GAN-Based Domain Translation for Hand Pose Estimation and Face Reconstruction

by

FARNAZ FARAHANIPAD

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2022

ACKNOWLEDGEMENTS

First and foremost I would like to express my sincere gratitude to my advisor Professor Farhad Kamangar. It has been an honor to be his Ph.D. student. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. I am also thankful for the excellent example he has provided by his respectful character and personality. I would like to give a special thank you to my co-advisor, professor Vassilis Athitsos for answering my many questions throughout the years. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. Besides my advisor, I would like to thank the rest of my dissertation committee members (Dr. David Levine, and Dr. Manfred Huber) for their insightful comments and invaluable advice. I am also grateful to my colleagues at VLM lab, Mohammad Rezaei, Alex Dillhoff, Reza Ghoddoosian, Saif Sayed, and Marnim Galib for discussions I had and topics I learned from.

I would also like to extend my deepest gratitude to Dr. Fillia Mekedon and Dr. Maria Kyrarini who supported and nurtured me during my Ph.D. process. I am very fortunate to have been included in such an incredible and meaningful project.

My gratitude goes to the faculty and staff of Computer Science department for their support throughout my Ph.D. study. Special thanks to Dr. Bahram Khalili and Ms. Ginger Dickens for all their help and encouragement from the day I arrived at UTA to the last day.

Finally, my deep and sincere gratitude to my family and friend for their continuous love, help and support.

To my parents, Ashraf and Manouchehr, whose words of encouragement and push for tenacity ring in my ears.

To my husband, Armin. You have been my inspiration and my soul mate. I am incredibly grateful to have you in my life.

To my brother, Farbod, who has always been my greatest supporter throughout this process.

This journey was not bearable without their encouragement, feedback, and constant inspirations.

April 15, 2022

ABSTRACT

GAN-Based Domain Translation for Hand Pose Estimation and Face Reconstruction

FARNAZ FARAHANIPAD, Ph.D.

The University of Texas at Arlington, 2022

Supervising Professor: Farhad Kamangar

Deep learning solutions for hand pose estimation are now very reliant on comprehensive datasets covering diverse camera perspectives, lighting conditions, shapes, and pose variations. Since, acquiring such datasets is a challenging task that may be infeasible for many novel applications, several studies aim to develop semi/self supervised learning methods, that learn to estimate hand pose from a few labeled/unlabeled data. Therefore, in this dissertation, we investigate new advances in semi/self supervised learning which will remove the bottleneck of obtaining time-consuming frame-by-frame manual annotations through generative adversarial networks (GANs).

To handle above mentioned challenges, this thesis makes the following contributions. First, we present a comprehensive study on effective hand pose estimation approaches, which are comprised of the leveraged generative adversarial network (GAN), providing a comprehensive training dataset with different modalities. We also, evaluate related hand pose datasets and performance comparison of some of these methods for the hand pose estimation problem. The quantitative and qualitative results indicate that these methods are able to beat the baseline approaches

with better visual quality and higher values in most of the metrics (PCK and ME) on benchmark hand pose datasets.

The second contribution is based on the progress of the Generative Adversarial Network (GAN) and image-style transfer. We propose a two-stage semi-supervised pipeline which is able to accurately localize the fingertip position even in severe self occlusion on depth images using Cycle-consistent Generative Adversarial Network (Cycle-GAN). Due to need for huge amount of labeled data for training neural networks, semi/self-supervised learning is very appealing for CNN training. Experiments on the challenging NYU hand dataset have demonstrated that our approach outperforms state-of-the-art approaches on 2-D fingertip estimation by a significant margin even in the presence of severe self-occlusion and irrespective of user orientation.

Moreover, we develop a GUI in MATLAB R2020a, to obtain 12-joints hand pose annotations of depth images. We prepare a comprehensive dataset of 10000 depth hand images collected by Microsoft Kinect V2 along with 7 keypoints on depth hand dataset.

Third, we present a novel framework and formulate 2D hand keypoint localization in sequenced data as a problem of conditional video generation. We aim to learn a mapping function from an input depth video in the source domain to target depth video by enforcing temporal consistency constraints. To the best of our knowledge, this is the first work ever performed on fingertip localization on depth videos through domain adaptation. Our comparative experimental results with the state-of-the-art single-frame hand pose estimation on the challenging NYU dataset demonstrates that by exploiting temporal information, our model manifests better hand appearance consistency in video-to-video synthesis stage which leads to accurate estimations of 2D hand poses under motion blur by fast hand motion.

In addition, we design and develop a novel game-based system for wrist rehabilitation, called HandReha. This is a unique and novel approach because the gestures are selected from a set of human gestures suitable for wrist rehabilitation and implemented to control a game built in a 3D environment as compared to previous works where most of the games designed for rehabilitation purposes are built in a 2D environment.

Finally, we propose a general domain translation framework that can be used to reconstruct the hidden part of face concealed by mask. We have employed GAN-based unpaired domain translation technique to translate masked face images from the source to the unmasked images in the destination domain which can be used for facial identification and secure authentication in human-computer interaction. The obtained results demonstrate that our model outperforms other representative state-of-the-art face completion approaches both qualitatively and quantitatively.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

Introduction

Hand pose estimation is getting a lot of attention in many areas such as Human-Computer Interaction, Virtual Reality (VR) and Augmented Reality (AR) device, and Sign Language Recognition. A fundamental step to accurately estimate the hand pose involves detecting and localizing fingertips. 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, varying lightning condition, self-similar parts, and severe self-occlusions.

In recent years, Convolutional Neural Networks (CNN) have been demonstrating the power of learning real world knowledge from images with supervision from labels. However, fully supervised models for such tasks is challenging, due to the difficulty and extent of effort involved in obtaining large amounts of training data. Getting such data involves manually specifying hand pose information for thousands or millions of images, and this has been a big bottleneck for progress on this topic.

Semi-supervised learning provides a solution by learning the patterns present in unlabelled data, and combining that knowledge with the generally, fewer labeled training samples in order to accomplish a supervised learning tasks.

This dissertation considers several problems related to hand pose estimation and face reconstruction in semi-supervised learning based on generative adversarial networks (GANs) . Leverage Generative Adversarial Networks(GANs), this dissertation addresses the problem of 2D/3D hand pose estimation on depth image/video whose challenges are mentioned above. With semi-supervised learning, the dataset

may contain millions of images, but we only need to specify hand pose information for a very small fraction of those images.

Moreover, this dissertation proposes solution for face reconstruction using GANs. The proposed framework, employed GAN-based unpaired domain translation technique to translate masked face images from the source to the unmasked images in the destination domain. It can be used for facial identification and secure authentication in human-computer interaction during a scenario when the person wears a mask.

1.1 Contributions

This dissertation makes contributions towards Hand Pose Estimation and Face Reconstruction using Generative Adversarial Networks. This section highlights these contributions as the pertain to each topic.

1. Reviewing all existing models for hand pose estimation using Generative Adversarial Networks.
2. Enriching the field with a challenging public finger tapping dataset to address the lack of having appropriate dataset.
3. Introducing semi-supervised 2D hand keypoint localization framework on depth images.
4. Introducing semi-supervised 2D hand keypoint localization method on depth video, leveraged temporal information.
5. Proposing a face reconstruction pipeline using paired image-to-image translation for masked-face reconstruction.

1.2 Dissertation Structure

Each following chapter in this dissertation is meant to be self-contained and it includes all related works as necessary. In chapter 2, an overview of related works in hand pose estimation leveraged generative adversarial networks or GANs in short is presented.

Chapter 3, describes the Generative Adversarial Networks and image-to-image translation technique and then presents a novel cycle-consistent framework for 2D hand pose estimation on depth images which works without paired supervision. Experiments on the challenging NYU hand dataset have demonstrated that proposed approach outperforms 2-D fingertip localization state-of-the-art by a significant margin in terms of Mean Error (ME) in pixel and Percentage of Correct Keypoints (PCK), even in the presence of severe self-occlusion and varying lighting conditions and irrespective of user orientation.

In chapter 4 and inspired by work presented in 3, 2D hand keypoint localization is formulated as a problem of conditional video generation, where the goal is to learn a mapping function from an input source video to an output photo-realistic depth video by enforcing temporal consistency constraints.

Chapter 5 presents the work towards an application of hand gesture recognition for game-based wrist rehabilitation. This is a unique and novel approach because the gestures are selected from a set of human gestures suitable for wrist rehabilitation and implemented to control a new designed game built in a 3D environment.

In chapter 6, a paired datasets of real face images and synthesized correspondence's with face-masks is presented. This is used towards training of a proposed GAN-based facial reconstruction system which can be used for facial identification and secure authentication in human-computer interaction.

This dissertation concludes with a discussion on the key components of future research towards GAN-based hand pose estimation.

CHAPTER 2

GAN-Based Data Augmentation for Hand Pose Estimation Problem

2.1 Introduction

Hand pose estimation, which is a problem of predicting the 2D/3D position of hand joints, given an RGB/depth input, is receiving a lot of attention in the computer vision field. It has been applied in many applications, such as human–computer interaction (HCI) [12], gesture recognition [13, 14, 15], sign language recognition [16, 17, 18, 19], interactive games [20, 21, 22], user interface controls [23], computer-aided design (CAD) [24], etc. In recent years, by the advancements in deep learning algorithms, data-driven approaches have become more advantageous and have led to significant improvements in 2D/3D hand pose estimation, as a large number of annotated datasets have become available [25, 26, 27]. However, acquiring accurate 3D labeled data requires an expensive marker-based motion capture system or a massive multi-view camera setting. Therefore, to avoid annotating such large datasets, which is costly, time consuming and labor intensive, researchers are trying to find alternative approaches that can leverage them. One upcoming solution is to use synthetic data for training, where data are automatically annotated and convenient for generating a large scale of data with accurate ground truth. Although image synthesis can be generated using a physical renderer, there are usually a few differences between real and synthetic data, without consideration of depth sensor noise in a realistic way. Therefore, models trained on the synthetic data suffers from the domain shift problem, and they fail to perform well on real datasets, due to the domain gap between the real and synthetic datasets.

The most promising approach is to use generative models that learn to discover the essence of data and find a best distribution to represent it. Generative adversarial networks [28], or GANs in short, are a class of generative models, where two neural networks, generator and discriminator, contest with each other in a zero-sum game, where one agent's gain is another agent's loss. Given a training set, the generator learns to generate new data with the same statistics as the training set, while the discriminator's goal is to distinguish between real and generated samples. GANs have the ability to translate source synthetic images into realistic target-like images for training purposes. This is known as domain transfer learning. Several state-of-the-art transfer learning research works used GANs to enforce the alignment of the latent feature space. The conditional generative adversarial networks (CGANs) [29], which is an extension of GAN, has the ability to train synthetic models to generate images based on auxiliary information. Due to the popularity of the framework, it has become the foundation for many successful architectures, such as CycleGAN [30], StyleGAN [31], PixelRNN [32], DiscoGAN [33], etc.

The great success of these methods inspired more researchers to apply the generative adversarial networks to the hand pose estimation problem and train deep learning models either with a synthesized comprehensive dataset or few existing datasets in a semi-supervised setup or benefit from unlabeled data in a self-supervised manner to mitigate the burden of labeled-data acquisition.

Despite the large body of works that have been conducted on hand pose estimation using generative adversarial networks, no recent all-round survey has been conducted on it. As far as we know, this is the first survey among current publications which focused on GAN-based data augmentation for hand pose estimation problem. Moreover, different from existing review studies on the hand pose estimation problem which mainly discuss depth-based methods [34, 35], in this chapter, we present

a comprehensive study on the most recent GAN-based methods based on input data modality, i.e., RGB, depth, or multi-modal information. Another point of motivation of our work is that researchers do attach much importance to semi/unsupervised learning using GANs.

In what follows, in Section 2.2, we discuss the challenge followed by a comprehensive study of the most representative GAN-based data augmentation studies in solving the hand pose estimation problem in Section 2.3. Additionally, the existing hand pose datasets, the evaluation metrics, and the state-of-the-art results on two common datasets are summarized in Section 2.4.

Finally, potential research directions in this rapidly growing field and conclusions are highlighted in Sections 2.5 and **??**, respectively.

2.2 Challenge Analysis

Despite the rapid progress in hand pose estimation, it conventionally struggles from many difficulties, such as an extensive space of pose articulations, self-occlusions, and limited number of manually annotated data. The most important challenges in hand pose estimation are the following:

- **Annotation difficulties:** Existing learning-based methods require a large number of labeled data to accurately estimate hand poses. However, acquiring precise labels is costly and labor intensive.
- **Lack of various modalities:** Most of the existing hand pose datasets only contain RGB images, depth frames or infrared images instead of paired modalities.
- **Requirement for variety and diversity:** The real datasets are limited in a quantity and coverage, mainly due to the difficulty of annotations, annotation accuracy, hand shape and viewpoint variations, and articulation coverage.

- **Occlusions:** Due to the high degree of freedom (DoF), the fingers can be heavily articulated. In particular, hand–object and hand–hand interaction scenarios are still a big challenge, due to object occlusion and the lack of a large annotated dataset. Severe occlusion might lead to loose information on some hand parts or different fingers mistakenly. To handle occlusion, several studies resorted to a multi-camera setup from different viewpoints; however, it is expensive and complex to set up a synchronous and calibrated system with multiple sensors.
- **Rapid hand and finger movements:** Most conventional RGB/depth cameras cannot capture the speed of the hand motions and, thus, cause blurry frames or uncorrelated consecutive frames, which directly affect the hand pose estimation results.

Although many existing methods try to address these challenges with powerful learning-based approaches, as the effectiveness of generative deep learning aroused, many researchers try to address these through generative adversarial networks. Such methods dominate the benchmarks on large public datasets, such as NYU [36], ICVL [37], and FreiHAND [38]. In what follows, we first explain GANs, then we discuss studies on hand pose estimation, focusing on addressing the above challenges through data augmentation using GANs.

2.3 GAN-Based Hand Pose Data Augmentation

The generative adversarial network (GAN) is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data such that the model can be used to generate new examples as similarly as possible to the original dataset. GAN consists of two networks called the generator and discriminator; Figure 2.1a. The generator takes a simple random variable and generates new examples, and the discriminator tries to distinguish real

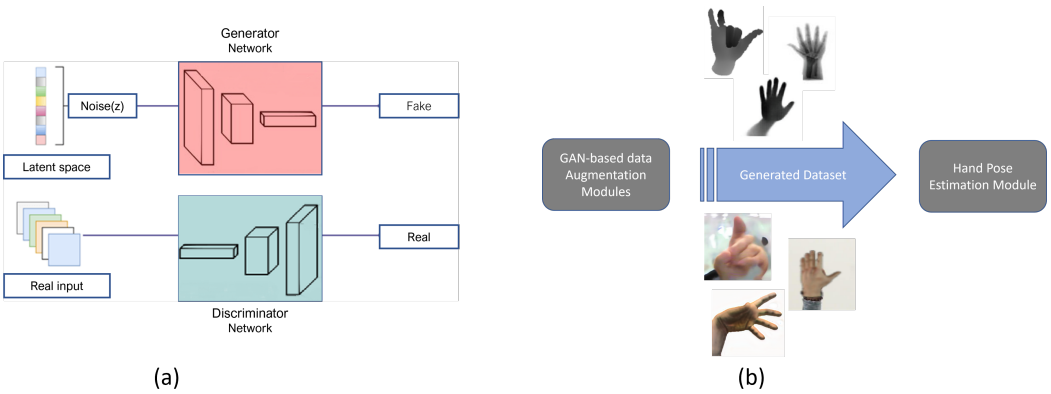

Figure 2.1. GAN-based hand pose data augmentation. (**a**) Overview on generative adversarial network, (**b**) procedure illustration of using generated data in HPE problem.

samples from the generated ones. The two models are trained together in a zero-sum game—adversarial—until the discriminator model is fooled about half of the time, meaning that the generator model generates plausible examples. Although the original framework [28] has no control of what is to be generated and it is only dependent on random noise, in a later study [29], the authors introduced conditional-GAN, where they add the conditional input vector c concatenated with noise vector z and feed the resulting vector into the generator. This conditional GAN can be used to generate examples from a domain of a given type. This allows for some of the more impressive applications of GANs, such as image-to-image translation, style transfer, photo colorization, and so on.

GANs are perhaps best known for their contributions to realistic image synthesis and model patterns of motion in video. GANs are able to enhance synthetic datasets such that the statistical distribution resembles a real-world dataset. Many approaches explore how to better manipulate images by applying GAN models [39, 40, 30]. Although image synthesis can be generated using a physical renderer, the difference between real and synthetic data is not considered in the image synthesis process.

Moreover, GANs' successful ability to model high-dimensional data, handle missing data, and the capacity of GANs to provide multi-modal outputs or multiple plausible answers made researchers more ambitious to the extent that they use GANs for the hand pose estimation problem either by generating data in new modalities or by realistic image synthesis through eliminating the domain gap between the synthetic and real data (Figure2.1b). Below is a comprehensive survey on GAN-based hand pose data augmentation based on GANs' application.

2.3.1 Image Style Transfer and Data Augmentation

To achieve high accuracy, much annotated data are required in data-driven methods, which are a labor-intensive and expensive process. Therefore, a few works aimed at improving the accuracy of pose estimation by using a synthetic image for data augmentation. However, using a physical renderer cannot embed the realistic noise in real data into data augmentation. To this end, several recent methods enrich existing training examples with style transfer by modeling real data noise realistically. Transferring the style from one image onto another has been a trendy subject in computer vision for the last few years.

In [1], they proposed a data-driven approach to generate depth hand images given ground-truth hand poses using a generative model. The style transfer is applied to generate the image with the style equivalent to the style image and the content from the content image. The style and content are defined based on the loss functions by measuring how far away the synthesized images are from the perfect style transfer. The proposed style-transfer network aims to transform the smooth synthetic images to become depth hand images more similar to the real ones. Figure 2.2 shows the architectural structure of the developed method. It contains three parts: a generator to transform the 3D hand pose into a deep hand image, and a discriminator which

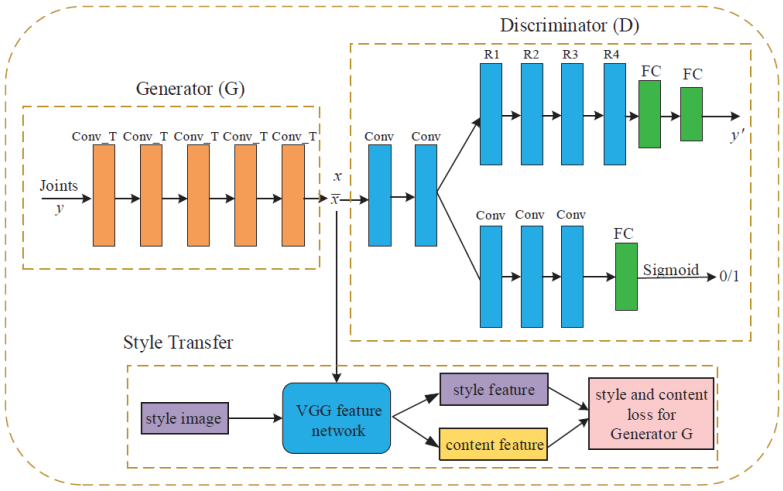

Figure 2.2. Flowchart of the proposed method in [1], covering the generator, the discriminator, and style-transfer networks in detail. Originally used in [1].

determines the authenticity of the generated image and the style-transfer network. At the end, they performed 3D hand pose regression on generated depth hand images based on the residual convolutional neural network. Their approach was evaluated and analyzed on three publicly available datasets NYU [36], ICVL [37], and MSRA gesture [41] datasets—and it was shown to boost hand pose estimation performance when used as training images.

