Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Procedural Generation of Datasets for Training Hand Pose Estimation Systems

Borja García Quiroga, BEng

A dissertation

Presented to the University of Dublin, Trinity College in partial fulfilment of the requirements for the degree of

MSc in Computer Science (Augmented and Virtual Reality)

Supervisor: Michael Manzke, PhD

Trinity term, 2023

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

The use made of AI tools in the preparation of this document is outlined in an appendix, as per the School guidelines (see below).

Borja García Quiroga
Dublin, August 18, 2023

Abstract

The rise of computer vision systems has reshaped many industries, driven by the powerful capabilities of deep neural networks. However, as the complexity of these systems grows, so does the demand for larger datasets. However, the manual annotation of large-scale datasets comprising the diversity required by these systems is labour-intensive and time-consuming. This dissertation delves into the exploration of procedural generation techniques for hand pose datasets to overcome this challenge while also investigating the impact of controlled variations in detection quality and reliability, encompassing joint angles, wrist orientations, texture, lighting, and background variations, aiming to make it capable of handling diverse real-world settings. To assess the efficacy of the generated datasets, a state-of-the-art computer vision system is trained to detect key points in hand images using both the procedurally generated dataset and traditionally annotated datasets. Comparative analyses evaluate the trained system's performance on real-world data, comprehending the influence of procedural variations on its accuracy, robustness, and generalisation capabilities. In conclusion, this dissertation contributes to hand pose estimation by integrating innovative approaches for procedurally generating datasets. The findings underscore the importance of automated variations in dataset generation and offer insights into their impact on the quality of trained computer fabrication systems.

Keywords — Synthetic data generation, computer vision, hand pose estimation, automated variations

Aknowledgements

I want to express my gratitude to the people who have helped me complete this dissertation and Master of Science:

First, I sincerely thank Dr Michael Manzke for supervising this dissertation and for his invaluable guidance and advice. His comments and explanations have steered this dissertation positively throughout the whole process.

I also express my gratitude to the rest of the Trinity Professors whose modules have played an important role in shaping the foundations of this project.

I extend my appreciation to Albert for his support and priceless proofreading. His keen eye and feedback have been instrumental in refining the clarity and coherence of this dissertation.

Also, I want to thank my friends and classmates for their support and the camaraderie we shared throughout the days and nights of stress, making this experience manageable and memorable.

Last, I express my heartfelt gratitude to my parents and siblings. Without their support, I would not have gotten here.

Borja García Quiroga
University of Dublin, Trinity College
August 2023

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, the proliferation of computer vision systems has witnessed a substantial rise, leading to a profound impact on the tasks computers can perform with little human interaction [13, 14]. These technologies have impacted nearly every sector, encompassing industries ranging from healthcare and automotive to retail, surveillance, and entertainment.

These systems aim to replicate human vision and understanding capabilities, allowing machines to recognise and analyse objects, scenes, and patterns within graphic data. Neural networks—particularly deep neural networks—have become the crucial technology in most complex modern computer vision systems [15, 16].

Neural Networks (NNs)—initially inspired by the structure and functioning of human neurons—consist of interconnected nodes organised in layers [17, 18]. Each node applies mathematical operations to its inputs and produces outputs using a series of trained parameters. These outputs are subsequently sent to adjoining layers that will take them as inputs. The depth of a Neural Network—the number of layers it possesses—plays a critical role in determining the complexity of tasks the system can accomplish. As their depth increases, networks become capable of learning increasingly intricate features and patterns, as they have higher capacities to model complex relationships between data [19]. Deep neural networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have remarkably performed in handling computer vision tasks [15].

Nevertheless, using deep neural networks presents a noteworthy drawback whereby larger datasets are needed to obtain generalisable models. Although other challenges associated with network depth—such as the vanishing/exploding gradient problem—can be mitigated with specialised training techniques and methods [20, 21], the need to enlarge the training dataset persists as the depth increases. As discussed by Zhang et al., the number of parameters in a deep network can easily present enough expressivity to overfit the training data if not big-enough datasets are used [22].

Training models with such big datasets has become more attainable in recent years due to the greater availability of powerful computational resources. Modern high-end machines can leverage their computing power to train deep neural networks with enormous amounts of data in reasonable training times to obtain robust models. The advance of specialised parallelisation software and hardware—GPUs or TPUs—has significantly reduced training times [23, 24].

The inherent challenges of training such systems are not eradicated despite the technical feasibility. One of the most significant challenges researchers and developers face is the obtention of large amounts of data [25]. While publicly and commercially available labelled datasets exist [26], the scarcity or absence of desired data for specific problems forces developers to seek alternative means of data acquisition.

Obtaining and labelling real-life data usually involves significant time and manual labour costs [26, 27]. Often, these costs make building these large datasets unfeasible or impractical. Therefore, alternative approaches for data augmentation and procedurally generating datasets have been introduced in diverse scenarios [28, 29]. Substantial results have been obtained in recent years using synthetic data—both for data augmentation [30, 31, 32] and for training models exclusively on synthetic data [33, 34].

This dissertation aims to demonstrate the design process of a synthetic dataset generation system using computer graphics and investigate the impact of different variations on a model's capacities. By exploring the correlation between different variations and model performance, this research deepens our understanding of how different techniques affect the model's ability to acquire knowledge.

The scope of synthetic dataset generation is vast, as these techniques can be applied to train virtually any deep learning system. This dissertation acknowledges the extensive nature of this field and recognises the necessity of narrowing the focus. Doing so makes it possible to focus on specific aspects of a more targeted problem and explore the effects of different techniques on the results. Defining the boundaries of this research to one specific problem allows for meaningful experiments that offer valuable insights, allowing for a deeper understanding of the corresponding intricacies and contributing to future research.

Therefore, the scope of this dissertation focuses on the problem of hand pose detection [35, 36, 37, 38]. It seeks to explore the creation of a synthetic dataset and conduct experiments to evaluate the influence of controlled variations, including joint angles, wrist orientations, texture, lighting, and background, on the quality and reliability of hand pose detection. By studying the effects of these variations, this research assesses dataset robustness.

To assess the effectiveness of the produced datasets, state-of-the-art computer vision systems are trained to detect key points in hand images using both the procedurally generated dataset and traditionally annotated datasets. Performance metrics are used to assess the suitability to infer real-world data, examining the influence of variations on its accuracy, robustness, and generalisation capabilities. In conclusion, this dissertation aims to contribute to the ongoing research on procedurally generating datasets and to assess their suitability for hand pose estimation.

Chapter 2

Background

This dissertation focuses on the synthetic generation of hand pose datasets to study the intricacies and specifics that impact such a system when applied to a definite, quantifiable problem. Therefore, this section encompasses a dual background review exploring synthetic dataset generation and hand pose detection, establishing a solid theoretical foundation for both the overarching theme and the experimental design.

2.1 Synthetic image dataset generation

Machine Learning (ML) is the field of Computer Science that fundamentally encompasses all algorithms and models created to extract and assimilate knowledge, relationships, and patterns from data and use them to make decisions [39]. Contrary to conventional programming principles, wherein developers craft specific instructions to direct computational tasks, ML uses mathematical foundations to allow computers to 'uncover' the most appropriate algorithms to perform the necessary task independently.

The origin of the term "Machine Learning" is commonly linked to Arthur L. Samuel, an IBM engineer and AI pioneer. This association stems from his 1959 publication titled "Some Studies in Machine Learning Using the Game of Checkers" [40]. During the 1960s, experiments revolved around pattern matching decision-making, including methods such as trial-and-error and the nearest neighbor algorithm [41, 42, 43]. The subsequent decades, particularly the 1970s and 1980s, witnessed increasing interest in research directed towards Machine Learning [44, 45, 46], a momentum that has exponentially grown until today.

Today, the ambit of Machine Learning has found its applications across a vast spectrum of domains where manual curation of algorithms by human programmers would prove inefficient, deeply imprecise and often economically impractical. Machine learning has significantly impacted many sectors, including healthcare, finance, manufacturing, transportation, energy, advertising, entertainment, and others [47, 48, 49].

