Ego2Hands: A Dataset for Egocentric Two-hand Segmentation and Detection

Fanqing Lin Tony Martinez
Brigham Young University, Provo UT 84602, USA

Abstract

Hand segmentation and detection in truly unconstrained RGB-based settings is important for many applications. However, existing datasets are far from sufficient both in terms of size and variety due to the infeasibility of manual annotation of large amounts of segmentation and detection data. As a result, current methods are limited by many underlying assumptions such as constrained environment, consistent skin color and lighting. In this work, we present a large-scale RGB-based egocentric hand segmentation/detection dataset Ego2Hands that is automatically annotated and a color-invariant compositing-based data generation technique capable of creating unlimited training data with variety. For quantitative analysis, we manually annotated an evaluation set that significantly exceeds existing benchmarks in quantity, diversity and annotation accuracy. We show that our dataset and training technique can produce models that generalize to unseen environments without domain adaptation. We introduce Convolutional Segmentation Machine (CSM) as an architecture that better balances accuracy, size and speed and provide thorough analysis on the performance of state-of-the-art models on the Ego2Hands dataset.

1. Introduction

With the rapid growing usage of wearable technologies generating massive volumes of egocentric image data [4, 3, 2, 1], the ability for machines to understand the human hands becomes crucial for applications such as human-computer interaction (HCI), activity logging, gesture/sign language recognition and VR/AR. Consequently, hand detection and segmentation are fundamental in areas such as 2D/3D hand pose estimation [36, 27, 23] and gesture recognition [7, 17]. However, hand segmentation on images in the wild is extremely challenging due to numerous factors: vastness of the color space, different skin color/texture, complex background noise, motion blur, lighting type/color, shadow features, speed and model size requirement, etc. As a result, existing color-based approaches can only perform in constrained environments with proper lighting and skin color consistent with the training data. These limitations are largely due to the lack of annotated segmentation data, a common limiting factor for segmentation tasks because manual annotation is oftentimes required but infeasible for large-scale data generation.

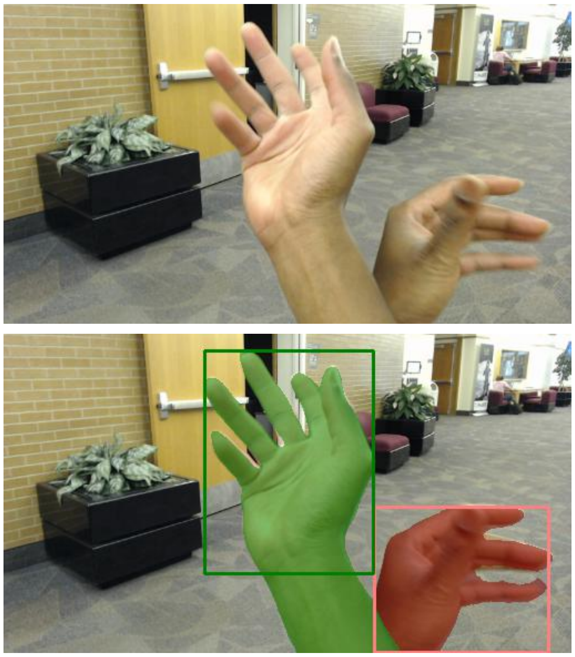

Figure 1: Our proposed dataset and training scheme enables domain generalization for two-hand segmentation and detection. Given an image with a new environment and hands not present in the training data (top), the trained models can provide accurate segmentation and detection results for both hands with free interaction (bottom).

In this work, we aim to push the boundary for the task of real-time egocentric two-hand segmentation and detection on images in the wild (Fig. 1). Since hand segmentation and detection are highly correlated and both imperative for subsequent applications, we find it natural to tackle both tasks simultaneously.

We first address the issue of the lack of annotated data. In general, real-world RGB data with segmentation ground truth is very labor-intensive to annotate. For this reason, existing hand segmentation datasets [21, 19, 7, 17, 20, 12] lack the quantity and sufficient variety necessary for learning-based approaches. Although synthetic data [23] with perfect ground truth can be generated with little cost, methods trained on synthetic image data cannot be directly applied on real-world data as CNNs are sensitive to even small textural differences between domains. We propose a novel segmentation data collection method for egocentric hands that can automatically annotate massive amounts of data for only the right hand in a green screen setting, and a corresponding data generation technique that composites training instances by combining a pair of randomly selected right hands with one horizontally flipped as the left hand. In order to develop a color-invariant approach, we explore the grayscale image space coupled with an edge map as input space and show successful generalization to unseen environments. This data generation method can push segmentation models beyond the limitation of a fixed-sized training set and evaluation set and enable models to produce accurate segmentation and detection results in unseen environments without domain adaptation, which can also be easily applied to further improve model accuracy for specific environments.

We introduce Ego2Hands which includes a training set with ∼180,000 unique right hand instances and an evaluation set with 2,000 manually annotated frames from diverse video sequences. In-depth comparison between Ego2Hands and previous datasets shows the superiority of our dataset in quantity and diversity (Section. 3). For quantitative analysis, we provide comprehensive comparison between the state-of-the-art approaches on our dataset and find that existing architectures lack the proper balance of accuracy, model size and speed, which is necessary for real-world applications. To this end, we introduce a well-balanced architecture *Convolutional Segmentation Machine* (CSM), where the 1st stage outputs fast and accurate predictions and the 2nd stage provides refinement with increased resolution. Our work opens up promising directions for two-hand gesture control systems using only low-cost RGB input devices.

2. Related Works

Depth-based methods. Early works [37, 35, 33] utilized Randomized Decision Forests (RDF) on depth image to obtain the hand segmentation, which allows multicore parallelization with fast inference time suitable for real-time applications. [36] introduced a Fully Convolutional Network (FCN) that segments the left and right hand for fast tracking of two interacting hands in egocentric viewpoint. Similarly, [9] proposed a hybrid encoder-decoder architecture with skip-connections for two-hand segmentation from a third-person viewpoint. Recently, [22] extended the segmentation task to 8 classes to include arms and objects and trained a FCN on synthetic data with a level of generalization on real depth data. Following [36], [27] used a Correspondence Regression Network to estimate two-hand segmentation prior to hand pose estimation, which shows the significance of separate segmentation of the two hands for pose estimation as it provides information on how interacting hands occlude each other. Note that for depth-based approaches, segmentation ground truth is obtained by color thresholding, requiring subjects to wear thin colored gloves. Therefore, datasets for depth-based hand segmentation are not suitable for training RGB-based approaches.

