A Pipeline for Hand 2-D Keypoint Localization Using Unpaired Image to Image Translation

Farnaz Farahanipad
The University of Texas at Arlington
Arlington, Texas, USA
farnaz.farahanipad@mavs.uta.edu

Mohammad Rezaei
The University of Texas at Arlington
Arlington, Texas, USA
mohammad.rezaei@mavs.uta.edu

Alex Dillhoff
The University of Texas at Arlington
Arlington, Texas, USA
alex.dillhoff@uta.edu

Farhad Kamangar
The University of Texas at Arlington
Arlington, Texas, USA
kamangar@cse.uta.edu

Vassilis Athitsos
The University of Texas at Arlington
Arlington, Texas, USA
athitsos@cse.uta.edu

ABSTRACT

Hand pose estimation is getting a lot of attention in many areas such as Human-Computer Interaction and Sign Language Recognition. A fundamental step to accurately estimate the hand pose involves detecting and localizing fingertips in an image. Despite the progress of 2-D hand pose estimation in recent studies, accurate and robust detection and localization of fingertips still remains a challenging task due to low resolution of a fingertip in images and varying lightning condition.

Inspired by the progress of the Generative Adversarial Network (GAN) and image-style transfer, we propose a two-stage pipeline to accurately localize the fingertip position even in varying lighting and severe self occlusion on depth images. The idea is to use a Cycle-consistent Generative Adversarial Network (Cycle-GAN) to apply unpaired image-to-image translation and generate a depth image with colored predictions on the fingertips, wrist, and palm given a real depth image. The model is trained in a semi-supervised manner using a collection of images from source and target domains that do not need to be related in anyway. Then, by applying color segmentation techniques, we localize the center of each colored area which results in finding the location of each fingertip along with center of the wrist and the palm. The proposed method achieves visually promising results on noisy depth images captured using the Microsoft Kinect. Experiments on the challenging NYU hand dataset have demonstrated that our approach not only generates plausible samples, but also outperforms state-of-the-art approaches on 2-D fingertip estimation by a significant margin even in the presence of severe self-occlusion and varying lighting conditions. Moreover, fingertips would be detected irrespective of user orientation using this method.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Hand pose estimation → Fingertip localization**; • **Generative Adversarial Networks → Domain transfer**; • **Computing methodologies → Image segmentation**.

KEYWORDS

2-D hand pose estimation, fingertip detection and localization, generative adversarial networks, human-computer interaction, domain transfer

ACM Reference Format:
Farnaz Farahanipad, Mohammad Rezaei, Alex Dillhoff, Farhad Kamangar, and Vassilis Athitsos. 2021. A Pipeline for Hand 2-D Keypoint Localization Using Unpaired Image to Image Translation. In *The 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), June 29-July 2, 2021, Corfu, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3453892.3453904

1 INTRODUCTION

Accurate fingertip localization from depth images plays an essential role in many computer vision applications such as sign language recognition [2] and human-computer interaction [10] when image-based models are used. Many proposed approaches such as [42] and [23] involve a two-stage architecture, i.e. first performing 2-D hand pose estimation and then lifting the estimated pose from 2-D to 3-D, which makes 2-D hand pose estimation itself still an important task.

In recent studies, deep learning methods have dominated state-of-the-art semantic keypoint detection methods. Mask RCNN [12] and PifPaf [17] are two representative methods for detecting semantic key-points using supervised learning. However, supervised training of a keypoint detection network requires extensive and expensive annotated data. To eliminate the need of human annotation, Shotton et al. [27] and Wetzler et al. [39] use markers, which, in some cases, are not visible in the sample due to self-occlusion and varying articulations.

Challenges in obtaining keypoint annotations have led to the rise in self/semi-supervised landmark localization research. Self-supervised learning is a re-emerging topic as of early 2020 which does not require expensive and task-specific human annotation.

PETRA 2021, June 29-July 2, 2021, Corfu, Greece

ACM ISBN 978-1-4503-8792-7/21/06…$15.00
https://doi.org/10.1145/3453892.3453904

Although unsupervised detection of landmarks can extract useful features, it is not able to detect perceptible landmarks with out supervision [14, 35]. In [8], to learn without explicit annotations, Dong et al. build on the pseudo-labeling technique which uses a teacher model and two students to generate more accurate pseudo-labels for unlabeled data. In another study by Jakab et al. [14], additional class attributes were utilized for semi-supervised keypoint detection.

In this paper, we investigate the problem of 2-D fingertip localization from depth images using a semi-supervised approach. The idea can be more broadly described as unpaired image-to-image translation techniques which involves transforming an image from domain A (real depth image) to domain B (annotated depth image) in the absence of paired data(Figure 1). This is especially useful when obtaining paired data can be expensive or, in some cases, impossible. To the best of our knowledge, this is the first study using unpaired image to image translation for 2-D keypoint detection and localization. Using a Cycle-Consistent Adversarial Network [41], we map the features learned from training samples to salient features of the real data set. Once the model is trained and a mapping between two domains is established, it is able to translate the real depth image input to the target domain, which is the input depth image along with colored markers on fingertips and two more keypoints (center of the palm and the wrist), as shown in Figure 2. Afterwards, using Hue, Saturation, and Value features of the output markers, we apply color segmentation techniques to localize and extract the 2-D coordinates of the center of each colored area for the fingertips, center of the palm, and the wrist.