To increase the amount of training data, Shrivastava et al. [2] proposed a framework which uses simulated and unsupervised learning to fit a model that uses unlabeled real data to improve the realism of a simulator's rendered data. They performed an experiment using real hand depth maps from the NYU [36] hand pose dataset in an extended version of SimGAN [2], and successfully added realistic noise to synthetic frames to better imitate imperfect real frames that are captured by depth cameras. Figure 2.3 gives an overview of the proposed model. Once the synthetic data are generated by a black box simulator, they are refined using a neural network

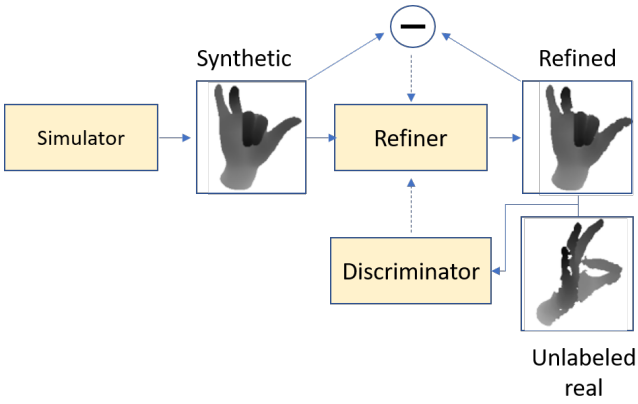

Figure 2.3. Overview of SimGAN; the self-regularization term minimizes the image difference between the synthetic and the refined images. Adapted from [2].

called the 'refiner network'. The refiner network is trained using adversarial loss from [28] to enforce the refined images similar to the real ones.

2.3.2 Domain Translation

Although using synthetic data is a potential solution to acquire accurate and unlimited data, avoiding expensive annotated real data, it has the strong disadvantage that the network trained only on synthetic data has limited generalization to real images and fails to generalize to real-world imagery. This visual domain shift from non-photo-realistic synthetic data to real images presents an even more significant challenge. Although the classical domain adaptation methods can be used to eliminate the dissimilarity between the real and synthetic images, recent studies focus on using GANs to bridge the gap between image distributions through adversarial training. Using domain translation techniques, such as image-to-image translation, not only leads to generating realistic training images which can be used to train any machine learning model, but it is also useful for generating data in different modalities. Since collecting and preparing training data in different modalities is a challenging task and

it requires expensive tools and a complex setup, researchers focus on using GANs to translate data from one domain to another or to multiple domains to generate a large scale of data in different modalities for the hand pose estimation problem.

Image-To-Image Translation

Image-to-image translation can be considered a type of image synthesis which maps an image from one domain to a corresponding image in another domain. It can be viewed as a generalization of style transfer since it not only transfers the style but also manipulates the attributes of the objects. Pix2Pix [39] and CycleGAN [30] are the most popular ones in supervised and unsupervised image-to-image translation. Pix2pix makes the assumption that paired data are available for the image translation problem that is being solved. In Pix2pix, model G was trained to translate images from domain X to domain Y. Cycle GAN does the same, but additionally, it also trains a model F that translates images in the opposite direction—from domain Y to domain X. CycleGAN was created in order to support working with unpaired data since having paired data available is actually rather rare, and collecting such data can require a large amount of resources.

In [3], Chen et al. suggested blending a synthetic hand poses generated by an augmented reality (AR) simulator with real background images to generate more realistic hand images, which later served as training data. Inspired by the pix2pix [39] which leverages the shape map to constrain the output image, they proposed a tonality-alignment GAN (TAGAN) to take the color distribution and shape features into account. Evaluation on multiple hand pose datasets indicates that their proposed approach outperforms state-of-the-art methods in both 2D and 3D hand pose estimation. Figure 2.4 gives an overview of the proposed method.

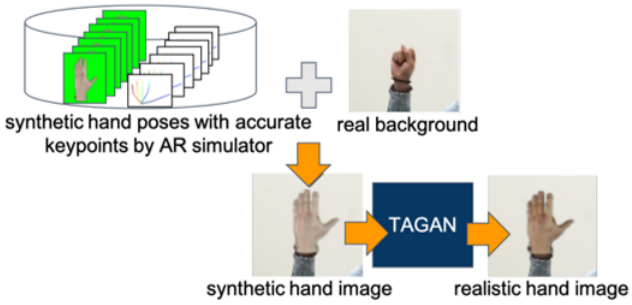

Figure 2.4. Overview of the TAGAN method for realistic hand image synthesis [3]. Synthetic pose by an AR simulator is blended with real background to yield a synthetic hand image, which is then fed to the proposed TAGAN to produce a more realistic hand image. Originally used in [3].

In another study by Wu et al., they proposed to directly generate realistic hand images from 3D pose and synthetic depth maps. However, unlike pose-guided person image generation, pose-guided hand generation is more challenging due to self-similarity and self-occlusion. To address these difficulties, they proposed a four-module model, MM-Hand, which contains 3D pose embedding, multi-modality encoding, progressive transfer, and image modality decoding [4]. They aimed to convert 3D hand poses to depth maps using a depth map generator. More specifically, in the 3D pose embedding module, they project the 3D hand pose onto a 2D image, given the projection matrix, which is followed by connecting the keypoints on each finger with an ellipsoid, using different colors. Then, a palm surrogate is formed through connecting a polygon from the base of each finger and wrist. Then, the depth map generator, which is a pix2pix-based model, is trained to synthesize depth maps based on any given 3D pose. Their experimental results show that the augmented hand images by their proposed approach significantly improved the 3D hand pose estimation results, even with reduced training data. The synthesized hand images using the proposed MM-Hand on the two benchmark datasets STB and RHP are shown in Figure 2.5.

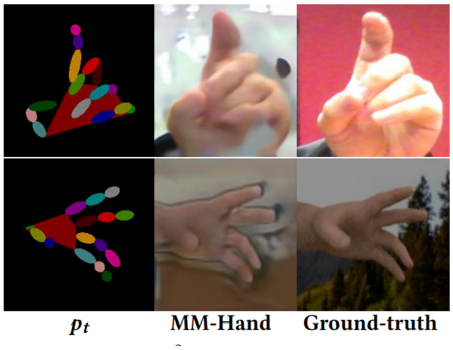

Figure 2.5. Qualitative results under MM-Hand model originally reported in [4]. Synthesized hand images using MM-Hand on two datasets, STB and RHP. From top to bottom: the STB dataset and the RHP dataset.

Moreover, to address the lack of various modalities problem, the authors in [7] presented a depth-image guided GAN model named DGGAN, which includes two sub-networks: a depth-map reconstruction module and a hand pose estimation module. Once the depth-map reconstruction module is trained using the GAN loss, it is able to generate a depth map of a hand based on the RGB input image. The second module trained using the task loss estimates hand poses from the input RGB and the GAN-reconstructed depth images. They aim to reconstruct the depth map from RGB hand images in the absence of paired RGB and depth training data. Once the depth maps are constructed from the RGB images, the hand pose estimation modules takes both RGB and depth images to estimate the 3D hand pose first by estimating the 2D hand keypoints on the RGB images followed by regressing the 3D hand poses from the inferred 2D keypoints. Next, exploiting the reconstructed depth map, it regularizes the inferred 3D hand poses. Experimental results on multiple benchmark datasets show that the synthesized depth maps produced by DGGAN are quite effective, yielding new state-of-the-art results in estimation accuracy by notably reducing the mean 3D end-point errors (EPE).

In another study [42], to generate new modalities, Haiderbhai et al. introduced a novel architecture based on the pix2pix model. They proposed a method of synthetic X-ray generation using conditional generative adversarial networks and created triplets for X-ray, pose, and RGB images of natural hand poses sampled from the NYU hand pose dataset . As a result, they introduced a two-module network. The first one aims to generate a 2D image of the pose, given the RGB input. Next, the output is stacked with the original RGB, which is used as input for the second module, which is identical to the pix2pix architecture. Their proposed model, pix2ray, has the advantages of creating X-ray simulations in situations where only the 2D input is available and generating more clear results, especially in occluded cases.

In [5], to improve hand pose estimation on weakly blurred infrared (IR) images under fast hand motion, the authors proposed a method based on domain transfer learning. The proposed model consists of a hand image generator (HIG), hand image discriminator (HID) and three hand pose estimators (HPE). The HIG synthesizes a depth image given input IR images. To train the HIG network, adapted by [39], they used the pair of unblurred depth and IR images with slow hand movement. The HID classifies whether the generated depth map conforms to the human hand depth map. The HPEs estimate the hand skeleton given an input depth image from the actual depth sensor, synthesized depth map, and IR image. It is worth mentioning that collecting depth and IR images from a single sensor eliminates the additional effort for depth image labeling. Moreover, since consistency loss is back propagated from the results of HPE, given the real depth image, the training is self-supervised. The proposed model is able to effectively improve hand pose estimation results in infrared images by generating un-blurred depth images as shown in Figure 2.6.

Since acquiring a large paired dataset can be difficult and expensive, inspired by CyclicGAN, Mueller et al. applied cycleGAN for realistic appearances of generated

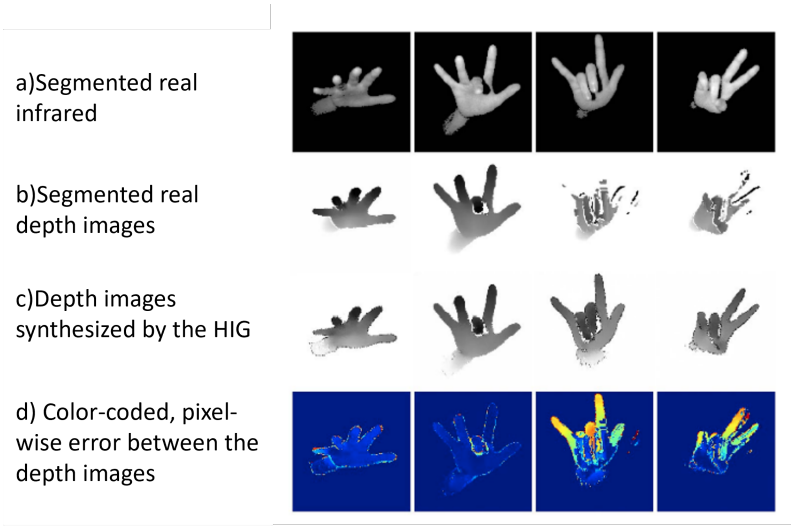

Figure 2.6. The HIG synthesizes a depth map from an infrared map. In the case of slow motion (the first and second column), the largest discrepancy is shown near the outline of the hand due to sensor noise. In the case of fast motion (the third and fourth column), the largest discrepancy is shown in blurry fingers. Originally reported in [5].

synthetic samples to reduce the synthetic-real domain gap [43]. They proposed a translation model, named GANerated, based on cycle-consistent adversarial networks (CycleGAN) to transfer the synthetic images to "real" ones so as to reduce the domain shift between them. Mueller et al. controlled the process through these two objectives: first converting synthesized image to real and calculating synth2real loss, and again converting the result to synthesized image and calculating real2synth loss. To make the images even more realistic, they also randomly put some background behind the hands. To simulate the occlusion, they artificially put some objects in front of the hand.

The proposed model obtains the absolute 3D hand pose by kinematic model fitting, which is more robust to occlusions, does not require paired data, and generalizes better due to enrichment of the synthetic data such that it resembles the distribution of real hand images.

In another study [6], inspired by cycleGAN [30], the authors applied a generative adversarial network to estimate hand poses through one-to-one relation between the disparity maps and 3D hand pose models. They aimed to enrich the existing dataset by augmenting them. Unlike other studies, they synthesized data in the skeleton space (instead of depth-map space), where data manipulation is intuitively controlled and simplified and, thereafter, automatically transfers them to realistic depth maps. Their proposed model consists of three networks: hand pose generator (HPG), hand pose discriminator (HPD), and hand pose estimator (HPE). The job of HPG is to generate a hand based on the 3D representation of joints while the HPD tries to determine how real or fake the generated samples are. The HPE is responsible for estimating the 3D hand pose based on the input depth map. During the training, these three networks are optimized to reduce the error of HPE. In the inference time, the algorithm refines the 3D model, which is guided by HPG to generate the most realistic depth map. More detailed architecture can be found in [6].

Although the recent studies try to solve an issue of lacking reliable RGB/depth datasets through generations of hand images, most of these works have focused on the generation of realistic appearances of hands without considering the temporal information. In [44], leveraged temporal information, they presented an unsupervised domain adaptation strategy based on CycleGAN for 3D hand–object joint reconstruction. Exploited by 3D geometric constraints and cycle consistency, their approach is able to effectively transfer annotation from the synthetic source images to an unlabeled real target domain. Moreover, by embedding short-term and long-term temporal consistency loss, the proposed model leverages unlabeled video to fine tune the model and outperforms the state-of-the-art models on benchmark datasets.

2.4 Results and Discussion

Although earlier hand pose datasets contain only depth data, more datasets that contain both RGB and depth images have been introduced due to the robustness of methods that leverage the RGB image. Since the performance of the DNN-based methods is closely tied to both the quality and quantity of the training data, in the following paragraphs, we compiled and described the most frequently used datasets in GAN-based data augmentation studies.

2.4.1 Benchmark Datasets

- **NYU Hand Pose Dataset** It has 72,000 images as training and 8000 as testing data. Data are collected by 3 Microsoft Kinect cameras from 3 different views with 36 3D annotations. It is the most commonly used dataset in the hand pose estimation problem since it covers a variety of poses in RGB and depth modalities.
- **Imperial College Vision Lab Hand Posture Dataset (ICVL)** The ICVL contains 300,000 training and 1600 images as testing images. All depth images are captured by Intel RealSense and, in total, 16 hand joints are initialized by the output of the camera and manually refined.
- **MSRA15** This includes 9 subjects with 17 different gestures. In total, it has 76,000 depth images with 320 $\times$ 240 resolution, collected by Intel's Creative Interactive Camera, with 21 annotated joints.
- **BigHand2.2M** It contains 2.2 million real depth maps collected from 10 subjects. Since it is collected by six magnetic sensors, it has precisely 6D annotations.

Table 2.1. Summary of hand pose estimation datasets commonly used in data augmentation using GANs.

Dataset	Modality	Type	Number of Joints	Number of Frames
NYU	D	Real	36	81 k
ICVL	D	Real	16	332.5 k
MSRA15	D	Real	21	76.5 k
BigHand2.2M	D	Real	21	2.2 M
STB	RGB+D	Real	21	18 k
RHD	RGB+D	Synthetic	21	44 k

- **Stereo Hand Pose Tracking Benchmark (STB)** STB includes 18,000 frames, 15,000 for training and 3000 for testing with 640 $\times$ 480 resolution. The 2D keypoint locations are obtained using the intrinsic parameters of the camera.
- **Rendered Hand pose Dataset (RHD)** It has 43,986 rendered hand images from 39 actions performed by 20 characters. Each depth image comes with segmentation mask, 3D and 2D keypoint annotations.

Modality, the type of data (i.e., synthetic or real data), the number of joints and the number of frames, are summarized in Table 2.1.

2.4.2 Evaluation Protocol

The most common evaluation metrics that are used to gauge the performance of these methods are end-point error (EPE) and percentage of correct keypoints (PCK). The former one is the average 3D Euclidean distance between the ground truth and predicted joints, and the latter one measures the mean percentage of the predicted joint locations that fall within a certain error threshold.

2.4.3 Quantitative and Qualitative Results

It should be noted that since not all these works evaluate their performance using both metrics and on the same dataset, we summarized the reported results for methods on NYU and STB hand pose datasets.

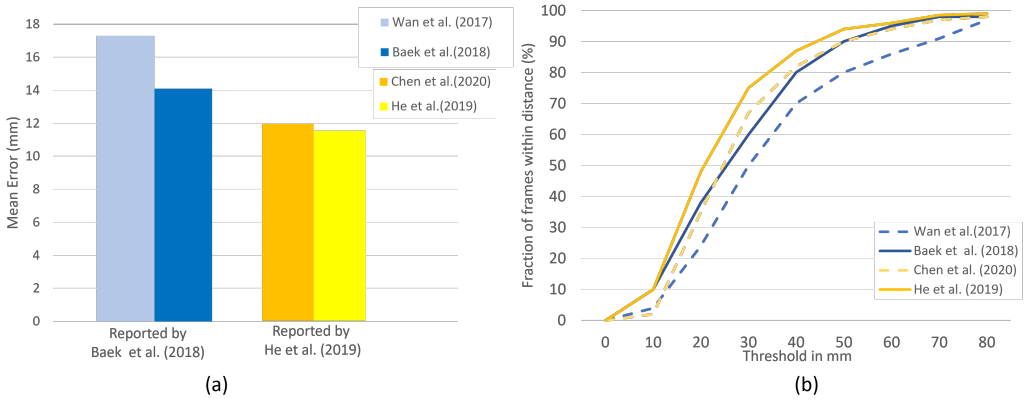

Figure 2.7. On NYU dataset, the contribution of the [1, 6] methods to the accuracy are compared. (**a**) Mean error. (**b**) The fraction of frames over different maximum Euclidean distance error threshold. The larger the area under each curve, the better. (Best viewed on screen).

For the NYU hand dataset, we choose Refs. [1, 6] since the other studies with NYU do not provide the quantitative results and only compare the quality of synthesized images. In Figure 2.7, the results are illustrated with and without the use of synthetic images for training on the NYU dataset. As it is reported in [6], the developed method obtains 0.4 mm reduction of the average 3D joint error, compared with the current best performance by Pose-REN [45]. Moreover, the results from Ref. [1] also indicate the 3.2 mm reduction in mean error due to the increase in training samples from the proposed GAN-based data augmentation model. Furthermore, the developed methods are compared by the percentage of frames at different maximum

error thresholds in Figure 2.7b. It has shown that both studies [1] and [6] achieved higher accuracy compared to the HPE base lines, [46] and [45], respectively.

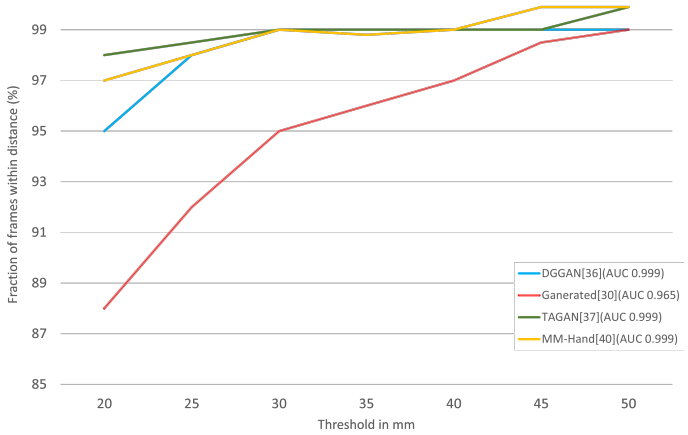

Figure 2.8. Comparison of [7, 3, 8, 4] approaches for 3D pose estimation on the STB dataset. The fraction of frames over different maximum Euclidean distance error threshold. The larger area under the curve (AUC) represents better results. (Best viewed on screen) .

For the STB dataset, we compare DGGAN [7], GANerated [8], TAGAN [3], and MM-Hand [4] based on the reported PCK value in Figure 2.8. As it is mentioned, the larger the area under the curve, the higher the represented accuracy. The GANerated [8] has the lowest value of 0.965, compared to the others.

2.5 Discussions and Future Directions

We provide a detailed discussion of the most recent studies on image synthesis and image-to-image translation in HPE, where they aim to alleviate the burden of the costly 3D annotations in a real-world dataset. The explosion of interest in GANs is driven not only by their potential to learn deep, highly nonlinear mappings from a latent space into a data space and back, but also by their potential to make use of the vast quantities of unlabeled image data that remain closed to deep representation

learning. However, due to the lack of robust and consistent metrics, coming up with good evaluation metric is still an open challenge to compare different GAN variants based on the visual assessment of the generated images. Moreover, despite the great performance of the current methods on hand pose estimation using GANs, still there remain difficulties in generalizing them to multi-hand interaction. Moreover, because of the interest of big technology companies in this field, perhaps in the near future, we can acquire much bigger and more generalized datasets generated by GAN and, therefore, very well-performing models on different modalities.