Machine learning's versatility arises from its independent approach to problem-solving, which revolves around transforming any problem into a common framework of numeric data [50]. Machine Learning is fundamentally rooted in mathematical models with adjustable internal parameters, which map input

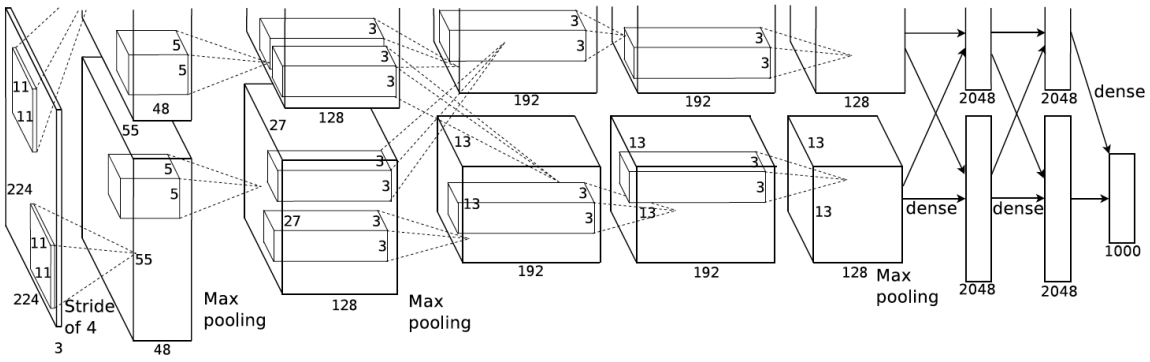


Figure 2.1: An illustration of the AlexNet architecture, an example of a deep neural network. AlexNet is considered one of the most influential contributions in the field of deep learning. As of August 2023, it has been cited over 135,000 times. [1]

features to outputs. The essence of the learning process entails finding optimal internal parameters or weights that can effectively address the given problem, achieving minimal possible loss. Consequently, mathematical optimisation stands as the cornerstone of machine learning [50]. In pursuit of determining these weights, various optimisation strategies are available, including but not limited to gradient descent, Evolutionary algorithms, and Adaptive Moment Estimation (ADAM) [51]. The principal distinction between conventional optimisation and machine learning lies in the concept of generalisation. Traditional optimisation algorithms excel in minimising loss within a training dataset, yet machine learning extends its focus toward minimising loss to previously unseen samples.

Understandably, models with greater numbers of trainable parameters can model more complex relationships between input features and outputs, being able to handle more complex problems [52]. Neural Networks—initially conceptualised to emulate the architectural configuration of the human brain [53]—constitute a genre of these complex machine learning systems. They comprise interconnected neurons organised into input, hidden, and output layers. Their proficiency in capturing complex patterns, leveraging substantial datasets, and executing intricate computations through layered connections establishes them as a dominant force in resolving multifaceted issues across diverse domains. This "deep" nature refers to the depth of transformation layers, further extending their capability to process data.

Neural Networks have proven effective in addressing many intricate Computer Vision challenges. Their ability to automatically learn and extract complex visual features from raw data, coupled with the hierarchical learning facilitated by deep architectures, has led to remarkable advancements in various domains of Computer Vision [54]. Tasks such as classification, detection, segmentation and attribute extraction have all seen significant improvements in accuracy and robustness by applying these technologies in contrast to the alternative methods previously used [55].

However, while deeper neural networks excel at handling more challenging tasks, they introduce additional complexities to the training process. Firstly, the issue of vanishing and exploding gradients [56] arises during backpropagation in deep networks, where gradients can dwindle or surge uncontrollably as they traverse the layers, potentially resulting in lousy convergence or complete training failure. Secondly, the number of parameters impacts the resources needed for training and inference [57]. Still, it also extends to the optimisation methods, as traditional methods might not work as expected in high-dimensional-spaced problems [58]. Finally, as discussed by Zhang et al., the number of parameters in a deep network can

easily present enough expressivity to perfectly overfit the given data if not enough data is used [22].

Many of these challenges have been effectively addressed by implementing advanced training techniques. Thoughtful weight initialisation, activation functions like ReLU, and normalisation methods such as batch normalisation mitigate issues related to vanishing and exploding gradients [56]. Furthermore, using sophisticated optimisers like Adam, RMSprop, and similar alternatives demands careful tuning to ensure reliable convergence [58]. However, the critical problem of achieving model generalisability hinges mainly on the availability of balanced, quality, and sufficiently large datasets. Therefore, obtaining labelled data has become one of the main obstacles in complex Machine Learning problems.

While real-world image data collection can be automated [59, 60] to a limited extent for applications involving fixed or vehicle-mounted cameras, this automation only applies to some computer vision scenarios. Challenges intensify when intricate data annotation is required. Manual data annotation can be a sluggish and arduous process, influenced by the specificity and accuracy demands of the annotations. Some types of datasets are relatively simple to annotate, like classification datasets. However, the undertaking becomes exceedingly time-consuming when confronted with annotating over a million images, even with a substantial workforce. As the annotations grow more complex, encompassing elements such as crowd headcounts, object poses, and depth perception, the cost-effectiveness of manual annotation diminishes significantly. Beyond temporal and financial concerns, manual annotation quality often deteriorates with large datasets due to inherent human errors [61]. Indeed, there are certain types of problems and applications for which individuals might not be capable of realistically and extensively providing reliable annotations.

In this context, generating custom training data receives significant interest from the research community. Synthetic data entails the creation of artificially generated datasets that replicate the statistical properties, patterns, and characteristics of real-world data without gathering actual data from the real world [62]. It has emerged as a long-term goal within machine learning systems, allowing researchers to create data that directly suits their precise problem scenarios. Synthetic data is generated through algorithms, models, or computational techniques and is extremely useful for tackling data scarcity, resource limitations and many other constraints.

Synthetic data encompasses a broad spectrum of techniques and purposes originating from sources other than real-world data. This realm is subdivided into three categories: data augmentation, data generation, and data completion [62]. Data augmentation involves techniques that enrich existing datasets by applying transformations or modifications to real-world data, enhancing their diversity and robustness. This means creating artificial variations of real-world instances that help achieve model generalisation. Data generation, on the other hand, entails creating entirely new data that emulate the statistical attributes and patterns found in real-world data. In Computer Vision, this often includes realistic rendering for creating images that look virtually like the real world. Lastly, data completion refers to filling in missing or incomplete data points, effectively reconstructing the dataset while adhering to its underlying structure.

The adaptability of synthetic data generation extends to the variety of data types that can be used for input and output [63]. It encompasses a diverse range of information modalities, including but not restricted to RGB images, depth maps, segmentation masks, and key point annotations. RGB images provide visual cues akin to human perception, emulating true-to-life scenes recorded with conventional cameras. This data type requires great details and advanced rendering techniques to obtain highly realistic images equivalent to real-world sampled images. Depth maps introduce spatial dimensions, capturing

Figure 2.2: Example of data augmentation through colour and blur variations from the Albumentations library. Data augmentation is a widely used technique for generating synthetic data. [2]

the distances between objects and facilitating the replication of three-dimensional contexts. While the accuracy of rendering techniques for such information isn't always necessary, achieving excessively flawless simulations can pose a challenge, given that real-world sensors do not operate in such an ideal manner [64]. Segmentation masks define object boundaries, enabling the isolation and categorisation of distinct elements within a scene. Segmentation problems are particularly interesting for synthetic data, as generating segmentation maps is very time-consuming to perform manually. Keypoint annotations, conversely, offer accurate positional information, facilitating the recreation of intricate structural relationships.

Synthetic data offers a multitude of compelling advantages. Firstly, it addresses privacy concerns, as the resulting data does not belong to actual individuals, ensuring total confidentiality [65]. Moreover, synthetic data can reduce the bias found in many real-world datasets. However, this presents a dual-faceted scenario. While a procedurally generated dataset can be constructed to address bias intentionally, it's equally susceptible to succumbing to the same diversity deficits that plague traditional datasets [66]. These biases are deeply ingrained within the human psyche and frequently go unnoticed, causing them to be inadvertently overlooked. By meticulously crafting data, skewed patterns can be deliberately rectified, fostering fairer and unbiased model outcomes. Additionally, synthetic data's volume can be tailored to the specific requirements of models, obtaining control over data quantity.

However, even though synthetic data presents notable theoretical advantages over real-world data acquisition, it still poses significant challenges for models to generalise to real-world settings effectively [67]. Mainly, these approaches demonstrate superior performance when dealing with more straightforward

problems of Computer Vision. This phenomenon is known as the "synthetic-to-real gap" [68], where models trained on synthetic data might not achieve optimal performance when faced with the intricacies, uncertainties, and unmodeled factors of real-world environments. Despite advancements in generating more diverse and realistic synthetic data, the inherent divergence between synthetic and real-world data remains a hurdle. As a result, bridging this gap is the focal point in research, involving techniques like domain adaptation, transfer learning, and fine-tuning to enhance model robustness and adaptability across the transition from synthetic to real data domains.