Color-based methods. Depth cameras have additional setup overhead and indoor requirements with higher power consumption and cost, making its applicable applications more limited compared to ubiquitous RGB cameras. Before the revolution of deep convolutional networks in the field of computer vision, [30, 15, 34] proposed motion-based approaches for binary foreground segmentation with the assumption that the motion pattern is different for the foreground and background. Some methods [5, 16, 32] rely on consistent skin color for hand segmentation. Being aware of the possible illumination difference in scenes, [19, 18] trained multiple hand detectors on a mixture of local and global appearance features from various scenes and adaptively selected detectors based on the test images. To address two-hand segmentation with possible slight inter-hand occlusion, [8] performed binary hand segmentation and left-right hand splits based on the distribution of angle/position of hands as well as temporal superpixels.

Recent approaches utilize convolutional deep networks as stronger appearance models. [7] used a CNN to classify proposed bounding boxes and performed hand segmentation using Grabcut inside the bounding boxes. They also demonstrated simple static gesture recognition using the obtained segmentation masks. Although the CNN is designed to classify detected hands as one of four interacting hands, the window proposal and classification algorithm do not address the issue of similar-object occlusion between hands. [17] later proposed to segment the hands directly using RefineNet [25] and performed extensive evaluation on multiple datasets for binary hand segmentation in less constrained environments. In order to better generalize the models to unseen scenes without requiring annotated data for training in the new domain, [12] proposed a Bayesian CNN-based approach to estimate pseudo-labels in the target domain with a hand shape discriminator for unsupervised domain adaptation. Despite achieving promising cross-dataset accuracy, their domain adaptation technique is valid for binary-label segmentation only. [23] introduced the first color-based two-hand segmentation method for complex interactions using an encoder-decoder residual network that also estimates the hand heatmap energy for detection. However, their model was only able to train on synthetic data and cannot be applied to images in the real-world domain. [38] trained a UNet [31] on a mixture of synthetic and noisy real-world data for two-hand segmentation in a laboratory environment from a third-person viewpoint. To perform both hand segmentation and detection, we adopt the method of [23] that adds the hand heatmap energy channels to the segmentation output channels for the segmentation models in our studies. Our experiments show that the addition of hand heatmap energy output channel complements the segmentation task and does not negatively impact segmentation accuracy.

3. Hand Segmentation Datasets

3.1. Existing Datasets

Pioneering work [19] contributed three egocentric videos (EDSH1, EDSH2 and EDSH-kitchen) with varying illumination for training and evaluation of binary hand segmentation. For activity recognition, [21] proposed the Georgia Tech Egocentric Activity Dataset (GTEA) with 663 annotated frames consisting of two-hand labels (no interhand occlusion). [20] later published an extended version (EGTEA) with 13,847 binary-label annotated frames. To enable hand segmentation in more unconstrained settings, [7] introduced EgoHands as the first large-scale hand segmentation dataset with 4,800 annotated frames consisting of a maximum of 4 interacting hands. For the same purpose, [17] additionally introduced EgoYouTubeHands (EYTH) with ∼1290 annotated frames from three Youtube videos and HandOverFace (HoF) with 300 annotated frames from third-person Web images. To demonstrate cross-dataset adaptation performance, [12] annotated 855 and 488 frames for human grasping datasets UTG [11] and YHG [10] respectively. To address the issue of data scarcity, [23] introduced two large-scale synthetic dataset (Ego3DHands) with a total of over 100,000 annotated frames on two hands.

3.2. Ego2Hands

As existing datasets with real-world data require manual annotation, they lack the quantity and variety needed for learning-based hand segmentation on images in the wild for real-world applications. Synthetic datasets consist of data in a different statistical distribution from real-world data and therefore can only be used for theoretical research analysis or mixed training with real-world data for limited knowledge transferral.

To solve the problem of data scarcity, we introduce a large-scale dataset Ego2Hands that consists of 188,362 annotated frames for only the right hand. Segmentation masks are obtained by automatically removing the background in a green screen setting. 22 participants with diverse skin colors and hand features are selected and instructed to perform free one-hand motion while recording using a head-mount webcam (Logitech C922) at 30 fps. This process allows simple and fast data collection for segmentation data. During training, we composite images online by randomly selecting two right hand images, flipping one horizontally to create the left hand, and inserting a random background image. For background images, we use the 19,216 images provided by [23] with the additional 14,997 high-quality images in the DAVIS datasets [28, 29], which results in approximately 1.21×10^{15} unique hand-scene combinations prior to data augmentation.

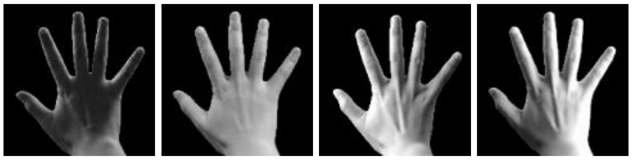
Figure 2: Hand images (grayscale) with different visual shadow features from different lighting directions.

Despite the massive quantity in training instances, it is still unrealistic for deep networks to learn the complete RGB space. For instance, for hands under a particular colored lighting, learning-based models would need sufficient training data with hands in that specific color. This is an important issue rarely addressed by previous works as their proposed datasets only contain light skin color under normal lighting. Consequently, we explore the grayscale image space coupled with an image edge map as inputs for a truly color-invariant approach. In the grayscale domain, we find two major factors crucial for generalization in the real-world domain: brightness and shadow features. For diversity in brightness, we scale the pixel values of both hands to shift the means to a randomly selected value $\beta \in [15, 240]$ while keeping the image values always clipped within [0, 255]. Variation in the brightness of the hands also contributes significantly to diversity in skin colors. For different shadow features, we include light sources from various directions during data collection (Fig. 2 shows the visual difference in shadow features due to the direction of the light source).

To obtain the hand energy for detection, we follow [23] by generating an internal synthetic dataset Ego3DHands$_R$ with only the right hand, and jointly train their proposed HandSegNet on Ego3DHands$_R$ and Ego2Hands using the following loss,

$$\mathcal{L}_{combined} = \mathcal{L}_{seg}^{synth} + \mathcal{L}_{seg}^{real} + \mathcal{L}_{energy}^{synth} \tag{1}$$

where $\mathcal{L}_{seg}^{synth}$ and $\mathcal{L}_{seg}^{real}$ are the Cross Entropy Loss as the segmentation loss of the synthetic and real right hands, and $\mathcal{L}_{energy}^{synth}$ is the MSE loss for the heatmap energy of the synthetic right hands obtained using the ground truth 2D joint locations. We exploit the feature that both domains contain right hands with no background noise. In doing so, we successfully transfer the knowledge of the hand energy from

