社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Style Adapted DataBase: セマンティクスを考慮したスタイライゼーションによる 手セグメンテーションの汎化

大川　武彦† 　八木　拓真† 　佐藤　洋一†

† 東京大学生産技術研究所　〒153-8505 東京都目黒区駒場 4–6–1
E-mail: †{ohkawa-t,tyagi,ysato}@iis.u-tokyo.ac.jp

あらまし　ウェアラブルカメラから得られる一人称視点映像におけるドメインシフトは，照明条件や背景のアピアランスの変化によって生じ，手セグメンテーションの精度を低下させる．本論文では，このようなドメインシフトの問題に対応するために，少数のターゲットラベルを用い，セマンティクスを考慮したスタイル変換によるドメイン適応手法を提案する．具体的には，ソース画像とターゲット画像をそれぞれコンテンツとスタイルとしてスタイル変換ネットワークに与え，これらのラベルによって前景と背景を分離した後，ネットワークは各領域ごとにソースデータにターゲットのスタイルを転移する．提案手法は，スタイル変換を施したソースデータセットに複数のスタイルを導入できることから，このデータセットで学習したモデルは複数のターゲットドメインに一度で同時に汎化する．提案手法は，最新の手セグメンテーションのためのドメイン適応手法と同等かそれ以上のクロスデータセット汎化性能を達成した．
キーワード　ドメイン適応，スタイル変換，手セグメンテーション，一人称ビジョン

Style Adapted DataBase: Generalizing Hand Segmentation via Semantics-aware Stylization

Takehiko OHKAWA†, Takuma YAGI†, and Yoichi SATO†

† Institute of Industrial Science, The University of Tokyo, 4–6–1 Komaba, Meguro-ku, Tokyo 153–8505, Japan
E-mail: †{ohkawa-t,tyagi,ysato}@iis.u-tokyo.ac.jp

Abstract　Domain shift in first-person vision degrades the performance of hand segmentation, which is caused by changes in lighting conditions and background appearances. In this paper, we propose a semantics-aware stylization approach for domain adaptation using an image style transfer with only a few target labels. Specifically, given a source image as content and a target image as style, foreground and background are separated by their labels, and the network transfers the styles of the target image to the source image separately for the foreground and background. Multiple styles can be fed into a stylized source dataset, thus the model trained on the dataset simultaneously generalizes to multiple target domains at once. Our method achieves the best cross-dataset generalization against the state-of-the-art domain adaptation methods for hand segmentation.
Key words　Domain Adaptation, Style Transfer, Hand Segmentation, First Person Vision

1. Introduction

The growth of wearable devices brings a large amount of egocentric videos, which records persons' daily interactions with their surrounding environments. To understand camera wearer's activities, hands are crucial entity in egocentric videos. Detection and segmentation of hand regions have a vital role in several computer vision tasks (*e.g.*, hand pose estimation, 3D hand shape reconstruction, and hand-object interaction recognition) and their applications such as robotics, human-machine interaction, and augmented reality.

Domain shift, a distribution mismatch between training and testing data, degrades the performance of a model trained in a domain. In first-person vision, this shift is caused by changes in appearances. Since egocentric videos are collected in a myriad of environments, illumination, background, and context are significantly diverse. Additionally, camera properties cause appearance-level differences such as brightness, white balance, and resolution. Domain adaptation (DA) is one of the solutions for aligning the distribution shift between source and target data. Cai et al. [1] proposed a Bayesian CNN-based model adaptation framework for hand segmentation,

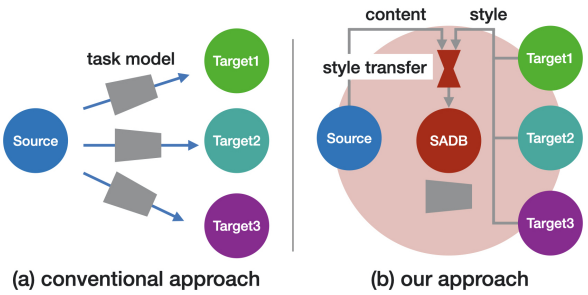

Figure 1: Illustration of our approach.

which constructs reliable pseudo-labels for target data.