2.2 Hand pose estimation

A hand pose refers to the intended or unintended configuration or arrangement of the hand's elements. These elements' relative position and orientation in a specific instant define a hand pose. Conversely, a hand gesture is a necessarily intentional action involving a sequence of poses, and their relative position to the body, intended to communicate a message [38].

The human hand exhibits a distinctive anatomical composition comprising three primary components ruling its configuration: the palm, fingers, and thumb. Each component comprises multiple bones, giving them unique anatomical, kinematic, and positional attributes restricting the range of feasible hand poses. In this section, the term "palm" encompasses both the front and back regions, which serve as the central connection between the wrist and the fingers. The palm comprises five metacarpal bones, each aligned with a corresponding finger [69]. Notably, these metacarpal bones are longer than any other bones in the hand, possess a cylindrical shape, and articulate with both the wrist bones and fingers. The fingers consist of three phalanges—proximal, middle, and distal—arranged based on their proximity to the palm, from inner to outer. The proximal and middle phalanges are articulated on both ends, while the distal phalanx (the fingertip) lacks any continuation by another bone. As for the thumb, it solely comprises two phalanges, missing the middle phalanx present in the other fingers. These bone configurations establish the rigid structural framework of the hand and, barring malformation cases, dictate the general proportions observed in adult human hands.

The bones in the hand are connected by a series of joints that allow its intricate movements. Finger joints are classified into two main categories: interphalangeal and metacarpophalangeal. Metacarpophalangeal joints (MCP) are situated between metacarpal bones and proximal phalanges. MCP joints allow flexion, extension, and some degree of abduction and adduction—lateral rotation spreading and bringing together the fingers. Secondly, the carpometacarpal joint (CMC) is a unique joint connection to the thumb that provides a wide range of motion and enables opposition. Finally, the wrist allows for a much greater range of movements, allowing flexion and extension, abduction and adduction, and pronation and supination.

The hand's degrees of freedom (DOFs) refer to the number of independent ways the joints can move to form different poses. Each degree of freedom is defined by one type of motion in one joint. The human hand possesses over 20 DOFs—the exact number depends on the level of abstraction—contributing to its dexterity [69]. Flexion and extension refer to a joint's bending (flexion) and straightening (extension) [70]. Abduction and adduction refer to the joint's spreading apart (abduction) and bringing it together. Opposition is the unique ability of the thumb to pivot across the palm, allowing it to meet the tips of the other fingers. Finally, pronation and supination refer to the motions that involve rotating the hand to turn the palm downward (pronation) or upward (supination).

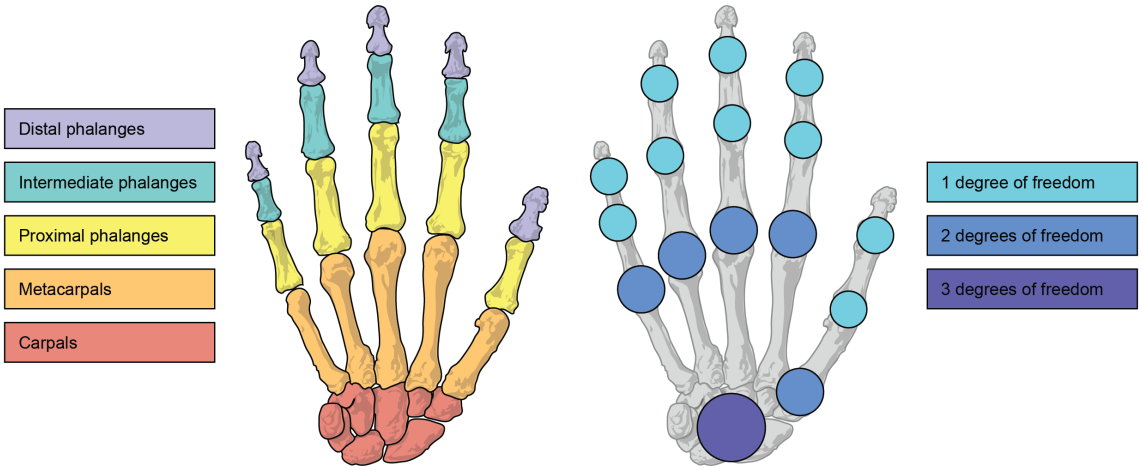


Figure 2.3: Illustration of the anatomy and degrees of freedom of the human hand. This model of dexterity is used for hand pose estimation. Adapted from [3]

Hand pose estimation is the process of determining the position of these elements by defining the 3D or 2D coordinates of specific key points of the hand from given data. This information is subsequently used to infer the hand's specific gesture or pose. The domain of hand pose estimation encloses diverse techniques employing distinctive input data, such as individual static images, video sequences, or specialised sensors capturing other types of data.

In recent years, hand pose estimation has been a relevant discussion topic in computer vision because of its wide range of real-world applications, necessary in many fields, such as human-computer interaction [71], sign language recognition [4], augmented and virtual reality [72], robotics, and gesture-based control systems. Given the fast, natural and organic means of communication that hand gestures can be, this topic has been of particular interest for researchers in interaction with other fields, such as accessibility [73], special education [38], anthropology and sociology.

Sign languages have garnered significant attention in this field over the past decade, driven by the goal of providing equal access to services and applications for their user communities. These languages cater to various users, such as deaf and hard-of-hearing people, individuals with autism facing communication challenges, and other users with special needs. As a result, organisations and institutions have shown increased interest and investment in researching this field [74, 75], leading to substantial growth in published literature. The diversity of sign languages, both in their purpose and lexicon, [76] has motivated researchers to develop efficient and precise systems capable of distinguishing hundreds of different signs and poses. This pursuit has made sign languages a driving force in advancing pose estimation research using computer vision.

Within the domain of hand pose estimation, a wide array of systems employ diverse techniques to detect and infer hand poses, catering to the vast range of applications and domain-specific requirements. These techniques encompass Machine Learning, depth-based, model-based, and hybrid systems. Depth-based methodologies use sensors such as LIDAR [77], Kinect [78], or stereo cameras to capture three-dimensional (3D) information. On the other hand, model-based approaches utilise hand models and optimisation techniques to infer hand poses by fitting the model to observed data. Hybrid approaches constitute a fusion of multiple techniques, such as integrating RGB images with depth information or combining model-based and data-driven methods. Ultimately, the choice of techniques depends on the specific use

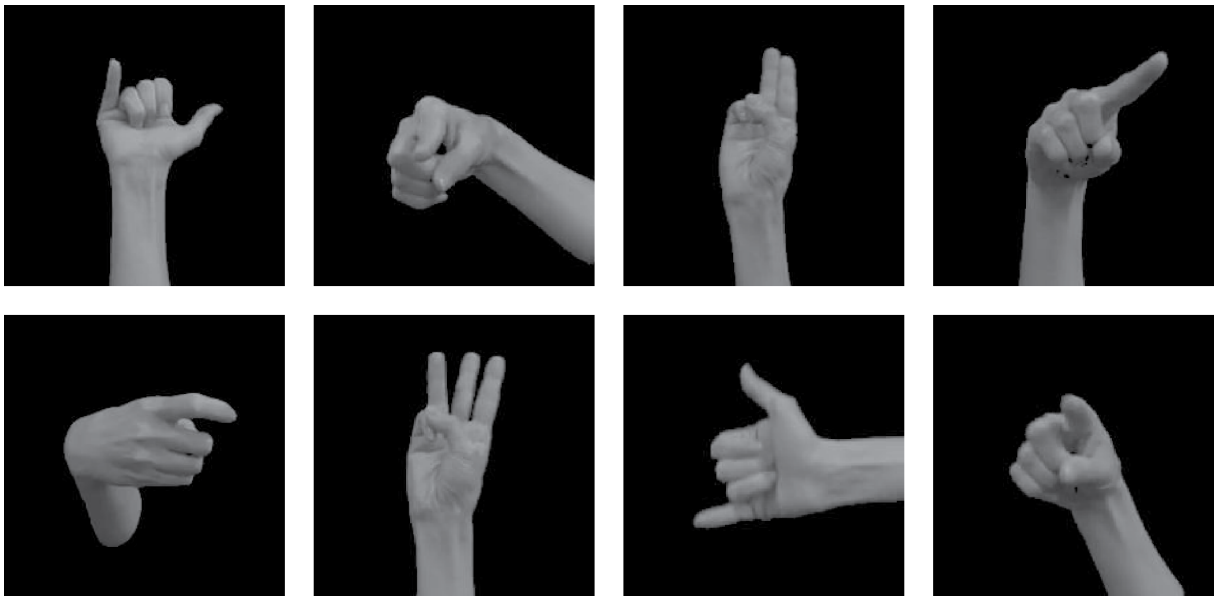


Figure 2.4: Examples of different configurations a human hand can take. The examples are based on the Irish Sign Language. [4]

case, available hardware, and the desired levels of accuracy and real-time performance.