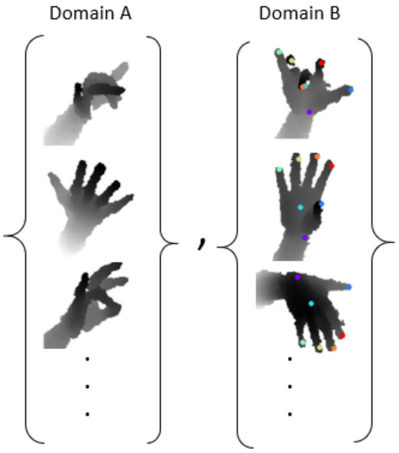


Figure 1: Unpaired training data, consisting of a source set and a target set, with no information provided as to which x_i in domain A matches which y_j in domain B.

The quantitative and qualitative results on the NYU hand dataset show that our proposed approach outperforms state of the arts methods and can handle severe occlusion and varying light conditions independent of user hand orientation. It is worth mentioning that our proposed method outperforms these state of the art methods even when significantly reducing the size of the training set. Further, unlike [22] and [15], which use multi-view camera to tackle severe self-occlusion, we only use a single depth sensor. Using this approach, we are able to generate corresponding pairs of images which can be used for unsupervised domain adaptation and, since we applied unpaired image to image translation, we eliminate the burden of requiring a large annotated dataset. Finally, since we segment the image in the HSV domain, our proposed method is more robust to varying lighting conditions.

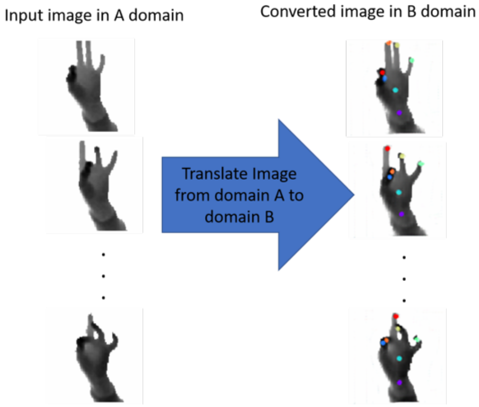

Input image in A domain
Converted image in B domain
Translate Image from domain A to domain B

Figure 2: Examples of translated input image from A domain to B domain

This paper is structured as follows; Section2 describes the summary of background and related work in 2-D hand pose estimation; section3 presents the proposed method; section 4 explained experimental details followed by results and discussion in section 5. Finally, section 6 discusses the conclusion and future work to be done in this system respectively.

2 BACKGROUND AND RELATED WORK

2.1 Hand pose estimation

In many hand pose estimation studies, such as [25] and [16], 2-D fingertip localization is an initial step for 3D hand pose estimation. However, fingertip detection is a challenging task due to self occlusion and high rotational variability. Fortunately, due to the progress of optical technologies, such as depth cameras, it is possible to capture more accurate information of our 3-D world. Several studies have been introduced which use depth images to estimate the hand poses [20, 26, 40]. Malassiotis and Strintzis extract PCA features from depth images of synthetic 3D hand models for training [19]. Suryanarayan et al. [32] use depth information to recognize scale and rotation invariant poses dynamically. Sinha et al. [29] used a regression-based model to find the 21 joints in the hand. They trained a separate network for each finger which regress three joint keypoints on each finger. To minimize the dependency on large hand pose datasets and to improve the generalization ability to unseen situations, data-efficient methods such as weakly supervised learning or hybrid methods are needed. By fast progress of Generative Adversarial Networks (GAN), several studies have been