CHAPTER 3

Semi-Supervised 2-D Hand Keypoint Localization using Unpaired Image-to-Image Translation

3.1 Introduction

Accurate fingertip localization from depth images plays an essential role in many computer vision applications such as sign language recognition [47] and human-computer interaction [48] when image-based models are used. Many proposed approaches such as [49] and [50] 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 [51] and PifPaf [52] 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. [53] and Wetzler et al. [54] 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. Although unsupervised detection of landmarks can extract useful features, it is not able to detect perceptible landmarks with out supervision [55, 56]. In [57], to learn without explicit annotations, Dong et al. build on the pseudo-labeling technique

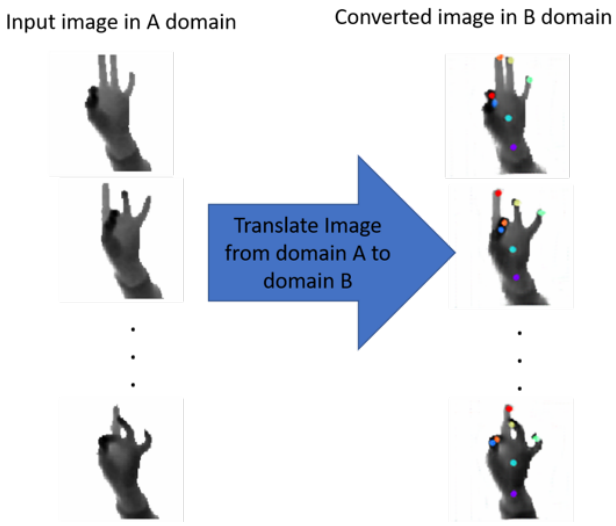

Figure 3.1. Examples of translated input image from A domain to B domain..

which uses a teacher model and two students to generate more accurate pseudo-labels for unlabeled data. In another study by Jakab et al. [56], additional class attributes were utilized for semi-supervised keypoint detection.

Therefore, inspired by the progress of the Generative Adversarial Network (GAN) and image-style transfer, we propose a semi-supervised 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 as shown in Figure 3.1. 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(See Figure 3.2).

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

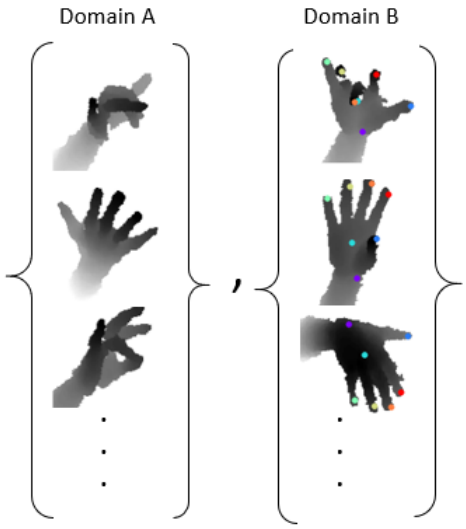

Figure 3.2. Unpaired training data, consisting of a source set and a target set, with no information provided as to which x_i in domain A matches which y_j in domain B.

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. Moreover, fingertips would be detected irrespective of user orientation using this method.

3.2 Hand Pose Estimation

In many hand pose estimation studies, such as [58] and [59], 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 [60, 61, 62]. Malassiotis

and Strintzis extract PCA features from depth images of synthetic 3D hand models for training [63]. Suryanarayan et al. [64] use depth information to recognize scale and rotation invariant poses dynamically. Sinha et al. [65] 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 semi/self-supervised learning or hybrid methods are needed. By fast progress of Generative Adversarial Networks (GAN), several studies have been performed to model the statistical relationship of the 3D pose space and corresponding space of the input data in semi/self-supervised manner [66, 46, 67]. Chen et al. [7] proposed a conditional Generative Adversarial Network (GAN) model called Depth-image Guided GAN (DG-GAN) 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 [1], He at al. proposed a data-driven method to generate deep hand images closer to real ones during training. In Chen et al. [68], they propose tonality-alignment generative adversarial networks (TAGAN) to align the tonality and color distribution between synthetic hand poses and real backgrounds.

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 [69], 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 [70]. 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,[71] and [72], authors use a weight sharing strategy to learn the most common representation between domains. In [2] and [73], 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[74]. In [75] 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[76].

3.3 Our 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 [30]. 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 3.3). The key idea behind CycleGAN is that it allows the model to use two unpaired collection of the images rather than two specific images. 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

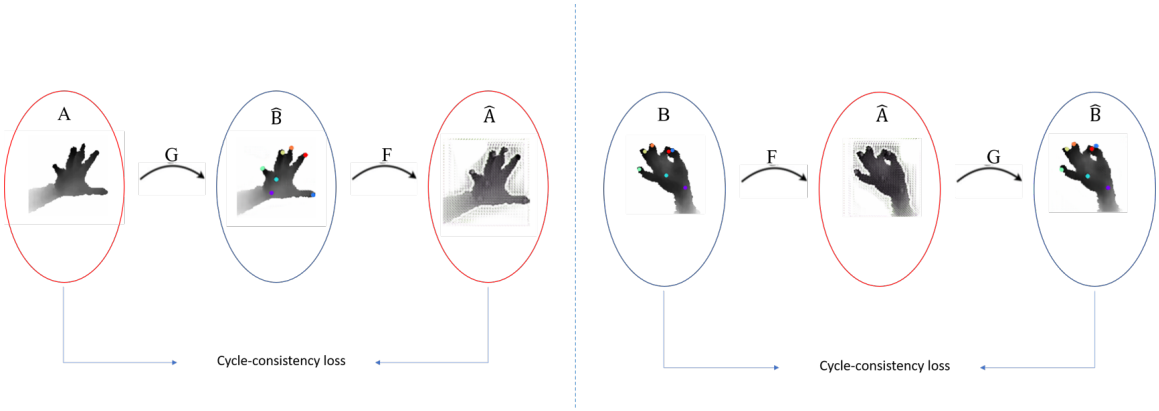

Figure 3.3. The simplified architecture of first stage to perform unpaired image to image translation using Cycle-Consistent Adversarial Networks.

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 3.4, 3.3 and Figure 3.6 respectively.

3.3.1 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.

3.3.1.1 Adversarial Loss

Adversarial loss[28] is a powerful loss specifically for image generation task. It, enforces the generated image to be indistinguishable from real photos. Since the

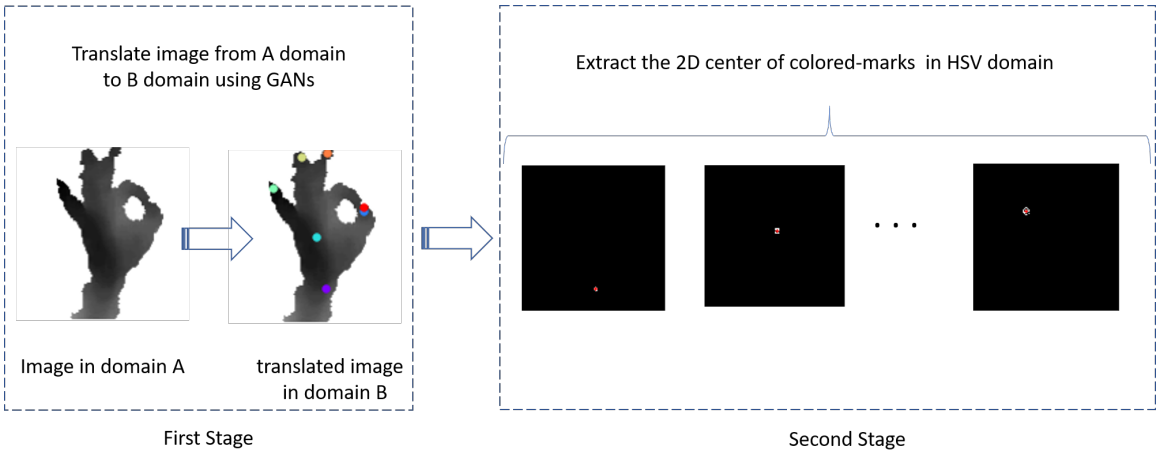

Figure 3.4. The overview of proposed model.

model has two mapping functions G and F, an adversarial loss is defined for each mapping function as:

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_B, A, B) = \mathbb{E}_{b\sim pdata(b)}[logD_B(b)] \\ + \mathbb{E}_{a\sim pdata(a)}[log(1 - D_B(G(a)))],\end{aligned} \tag{3.1}$$

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:

$$\begin{aligned}\mathcal{L}_{GAN}(F, D_A, B, A) = \mathbb{E}_{a\sim pdata(a)}[logD_A(a)] \\ + \mathbb{E}_{b\sim pdata(b)}[log(1 - D_A(F(b)))],\end{aligned} \tag{3.2}$$

3.3.1.2 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[30]. 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[30]:

$$\begin{aligned}\mathcal{L}_{cyc}(G,F) = \mathbb{E}_{a\sim p_{data}(a)}\left[||F(G(a)) - a||_1\right] \\ + \mathbb{E}_{b\sim p_{data}(b)}\left[||G(F(b)) - b||_1\right]\end{aligned} \tag{3.3}$$

The same process is done in opposite direction as shown in Figure 3.3.

3.3.1.3 Full Objective

The final loss function for training Cycle-GAN is defined as [30] :

$$\begin{aligned}\mathcal{L}(G,F,D_A,D_B) = \mathcal{L}_{GAN}(G,D_B,A,B) \\ + \mathcal{L}_{GAN}(F,D_A,B,A) \\ + \lambda\mathcal{L}_{cyc}(G,F)\end{aligned} \tag{3.4}$$

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

$$G^*, F^* = arg\ min_{G,F}\ max_{D_a,D_b}\ \mathcal{L}(G,F,D_A,D_B) \tag{3.5}$$

3.3.2 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 3.5. Hue represents an angle in range [0,2π]

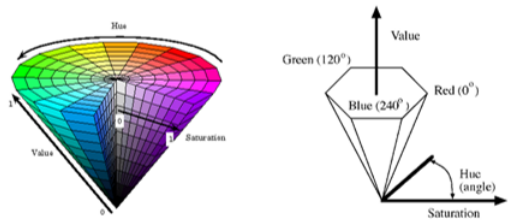

Figure 3.5. HSV color space representation.

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 [77]. 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. Therefore, in the second stage, to extract region of interest from the generated colored annotated depth image from the previous stage we perform color segmentation in HSV color space. 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 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 3.6).

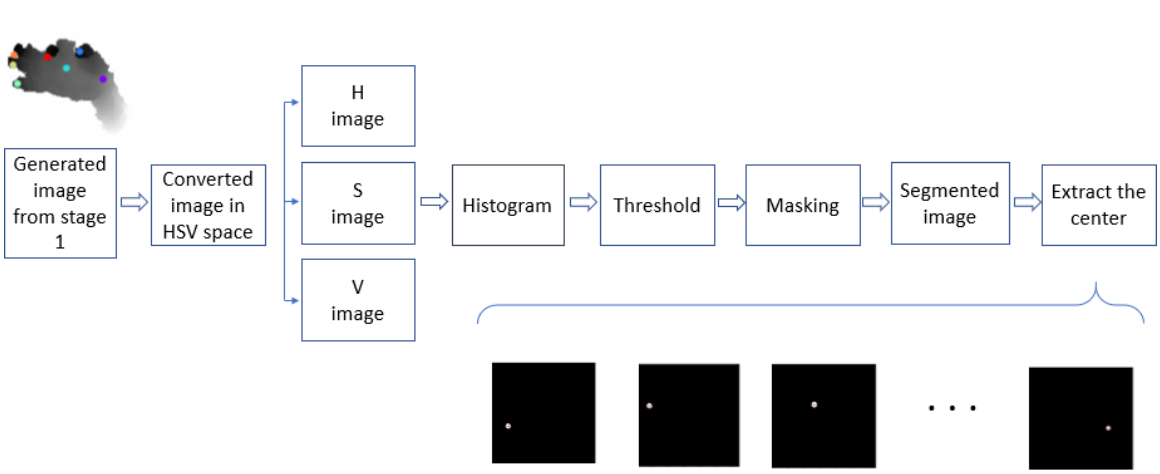

Figure 3.6. Second stage overview.

3.4 Experimental Details

Although there are some datasets like ICVL [37] and MSRA14 [58] 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.

3.4.1 Data preparation

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

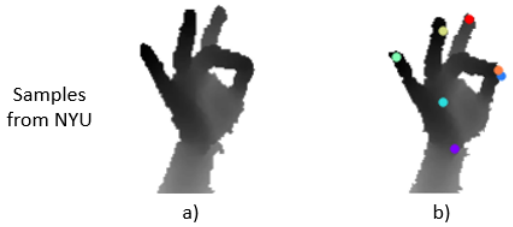

Figure 3.7. Example of test data; real depth image (a) and annotated depth (b) sample from NYU hand dataset.

1(front view). For test data, we randomly chose 300 real depth images of the same view from test set of 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 3.7 shows examples of customized NYU hand dataset.

3.4.2 Model Architecture and Internal Parameters

The general architecture of CycleGAN [30] 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 3.3.2.

3.4.3 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.

3.5 Results and Discussion

Since, most of previous methods , [54] and [78], 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 with almost 0.03 of NYU dataset while previous methods are trained over the entire dataset. Furthermore, 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.

Table 3.1. Quantitative evaluation on NYU (Fingertips only).

Methods	Mean error (Pixels)
Ours	7.2
Duan paper[78]	12.2
Mask RCNN(kpt and mask)[78]	13.6
Mask RCNN(kpt)[78]	24.5

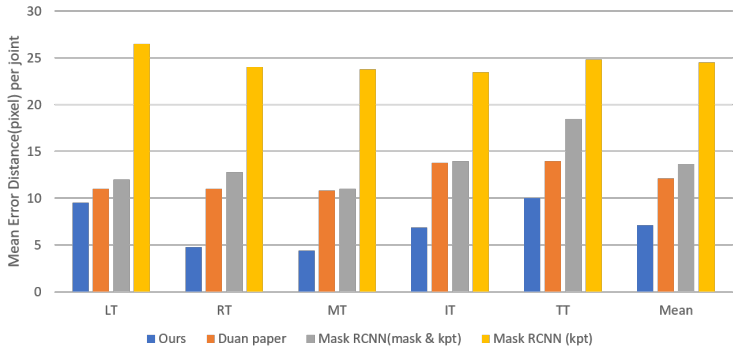

Figure 3.8. Comparison on per-joints mean error distance in pixels on NYU hand dataset.

3.5.1 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 [78]) and summarized the 2D localization results for 5 fingertips in Figure 3.8 and Table 1.

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 [78]. Moreover, the comparison of our methods with extracted

results from [78], on each joint for right hand in the NYU hand dataset is shown in Figure 3.8.

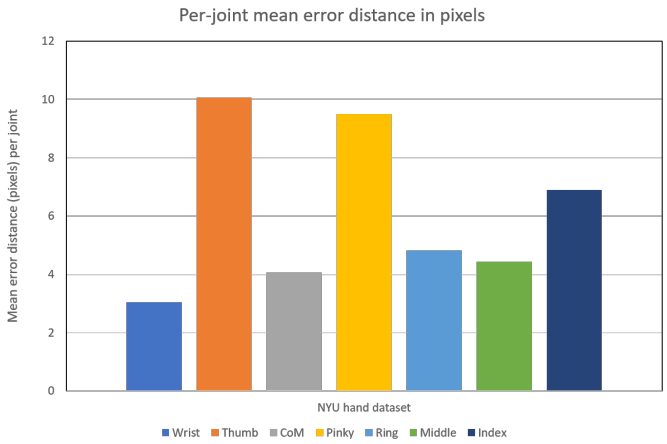

Figure 3.9. Per-joint mean error distance in pixels on NYU test dataset.

Moreover, Figure 3.9 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 3.10.

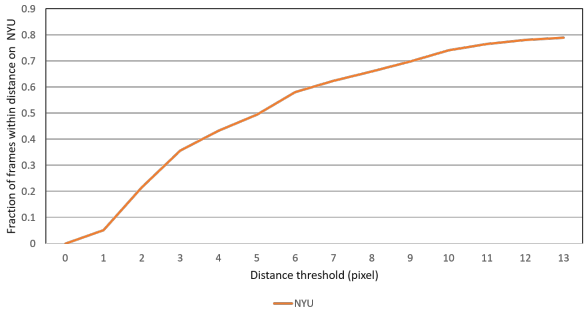

Figure 3.10. Fraction of frames within distance on NYU test datasets.

Figure 3.11. 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.

3.5.2 Qualitative Results

As can be seen in Figure 3.11, 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.

3.6 Conclusion and Future Works

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. Moreover, we plan to apply a similar pipeline with small changes to RGB images.

CHAPTER 4

2-D Fingertip Localization on Depth Videos Using Paired Video-to-Video Translation

4.1 Introduction

Despite the fact that convolutional neural networks have brought increasingly effective solutions for hand pose estimation on RGB and depth images, hand pose estimation on depth videos remains difficult to deploy in practice, due to the lack of large annotated datasets that cover a wide enough range of scenarios and lighting conditions. Although a common remedy is to exploit synthetic data, unfortunately a network trained with this type of data can easily fail when dealing with real data due to the domain shift problem. Moreover, complex backgrounds, self-occlusion, large view-point variation, and rapidly changing illumination in realistic environments always make it arduous to generalize the parameters of a model to the different domain data. A promising direction to address the domain gap is to use domain translation techniques on images and videos, such as Pix2PixHD [79] and video-to-video translation [80], to turn the source domain data close to the target domain.

Inspired by previous application-specific video synthesis methods[81, 75, 82], in this study we investigate the problem of 2D fingertip localization on depth videos using video-to-video translation. Our model consists of two stages: 1) video-to-video translation model 2) color segmentation in HSV color space and histogram threshold. At training time, we use a video-to-video translation model to translate a source depth video into a target-style video (depth video along with color markers on fingertips). We minimize an adversarial loss to ensure the statistical similarity

of the translated videos and the real target videos. Leveraged by rich temporal information, our proposed framework successfully localize 2D hand keypoints on depth video especially when the existing frame-based HPE provides inaccurate estimations due to motion blur.

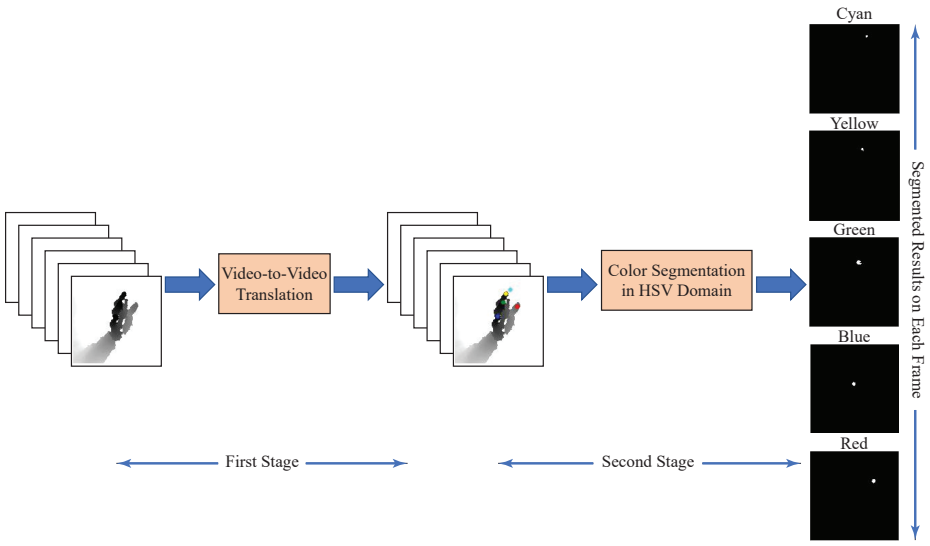

Figure 4.1. Model overview.

Quantitative and qualitative results on the NYU benchmark dataset show that our proposed approach outperforms the state-of-the-art methods and can handle motion blur and severe occlusion independent of user hand orientation. In summary, this chapter presents the following contributions:

- To the best of our knowledge, this is the first work that estimates hand poses from depth videos.
- We present a novel end-to-end framework based on the domain adaptation for 2D hand keypoint localization on depth video. The method generalizes well and improves the accuracy of 2D hand pose estimation under motion blur and severe occlusion.

- Our method works for other input video formats such as RGB, enabling many applications from generating RGB from depth video to hand keypoint localization on RGB videos.

4.1.1 2D/3D Hand Pose Estimation

Recently, there has been a big increase in the number of papers that exploit convolutional and deep networks to estimate hand pose on RGB-D data [83, 49, 84, 85, 25, 46]. While one line of the studies tries to directly estimate 3D pose, the other group first estimates 2D pose before lifting the 2D pose to 3D pose through structured learning or a kinematic model. In the latter case, 2D hand pose estimation is an important and fundamental task. Zimmermann and Brox [49] proposed combining hand segmentation and 2D hand pose estimation through CPM [86], followed by estimating 3D hand pose relative to a canonical pose. In [84], the authors estimate 3D hand pose by applying inverse kinematic on estimated 2D hand pose.

Compared to RGB images, depth images are robust to texture and light intensity variations. Moreover, since depth images contain surface geometry information, several studies [87, 34, 88] try to estimate hand poses on depth images or use depth map as some intermediate guidance for hand pose estimation network. In [89], the authors use a depth map as intermediate guidance and conduct an end-to-end training framework. In another work, they proposed a weakly supervised approach leveraging the depth map for regularization [90].