Despite ongoing research on hand pose estimation, this topic has not been perfectly resolved due to its non-trivial nature and inherent challenges. These challenges present obstacles to achieving high accuracy and reliability in hand pose estimation. Some of the most prominent challenges include occlusions, where parts of the hand are obscured, and self-occlusions, where one part overlaps with another, resulting in ambiguity in the positions of key points. Additionally, the varying shapes of hands introduce complexity, as hands can assume diverse configurations, leading to variations in keypoint appearance. Another issue arises from ambiguities caused by multiple hand poses producing similar visual cues [79]. Moreover, regular RGB cameras add significant difficulty in accurately estimating hand pose, as they cannot directly measure the distance or depth of objects in the scene, unlike depth cameras or specialised sensors.

In contrast, the inherent attributes of the hand can be used to ease hand pose estimation and augment its accuracy. Mechanical constraints and limitations inherent to the human hand delimit possible hand poses. Furthermore, integrating joint limits, kinematic constraints, and physical plausibility engenders heightened realism within the estimations. Nevertheless, despite the improvements achieved by anatomical constraints, they can only partially obviate challenges arising from occlusions, varying hand shapes, and ambiguities.

Abstracting the full complexity of this problem within a computationally solvable model introduces an added layer of considerations. Hand pose representation encompasses three prevalent methods: 2D Keypoint Coordinates, 3D Joint Locations, and Skeletal Representations. The 2D coordinate approach stands out for its simplicity and computational efficiency, albeit it suffers from the absence of depth information, leading to ambiguities in specific contexts. Conversely, the 3D representation exhibits heightened accuracy and diminished ambiguity, albeit at the expense of intricate setups and increased computational demands. As for Skeletal Representations, they balance simplicity and robustness yet relinquish fine-grained detail. Consequently, selecting a suitable representation hinges upon specific application requirements, data availability, and the desired levels of accuracy and expressiveness.

Several widely used datasets and benchmarks have significantly contributed to advancing hand pose estimation algorithms. Well-known datasets include NYU Hand Pose Dataset [80, 81], MSRA Hand Pose Dataset [82, 83], BigHand2.2M Benchmark [84], FreiHAND Dataset [12, 85], and Multiview 3D Hand Pose Dataset [86, 87], and GANerated Hands Dataset [88, 9]. Each dataset has unique characteristics, such as the type of data (RGB, depth, or both) and the number of annotated 3D hand joint positions. Evaluation of algorithms on these datasets is typically done using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Percentage of Correct Keypoints (PCK) to measure the accuracy of the predicted joint locations compared to the ground truth annotations.

Chapter 3

Literature review

Synthetic dataset generation and hand pose estimation have each evolved within distinct scientific domains, with sporadic intersections in studies where synthetic data has been used for hand pose estimation techniques but mostly maintaining independent trajectories. Hence, reviewing these two areas along separate timelines, highlighting common points and common trends when they arise, allows for a better understanding of each field's trends. In addition, synthetic data generation for hand pose estimation systems is explicitly examined in its corresponding subsection.

3.1 Synthetic image dataset generation

Computer Vision systems rely heavily on datasets for training, evaluation, and validation. However, dataset creation remains a major obstacle due to its demanding resource requirements. In response to this challenge, the research community has dedicated substantial efforts to produce openly accessible datasets for these domains. Notable examples include PASCAL VOC [89], Microsoft COCO [5], ImageNet [90], and NYU-Depth V2 [91], SUN RGB-D [92]. While these contributions have undeniably driven progress in numerous computer vision problem areas, they cannot cover the totality of challenges, scenarios, and classes tackled by current research. Consequently, the academic focus has moved to include dataset creation as one of the main issues to resolve to advance in the field. These new efforts have often focused on synthetic dataset generation methodologies as one of the ideal solutions.

The exploration of synthetic data generation for computer vision traces its origins to the late 1980s, producing a long trajectory that has interacted with many of the main goals within this field. Several review papers and books like Nikolenko [93] or Man and Chahl [94] extensively review the evolution in the field, starting with early explorations focusing on what the literature defines as low-level computer vision problems.

Figure 3.1: Samples from the Microsoft COCO dataset. The COCO dataset is significant due to its comprehensive and diverse collection of images, serving as a benchmark for training and evaluating object detection and segmentation models. Adapted from [5].

3.1.1 Early approaches and low-level computer vision

Ground truth datasets are hard to produce and label but are simple to simulate. Therefore, low-level computer vision problems were some of the earliest fields where synthetic data was successfully used. Synthetic datasets successfully evaluated optical flow estimation algorithms during the late 1980s and early 1990s [95, 96]. These first approaches primarily focused on the problem at hand, considering synthetic data as an auxiliary tool rather than the main focus of research.

Throughout the 1990s, more researchers recognised synthetic data generation as a valid approach for assessing general and cross-cutting topics in computer vision. For instance, MINPRAN [97] introduced an estimator capable of effectively fitting models with considerable amounts of outliers, validating the system using synthetic data. Similarly, Leedan and Meer focused on techniques for estimating solutions to bilinear forms, frequently found in various computer vision problems involving intricate variable relationships. Their approaches were consistently validated using synthetic data in several works [98, 99, 100]. By the decade's end, Freeman and Pasztor [101] discussed synthetically generated datasets as a comprehensive methodology for training networks to address various low-level problems, including motion analysis, shape estimation, and image resolution enhancement.

The growing trend of employing synthetic data for validating low-level computer vision challenges remained consistent throughout the 2000s. Several works introduced estimators based on the Random Sample Consensus (RANSAC) algorithm [102] for handling data with significant amounts of inliers [103, 104]. Wang et al. [105] achieved precise eye gaze estimation by focusing on a single eye, employing both synthetic and real-world data for validation. Similarly, Von Neumann-Cosel et al. [6] assessed Audi AG's lane tracking algorithm by comparing its results with a synthetic ground truth dataset.

However, regardless of the growing weight of synthetic data in computer vision research, concurrent research questioned the assumption that synthetic ground truth datasets should be considered reliable. This scepticism arose as different studies around the same time revealed notable differences between evaluations conducted using synthetic data and real-world data [106]. This topic has remained an ongoing debate as successive research has delved into synthetic data generalisation capabilities, as discussed in the sections below.

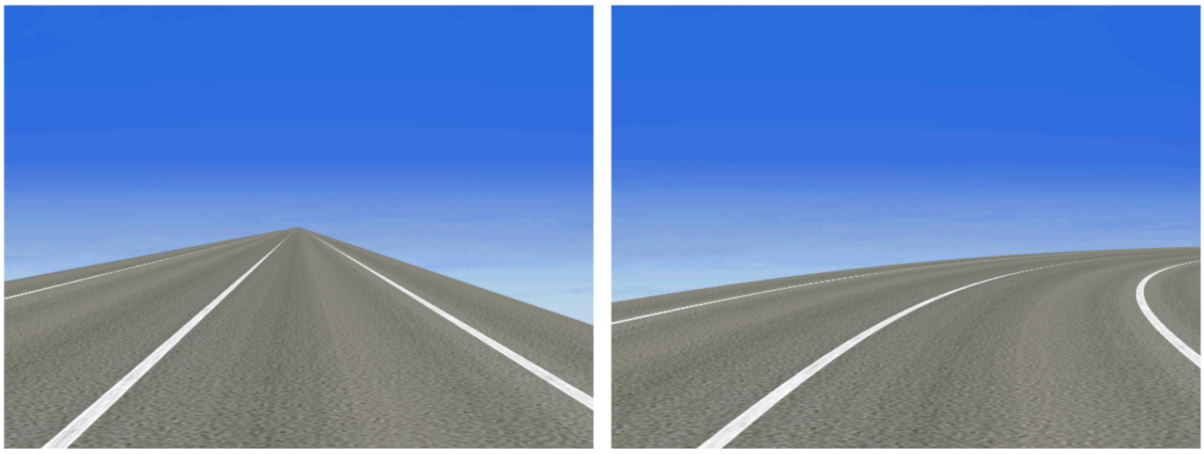


Figure 3.2: Von Neumann-Cosel et al. evaluated a lane tracking algorithm by comparing its results with synthetic ground truth data. [6]

More relevant works added meaningful contributions as the topic became a primary research goal. The Middlebury dataset [107] is considered a milestone in synthetic data for its meaningful advances in low-level computer vision. This dataset integrated real-life ground truth lighting data with realistic synthetic images. Other datasets for low-level problems were also introduced around the same time, tackling other data types and specific problems. For instance, the Tsukuba CG Stereo Dataset [108] presented synthetic data alongside ground truth disparity maps, demonstrating enhancements in the quality of disparity classification. Another example is MPISintel [109], which offers a synthetic optical flow dataset.