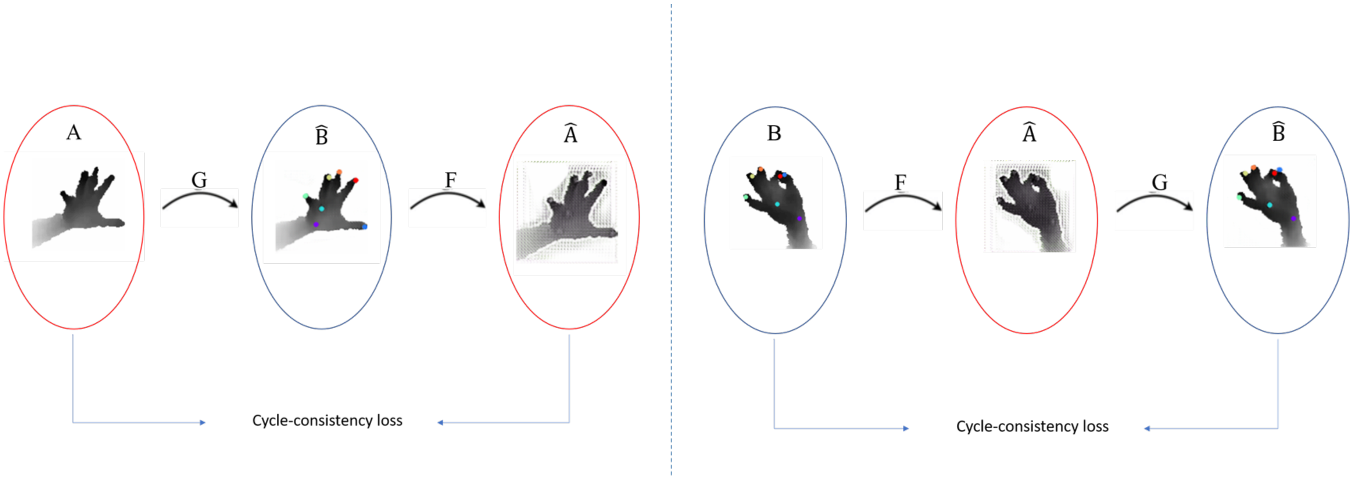


Figure 3: The simplified architecture of first stage to perform unpaired image to image translation using Cycle-Consistent Adversarial Networks

performed to model the statistical relationship of the 3D pose space and corresponding space of the input data in semi/self-supervised manner [1, 5, 38]. Chen et al. [6] proposed a conditional Generative Adversarial Network (GAN) model called Depth-image Guided GAN (DGGAN) to generate realistic depth maps conditioned on the input RGB image and use the synthesized depth image to refine the 3D hand pose estimation. In [13], He at al. proposed a data-driven method to generate deep hand images closer to real ones during training. In Chen et al. [7], they propose tonality-alignment generative adversarial networks (TGAN) to align the tonality and color distribution between synthetic hand poses and real backgrounds.

2.2 Unpaired image to image translation

Despite the easy generation and annotation of synthesized dataset, they lack the generalization power and they will not perform well on real-world hand images. To eliminate the domain gap between synthesized data and real dataset, in [24], they used conditional GAN called GeoConGAN to transfer the generated images to real images. Image to image translation is a concept from machine translation where a phrase translated from English to French should translate from French back to English and be identical to the original phrase. The reverse process should also be true [37]. However, traditionally, paired image to image translation requires a dataset of paired examples which is challenging and expensive to prepare. As such, there is a huge interest in unpaired image to image translation approaches. Unpaired image to image translation uses extra terms along with adversarial networks to enforce the output to be close to input in a specified way, such as labels space, image pixels space or image features space. In recent studies,[3] and [18], authors use a weight sharing strategy to learn the most common representation between domains. In [28] and [33], to perform unpaired image to image translation, the proposed models share the specific "content" features between the two domains even though they may differ in "style". Baek et al. used a CyclicGAN to transfer the depth map of the hand to the 3D representation of the hand joints[4]. In [21] authors, proposed a strategy that exploits the unpaired image style transfer capabilities of CycleGAN in semi-supervised semantic segmentation. Spurr et al. also applied similar approach to make one to one relation between RGB images to 3D hand joints pose[30].

3 PROPOSED METHOD

In this study we propose a two-stage pipeline for fingertip localization in 2-D plane; first we reduce the problem to an unpaired image to image translation using Cycle-consistent Generative Adversarial Network [41]. It is a general-purpose network for unpaired image to image translation and does not require paired image and uses the concept of cycle consistency to enforce the model to map between domain A and domain B and vice versa with the inverse mapping (see Figure 3). The key idea behind CycleGAN is that it allows the model to use two unpaired collection of the images rather than two specific images. A detailed structure is explained in section 4.2. Applying unpaired image to image translation, the model is able to translate the input real depth image to depth map with colored marks corresponds to fingertip locations. Using these colored mark, we extract the location of the fingertips along with two other points (center of the palm and wrist) using color segmentation techniques in HSV color space. An overview of the proposed pipeline, detailed architecture of the unpaired image to image translation for first stage and detailed overview of second stage are demonstrated in Figure 4, 3 and Figure 5 respectively.

3.1 Formulation

We aim to learn the mapping between real depth images and depth images with colored marks corresponding to fingertip locations without paired example. This can be done using general adversarial loss however, the model ignores the input image completely and keeps generating the same depth image from the domain B. To ensure that the model considers the input image, Cycle-GAN, uses two objectives: adversarial loss and cycle-consistency loss.