Despite these advances, 2D/3D hand pose estimation on temporal data remains a challenging problem due to the lack of accurate, large-scale 2D/3D pose annotations, fast movement of the hand and motion artifacts. The methods discussed above estimate frames in isolation and may not perform well on sequential data. Since they do not leverage the temporal consistency between individual frames, the output may

produce noisy or infeasible changes in the estimated poses. While there are several studies on hand pose estimation using RGB videos [91, 92, 93, 94], to our knowledge, hand pose estimation using depth video has not yet been explored by prior work.

4.1.2 Domain Transfer Learning

Domain transfer learning algorithms attempt to transfer image/video from one domain to a corresponding image/video in another domain by bringing the distribution of the source closer to that of the target. Several recent works on hand pose estimation use different modalities (RGB, depth, and synthetic images) for domain transfer learning [95, 96, 97, 46]. Rad [95] maps the features from the RGB space to the depth space to avoid training using labeled color images. The authors in [98] proposed a framework for 3D hand pose estimation from RGB images aided from depth information. Mueller [43] use an image-to-image translation network to create a large amount of RGB training images and combine a CNN with a kinematic 3D hand model for pose estimation. In [76], they used a cyclic concept for making a one-to-one relation between RGB images and 3D hand joints. They combined a GAN and a Variational Auto encoder (VAE) to transfer images to a latent space before transferring them to a target domain which is defined as a 3D pose. However, these methods only focus on generating RGB/depth images and none of them utilize the temporal information in order to generate sequenced data for assisting hand pose estimation on videos. Moreover, directly applying existing image synthesis approaches to an input video often results in temporally incoherent videos of low visual quality. To this end, we came up with a new approach for 2D hand pose estimation on depth videos based on video to video translation models which, unlike the other methods, learns not only the appearance of objects and scenes but also realistic motion and transitions between consecutive frames.

4.1.3 Optical Flow

Optical flow is the pattern of apparent motion of image objects between two consecutive frames caused by the movement of object or camera. Consider a pixel $I(x, y, t)$ in first frame, it moves by distance (dx, dy) in next frame taken after dt time. So since those pixels are the same and intensity does not change, we can say,

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{4.1}$$

then take Taylor series approximation of right-hand side, removing common terms and divide by dt to get the optical flow equation:

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \tag{4.2}$$

where $u = \frac{d_x}{d_t}$ and $v = \frac{d_y}{d_t}$.

$\frac{d_I}{d_x}$, $\frac{d_I}{d_y}$ and $\frac{d_I}{d_t}$ are the image gradients along the horizontal axis, the vertical axis, and time. Hence, we conclude with the problem of optical flow, that is, solving $u(\frac{d_x}{d_t})$ and $v(\frac{d_y}{d_t})$ to determine movement over time. However, as this equation is under-constrained, there are several methods such as supervised learning method [99] to solve for u and v.

Optical flow is utilized in one form or another in most video-processing algorithms especially it is useful in providing smoothing in Generative Adversarial Networks such as [80], so that generated output can appear to be temporally coherent.

4.2 Methodology

We propose a two-stage pipeline for fingertip localization on depth video. An overview of our approach is shown in Figure 4.1. We propose to reduce the 2D hand keypoint localization problem to video-to-video translation using generative adversarial networks. In the first stage, inspired by [80], we use a video-to-video

translation model, which is conditioned on the previous video sequence as well as the corresponding target video. In the second stage, we apply color segmentation techniques on the translated video in the target domain to extract the center of each colored marks on the fingertips. Based on our experiments with the state-of-the-art frame-based hand pose estimation methods, our results show that per-frame approaches cannot capture the essential properties of videos, such as global motion patterns and shape and texture consistency of translated objects. To the best of our knowledge, our method is the first study on both hand pose estimation on depth video and the first work to introduce the concept of using domain translation to support 2D hand keypoint localization in hand pose estimation on depth videos. In the remainder of this chapter, we will describe each module of our model in detail.

4.2.1 Mathematics

Given a real depth hand video, we aim to learn the mapping function that translates the source video to a new realistic depth hand video along with 5 color marks on each fingertip. In principle, one can apply image-to-image translation on each frame. However, our experiments show that frame-wise translation only produces realistic results on a single frame, and under performs significantly on the scale of the whole video. The main reason is that videos have a temporal structure in addition to the spatial structure found in images. This means that information in a video is encoded not only spatially, but also sequentially.

4.2.1.1 Sequential Generator Exploiting Optical Flow

Similar to [80], given a pair of real depth videos in the source domain $s_1^T \equiv s_1, s_2, ..., s_T$ and corresponding depth video along with 5 color marks on each fingertips $x_1^T \equiv x_1, x_2, ..., x_T$ in the target domain, we aim to generate videos $\tilde{x}_1^T \equiv \tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_T$

conditioned on the source domain such that the conditional distribution of generated frames given the source domain video is identical to the conditional distribution of the target domain videos given in the source domain videos.

$$p(\tilde{x}_1^T|s_1^T) = p(x_1^T|s_1^T). \tag{4.3}$$

To learn the desired temporal coherence, the problem is expressed as conditional sequence distribution matching. In other words, the problem is formulated by factorizing $p(\tilde{x}_1^T|s_1^T)$ as the product of the probabilities of the last two time steps.

$$p(\tilde{x}_t^T|s_t^T) = \prod_{t=3}^{T} p(\tilde{x}_t|\tilde{x}_{t-2}^{t-1}, s_{t-2}^t). \tag{4.4}$$

To simplify the equation for $t = 3$ we followed the below notations;

- current source frame (s_3)
- the past two frames in source domain (s_1 and s_2)
- the past two generated frames in target domain ($\tilde{x}_1$ and $\tilde{x}_2$)

Based on this conditional assumption which is called the Markovian assumption, authors in [80] used the deep network to learn the mapping function F from source to target domain. The consecutive frames contain a remarkable amount of redundant information. To optimize the process, we can estimate the next frame using a frame warping process assuming the optical flow information (W) is available. The optical flow is estimated using both input source images s_{t-2}^t and previously translated images $\tilde{x}_{t-2}^{t-1}$ (see equation 4.4). Moreover, to avoid failures of Optical flow warping in occluded area, an additional mask (M) is used to determine the occluded and non-occluded pixels. For W and M networks, we used pre-trained models Flownet2 [99] and Mask-RCNN [51], respectively. These models, together produce W(t) -warped images of $\tilde{x}$. Finally, to generate frame (h) in occluded pixels, an image generator (H) is added.

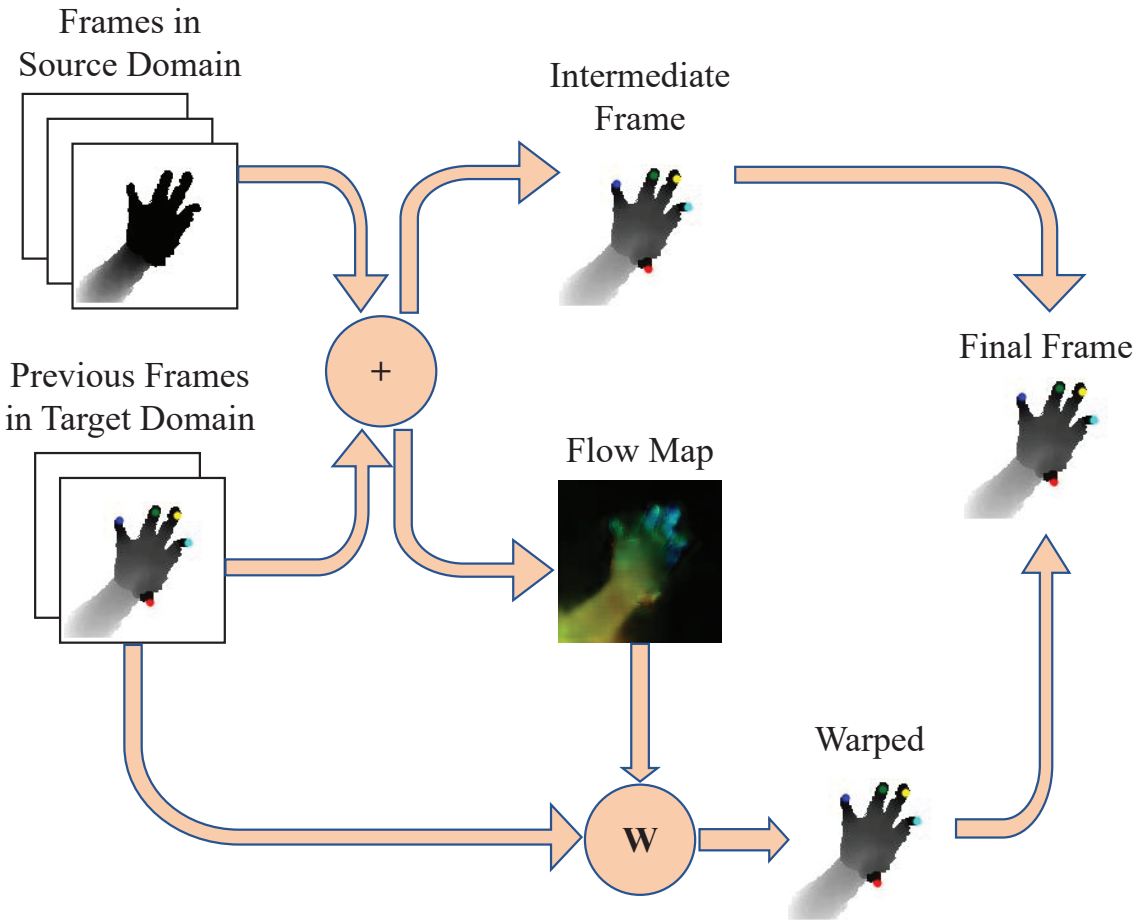

Figure 4.2. An illustration of the generator components.

As a result, the network slowly adds details by blending the warped pixels and the generated pixels as follow:

$$\begin{aligned} \tilde{x}_t &= F(\tilde{x}_{t-1}, \tilde{x}_{t-2}, s_t, s_{t-1}, s_{t-2}) \\ &= (1 - \tilde{m}_t) \odot \tilde{w}_{t-1}(\tilde{x}_{t-1}) + \tilde{m}_t \odot \tilde{h}_t. \end{aligned} \tag{4.5}$$

where $\odot$ represents the element-wise product operator and 1 is an image of all ones. The first part refers to pixels warped from the previous frame and the second part is the hallucinated image, synthesized directly from the generator H for the

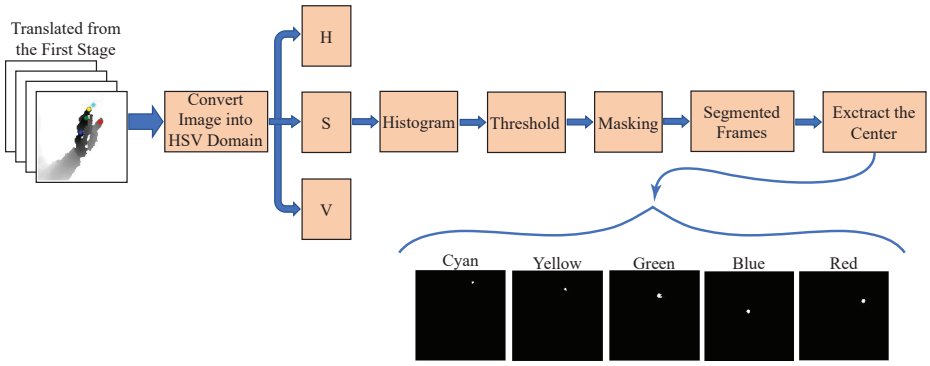

Figure 4.3. Second stage overview.

occluded part. The generator components of the first stage, which are similar to Pix2PixHD [79], are illustrated in Figure 4.2. All the models M,W, and H use short residual skip connection.

4.2.1.2 Discriminator: Exploiting Spatio-Temporal Information

To solve the problem considering spatio-temporal information, two discriminators are used (similar to [80]); one as image discriminator D_I, and one as video discriminator D_V. Both follow PatchGAN from Pix2Pix [39] to make computations more tractable by lowering the time and space complexity. The image discriminator tries to distinguish the generated image (video slice) from each pair of real depth and a depth image in target domain. The video discriminator ensures that consecutive output frames resemble the temporal dynamics of a real video given the same optical flow.

4.2.1.3 Final Learning Objective Function

The video-to-video translation model is trained by solving

$$\min_{F}(\max_{D_I} L_I(F, D_I) + \max_{D_V} L_V(F, D_V)) + \lambda_W L_W(F), \tag{4.6}$$

where L_I is the adversarial loss for conditional image discriminator D_I, L_V is the adversarial loss on K consecutive frames for video discriminator D_V, and L_W is the flow estimation loss with $\lambda_W = 10$, based on a grid search. Moreover, the discriminator feature matching loss and the VGG feature matching loss are added, similar to Pix2PixHD [79]. All the above are combined to achieve the total optimization criterion to tackle video synthesis.

4.2.2 Implementation Details on Video-to-Video Translation

For the video-to-video translation model in the first stage, we start by generating a few low resolution frames and gradually improve it to full resolution with all 30 frames. We trained the model for 80 epochs using ADAM optimizer [100] and we set learning rate, $\beta_1$and β_2 as 0.0002 , 0.5 and 0.999 respectively.

4.2.3 Color Segmentation in HSV Color Space

The objective of the second stage is to find out the colored regions on each frame in the translated sequence. This color detection is the process of separation between fingertips and non-fingertips pixels. However, RGB color space is not preferred for color based detection and color analysis because of mixing of color (chrominance) and intensity (luminance) information and its non-uniform characteristics. Thus, we perform color segmentation in HSV domain. HSV color space is a collection of three different components as hue, saturation and value. Geometrically, it can be pictured as a cone or cylinder with H being the degree, saturation being the radius, and value being the height. Luminance and Hue-based approaches discriminate color and intensity information even under uneven illumination conditions and are best suited for image with uniform background.

The algorithm for the detection of colored marks is explained as follows (Figure 4.3):

1. Each frame on a translated sequence is transformed into HSV color space.
2. Transformed frame is split into three components.
3. The Histogram is computed for all three components followed by determination of lower and upper threshold value for each component. The threshold values are set based on converting of predefined colors in RGB to HSV value before fine tuning.
4. Then, to isolate the desired colored areas, we applied multiple masks. A low threshold and high threshold mask for hue, saturation and value. Pixels within the threshold's range will be set to 1 and the remaining pixels will be zero. Next by computing the connected component of Boolean image, the colored areas are segmented.
5. Finally, the centers of the segmented parts are extracted as 2-D coordinates that represent the locations of the desired points on each frame of the input video.

4.3 Experiments

In recent years, the NYU hand pose dataset [36] has become a standard dataset for depth-based methods. The ground truth for this dataset is computed using a generative method [101] instead of manual labeling.

4.3.1 Data Preparation

The creators of this dataset provide RGB images that are warped onto the depth map which makes the dataset unusable by RGB-only methods. To get around this, we used real depth frames from the training data of the NYU dataset, which all

are center-cropped around the hand with a resolution of 128 × 128. To prepare the training data for both source and target domains, we grouped every 30 real depth frames from the training set of NYU into a single depth video for the source domain and created corresponding frames with color marks on each fingertips for the target domain as shown in the Figure 4.4. Totally, there are 2425 videos of 30 frames for source and 2425 videos of the same length for target domain. For testing, we did the same on test set of NYU and prepared 275 videos of 30 frame for each domain.

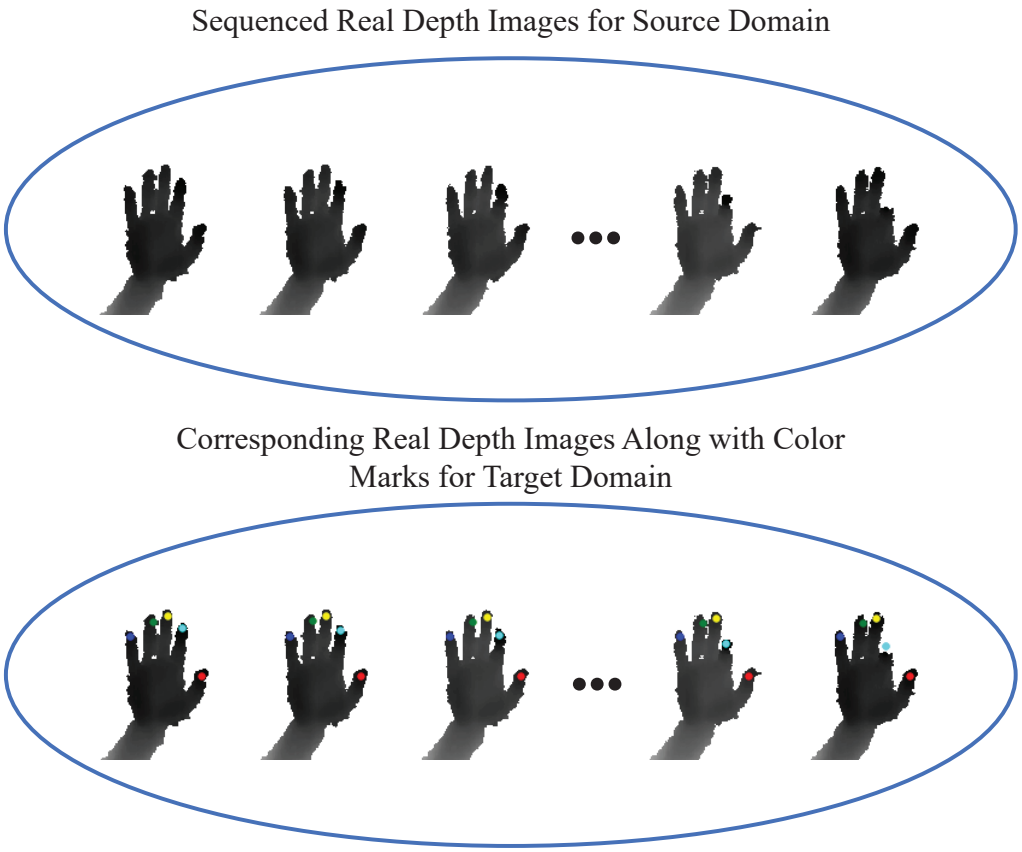

Figure 4.4. Example of training video data for video to video translation.

4.3.2 Evaluation Metrics

We quantitatively evaluate our method on the NYU test set using Mean Error (ME) as metric. ME is defined as the average Euclidean distance between the predicted and the ground-truth joint locations, measured in pixels.

4.4 Qualitative and Quantitative Results

To the best of our knowledge, this is the first marker-less 2D hand keypoint localization method on depth videos. However, to have a fair evaluation, we compared results of our pipeline with the state-of-the-art hand pose estimation methods on single frames. These methods have published their results in terms of 3D hand pose estimation. We have computed their 2D pose estimation performance from their prediction files that are made publicly available.

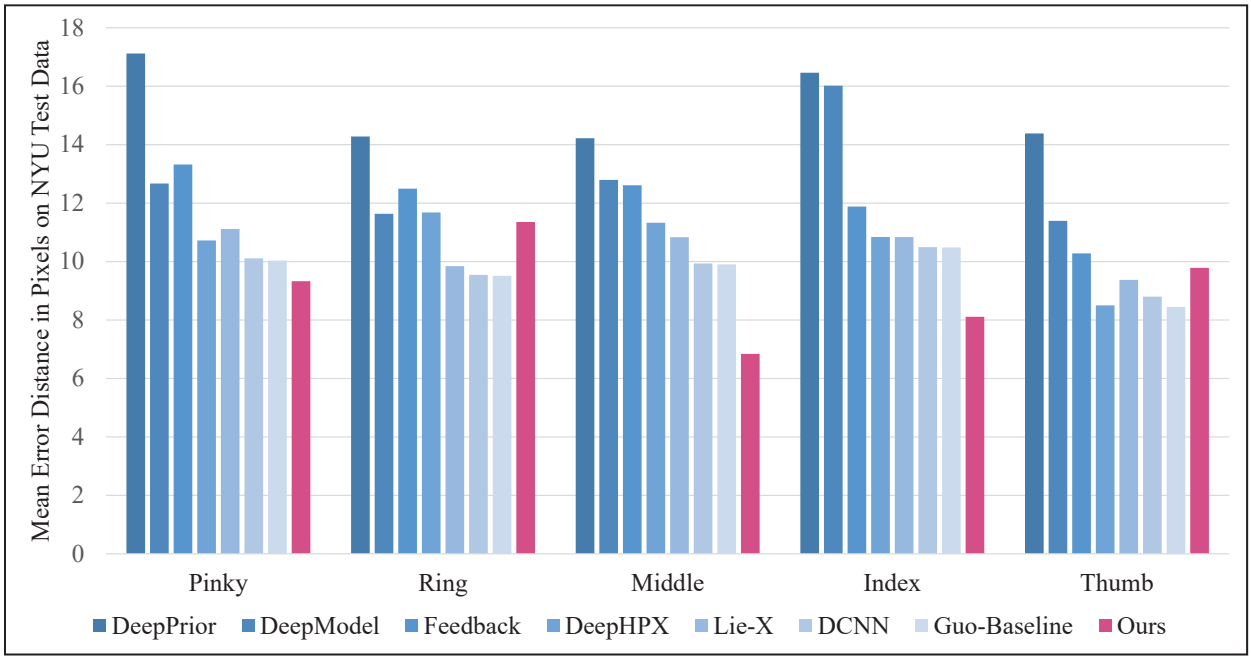

Figure 4.5. Per-joint mean error distance in pixels on NYU hand dataset..