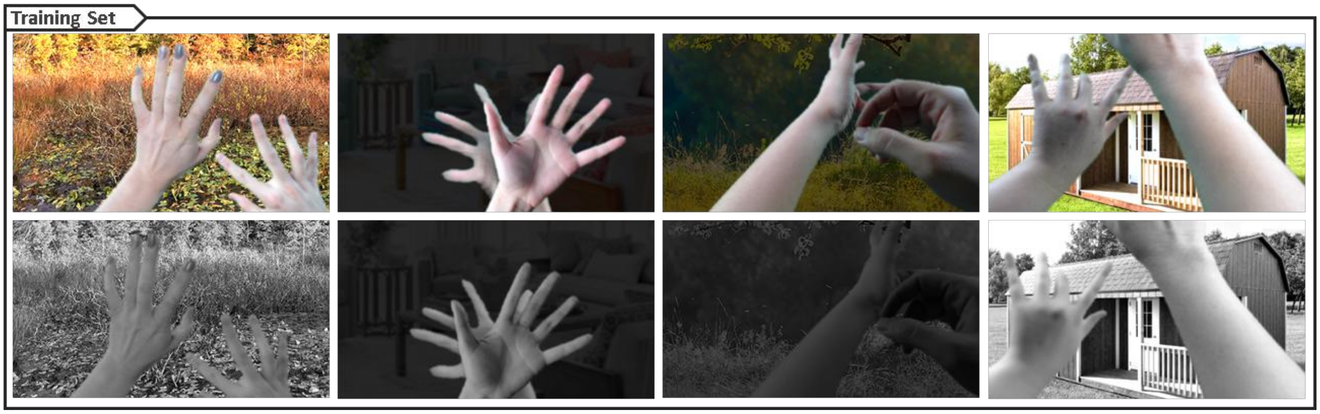

Figure 3: The top row shows the original composited images using the training set of Ego2Hands and the available background images. The bottom row shows the data augmented version in grayscale for training.

synthetic data to real-world data and generate the hand energy for all training instances in Ego2Hands. This novel knowledge transferral method that automatically generates the heatmap energy data in the target domain is very efficient as ground truth data for object detection commonly requires extensive manual annotation [26, 14]. Sample training instances are provided in the supplementary document.

Some datasets [7, 17] contain annotated hand segmentation that excludes the arm, therefore combining the task of segmentation and detection into one, which is valid in absence of occlusion. Others [21, 19, 20, 12] include the arm in segmentation and neglect the task of hand detection. We argue that it is more natural to keep the two tasks separate by including the arm for segmentation because the boundary line is ambiguous. In addition, we can address hand detection in the form of heatmap energy. Note that removing the arm in hand segmentation also requires manual annotation infeasible for large-scale data generation. We show in Section. 5 that our training data enables models to achieve high accuracy in both tasks simultaneously.

With the obtained hand energy, we are able to composite more realistic training images by selecting the proper overlaying order. After random selection of the left and right hand from Ego2Hands, the hand with the larger energy sum is selected to be overlaid on top of the other hand. We discover that naive overlaying creates the unrealistic feature of green color bleeding at the hand boundaries, which leads to a noticeable decline in the model's ability to generalize to real-world data in our experiments. Accordingly, we apply dilation and gaussian blur on the original alpha-channel to create smoother hand boundaries for overlay. The green color bleeding also becomes unrecognizable with smooth-edged overlaying in the grayscale domain.

For each composited image, we further data augment by applying 1) random horizontal and downward vertical translation within reasonable ranges on each hand, 2) random smoothing with various kernel sizes to simulate blur from motion or auto-focus, 3) random brightness on the hands and background images, 4) Random horizontal flips and cropping on background images, and 5) 10% drop rate for each hand (mutually exclusive). Thus our compositing-based approach can generate unlimited training images with variety. For domain adaptation on specific environments, we simply use the background images collected from that scene for compositing training images. See Fig. 3 for generated sample images for training.

To support quantitative evaluation, we introduce an evaluation set that includes 8 videos each with 250 annotated frames. We select 4 additional participants with diverse skin tones to perform free two-hand motion in 8 different scenes with various lighting conditions. We include more details in the supplementary document to demonstrate the diversity in our collected sequences. Inspired by the video object segmentation method [24], we exploit the temporal consistency in video sequences by obtaining manual annotation with the help of HandSegNet [23] pretrained on Ego2Hands to minimize annotation time. For frame $\boldsymbol{F}_i$, our pretrained model produces a semi-accurate segmentation, which we manually refine with the support of Grabcut to create the ground truth $\boldsymbol{S}_i$. The model is then finetuned on $\boldsymbol{S}_i$ for a more accurate prediction on $\boldsymbol{F}_{i+1}$. We also annotate the hand energy $\boldsymbol{E}_i$ for approximate location of both hands. Fig. 4 shows a comparison of annotation accuracy between Ego2Hands and other datasets. Due to the gap between the gesture space of Ego3DHands$_R$ and Ego2Hands, some of the generated energy data also needs refinement. We use a similar tool to refine the generated energy for the entire training set of Ego2Hands semi-automatically with human supervision.

We show in Table. 1 a detailed comparison between Ego2Hands and existing benchmark datasets. Ego2Hands consists of significantly more annotated hand frames capable of generating unlimited training data with pixel-

Datasets	Type	#Annotated Frames	#Hand Instances	#Subjects	#Scenes	Objects	#Classes	Resolution
GTEA [21]	Real	663	1231	4	1	Yes	3	720 × 405
EDSH [19]	Real	743	-	1	3	Yes	2	1280 × 720
EgoHands [7]	Real	4800	15053	4	3	Yes	5	1280 × 720
EYTH [17]	Real	1290	2600	-	-	Yes	2	384 × 216
HoF [17]	Real	300	507	-	-	No	3	384 × 216
EGTEA [20]	Real	13847	-	32	1	Yes	2	960 × 720
UTG [12]	Real	855	-	5	2	Yes	2	480 × 360
YHG [12]	Real	488	-	4	-	Yes	2	480 × 360
Ego3DHands [23]	Synth	110,000	∼214,500	1	-	No	3	960 × 540
Ego2Hands (Ours)	Real	188,362 (train) 2,000 (test)	∞ (train) 4,000 (test)	22 (train) 4 (test)	- (train) 8 (test)	No	3	800 × 448

Table 1: Statistics of available hand segmentation datasets. Datasets with #Classes = 2 only support binary segmentation.