However, since we cannot assume that test data are drawn from an identical distribution, single-target domain adaptation (STDA) that assumes a single target domain is limited in the real world. Recent work [2], [3] focuses on multi-target domain adaptation (MTDA), where target data underlie multiple distributions, and the training aims simultaneous adaptation to multi-target domains. Here, MTDA can be employed in the following scenarios of first-person vision. A camera wearer can move across environments (*e.g.*, *kitchen→dining room*, *laboratory→outside*). Moreover, when the time has passed by, climate and lighting conditions in rooms drastically change in daily life. Therefore, multi-target adaptation is worth considering for real-world applications of first-person vision.

In hand segmentation, the appearance of the foreground (hands) is based on the skin color of camera wearers or reflected light on their hands, which has less dependency on the background appearances. To address the appearance gap separately for the semantically corresponding areas, we propose a simple yet effective weakly-supervised and semantics-aware domain adaptation approach using a style transfer. Given a source image as content and a target image as style, the network transfers the target style to the source domain while preserving its content. Concretely, we split the foreground and background of both inputs by their labels, and then stylize the source image for each region. Owing to it, we obtain a style aligned source dataset, named as **Style Adapted DataBase (SADB)**, which is utilized for the training of a segmentation model. Since the dataset-level adaptation and segmentation training are disjoint, our method can be applied to other downstream tasks and networks.

In the MTDA setting, we can extend the SADB to the one with multiple target appearances since the style transfer enables to represent the multi-style, as illustrated in Figure 1. The model trained on the SADB with the multi-target styles once generalizes simultaneously to the multi-target domains.

In our experiments, we quantitatively demonstrate the further improved cross-dataset generalization against Bayesian CNN and the state-of-the-art DA methods without specifying a source model to a test domain. The proposed method improves the adaptation performance of hand segmentation by 13.56% on average. We also show that the appearance-level adaptation from different domains is supportive by exploring the performance on the SADB excluding the style of a test domain. Qualitatively, we find the proposed mask aligned stylization transfers lighting on hands to other domains, which bridges the hand-level appearance gap. The model is more robust to motion blur, which frequently occurs in first-person vision.

2. Related Work

2.1 Hand Segmentation in Egocentric Videos

Li and Kitani [4] categorized classical hand detection approaches into three groups: (1) local appearance-based detection, (2) global appearance-based detection, and (3) motion-based detection. Recently, Urooj and Borji [5] adopted an end-to-end CNN approach (RefineNet [6]), and achieved state-of-the-art results. Nevertheless, generalization to unseen datasets is sensitive to the choice of the source dataset [5] due to the gap of illumination, lighting conditions, and camera lens properties between training and testing environments.

2.2 Domain Adaptation

Single-target domain adaptation. Domain shift or domain gap is the problem that a typical model trained on a specific distribution of data from a particular domain will not generalize well to other datasets not seen during training. A way of addressing domain shift is domain adaptation that usually indicates single-target domain adaptation (STDA). There are mainly three categories: (1) minimizing the distance between the source and target feature distributions, (2) generative (pixel-level) approach [7], [8], and (3) self-training with pseudo-labels [1], [9].

Recently, Cai et al. [1] applied self-training-based unsupervised domain adaptation to hand segmentation in egocentric videos and proposed an uncertain-guided model adaptation (UMA) framework using a Bayesian CNN. To estimate model uncertainty, the UMA must conduct stochastic forward calculation many times in each iteration. Although such domain adaptation methods can produce impressive results, its functionality and scalability are limited in multi-target settings because it requires the same number of training for adaptation as targets.

