conclusion **34**

References **35**

List of Figures

1.1 (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. 3

3.1 Architecture of U-Net. The left path is encoder in composition of down-sampling convolutional blocks, while the similar right path is decoder, replaced by up-sampling convolutional blocks. Skip connection (mid bridges) means directly propagate these embedded features from encoder to corresponding decoding layers, where Mean Normalization is applied to. 9

4.1 Examples of unlabeled frames from sequences of different scenes and modalities in our dataset. From first to third rows show Gym-IR, Gym-RGB, and Store-RGB frames in each video. 15

4.2 Examples of unlabeled frame from sequences in one of our datasets which contains visual data of multiple scenes captured by infrared sensors. Here are some video examples of restaurant, walkway, and playground for each row. . 15

5.1 Training Procedure of Policy Model via reinforcement learning. The policy model ϕ (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 θ, and then the improvement shown on a hold-out evaluation set will become a reward to update the policy model ϕ. 18

5.2 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 "good prior" selection. Bottom-half shows selected priors for fine-tuning segmenter on target domain. 22

5.3 Overview of unsupervised adversarial domain adaptation (ADA) framework with additional finetuning loss $\mathcal{L}_{\mathcal{T}}$, introduced by policy-selected samples on target domain. 24

6.1 The performance of human segmentations on target domain using our weakly-supervised active learning methods, comparing to other baselines: Random, Limited flow, and using All patches. The policy-based active learning is trained on Gym-RGB and Store-RGB (source domain), respectively, and is applied to Gym-IR, Multi-Scene-IR, and UrbanStreet (target domain). Note that only motion prior is used for fine-tuning on target domain. 29
6.2 The two sets show examples of image-prior pairs on target domain which are"selected" or "unselected" by the policy. The binary masks represent foreground (white) / background (black) priors generated by magnitudes of optical flow results. 30
6.3 "Before" vs. "After" show improved active learning results on target data of following source-target domain settings: Store-RGB→Gym-IR (first two rows), and Store-RGB→Multi-Scene-IR (last row). Bounding-boxes in dash-line highlight the improvement. 30
6.4 Qualitative results of improving human segmentation on target domain of the following four source-target settings: Store-RGB→Gym-IR (top-left 6 images), Gym-RGB→Multi-Scene-IR (top-right 6 images), Store-RGB→Multi-Scene-IR (images in third row), Gym-RGB→Gym-IR (bottom-left 3 images), and Gym-RGB→UrbanStreet (bottom-right 3 images). In each set, from left to right, we show "before", "after" and "ground truth", respectively. The columns "after" show the prediction of segmenter improved by **PAL+NMD**. Bounding-boxes in dash-line highlight the significant change. Please see supplementary materials for more examples. 33

List of Tables

4.1 Source domain datasets. "Images" refers to the number of images that are labeled. "Videos" refers to the number of videos that contain unlabeled frames. 16

4.2 Target domain datasets. "Images" refers to the number of images that are labeled. "Videos" refers to the number of videos consist of unlabeled frames. Note that there are no labeled training images in the target domain. 16

6.1 Motion Analysis for one RGB dataset and one IR dataset. "foreground" and "background" denote human oracles, "Motion Magnitude" denotes the results of optical flow [1]. The analysis shows $\sim$ 70% foreground has significant motions and $\sim$ 95% background is static. 26

6.2 Cross-domain human segmentation performance (IoU) comparison of the proposed weakly-supervised active learning method "PAL" with other methods (Random, Human Selection). 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. Our method achieves performance close to "Human Selection" which is treated as the upper bound. 28

6.3 Cross-domain human segmentation performance (IoU) comparison of the proposed method (**bold**) with other baselines in 6 diverse source-target domain pairs. Top row "Source Only" is direct application of pre-trained model on target domain data. The third and fourth rows (DSN and NMD) denote the performance of adversarial-based domain adaptation baselines. And 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. 31

Chapter 1

Introduction

1.1 Motivation and Problem Description

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. 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 [2] which have been proven to be effective when applied to classical machine learning models. However, Sener and Savarese [3] have shown that these heuristics are less effective when applied to CNN. To overcome the limitation, Sener and Savarese [3] propose a new active learning method specifically designed for Convolutional Neural Networks

(CNNs). Despite recent advances, all active learning methods mentioned above 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. We refer these candidate segments as "motion prior". Due to the nature of imperfect optical flow, a majority of the segments are quite noisy (see examples in Fig. 1.1). 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

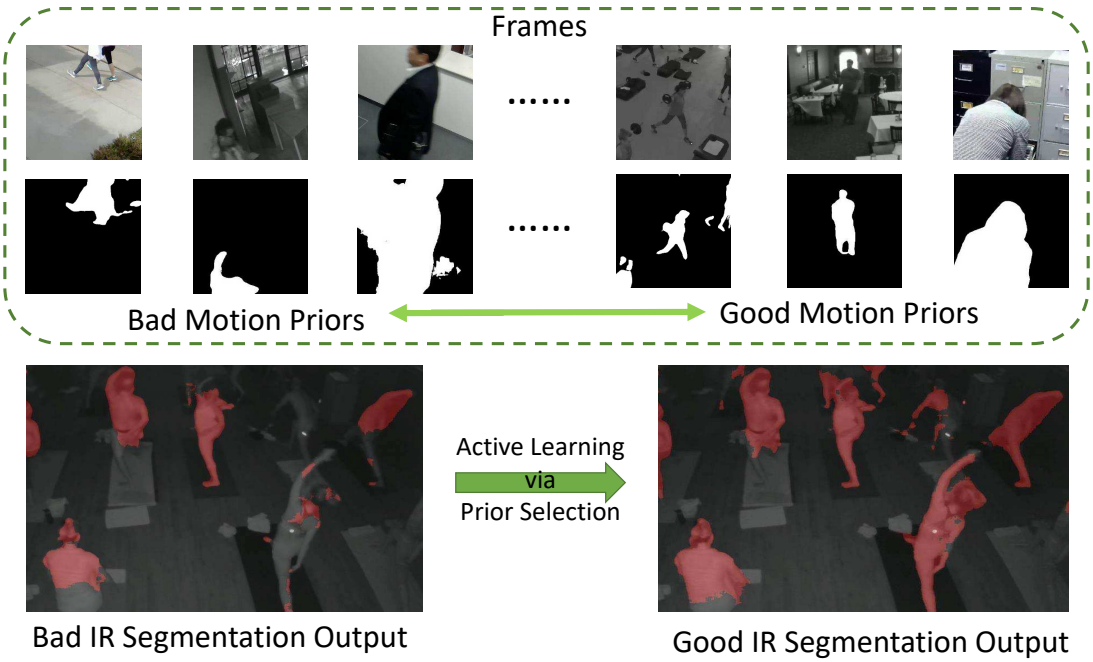


Figure 1.1: (top): RGB patches and their corresponding patch-based motion priors extracted from videos. The priors can be classified into "good" and "bad" ones. (bottom): Our proposed active learning strategy can select good motion priors to improve performance in a cross-modality (RGB to IR) segmentation scenario.

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.

1.2 Main Contribution

The contributions of the paper are summarized below:

- We propose a novel policy-based active learning model for cross-domain setting of human segmentation. It utilizes the property of motion information in videos which is invariant across different modalities.
- We propose to utilize the motion prior of foreground pixels in the target dataset during domain adaptation to further improve the performance.
- Our active learning model is complementary to most existing adversarial domain adaptation methods, achieving significant improvement on our collected datasets and one public pedestrian dataset—UrbanStreet [4]).

