Generalizing Hand Segmentation in Egocentric Videos with Uncertainty-Guided Model Adaptation

Minjie Cai[1,*], Feng Lu[2,3,4,*], and Yoichi Sato[5]
[1]Hunan University, [2]State Key Lab. of VR Technology and Systems, Beihang University,
[3]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University,
[4]Peng Cheng Laboratory, [5]The University of Tokyo
caiminjie@hnu.edu.cn, lufeng@buaa.edu.cn, ysato@iis.u-tokyo.ac.jp

Abstract

Although the performance of hand segmentation in egocentric videos has been significantly improved by using CNNs, it still remains a challenging issue to generalize the trained models to new domains, e.g.*, unseen environments. In this work, we solve the hand segmentation generalization problem without requiring segmentation labels in the target domain. To this end, we propose a Bayesian CNN-based model adaptation framework for hand segmentation, which introduces and considers two key factors: 1) prediction uncertainty when the model is applied in a new domain and 2) common information about hand shapes shared across domains. Consequently, we propose an iterative self-training method for hand segmentation in the new domain, which is guided by the model uncertainty estimated by a Bayesian CNN. We further use an adversarial component in our framework to utilize shared information about hand shapes to constrain the model adaptation process. Experiments on multiple egocentric datasets show that the proposed method significantly improves the generalization performance of hand segmentation.*

1. Introduction

The popularity of wearable cameras in recent years is accompanied by a large amount of first-person view (egocentric) videos recording persons' daily interactions with their surrounding environments [27, 5, 46]. Since hands are among the most common objects in a user's field of view, hand segmentation is critically important for various objectives of egocentric video analysis [10, 12, 14]. Hand segmentation in egocentric videos is challenging due to rapidly changing imaging conditions and the lack of body cues [30]. Although recent researches have shown significant performance improvement by using various CNN-based models [43], how to generalize such models to new domains, *e.g.*, egocentric videos taken in unseen environments, remains a challenging issue.

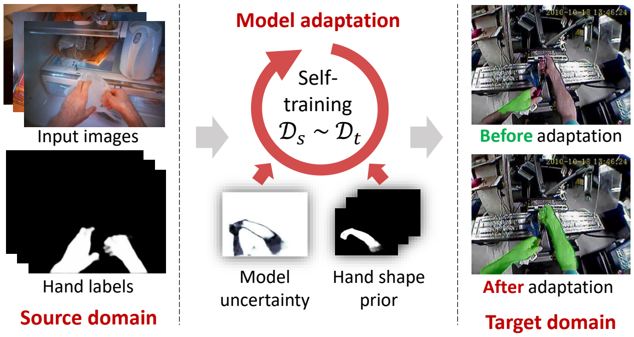

Figure 1. Illustration of the proposed model adaptation framework for hand segmentation in a new domain.

This work aims to generalize hand segmentation in egocentric videos in an unsupervised manner. The task can be viewed as unsupervised domain adaptation for hand segmentation and is challenging since the lack of annotated data in the new domain prohibits conventional approaches of fine-tuning models. Furthermore, the unique characteristics of egocentric videos (*e.g.*, rapidly changing illumination and background, lack of contextual information from body part) make it difficult to adapt model parameters to the new domain. As shown in Figure 1, the images in the target domain have different hand appearance and background compared with the images in the source domain. Consequently, a hand segmentation model trained in the source domain would have poor performance by directly applying to the target domain.

Based on such observation, we identify two major factors that are important for improving generalization performance of hand segmentation. The first factor is model uncertainty which measures how confident a model is with its prediction. The model uncertainty provides a good mea-

[*] Corresponding authors.

surement of the gap between data of the source and target domains. Commonly speaking, more similar an image (or image region) is from the training data, more confident a model becomes with its prediction and vice versa. Therefore, model uncertainty can be used to guide model adaptation in the target domain. The second factor is hand shape prior. Although egocentric videos may be captured with varying illumination and backgrounds leading to large variation on hand appearance, the shape of a hand tends to be consistent from the user's first-person point of view. Therefore, a common hand shape learned from the training data is expected to provide good prior information for promoting the model adaptation in the new domain.

In this paper, we propose a novel model adaptation framework for generalizing a hand segmentation model trained with source domain data to an unseen target domain without additional hand labels. Specifically, we formulate the CNN-based hand segmentation model in a Bayesian framework (Bayesian CNN) which is robust to overfitting and can provide more reliable estimation of model uncertainty than conventional deterministic CNN models. The core component of the framework is uncertainty-guided model adaptation which conducts self-training in the target domain iteratively by constructing reliable pseudo-labels based on the model uncertainty estimated with the Bayesian CNN. Furthermore, we compose prior information of hand shapes for model adaptation by enforcing the shape of a predicted hand region in the target domain to become similar to hand shapes in the source domain.

The main contributions of this work include:

- We propose a new Bayesian CNN-based model adaptation framework for generalizing hand segmentation in egocentric videos. To the best of our knowledge, this is the first effort to generalize hand segmentation with unsupervised model adaptation.
- We demonstrate the effectiveness of using uncertainty prior and hand shape prior to assist generalization of a hand segmentation model for egocentric videos.
- We demonstrate via thorough experiments that the proposed method improves the generalization performance of hand segmentation significantly compared with state-of-the-art CNN-based methods.