Although the initial success of synthetic dataset generation for addressing low-level vision problems was satisfactory, research in the field has remained active through the years. Hence, novel approaches have been presented to previously resolved issues, seeking higher efficiency and precision. One example is Dosovitsky et al. [10]. This paper introduced a sizable synthetic dataset, called Flying Chairs, constructed using a public 3D chair model database with real-life backgrounds. This dataset was originally used to train a CNN-based model for optical flow estimation. Subsequent research [110] built upon this dataset to address image disparity. Besides, more recent work [111] leveraged these datasets alongside other ones to assess and compare their performance.

3.1.2 Synthetic data for high-level deep learning systems

As computer vision research shifted towards deep learning, state-of-the-art models required larger datasets, saturating the capacity and resources of real-world dataset production. This phenomenon increased the prevalence of synthetic datasets and triggered a shift from evaluation to training datasets. This transition made sense, as high-level computer vision researchers were interested in solving data scarcity and synthetic data's capacity to deliver flawless annotations.

During the initial surge of deep learning during the mid-2010s, the suitability of synthetic datasets for effectively training high-level computer vision problems was uncertain [93]. Nevertheless, Peng et al. [112] proved there was space for promising results by employing AlexNet to compose a basic detection system trained with synthetic data. Although those results were never state-of-the-art technology, the paper gained attention for its significance for synthetic data research. Parallelly, using synthetic videos,

Figure 3.3: Summary of the synthetic dataset produced by Rajpura et al. This study focuses in a scenario optimal for synthetic data due to its reduced domain variability. a) 3D models used in the system. b) resulting synthesised images. Adapted from [7]

Bochinski et al. [113] expanded that approach. These publications also pioneered their methodology, which has become usual in the field, as they were among the first to use video games for synthetic data generation.

Meanwhile, Hinterstoisser et al. [114] proposed an innovative approach by engineering neural networks to use synthetic data effectively. They pointed out that training exclusively on synthetic data did not yield optimal results because of the differences between computer-generated and real-world images. Thus, they proposed not training the entire network on synthetic data. Instead, they pre-trained a model using existing real-world datasets, freezing the lower layers of the model and only using synthetic data to train the upper layers. This way, the basic image features that work well for real photos remain intact, while parts responsible for classifying objects would adjust to recognise the desired classes.

Contrarily, instead of focusing on producing the most technologically novel approach, Rajpura et al. [7] presented the quintessence of the perfect fit for a synthetic dataset. In their work, they built a dataset to recognise multiple objects on supermarket and fridge shelves. As the scenes and backgrounds for this problem are not common in existing datasets, and backgrounds and scene compositions are pretty standardised, this domain proved synthetic datasets conceptually useful when real-world datasets did not match. This perspective allowed them to focus on synthetic data's tangible benefits, making this field worthy of further research.

Continuing with systems with real-world applications, Nowruzi et al. [115] worked on the specific domain of urban outdoor environments, with a clear focus towards autonomous driving. They proposed using multiple datasets, mixing synthetic and real-world data, and a simulation tool to create more cheaply annotated synthetic data. They outlined a methodology for training neural networks using such datasets, and their conclusions have been applied successfully to other domains, becoming a go-to source in synthetic data literature.

Concurrently, synthetic data research started to obtain profitable results. The work by Hinterstoisser et al. [116] stands out for achieving outstanding object detection results using only synthetic data. Their research focused on everyday object detection, like food and medicine packages. A compelling matter discussed in this paper is their work on domain randomisation, particularly concerning background images. Importantly, their results show the potential of synthetic data when meticulously generated and all the data completion requisites are met. This purely synthetic dataset marked a significant milestone, outperforming a conventional real-world dataset of 2000 images.

However, despite its long trajectory and the importance of detection in the greater field, the potential of synthetic datasets goes far beyond this domain. Segmentation is another computer vision problem that can immensely benefit from pixel-perfect synthetic annotations. In this spirit, ShapeNet [117] indexed over 3 million 3D models classified into 3,135 categories and produced valuable annotations, including geometric, functional, and physical attributes. This paper became the basis for later efforts to automate the obtention of more complex annotations. Further work has developed ShapeNet to include hierarchically segmented parts and other enhancements [118, 119, 120].

These advantages were exploited by McCormac et al. [121] to smartly leverage flawless segmentation maps without constraining to synthetic data only. Instead, they proposed a system trained with purely synthetic data, pre-trained with ImageNet. The RGB Convolutional Neural Network trained using their synthetic dataset marked a significant milestone, as it represented the first instance where synthetic data yielded such improvement. Their dataset was an expansion of SceneNet [122], an annotated model generator for indoor scenes. This achievement contributed to the common practice of pre-training segmentation models using synthetic data.

Regardless of the splendid results of McCormac et al., the conclusions did not provide an infallible technique capable of successfully training any model with synthetic data. Contrarily, Saleh et al. [123] presented their work on how some object classes are not equally suited for segmentation synthetic dataset generation due to differences in their textures.

Consequently, they propose a resourceful approach combining detection and semantic segmentation masks. Specifically, Mask R-CNN [124] is used for detecting foreground classes and DeepLab [125] for segmenting background classes. Similarly to [114], this paper incorporated modifying the system's pipeline to optimise synthetic data as a pragmatic strategy.

In addition to detection and segmentation, other data types have received attention recently. 3D data, viewpoint, and depth are becoming increasingly relevant topics for their versatility and the more robust models produced. They are rapidly being incorporated into the field of synthetic data for being exceptionally challenging to label manually. Several works have dived into this issue. Aubry et al. [126] focused on chairs, while Liu et al. [127] did it on indoor objects. Conversely, Gupta et al. [128] produced synthetic renderings of different objects to train a CNN to detect and segment object instances to align 3D models.

Estimating 3D position and instance orientation is a common goal for 3D data. This problem is often known as the 6-DoF (degrees of freedom). Hodan et al. [129] have introduced a dataset containing real-world sensor data alongside 3D models of the object to provide the ground truth poses. However, it was only with the work of Tremblay et al. [130] that the first state-of-the-art network for 6-DoF pose estimation trained exclusively on synthetic data was introduced.

3.1.3 Methods and frameworks for synthesising datasets

All modern synthetic dataset generation techniques for computer vision problems lie in one of four main categories: manual generation, Generative Adversarial Networks (GANs), parametric models, and video game or 3D engines. Unlike low-level problem datasets, often generated using mathematical models, high-level datasets require high-quality 3D models for producing the desired datasets.

Manual generation processes are a direct and uncomplicated approach involving a person arranging scenes, objects, and environments to create datasets. These techniques are straightforward regarding needed tools, as they can be executed using any 3D modelling software. However, these methodologies demand hefty time, particularly for data labelling. As a result, manual generation frequently diminishes the inherent benefits of data synthesis, namely, primarily dataset scalability and automatic annotation. However, it is relevant to note that some works, particularly those from the initial stages of this field, have produced datasets generated through manual approaches [10, 131, 132].

Transitioning from manual data synthesis to automated methods, Generative Adversarial Networks (GANs) [133] have emerged as the most prominent technique among novel approaches to generate 2D synthetic images. GANs are a system consisting of a generator and a discriminator confronting each other to produce optimal synthetic images. GANs find applications beyond image generation, including AI art [134, 135].