In this work, we adopt the generative approach to synthesize a stylized source dataset, but the adaptation and segmentation training are independent. Therefore, our dataset-level adaptation can be compatible with other downstream tasks and task models. Furthermore, compared to the Bayesian CNN and STDA methods, our approach is more practically efficient since we only need one-time training on the SADB without the stochastic forward.

Multi-target domain adaptation. Even though STDA assumes they are drawn from an identical distribution, all target instances stem from multiple distributions in real-world scenarios, whereas we can access the domain category. Multi-target domain adaptation (MTDA) aims to adapt simultaneously to multiple and unlabeled target domains. Gholami et al. [2] proposed a MTDA approach

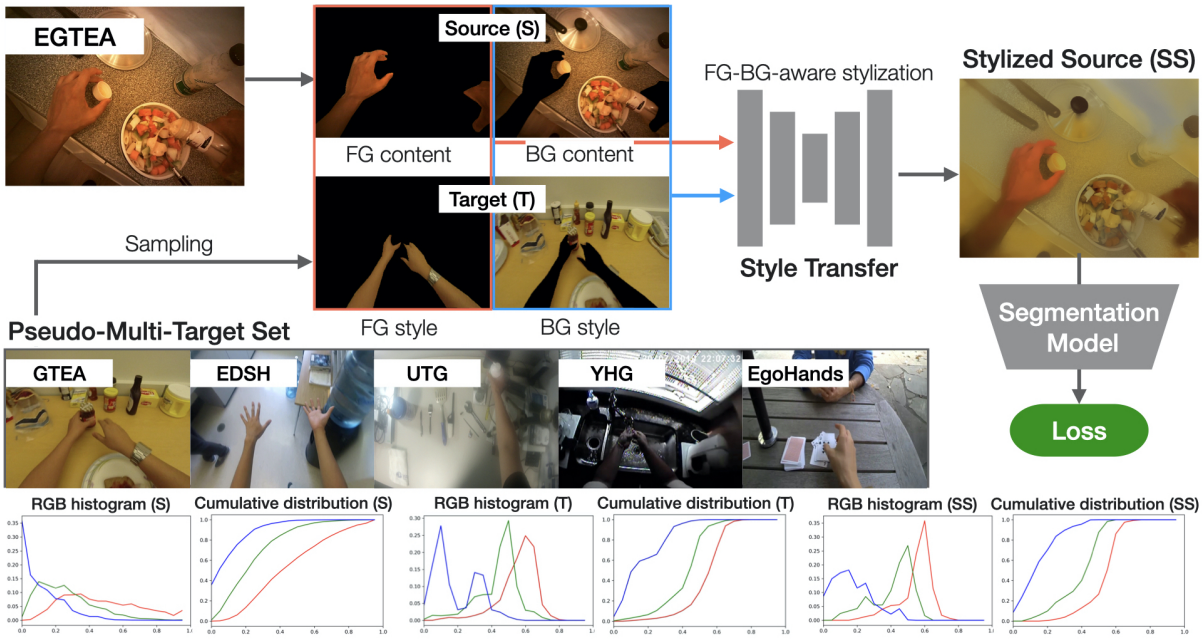

Figure 2: Method overview.

by maximizing the mutual information between domain labels and domain-specific features while minimizing the mutual information between the shared features. Recently, Chen et al. [3] proposed to blend multiple target domains and minimize the discrepancy between the source and the blended targets.

In the paper, by making use of a style transfer that enables multiple style representations, our method simultaneously transfers the appearances, and the segmentation model adapts to multiple targets.

2.3 Style Augmentation & Style Adaptation

Style transfer is a class of image processing algorithms that modify the visual style of an image while preserving its semantic content. For data augmentation, the style transfer extends lighting variations and synthesizes different texture as well. Geirhos et al. [10] created a Stylized ImageNet (SIN) using the style transfer trained on *Painter By Numbers (PBN)*, which provides shape-based representation of CNN for visual recognition. Jackson et al. [11] defined the randomization of color, texture, and contrast using a style transfer trained on the PBN while preserving geometry as **Style Augmentation**. This style randomization improves robustness to unseen domains. However, its drawback is to require a large amount of artistic images of PBN as style images.