2. Related works

2.1. Hand segmentation in egocentric videos

Detecting or segmenting hands in egocentric videos with changing illumination and backgrounds is challenging for traditional color statistics-based methods such as [26], and many attempts have been made in recent year to overcome the challenge [39, 16, 30, 31, 47, 2, 4, 43, 32]. Ren and Gu [39] posed the task of hand segmentation as a figure-ground segmentation problem based on the assumption that motion patterns of hands are different from that of a background. Li and Kitani [30, 31] proposed a scene-adaptive method by training multiple hand detectors for different groups of images and choosing suitable hand detectors for different test images. Bambach et al. [2] proposed a two-stage hand segmentation method by first detecting hand bounding boxes with a convolutional neural network and then segmenting a hand region through Grabcut [40] in each detected bounding box. Recently, Urooj and Borji [43] used fully convolutional networks (RefineNet-ResNet101 [35] originally proposed for semantic segmentation) for hand segmentation and achieved state-of-the-art performance. However, existing methods have poor performance when applied to unseen datasets that are quite different from the dataset on which they are trained.

2.2. Unsupervised domain adaptation

Unsupervised domain adaptation [15] is a well studied topic which aims to reduce the domain gap for visual tasks and has attracted much research attention for semantic segmentation. Traditional approaches of unsupervised domain adaption try to learn feature representations that can minimize the discrepancy between source and target domains [19, 36]. Recently, the idea of adversarial learning was employed to learn general feature representations between source and target domains through an adversarial objective [24, 8, 25, 41, 44, 34]. In [23], a two-stage approach was proposed for domain adaptation which consists of an image-to-image translation network and a segmentation adaptation network. Li et al. extended the approach further with a bidirectional learning between the two stages [34].

Another line of work for unsupervised domain adaptation is based on the idea of self-training where predictions from a previously trained model are exploited as pseudo-labels for training a model of focus [49, 48]. In [49], a self-training based approach is proposed for adapting semantic segmentation models to new domains with class balancing and spatial prior. In this work, we adopt the idea of self-training and propose an uncertain-guided model adaptation framework based on a Bayesian CNN. Besides, we incorporate the hand shape prior for hand segmentation and formulate it in our model adaptation framework.

2.3. Bayesian deep learning

Bayesian inference has a long history in machine learning [6]. It provides uncertainty estimates with a posterior distribution. To overcome the difficulty of Bayesian inference in large models such as neural networks, early works explored a variety of methods such as Markov Chain Monte Carlo (MCMC) [37] and variational inference [22, 3]. Many other works have also been proposed to enable scalable variational inference in large Bayesian deep learning problems [18, 21, 29, 1]. Recently, approaches have been seen to ex-

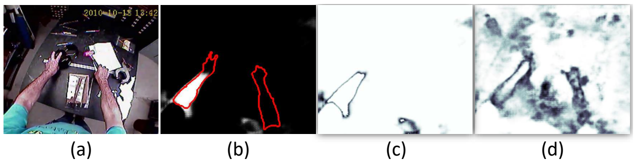

Figure 2. Comparison of different uncertainty maps: (a) input image (b) prediction probability (softmax output) from a standard CNN and ground-truth hand region (in red boundary) (c) uncertainty map obtained based on softmax output (d) uncertainty map obtained with a Bayesian CNN. The darker means less certain.

ploit uncertainty estimated from Bayesian deep learning for unsupervised domain adaptation [20, 45]. In [45], Bayesian uncertainty is matched to approximately reduce the domain-shift of a classifier.

In this work, we utilize Bayesian uncertainty to guide the adaptation of a pre-trained hand segmentation model to unseen environments.

3. Model uncertainty in hand segmentation

Before explaining the proposed method of uncertainty-guided model adaptation in Section 4, we briefly describe model uncertainty in hand segmentation.

3.1. Model uncertainty

Model uncertainty measures the confidence of a model with its prediction and is indispensable for many practical deep learning applications [42]. For example, if a model returns a classification result with high uncertainty, we might better be careful when using the result. In this work, we rely on model uncertainty to guide the adaptation of a pre-trained hand segmentation model to new domains. Briefly speaking, if a model is confident with its predictions from a part of the data in the target domain, such predictions then can be used as pseudo-labels for adapting model parameters to the target domain. The details of uncertain-guided model adaptation is described in Section 4. Here we first describe how to estimate model uncertainty for hand segmentation.

Standard CNN models do not capture model uncertainty. Alternatively, a prediction probability, *e.g.*, softmax output of the last layer of the model in the case of classification, is often erroneously used to interpret model uncertainty. Indeed, it is known that a model might be uncertain with its prediction even with a high prediction probability [17]. A Bayesian CNN provides a probabilistic interpretation of a CNN model by considering a distribution over model parameters and therefore provides a more reliable way of estimating model uncertainty. As shown in Figure 2, the uncertainty map obtained through a prediction probability is over-confident with the region of the right hand as we can see the region having very low values in the map. On the contrary, the uncertainty map obtained through a Bayesian CNN correctly identifies the region of the right hand being uncertain. In this work, we propose to use a Bayesian CNN for estimating model uncertainty for hand segmentation, and the details of uncertainty estimation are given in the following section.

3.2. Uncertainty estimates with Bayesian CNN

In a Bayesian CNN, model parameters are considered as random variables. Given training data $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ with inputs $\mathcal{X}$ and corresponding outputs $\mathcal{Y}$, the posterior distribution of the model parameters w is defined by invoking the Bayes' theorem:

$$p(w|\mathcal{D}) = \frac{p(\mathcal{Y}|\mathcal{X}, w)p(w)}{p(\mathcal{Y}|\mathcal{X})} \tag{1}$$

Computing the posterior distribution $p(w|\mathcal{D})$ is often intractable, and approximate inference is needed. As an active area of research in Bayesian deep learning, variational inference [7] approximates the complex posterior distribution $p(w|\mathcal{D})$ with an approximating variational distribution $q(w)$ by minimizing the Kullback-Leibler (KL) divergence between the two distributions. During the testing phase, the predictive distribution of output y given a new input x could be obtained through multiple stochastic forward passes with network parameters sampled from $q(w)$:

$$\begin{aligned} p(y|x) &= \int p(y|x, w)q(w)\, dw \\ &\approx \frac{1}{T}\sum_{i=1}^{T} p(y|x, w_i), \quad w_i \sim q(w) \end{aligned} \tag{2}$$

where T is the number of stochastic forward passes, w_i denotes one realization of model parameters sampled from $q(w)$. In practice, we follow the Bayesian approximation method in [18] which approximates the sampling of model parameters with dropout that has been widely used as a regularization tool in deep learning. Such approximation has the benefit that existing CNN models trained with dropout can be cast as Bayesian models without changing the original models.