1.3 Thesis Structure

In the following chapters, this thesis first describes the related works in Chapter 2 and then presents the preliminaries for model architecture, training algorithms and our previous work in Chapter 3. Second, a new collected surveillance cameras dataset is introduced in Chapter 4. The third part of this thesis, our main technical contribution—policy-based weakly-supervised active learning for strong motion prior selection—is introduced in Chapter. 5. The last part (Chapter 6) provides an extensive discussion in terms of implementation details (Sec. 6.2), quantitative comparisons with other baseline methods, analysis of aggregated system, and qualitative results for our proposed methods (see in Sec. 6.3 and Sec. 6.4).

Chapter 2

Related Work

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

2.1 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 [5, 6]. Spina et al. [7] demonstrate applications in pose estimation and behavior study. On the other hand, in many applications, real-time performance is critical. Song et al. [8] achieve 1000 fps using a CNN-based architecture which outperforms traditional methods in both speed and accuracy.

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. [9] propose a background subtraction based method for human segmentation on thermal infrared images. He et al. [10] 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.

2.2 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 [11, 12] and subspace-based method [13, 14]. Some of the works utilize properties of trajectory data. For example, Yan and Pollefeys [15] use geometric constraint and locality to solve the problem. Other work [16] uses motion segmentation to improve video segmentation results to solve disocclusion caused from hamper frame-to-frame propagation. Recently, [17, 18] propose to jointly tackle the motion segmentation and optical flow tasks. 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.

2.3 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 for the selection strategies. They find hard examples by MC dropout sampling [19], using heuristics like highest entropy [20], or geometric distance to decision boundaries [21, 22].

Other approaches consider the diversity of samples, using k-means algorithms [3,23] or sparse representation for subset selection [24]. Still other important concepts 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], also help the performance of active learning.

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 DA methods focus on learning deep neural network 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 [35–37]. 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.

Chapter 3

Preliminaries

In this chapter, we will give a comprehensive overview of the main topics, including (Sec. 3.1) U-Net structure for semantic segmentation, (Sec. 3.2) a general idea of adversarial domain adaptation, as well as a more specific method as our previous work, and last, (Sec. 3.3) policy gradient method as reinforcement learning which is used as the training algorithm in this work.

3.1 U-Net for Semantic Segmentation

Semantic segmentation is one of important topics in computer vision which predicts the label of each pixel in an image. Fully convolutional network [43] is a popular structure for CNN-based segmentation model, using the spacial-invariant property to reducing parameters of a neuron network. U-Net [44] has been further proposed by Ronneberger et al., with a symmetric encoder-decoder architecture based on fully convolutional layers. See in fig. 3.1, via skip connections between specific encoding and decoding layers, the u-shaped architecture can propagate context information from successive layers to the following layers. That is, the features of encoding layers are directly passed (by the path of skip connections) and simply concatenated with features of decoding layers, the additional information to the encoder can enhance the precision of border region in the output.

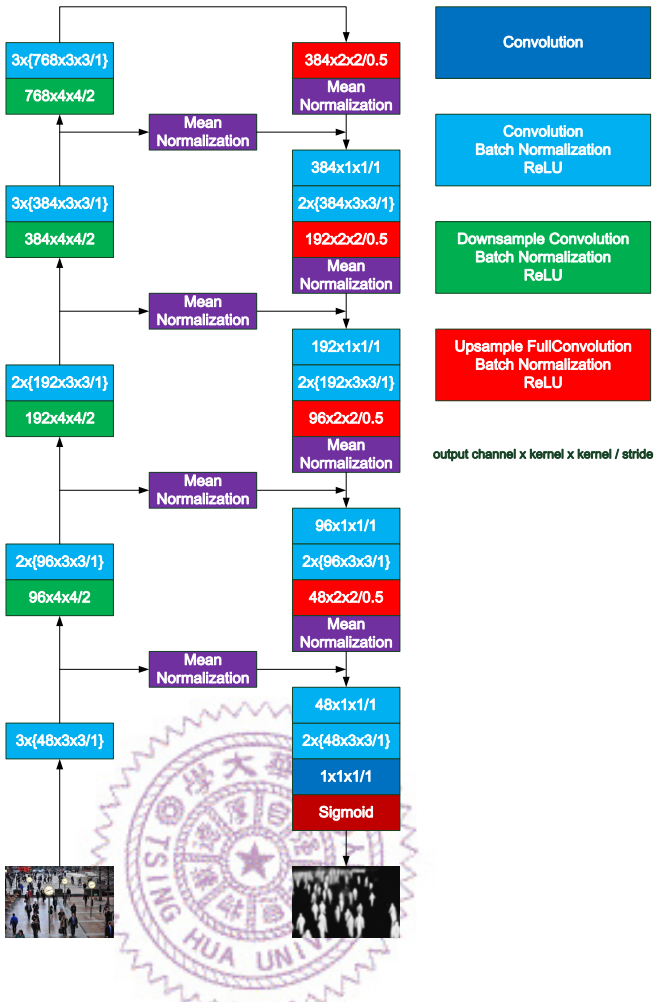


Figure 3.1: Architecture of U-Net. The left path is encoder in composition of down-sampling convolutional blocks, while the similar right path is decoder, replaced by up-sampling convolutional blocks. Skip connection (mid bridges) means directly propagate these embedded features from encoder to corresponding decoding layers, where Mean Normalization is applied to.

3.2 Adversarial Domain Adaptation

In recent years, adversarial-based domain adaptation methods outperform other baseline methods in deep neuron networks. Traditionally, the difference between source and target data on feature space is usually defined by the hand-crafted distance (such as MMD, KL). However, inspired by Generative Adversarial Network (GAN), the idea of competing between two models (generator and discriminator) can also be used to minimize the difference of features between two domains. The propose of the discriminators is to distinguish the resource of data distributions (from source or target). On the

other hand, generators have to "fool" discriminators. More and more researchers use the basic technique to learn the domain-invariant feature, which has been introduced in Sec. 2.4. Then, we explain the resources of domain shift as well as the corresponding model adapting solution.

3.2.1 Global and Class-wise Domain Shift

As adapting image segmenters across domain, two different types of domain shifts (or dataset biases) can be expected: *global* and *class-wise domain shift*. The former comes from the overall differences in appearances between data sources, while the latter is due to distinct compositions of various categories in each domain.

In this section, firstly, we give the preliminary about global Adversarial Domain Adaptation (ADA). Then, we describe our previous work [40], introducing a unified framework which both global and class-wise adaptation use adversarial-based approach.

Firstly, we define some notations in common use.

Indice: we use i, n, c to index pixel, patch, and semantic class . **2D feature map:** g indicate soft (pseudo) class label map in feature level, and g_n^c denotes accessing the soft (pseudo) label for class c of the n^{th} patch.