For domain adaptation, the style transfer can be seen as a special domain adaptation problem with each style as a domain [12]. Here, the adaptation aligning style feature distributions of CNN can be defined as **Style Adaptation**. DCAN [7] presented channel-wise feature alignment for matching style feature statistics from two different domains, where the network jointly stylizes images and performs segmentation.

In this work, we adopt a style adaptation approach to align style feature distributions requiring a few images per target.

3. Proposed Method

3.1 Preliminaries

We first consider the problem of domain adaptation (DA). Let $\mathcal{S} = \{\boldsymbol{x}^{(s)}, \boldsymbol{y}^{(s)}\}$ be a source domain where $\boldsymbol{x}^{(s)}$ and $\boldsymbol{y}^{(s)}$ denote source data and pixel-wise labels, respectively. Target domain $\mathcal{T} = \{\boldsymbol{x}^{(t)}\}$ includes unlabeled target data $\boldsymbol{x}^{(t)}$. $\mathcal{S}$ and $\mathcal{T}$ underlie distributions $P_{\mathcal{S}}(\boldsymbol{x}^{(s)}, \boldsymbol{y}^{(s)})$ and $P_{\mathcal{T}}(\boldsymbol{x}^{(t)})$, in which $P_{\mathcal{T}}(\boldsymbol{x}^{(t)}) = \int P_{\mathcal{T}}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) d\boldsymbol{y}^{(t)}$ indicates the marginal distribution over target labels. We also define a segmentation network M. The goal is to learn the model M that is capable of correctly predicting labels $\boldsymbol{y}^{(t)}$ for target data $\boldsymbol{x}^{(t)}$. Domain adaptation assumes all unlabeled data derived from a single target domain, namely single-target domain adaptation (STDA).

Here we turn to consider multi-target domain adaptation (MTDA). Every unlabeled target instance $\boldsymbol{x}^{(t)}$ underlies k distributions $\{P_{\mathcal{T}_j}(\boldsymbol{x}^{(t)})\}_{j=1}^{k}$. Existing domain adaptation algorithms can address the problem by training k target-specific models $\{M_j\}_{j=1}^{k}$, respectively, and using the j-th target model M_j to classify the examples from the j-th target. In contrast, the goal of MTDA is to simultaneously adapt k targets $\{\mathcal{T}_j\}_{j=1}^{k}$.

3.2 Style Adapted DataBase

To achieve generalization to multi-target domains by a single model, we propose a dataset-level style adaptation to multiple domains using a style transfer. Our approach aligns style distributions between source and target domain via a weakly-supervised foreground-background separated stylization with a few target labels. The stylization produces a stylized source image with the source content and target style. We then create a stylized source dataset for the training of a segmentation model, which is called a

Style Adapted DataBase (SADB). The model trained on the SADB simultaneously generalizes to multiple target domains at once. The overview of creating the SADB is shown in Figure 2, which is divided into three steps: (1) preparing a pseudo-multi-target set, (2) building a foreground-background separated stylization, and (3) creating a stylized source dataset using the pseudo-multi-target set as a set of style images.

Step1: Prepare a pseudo-multi-target set. To begin with, we randomly collect m_j images in the target domain $\mathcal{T}_j$ $(1 \leq j \leq k)$, and we have a total of n $(= \Sigma_{j=1}^{k} m_j)$ images from all the target domains. We annotate only the n images for the following stylization. Here, we define a pseudo-multi-target set as:

$$\mathcal{PMT} = \{\{\boldsymbol{x}_i^{(t_j)}, \boldsymbol{y}_i^{(t_j)}\}_{i=1}^{m_j}\}_{j=1}^{k}, \tag{1}$$

where $\boldsymbol{x}_i^{(t_j)}$ $\left(\boldsymbol{y}_i^{(t_j)}\right)$ denotes the i-th target image (label) in the j-th target domain. We utilize the pseudo-multi-target set as a style set, one of which is fed into the style transfer to stylize a content image.