Here we describe how to perform Bayesian inference and estimate model uncertainty for hand segmentation. Suppose we have trained a hand segmentation model $H(\mathbf{I}, w)$ which outputs a hand probability (softmax output) map $\mathbf{P}$ given an input image $\mathbf{I}$. The mean probability map $\bar{\mathbf{P}}$ and uncertainty map $\mathbf{U}$ are computed as:

$$\begin{aligned} \bar{\mathbf{P}} &= \frac{1}{T}\sum_{i=1}^{T} H(\mathbf{I}, w_i), \quad w_i \sim dropout(w) \\ \mathbf{U} &= \frac{1}{T}\sum_{i=1}^{T} \mathbf{P}_i^2 - \bar{\mathbf{P}}^2 \end{aligned} \tag{3}$$

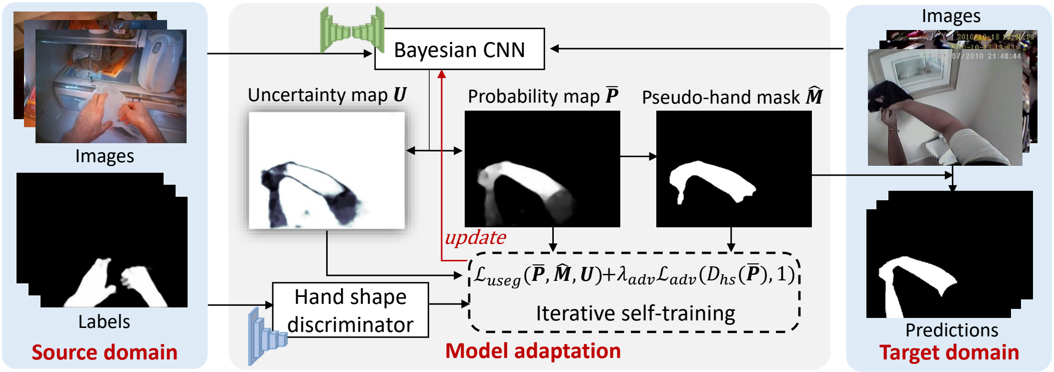
Images
Labels
Source domain
Bayesian CNN
Uncertainty map $\boldsymbol{U}$
Probability map $\bar{\boldsymbol{P}}$
Pseudo-hand mask $\hat{\boldsymbol{M}}$
update
$\mathcal{L}_{useg}(\bar{\boldsymbol{P}},\hat{\boldsymbol{M}},\boldsymbol{U})+\lambda_{adv}\mathcal{L}_{adv}(D_{hs}(\bar{\boldsymbol{P}}),1)$
Iterative self-training
Hand shape discriminator
Model adaptation
Images
Predictions
Target domain

Figure 3. Overview of the proposed uncertainty-guided model adaptation.

where $\mathbf{P}_i = H(\mathbf{I}, w_i)$ denotes a hand probability map obtained after one stochastic forward pass, and the square operators in Equation 3 are element-wise. It is noted that $\bar{\mathbf{P}}$ and $\mathbf{U}$ have the same spatial size with the input image, and the estimation of $\mathbf{U}$ essentially equals calculating the variance of a hand probability at each pixel. By thresholding $\bar{\mathbf{P}}$, we obtain a predicted hand segmentation mask $\hat{\mathbf{M}}$.

4. Proposed method

4.1. Task definition

Suppose we have a baseline hand segmentation model $H(\mathbf{I}, \theta_s)$ with parameters θ_s learned by using training data from the source domain $\mathcal{D}_s = \{\mathbf{I}_i, \mathbf{M}_i\}_{i=1}^{n_s}$, in which $\mathbf{I}_i$ denotes a RGB image and $\mathbf{M}_i$ denotes a binary hand segmentation mask. While the pre-trained baseline model can perform well as long as test data have a similar distribution as the training data $\mathcal{D}_s$, it may not generalize to data with a different distribution. Our task is to adapt the pre-trained model to a new target domain $\mathcal{D}_t = \{\mathbf{I}_i\}_{i=1}^{n_t}$ without newly annotated hand segmentation masks.

4.2. Uncertainty-guided model adaptation

We adopt the idea of self-training from semi-supervised learning [13] for model adaptation. Although no hand segmentation label is given for the target domain, by exploiting pseudo-labels from confident model predictions, the model could be updated and adapted to the target domain. As discussed in Section 3.1, a prediction probability from a deterministic CNN model cannot provide reliable uncertainty estimates. Different from previous approaches which construct pseudo-labels based on such prediction probabilities, we utilize uncertainty estimated based on Bayesian deep learning to construct more reliable pseudo-labels.