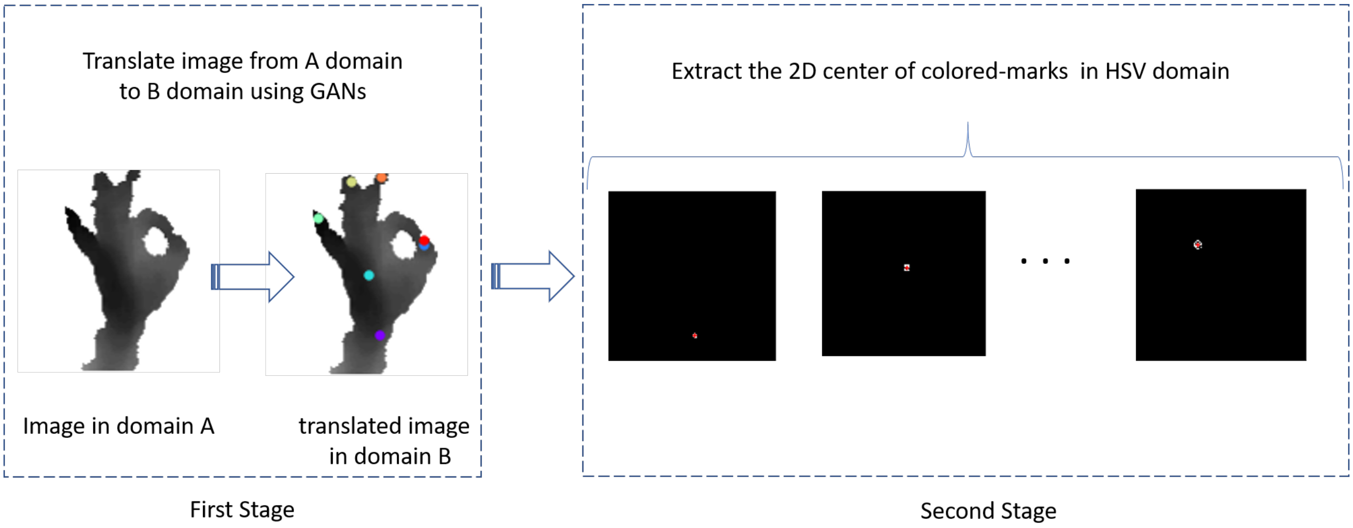


Figure 4: The overview of proposed model

3.1.1 Adversarial loss . Adversarial loss[11] is a powerful loss specifically for image generation task. Adversarial loss, in GAN, enforce the generated image to be indistinguishable from real photos. Since the model has two mapping functions G and F, an adversarial loss is defined for each mapping function as:

$$\mathcal{L}_{GAN}(G, D_B, A, B) = \mathbb{E}_{b\sim pdata(b)}[logD_B(b)] + \mathbb{E}_{a\sim pdata(a)}[log(1 - D_B(G(a))], \tag{1}$$

where G tries to generate images $G(a)$ that look similar to images from domain B , while D_B aims to distinguish between translated samples $G(a)$ and real samples b. Similarly for mapping function F, it is defined as:

$$\mathcal{L}_{GAN}(F, D_A, B, A) = \mathbb{E}_{a\sim pdata(a)}[logD_A(a)] + \mathbb{E}_{b\sim pdata(b)}[log(1 - D_A(F(b))], \tag{2}$$

3.1.2 Cycle consistency loss . Although adversarial loss can enforce the model to learn the mapping G and F and produce outputs identically distributed as target domain, however, the network might map the same set of input image to any random permutation of image in target domain. Therefore, Zhu et al. use cycle consistency loss for generative adversarial networks to perform unpaired image to image translation[41]. Given an input image from domain A, they apply mapping G to translate image to domain B followed by inverse mapping F to reconstruct the input image in domain A. Cycle-consistency loss compares the reconstructed image and input image using L1-norm distance and it can be written as[41]:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{a\sim p_{data}(a)}[||F(G(a)) - a||_1] + \mathbb{E}_{b\sim p_{data}(b)}[||G(F(b)) - b||_1] \tag{3}$$

The same process is done in opposite direction as shown in Figure 3.

3.1.3 Full Objective. The final loss function for training Cycle-GAN is defined as [41] :

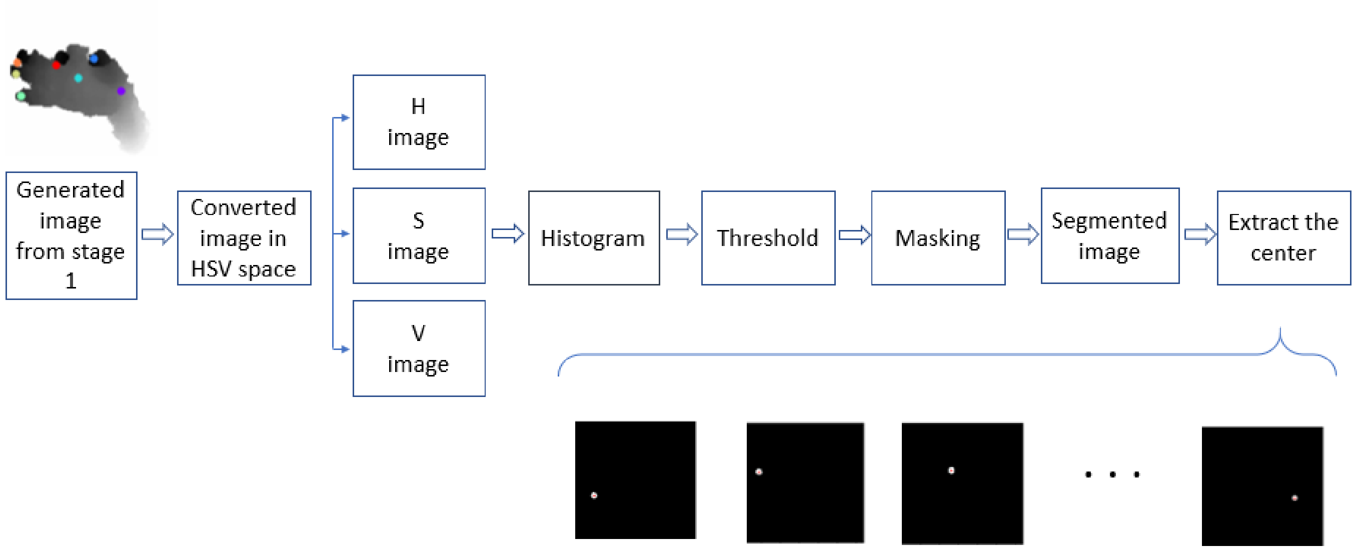
$$\mathcal{L}(G, F, D_A, D_B) = \mathcal{L}_{GAN}(G, D_B, A, B) + \mathcal{L}_{GAN}(F, D_A, B, A) + \lambda\mathcal{L}_{cyc}(G, F) \tag{4}$$