3.2.2 Global Adversarial Domain Adaptation

There are three main components in global ADA: feature extractor φ_F, label predictor φ_Y, and domain classifier φ_D. Note that we take semantic segmentation model for example such that the output after extractor φ_F should be a map, instead of single instance. That is, the extractor's task is to compute a feature map $\varphi_F(I;\theta_F)$, where I is an input image, θ_F is the parameters of the extractor. Then, the label predictor is used to predict human probability map $\mathbf{y} = \varphi_Y(\varphi_F(I;\theta_F);\theta_Y)$, where $\mathbf{y} = \{y_i \in [0,1]\}_i$ is a set of per pixel human probability and θ_Y is the parameters of the predictor.

In order to learn invariant features across source and target domains, the domain classifier is introduced to predict the probability $p_n(I)$ that a neuron on a feature map is

from the source domain as follows,

$$p_n(I) = \varphi_D(\varphi_F(I, \theta_F)_n, \theta_D) \in [0, 1] \ , \tag{3.1}$$

where n is the neuron index, $\varphi_F(I, \theta_F)_n$ indicates the feature corresponding to the n^{th} neuron, θ_D is the parameters of the domain classifier. During adaptation, the goal of the extractor is to confuse the domain classifier, whereas the goal of the classifier is to differentiate domains. Hence, the following optimization problem is defined:

$$\begin{aligned} &\max_{\theta_F} \min_{\theta_D} \ \mathcal{L}_{global}(\theta_F, \theta_D) = \\ &- \sum_{I^{\mathcal{S}}} \sum_{n \in N} log(p_n(I^{\mathcal{S}})) - \sum_{I^{\mathcal{T}}} \sum_{n \in N} log(1 - p_n(I^{\mathcal{T}})) \ , \end{aligned} \tag{3.2}$$

where $I^{\mathcal{S}}$ and $I^{\mathcal{T}}$ are images from the training set of the source domain and the target domain, respectively. The max-min optimization problem can be solved by an inserted gradient reversal layer proposed in [36].

3.2.3 Class-wise Adversarial Domain Adaptation

In previous work, with Chen et al. [40], we consider Class-wise ADA as an extension of global ADA by introducing a domain classifier for each semantic class. The formulation is the following optimization problem. Given the soft class assignment $\{g_n^c\}_n$ for all neurons,

$$\begin{aligned} &\max_{\theta_F} \min_{\{\theta_D^c\}_c} \ \mathcal{L}_{class}(\theta_F, \{\theta_D^c\}_c) = \\ &- \sum_{I^{\mathcal{S}}} \sum_{c \in \mathcal{C}} \sum_{n \in N} g_n^c(I^{\mathcal{S}}) log(p_n^c(I^{\mathcal{S}})) - \sum_{I^{\mathcal{T}}} \sum_{c \in \mathcal{C}} \sum_{n \in N} g_n^c(I^{\mathcal{T}}) log(1 - p_n^c(I^{\mathcal{T}})) \ , \end{aligned} \tag{3.3}$$

where $\{\theta_D^c\}_c$ is the set of parameters of the class-wise domain classifiers, $\mathcal{C}$ is the set of semantic classes, and p_n^c is the probability that a neuron on a feature map is from the source domain according to the c^{th} class-wise domain classifier. We define p_n^c as

follows,

$$p_n^c(I) = \varphi_D^c(\varphi_F(I, \theta_F)_n, \theta_D^c) \in [0, 1] \ , \tag{3.4}$$

where n is the neuron index, $\varphi_F(I, \theta_F)_n$ indicates the feature corresponding to the n^{th} neuron. Note that $g_n^c(I^{\mathcal{S}})$ is the class assignment for a source domain image which is computed using ground truth class label as described in [40]. In contrast, $g_n^c(I^{\mathcal{T}})$ is the class distribution assigned for a target domain image by the prediction of segmentation model. Similar to global ADA, the class-wise ADA problem is optimized using gradient reversing method as well.

We also observe that background class dominates most pixels in our scenes. To overcome the class unbalance problem, we normalize the pseudo labels in Eq. (3.5) as below,

$$\hat{g}_n^c(I) = \frac{g_n^c(I)}{\sum\limits_{n \in N} g_n^c(I)} \ . \tag{3.5}$$

For background class, the denominator is large since it dominates most pixels in each image. If we treat the soft assignments as weights, the normalized weights for background class will be attenuated. Eq. (3.5) is used in both source and target domains. The modified weighting $\hat{g}_n^c$ replaces g_n^c in Eq. (3.3) as the final class-wise domain loss.

3.3 Policy Gradient

In this work, we try to train a policy model to select useful data-prior pairs. The sequence of selection can be formulated as a Markov Decision Process (MDP) and solved by reinforcement learning [45], where the policy model can be viewed as an *agent*, interacting with the external environment (the image and the prior received as input at every time step).

At time t, The *agent* takes an *action* a according to the current *policy* π_θ and current *state* s. Here *policy* is defined by the parameters of *agent* (replaced by a DNN) and

an action is sampled from the probability of predicted actions at each time step. $a_t \sim \pi(a_t|s_t;\theta_t)$. The agent updates the state after choosing an action.

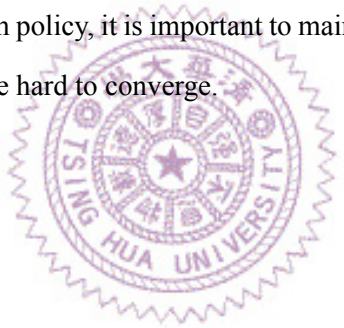
Once the agent has reached the end of a sequence, called an episode, it receives a *reward* r. The reward is used to train our policy. Since these steps of sampling actions are in-differentiable, REINFORCE [46] algorithm can be applied to solve it. The training objective is to find the optimal policy $\pi_\theta^\star$ that can maximize the expected reward in an episode:

$$J(\theta) = \mathbb{E}_{(s,a)\sim\pi_\theta(s,a)} \sum_t r(s_t, a_t) \ , \tag{3.6}$$

By the derivation in REINFORCE algorithm, we can update the parameter by the approximated gradients:

$$\nabla_\theta J(\theta) = \mathbb{E}_{(s,a)\sim\pi_\theta(s,a)}[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t) \cdot \nabla_\theta \log \pi_\theta(a_t \mid s_t)] \ . \tag{3.7}$$

Since we start with a random policy, it is important to maintain the size of searching space at the start of learning, or it will be hard to converge.



Chapter 4

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. The drastic appearance shift from RGB to IR has lead to significant performance drop in both detection and segmentation. In our qualitative test, none of the commercial detectors perform well for human detection or segmentation in infrared domain. We speculate it is because most of them are trained from images collected from the Internet which are dominated by color(RGB) images. In addition to sensor modalities, performance drops are often observed between different scenes. For example, a model trained from outdoor images often performs poorly for indoor scenes. 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 (See examples in Fig. 4.1 and Fig. 4.2.): Gym-RGB, Gym-IR, Store-RGB , 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

Figure 4.1: Examples of unlabeled frames from sequences of different scenes and modalities in our dataset. From first to third rows show Gym-IR, Gym-RGB, and Store-RGB frames in each video.

Figure 4.2: Examples of unlabeled frame from sequences in one of our datasets which contains visual data of multiple scenes captured by infrared sensors. Here are some video examples of restaurant, walkway, and playground for each row.

collected by stationary cameras, we label a subset of frame sparsely sampled from each video.

4.1 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. 5.1). 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. 5.2). 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. 4.1 and 4.2, respectively.

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

Table 4.1: 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 4.2: Target domain datasets. "Images" refers to the number of images that are labeled. "Videos" refers to the number of videos consist of unlabeled frames. Note that there are no labeled training images in the target domain.