The model adaptation is formulated as an iterative self-training procedure in which the hand probability maps and uncertainty maps obtained from the model at the previous iteration are used for training current model. The loss function to learn $H(\mathbf{I}, \theta_t)$ for the target domain is defined as:

$$\mathcal{L}_{H_t^{(k)}} = \mathcal{L}_{useg}\left(\bar{\mathcal{P}}_t^{(k)},\ \hat{\mathcal{M}}_t^{(k-1)},\ \mathcal{U}_t^{(k-1)}\right) \tag{4}$$

where k denotes the iteration index, $\bar{\mathcal{P}}_t = \{\bar{\mathbf{P}}_i\}_{i=1}^{n_t}$ and $\mathcal{U}_t = \{\mathbf{U}_i\}_{i=1}^{n_t}$ denote the mean hand probability maps and uncertainty maps of the target domain obtained through Equation 3, $\hat{\mathcal{M}}_t = \{\hat{\mathbf{M}}_i\}_{i=1}^{n_t}$ denotes the predicted hand segmentation masks that are obtained by thresholding $\bar{\mathcal{P}}_t$ with 0.5. $\mathcal{L}_{useg}$ denotes the uncertainty-guided hand segmentation loss and is defined as:

$$\begin{aligned}\mathcal{L}_{useg}(\bar{\mathbf{P}}, \hat{\mathbf{M}}, \mathbf{U}) = -\frac{1}{M}\sum_{m=1}^{M}(1-\mathrm{U}_m)\big(\hat{\mathrm{M}}_m \log \bar{\mathrm{P}}_m \\ + (1-\hat{\mathrm{M}}_m)\log(1-\bar{\mathrm{P}}_m)\big)\end{aligned} \tag{5}$$

where the iteration index and sample index are omitted for simplicity, and m denotes the pixel index of $\bar{\mathbf{P}}, \hat{\mathbf{M}}, \mathbf{U}$. It is noted that, instead of selecting pixels of low uncertainty as pseudo-labels with a manually specified threshold, we use uncertainty as a soft weight on the whole predictions. In other words, pixels with high confidence contribute more to model adaptation and vice verse. $\mathbf{U}$ is normalized to a range of [0,1] before being used.

The model uncertainty is also used to determine when the iterative adaptation procedure is terminated to avoid overfitting. Specifically, we terminate the iteration when the reduction of the average uncertainty score is smaller than 10%. The overall iterative adaptation procedure is summrized in Algorithm 1.

4.3. Hand shape constraint

To improve generalization performance of hand segmentation, it is also important to explore common information of human hands shared between the source and target domains. In this work, we propose to exploit hand shapes as such common information to help promote the adaptation of hand segmentation models to the target domain. Although imaging conditions and backgrounds can be very different

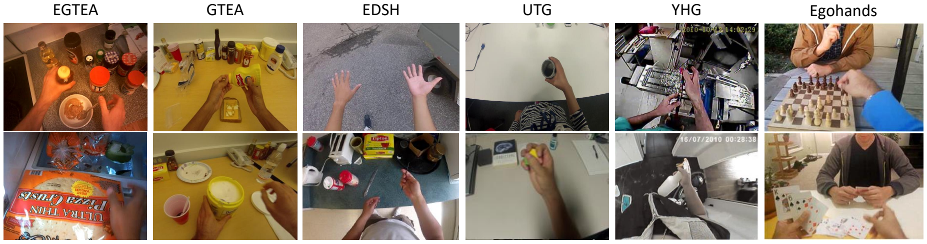

Figure 4. Image samples of six datasets. Large variation on illumination and background could be observed across different datasets.

Algorithm 1: Procedure of model adaptation

```
Input: $\mathcal{D}_t$ and $H_s$ trained on $\mathcal{D}_s$
Output: $H_t$
Initialize: $\hat{\mathcal{M}}^{(0)}$, $\mathcal{U}_t^{(0)} \leftarrow H_s(\mathcal{D}_t)$ with Equation 3
for $k \leftarrow 1$ to $K$ do
    Train $H_t^{(k)}$ with Equation 4 or 8
    $\hat{\mathcal{M}}^{(k)}$, $\mathcal{U}_t^{(k)} \leftarrow H_t^{(k)}(\mathcal{D}_t)$ with Equation 3
    if $|\bar{\mathcal{U}}_t^{(k)} - \bar{\mathcal{U}}_t^{(k-1)}| < \frac{1}{10}\bar{\mathcal{U}}_t^{(k-1)}$ then
        Stop iteration
```

across different egocentric datasets, there is consistency in the shape of hands from user's first-person point of view. Therefore, the information of hand shape learned from the source domain could be used as useful prior information for model adaptation in the target domain.

To be more concrete, the hand shape prior is learned by adding a hand shape discriminator D_{hs} in the training of hand segmentation in the source domain, and the loss function is formulated as:

$$\mathcal{L}_{H_s} = \mathcal{L}_{seg}\big(\mathcal{P}_s, \mathcal{M}_s\big) + \mathcal{L}_{adv}\big(D_{hs}(\mathcal{P}_s), 1\big) \tag{6}$$

$$\mathcal{L}_{D_{hs}} = \mathcal{L}_{adv}\big(D_{hs}(\mathcal{M}_s), 1\big) + \mathcal{L}_{adv}\big(D_{hs}(\mathcal{P}_s), 0\big) \tag{7}$$

where $\mathcal{L}_{seg}$ denotes standard hand segmentation loss and $\mathcal{L}_{adv}$ denotes image-level binary cross-entropy loss. After the above adversarial learning, information of hand shapes is encoded in D_{hs} and can be used for model adaptation.

During adaptation, the loss function to learn $H(\mathbf{I}, \theta_t)$ with the obtained prior of hand shapes is modified as:

$$\begin{aligned} \mathcal{L}_{H_t^{(k)}} = \mathcal{L}_{useg}\big(\bar{\mathcal{P}}_t^{(k)}, \hat{\mathcal{M}}_t^{(k-1)}, \mathcal{U}_t^{(k-1)}\big) \\ + \lambda_{adv}\mathcal{L}_{adv}\big(D_{hs}(\bar{\mathcal{P}}_t^{(k)}), 1\big) \end{aligned} \tag{8}$$

where the second term with weighting factor λ_{adv} is used to enforce the shape of a predicted hand segmentation to be similar to that learned from the source domain.