4.4.1 Quantitative Comparison

Figure 4.5 shows the comparison results of our method with the state-of-the-art frame-based hand pose estimation method on NYU hand dataset. In all cases, we can see that our method produce fewer pixels errors on each fingertip, except ring fingertip, by enforcing temporal information.

The quantitative results of the evaluated methods are summarized in Table 4.1. It is worth mentioning that, in our proposed pipeline, besides generating high quality frames, including temporal information in video synthesis leads to improvement of 2D localization in the second stage. This ability to generalize and generate high-quality sequenced frames along with color markers from depth hand images is key to the goals of this work.

Methods	Mean Error (pixels)
Ours	9.08
Guo-baseline [102]	9.68
3DCNN [25]	9.77
Lie-X [103]	10.4
DeepHPS [104]	10.61
Feedback [105]	12.11
DeepModel [106]	12.9
DeepPrior [107]	15.29

Table 4.1. Quantitative results on NYU hand dataset

4.4.2 Qualitative Result

Figure 4.6 illustrates our generated video using the NYU dataset in the first stage which leveraged by temporal information. Experiments demonstrate that, leveraged by the spatio-temporal information in the first stage, the model is able to gener-

ate high-quality videos with precise color markers, resulting in accurate 2D keypoint localization in the second stage.

Figure 4.6. Qualitative results on the NYU hand. **First row:** shows the sequenced data in source domain. **Second row:** represent the ground truth in target domain followed by **Third row** which represents translated video obtained by our proposed pipeline. Best viewed in color with zoom..

4.5 Conclusion

In this study, we propose a general frame work for 2D fingertip localization on depth video. We transform the problem of 2D localization to the problem of video to video translation. To best of our knowledge, this is the first work on hand pose estimation on depth video. Through well-designed generators and discriminators coupled with spatio-temporal adversarial loss functions, we achieve high-resolution, temporally coherent video in the first stage. This leads to a significant improvement in performance of the localization problem in the second stage. Experimental results on the NYU benchmark dataset demonstrate that the proposed method achieves the best performance of fingertip localization on depth video in both qualitative and quantitative evaluations. Most importantly, our pipeline can be extended to estimate

depth value for 3D hand pose estimation problem. Secondly, the proposed pipeline can be applied on RGB videos as well with small changes. In other words, we only need to define colors for the region of interest such that they do not exist in background.

CHAPTER 5

HAND-REHA: Dynamic Hand Gesture Recognition for Game-Based Wrist Rehabilitation

5.1 Introduction

Rehabilitation is the recovery from the lost ability to control or coordinate a body part of the patient through a repetitive task. Based on worldwide statistics annually, more than 15 million people suffer from stroke, which is one of the main causes for people to lose their control over their body-part such as hands or feet [108]. Though physical rehabilitation can help patients regain the ability to control the affected body-parts for many decades, conventional rehabilitation exercises tend to not engage stroke patients due to tedious and repetitive nature [109]. Moreover, it is important to perform these rehabilitation exercises at home to improve their recovery.

Recently, there is a growing interest in research to utilize gesture recognition as a viable interface between humans and computers[110],[111]. Gesture-based interfaces can be an alternative to several physical peripherals that we use with our computers[112]. Additionally, a new area of research focuses on developing game-based approaches for various purposes other than entertainment or recreation. There are some works such as [113], [114], [115] and [116] that focus on using games as a mode for rehabilitation. In some of these studies they focus only on interacting with 2D games or the game is developed to perform only a specific action such as selecting an answer for a question repeatedly or set of actions that is performed repeatedly without any change in order. In this chapter, a game-based wrist rehabilitation sys-

tem was developed. It recognizes pre-defined hand gestures performed in front of a web camera. The recognized gestures are used to control the actions of an avatar in a three-dimensional maze run game. The system runs on a personal computer equipped with a web-camera, which enables patients to perform their hand rehabilitation exercises at the comfort of their home. Additionally, as the hand recognition is vision-based, no additional sensors, such as Armbands [117], special suits [116], gloves, motion sensors or depth-cameras [118][115] are needed. The system uses images from the web-camera to recognize the gestures. The raw images are first pre-processed using image processing techniques followed by a background subtraction for hand segmentation. Once the hand in the processed image is segmented, it is sent as an input to a three-layered convolutional neural network that classifies the type of performed gesture. Based on the classified gesture type, the avatar in the designed maze run game performs actions such as moving, rotating and shooting hostile drones in the 3D environment. Five different gestures were selected from a well-researched pool of medically-approved gestures suitable for wrist therapies[119]. Unlike [113] and [120], where they used gestures for simulating mouse and keyboard events for system interaction, we focused on using gestures to directly interact with the system and navigate the avatar in the 3D game. Besides, in many studies, they try to collect hand gestures through a variety of sensors, including gloves, electromagnetic or optical position and orientation sensors for the wrist. Yet wearing gloves or trackers, as well as other mentioned sensors, is uncomfortable and time-consuming approaches due to setup time or calibration process. Besides, due to our computer-vision-based interfaces, there are several notable advantages such as providing a non-intrusive approach where the hardware is commercially available at a lower cost compared to other approaches. Furthermore, by applying the background subtraction method to segment hand we were able to achieve a robust gesture classifier model that can classify the performed

gesture in a real-time manner with the highest accuracy of 98.8, which is in the same range of the state of the art methods.

The following sections of the chapter are organized as follows. Section 2 describes the background and related work is done in gesture recognition. Section 3 explains the methodology and architecture of the entire system. Section 4 describes the gesture classifier, the data-set obtained for this system, followed by the achieved result in detail. Section 5 describes the game design and the development of the game. Section 6 explains the experimental user study and the devices used for experimentation and finally the results obtained through pre and post-survey results for our experiment. Section 7 discusses the conclusion and future work to be done in this system respectively.

5.2 Background and Related Work

In recent years, as the percentage of older adults increased, the need for medical and rehabilitation dramatically raised. Furthermore, motivating game-based training improved therapy for people with physical impairments. As a result, research works that are performed in the area of developing games for rehabilitation purposes became extremely popular. Also, some research has been done in integrating gestures with a human-computer interaction system. [120] focused on integrating human gesture recognition in a human-computer interaction system by using gestures for certain mouse events such as mouse hold, mouse drag and mouse click along with certain keyboard events such as up and left keyboard press and used those gestures for certain scenarios such as playing angry birds game and working in Robot Operating System (ROS). In [121], they developed a customized augmented reality system for stroke rehabilitation.

We were inspired by the work done in [122] that implemented the use of gesture recognition for therapy. There has been compelling research on hand gestures that help in increasing a joint's range of motion or lengthen the muscle and tendons via stretching. Some of the popular motions like wrist ulnar and radial deviation gestures that are effective for improvements in hand motion are included as part of gestures in our system. This study has a focus on assisting patients to carry out rehabilitation with hand gesture recognition[115]. The process is supported by the patient playing a game on their computer to make the process interesting and make it as a rehabilitation session.

As noted in the introduction, the proposed model has two main parts for the vision-based hand gesture recognition; background subtraction along with hand detection and gesture classification. For the image processing part, we apply a background removal technique by taking a continuous average of the background at the start of the game, which then acts as a threshold throughout the game. Any object or obstacle that forms a contour in the image after the background is subtracted by the system. This detected contour is passed to our classifier which predicts one of the five possible gestures. We utilized some naive approaches explained in [123] and [124] to apply background subtraction and extract relevant information about the hand. By doing so, we were able to remove unnecessary noise from the background making it easier for the neural network to perform the classification. However, hand detection and background removal using only an RGB camera can be relatively tough depending on the lighting conditions and objects present in the background. So, we used a fixed bounded box that would enclose the person's hand. Therefore, the recognition task is done faster, processing the smaller area, which is a key factor to have a real-time hand gesture classifier. Since convolution neural network (CNN, or ConvNet)[125] gained popularity back in 2012 after the revolutionary success of AlexNet on the

public dataset ImageNet, they have come into the hype and been implemented in a lot of research areas such as classification. To classify the gestures we applied various methods to our processed datasets. Since CNNs are mostly applicable for image data, we first started with a simple 3-layer convolutional neural network in PyTorch [126] (a popular open-source deep-learning library). Besides, We implemented transfer learning with pre-trained models like ResNet50 and VGG16 [127] [128]. On comparing the results, the CNNs trained from scratch performed better than other methods for our use-case.

Finally, to evaluate our implemented system, unlike [113], [109] and [129] which only focus on evaluating their system in terms of speed and accuracy, we also examined how the user feels while performing the gestures to play the game.

5.3 Methodology

We design and develop a unique Maze run game such that the user can perform specific hand movements (according to their therapy treatment) in front of the web-camera and control the character in the game through the maze. As a result, doing the therapy through the game can motivate and interest the patient to continue their treatment without getting bored. As soon as the user starts performing the predefined gestures in front of the web-camera, the avatar in the game will perform several actions such as starting the game, moving forward, rotating right/left and shooting hostile drones in the maze. The conceptual model for our proposed method is shown in figure 5.1.

In summary, the network consists of two stages, one for hand detection and the second stage for gesture classification. Once the gesture is shown by the candidate the hand is detected and then the gesture classification process is performed on the hand image. The gesture classification is explained in detail in the next section. Once

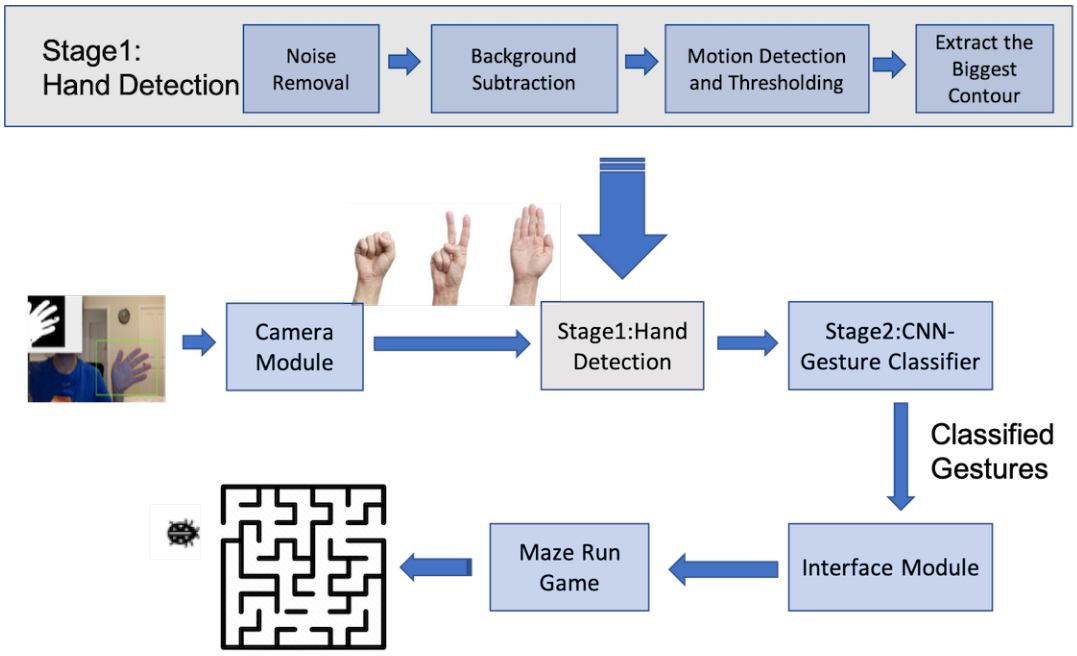

Figure 5.1. Overview of HandReha System.

the gesture classification is done, the classified gesture is sent to the interface module that interconnects the gesture classifier and the game. Based on the gesture, the corresponding action is performed by the avatar in the game.

5.4 Building the Gesture Classifier

To build a real-time gesture recognition model, we propose a two-stage model. In the first stage, the user's hand region is segmented from the static background and then, in the second stage, these segmented hand gestures will be used as input to the Convolutional Neural Network (CNN), to be classified.

5.4.1 Hand Detection and Tracking

One of the main challenges and essential processes for information extraction in many computer vision applications is the detection and tracking of the moving objects

in video streams or image sequences [130]. Generally, there are three approaches for this moving object detection task; Background subtraction, temporal referencing and optical flow [131]. Among these, we chose a background subtraction method because it is one of the most popular ones in motion object detection and it takes less computational time and space. In this method, the foreground mask for every frame is generated by subtracting the background frame from the current frame. In other words, it has two major steps. First, constructing a good statistical representation for the background which is robust to noise; second, building a statistical model for foreground object which represents the changes that take place in the current frame [132]. To build a background frame which is less affected by noise, we applied some pre-processing steps. First, all the captured frames were resized to 128 by 128, then flipped to avoid the mirroring problem. Afterward, to simplify the process and decrease the processing time, we cropped the images from the pre-defined Region Of Interest (ROI window) around the hand in the original frames. Figure 5.2 displays the output images for this step.

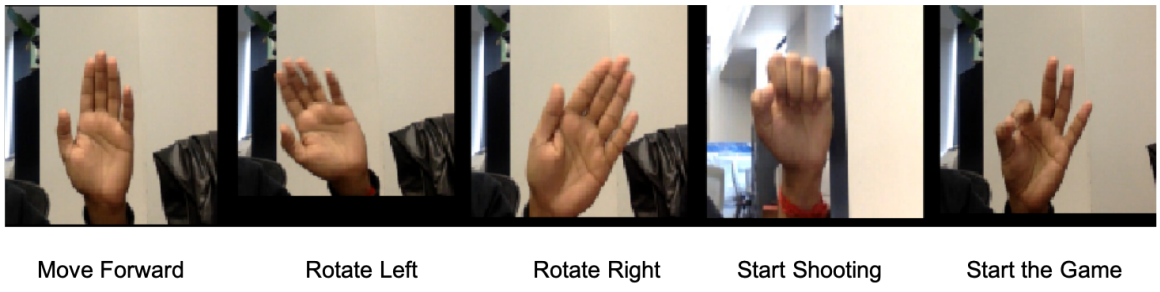

Figure 5.2. Input frames after applying resize, flip and crop operations.

Then, the cropped images are converted to gray scale to avoid the long processing time for color image analyses. The converted images are then followed by

Gaussian blurring filter[133] to remove noise. Afterward, to create a smooth background image, we calculate the average frame for the first k stationary background frames where k equals 30 and it is selected empirically(clean plates).

$$AvBg = 1/n \sum_{i=1}^{n} BG(i) \tag{5.1}$$

where n is total number of frames, $AvBg$ represents the smooth-average background image(clean plate) and $BG(i)$ is i-th background frame. As a result, noise is suppressed as much as possible and the model will become more robust to changes in lightening. Second, to build a proper representation of the foreground object, we need to subtract the background and segment the image such that it has two components; hand as foreground segment and stationary background segment. To eliminate background there are two approaches: one with a known background called a clean plate and the other one is without known background[134]. In this study, we consider the first approach as we already created a clean plate in the first step.

Then, as the candidate starts performing desired gestures the absolute value of frames and clean plates are calculated and saved as different images.

$$diff(i) = |I(i) - AvBg| \tag{5.2}$$

Where i represents the current frame with and $AvBg$ as define in equation (1) is the clean plate image as a background frame. In other words, *diff* is the absolute difference between these two frames. Afterward, we applied a threshold value to get the foreground object (hand) for the difference image.

$$segmented = \begin{cases} foreground, & \text{if } diff(x,y)(i) > \tau \\ background, & \text{otherwise} \end{cases} \tag{5.3}$$

τ is the threshold value that was selected empirically as 25. It means that all the pixels in $d$iff(i-th) frame which has a value larger than τ will be assigned a value equal to one and the remaining pixels belong to the background segment with a value equal to zero. Although the extracted hand in our case, will be segmented from a black background after applying the threshold, the final segmented hand might have some missing pixels. The resulting image is a black and white segmented image with a hand segment as white and background as black. Finally, these segmented images are resized into a fixed size keeping the same aspect of ratio and ready to be used as input to our CNN classifier in the second stage.

Figure 5.3. Output of the Hand Detection Stage.

In figure 5.3 the final processed frames are depicted. Besides, we get the contours in these segmented images and consider the contour with maximum length as hand's contours. These contours are then displayed around the user's hands in each frame to evaluate the performance of the hand detection system.

5.4.2 Hand Gesture Classifier Model

In the second stage, to build the gesture classifier model we propose a 2D Convolution Neural Network (CNN) as a gesture recognition model. The architecture of the model is depicted in figure 5.4. The CNN feeds on binary images so that the

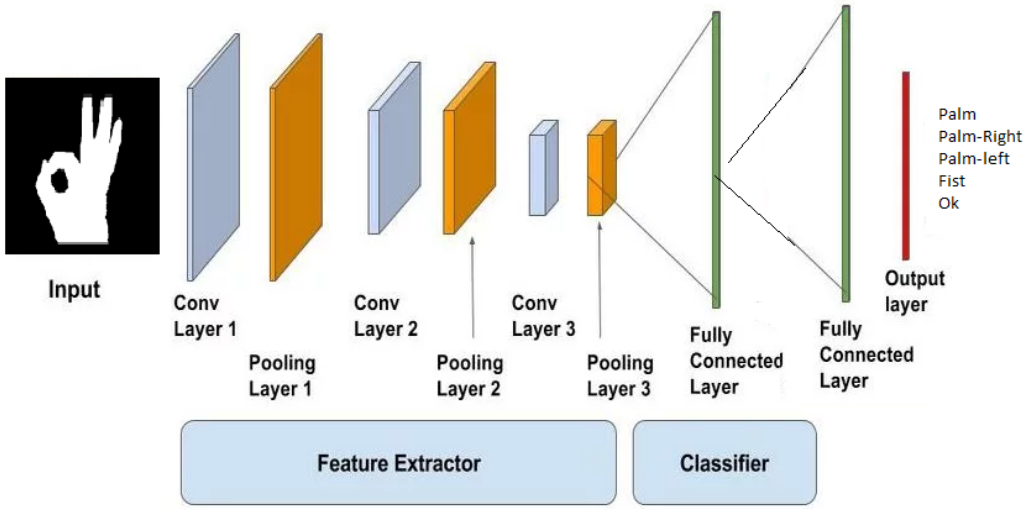

Figure 5.4. Overview of the CNN classifier.

color features do not affect the classifier. All the frames are first pre-processed, as explained in section 4.2. The pre-processing steps include resizing, converting to grayscale and applying Gaussian blurring, background subtraction, thresholding, and contour extraction. The model includes three 2D-convolutional layers for feature extraction, each followed by a max-pooling layer, two fully connected layers and a softmax layer for predicting gestures probabilities. The model classifies the gesture as the one with the highest probability. The input to this classifier is the processed frames from the hand detection stage which have the size of 128 by 128. Compared to other architecture such as RESNET50 [127]and VGG16 [128], the proposed CNN model only has 3 layers which makes the model much faster compared to above architectures.