Figure 3.4: Synthetic bird images generated by StackGAN. The examples in the row below are obtained by running the ones in the top through the Stage II of StackGAN. Adapted from [8]

In GANs, the generators craft instances while the discriminators evaluate their authenticity relative to authentic data. Thus, the quality of the synthetic data steadily advances, progressively reaching seemingly-authentic outcomes. In an ideal scenario, the generator would eventually obtain images indistinguishable from authentic data. Nonetheless, practical outcomes deviate from this ideal, as the generator and discriminator always reach an equilibrium. However, this technology currently stands at the core of a fair share of the state-of-the-art approaches to synthetic data for computer vision [8, 136, 137, 138]. The vast landscape of GANs, which exceeds 500 proposed variations, [139, 140] has produced many approaches to generate synthetic data.

On the opposite extreme of data generation control lie parametric models. These synthesising techniques involve utilising parameterised variables within 3D models to alter elements of rendered scenes. Although crafting these models is costly and frequently limited by the available data to build the models, they offer excellent domain control and are easily applicable across diverse domains. Nevertheless, encompassing a wide enough range of parameters is critical for producing usable datasets. Parametric models have been

used in several works such as Alken et al. [141], producing a dataset of fish in underwater backgrounds with variations in positions, rotations, and sizes; or Dahmen et al. [142], who simulated measurements based on parametric models of real-world objects. Semi-parametric image synthesis has also been explored with datasets such as Cityscapes [143]. The utility of parametric model data generators lies in their capability to precisely regulate specific attributes, facilitating the execution of quantitative investigations.

An eminent subset within this category comprises 3D morphable models (3DMMs) [144]. These systems rely on statistical models to generate 3D features. Since they were first introduced in the late 1990s, 3DMMs have evolved substantially, now integrated into more complex parametric models. They have had a notable impact on the research of human synthetic dataset generation [145, 146, 147]. Nonetheless, the use of this technology depends on amassing real-world data. Regrettably, there are no extensive public datasets fitting these requisites. Moreover, similar real-world datasets tend to produce notably biased models due to the inherent bias in the base dataset. This problem has already been discussed in the background chapter of this dissertation. Works have been published discussing how to reduce the impact of these biases [148, 149].

Lastly, a common approach to generating synthetic data receiving attention and showing promising results is to reuse existing engines and environments to create synthetic datasets. Nikolenko et al. [150] offer a comprehensive review of such methods.

Game engines like Unreal Engine and Unity, alongside 3D modelling software such as Blender, have gained substantial traction, offering a simple alternative to constructing 3D virtual environments [151, 152]. These systems offer simple integration with additional tools—i.e. modelling or video editing—and interactions with compatible software [153]. Among these solutions, UnrealCV [154, 155] stands out as an open-source plugin designed for Unreal Engine 4. This tool covers a spectrum of functionalities for adding variations to the datasets. Unity [156] and Blender [157] offer alternatives for these purposes too.

Video games have also been exploited for producing synthetic image datasets, presenting the advantage of including complex and increasingly photorealistic ready-to-use environments. This approach has been used for many different domains such as human detection [158], plane detection [159] or vehicle identification [160]. Besides, Richter et al. [161, 162] presented an interesting photorealistic dataset generated using Grand Theft Auto V, known for its large and realistic urban environment. Modifying software can be inserted into video games to automatically annotate data being able to obtain image and video data with pixel-perfect labels.

3.1.4 Synthetic data for hand pose estimation

Generating synthetic datasets of hand poses entails specific issues and constraints that require particular attention. While it might have received less attention than other prominent topics, this subject has seen several valuable works published in recent years. This section discusses papers tackling hand pose dataset generation and other related works that may offer practical insights that can be adapted and applied to hand poses.

Keskin et al. [163, 164] were among the first works to introduce hand pose synthetic data for training a hand pose estimation system. Their work focuses mainly on the estimation task—tackled with random decision forests—using depth synthetic data to simulate the entire pipeline. Although their focus was not

to produce a synthetic dataset, their early work using parametric models pioneered in this field. Early attempts did not yield outstanding performance in systems trained solely on synthetic data. However, Tang et al. [165] and Molina et al. [166] already proved the benefits of employing both real-world and synthetic data in supervised learning to boost generalisation. Similarly, Deng et al. [167] also used 3D hand synthetic data to augment the existing real-world dataset. In short succession, Madadi et al. [168] propose a system for estimating hand poses in videos using a two-step approach consisting of a part-based model initialisation and temporal data constraints. Their system was evaluated using a newly created dataset along with NYU and MSRA datasets. Their results demonstrated better results than most approaches at the time.

Zimmerman and Brox [169] presented the first work primarily focused on producing a large-scale synthetic dataset of hand poses. Their results were validated by training a neural network and comparing its performance against state-of-the-art models. Malik et al. [170] and Mueller et al. [88] followed by presenting their datasets SynHand5M and GANerated Hands, respectively. Both datasets provided automatically generated annotations obtained using synthetic generation. However, the latter proposed a GAN to translate the domain of the synthetic images to that of real-world images. Starting from a common ground, Cai et al. [171, 36] delves into estimating 3D hand pose from monocular RGB images. In this paper, synthetic data is produced to solve the issue of depth ambiguity. Their strategy reduces the need for real-world data by capitalising on transferring knowledge from fully-annotated synthetic datasets to weakly-labelled real-world datasets.

More recent work has proposed pre-training models using synthetic data and training it on unlabelled real-world data [172]. This approach, similar to what had been previously proposed by Hinterstoisser et al. [114], used pseudo-labelling to complete the training process. Conversely, Park et al. [173] has recently discussed training hand pose estimation models solely with synthetic data, focusing on realism and physical constraints as the tool to achieve generalisation.

Besides hand pose estimation, other synthetic datasets tackling human images have been proposed for other computer vision problems. For instance, human-face datasets have been widely discussed due to their appealing applications, greater variability, and privacy concerns. However, some of these publications tackle matters that are very relevant for hand pose datasets. For instance, Kortylewski et al. [33, 149] focus vast portions of their research on the impact of randomisation of pose, camera, lighting, and background using 3D morphable models. Moreover, they intensely discuss the problem of dataset

Figure 3.5: Samples from the GANerated Hands Dataset. Procedural models were employed to synthetically generate the dataset, which was subsequently processed through a GAN for image-to-image translation, enhancing the resemblance of features to those found in real hands. Adapted from [9].

bias and the threats and opportunities that synthetic data arise.

Similarly, synthetic human pose estimation provides conclusions that can easily be imported to this field. Ragheb et al. [174] demonstrated the viability of recognising human poses through a system trained with synthetic silhouettes obtained in a virtual environment. Khodabandeh et al. [175] presented GAN approaches to generate frame sequences with human skeletons. Besides, the PeopleSansPeople [176] project uses Unity Perception to create labelled synthetic datasets, allowing for great customisation in a way that could be exploited for hand pose data.

3.1.5 Image realism and generalisation

As previously addressed in this dissertation, generalisation is a prominent issue within the realm of deep learning systems. This concern notably amplifies when dealing with synthetic data, where the central objective of these systems lies in achieving effective generalisation to real-world contexts.

Intuition often leads to expect that extreme realism is the optimal answer for achieving real-world generalisation. However, while this belief might be true for some applications, the answer varies from problem to problem. Mayer et al. [111] came to three fundamental conclusions about synthetic datasets regarding this issue. Firstly, additional realism is not always critical, as their not-realistic Flying Chairs dataset performed correctly. Secondly, although realistic environments are unnecessary, realistic camera parameters are critical. Finally, they argue that domain randomisation has a more significant impact on generalisation than realism.

These conclusions are supported and extended by other relevant works. Firstly, it is elemental to note that what is perceived as "extreme realism" does not necessarily translate to absolute equivalence with reality. This concept was explored by Meister and Kondermann [177], who delved into this matter by training two systems with real-world data and seemingly equivalent synthetic data produced with ray tracing. Their conclusions showcased that, although both systems yielded approximately equal performance results, the spatial distribution of errors differed significantly.

These conclusions, however, do not mean that realism does not help generalisation but that researchers must recognise other important aspects of dataset optimality in balance with realism. This topic was extensively studied by Movshovitz-Attias et al. [178] surrounding viewpoint estimation. Their conclusions

Figure 3.6: The flying chairs dataset demonstrates strong performance in flow estimation, despite its non-realistic imagery, highlighting that strict realism isn't always imperative for achieving favourable outcomes. [10].

showed that realistic rendering does help and that performance gaps between models trained with real-world and synthetic data can be explained by domain adaptation. They argued that any attempt to generalise models to a different real-world dataset would be just as hard as adapting from a synthetic one. Tsirikoglou et al. [179] pointed in the same direction by presenting the outstanding performance of a model trained with a hyperrealistic synthetic dataset produced using Monte Carlo-based lighting and optics simulation.