4.4. Network architecture and training details

Network architecture. We adopt RefineNet [35] as our baseline hand segmentation network considering the state-of-the-art performance achieved by it in recent work [43]. It is noted that the segmentation network itself is not our contribution and our proposed model adaptation method could be applied to any segmentation networks with dropout. To formulate a Bayesian CNN, we simply train the hand segmentation network in the source domain with one dropout layer (dropout probability $p = 0.5$) added after each residual convolutional unit of RefineNet, and the dropout layers are also applied during testing. The hand shape discriminator D_{hs} has the same architecture as the one used in [38].

Training details. We employ PyTorch for implementations[1]. All experiments are run on a single NVIDIA 2080TI GPU. We use Adam optimizer [28] with learning rate 10^{-5} to train the hand segmentation network and hand shape discriminator in the source domain for 20 epochs. For iterative uncertainty-guided model adaptation, we use RMSProp with learning rate 10^{-5}, and within each iteration the network is trained for one epoch with pseudo-labels. To estimate model uncertainty with Bayesian CNN, we conduct $T = 10$ times of stochastic forward passes. The weighting factor for adversarial loss is set as $\lambda_{adv} = 0.1$.

5. Experiments

5.1. Datasets

EGTEA dataset [33]. The Extended GeorgiaTech Egocentric Activity (EGTEA) dataset contains 29 hours of egocentric videos with a resolution of 1280×960. These videos record meal preparation tasks performed by 32 subjects in a naturalistic kitchen environment. Within the dataset, 13847 images are labeled with hand masks. We use this dataset to train the initial hand segmentation network.

GTEA dataset [16]. This dataset consists of 28 egocentric videos with a resolution of 720 × 405 recording 7 daily activities performed by 4 subjects. 663 images are annotated with hand masks. We follow the data split as in [43] that images from subject 1, 3, 4 are used as a training set and the rest as a test set.

EDSH dataset [30]. This dataset contains 3 egocentric videos (*EDSH1*, *EDSH2* and *EDSH-Kitchen*) with a resolution of 1280 × 720 recorded in both indoor and outdoor

[1] Code available at https://github.com/cai-mj/UMA.

Table 1. Cross-dataset hand segmentation performance of different model components. EGTEA dataset is used as source domain. Mean Intersection over Union (mIoU) and mean F1 score (mF1) are used as evaluation metric.

Method	GTEA		EDSH-2		EDSH-K		UTG		YHG		Egohands	
	mIoU	mF1	mIoU	mF1	mIoU	mF1	mIoU	mF1	mIoU	mF1	mIoU	mF1
CNN	0.8845	0.9257	0.6936	0.8030	0.7205	0.8078	0.5481	0.6859	0.2831	0.3870	0.4019	0.5357
CNN+uma	0.8766	0.9127	0.7141	0.8170	0.7723	0.8472	0.6089	0.7284	0.3159	0.4257	0.4252	0.5632
Bayesian CNN	0.8896	0.9362	0.7632	0.8553	0.7576	0.8356	0.5832	0.7174	0.3619	0.4987	0.4235	0.5619
Bayesian CNN+uma	0.8945	0.9391	0.7965	0.8819	0.7812	0.8599	0.6762	0.7892	0.5223	0.6608	**0.4665**	**0.6134**
Bayesian CNN+uma+hs	**0.8990**	**0.9417**	**0.8025**	**0.8856**	**0.7951**	**0.8674**	**0.6827**	**0.7922**	**0.5596**	**0.7048**	0.4660	0.6123

environments. We adopt the same data split as in [30]. 442 labeled images from *EDSH1* are used as a training set. 104 labeled images from *EDSH2* and 197 labeled images from *EDSH-Kitchen* are used as two separate test sets.

UTG dataset [11]. The University of Tokyo Grasping (UTG) dataset consists of 50 egocentric videos with a resolution of 1920 × 1080. This dataset captures 17 different types of hand grasps performed by 5 subjects. To facilitate our study, we mannually annnotated hand masks on 872 images and randomly split them into training and test set with the ratio of 75% and 25% respectively.

YHG dataset [9]. The Yale Human Grasping (YHG) dataset provides daily observation of human grasping behavior in unstructured environments. It consists of 27.7 hours of egocentric videos with a resolution of 640 × 480 recorded by two machinists and two house keepers during their daily work. We manually annotated hand masks on 488 images and randomly split them into training and test set with the ratio of 75% and 25% respectively.

Egohands dataset [2]. This dataset consists of 48 egocentric videos with a resolution of 1280×720 which records social interactions between two persons in both indoor and outdoor environments. 4800 randomly sampled images are labeled with hand masks. Following [2] and [43], we split the data into training, validation and test set with the ratio of 75%, 8% and 17%.

Image samples of these datasets are shown in Figure 4. It is noted that we only use hand mask labels in the training set of EGTEA dataset for training our hand segmentation network, and the labels in other datasets are only used for performance evaluation.

5.2. Performance analysis

5.2.1 Ablation study of the proposed method

We first conduct an ablation study on the effectiveness of different components of the proposed method as follows:

- CNN: a standard CNN-based hand segmentation model using architecture of RefineNet [35].
- CNN+uma: uncertainty-guided model adaptation in which the model uncertainty is estimated based on a standard CNN.
- Bayesian CNN: a Bayesian version of a CNN-based hand segmentation model.
- Bayesian CNN+uma: uncertainty-guided model adaptation in which the model uncertainty is estimated based on Bayesian CNN.
- Bayesian CNN+uma+hs: Bayesian CNN+uma with the hand shape constraint for model adaptation.