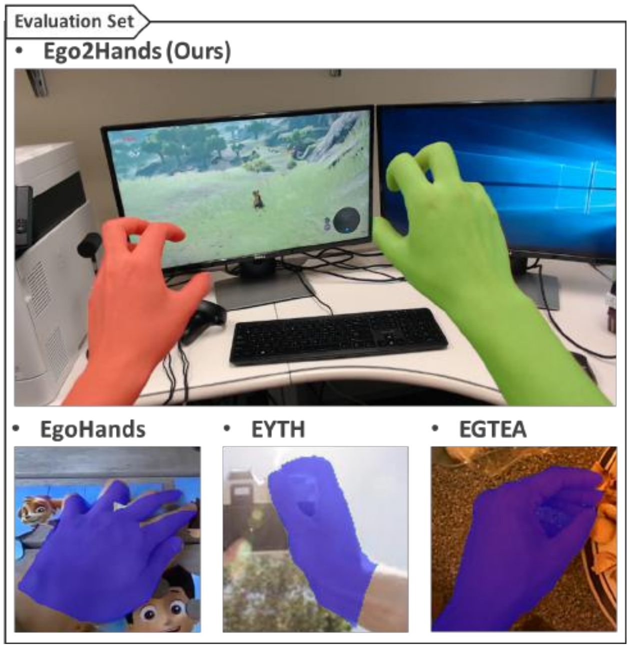

Figure 4: Illustration of the difference in annotation accuracy between datasets. Existing datasets contain false positive labeling for gaps and holes and potentially inaccurate boundaries (**Best viewed in magnification. Annotated masks are overlaid in colors.**)

accurate segmentation annotation on both hands using our compositing-based approach. We also provide the largest and most diverse evaluation set necessary for comprehensive evaluation in various real-world practical settings. Ego2hands provides significant improvements in annotation quality, quantity and data diversity.

4. Convolutional Segmentation Machine

Inspired by the Convolutional Pose Machine [40] that sequentially improves the heatmap estimation for 2D keypoints in six stages, we introduce Convolutional Segmentation Machine (CSM) that outputs segmentation prediction in two stages. The task of 2D keypoint estimation is fundamentally different from semantic segmentation as the former estimates the approximate location of the target keypoints in form of heatmap energy and does not require high-resolution output. On the other hand, semantic segmentation requires higher precision and resolution for classification of each pixel in the image.

Accordingly, we propose an encoder-decoder architecture for the task of hand segmentation as shown in Fig. 5. We couple the grayscale image and the edge map obtained from the original RGB image for a 2-channel input image and resize the image size to 288 × 512. Each downsampling and upsampling residual layer consists of 3 bottleneck and 3 deconvolutional bottleneck blocks respectively. For stage 1, the network output has 1/4 of the original resolution and is concatenated to the intermediate network feature. Stage 2 refines the results from stage 1 with increasing resolution that is half of the original resolution. Empirically we find that an additional stage 3 that returns the output to the original resolution does not produce higher segmentation accuracy. At test time, our 2-stage design gives users the flexibility of choosing between speed and accuracy in real-world applications. Section. 5 shows that both stages are capable of producing competitive results.

The segmentation output is trained on the 3 classes (left hand, right hand and background) with cross-entropy loss. To perform hand detection as well as segmentation, we add 3 additional output channels with sigmoid activations trained using mean squared error loss for the energy regression of the 3 classes. Quantitative analysis in Section. 5.2 shows that segmentation models can perform both tasks well without compromising accuracy.

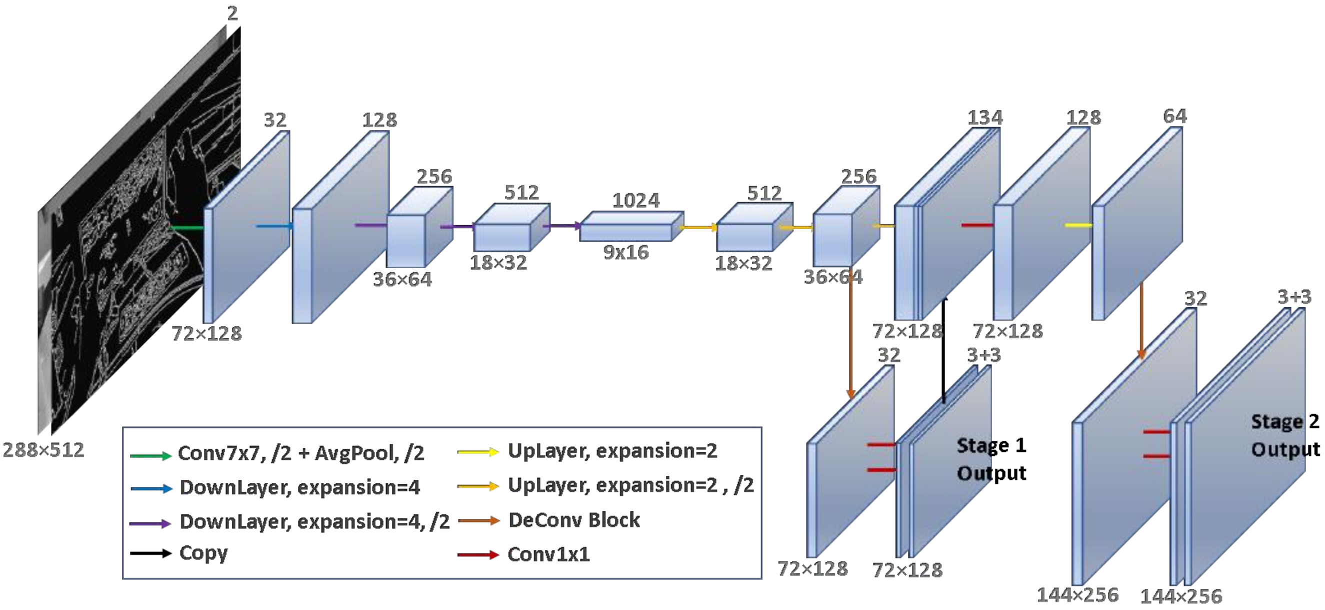

Figure 5: Architecture of the proposed Convolutional Segmentation Machine. #channels is denoted at the top of the boxes.

5. Experiments

We evaluate existing state-of-the-art methods on the proposed evaluation set of Ego2Hands (8 sequences each with 250 annotated frames) and compare the two-hand segmentation and detection accuracy as well as the corresponding model sizes and inference speed. We use the mean Intersection over Union (mIoU) as the metric for the segmentation task. For hand detection, we use the conventional metric of Average Precision that classifies a detection bounding box as correct if its IoU between the ground truth bounding box exceeds 50% ($AP_{0.5}$). The predicted bounding boxes are obtained using the output energy thresholded at 0.5. The closing operation with a kernel size of 7 is performed on the energy for noise removal. The ground truth bounding boxes are obtained using the annotated hand energy heatmaps.