5.4.3 HandReha Dataset

After researching on which types of gesture have been used in wrist therapy, we create our dataset of five gestures; “Fist”, “Ok”, “Open Palm”, ”PalmRight” and ”PalmLeft” that are widely used in hand therapy procedure especially for wrist hand injuries[119]. A total of 7405 images are taken from three persons (two males, one female). We ask the participants to perform the gestures at a distance of 40 to 50 centimeters from the web-camera. Each class has around 1400 images as we tried to create a balanced data set. While collecting data all the mentioned pre-processing steps are applied to the captured images and finally, the processed images are saved into the separated folders. In our designed game, we assigned different tasks to each of these gesture classes; for instance, the character in the game should go straight if the model recognizes the ”Palm” gesture or it should start shooting after recognizing ”Fist” gesture.

5.4.4 Training the CNN Gesture Classifier

For the training process, we used a HAND-REHA dataset which includes 7405 images. We will explain the reason why we choose these special gestures in the data collecting section. For each gesture, we used 0.1 percent of each class as the validation set and the remaining as the training set. The model uses input images of size 128 by 128. The loss function set as MSE and ADAM[100] as our optimizer. We trained the model for 5 epochs with processed images with various values of the learning rate, α, and realized α= 0.001 and image size as 128*128 provides the best accuracy for the classification task. Figure 5.5 and figure 5.6 show the loss and accuracy on training and validation set. The model achieves 100 and 98.8 percent accuracy for training and test set respectively.

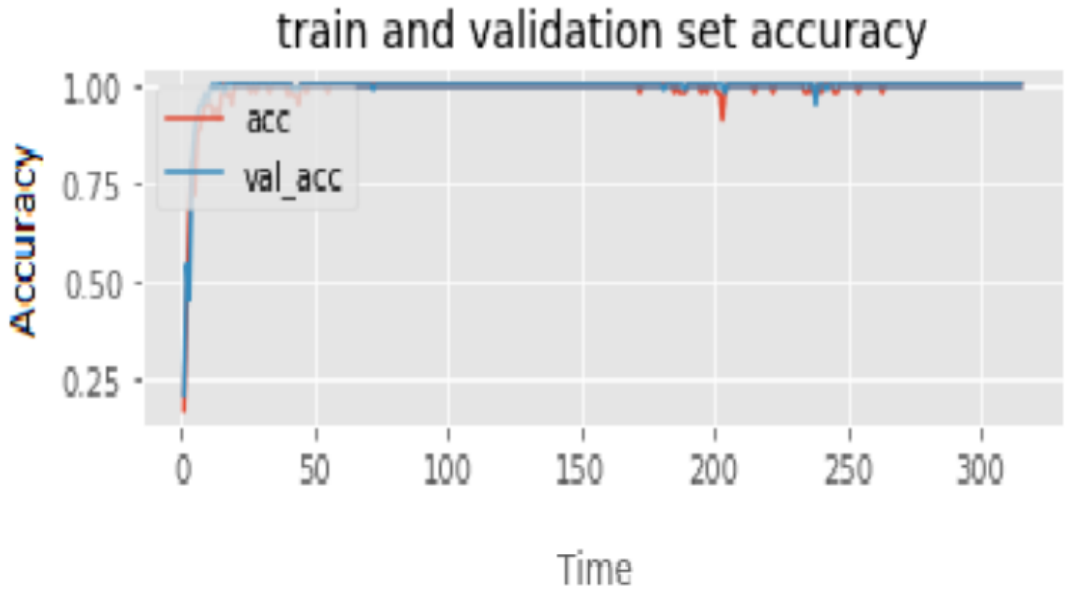

Figure 5.5. Train and validation accuracy .

5.5 Game Design and Development

We developed a 3D maze run game where an avatar navigates inside the maze surrounded by hostile drones that shoot once the avatar is near to its location. The design for the main avatar, drones and the surrounding environment was done using Blender 2.8[135]. The game engine used for developing the game is Godot 3.1[136]. Godot is a free and open-source game engine under the MIT License. The avatar, drones and several other aspects of the game designed in Blender were then imported into the Godot game engine. The stage for the game along with the avatar and drone navigation was implemented in the Godot game engine.

We built two levels in the game. The first level is a normal maze without any drones and the avatar only navigates around the maze. This level is built to familiarize the users with hand gestures to navigate the avatar around the maze. The

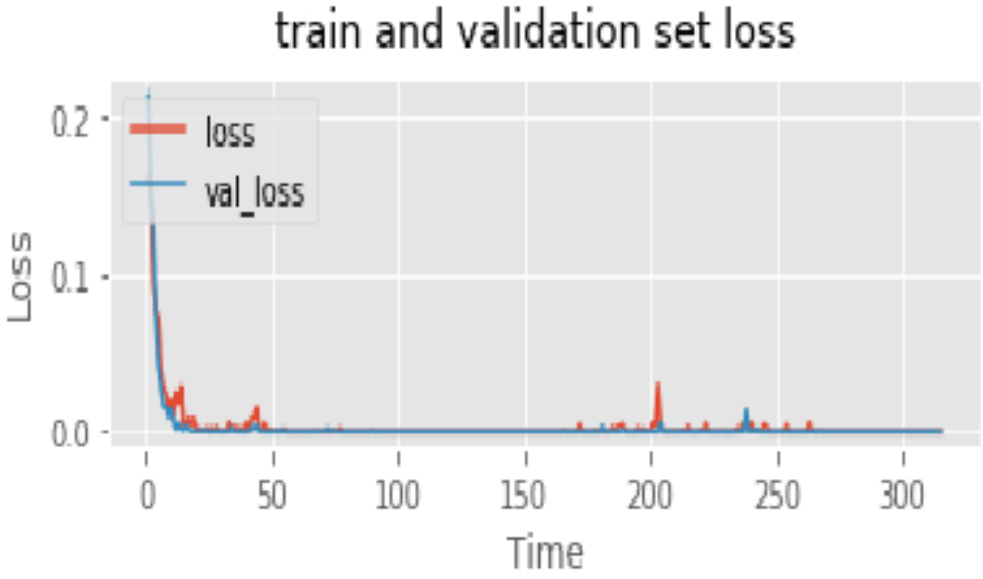

Figure 5.6. Train and validation loss .

next level is the main level where the avatar navigates around the maze while drones are present in it. The avatar has to navigate around the maze and shoot the drones when they encounter them.

Figure 5.7 and 5.8 show the navigation and shooting performed by the avatar in the game. Figure 5.9 shows the game played with gestures. You can find the video of playing the actual game using gestures at https://www.youtube.com/watch?v=V7X4CCbExmc.

5.6 Evaluation of the HAND-REHA System

Twelve healthy participants from the Computer Science and Engineering department at the University of Texas, Arlington participated in the user study to evaluate the Hand-Reha system. Of those twelve participants, six were male and six were female participants. Eight participants were aged between 25 and 30 years. Two

Figure 5.7. Navigation of avatar in the Game.

Figure 5.8. Shooting in the Game.

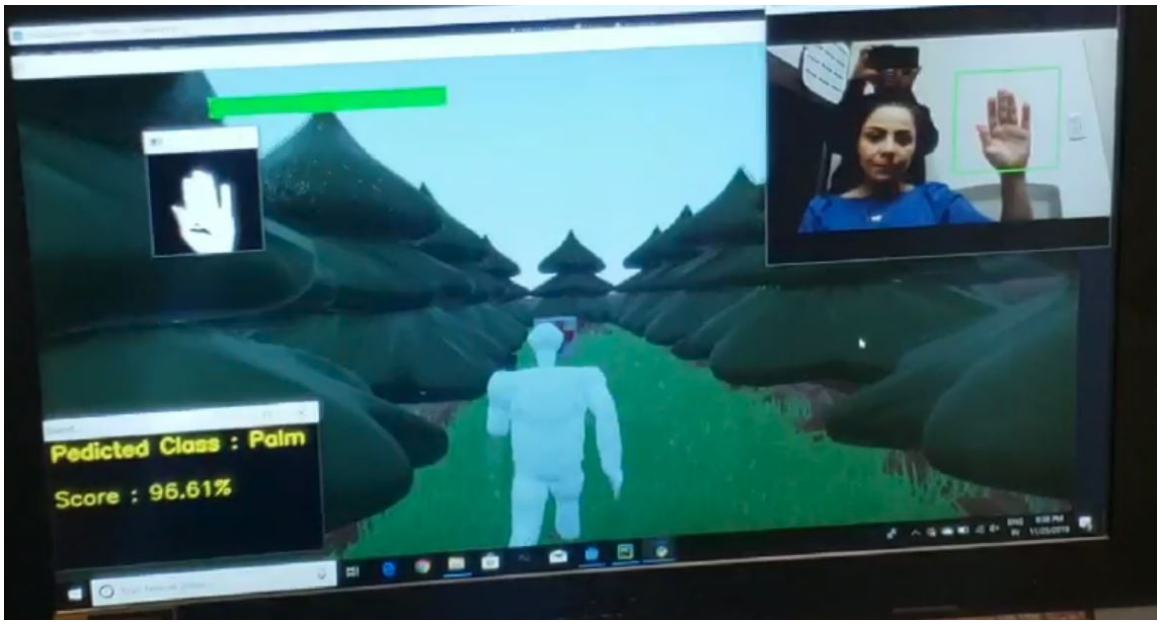

Figure 5.9. Game played with gestures.

participants were aged between 31 and 50 years. Two participants were aged between 18 and 24 years.

5.6.1 Hardware

For the experiment, we used an Acer NITRO 5 Laptop with a Windows 10 64-bit Operating system. The laptop has an 8GB RAM along with an NVIDIA Geforce GTX 1050 Ti GPU. The laptop has a built-in camera which was used for gesture recognition. The Game runs on Godot Game Engine and the gesture recognition model runs on PyCharm IDE. Both the game and the model run on the same laptop.

5.6.2 Results of Vision-Based Gesture Classification

As mentioned in the previous section the accuracy for the validation set after 5 epochs reaches 98 percent. To test and evaluate the model we tried both testings

offline and in real-time prediction. Although the model is not 100 percent accurate and has some flaws detecting gestures performed at far distance compared to training data, or in some cases the size of the hand matters(it defines how far the hand is located from the camera), our model still shows promising results while doing real-time predictions compared to other studies. Considering that we only use a shallow CNN model(only 3 layers) compared to other states of the art models and the fact that we only have around 1400 images per class will support our efficiency of the proposed model. Besides, the other issue was how perfect the participants perform the gestures, as some of them were focused on the game so they did not perform the gestures well, especially in the first couple of minutes starting to play the game. Yet considering all the mentioned reasons, we proved that our gesture classification model has acceptable performance in terms of accuracy and processing time.

5.6.3 User Study Methodology

Before the study begins, each participant filled a pre-study survey form. The form contained questions that asked whether the participant had any previous experiences of hand pain or difficulty in hand movement and their preferences in the kind of therapy if they had any such pain.

After filling the pre-study survey form, we explained the user study process of our system in detail. Afterward, the participants played the first level of the game, a plain maze where the avatar can only move around, to familiarized the participants with the gestures. Then they played the main game that included the avatar and the drones along with shooting capability for both drones and the avatar.

After playing the game the participants filled the post-study survey form. The form had questions which asked whether the participants were comfortable with using gestures while playing the game. The form was used to get feedback from the par-

In case of pain where do you prefer to receive treatment?	
Options	Number of responses
Clinic	2
Home/Private Assistant	5
Hospital	4
Rehabilitation Center	0
Any Place is fine	3

Table 5.1. Table focusing on response regarding preferred place for receiving treatment

ticipants regarding the efficiency of the system along with their opinion about using gestures for gaming and suggestions for future work.

5.6.4 Pre-study Survey Responses

In the pre-study survey form, apart from the name, age, and gender of participants we also asked a few questions regarding whether they faced any problem with hand movements and their preferences. With respect to the type of therapy, in case of any pain or difficulty in hand movement, we gave two options in the survey questionnaire: Individual therapy and Group therapy. Everyone chose Individual therapy as their preferred type of therapy.

Table 5.1 shows the response for the survey question regarding the preferred place for receiving treatment in case of any pain. For this question, 5 participants prefer treatment either at home or through a private assistant, 4 of the participants prefer to receive treatment in hospitals, 2 of the participants prefer clinic treatment and 3 of the participants were fine to receive treatment in any place.

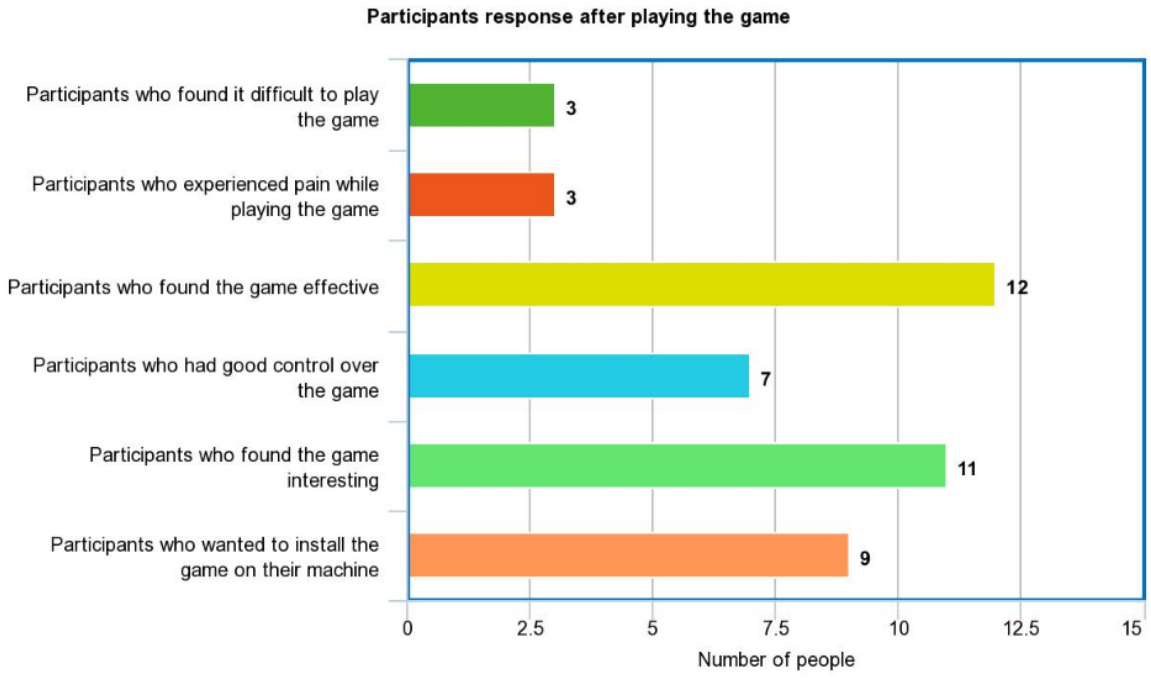

Figure 5.10. Graph representing participants feedback on HandReha.

5.6.5 Post-Study Survey Responses

After conducting the experimentation we had a post-study survey that had a set of questions regarding comfort, difficulty, effectiveness, and excitement while playing the game with hand gestures.The answers are based on the 5-point Likert scale rating with a range from 1 ("least") to 5 ("most"). As shown in Figure 5.10 and 5.11 most of the participants gave positive responses in each categories.

5.6.6 Discussion of the User Study Results

Overall most of participants reported that the hand gesture-based game was effective, exciting and comfortable. Most of the participants reported that they felt less pain and difficulty in controlling the game with gestures compared to traditional method such as using a controller. However, some participants reported that they

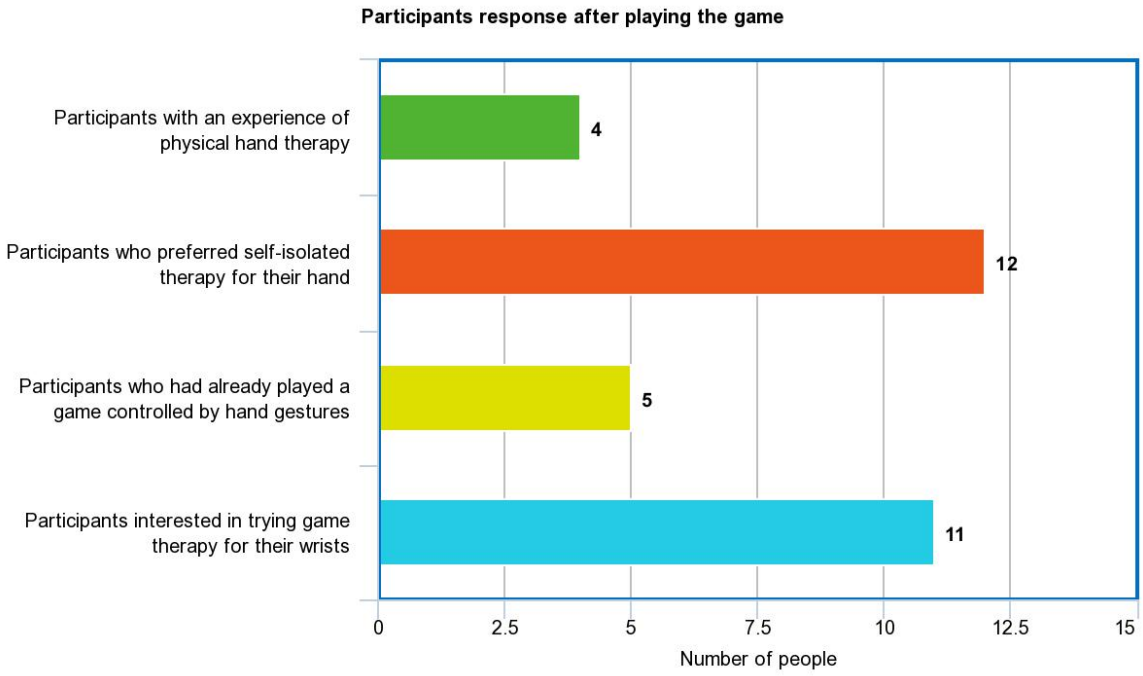

Figure 5.11. Graph indicating responses from the participants before using HandReha.

had some difficulty while playing the game due to a time lag between gesture recognition and character movement in the game which is probably because of lower RAM capacity (8 GB) on the laptop.

5.7 Conclusion and Future Work

We have designed a game-based wrist rehabilitation system, which enables the user to control an avatar in a three-dimensional maze-run game using hand gestures. This is a unique and novel approach because the gestures are selected from a set of human gestures suitable for wrist rehabilitation and implemented to control a game built in a 3D environment as compared to previous works where most of the previous game-based rehabilitation works are built in a 2D environment. Moreover, the game

is built with an avatar performing multi actions based on 5 types of gestures. This is a different from previous works where the game is built based on a single action and the gestures are implemented to simulate mouse or keyboard events.

The results from the user study showed a good and favorable outcome where almost all participants gave moderate to high ratings in terms of effectiveness and interest in playing the game with gestures. Most participants reported that they had less pain and difficulty while playing the game with gestures compared to traditional therapies.

As of future research, we aim to improve the overall efficiency, accuracy, functionality, and usability. We plan to extend the HandReha system to be compatible with other everyday devices such as smartphones and tablets. Furthermore, additional levels will be added to the game to increase user engagement and include additional gestures. Moreover, we plan to extend the HandReha dataset by including additional gestures and capturing data from a larger number of people to improve accuracy during gesture recognition. Finally, we plan to work on different ways of integrating gestures with the game.

CHAPTER 6

GAN-based Face Reconstruction for Masked-Face

6.1 Introduction

In modern technology, face recognition is becoming a new trend for the security authentication systems and human-computer interaction (HCI)[137, 138]. However, with the recent world-wide COVID-19 pandemic, the use of these face masks has raised a serious question on the accuracy of the facial recognition system. Many HCI applications based on face recognition techniques, such as face access control, and face authentication based mobile payment, have nearly failed to effectively recognize the masked faces. Moreover, touch-less verification which allows individuals to perform photo ID checks with their mask on has become extremely important in public places due to the impact of the coronavirus.

Despite the rapid growth in the amount of research works in face identification, the problem of occluded face images, including masks, has not been completely addressed due to the lack of the masked face dataset, large size and complex nature of the mask, and variation in face.

Therefore, recognizing and authenticating people wearing masks will be a long-established research area, and more efficient methods are needed for real-time face recognition. In this chapter, we are going to attempt to tackle the problem of getting rid of face-masks in facial image by using Generative Adversarial Networks (GANs). The problem we are trying to solve can be viewed as image-to-image translation, which is generally considered to be the process of translation of images from the source to the destination domain. In other words, given a masked face image, we apply unpaired

image-to-image translation [30], to remove the mask and synthesizes the affected region with fine details while retaining the global coherency of face structure. More details of the proposed model are discussed in the section 3.