The cross-dataset hand segmentation performance of different models is shown in Table 1. We first analyze the results based on IoU. It can be seen that the Bayesian CNN has better generalization ability than the standard CNN. With Bayesian CNN, the uncertain-guided model adaptation (Bayesian CNN+uma) improves the segmentation performance for all the datasets. In particular, the improvement is significant for datasets of UTG and YHG which have very different imaging conditions from the source domain dataset. Besides, the effectiveness of uncertain-guided model adaptation with Bayesian CNN is much better than that with standard CNN, indicating that Bayesian CNN provides a better way of estimating model uncertainty than standard CNN. Adding hand shape constraint (Bayesian CNN+uma+hs) further improves the segmentation performance, verifying our hypothesis that hand shape is consistent in egocentric videos and could be used to promote segmentation adaptation. It is noted that the generalization performance in Egohands is limited even with model adaptation. The reason is that the hands in Egohands are recorded in a mixture of first (egocentric) and second-person views and the segmentation model (as well as the hand shape prior) learned in first-person view could not adapt well to the second-person view. This indicates that to adapt hand segmentation across different views, new labels might be needed. Similar results are seen with the mean F1 score.

5.2.2 Evaluation of iterative adaptation

Here we evaluate how the segmentation performance varies during the iteration procedure of our model adaptation method. In Figure 5, we demonstrate the performance variation of two versions of our method: Bayesian CNN+uma and Bayesian CNN+uma+hs. Since the iteration terminates (illustrated by the vertical dashed line) before five iterations based on our stop criterion for all datasets, we only demonstrate results of five iterations. It can be seen from the figure that with iterative adaptation the segmentation performance tends to improve and then degrades after a certain number

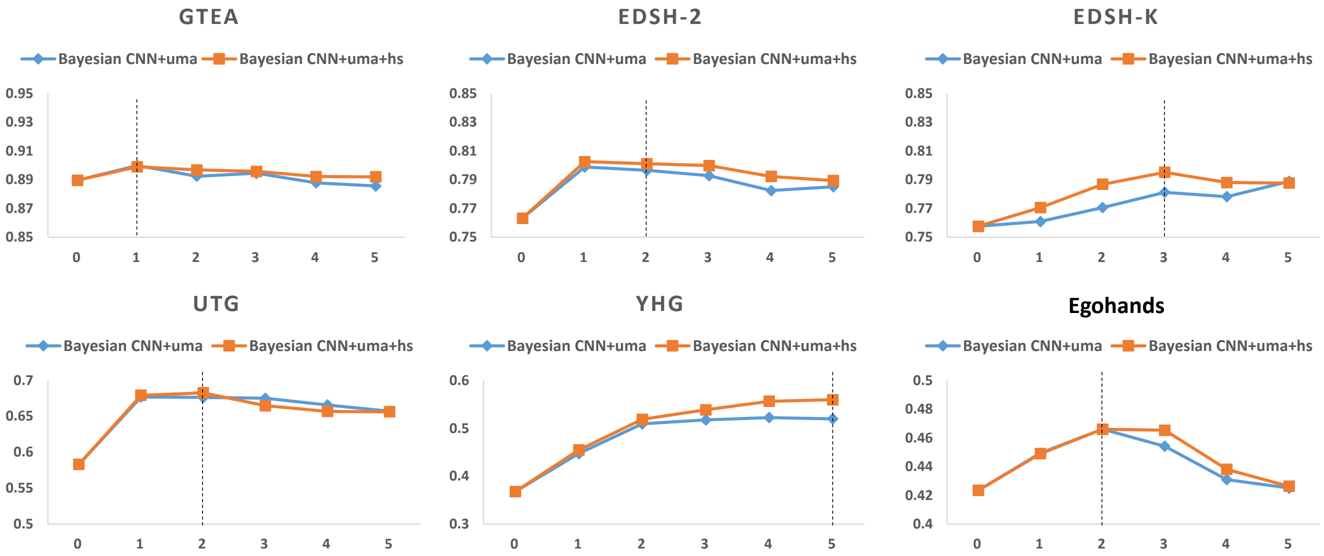

Figure 5. Performance variation of the iterative model adaptation. The horizontal axis shows the number of iterations, with "0" denoting the initial prediction before model adaptation. The vertical axis shows the segmentation performance (IoU).

of iterations. The reason is probably that as model adaptation iterates the model becomes more confident (possibly over-confident) with the data of the target domain and might overfit to its false predictions. This indicates that a proper stop criterion is needed to prevent overfitting. The results show that based on our stop criterion, the adaptation procedure could terminate before performance degradation.

Qualitative results of our method on YHG dataset are shown in Figure 6. It can be seen that at the beginning, the segmentation performance with initial model is rather poor and correspondingly the area of uncertain region is relatively large. With uncertainty-guided model adaptation, segmentation performance improves and the area of uncertain region decreases progressively. More qualitative results on other datasets are given in the supplementary material.