We compare the models' performance w/wo the input edge map and the additional output energy channel to justify our design choices. As it is impossible for static pretrained models to produce highly accurate results in all possible scenes, we also perform experiments to study the impact of domain adaptation. To support scene-specific adaptation, we include in the evaluation set a collected background sequence ($\sim$30 seconds) for each evaluation sequence. The background collection process simulates an environment scanning procedure using prospective egocentric color-based hand tracking devices.

The following architectures are selected for evaluation:

- UNet and UNet$_{1/8}$ [31]. We evaluate using the standard UNet and a version with 1/8 of the original network width. Previous work [39] has shown that reducing the number of UNet input feature channels from 64 to 8 results in much more compact model while preserving its ability for binary-label hand segmentation.
- RecUNet and DRU-Resnet50 [39]. It is proposed that integrating recursions on the internal state of UNet$_{1/8}$ can produce higher segmentation accuracy. We select RecUNet-DRU(4) and DRU-Resnet50 with Dual-gated Recurrent Unit (DRU) and step size = 3 for evaluation as these two models achieved state-of-the-art results on multiple datasets for binary-label hand segmentation.
- SegNet [6]. Primarily motivated by scene understanding applications, SegNet is a popular semantic segmentation architecture with balance in model size and accuracy that fits well with the task of two-hand segmentation on images in the wild.
- ICNet [41]. Proposed for real-time semantic segmentation, ICNet with multi-resolution image cascade achieves high accuracy and impressive generalization with 1/4 of the output resolution. It is apparent that lower output resolution can avoid unnecessary deconvolution layers and result in faster inference speed.
- DeepLab V3+ [13]. Targeting high-quality semantic segmentation, DeepLab v3+ improves its predecessor by adding a decoder module to further refine segmentation results. We use Resnet-101 as the encoder for this model in our experiments.
- RefineNet [25]. Using Resnet-101 as encoder, RefineNet uses multi-path refinement to exploit features available in the down-sampling process for high-quality segmentation. [17] adopted it for the task of

Model	#Params	Inference time (ms)	Pretrained				Adapted	
			edge ✗ energy ✗	edge ✓ energy ✗	w/ edge & energy		w/ edge & energy	
					mIoU	$AP_{0.5}$	mIoU	$AP_{0.5}$
UNet$_{1/8}$ [31]	0.2M	9.5	0.722	0.749	0.754	0.633	0.844	0.739
RecUNet [39]	1.1M	78.1	0.812	0.834	**0.844**	0.805	0.874	0.839
CSM-stage1 (Ours)	9.7M	25.4	0.721	0.802	0.792	0.836	0.878	0.917
CSM-stage2 (Ours)	10.0M	35.9	0.728	0.811	0.803	0.834	**0.889**	0.919
UNet [31]	13.4M	10.3	0.652	0.631	0.651	0.655	0.775	0.737
ICNet [41]	28.3M	43.1	0.828	0.824	0.823	**0.886**	0.885	**0.945**
SegNet [6]	29.4M	12.3	0.687	0.668	0.645	0.670	0.789	0.787
DeepLabV3+* [13]	59.3M	42.2	0.729	0.777	0.777	0.821	0.866	0.906
RefineNet* [25]	113.9M	50.5	0.825	0.847	0.836	0.874	0.884	0.903
DRU-Resnet* [39]	145.5M	81.0	0.001	0.275	0.306	0.325	0.550	0.525

Table 2: Evaluation of state-of-the-art models on Ego2Hands. Only mIoU is reported for experiments without the energy output channel (denoted as "energy ✗"). Models with * are trained using pretrained Resnet encoders.

binary-label hand segmentation.

5.1. Training Details

We divide the training process into the pretraining and the adaptation phase. In the pretraining phase, all models are trained for 100k iterations with a batch size of 4 and an initial learning rate of 1.0×10^{-4} decreased with a ratio of 0.5 every 20k iterations. In the adaptation phase, we train the pretrained (w/ input edge map & output energy channel) models for 10k iterations using an initial learning rate of 1.0×10^{-5} decaying with the same ratio every 5k iterations. We use the Adam optimizer and find general convergence from all models using this training setup. To ensure accurate estimation, averaged results from 3 trained model instances are reported for every architecture. GeForce RTX 2080 Ti is used as the GPU in our experiments.

As illumination (not skin tone) is the dominant factor for the brightness of the hands in input images and is known for specific environments, we perform brightness augmentation within ranges specific to the scenes in the adaptation phase. The mean brightness value β of the composited hands is scaled to be in the range [0, 55], [55,200], [55, 255] for scenes with dark (seq5), normal (seq1, 3, 4, 6, 7) and bright (seq2, 8) illumination respectively. Bright scenes have a wider brightness range due to the possibility of shadow. β for background images is jittered by ± 50.

5.2. Quantitative Analysis

Table. 2 provides detailed quantitative results for the selected models on different settings. We point out that a comprehensive comparison involves various factors including model size, inference speed, segmentation/detection accuracy, the ability to generalize and adapt.

Firstly, we justify our design choice of the additional input edge map and energy output. Models with an input edge map show overall improvement with the exceptions of UNet, ICNet and SegNet. UNet and SegNet underperform in general on the target dataset with lower generalization and adaptation accuracy. The addition of the energy output shows minimum impact on the segmentation accuracy while providing hand detection output information essential for many applications.

With a small number of parameters, UNet$_{1/8}$ and RecUNet achieve good segmentation accuracy. However, they have worse adaptation accuracy and detection accuracy. Additionally, the recursion on internal network states improves accuracy while notably increasing the inference time, making RecUNet and DRU-Resnet the slowest models in our analysis. It is worth mentioning that inference speed for segmentation and detection is crucial in real-world applications as there are oftentimes subsequent pose estimation/gesture recognition modules. Heavy models (DeepLabV3+, RefineNet) generally achieve high pretrained accuracy as well as adapted accuracy in both segmentation and detection. We find that heavy models are dependent on pretrained encoders for optimal performance. Interestingly, DRU-Resnet with the largest model size has the lowest test accuracy despite high training accuracy.

We argue that it is advantageous for an architecture to be well-balanced with high generalization/adaptation accuracy, compact model size and fast inference speed for practical applications. Existing models struggle to satisfy all the aforementioned qualities. Though not the major contribution of this paper, CSM seeks to fill in the gap and achieves a high adaptation accuracy of mIoU = 0.889 in segmentation and AP$_{0.5}$ = 0.913 in detection with only 10.0M parameters. At the same time, CSM-stage1 achieves an inference time of 0.0254s with slightly lower accuracy compared to stage2.