where λ controls the relative importance of the two objectives. The Cycle-GAN model is trained by minimizing the following loss:

$$G^*, F^* = arg\ min_{G,F}\ max_{D_a,D_b}\ \mathcal{L}(G, F, D_A, D_B) \tag{5}$$

3.2 Color segmentation in HSV color space

HSV is a cylindrical color model that remaps the RGB primary colors into dimensions that are easier for humans to understand. These dimensions are hue, saturation and value as shown in Figure 6. Hue represents an angle in range $[0,2\pi]$ relative to the Red axis with red at angle 0, green at $2\pi/3$, blue at $4\pi/3$ and red again at 2π. Saturation defines the depth or purity of the color and is measured as a radial distance from the central axis with value between 0 at the center to 1 at the outer surface [31]. Finally, the value of Intensity determines the particular gray shade to which this transformation converges. It is seen that, HSV based approximation can determine the intensity and shape variations near the edges of an object which result in sharpening the boundaries and retraining the color information of each pixel. Furthermore, the approximation done by the RGB features blurs the distinction between two visually separable colors by changing the brightness. In the second stage, we develop a new framework , in HSV color space, to extract region of interest from the generated colored annotated depth image from the previous stage. First, the images are converted to HSV color space to have all components quantities with same precision. Afterwards, the converted images are split into three different sub images as hue, saturation and value. Histogram for all three components is computed and plotted and the threshold value for each



,

Figure 5: Second stage overview

component is selected accordingly. Finally by masking operation a desired colored area is segmented and the center of the segmented part extracted as 2-D coordinates of the desired points (Figure 5).

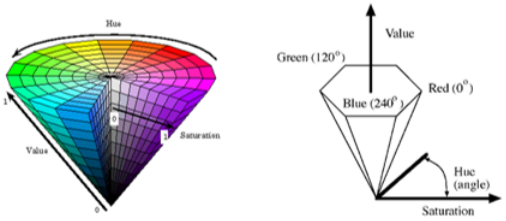


Figure 6: HSV color space representation

4 EXPERIMENTAL DETAILS

4.1 Data preparation

Although there are some datasets like ICVL [34] and MSRA14[25] for hand pose estimation, we chose New York University (NYU) dataset [36]. NYU is a challenging hand pose dataset and it is more commonly used in recent studies due to its accurate annotation and variety of poses. It contains RGBD dataset captured from 3 views and it has 72,757 frames from a single user in train set and 8,252 frames from two different user in test set. It uses 36-joints model to annotate the hand images.

To prepare training data from NYU hand dataset for Cycle-consistency model, we prepare two sets of data: train data for domain A which includes 3000 cropped real depth images of hand and train data for domain B, which contains 3000 cropped images around hand with color markers on 7 points (5 fingertips and center of the wrist and center of the palm). To simulate unpaired supervision, these two set of data do not have one to one mapping and are selected randomly from the view-point 1(front view).

For test data, we randomly chose 300 real depth images of the same view from test set of NYU hand dataset.

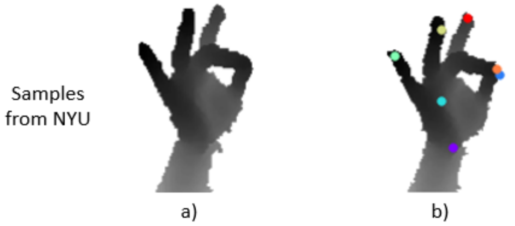


Figure 7: Example of test data; real depth image (a) and annotated depth (b) sample from NYU hand dataset

All the images are of size 128 x 128 and they only contained cropped image of hand. There are 7 keypoints, which are annotated using 7 different predefined colors and corresponds to pinky fingertip, ring fingertip, middle fingertip, index fingertip, thumb fingertip, center of the palm and center of the wrist. Figure 7 shows examples of customized NYU hand dataset.