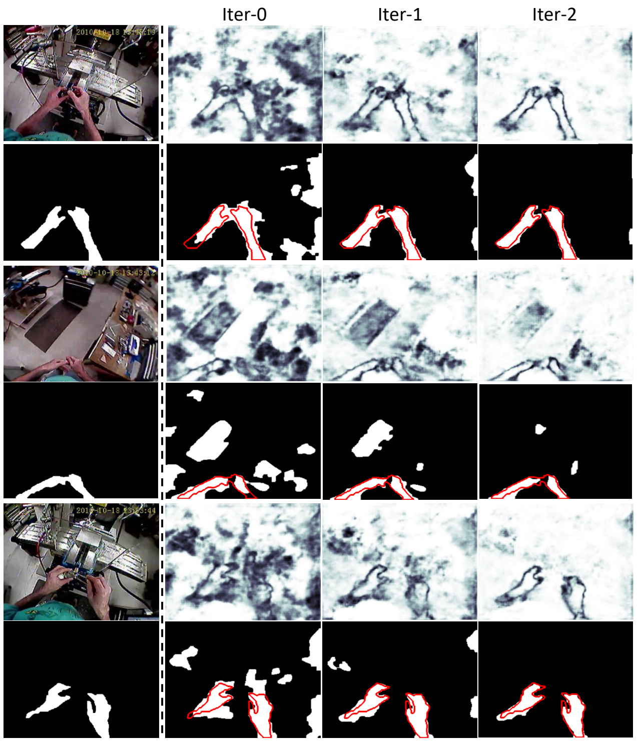

Figure 6. Qualitative results with iterations of uncertainty-guided model adaptation. The left column shows original images and hand masks of three samples from YHG dataset. The other part of the figure shows hand segmentation results and estimated model uncertainty at different iterations.

5.2.3 Evaluation of stochastic forward passes

In previous sections, we have shown that the proposed uncertain-guided model adaptation improved the generalization performance of hand segmentation significantly. In particular, by sampling model parameters through multiple stochastic forward passes, the Bayesian CNN works better for both inference and uncertainty estimation of hand segmentation compared with standard CNN. In this part, we study how the number of stochastic forward passes affects the final performance. Figure 7 shows the segmentation performance of Bayesian CNN+uma with different numbers of stochastic forward passes. The performance improves at the beginning (before 15), and then fluctuates around IoU of 0.525 when the number of stochastic forward passes increases. The reason of performance fluctuation shown in the figure might be that current dropout-based sampling could not well approximate the posterior distribution of model parameters without enough number of sampling. This indi-

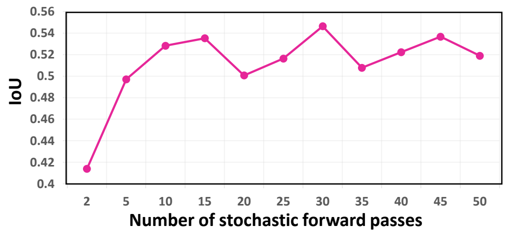

Figure 7. Evaluation of stochastic forward passes with the proposed method on YHG dataset.

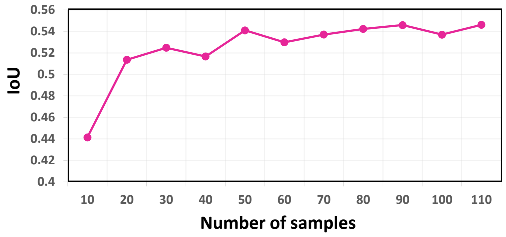

Figure 8. Simulation of online hand segmentation performance with the proposed method on YHG dataset.

cates that more thorough study on the impact of different sampling strategies is needed in the future work.

5.2.4 Simulation of online hand segmentation

Suppose we need a hand segmentation system which could be practically usable for different real-world environments. The proposed method could serve as an online model adaptation tool for such a system since it could adapt a pre-trained hand segmentation model to unseen environments without labels. To simulate our method's ability for online hand segmentation, we gradually sample more raw images (without labels) from the training set of YHG dataset for model adaptation and evaluate the segmentation performance on the testing set of the same dataset. Figure 8 shows the segmentation performance as a function of number of samples. The results indicate that with only a few unlabeled data (20 raw images), the model could be well adapted to the target domain, and the performance could keep improving as we collect more data.

5.3. Comparison with state-of-the-art models

We compare the cross-dataset performance with state-of-the-art methods on hand segmentation and unsupervised domain adaptation for semantic segmentation.

- RefineNet [43]: a state-of-the-art hand segmentation model using RefineNet [35] as the network architecture. It is also used as a baseline model in the ablation study (Section 5.2.1).
- CBST [49]: a self-training method for semantic segmentation. It generates pseudo-labels for model adaptation based on softmax output and further improves the performance with spatial prior information.
- BDL [34]: a state-of-the-art unsupervised domain adaptation method for semantic segmentation. It combines self-training in [49] with adversarial learning to decrease the domain gap.

CBST [49] and BDL [34] were originally proposed for semantic segmentation and are compared here to show how state-of-the-art domain adaptation methods could help improve the generalization performance of hand segmentation. We adapt their methods to solve the hand segmentation task. To give a better comparison, we replace their original segmentation networks with RefineNet.

Table 2. Cross-dataset hand segmentation performance of different methods. EGTEA dataset is used as source domain. Intersection over Union (IoU) is used as evaluation metric.

Method	GTEA	EDSH-2	EDSH-K	UTG	YHG	Egohands
RefineNet [43]	0.8845	0.6936	0.7205	0.5481	0.2831	0.4019
CBST [49]	0.8766	0.7353	0.7207	0.5627	0.3539	0.4293
BDL [34]	0.8609	0.7240	0.7360	0.6210	0.4170	0.4390
Ours	**0.8990**	**0.8025**	**0.7951**	**0.6827**	**0.5596**	**0.4660**

Quantitative results of different methods are shown in Table 2. Our method achieves the best performance on all the target datasets and significantly outperforms the state-of-the-art hand segmentation method [43] without domain adaptation. The superior performance of our method over CBST [49] and BDL [34] verifies the effectiveness of the proposed method for generalizing hand segmentation.