Step2: Build a semantics-aware stylization. Next, we design a foreground-background separated stylization aligning semantically corresponding areas between content and style images. We use a photo-realistic style transfer [13] that preserves the geometrical structure of the image introducing a smoothing step after the stylization. The network contains a style transform function $\mathcal{F}_1$ called PhotoWCT, and a photo-realistic smoothing function $\mathcal{F}_2$. Given a style image I_S and a content image I_C as inputs, the whole algorithm can be written as:

$$\mathcal{F}_2\Big(\mathcal{F}_1(I_C, I_S), I_C\Big). \tag{2}$$

$\mathcal{F}_1$ transfers the style of I_S to the content image I_C while minimizing structural artifacts in the output image. The key idea of PhotoWCT is to directly match feature correlations of the content image to those of the style image via the two projections. $\mathcal{F}_2$ processes the output to eliminate artifacts of the stylization and produce a more spatially consistent image. In our stylization pipeline, we separate the foreground and background of the content and style images using their labels, respectively. The pair is then employed to stylize these image regions. Finally, we concatenate the stylized foreground and background. The mask alignment promotes style features for semantically relevant regions.

Step3: Create a stylized source dataset. Here, we utilize a source dataset $\mathcal{S}$ as content inputs and the pseudo-multi-target set $\mathcal{PMT}$ as style inputs. Let $\mathcal{SS} = \{\boldsymbol{x}^{(ss)}, \boldsymbol{y}^{(ss)}\}$ be a stylized source domain with the source contents and the target styles. The stylized source image set $\{\boldsymbol{x}^{(ss)}\}$ contains the same number of images with the original dataset. Given a style input pair $(\boldsymbol{x}^{(t_j)}, \boldsymbol{y}^{(t_j)})$ in $\mathcal{PMT}$ and a content input pair $(\boldsymbol{x}^{(s)}, \boldsymbol{y}^{(s)})$ in $\mathcal{S}$, $\boldsymbol{x}^{(ss)}$ and $\boldsymbol{y}^{(ss)}$ are written as:

$$\boldsymbol{x}_i^{(\mathrm{ss})} = \mathcal{F}_2\left(\mathcal{F}_1\left(\boldsymbol{x}_i^{(s)}, \boldsymbol{y}_i^{(s)}, \boldsymbol{x}_l^{(t_j)}, \boldsymbol{y}_l^{(t_j)}\right), \boldsymbol{x}_i^{(s)}\right), \tag{3}$$

$$\boldsymbol{y}_i^{(\mathrm{ss})} = \boldsymbol{y}_i^{(\mathrm{s})}, \tag{4}$$

where i denotes the index of the source data, and l, j are uniformly sampled by $l \sim U(1, m_j)$, $j \sim U(1, k)$. We use the ground truth labels in the source domain for the stylized dataset because the content is shared between the two datasets. To validate the effect of style adaptation, we show RGB distributions and their cumulative distributions in Figure 2. Both distributions in the stylized source domain (right) are close to ones in the target domain (middle), and accordingly our approach leads to domain adaptation in color space.

3.3 Training of Hand Segmentation

Suppose we have a hand segmentation network as the model M, hand segmentation can be defined as a binary semantic segmentation task. The model M is trained on the stylized source dataset $\mathcal{SS}$ and a mini-batch used per iteration is randomly sampled from the dataset. A task loss can be written as a binary cross-entropy loss:

$$L_{seg}(\boldsymbol{x}^{(ss)}, \boldsymbol{y}^{(ss)}) = -\sum_{h,w} y_{h,w}^{(ss)} \log P_{h,w}^{(ss)} + (1 - y_{h,w}^{(ss)}) \log(1 - P_{h,w}^{(ss)}), \tag{5}$$

where $P^{(ss)} = M(\boldsymbol{x}^{(ss)})$ is the final output of the model M given a stylized source image $\boldsymbol{x}^{(ss)}$, and the sample index is omitted for simplicity. Besides the pseudo-multi-target set $\mathcal{PMT}$, we do not use any target image for refining the model prediction on the target data.