The main contributions of this work are:

- Leveraged by GANs, we propose a novel approach that automatically removes mask object from face and reconstruct the affected region with delicate details.
- To overcome the data scarcity problem, we have collected a 10249 real face images of 12 people and add synthetic mask on the real faces in order to create a paired dataset of with and without mask faces.

The rest of this chapter is organized as follows. Section 2 presents related studies. The proposed model is detailed in Section 3. Sections 4 and 5 describe experimental setup and results, respectively.

6.2 Related Work

In order to remove undesired object in the images two main problem should be tackled: a) object detection and removal, b) image completion. There has been a considerable amount of learning and non-learning based object removal algorithm to tackle object removal in an image.

Recently, due to the GAN's nature of unsupervised learning, ability to generate high-quality, natural and realistic images, and the power of adversarial training, deep learning GAN-based approaches have merged as a promising paradigm for variety of application such as data augmentation [139], pose estimation [140], and image inpainting [141]. However, due to the plethora of related literature, we only review some representative works related to undesired object such as sunglasses, microphone, hand, and face masks.

Non-learning based object removal algorithms tried to solve the problem by removing the undesired object such as sunglasses, and random objects and synthesize the missing content by matching similar patches from other part of the image [142, 143]. In [144], they introduced a regularized factor to adjust the path priority function in computing function to remove eyeglasses from facial images. However, these methods can only handle relatively small holes, where the color and texture variance are small.

On the other hand, learning-based method mainly describe image inpainting with the main application of object removal and outperform the traditional methods both quantitatively and qualitatively. In [10], Iizuka et al. proposed a GAN-based model, that removes an object and reconstruct the damage part. Their proposed method, leveraged two discriminators (global and local) to ensure local and global realism of the reconstructed image. They also apply Poisson blending as a post-processing. Poisson blending technique is an image processing operator that allows the user to insert one image into another, without introducing any visually unappealing seams. Despite the ability to complete the random damaged part, this meth, is not capable to complete high resolutions images and it is resulting artifacts when damaged part is around the margin of the image. Yu et al.[11], presented a two-stage image inpainting network. First, stage includes a dilated convolutional network which is trained with reconstruction loss to rough out the missing parts. The contextual attention is integrated in the second stage to encourage spatial coherency of attention. In another study by Khan et al., [145], a coarse-to-fine GAN-based approach to remove object from facial images was introduced. For mask removal, Boutros et al. [146] introduced an embedding unmasking model which takes a feature embedding extracted from the masked face as input and generates a new feature embedding similar to an embedding of the unmasked face. Moreover, Din et al. [147, 148]

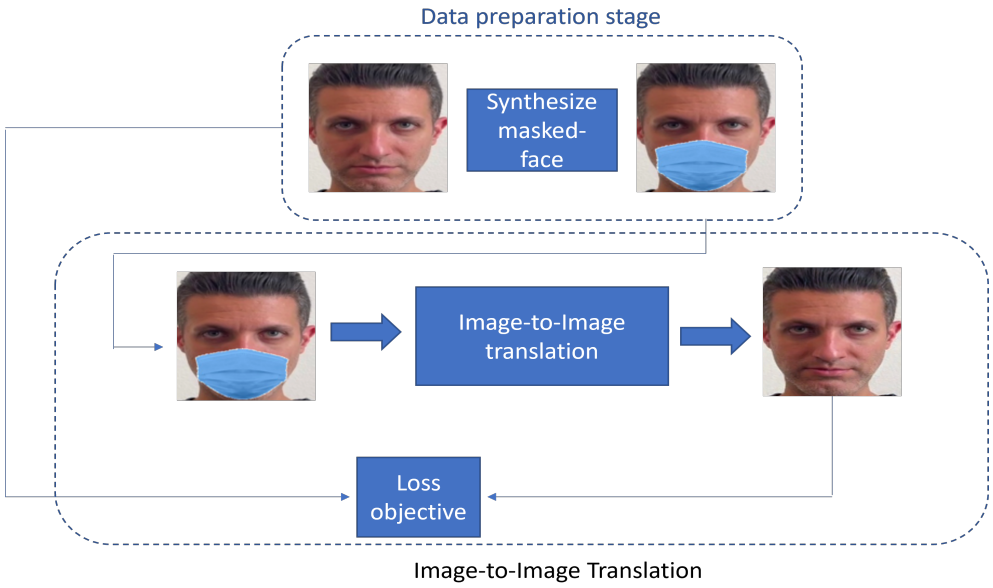

Figure 6.1. Overview of our proposed model..

used GAN-based image inpainting for image completion through an image-to-image translation approach to automatically remove face masks.

Due to the great success of learning methods to recover missing part of facial image, we proposed a novel framework which aims to automatically reconstruct hidden part of the masked-face through Image-to-Image translation, and it is able to remove masks regardless of facial angle or underlying facial expression.

6.3 Proposed Method

In this section, we provide the details of our proposed GAN-based framework that automatically removes mask and completes the missing hole through image-to-image translation so that the completed face not only looks natural and realistic but also has consistency with the rest of the image. The overall structure of our framework is illustrated in Figure 6.1.

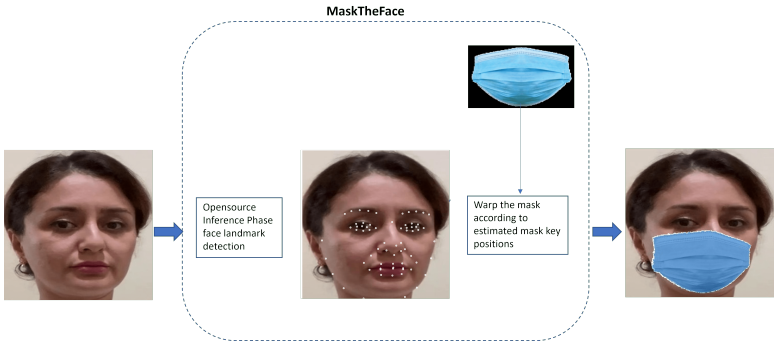

Figure 6.2. Dataset preparation: To create paired-face dataset with and without mask, "MaskTheFace" [9], tool warps the mask template based on the key face landmark positions of the face.

6.3.1 Translation using a Cycle-consistency Constraint

Unsupervised Image-to-Image Translation (UI2I) [30], composed of 2 GANs which uses two large but unpaired sets of training images to convert images from one representation to another and vice versa. The data distributions are denoted as $a \sim pdata(a)$ and $b \sim pdata(b)$. More specifically, given an input masked face image, the unpaired image-to-image translation model aims to generate a complete image without the mask using unpaired collections of facial images with and without mask. CycleGAN loss primarily consists of adversarial loss [28] and cycle-consistency loss. Adversarial loss, in GAN, enforce the generated image to be indistinguishable from real photos. While generator, G, tries to find the mapping $G : A \rightarrow B$, its discriminator's D_b objective function is defined as:

$$\begin{aligned} \mathcal{L}_{GAN}(G, D_B, A, B) = \mathbb{E}_{b \sim pdata(b)}[logD_B(b)] \\ + \mathbb{E}_{a \sim pdata(a)}[log(1 - D_B(G(a))], \end{aligned} \tag{6.1}$$

where, G generates images $G(a)$ that appears like images from field B and D_b observes between translated samples $G(a)$ and original samples b. A similar adver-

sarial loss is postulated for the second generator, $F : B \rightarrow A$ and its discriminator D_a.

However, the adversarial loss alone is not sufficient to produce good images, as it leaves the model under-constrained. Adversarial loss, enforces the generated output be of the appropriate domain but does not enforce that the input and output are recognizably the same. The cycle consistency loss addresses this issue. It relies on the expectation that if you convert an image to the other domain and back again, by successively feeding it through both generators, you should get back something similar to what you put in. In other words, it compares the reconstructed image and input image using L1-norm distance and enforces that $F\big(G(a)\big) \approx a$ and $G\big(F(b)\big) \approx b$.

$$\begin{aligned} \mathcal{L}_{cyc}\left(G, F\right) = \mathbb{E}_{a \sim p_{data}(a)}\left[||F\left(G\left(a\right)\right) - a||_1\right] + \\ \mathbb{E}_{b \sim p_{data}(b)}\left[||G\left(F\left(b\right)\right) - b||_1\right] \end{aligned} \tag{6.2}$$

The full loss with cycling parameter *lambda* is:

$\mathcal{L} = \mathcal{L}_{GAN}(G, D_B, A, B) + \mathcal{L}_{GAN}(F, D_A, B, A) + \lambda \mathcal{L}_{cyc}(G, F)$ which is used for training the model with an Adam optimizer [149]. *lambda* controls the relative importance of the two objectives.

6.4 Experimental Details

6.4.1 Data Preparation

To address the lack of the masked face dataset, this study firstly contributes a new dataset of high-quality, paired face images with and without mask simultaneously which in real world is not possible. To this end, we propose to superimpose artificial face masks onto real face images based on key face landmarks position, through an open-source masking tool "MaskTheFace" [9], as shown in Figure 6.2. To this

end, first, we collected 10249 high quality face images from 12 people (10 user for training and 2 users for testing) and all faces in the images were detected by YOLO algorithm[150]. Then, we used the dlib library-based face landmarks detector to identify the face tilt and six key features (eyes, nose, lips, face edges etc.,) of the face necessary for applying mask. The template mask is then transformed based on the six key features to fit perfectly on the face. This results in creation of a large, paired dataset of face images with and without masks which can be used as real world masked-face test data and the ground truth data without mask.

6.4.2 Implementation Details

A CycleGAN model was trained to unmask the masked-face. We adapt the architecture for our generative networks from [30], it utilizes two parts Generators and Discriminators. Each generator composed of three initial convolutional, nine 64-channel convolutional ResNET block, two fractionally strided convolutions, and a final convolution to reduce the output's channel. Furthermore, each 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 150 epochs with 8131 unpaired facial images of size 256x256 with and without mask with learning rate of 0.0002 and lambda value of 10 to calculate cycle loss. Once the model is trained, we evaluate it using 2118 masked face test images of from our created dataset.

6.4.3 Evaluation Metric

We compared the results between our method and the other method using the Structural SIMilarity (SSIM) quantitative metrics [151]. However, as reported by many other works [152] and [153], we argue that quantitative analysis may not be the most effective measure of the image editing task.

Figure 6.3. Output examples generated by our model for test samples of our created dataset. **First column**, masked face image in source domain, **second column**, generated unmasked face in target domain and, **third column**, ground truth unmasked face in target domain..

6.5 Completion Results

We now discuss the qualitative and quantitative performance of our method and its comparison with other previous state-of-the-art image manipulation methods on real world images with mask.

6.5.1 Qualitative Comparison

Figure 6.3 shows the sample generated by our model for masked face test images. We also present a qualitative comparison with Iizuka et al. [10] and Yu et al. [11] on real world test images as can be seen in Figure 6.4. Although, proposed model by Iizuka et al. [10], completes the image for random damaged region in facial images, it is limited to relatively low resolutions (178x218) and produces artifacts when damaged part is at the margins of an image. Moreover, Yu et al. [11] reduces the artifacts at margins but is unable to recover a complex face structure. Moreover, although in each test image, almost half of key facial semantics are covered by face mask, our model offers significantly improved results for real world data than the other previous state-of-the-art image manipulation methods and successfully removes the mask object and generates natural looking outputs with structural consistency.

Table 6.1. Performance comparison in term of Structural SIMilarity (SSIM).

Methods	SSIM
Yu et al.[11]	0.86
Ours	**0.89**

Figure 6.4. Visual comparison of our proposed method with representative image completion methods on real world images. From left to right: Input image, [10], [11], and ours. **Note:** There is no ground truth since all samples are real world images collected from the Internet. .

6.5.2 Quantitative Comparison

To have a fair comparison, we have created a synthetic dataset of 6446 images using the publicly available, CelebA-HQ [154], celebrity face dataset and trained Yu et al. [11] and our model. Then, we evaluate model performance and training effectiveness by Structural SIMilarity (SSIM) [151]. It is a full reference metric that requires two images from the same image capture and it measures the perceptual difference between two similar images. Since real images with mask do not have corresponding ground truth without mask object, we have evaluated SSIM on 2459 Synthetic test data created using CelebA-HQ. Table 6.1 provides a quantitative comparison with Yu et al. [11].

6.6 Conclusion

Partially concealed faces by mask in situation like pandemics, or air pollution has exerted dramatic influences and reduce the performance of existing security and authentication systems due to the absence of large-scale training data and the presence of large intra-class variation between masked faces and full faces. This imposes the demand to tackle such authentication concerns using more robust and reliable facial recognition systems under different settings.

To this end, we proposed a novel method for interaction-free mask removal from facial images. The hidden parts of the face are regenerated in the most realistic way by GAN-based image-to-image translation. Our proposed pipeline could be involved in various areas such as criminal face recognition, and secure authentication. Moreover, due to the lack of public datasets containing real masked face images, we create a high-quality paired dataset of real faces along with their simulated masked one by placing synthetic masks over the real face images for training. To the best of our knowledge, this is the first effort to create high-quality, well-established face benchmarks paired dataset of face images with and without masks. The proposed dataset is part of an ongoing effort to gather a larger scale database with realistic variations of masks and will be available upon request. Moreover, both qualitative and quantitative comparison show that our model demonstrates superior performance for large facial object (face mask) as compared to the state-of-the art.

Several future steps could be taken to improve the results as well as put the trained model into practical usage. First, we plan to collect and expand our masked-face dataset to improve our face reconstruction model. We also plan to develop a user-friendly interface for unmasking the masked-face. Furthermore, future research will continue to leverage these reconstructed face images in state-of-the-art face recognition models in automobile security, secure authentication, and access control.

CHAPTER 7

Conclusions and Future Work

Reliable estimation of hand pose using computer vision remains an elusive goal, due to the difficulty and extent of effort involved in obtaining large amounts of training data. Getting such data involves manually specifying hand pose information for thousands or millions of images, and this has been a big bottleneck for progress on this topic. However, with semi-supervised learning, the dataset may contain millions of images, but we only need to specify hand pose information for a very small fraction of those images. We conclude this dissertation with summarizing the work and contributions therein. Discussions of future work are included as these challenges are ongoing and require further investigation before robust and accessible solutions can be provided.

This dissertation investigated semi-supervised learning for hand pose estimation and, face reconstruction problem through generative adversarial networks in many perspectives:

(a) a comprehensive survey of available hand pose estimation works through generative adversarial networks is presented. (b) a semi-supervised GAN-based approach for 2D hand pose estimation is presented. (c) a paired domain translation approach is proposed for hand pose estimation on depth video leveraged by temporal information. (d) as an example application, we designed a game-based wrist rehabilitation system, which enables the user to control an avatar in a three-dimensional maze-run game using hand gestures. (e) finally, a general domain translation frame-

work that can be used to reconstruct the hidden part of face concealed by mask is proposed. The contributions in this work are as follow:

1. **A survey of all existing hand pose estimation studies through generative adversarial networks is presented.** We present a comprehensive study on effective hand pose estimation approaches, which are comprised of the leveraged generative adversarial network (GAN), providing a comprehensive training dataset with different modalities. Benefiting from GAN, these algorithms can augment data to a variety of hand shapes and poses where data manipulation is intuitively controlled and greatly realistic.
2. **New public depth fingertip dataset of 10K is created.** A comprehensive depth hand images are collected by Microsoft Kinect V2 camera. The collected dataset is then annotated for 7 joint with implemented GUI in Matlab 2020a. The GUI is designed to read all image formats and it supports depth image illustration to ease the annotation of depth maps.
3. **Introduced a novel semi-supervised frame work for 2D hand pose estimation.** To avoid using large labeled dataset and avoid the annotation difficulties, we propose 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.
4. **Leveraged by temporal information and domain translation, we addressed 2D hand pose estimation on depth videos.** We transform the problem of 2D localization to the problem of video-to-video translation. To

the best of our knowledge, this is the first work on hand pose estimation on depth video. Through well-designed generators and discriminators coupled with spatio-temporal adversarial loss functions, we achieve high resolution, temporally coherent video in the first stage. This leads to a significant improvement in performance of the localization problem in the second stage. Experimental results on the NYU benchmark dataset demonstrate that the proposed method achieves the best performance of fingertip localization on depth video in both qualitative and quantitative evaluations.

5. **A novel 3D game-based gesture recognition is developed for wrist rehabilitation.** We aim to deign and develop an effective, exciting and easy to access 3D game for wrist rehabilitation at the comfort of patient's home. The idea is to automatically recognize pre-defined hand gestures using a web-camera, so to control an avatar in a three dimensional maze run game. The pre-defined gestures are picked from a pool of well-defined gestures suitable for wrist rehabilitation. Deep learning techniques were utilized to perform real-time hand gesture recognition from the images. The user study showed that the developed wrist rehabilitation system is intuitive and engages the user, which is crucial for rehabilitation purposes.
6. **A new paired-synthetic dataset of face with and without mask is created.** Due to the lack of public datasets containing real masked face images, we create a high-quality paired dataset of real faces along with their simulated masked one by placing synthetic masks over the real face images for training. To the best of our knowledge, this is the first effort to create high-quality, well-established face benchmarks paired dataset of face images with and without masks.

7. **Realistic face reconstruction approach based on domain translation is proposed.** The hidden part of the face are reconstructed in the most realistic way by GAN-based image-to-image translation. They can be used for facial identification and secure authentication in human-computer interaction. The obtained results demonstrate that our model outperforms other representative state-of-the-art face completion approaches both qualitatively and quantitatively

7.1 Future Works

- **Explore good evaluation metrics for GAN's evaluation.** Due to the lack of robust and consistent metrics, coming up with good evaluation metric is still an open challenge to compare different GAN variants based on the visual assessment of the generated images.
- **Extend and improve the 2D hand pose estimation pipeline to be applied on RGB hand dataset in our next study.** The proposed pipeline can be applied to RGB input images as well with small changes. In other words, we only need to define colors for the region of interest such that they do not exist in the background.
- **Extend the 2D hand pose estimation pipeline for 3D hand pose estimation.** To this end, the estimated 2D joints along with their depth value can be used for 3D pose estimation.
- **Enrich the synthesized dataset for hand pose with different modalities using GANs.** Apply the ability of GAN's to generate realistic samples for situation where data collection is difficult or impossible.
- Develop a user friendly interface for unmasking the masked-face in the future. the proposed pipeline can be used in user-friendly interface for unmasking the

masked-faces. Moreover, the reconstructed face from the proposed pipeline can be used in facial recognition application for secure authentication, criminal face recognition,etc.

REFERENCES

[1] W. He, Z. Xie, Y. Li, X. Wang, and W. Cai, "Synthesizing depth hand images with gans and style transfer for hand pose estimation," *Sensors*, vol. 19, no. 13, p. 2919, 2019.

[2] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2107–2116.

[3] L. Chen, S.-Y. Lin, Y. Xie, H. Tang, Y. Xue, Y.-Y. Lin, X. Xie, and W. Fan, "Tagan: Tonality aligned generative adversarial networks for realistic hand pose synthesis," in *BMVC*, 2019.

[4] Z. Wu, D. Hoang, S.-Y. Lin, Y. Xie, L. Chen, Y.-Y. Lin, Z. Wang, and W. Fan, "Mm-hand: 3d-aware multi-modal guided hand generative network for 3d hand pose synthesis," *arXiv preprint arXiv:2010.01158*, 2020.

[5] G. Park, T.-K. Kim, and W. Woo, "3d hand pose estimation with a single infrared camera via domain transfer learning," in *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2020, pp. 588–599.

[6] S. Baek, K. I. Kim, and T.-K. Kim, "Augmented skeleton space transfer for depth-based hand pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8330–8339.

[7] L. Chen, S.-Y. Lin, Y. Xie, Y.-Y. Lin, W. Fan, and X. Xie, "Dggan: Depth-image guided generative adversarial networks for disentangling rgb and depth

images in 3d hand pose estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 411–419.

[8] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," *arXiv preprint arXiv:2008.11104*, 2020.

[10] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–14, 2017.

[11] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514.

[12] P. Krejov, A. Gilbert, and R. Bowden, "Guided optimisation through classification and regression for hand pose estimation," *Computer Vision and Image Understanding*, vol. 155, pp. 124–138, 2017.

[13] Y. Zhou, G. Jiang, and Y. Lin, "A novel finger and hand pose estimation technique for real-time hand gesture recognition," *Pattern Recognition*, vol. 49, pp. 102–114, 2016.