4.2 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.

Chapter 5

Policy-based Active Learning in Cross-domain Setting

We describe how to obtain motion prior from optical flow (Sec. 5.1) and select a set of *strong* motion prior (Sec. 5.2). Last, the usage on cross-domain setting is illustrated in Sec. 5.3.

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. **m** indicates motion prior, and m_i denotes the motion prior of the i^{th} pixel.

5.1 Motion Priors from Video Frames

Our goal is to obtain a set of motion prior **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 [1]. Then, we obtain **m** as the binarized flow map such that $m_i = 1$ if its flow magnitude is larger than a threshold τ. 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., [47]) 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. 5.2.

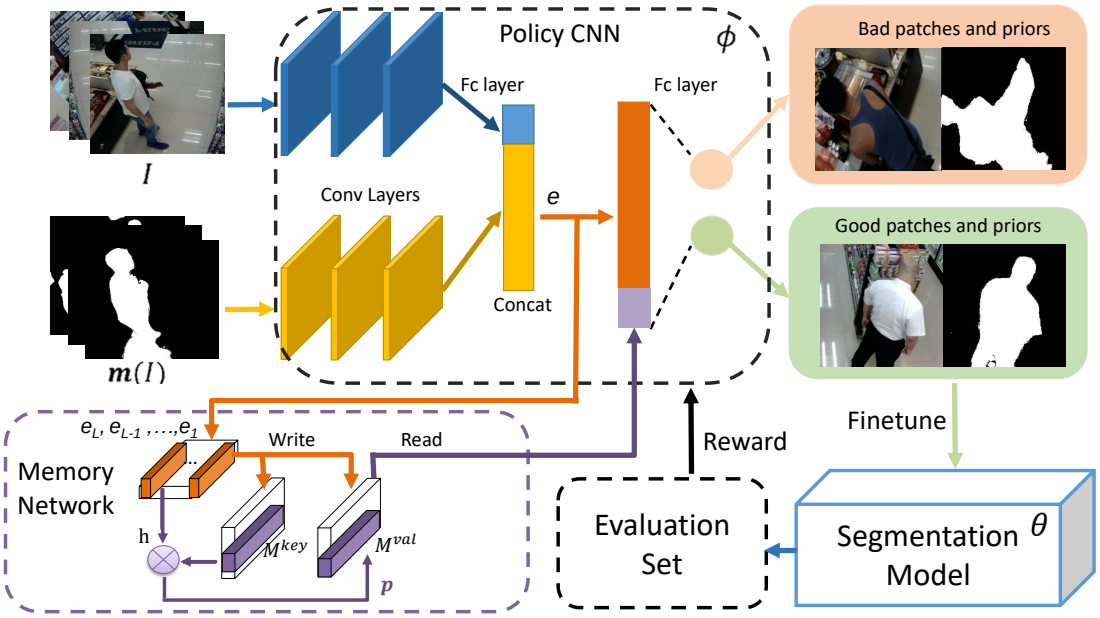


Figure 5.1: Training Procedure of Policy Model via reinforcement learning. The policy model ϕ (consist of policy CNN and memory network) takes both the image I and the motion prior $\mathbf{m}(I)$ as inputs and predicts an action, selecting $\mathbf{m}(I)$ as a good prior or not. The selected priors are further used to improve segmenter θ, and then the improvement shown on a hold-out evaluation set will become a reward to update the policy model ϕ.

5.2 Motion Priors Selection

We train a policy model π which learns to select a set of *strong* motion priors. Further, these *strong* motion priors are treated as ground truth to fine-tune our model using cross-entropy loss directly. In our purpose, The meta-model π is able to "actively" select the useful data. The active learning process is viewed as a sequence of decisions (select or not) and formed as a Markov Decision Process (MDP).

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. 5.1.

Policy model. We define the policy π as the following probability function:

$$\pi(a|I, \mathbf{m}(I); \phi) \,, \tag{5.1}$$

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 ϕ is the model parameters.

5.2.1 Network Architecture

Inspired by the ideal using Memory Network [48] in Deep Q-Network (DQN) proposed by Oh at el. [49], 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 embedded features from two streams. Then, we apply a linear transformation on the concatenated feature to mix the features. Note that we design the motion prior feature to have a larger dimension than the appearance feature since there is less domain shift in motion priors than in appearance.

Memory network. There are two operations, "write" and "read", in memory network, which is similar to the architecture proposed in [49].

- 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,

  $$M_k^{key} = W^{key} E_k \tag{5.2}$$

  $$M_k^{val} = W^{val} E_k, \tag{5.3}$$

  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. $E_k = \{e_{k-i}\}_{i=1,2,...,L} \in \mathbb{R}^{e \times L}$ 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}.

  $$p_{k,\ell} = \frac{\exp(h_k^\intercal M_k^{key}[\ell])}{\sum_{j=1}^{L} \exp(h_k^\intercal M_k^{key}[j])}, \tag{5.4}$$

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

$$o_k = M_k^{val} p_k, \tag{5.5}$$

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.

5.2.2 Reinforcement Learning

Reward.

The training objective for learning the policy takes the form of a scalar "reward", which gives feedback on the quality of the actions made by the agent. We use the improved segmentation accuracy on a hold-out set in the source domain as the reward r as follows,

$$r = \text{IoU}(\mathcal{I}_E^{\mathcal{S}}; \theta) - \text{IoU}(\mathcal{I}_E^{\mathcal{S}}; \theta^0) \ , \tag{5.6}$$

where IoU is the Intersection over Union (IoU) metric which is standard for semantic segmentation, θ^0 is the initial parameters of the human segmentor, θ 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, $\text{IoU}(\mathcal{I}_E^{\mathcal{S}}; \theta^0)$ is replaced with other estimated baseline value such as averaged reward in near episodes, in order to maintain learning efficiency.

Policy Gradient.

According to above reward function, we compute the policy gradient to update the model parameters ϕ, represented as below,

$$\nabla_\phi \frac{1}{K} \sum_{k=1}^{K} r \cdot \log \pi(a_k \mid I_k, \mathbf{m}(I_k); \phi) \ ; I_k \in \mathcal{V}_T^{\mathcal{S}} \ , \tag{5.7}$$

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.

Training Procedure.

We conduct the following steps iteratively until the reward and policy loss converge.

- Given ϕ, 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 θ.
- Given the new θ, we compute the reward r in Eq. (5.6).
- Given r, we compute policy gradient in Eq. (5.7) and update the policy parameters ϕ using Gradient Decent (GD).
- A budget of used data for training the segmenter θ 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.