4. Experiments

This section validates our proposed SADB in the real-world domain adaptation of hand segmentation in egocentric videos. We first conduct an experiment to verify cross-dataset generalization ability across the target domains. In ablation studies, we explore the support of the adaptation from the target domains excluding a test one. We also investigate the effect of the mask alignment, and conduct a blur robustness test.

4.1 Experimental Details

Datasets. Following [1], we set first-person vision datasets containing various illuminations. EGTEA [14] is used as the source domain, and GTEA (G) [15], EDSH (E) [4], UTG (U) [16], YHG (Y) [17], and EgoHands (EH) [18] are utilized as the target domains. EDSH-1 (E1) is a training dataset, and EDSH-2 (E2) and EDSH-K (EK) are testing ones. In our experiments, we resize these images to 256×256 pixels for training the segmentation network.

For the proposed SADB, we prepare the pseudo-multi-target set $\mathcal{PMT}$-50 from training images of the target domains, where $m_j = 10$ $(1 \leq j \leq k)$, $k = 5$. We also create pseudo-target sets $\{\mathcal{PT}_j\text{-}10\}_{j=1}^{5}$. $\mathcal{PMT}$-50 has a total of 50 images where we randomly collect 10 images per target. $\mathcal{PT}_j$-10 contains a total of 10 images only from the target domain $\mathcal{T}_j$ corresponding to the test one. To confirm the effect of style adaptation from the target domain used in testing and the other target domains, we create pseudo-multi-target sets $\{\mathcal{PMT}_{\setminus j}\text{-}40\}_{j=1}^{5}$ where 10 images are randomly sampled per target, but the test domain $\mathcal{T}_j$ is excluded.

Table 1: Cross-dataset generalization ability. Mean IoU (%) is used for the evaluation. # denotes the number of models used in training. Bold and blue letters indicate the best value and the second best one, respectively.

Method	Source	Style set	GTEA	EDSH-2	EDSH-K	UTG	YHG	EgoHands	Avg.	#
RefineNet [6]	EGTEA	-	88.45	69.36	72.05	54.81	28.31	40.19	58.86	1
Bayesian RefineNet	EGTEA	-	88.96	76.32	75.76	58.32	36.19	42.35	62.98	1
CBST [9]	EGTEA	-	87.66	73.53	72.07	56.27	35.39	42.93	61.31	5
BDL [8]	EGTEA	-	86.09	72.40	73.60	62.10	41.70	43.90	63.30	5
Bayesian RefineNet + UMA [1]	EGTEA	-	89.45	79.65	78.12	67.62	52.23	46.65	68.95	5
Bayesian RefineNet + UMA + HS [1]	EGTEA	-	89.90	**80.25**	**79.51**	68.27	55.96	46.60	70.01	5
Ours-MTDA	SADB	$\mathcal{PMT}$-50	89.06	75.71	77.29	74.35	**59.27**	49.32	70.83	1
Ours-STDA	SADB	$\mathcal{PT}_j$-10	**91.00**	74.67	78.77	**81.83**	55.11	**53.15**	**72.42**	5
Target only	-	-	91.97	84.23	76.85	90.81	81.84	79.99	84.28	5

Table 2: Style adaptation from a test domain and the others.

Style set	G	E2	EK	U	Y	EH	Avg.
-	88.45	69.36	72.05	54.81	28.31	40.19	58.86
$\mathcal{PMT}_{\setminus j}$-40	88.40	**75.57**	77.04	75.89	43.77	47.81	68.08
$\mathcal{PT}_j$-10	**88.88**	74.67	**78.77**	**81.83**	**55.11**	**53.15**	**72.07**

Table 3: Mask aligned stylization vs Unaligned stylization.