6. Conclusion

We proposed a novel method to generalize hand segmentation across different environments. With model uncertainty estimated from a Bayesian CNN, the proposed method could adapt a pre-trained hand segmentation model to a new environment without labels. Thorough experiments show significant improvements on the generalization performance of hand segmentation compared with existing CNN-based methods and enables flexible online adaptation of hand segmentation to new environments. As for our future work, we would like to study the effectiveness of different quantitative measurement of model uncertainty based on Bayesian CNN. In addition, as current experiments show fluctuating performance with different number of stochastic forward passes, we would like to study more deeply on the impact of different sampling strategies.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (NSFC) under Grant 61906064 and Grant 61972012, and by CREST, JST.

References

[1] A. K. Balan, V. Rathod, K. P. Murphy, and M. Welling. Bayesian dark knowledge. In *Advances in Neural Information Processing Systems*, pages 3438–3446, 2015.

[2] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *IEEE International Conference on Computer Vision*, pages 1949–1957, 2015.

[3] D. Barber and C. M. Bishop. Ensemble learning in bayesian neural networks. *Nato ASI Series F Computer and Systems Sciences*, 168:215–238, 1998.

[4] A. Betancourt, P. Morerio, E. Barakova, L. Marcenaro, M. Rauterberg, and C. Regazzoni. Left/right hand segmentation in egocentric videos. *Computer Vision and Image Understanding*, 154:73–81, 2017.

[5] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg. The evolution of first person vision methods: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):744–760, 2015.

[6] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[7] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[8] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3722–3731, 2017.

[9] I. M. Bullock, T. Feix, and A. M. Dollar. The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *The International Journal of Robotics Research*, 34(3):251–255, 2015.

[10] M. Cai, K. Kitani, and Y. Sato. A scalable approach for understanding the visual structures of hand grasps. In *IEEE International Conference on Robotics and Automation*, pages 1360–1366, 2015.

[11] M. Cai, K. Kitani, and Y. Sato. An ego-vision system for hand grasp analysis. *IEEE Transactions on Human-Machine Systems*, 47(4):524–535, 2017.

[12] M. Cai, F. Lu, and Y. Gao. Desktop action recognition from first-person point-of-view. *IEEE Transactions on Cybernetics*, 49(5):1616–1628, 2018.

[13] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[14] N. Charoenkulvanich, R. Kamikubo, R. Yonetani, and Y. Sato. Assisting group activity analysis through hand detection and identification in multiple egocentric videos. In *International Conference on Intelligent User Interfaces*, pages 570–574, 2019.

[15] G. Csurka. A comprehensive survey on domain adaptation for visual applications. In *Domain Adaptation in Computer Vision Applications*, pages 1–35. Springer, 2017.

[16] A. Fathi, A. Farhadi, and J. Rehg. Understanding egocentric activities. In *IEEE International Conference on Computer Vision*, pages 407–414. IEEE, 2011.

[17] Y. Gal. *Uncertainty in deep learning*. PhD thesis, University of Cambridge, 2016.

[18] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[19] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[20] L. Han, Y. Zou, R. Gao, L. Wang, and D. Metaxas. Unsupervised domain adaptation via calibrating uncertainties. In *CVPR Workshops*, 2019.

[21] J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

[22] G. Hinton and D. Van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *ACM Conference on Computational Learning Theory*, 1993.

[23] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1994–2003, 2018.

[24] 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.

[25] W. Hong, Z. Wang, M. Yang, and J. Yuan. Conditional generative adversarial network for structured domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2018.

[26] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.

[27] T. Kanade and M. Hebert. First-person vision. *Proceedings of the IEEE*, 100(8):2442–2453, 2012.

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

[29] C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *AAAI*, pages 1788–1794, 2016.

[30] C. Li and K. Kitani. Pixel-level hand detection in ego-centric videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2013.

[31] C. Li and K. M. Kitani. Model recommendation with virtual probes for egocentric hand detection. In *IEEE International Conference on Computer Vision*, pages 2624–2631, 2013.

[32] M. Li, L. Sun, and Q. Huo. Flow-guided feature propagation with occlusion aware detail enhancement for hand segmentation in egocentric videos. *Computer Vision and Image Understanding*, 187:102785, 2019.

[33] Y. Li, M. Liu, and J. M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference on Computer Vision*, pages 619–635, 2018.

[34] Y. Li, L. Yuan, and N. Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[35] G. Lin, A. Milan, C. Shen, and I. D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1925–1934, 2017.

[36] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[37] R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[38] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[39] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3144, 2010.

[40] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics*, volume 23, pages 309–314. ACM, 2004.

[41] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[42] J. Snoek, Y. Ovadia, E. Fertig, B. Lakshminarayanan, S. Nowozin, D. Sculley, J. Dillon, J. Ren, and Z. Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980, 2019.

[43] A. Urooj and A. Borji. Analysis of hand segmentation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4710–4719, 2018.

[44] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[45] J. Wen, N. Zheng, J. Yuan, Z. Gong, and C. Chen. Bayesian uncertainty matching for unsupervised domain adaptation. In *International Joint Conference on Artificial Intelligence*, pages 3849–3855, 2019.

[46] H. Yu, M. Cai, Y. Liu, and F. Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *ACM International Conference on Multimedia*, pages 1358–1366, 2019.

[47] X. Zhu, X. Jia, and K.-Y. K. Wong. Pixel-level hand detection with shape-aware structured forests. In *Asian Conference on Computer Vision*, pages 64–78, 2014.

[48] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang. Confidence regularized self-training. In *IEEE International Conference on Computer Vision*, pages 5982–5991, 2019.

[49] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *European Conference on Computer Vision*, pages 289–305, 2018.