5.2.3 Patch-based Selection.

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

$$\mathbf{m}_n = \{m_i; i \in R_n\} \,. \tag{5.8}$$

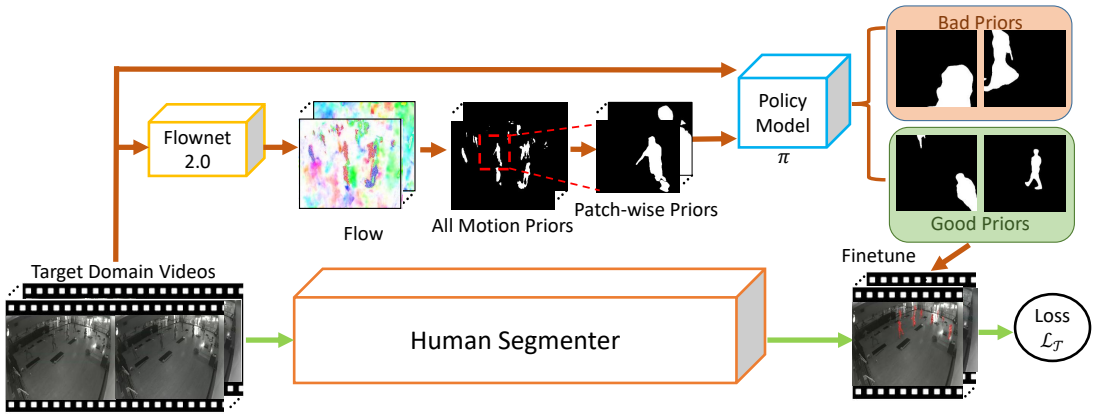


Figure 5.2: The figure illustrates the extraction and usage of motion prior. Top-half shows the path to generate motion priors from videos, followed by policy model based "good prior" selection. Bottom-half shows selected priors for fine-tuning segmenter on target domain.

The image-based policy gradient in Eq. (5.7) is modified to,

$$\nabla_\phi \frac{1}{KN} \sum_{k=1}^{K} \sum_{n=1}^{N} r \cdot \log \pi(a_{k,n} \mid I_{k,n}, \mathbf{m}(I_k)_n; \phi) \,, \tag{5.9}$$

where $I_{k,n}$ denotes the appearance of the n^{th} patch on the k^{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 $i \in R(n)$).

5.2.4 Inference on Target Domain.

We apply the trained policy π 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,

$$\mathcal{K}_\mathcal{T} = \{(k, n); a_{k,n} = 1\} \,. \tag{5.10}$$

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. 5.2.

5.3 Combined with Adversarial Domain Adaptation

We turn to the question of learned policy transferring to another target dataset. Firstly, the adversarial domain adaptation has been introduced in Preliminary (Chap. 3). In this section, we show the proposed active learning policy can cooperate with adversarial domain adaptation (ADA) approaches as a complete system. They share the same goal that utilizes the sufficient training data on source domain to improving performance on target domain, which is lack of human label.

5.3.1 Fine-tuning in Both Domains

As we have described, unsupervised ADA setting is lack of ground truth label on target domain, so it's hard for fine-tuning a classifier for target data. However, the strong motion priors obtained from policy model (in Sec. 5.2) can be used as ground truth for fine-tuning while conducting ADA.

Given pairs of images and ground truth segment data, the segmentation model can be fine-tuned by minimizing the cross-entropy loss. Recall that we have obtained patch-wise *strong* motion prior $\mathcal{K}_\mathcal{T}$ in the target domain as defined in Eq. (5.10). In the source domain, we already have pairs of image $I_\mathcal{S}$ and ground truth class label $\mathbf{c}$. We combine these data from both source and target domains and define the finetuning loss below,

$$\begin{aligned} \max_{\theta_F,\theta_Y} \mathcal{L}_\mathcal{S}(\theta_F,\theta_Y) + \mathcal{L}_\mathcal{T}(\theta_F,\theta_Y)\ , \\ \mathcal{L}_\mathcal{S}(\theta_F,\theta_Y) = \sum_{I^\mathcal{S}\in\mathcal{S}} \sum_{i\in P} \log(y_i^{c_i}(I^\mathcal{S})) \\ \mathcal{L}_\mathcal{T}(\theta_F,\theta_Y) = \sum_{(k,n)\in\mathcal{K}_\mathcal{T}} \sum_{i\in R_n} \log(y_i^{m_i}(I_{k,n}))\ , \end{aligned} \tag{5.11}$$

where both $\mathcal{L}_\mathcal{S}$ and $\mathcal{L}_\mathcal{T}$ are the minus cross-entropy loss to be maximized, c_i is the ground truth class label in $\mathbf{c}$, R_n is the set of pixel indice corresponding to the n^{th} patch, and m_i is the motion prior of the i^{th} pixel. Note that $\mathcal{L}_\mathcal{S}$ and $\mathcal{L}_\mathcal{T}$ are the only two loss functions related to the parameters θ_Y of the predictor. Hence, only via finetuning, the predictor can be directly adapted.

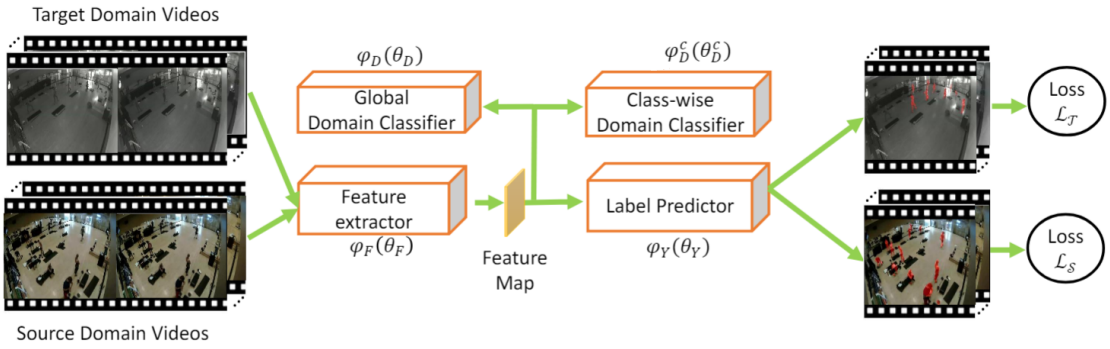


Figure 5.3: Overview of unsupervised adversarial domain adaptation (ADA) framework with additional finetuning loss $\mathcal{L}_\mathcal{T}$, introduced by policy-selected samples on target domain.

5.3.2 Full Optimization Problem

We bring together all the loss functions in Eq. 3.2, 3.3 (Global and Class-wise domain loss), and Eq. 5.11 (finetuning loss) and rewrite the complete optimization problem below,

$$\max_{\theta_F}\{\max_{\theta_Y}(\mathcal{L}_\mathcal{S}(\theta_F,\theta_Y)+\lambda_T\mathcal{L}_\mathcal{T}(\theta_F,\theta_Y))+ \\ \min_{\theta_D}\lambda_G\mathcal{L}_{global}(\theta_F,\theta_D)+\min_{\{\theta_D^c\}_c}\lambda_C\mathcal{L}_{class}(\theta_F,\{\theta_D^c\}_c)\}\ , \tag{5.12}$$

where $\mathcal{L}_{global}$ and $\mathcal{L}_{class}$ are domain losses for global and class-wise parts, and λ_T, λ_G, λ_G are relative ratios to supervised loss on source domain. The complete optimization problem is solved using stochastic gradient descent (SGD), along with gradient reversal for min-max problem. The overview of our framework shows in Fig.5.3. In order to show the proposed policy-based active learning is complementary to general ADA method, not only try on ours, we choose more than one ADA method as baselines. More details will be given in the chapter of experiments (Sec. 6.4).

Chapter 6

Experiments

6.1 Introduction

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 samples on a new target domain in Sec. 6.3. 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. 6.4.

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 [4]. Moreover, we do analysis to show how the efficiency of motions are used on our datasets.