The presented conclusions align with established principles of domain randomisation. The underlying concept suggests that by training on an extensive and diverse dataset, the network will more likely perform effectively on actual data [180]. This rule remains true even when individual instances within this artificially generated dataset are unrealistic.

Backgrounds in synthetic datasets have been consistently discussed in the literature concerning this issue. Understandably, if all instances of a synthetic dataset were positioned in front of a white backdrop, it would become straightforward for the network to learn how to segment it. However, relying solely on that dataset would make the resulting model impractical for real-world use. Instead, the learning process becomes more challenging by incorporating additional objects unrelated to the task, possibly inducing confusion but making the resulting model more robust. Abu Alhaija et al. [132] and Georgakis et al. [181] propose methods to add randomisation to backgrounds. The first paper uses random outdoor scenes and 3D models to create confusion. Conversely, the latter uses indoor scenes.

3.2 Hand pose estimation

Hand pose estimation has witnessed a lengthy trajectory in research due to its extensive potential applications. Since the earliest approaches, the field has been characterised by continuous exploration and refinement.

Early stages of research during the 1990s and 2000s focused on 3D hand pose estimation from single-camera RGB inputs. Resulting works were able to estimate hand poses through complex model-fitting approaches [182, 183, 11]. These methodologies used models and optimisation methods to approximate hand poses by aligning the model with observed data. However, these methods required a deep understanding of physics, dynamics and kinetics and heavily relied on multiple questionable pre-established hypotheses. Therefore, despite their conceptual complexity, these methods often obtained poor precision and a very constrained scope, limiting their viability for real-world scenarios. Subsequent multi-camera approaches [184, 185] addressed issues related to occlusion and exhibited more satisfactory accuracy levels. However, they relied on intricate models and expensive optimisation strategies, making them unsuitable for real-time use.

The arrival of affordable depth sensors made depth-based methods more enticing since depth images provide richer context that significantly reduces depth ambiguity and helps handle occlusions. Cai et al. [36] categorise approaches into three primary groups: generative, discriminative, and hybrid. Generative methods involve model-fitting approaches akin to those previously discussed. However, in this context, they aim to fit the model to depth data instead of RGB images. Oikonomidis et al. [186] introduced Particle Swarm Optimisation, an optimisation technique to be used alongside a Kinect sensor. Similar strategies were proposed by others, producing favourable outcomes [187, 188, 38]. In contrast, discrimi-

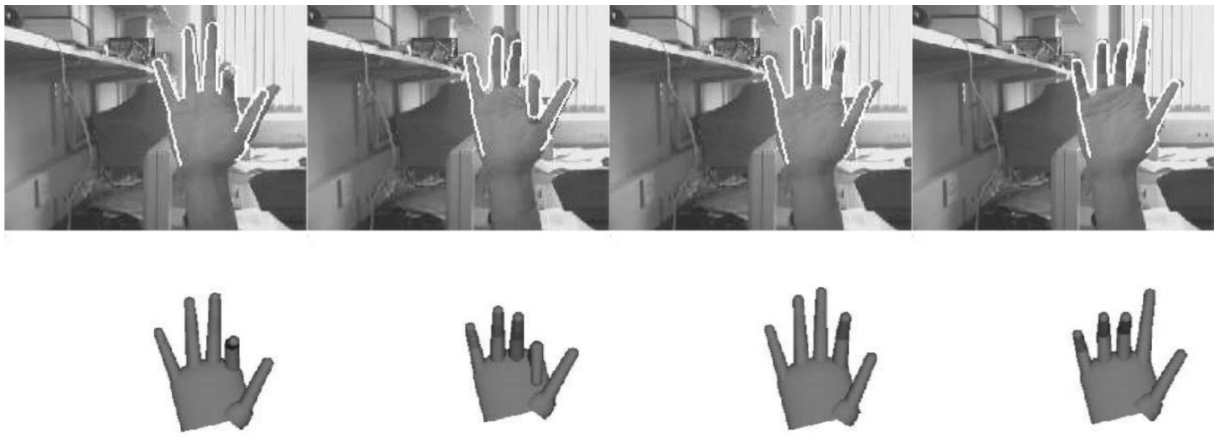


Figure 3.7: Fitting examples presented by Stenger et al. Early research during the 1990s and 2000s focused on fitting complex models to observed data. [11].

native methods directly predict the 3D joint positions from depth data. For example, Keskin et al. [164] proposed leveraging Randomised Decision Forests, while Xu and Cheng [189] introduced an intricate algorithm for selecting and classifying candidates. Hybrid methods have also emerged, amalgamating generative and discriminative techniques into single systems [190, 191].

The consolidation of neural networks as the central technology for all computer vision tasks rapidly impacted this field, leading to improvements in performance with no need for depth sensors. These approaches use different neural network architectures, sharing standard methodologies [192, 193, 194]. Zimmermann and Brox [169] introduced an innovative technique that involves training a neural network to learn the physical constraints of joints before estimating keypoint locations. Instead, Spurr et al. [195] proposed using a statistical hand model to ponder likely configurations. Other works have explored the feasibility of reconstructing 3D hand meshes from RGB inputs and deep learning techniques [196, 197, 198]. More recently, Zheng et al. [199] have shown effective results from 2D RGB data using U-Net, an architecture developed initially for biomedical computer vision [200].

Despite RGB images regaining prominence due to the enhanced performance of neural networks, the exploration of depth information has continued to progress concurrently. Ge et al. introduced innovative methods that leveraged the power of CNNs and RGB-D data [201, 202, 203, 204]. Similarly, other authors delved into estimation based exclusively on depth information [205, 206]. Furthermore, Wan et al. [207] explored the outcomes of variational autoencoder (VAE) and generative adversarial network (GAN) for synthesising depth data, while Mueller et al.[208] proposed a hybrid approach rooted in RGB-D data.

Chapter 4

Experimental design

This section outlines the experimental methodology used to evaluate the current state of synthetic dataset generation, as highlighted in the preceding chapter. This dissertation acknowledges the expansive nature of this field and recognises the need to constrain the experiments to a scale that aligns with the scope of this research. Consequently, the general goal of this methodology is to determine whether the findings and assertions discussed in the synthetic dataset generation literature remain valid or exhibit variability when applied to a specific problem—namely, hand pose estimation, which has unique characteristics and outcomes. By exploring the correlation between different variations and model performance, this approach aims to understand how diverse techniques influence models' performance and ability to generalise.

Consistent with this, the concrete methodology used in this dissertation to assess the feasibility of synthetic data generation is to build a system capable of automatically generating entirely new synthetic datasets devoid of initial real-world data. The details surrounding the choices for the different components of this experimental environment are discussed in subsequent sections.

4.1 Methodology

This dissertation employs a methodology for assessing the automatic generation of hand pose datasets based on crafting distinct versions differing from each other in controlled variations. Initially, a model is trained solely on real-world data to serve as a benchmark for evaluating the performance of synthetic datasets. This real-world data is only used for benchmarking purposes and not to train any other model. Subsequently, diverse versions of the dataset are produced, incorporating the different variations. These distinct datasets are then utilised to train separate instances of the model. Ultimately, these models attempt to estimate hand pose from real-world images, obtaining the corresponding accuracy metrics for each. These metrics facilitate an evaluation of their capability to infer real-world data, enabling an examination of the impact of variations on accuracy, robustness, and generalisation potential.

This methodology presents a valuable method for assessing the validity of the proposed synthetic data generation system in training robust models. Leveraging deep learning systems for this evaluation offers a more precise vantage point than mere attempts at estimating disparities between generated data and

ground-truth images. As noted by Mayer et al. and Meister and Kondermann [111, 177], the fidelity of the generated data shall not be the sole determining factor, as seemingly unrealistic datasets might yield more favourable outcomes. This approach also enables quantitative analysis of the synthetic-to-real gap.

4.2 Hand pose estimation model

The literature review shows a plethora of distinct approaches that have been successfully tested for estimating hand pose data. Consistently with the motivation of this study discussed and arguments in previous chapters, the technology employed to estimate hand pose must be a state-of-the-art neural network. This constraint narrows the potential options, yet a rich array remains for consideration.