4.2 Model Architecture and Internal Parameters

The general architecture of CycleGAN [41] utilizes two parts Generators and Discriminators. Each generator has three parts; encoder,transformer and decoder. The encoder consists of 3 convolutional layers that reduces the representation by 1/4-th of actual image size. The transformer contains 6 or 9 residual blocks based on the size of input and the decoder uses 2 deconvolution block with fractional strides to increase the size of representation to the original size. The network uses instance normalization as opposed to batch normalization, and the discriminator is a 70x70 Patch GAN which penalizes images at the level of individual patches as opposed to per-pixel or per-image basis. We trained the model for 200 epochs for customized NYU with 3000 unpaired data with learning rate of 0.0002 and lambda value of 10 to calculate cycle loss.Once the model is trained, we evaluate it using 300 test images from NYU dataset to translate them from domain A to domain B which in turns are generated depth map along with colored markers. In the second stage, we use the HSV color space with emphasis on the variation in Hue and Saturation. Segmentation using this method shows better identification of fingertip localization in an image. The center of these segmented area are extracted as fingertip positions in 2-D as explained in section 3.2.

4.3 Evaluation metrics

The two most common metric utilized to quantitatively evaluate the localization method are Mean Error (ME) in pixel and Percentage of Correct Keypoints (PCK). ME is the average 2-D Euclidean distance between predicted and ground-truth joints and PCK measures the mean percentage of predicted joint locations that fall within certain error thresholds compared to correct location. To have a fair comparison we evaluate our proposed pipeline on NYU hand dataset with these two metrics.

5 RESULTS AND DISCUSSION

Since, most of previous methods , [39] and [9], on 2-D hand pose estimation, have primarily reported results on NYU hand dataset, we evaluate our method on NYU hand dataset. It is worth mentioning that we only trained our model over 0.03 of NYU dataset while previous methods are trained over the entire dataset. Unlike the previous methods where they use paired example for training, our pipeline uses unpaired supervision and receives no information about which labeled image corresponds to which image. Both qualitative and quantitative results indicate that our propose methods produce fewer pixel errors in each frame.

5.1 Quantitative results

We employ two metrics to evaluate the performance of our proposed method; the average Euclidean distance in pixels between the results and ground truth and the percentage of frames in which all joints error are within a certain threshold. However, since there is no result reported directly on the same joints as our study, to have a fair comparison, we extract the result for 5 fingertips from the reported results on right hand (Figure 9 in Duan's paper [9]) and summarized the 2D localization results for 5 fingertips in Figure 8 and Table 1.

Table 1: Quantitative evaluation on NYU Hands(Fingertips only)

Methods	Mean error (Pixels)
Ours	7.2
Duan paper[9]	12.2
Mask RCNN(kpt and mask)[9]	13.6
Mask RCNN(kpt)[9]	24.5

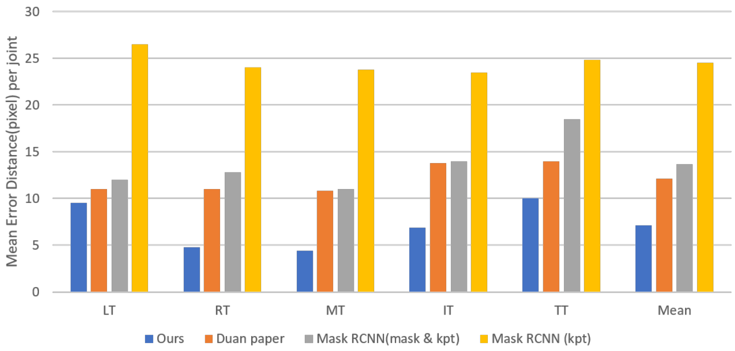


Figure 8: Comparison on per-joints mean error distance in pixels on NYU hand dataset

As shown in Table 1, the mean joint pixel error on subset of 300 images of NYU test data, is 7.2 which is better than reported average results on fingertips of right hand by Duan et al. in [9]. Moreover, the comparison of our methods with extracted results from [9], on each joint for right hand in the NYU hand dataset is shown in Figure 8.

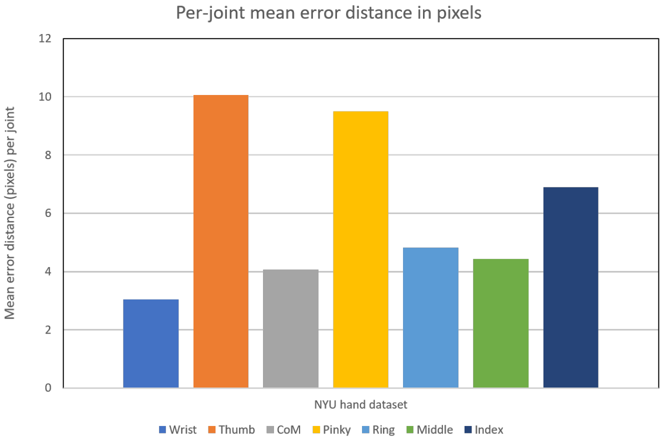


Figure 9: Per-joint mean error distance in pixels on NYU test dataset

Moreover, Figure 9 illustrates the mean joint pixel error for 7 keypoints on subset of NYU test data with our proposed pipeline. The Percentage of Correct Keypoint over a different threshold is shown in Figure 10.