6.1.1 Additional Dataset

UrbanStreet. A public available dataset [4] contains 18 stereo sequences of pedestrians taken from a stereo rig mounted on a car driving in the streets of Philadelphia. The image resolution is 516x1024. All pedestrians larger than 100 pixels in height are labelled

every 4 frames (0.6 seconds) in each video sequence.

Table 6.1: Motion Analysis for one RGB dataset and one IR dataset. "foreground" and "background" denote human oracles, "Motion Magnitude" denotes the results of optical flow [1]. The analysis shows $\sim 70\%$ foreground has significant motions and $\sim 95\%$ background is static.

	foreground		background	
Motion Magnitude	$<= 1$	> 1	$<= 1$	> 1
Store-RGB	22.8%	77.2%	96.0%	4.0%
Multi-Scene-IR	31.9%	68.1%	96.3%	3.7%

6.1.2 Motion Analysis

We provide some dataset statistic in Table. 6.1. Ideally, under the assumption of captured by stationary indoor surveillance cameras, areas with largest motions should be considered as foreground, and vise versa. Note that motions are captured by recent Optical Flow method [1]. In the motion analysis, we found most pixels follow the rule (reference to human oracles), to make sure the usage of motions is efficient enough.

6.2 Implementation Details

In all experiments, we use U-Net structure [44] 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) [50] as evaluation metrics for all experiments, where IoU $= \frac{TP}{TP+TF+FP}$. The quantitative results in Tables. 6.2 and 6.3 show the IoU scores of foreground class.

pre-training on source. The architecture of our segmentation model consists of CNN-based encoder and decoder, with hidden dimension 512. The image features are extracted using the MSCOCO-pretrained VGG-16 [51]. Firstly, we fintune the segmenter on source dataset via cross entropy loss, using Adam optimizer [52] with learning rate 1×10^{-3}, $beta1 = 0.9$, and $beta2 = 0.999$. A decay factor of 0.955 for learning rate is applied by every epoch. We select the best model via validation set and use it as initial performance during adaptation setting.

training the policy model. We train our policy-based active learning model by REINFORCE [45], using initial learning rate of 1×10^{-4} with Adam optimizer, and 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 [53] is further fine-tuned with the policy model. The segmenter is reset to initial weights at the start of each episode. Note that we only remain the learned policy model after this stage.

Adversarial training. We use Adam optimizer with learning rate of 1×10^{-5} for fine-tuning loss and set the hyper-parameters for each loss weight in Eq. (5.12): $\lambda_{global} = 1$, $\lambda_{class} = 0.1$ and $\lambda_T = 0.1$. The reverse coefficient u in gradient reversal layer is gradually changing from 0 to m in the schedule: $u_p = \frac{2m}{1+\exp(-\gamma \cdot p)} - m$, where $m = 0.1$ for global and 0.01 for class-wise part. γ was set to 10 in all experiments and p is the ratio of current step over expected total steps for adaptation.

6.3 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. 6.2. 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).

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.6.2. We also provide typical examples of policy-selected patches (Fig. 6.2) and qualitative results of active learning (Fig. 6.3) on target domains.

Table 6.2: Cross-domain human segmentation performance (IoU) comparison of the proposed weakly-supervised active learning method "PAL" with other methods (Random, Human Selection). 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. Our method achieves performance close to "Human Selection" which is treated as the upper bound.

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	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%**
Random	52.5%	26.5%	49.3%	29.3%	62.4%	30.2%
Human Selection	57.5%	34.6%	55.8%	32.5%	68.5%	41.0%

Cross-modalities in same scene. In our experiment, we change data in Gym from RGB images to infrared images. In Table. 6.2, 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. For the reason, 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 proposed method still works under the condition of great appearance change. We conduct two experiments: Gym-RGB $\rightarrow$ UrbanStreet and Store-RGB $\rightarrow$ UrbanStreet showed in Table. 6.2. 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 $\rightarrow$ Multi-scene, Store $\rightarrow$ Gym and Store $\rightarrow$ Multi-Scene in Table 6.2. Note that all settings are from RGB to IR. In all settings, the result shows that PAL offers significant improvement from "Source

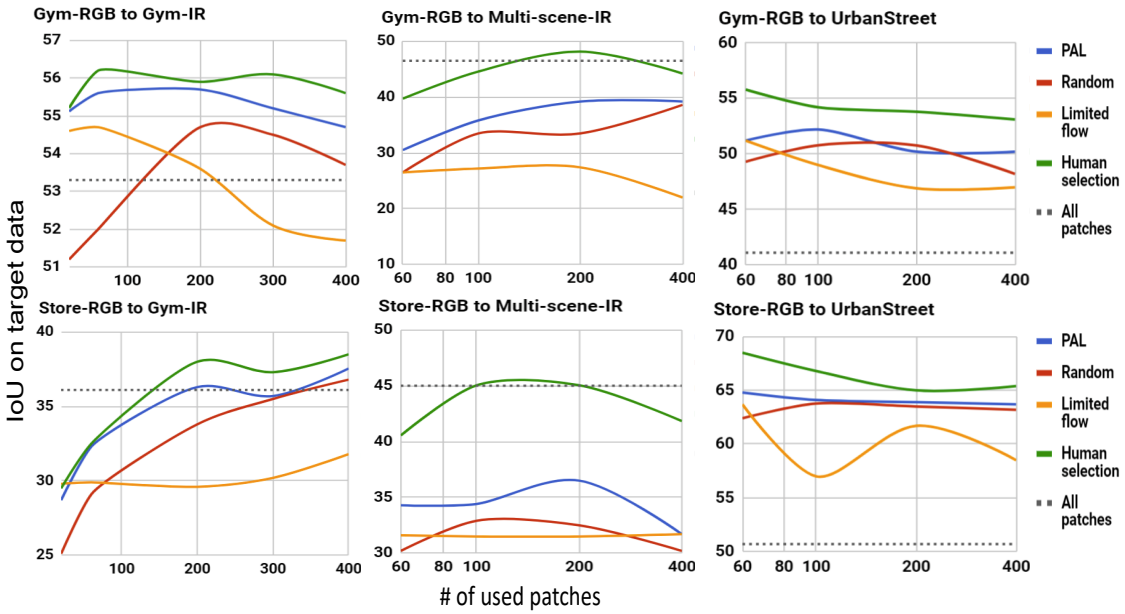


Figure 6.1: The performance of human segmentations on target domain using our weakly-supervised active learning methods, comparing to other baselines: Random, Limited flow, and using All patches. The policy-based active learning is trained on Gym-RGB and Store-RGB (source domain), respectively, and is applied to Gym-IR, Multi-Scene-IR, and UrbanStreet (target domain). Note that only motion prior is used for fine-tuning on target domain.