Our experiments on Ego2Hands provide valuable insights on the trade-off between architectures. Although

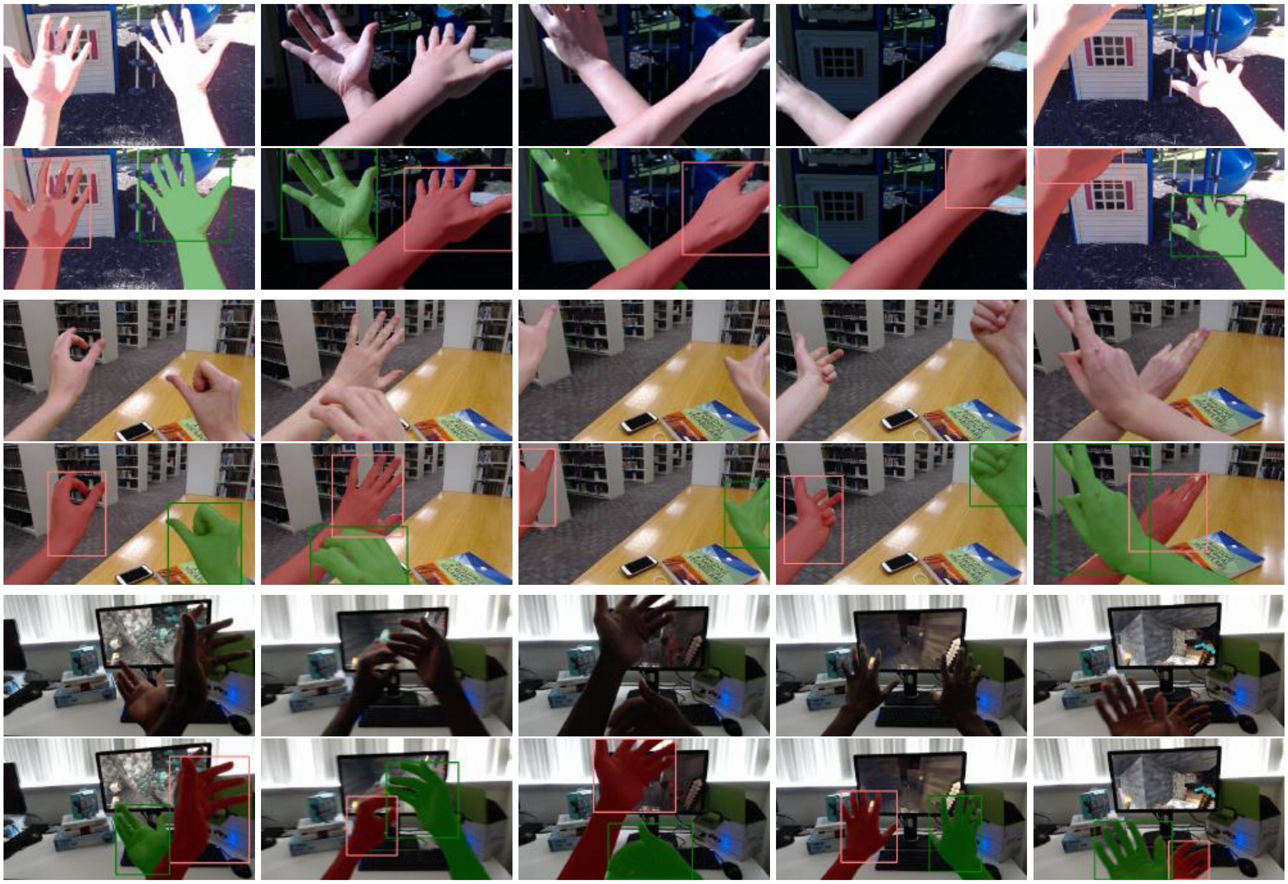

Figure 6: Qualitative results obtained using scene-adapted CSM-stage2. Odd rows show sample images in evaluation sequences with various skin tones and illumination. Even rows show the output visualization. **Best viewed in magnification.**

the proposed CSM has a good overall balance of the desired qualities, in applications where scene-specific adaptation is unrealistic and generalization accuracy is preferred (such as egocentric sign language recognition in unconfined environments), ICNet produces promising pretrained accuracy with relatively compact model size and fast inference speed. ICNet can also be favored in applications that focus more on hand detection. In cases where the model size is not the limiting factor, RefineNet achieves the highest overall pretrained segmentation/detection accuracy. In memory-constrained settings, RecUNet produces high accuracy by sacrificing inference speed. On the other hand, shallow models such as UNet and SegNet have faster inference speed with lower accuracy.

We reemphasize that the evaluation sequences cover various ranges of illumination and include hands (various skin tones) and scenes not present in the training set of Ego2Hands. Our quantitative results show that the proposed dataset and compositing-based training method enables models to generalize to the real-world image domain. To provide a proper perspective for our significantly increased level of generalization, as [12] very recently tried to address the problem of domain adaptation in a specific unseen environment for binary-label hand segmentation, we enable models to achieve high accuracy on two-hand segmentation and detection in a domain-invariant setting with the option to further improve using scene-specific adaptation. Note that our data collection process and data generation approach can also be applied to other vision-based tasks such as object segmentation/detection as long as the training instances can be composited with sufficient realism. We provide qualitative results in Fig. 6.

6. Conclusion

In this work, we introduce a color-based hand dataset and the corresponding training technique that helps deep convolutional networks to achieve domain generalization on the task of two-hand segmentation/detection. Validation and analysis of our new benchmark dataset is reported on state-of-the-art models, including our new CSM model. We hope this work can open up more directions for color-based hand tracking systems in the industry.

References

[1] Narrative Clip. *http://getnarrative.com/*, 2015. 1

[2] GoPro Camera Series. *https://gopro.com/en/us/*, 2019. 1

[3] HTC Vive. *https://developer.vive.com/resources/vive-sense/sdk/vive-hand-tracking-sdk/*, 2019. 1

[4] Oculus Quest. *https://www.oculus.com/quest/*, 2019. 1

[5] A. A. Argyros and M. I. Lourakis. Real-time Tracking of Multiple Skin-colored Objects with a Possibly Moving Camera. *In ECCV*, 2004. 2

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2481–2495, 1 2017. 6, 7

[7] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions. *In ICCV*, 2015. 1, 2, 3, 4, 5

[8] 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, 2016. 2

[9] A. K. Bojja, F. Mueller, S. R. Malireddi, M. Oberweger, V. Lepetit, C. Theobalt, K. M. Yi, and A. Tagliasacchi. HandSeg: An Automatically Labeled Dataset for Hand Segmentation from Depth Images. *arXiv preprint arXiv:1711.05944*, 2018. 2