Method	G	E2	EK	U	Y	EH	Avg.
Baseline	88.45	69.36	72.05	54.81	28.31	40.19	58.86
Unaligned	84.11	**77.97**	75.81	61.14	31.83	32.29	60.53
Aligned	**89.06**	75.71	**77.29**	**74.35**	**59.27**	**49.32**	**70.83**

Baseline & Comparison methods. We use RefineNet [6] as a backbone network for hand segmentation. We compare the performance of hand segmentation with its extensions to the Bayesian method and domain adaptation, and recent domain adaptation models for semantic segmentation. These methods are shown in Table 1.

Evaluation. For evaluating the performance, we report mean Intersection over Union (mIoU). For the robustness test, we add corruptions and perturbations to the test data as the method of [19]. We report how the performance degrades from the one without corruptions (Clean). Corruption mIoU (C-mIoU) is calculated with the formula:

$$\text{C-mIoU}_c^{db} = \left(\sum_{s=1}^{5} S_{s,c}^{db}\right) / \left(\sum_{s=1}^{5} S_{s,\text{Clean}}^{db}\right), \qquad (6)$$

where c and s denote the corruption type and the level of distortion severity ($1 \leq s \leq 5$), respectively, and db indicates the database used in training. $S_{s,c}^{db}$ is the value of mIoU.

4.2 Results

Cross-dataset generalization. The cross-dataset generalization performance of different methods is shown in Table 1. The method using $\mathcal{PMT}$-50 as a style set (Ours-MTDA) generalizes well to all the target domains without the model specifying to the test domain. Ours-MTDA achieves 70.83% on average, which is superior to Bayesian RefineNet, the recent domain adaptation methods (CBST [9], BDL [8]), and the state-of-the-art UMA [1] methods. Furthermore, the method specifying the SADB to a single target using $\mathcal{PT}_j$-10 (Ours-STDA) significantly improves the performance by 27.02% on UTG, 26.8% on YHG, and 12.96% on EgoHands. The average score over all the target domains achieves 72.42%. One finding is that the stylization is effective on the domains with smaller within domain gaps, especially in UTG, since a few style images can represent most of the target appearance.

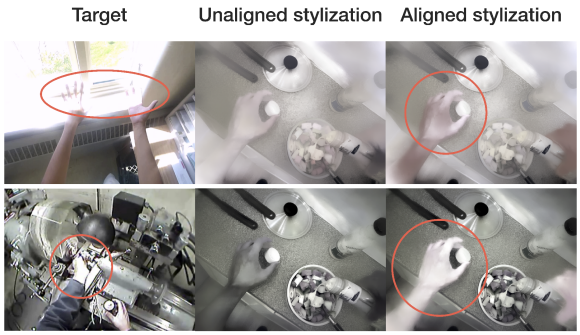

Figure 3: The effect of mask aligned stylization. (Top: EDSH1, Bottom: YHG)

Style adaptation from other target domains. To verify the support of style adaptation from different domains, we create test-domain-specific SADBs using $\mathcal{PMT}_{\setminus j}$-40, where the test domain $\mathcal{T}_j$ is excluded. We then train the RefineNet on them. Although the stylization using $\mathcal{PMT}_{\setminus j}$-40 as a style set does not exploit any target image in the test domain during training, the performance improves by 10% from the baseline. This can be explained that illumination or lighting conditions are partially shared across target domains. Hence, our multi-target style adaptation promotes each other.

Mask aligned stylization vs unaligned stylization. To verify the effectiveness of the mask alignment, we compare the segmentation performances trained on EGTEA only (baseline), SADB without the foreground-background separation (unaligned), and the proposed SADB with the mask alignment (aligned) in Section 3.. As shown in Table 3, the unaligned stylization provides little marginal gain from the baseline, but the aligned one significantly improves the performance across most target domains. Qualitatively, red circles in Figure 3 illustrate that the aligned stylization transfers the light on hands to the stylized image. In contrast, the unaligned one alters the style in a flat tone. Especially, the aligned stylization can reproduce

Table 4: Robustness test for different blur distortions.