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. 6.1. We show the effectiveness of PAL results, comparing with *Random, Limited flow, All patches,* and the upper bound "Human selection". *Limited flow* is a simple heuristics considering the motion statistics in our dataset: takes b patches with the largest average flow under a value of 0.5. In the intuition of picking salient patches (largest motion) but filtering out those caused by *light change*, since the *light change* dominates the most area of motion, introducing a large averaged value of flow. *All patches* denotes using all previous cropped patches without selecting. Interestingly, the curve in Store-RGB → Gym-IR and Store-RGB → Multi-Scene-IR imply 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. In the cases of transferring to Multi-Scene-IR as the target domain, using all patches achieves the highest

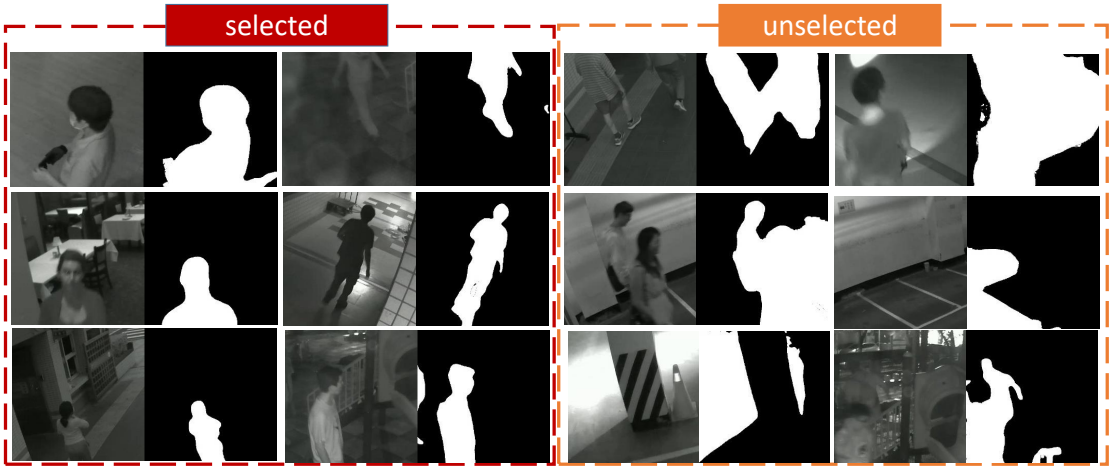


Figure 6.2: The two sets show examples of image-prior pairs on target domain which are"selected" or "unselected" by the policy. The binary masks represent foreground (white) / background (black) priors generated by magnitudes of optical flow results.

performance. We think that for a great amount of data, the segmentation model is able to distinguish the noise from ground truth. For instance, fake foreground segments caused by the light change or broken foreground segments caused by static human are *noise* we call here. For other cases, our method outperforms *using all patches* in a large margin.

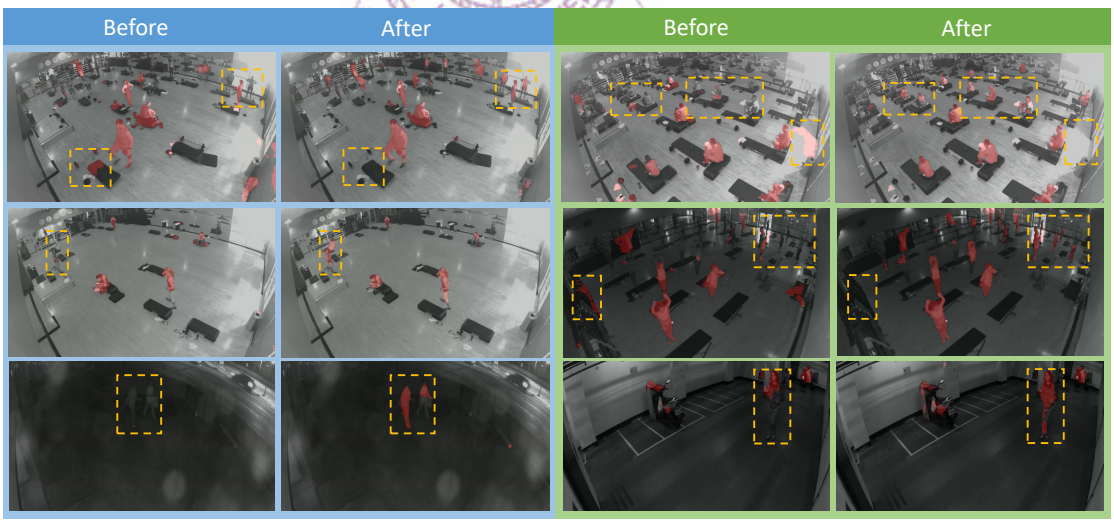


Figure 6.3: "Before" vs. "After" show improved active learning results on target data of following source-target domain settings: Store-RGB→Gym-IR (first two rows), and Store-RGB→Multi-Scene-IR (last row). Bounding-boxes in dash-line highlight the improvement.

Table 6.3: Cross-domain human segmentation performance (IoU) comparison of the proposed method (**bold**) with other baselines in 6 diverse source-target domain pairs. Top row "Source Only" is direct application of pre-trained model on target domain data. The third and fourth rows (DSN and NMD) denote the performance of adversarial-based domain adaptation baselines. And 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	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	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%

6.4 Combined with adversarial Domain Adaptation

In this part, we integrate the proposed weakly-supervised active learning with other existing unsupervised domain adaptation (UDA) methods, in the same goal of ZERO label cost on target domain. 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.

On the concern of performance and complexity, we combine proposed PAL with two of existing methods, DSN [38] and NMD [40].

DSN. Domain Separation Networks, a UDA method proposed by Bousmalis et al., (1) introduces additional reconstruction loss and (2) separates the representation into domain-transferable (-common) features and domain-private features in both domains. We simply extend the structure for segmentation tasks since the original model is only applied for classifications.

NMD. Our previous work, a UDA framework unifies Global and Class-wise adaptations in adversarial learning, especially for semantic segmentation.

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. 6.3.

For the direct comparison with UDA baselines (reference to 2^{th} sin 4^{th} rows in Table. 6.3), in most cases, our method has better performance, especially for the modality change. For instance, in experiment of Gym-RGB $\rightarrow$ Multi-Scene-IR, PAL outperforms DSN by 4.6% IoU gap. It also implies that motion informations are really helpful for improving IR prediction. In other hands, for the cases of scene change (ex. Gym-RGB to UrbanStreet-RGB), our method achieve comparable performance with UDA baselines.

For further experiments, we aggregate the proposed PAL methods with UDA baselines. The result (see last two rows vs . 2^{th} sin 4^{th} rows in Table. 6.3) shows that aggregated framework outperform each sub-approach (PAL or UDA) in visible margin. For instance, in the setting Gym-RGB $\rightarrow$ Multi-Scene IR (second column), the combined method "PAL+NMD" achieve about 6.2% IoU improvements from each sub-approach. The qualitative results adapted by aggregated framework can be found in Fig. 6.4.