[14] M. Murugeswari and S. Veluchamy, "Hand gesture recognition system for real-time application," in *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*. IEEE, 2014, pp. 1220–1225.

[15] C. Carley and C. Tomasi, "Single-frame indexing for 3d hand pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 101–109.

[16] J. Isaacs and S. Foo, "Optimized wavelet hand pose estimation for american sign language recognition," in *Proceedings of the 2004 congress on evolutionary computation (IEEE Cat. No. 04TH8753)*, vol. 1. IEEE, 2004, pp. 797–802.

[17] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3413–3423.

[18] S. Bilal, R. Akmeliawati, M. J. El Salami, and A. A. Shafie, "Vision-based hand posture detection and recognition for sign language—a study," in *2011 4th International Conference on Mechatronics (ICOM)*. IEEE, 2011, pp. 1–6.

[19] F. Kirac, Y. E. Kara, and L. Akarun, "Hierarchically constrained 3d hand pose estimation using regression forests from single frame depth data," *Pattern Recognition Letters*, vol. 50, pp. 91–100, 2014.

[20] J. Taylor, L. Bordeaux, T. Cashman, B. Corish, C. Keskin, T. Sharp, E. Soto, D. Sweeney, J. Valentin, B. Luff, *et al.*, "Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–12, 2016.

[21] H. Liang, J. Wang, Q. Sun, Y.-J. Liu, J. Yuan, J. Luo, and Y. He, "Barehanded music: real-time hand interaction for virtual piano," in *Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 2016, pp. 87–94.

[22] Y. Zhang and O. Meruvia-Pastor, "Operating virtual panels with hand gestures in immersive vr games," in *International conference on augmented reality, virtual reality and computer graphics*. Springer, 2017, pp. 299–308.

[23] H. Liang, J. Yuan, J. Lee, L. Ge, and D. Thalmann, "Hough forest with optimized leaves for global hand pose estimation with arbitrary postures," *IEEE Transactions on Cybernetics*, vol. 49, no. 2, pp. 527–541, 2017.

[24] R. Wang, S. Paris, and J. Popović, "6d hands: markerless hand-tracking for computer aided design," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 549–558.

[25] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "3d convolutional neural networks for efficient and robust hand pose estimation from single depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1991–2000.

[26] B. Doosti, S. Naha, M. Mirbagheri, and D. J. Crandall, "Hope-net: A graph-based model for hand-object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6608–6617.

[27] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid, "Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 571–580.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[29] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[30] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[31] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[32] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1747–1756.

[33] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 1857–1865.

[34] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Y. Chang, K. M. Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge, *et al.*, "Depth-based 3d hand pose estimation: From current achievements to future goals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2636–2645.

[35] J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-based hand pose estimation: data, methods, and challenges," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1868–1876.

[36] 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)*, vol. 33, no. 5, pp. 1–10, 2014.

[37] D. Tang, H. Jin Chang, A. Tejani, and T.-K. Kim, "Latent regression forest: Structured estimation of 3d articulated hand posture," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3786–3793.

[38] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "Freihand: A dataset for markerless capture of hand pose and shape from single rgb images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 813–822.

[39] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[40] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "Crdoco: Pixel-level domain transfer with cross-domain consistency," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1791–1800.

[41] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 824–832.

[42] M. Haiderbhai, S. Ledesma, S. C. Lee, M. Seibold, P. Fürnstahl, N. Navab, and P. Fallavollita, "pix2xray: converting rgb images into x-rays using generative adversarial networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 6, pp. 973–980, 2020.

[43] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 49–59.

[44] M. Qi, E. Remelli, M. Salzmann, and P. Fua, "Unsupervised domain adaptation with temporal-consistent self-training for 3d hand-object joint reconstruction," *arXiv preprint arXiv:2012.11260*, 2020.

[45] X. Chen, G. Wang, H. Guo, and C. Zhang, "Pose guided structured region ensemble network for cascaded hand pose estimation," *Neurocomputing*, vol. 395, pp. 138–149, 2020.

[46] C. Wan, T. Probst, L. Van Gool, and A. Yao, "Crossing nets: Dual generative models with a shared latent space for hand pose estimation," in *Conference on Computer Vision and Pattern Recognition*, vol. 7, 2017.

[47] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Quan Yuan, and A. Thangali, "The american sign language lexicon video dataset," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008, pp. 1–8.

[48] F. Farahanipad, H. R. Nambiappan, A. Jaiswal, M. Kyrarini, and F. Makedon, "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*, 2020, pp. 1–9.

[49] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4903–4911.

[50] P. Panteleris, I. Oikonomidis, and A. Argyros, "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, 2018, pp. 436–445.

[51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[52] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 977–11 986.

[53] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*. Ieee, 2011, pp. 1297–1304.

[54] A. Wetzler, R. Slossberg, and R. Kimmel, "Rule of thumb: Deep derotation for improved fingertip detection," *arXiv preprint arXiv:1507.05726*, 2015.

[55] J. Thewlis, S. Albanie, H. Bilen, and A. Vedaldi, "Unsupervised learning of landmarks by descriptor vector exchange," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6361–6371.

[56] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi, "Unsupervised learning of object landmarks through conditional image generation," in *Advances in neural information processing systems*, 2018, pp. 4016–4027.

[57] X. Dong and Y. Yang, "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)*, October 2019.

[58] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and robust hand tracking from depth," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1106–1113.

[59] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Real time hand pose estimation using depth sensors," in *Consumer depth cameras for computer vision*. Springer, 2013, pp. 119–137.

[60] J. Malik, I. Abdelaziz, A. Elhayek, S. Shimada, S. A. Ali, V. Golyanik, C. Theobalt, and D. Stricker, "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*, 2020, pp. 7113–7122.

[61] M. Rad, M. Oberweger, and V. Lepetit, "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*, 2018, pp. 4663–4672.

[62] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan, "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*, 2019, pp. 793–802.

[63] S. Malassiotis and M. G. Strintzis, "Real-time hand posture recognition using range data," *Image and Vision Computing*, vol. 26, no. 7, pp. 1027–1037, 2008.

[64] P. Suryanarayan, A. Subramanian, and D. Mandalapu, "Dynamic hand pose recognition using depth data," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3105–3108.

[65] A. Sinha, C. Choi, and K. Ramani, "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*, 2016, pp. 4150–4158.

[66] S. Baek, K. I. Kim, and T.-K. Kim, "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*, 2020, pp. 6121–6131.

[67] M. Abdi, E. Abbasnejad, C. P. Lim, and S. Nahavandi, "3d hand pose estimation using simulation and partial-supervision with a shared latent space," *arXiv preprint arXiv:1807.05380*, 2018.

[68] L. Chen, S.-Y. Lin, Y. Xie, H. Tang, Y. Xue, X. Xie, Y.-Y. Lin, and W. Fan, "Generating realistic training images based on tonality-alignment generative adversarial networks for hand pose estimation," *arXiv preprint arXiv:1811.09916*, 2018.

[69] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

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

[71] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba, "Cross-modal scene networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2303–2314, 2017.

[72] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," *Advances in neural information processing systems*, vol. 29, pp. 469–477, 2016.

[73] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016.

[74] S. Baek, K. I. Kim, and T.-K. Kim, "Augmented skeleton space transfer for depth-based hand pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[75] A. K. Mondal, A. Agarwal, J. Dolz, and C. Desrosiers, "Revisiting cyclegan for semi-supervised segmentation," *arXiv preprint arXiv:1908.11569*, 2019.

[76] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 89–98.

[77] S. Sural, G. Qian, and S. Pramanik, "Segmentation and histogram generation using the hsv color space for image retrieval," in *Proceedings. International Conference on Image Processing*, vol. 2. IEEE, 2002, pp. II–II.

[78] L. Duan, M. Shen, S. Cui, Z. Guo, and O. Deussen, "Estimating 2d multi-hand poses from single depth images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[79] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in

Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798–8807.

[80] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," *arXiv preprint arXiv:1808.06601*, 2018.

[81] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *German conference on pattern recognition*. Springer, 2016, pp. 26–36.

[82] H. Wang, M. Huang, D. Wu, Y. Li, and W. Zhang, "Supervised video-to-video synthesis for single human pose transfer," *IEEE Access*, vol. 9, pp. 17 544–17 556, 2021.

[83] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt, "Real-time hand tracking under occlusion from an egocentric rgb-d sensor," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1154–1163.

[84] P. Panteleris and A. Argyros, "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 (ICCV) Workshops*, Oct 2017.

[85] C. Choi, A. Sinha, J. H. Choi, S. Jang, and K. Ramani, "A collaborative filtering approach to real-time hand pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2336–2344.

[86] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[87] C. Wan, T. Probst, L. Van Gool, and A. Yao, "Dense 3d regression for hand pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5147–5156.

[88] L. Ge, Z. Ren, and J. Yuan, "Point-to-point regression pointnet for 3d hand pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 475–491.

[89] X. Wu, D. Finnegan, E. O'Neill, and Y.-L. Yang, "Handmap: Robust hand pose estimation via intermediate dense guidance map supervision," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 237–253.

[90] Y. Cai, L. Ge, J. Cai, and J. Yuan, "Weakly-supervised 3d hand pose estimation from monocular rgb images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 666–682.

[91] L. Chen, S.-Y. Lin, Y. Xie, Y.-Y. Lin, and X. Xie, "Temporal-aware self-supervised learning for 3d hand pose and mesh estimation in videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1050–1059.

[92] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with rgb-d videos and 3d hand pose annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 409–419.

[93] M. de La Gorce, D. J. Fleet, and N. Paragios, "Model-based 3d hand pose estimation from monocular video," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 9, pp. 1793–1805, 2011.

[94] N. Louis, L. Zhou, S. J. Yule, R. D. Dias, M. Manojlovich, F. D. Pagani, D. S. Likosky, and J. J. Corso, "Temporally guided articulated hand pose tracking in surgical videos," *arXiv preprint arXiv:2101.04281*, 2021.

[95] M. Rad, M. Oberweger, and V. Lepetit, "Domain transfer for 3d pose estimation from color images without manual annotations," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 69–84.

[96] L. Yang, S. Li, D. Lee, and A. Yao, "Aligning latent spaces for 3d hand pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2335–2343.

[97] F. Farahanipad, M. Rezaei, A. Dillhoff, F. Kamangar, and V. Athitsos, "A pipeline for hand 2-d keypoint localization using unpaired image to image translation," in *The 14th PErvasive Technologies Related to Assistive Environments Conference*, ser. PETRA 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 226–233. [Online]. Available: https://doi.org/10.1145/3453892.3453904

[98] S. Yuan, B. Stenger, and T.-K. Kim, "Rgb-based 3d hand pose estimation via privileged learning with depth images," *arXiv preprint arXiv:1811.07376*, 2018.

[99] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[100] D. Kingman and J. Ba, "Adam: A method for stochastic optimization. conference paper," in *3rd International Conference for Learning Representations*, 2015.

[101] O. Iason, K. Nikolaos, and A. Antonis, "Efficient model-based 3d tracking of hand articulations using kinect," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2011, pp. 101.1–101.11, http://dx.doi.org/10.5244/C.25.101.

[102] H. Guo, G. Wang, X. Chen, C. Zhang, F. Qiao, and H. Yang, "Region ensemble network: Improving convolutional network for hand pose estimation," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4512–4516.

[103] C. Xu, L. N. Govindarajan, Y. Zhang, and L. Cheng, "Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups," *International Journal of Computer Vision*, vol. 123, no. 3, pp. 454–478, 2017.

[104] J. Malik, A. Elhayek, F. Nunnari, K. Varanasi, K. Tamaddon, A. Heloir, and D. Stricker, "Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 110–119.

[105] M. Oberweger, P. Wohlhart, and V. Lepetit, "Training a feedback loop for hand pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3316–3324.

[106] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei, "Model-based deep hand pose estimation," *arXiv preprint arXiv:1606.06854*, 2016.

[107] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands deep in deep learning for hand pose estimation," *arXiv preprint arXiv:1502.06807*, 2015.

[108] "The internet stroke center," 2008. [Online]. Available: http://www.strokecenter.org/patients/about-stroke/stroke-statistics/

[109] Y.-X. Hung, P.-C. Huang, K.-T. Chen, and W.-C. Chu, "What do stroke patients look for in game-based rehabilitation: a survey study," *Medicine*, vol. 95, no. 11, 2016.

[110] S. Escalera, V. Athitsos, and I. Guyon, "Challenges in multi-modal gesture recognition," in *Gesture Recognition*. Springer, 2017, pp. 1–60.

[111] S. Gieser, A. Boisselle, and F. Makedon, "Real-time static gesture recognition for upper extremity rehabilitation using the leap motion," 07 2015, pp. 144–154.

[112] A. R. Babu, M. Zakizadeh, J. R. Brady, D. Calderon, and F. Makedon, "An intelligent action recognition system to assess cognitive behavior for executive

function disorder," in *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2019, pp. 164–169.

[113] Chiang Wei Tan, Siew Wen Chin, and Wai Xiang Lim, "Game-based human computer interaction using gesture recognition for rehabilitation," in *2013 IEEE International Conference on Control System, Computing and Engineering*, Nov 2013, pp. 344–349.

[114] E. R. Ramírez, R. Petrie, K. Chan, and N. Signal, "A tangible interface and augmented reality game for facilitating sit-to-stand exercises for stroke rehabilitation," in *Proceedings of the 8th International Conference on the Internet of Things*, 2018, pp. 1–4.

[115] R. Proffitt, M. Sevick, C.-Y. Chang, and B. Lange, "User-centered design of a controller-free game for hand rehabilitation," *Games for health journal*, vol. 4, no. 4, pp. 259–264, 2015.

[116] H.-Y. Chen, T.-Y. Lin, L.-Y. Huang, A.-C. Chen, Y.-C. Zheng, H.-M. Wang, S.-Y. Wei, and Y.-Y. Chou, "Hp2: Using machine learning model to play serious game with imu smart suit," in *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, ser. MUM 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 397–402. [Online]. Available: https://doi.org/10.1145/3282894.3289731

[117] S.-P. Lai, C.-A. Hsieh, T. Harutaipree, S.-C. Lin, Y.-H. Peng, L.-P. Cheng, and M. Y. Chen, "Fitbird: Improving free-weight training experience using wearable sensors for game control," in *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, ser. CHI PLAY '19 Extended Abstracts. New York, NY, USA: Association for Computing Machinery, 2019, p. 475–481. [Online]. Available: https://doi.org/10.1145/3341215.3356258

[118] R. Han, Z. Feng, T. Xu, C. Ai, W. Xie, K. Zhang, and J. Li, "Multi-sensors based 3d gesture recognition and interaction in virtual block game," in *2017 International Conference on Virtual Reality and Visualization (ICVRV)*, Oct 2017, pp. 391–392.

[119] H. H. Publishing, "5 exercises to improve hand mobility," 2019. [Online]. Available: https://www.health.harvard.edu/pain/5-exercises-to-improve-hand-mobility-and-reduce-pain

[120] P. Xu, "A real-time hand gesture recognition and human-computer interaction system," *CoRR*, vol. abs/1704.07296, 2017. [Online]. Available: http://arxiv.org/abs/1704.07296

[121] Y. Bouteraa, I. B. Abdallah, and A. M. Elmogy, "Training of hand rehabilitation using low cost exoskeleton and vision-based game interface," *Journal of Intelligent & Robotic Systems*, vol. 96, no. 1, pp. 31–47, 2019.

[122] A. D. Gama, T. M. Chaves, L. S. Figueiredo, A. Baltar, M. Ma, N. Navab, V. Teichrieb, and P. Fallavollita, "Mirrarbilitation: A clinically-related gesture recognition interactive tool for an ar rehabilitation system," *Computer methods and programs in biomedicine*, vol. 135, pp. 105–14, 2016.

[123] S. Karishma and V. Lathasree, "Fusion of skin color detection and background subtraction for hand gesture segmentation," *International Journal of Engineering Research and Technology*, vol. 3, no. 2, pp. 1835–1839, 2014.

[124] K. Benabderrahim and M. Bouhlel, "Detecting and tracking the hand to create an augmented reality system," *IOSR Journal of Computer Engineering*, vol. 16, pp. 31–35, 01 2014.

[125] A. Krizhevsky. (2012, Sept.) Alexnet: Convolutional neural networks. [Online]. Available: https://en.wikipedia.org/wiki/AlexNet

[126] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[127] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[128] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[129] Y. Hu, J. Xu, Z. Ma, and G. Cao, "Predictive hand gesture classification for real time robot control," in *Proceedings of the 10th International Conference on Internet Multimedia Computing and Service*, ser. ICIMCS '18. New York, NY, USA: ACM, 2018, pp. 28:1–28:5. [Online]. Available: http://doi.acm.org/10.1145/3240876.3240914

[130] S. Ghanem, A. Imran, and V. Athitsos, "Analysis of hand segmentation on challenging hand over face scenario," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 2019, pp. 236–242.

[131] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 10, pp. 1337–1342, 2003.

[132] M. A. Mofaddel and W. M. Abd-Elhafiez, "Fast and accurate approaches for image and moving object segmentation," in *The 2011 International Conference on Computer Engineering & Systems*. IEEE, 2011, pp. 252–259.

[133] E. S. Gedraite and M. Hadad, "Investigation on the effect of a gaussian blur in image filtering and segmentation," in *Proceedings ELMAR-2011*, Sep. 2011, pp. 393–396.

[134] H.-Y. Lai, H.-Y. Ke, and Y.-C. Hsu, "Real-time hand gesture recognition system and application," *Sensors and Materials*, vol. 30, no. 4, pp. 869–884, 2018.

[135] B. Foundation, "Home of the blender project - free and open 3d creation software," 2019. [Online]. Available: https://www.blender.org/

[136] G. Engine, "Free and open source 2d and 3d game engine," 2019. [Online]. Available: https://godotengine.org/

[137] A. Jain, D. Arora, R. Bali, and D. Sinha, "Secure authentication for banking using face recognition," *Journal of Informatics Electrical and Electronics Engineering*, vol. 2, no. 02, pp. 1–8, 2021.

[138] F. F. Goldau, T. K. Shastha, M. Kyrarini, and A. Gräser, "Autonomous multi-sensory robotic assistant for a drinking task," in *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2019, pp. 210–216.

[139] M. Z. Zadeh, A. Ramesh Babu, A. Jaiswal, M. Kyrarini, and F. Makedon, "Self-supervised human activity recognition by augmenting generative adversarial networks," in *The 14th PErvasive Technologies Related to Assistive Environments Conference*, 2021, pp. 171–176.

[140] F. Farahanipad, M. Rezaei, A. Dillhoff, F. Kamangar, and V. Athitsos, "A pipeline for hand 2-d keypoint localization using unpaired image to image translation," in *The 14th PErvasive Technologies Related to Assistive Environments Conference*, 2021, pp. 226–233.

[141] A. Lahiri, A. K. Jain, S. Agrawal, P. Mitra, and P. K. Biswas, "Prior guided gan based semantic inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 696–13 705.

[142] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on image processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[143] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding: Combining inconsistent images using patch-based synthesis," *ACM Transactions on graphics (TOG)*, vol. 31, no. 4, pp. 1–10, 2012.

[144] J.-S. Park, Y. H. Oh, S. C. Ahn, and S.-W. Lee, "Glasses removal from facial image using recursive error compensation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 805–811, 2005.

[145] M. K. J. Khan, N. Ud Din, S. Bae, and J. Yi, "Interactive removal of microphone object in facial images," *Electronics*, vol. 8, no. 10, p. 1115, 2019.

[146] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Unmasking face embeddings by self-restrained triplet loss for accurate masked face recognition," *arXiv preprint arXiv:2103.01716*, 2021.

[147] N. U. Din, K. Javed, S. Bae, and J. Yi, "A novel gan-based network for unmasking of masked face," *IEEE Access*, vol. 8, pp. 44 276–44 287, 2020.

[148] ——, "Effective removal of user-selected foreground object from facial images using a novel gan-based network," *IEEE Access*, vol. 8, pp. 109 648–109 661, 2020.

[149] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[150] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[151] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[152] K. Javed, N. U. Din, S. Bae, R. S. Maharjan, D. Seo, and J. Yi, "Umgan: Generative adversarial network for image unmosaicing using perceptual loss,"

in *2019 16th International Conference on Machine Vision Applications (MVA)*. IEEE, 2019, pp. 1–5.

[153] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6721–6729.

[154] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.