Blur	Method	G	E2	EK	U	Y	EH	Avg.
Clean	-	100	100	100	100	100	100	100
Gaussian	baseline	96	93	87	88	**106**	85	93
Gaussian	SADB	**97**	**100**	**96**	**95**	**106**	**97**	**99**
Defocus	baseline	95	92	84	86	103	82	90
Defocus	SADB	**97**	**99**	**96**	**95**	**106**	**96**	**98**
Glass	baseline	**93**	77	75	86	**89**	79	83
Glass	SADB	92	**89**	**80**	**92**	87	**83**	**87**
Motion	baseline	95	91	87	86	**111**	86	93
Motion	SADB	**96**	**96**	**95**	**92**	102	**99**	**97**
Zoom	baseline	83	67	74	69	**96**	57	74
Zoom	SADB	**88**	**78**	**84**	**76**	89	**82**	**83**
Total	baseline	92	84	81	83	**101**	78	87
Total	SADB	**94**	**92**	**90**	**90**	98	**91**	**93**

the blown-out highlight on the left hand in the bottom image of the YHG domain.

Robustness test. A common failure case in egocentric hand segmentation is motion blur due to dynamic and unpredictable camera motion by a camera wearer, as discussed in [5]. To reveal the property of the model trained on the SADB, we conduct a blur robustness test. Perturbations are enforced by 5 blur algorithms: gaussian, defocus, glass, motion, and zoom blur. The result is shown in Table 4. For all the blur algorithms, the proposed method prevents performance degradation against the model trained on EGTEA.

5. Conclusion

This paper presents a weakly-supervised foreground-background separated stylization from a few images per target using a style transfer. Through the semantics-aware stylization, we create a Style Adapted DataBase (SADB), which is a style aligned source dataset for adapting to multiple target domains. The model trained on the SADB achieves the best performance on the cross-dataset generalization benchmark of hand segmentation. In the ablation studies, we show that the SADB leverages the effect of style adaptation from different target domains, mask alignment can transfer highlight on hands to other domains, and the model demonstrates blur robustness, which is a desirable property in first-person vision.

This work focuses on the appearance-level domain shift, but does not explicitly cope with the spatial mismatching of the labels between source and target domain. Handling geometric domain gap is an important problem for future work.

Acknowledgment

This work was supported by JST AIP Acceleration Research Grant Number JPMJCR20U1, Japan.

References

[1] M. Cai, E. Lu, and Y. Sato. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14380–14389, 2020.

[2] O. Rudovic K. Bousmalis B. Gholami, P. Sahu and V. Pavlovi. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*.

[3] Z. Chen, J. Zhuang, X. Liang, and Liang Li. Blending-target domain adaptation by adversarial meta-adaptation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2243–2252, 2019.

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

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

[6] G. Lin, A. Milan, C. Shen, and I. D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5168–5177, 2017.

[7] Z. Wu, X. Han, Y.-L. Lin, M. Gokhan Uzunbas, T. Goldstein, S. Nam Lim, and L. S. Davis. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In *Proceedings of the European Conference on Computer Vision*, pages 518–534, 2018.

[8] Y. Li, L. Yuan, and N. Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6929–6938, 2019.

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

[10] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations*, 2019.

[11] P. T. Jackson, A. A. Abarghouei, S. Bonner, T. P. Breckon, and B. Obara. Style augmentation: data augmentation via style randomization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 83–92, 2019.

[12] Y. Li, N. Wang, J. Liu, and X. Ho. Demystifying neural style transfer. In *International Joint Conferences on Artificial Intelligence*, page 2230–2236, 2018.

[13] Y. Li, M. Y. Liu, X. Li, and J. Kautz. M. H. Yag. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision*, pages 468–483, 2018.

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

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

[16] 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.

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

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

[19] D. Hendrycks and T. Dietteric. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019.