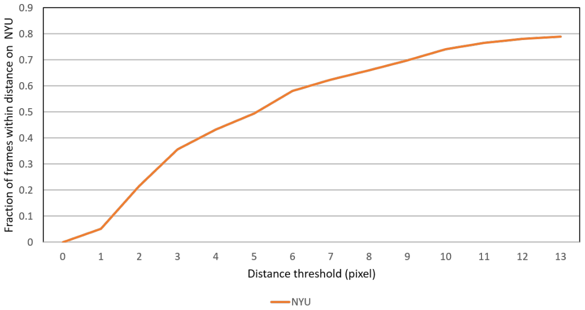


Figure 10: Fraction of frames within distance on NYU test datasets

5.2 Qualitative results

As can be seen in Figure 11, our proposed approach can improve the localization of fingertip positions and provide a more accurate estimation on NYU hand dataset, by better recovery of details, and generating more natural images by unpaired image to image translation independent of the hand orientation and in presence of severe self occlusion.

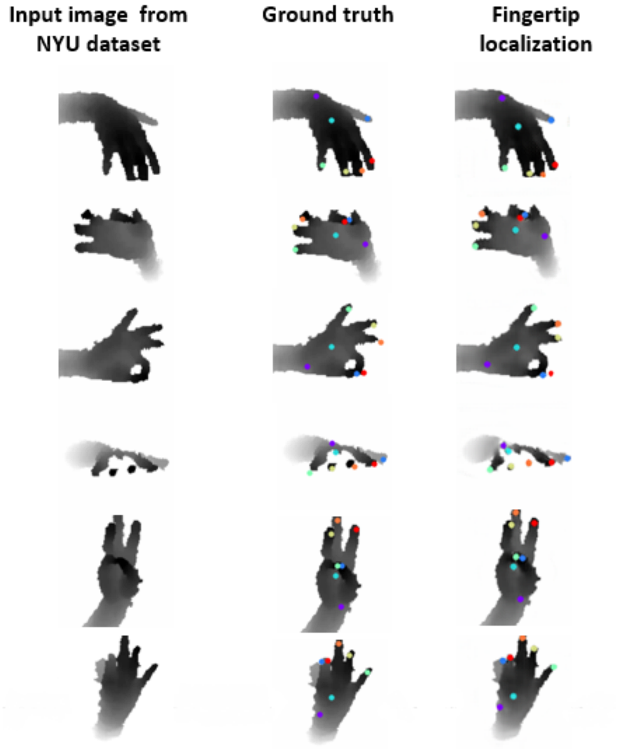


Figure 11: Qualitative results on examples of test data from NYU hand dataset; first column real depth image, second column ground truth locations and third column represents the translated image using Cycle-constituency approach

6 CONCLUSION AND FUTURE WORK

Since many 3D hand pose estimation methods perform a two-stage approach to obtain 3D joint locations based on 2-D positions of fingertip locations, obtaining accurate 2-D location of joints and fingertip has a great importance. Despite the advantage of using low cost depth-cameras, localizing the fingertip position accurately is a difficult and challenging task since, after depth-segmentation, hand contours are prone to erosion. Furthermore, self occlusion and varying lighting conditions are another challenging issues.

To tackle these issues, we implemented a pipeline for 2-D localization by reducing the problem to an unpaired image to image translation task followed by color segmentation in HSV domain and histogram threshold, to extract the fingertip positions. Evaluation of our pipeline with subset of NYU test detests, shows that our method can be used to localized 2-D fingertip positions which are also competitive to state of the arts even at presence of severe self occlusion and performs well independent of hand rotations.

The model was not completely successful to predict the fingertips in cases where part of fingers are out of the cropped ROI. Therefore, in the future, we plan to improve the performance of our model by having more accurate hand segmentation in prepossessing step, to accurately define the ROI around the segmented hand. More importantly, our system could be extended to be used in 3D hand pose estimation in our next study. In this study we have considered results only on depth images but we plan to apply a similar pipeline to RGB images.

REFERENCES

[1] Masoud Abdi, Ehsan Abbasnejad, Chee Peng Lim, and Saeid Nahavandi. 2018. 3d hand pose estimation using simulation and partial-supervision with a shared latent space. *arXiv preprint arXiv:1807.05380* (2018).

[2] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Quan Yuan, and A. Thangali. 2008. The American Sign Language Lexicon Video Dataset. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 1–8. https://doi.org/10.1109/CVPRW.2008.4563181

[3] Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2017. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence* 40, 10 (2017), 2303–2314.