[10] 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, 2015. 3

[11] M. Cai, K. Kitani, and Y. Sato. An Ego-vision System for Hand Grasp Analysis. *IEEE Transactions on Human-Machine Systems*, 47:524–535, 8 2017. 3

[12] M. Cai, F. Lu, and Y. Sato. Generalizing Hand Segmentation in Egocentric Videos with Uncertainty-Guided Model Adaptation. *In CVPR*, 2020. 1, 2, 3, 4, 5, 8

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *In ECCV*, 2018. 6, 7

[14] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015. 4

[15] A. Fathi, X. Ren, and J. M. Rehg. Learning to Recognize Objects in Egocentric Activities. *In CVPR*, 2011. 2

[16] P. Kakumanu, S. Makrogiannis, and N. Bourbakis. Real-time Tracking of Multiple Skin-colored Objects with a Possibly Moving Camera. *Pattern Recognition*, 40:1106–1122, 2007. 2

[17] A. U. Khan and A. Borji. Analysis of Hand Segmentation in the Wild. *In CVPR*, 2018. 1, 2, 3, 4, 5, 6

[18] C. Li and K. M. Kitani. Model Recommendation with Virtual Probes for Egocentric Hand Detection. *In ICCV*, 2013. 2

[19] C. Li and K. M. Kitani. Pixel-level Hand Detection in Ego-Centric Videos. *In CVPR*, 2013. 1, 2, 3, 4, 5

[20] Y. Li, M. Liu, and J. M. Rehg. In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video. *In ECCV*, 2018. 1, 3, 4, 5

[21] Y. Li, Z. Ye, and J. M. Rehg. Delving into Egocentric Actions. *In CVPR*, 2015. 1, 3, 4, 5

[22] G. M. Lim, P. Jatesiktat, C. W. K. Kuah, and W. T. Ang. Hand and Object Segmentation from Depth Image using Fully Convolutional Network. *In EMBC*, 2019. 2

[23] F. Lin, C. Wilhelm, and T. Martinez. Two-hand Global 3D Pose Estimation Using Monocular RGB. *arXiv preprint arXiv:2006.01320*, 2020. 1, 2, 3, 4, 5

[24] F. Lin, C. Yao, and T. Martinez. Flow Adaptive Video Object Segmentation. *Image and Vision Computing*, 2020. 4

[25] G. Lin, A. Milan, C. Shen, and I. Reid. RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. *In CVPR*, 2017. 2, 6, 7

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. *In ECCV*, 2014. 4

[27] F. Mueller, M. Davis, F. Bernard, O. Sotnychenko, M. Verschoor, M. A. Otaduy, D. Casas, and C. Theobalt. Real-time Pose and Shape Reconstruction of Two Interacting Hands With a Single Depth Camera. *ACM Transactions on Graphics (TOG)*, 2019. 1, 2

[28] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. *In CVPR*, 2016. 3

[29] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv:1704.00675*, 2017. 3

[30] X. Ren and C. Gu. Figure-Ground Segmentation Improves Handled Object Recognition in Egocentric Video. *In CVPR*, 2010. 2

[31] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *In MICCAI*, 2015. 3, 6, 7

[32] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Cucchiara. Hand Segmentation for Gesture Recognition in EGO-Vision. *In IMMPD*, 2013. 2

[33] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, and S. Izadi. Accurate, Robust, and Flexible Real-time Hand Tracking. *In CHI*, 2015. 2

[34] Y. Sheikh, O. Javed, and T. Kanade. Background Subtraction for Freely Moving Cameras. *In ICCV*, 2009. 2

[35] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and Robust Hand Tracking Using Detection-Guided Optimization. *In CVPR*, 2015. 2

[36] J. Taylor, V. Tankovich, D. Tang, C. Keskin, D. Kim, P. Davidson, A. Kowdle, and S. Izadi. Articulated Distance Fields for Ultra-Fast Tracking of Hands Interacting. *In SIGGRAPH Asia*, 2017. 1, 2

[37] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks. *ACM Transactions on Graphics (TOG)*, 2014. 2

[38] J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. A. Otaduy, D. Casas, and C. Theobalt.

RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video. *ACM Transactions on Graphics (TOG)*, 39(6), 12 2020. 3

[39] W. Wang, K. Yu, J. Hugonot, P. Fua, and M. Salzmann. Recurrent U-Net for Resource-Constrained Segmentation. *In ICCV*, 2019. 6, 7

[40] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. *In CVPR*, 2016. 5

[41] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. *In ECCV*, 2018. 6, 7

Supplementary Document:
Ego2Hands: A Dataset for Egocentric Two-hand Segmentation and Detection

1. Ego2Hands Qualitative Examples

1.1. Training set

We show sample collected images with the corresponding hand heatmap energy for the detection task. The energy map annotates the hand (w/o the arm) as the foreground and all other region as the background. Our training data covers a wide range of hand location, hand pose, hand size, skin tone and illumination.