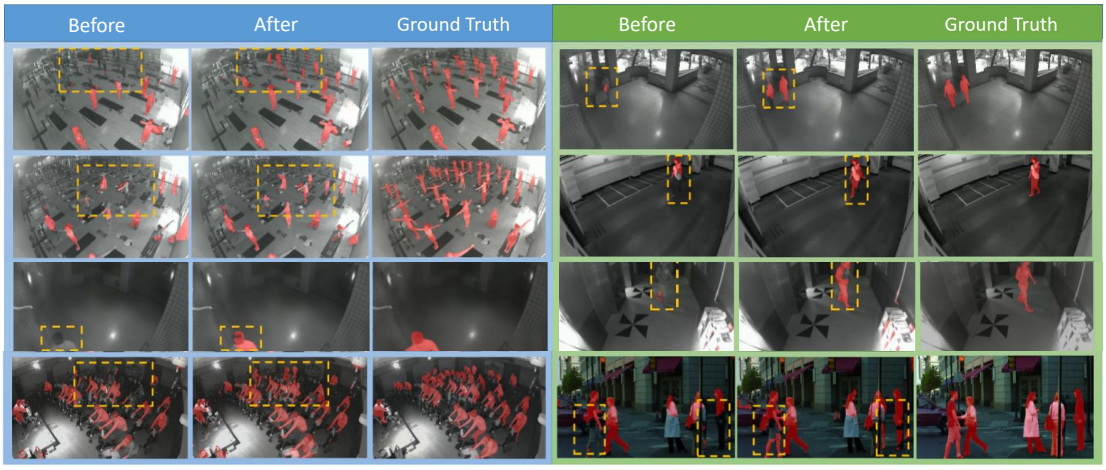


Figure 6.4: Qualitative results of improving human segmentation on target domain of the following four source-target settings: Store-RGB→Gym-IR (top-left 6 images), Gym-RGB→Multi-Scene-IR (top-right 6 images), Store-RGB→Multi-Scene-IR (images in third row), Gym-RGB→Gym-IR (bottom-left 3 images), and Gym-RGB→UrbanStreet (bottom-right 3 images). In each set, from left to right, we show "before", "after" and "ground truth", respectively. The columns "after" show the prediction of segmenter improved by **PAL+NMD**. Bounding-boxes in dash-line highlight the significant change. Please see supplementary materials for more examples.

Chapter 7

Conclusion

Acquiring labeled data to train a model in supervised learning can be difficult and expensive in a new target domain. Some domain adaptation methods have been proposed to deal with dataset shift issues.

In the same goal, we propose to leverage "motion prior" in videos to improve human segmentation with the cross-domain setting. Due to the motion prior given by optical flow results should contain noisy candidate segments, we propose a memory-network-based policy model to select "strong" motion priors through reinforcement learning. The policy-selected segments (which have high precision) are used to fine-tune the model on target domain for improving the target performance.

Furthermore, the proposed 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.

References

[1] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," *arXiv preprint arXiv:1612.01925*, 2016. ix, 17, 26

[2] B. Settles, "Active learning literature survey," 2010. 1

[3] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *ICLR*, 2018. 1, 6

[4] K. Fragkiadaki, W. Zhang, G. Zhang, and J. Shi, "Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions," in *ECCV*, 2012. 4, 25

[5] T. Zhao and R. Nevatia, "Stochastic human segmentation from a static camera," in *Motion and Video Computing, Workshop*, 2002. 5

[6] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *CVPR*, 2003. 5

[7] T. V. Spina, M. Tepper, A. Esler, V. Morellas, N. Papanikolopoulos, A. X. Falcão, and G. Sapiro, "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. 5

[8] C. Song, Y. Huang, Z. Wang, and L. Wang, "1000fps human segmentation with deep convolutional neural networks," in *ACPR*, IEEE, 2015. 5

[9] Y. Tan, Y. Guo, and C. Gao, "Background subtraction based level sets for human segmentation in thermal infrared surveillance systems," *Infrared Physics & Technology*, vol. 61, pp. 230–240, 2013. 5

[10] F. He, Y. Guo, and C. Gao, "Human segmentation of infrared image for mobile robot search," *Multimedia Tools and Applications*, pp. 1–14, 2017. 5

[11] R. Dragon, B. Rosenhahn, and J. Ostermann, "Multi-scale clustering of frame-to-frame correspondences for motion segmentation," in *ECCV*, Springer, 2012. 6

[12] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1187–1200, 2014. 6

[13] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *CVPR*, IEEE, 2009. 6

[14] M. Y. Yang, H. Ackermann, W. Lin, S. Feng, and B. Rosenhahn, "Motion segmentation via global and local sparse subspace optimization," *arXiv preprint arXiv:1701.06944*, 2017. 6

[15] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *ECCV*, Springer, 2006. 6

[16] N. Shankar Nagaraja, F. R. Schmidt, and T. Brox, "Video segmentation with just a few strokes," in *ICCV*, 2015. 6

[17] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *CVPR*, 2016. 6

[18] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," in *ICCV*, 2017. 6

[19] R. I. Yarin Gal and Z. Ghahramani, "Deep bayesian active learning with image data," in *ICML*, 2017. 6

[20] S. R. Colwell and A. W. Joshi, "Multi-item scale development for measuring institutional pressures in the context of corporate environmental action," in *IABS*, 2009. 6

[21] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *ICML*, 2003. 6

[22] M. Ducoffe and F. Precioso, "Adversarial active learning for deep networks: a margin based approach," *arXiv preprint arXiv:1802.09841*, 2018. 6

[23] R. G. Xianglin Li and J. Cheng, "Incorporating incremental and active learning for scene classification," in *ICMLA*, 2012. 6

[24] A. Y. Ehsan Elhamifar, Guillermo Sapiro and S. S. Sastry, "A convex optimization framework for active learning," in *ICCV*, 2013. 6

[25] Y. Yang and M. Loog, "A variance maximization criterion for active learning," *arXiv preprint arXiv:1706.07642*, 2017. 6

[26] E. R. A. P. Christoph Kading, Alexander Freytag and J. Denzler, "Large-scale active learning with approximations of expected model output changes," in *GCPR*, 2016. 6

[27] A. Kuwadekar and J. Neville, "Relational active learning for joint collective classification models," in *ICML*, 2011. 6

[28] J. H. B. Sujoy Paul and A. Roy-Chowdhury, "Non-uniform subset selection for active learning in structured data," in *CVPR*, 2017. 6

[29] M. Fang, Y. Li, and T. Cohn, "Learning how to active learn: A deep reinforcement learning approach," in *EMNLP*, 2017. 6

[30] A. S. Philip Bachman and A. Trischler, "Learning algorithms for active learning," in *ICML*, 2017. 6

[31] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *ICCV*, 2015. 7

[32] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *ICML*, 2015. 7

[33] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (cmd) for domain-invariant representation learning," in *ICLR*, 2017. 7

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014. 7

[35] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *NIPS*, 2016. 7

[36] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015. 7, 11

[37] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," *arXiv preprint arXiv:1702.05464*, 2017. 7

[38] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *NIPS*, 2016. 7, 25, 31

[39] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "Fcns in the wild: Pixel-level adversarial and constraint-based adaptation," *arXiv preprint arXiv:1612.02649*, 2016. 7

[40] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. F. Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *ICCV*, 2017. 7, 10, 11, 12, 25, 31

[41] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *ICCV*, 2017. 7

[42] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa, "Unsupervised domain adaptation for semantic segmentation with gans," *arXiv preprint arXiv:1711.06969*, 2017. 7

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015. 8

[44] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, Springer, 2015. 8, 26

[45] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge, 1998. 12, 27

[46] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3-4, pp. 229–256, 1992. 13

[47] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *TPAMI*, vol. 33, no. 3, pp. 500–513, 2011. 17

[48] S. C. J. Weston and A. B. M. networks., "Bordes. memory networks.," in *ICLR*, 2015. 19

[49] J. Oh, V. Chockalingam, S. Singh, and H. Lee, "Control of memory, active perception, and action in minecraft," in *ICML*, 2016. 19

[50] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *IJCV*, vol. 111, no. 1, pp. 98–136, 2015. 26

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. 26

[52] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015. 26

[53] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014. 27