[4] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. 2018. Augmented Skeleton Space Transfer for Depth-Based Hand Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[5] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. 2020. Weakly-Supervised Domain Adaptation via GAN and Mesh Model for Estimating 3D Hand Poses Interacting Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6121–6131.

[6] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, Wei Fan, and Xiaohui Xie. 2020. DGGAN: Depth-image Guided Generative Adversarial Networks forDisentangling RGB and Depth Images in 3D Hand Pose Estimation. In *The IEEE Winter Conference on Applications of Computer Vision*. 411–419.

[7] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Hui Tang, Yufan Xue, Xiaohui Xie, Yen-Yu Lin, and Wei Fan. 2018. Generating Realistic Training Images Based on Tonality-Alignment Generative Adversarial Networks for Hand Pose Estimation. *arXiv preprint arXiv:1811.09916* (2018).

[8] Xuanyi Dong and Yi Yang. 2019. Teacher Supervises Students How to Learn From Partially Labeled Images for Facial Landmark Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

[9] Le Duan, Minmin Shen, Song Cui, Zhexiao Guo, and Oliver Deussen. 2018. Estimating 2d multi-hand poses from single depth images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 0–0.

[10] Farnaz Farahanipad, Harish Ram Nambiappan, Ashish Jaiswal, Maria Kyrarini, and Fillia Makedon. 2020. HAND-REHA: dynamic hand gesture recognition for game-based wrist rehabilitation. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 1–9.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[13] Wangyong He, Zhongzhao Xie, Yongbo Li, Xinmei Wang, and Wendi Cai. 2019. Synthesizing depth hand images with GANs and style transfer for hand pose estimation. *Sensors* 19, 13 (2019), 2919.

[14] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. 2018. Unsupervised learning of object landmarks through conditional image generation. In *Advances in neural information processing systems*. 4016–4027.

[15] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[16] Cem Keskin, Furkan Kıraç, Yunus Emre Kara, and Lale Akarun. 2013. Real time hand pose estimation using depth sensors. In *Consumer depth cameras for computer vision*. Springer, 119–137.

[17] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2019. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11977–11986.

[18] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. *Advances in neural information processing systems* 29 (2016), 469–477.

[19] Sotiris Malassiotis and Michael G Strintzis. 2008. Real-time hand posture recognition using range data. *Image and Vision Computing* 26, 7 (2008), 1027–1037.

[20] Jameel Malik, Ibrahim Abdelaziz, Ahmed Elhayek, Soshi Shimada, Sk Aziz Ali, Vladislav Golyanik, Christian Theobalt, and Didier Stricker. 2020. HandVoxNet: Deep Voxel-Based Network for 3D Hand Shape and Pose Estimation from a Single Depth Map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7113–7122.

[21] Arnab Kumar Mondal, Aniket Agarwal, Jose Dolz, and Christian Desrosiers. 2019. Revisiting CycleGAN for semi-supervised segmentation. *arXiv preprint arXiv:1908.11569* (2019).

[22] Paschalis Panteleris and Antonis Argyros. 2017. Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 575–584.

[23] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. 2018. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 436–445.

[24] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2536–2544.

[25] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. 2014. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1106–1113.

[26] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. 2018. Feature mapping for learning fast and accurate 3d pose inference from synthetic images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4663–4672.

[27] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*. Ieee, 1297–1304.

[28] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. 2017. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2107–2116.

[29] Ayan Sinha, Chiho Choi, and Karthik Ramani. 2016. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4150–4158.

[30] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. 2018. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 89–98.

[31] Shamik Sural, Gang Qian, and Sakti Pramanik. 2002. Segmentation and histogram generation using the HSV color space for image retrieval. In *Proceedings. International Conference on Image Processing*, Vol. 2. IEEE, II–II.

[32] Poonam Suryanarayan, Anbumani Subramanian, and Dinesh Mandalapu. 2010. Dynamic hand pose recognition using depth data. In *2010 20th International Conference on Pattern Recognition*. IEEE, 3105–3108.

[33] Yaniv Taigman, Adam Polyak, and Lior Wolf. 2016. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200* (2016).

[34] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. 2014. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3786–3793.

[35] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. 2019. Unsupervised learning of landmarks by descriptor vector exchange. In *Proceedings of the IEEE International Conference on Computer Vision*. 6361–6371.

[36] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. 2014. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)* 33, 5 (2014), 1–10.

[37] Mark Twain. 1971. *The jumping frog: in English, then in French, then clawed back into a civilized language once more by patient, unremunerated toil.* Courier Corporation.

[38] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. 2017. Crossing nets: Dual generative models with a shared latent space for hand pose estimation. In *Conference on Computer Vision and Pattern Recognition*, Vol. 7.

[39] Aaron Wetzler, Ron Slossberg, and Ron Kimmel. 2015. Rule of thumb: Deep derotation for improved fingertip detection. *arXiv preprint arXiv:1507.05726* (2015).

[40] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. 2019. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE International Conference on Computer Vision*. 793–802.

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

[42] Christian Zimmermann and Thomas Brox. 2017. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*. 4903–4911.