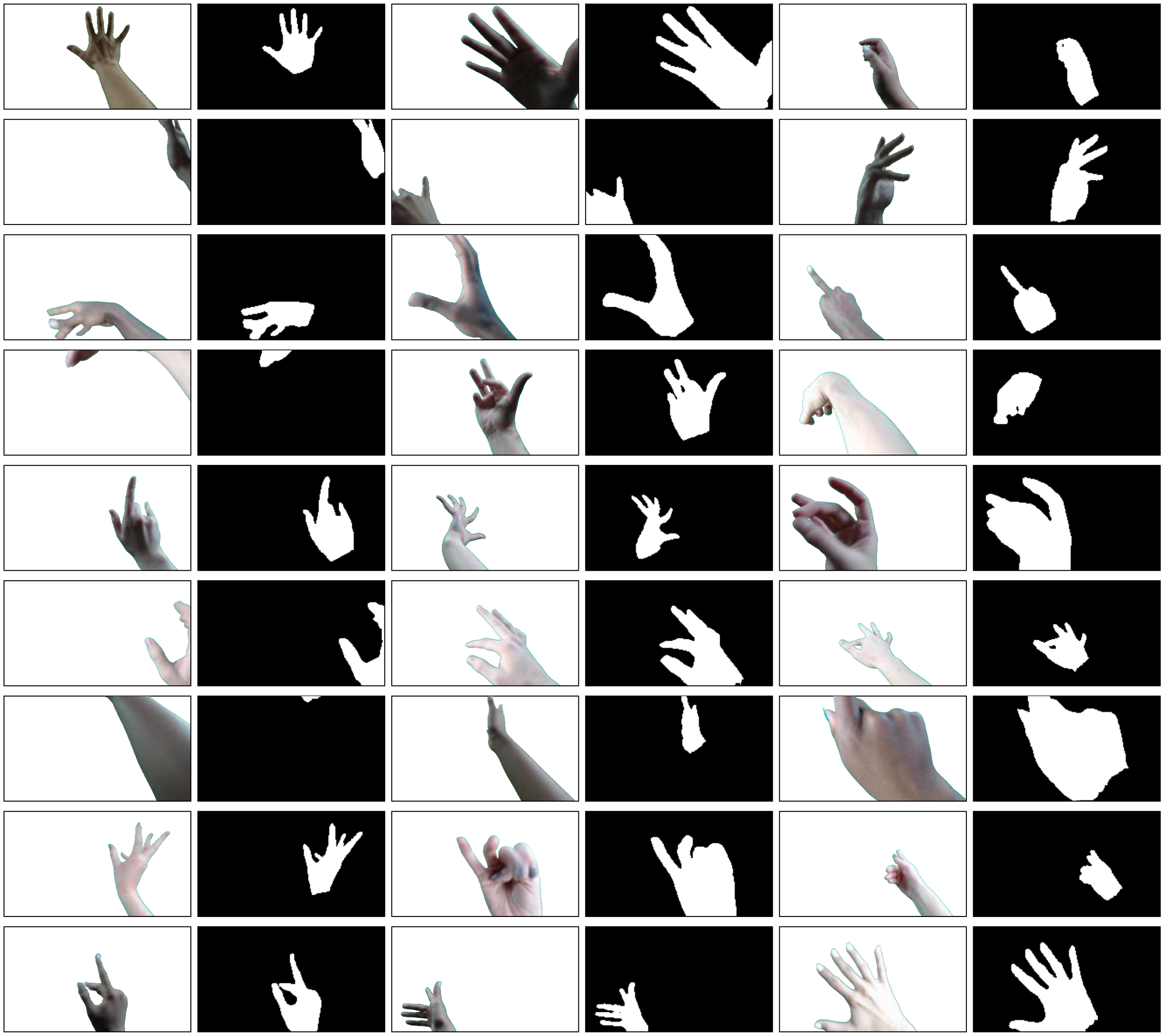

Figure 7: Sample images from the training set of Ego2Hands (odd columns). Energy for hand detection is annotated for the hand region (even columns). The rows contain annotated instances from subject 0 to 8 respectively.

1.2. Evaluation set

We show sample images from all 8 sequences below to demonstrate the diversity and annotation accuracy of our evaluation set. The evaluation sequences contain free two-hand motion with various skin tones and illumination, possible heavy occlusion and motion blur. All annotations are provided with the original image resolution of 800 × 448.

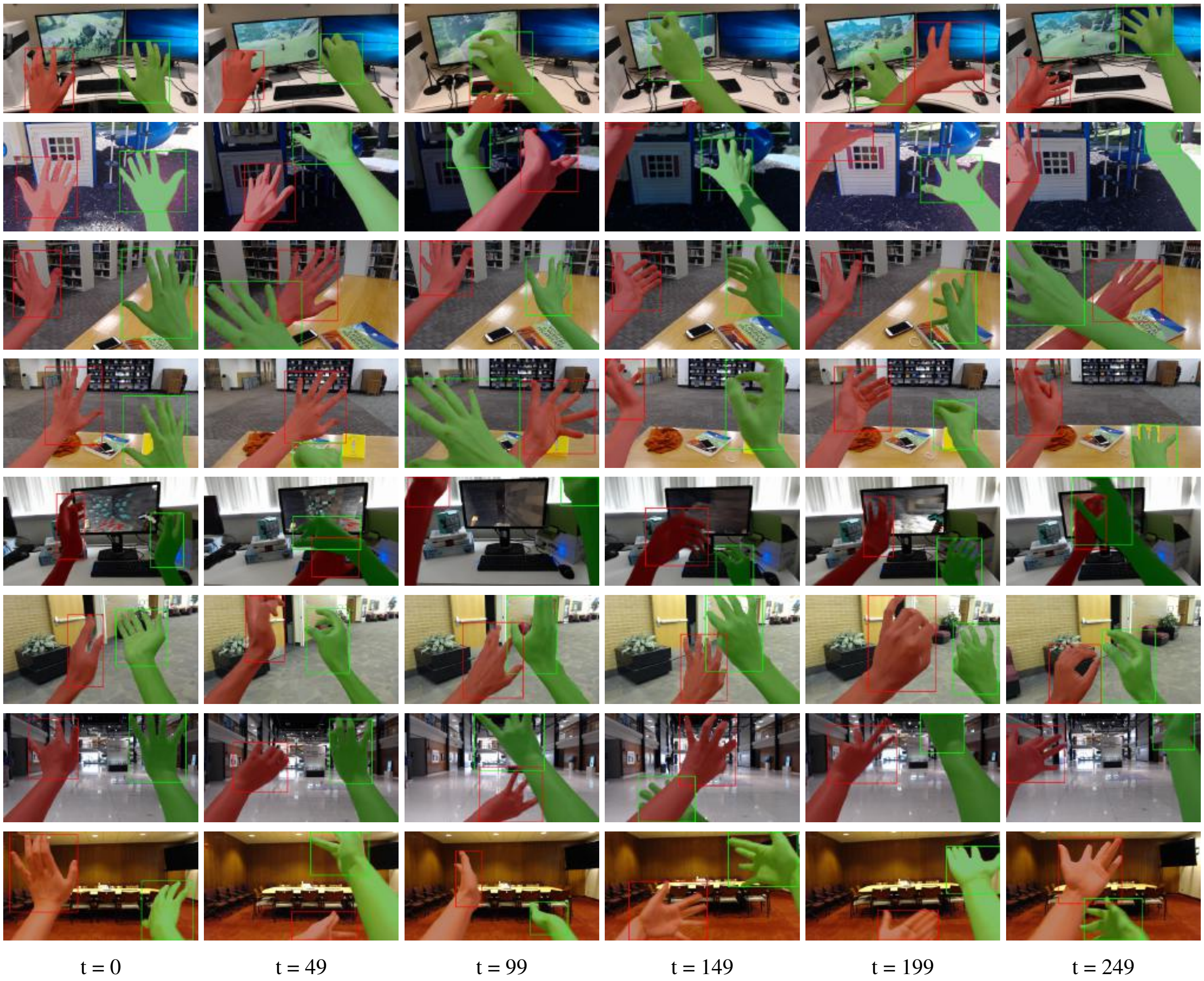

Figure 8: Sample annotated images from the 8 sequences of the evaluation set of Ego2Hands. Segmentation annotation is overlaid in colors. Detection is annotated as energy and visualized as bounding boxes. **Best viewed in magnification.**