While relevant literature explores hand pose estimation using depth data [205, 206], working solely on depth limits the inclusion of various variations that could present compelling conclusions. Moreover, a significant driver behind this dissertation is the creation of alternative datasets in scenarios where acquiring real-world datasets proves impractical. However, depth-only datasets can be automatically labelled using modern technologies like motion capture, as the absence of image data allows for the use of patterns purposely printed on the skin. Consequently, this approach would undermine the adequacy of the example.

Extensive research has also delved into RGB-D data for hand-pose estimation [203, 204, 208]. However, the reduced number of publicly available real-world datasets of this kind poses an obstacle to building the benchmark model. Interestingly, this situation perfectly aligns with the problem described as the research's motivation.

Nonetheless, predominant research currently focuses on RGB data, achieving promising outcomes [192, 195, 196]. Furthermore, the attributes of existing RGB datasets align more suitably with the scope of this dissertation, otherwise missing topics like image realism or texture variations.

Among the publicly available datasets referenced throughout this document, four options contain RGB images profitable for Neural Network training. The GANerated Hands Dataset [88] is unsuitable for this purpose due to its synthetic nature, hindering the assessment of the synthetic-to-real gap. Conversely, the NYU Hand Pose Dataset [80] does incorporate RGB data. Yet, it still falls short of being a benchmark for RGB estimators due to its limited environment randomisation compared to alternative datasets. The FreiHAND Dataset [12] and the Multiview 3D Hand Pose Dataset [86] meet the requirements for such a system. Nevertheless, the FreiHand Dataset has garnered greater use within literature and stands as a pinnacle among current state-of-the-art datasets, making it the optimal choice. This dataset also contains synthetic data for data augmentation, yet its structure enables the omission of these images.

Lastly, generating accurate ground truth annotations that match the images is imperative for such a system. Being FreiHand the selected dataset, the generated outputs shall align with the original characteristics of the dataset. Consequently, key point annotations will serve as the designated labels for the images. Akin to this approach, Zheng et al. [199] recently achieved satisfactory outcomes by utilising RGB data coupled with 2D annotations. Their proposed system employed a U-Net model [200], proving this architecture a reasonable experiment option.

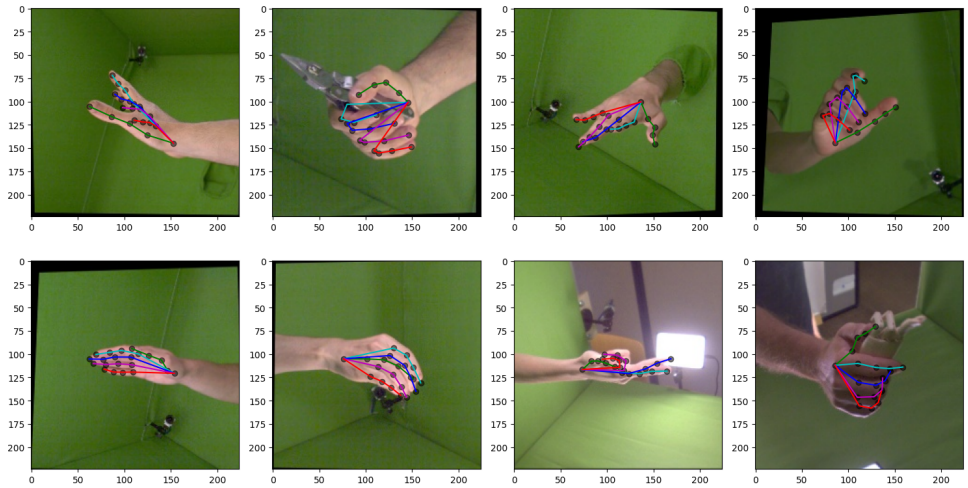


Figure 4.1: Samples of the FreiHAND dataset with their corresponding keypoint configurations. The suitability of this dataset for the purposes of this dissertation is further discussed in the Experimental design chapter. Data obtained from [12].

4.3 Synthetic data generation approach

The methodology and hand pose estimation choices presented in the sections above determine several characteristics the produced synthetic dataset must present. Namely, the resulting system must produce RGB input data and matching keypoint output data. However, as the review evidences, many distinctive approaches have been made to synthesise datasets, each with its advantages and particular considerations. However, all modern synthetic dataset generation techniques for computer vision problems lie in one of four main categories: manual generation, GANs, video-game-assisted, and parametric models.

Manual generation proves counterproductive, as it often undermines the intrinsic advantages of data synthesis, particularly the fundamental aspects of dataset scalability and automated annotation. As this study emphasises the importance of achieving automation, there are better fits for the experiments than these.

Generative Adversarial Networks (GANs) do not represent the ideal solution for this experimentation either, as they rely on real-world data to generate datasets. Furthermore, the model selection process within GANs is obscured and intricately tied to the evaluation network, not allowing for insights into the variations. Nonetheless, the underlying principles of these technologies draw some conceptual parallelism with the methodology described in this section, as both use a neural network trained with real-world data to quantify the quality of synthetic data.

Similarly, the characteristics of video-game-assisted generation make it unfit for this specific experiment. Although the utilisation of video games sometimes presents an inexpensive approach for synthesizing high-quality data [151, 152], no existing approach provides an environment compatible with hand pose estimation, nor do they facilitate the desired degree of control and customisation.

Conversely, contemporary 3D engines do indeed provide the customisation capacities desired [155, 156, 157], making them employable for the objectives of this dissertation. However, delving into the intricacies and challenges of constructing a generation system from scratch adds a compelling element to this

Figure 4.2: View of the baseline configuration of the synthetic dataset generator, showing the procedural hand model in an idle position. The purple cirles represent the projected keypoints.

dissertation. This exploration provides a greater understanding of how such a system can be built and adds greater control to add methods and ideas not contemplated by existing tools.

Hence, parametric models have been selected as the optimal approach for constructing the generation system—the implementation of the system is detailed in the Implementation chapter. Despite the cost and occasional constraints associated with crafting procedural models, they allow for superior domain control and facilitate seamless adaptation of this methodology to other domains.

As previously stated, the addition of proper domain randomisation is likely the most critical matter for obtaining datasets capable of generalising correctly to real-world environments [180, 149]. Therefore, based on the existing literature, the variations introduced in the datasets are arm position, arm orientation, joint angles, skin tone, background, lighting, and shininess. The implementation also regulates the number of variations present within the dataset. This feature holds the potential to offer compelling insights into the impact of each variation. The physical limitations in hand movement are controlled using ranges for randomising the 20 degrees of freedom in the human hand.

The generation system employs one single parametric model for producing the datasets. Although this condition may contrast with the notion of maximising domain randomisation, this decision is due to resource limitations. This condition delves heavily into the problem of bias introduction in datasets, which has been thoroughly discussed in machine learning literature and included in this dissertation. Nevertheless, the prevalent bias stemming from the utilisation of one single model persists across the entirety of the available options, rendering this concern unavoidable even with the inclusion of additional models. This circumstance arises due to the heavy predominance of male, caucasian hands among available models. Although skin tone is tackled by the variations procedurally introduced in the datasets, this lack of diversity may extend to other aspects like hand shape or physical proportions.

Unfortunately, these same biases extend beyond the 3D model realm. It is worth noting that FreiHand—the selected dataset—harbours identical biases. Consequently, these biases do not jeopardise the validity of this study as long as they are found on both sides of the experiment. Instead, it effectively narrows the gap in domain transference, enhancing the conclusiveness of the findings of this study regarding

the synthetic-to-real gap.

Lastly, the method employed for background randomisation holds significant importance when generating a synthetic dataset, as echoed in numerous articles. Mirroring a strategy that has demonstrated efficacy across many studies [132, 181, 10], this system adopts the methodology of integrating an additional dataset as a source of random background images. It uses the UASOL dataset [209] for this purpose. Initially developed for outdoor depth estimation through single and stereo RGB images, the RGB outdoor images within the dataset match well the purpose of background randomisation.

Chapter 5

Implementation

This chapter discusses the implementation aspects of the tools developed for conducting the experiments. The specific technology choices made throughout the project are detailed to guarantee experimental transparency and facilitate reproduction for other researchers, regardless of their familiarity with the subject. The replicability of the experiments described in this dissertation is crucial to its contribution.

It is worth highlighting that code, libraries, and components from external repositories have been incorporated into this implementation. These contributions are duly acknowledged in this chapter through proper citation.

The experiments were executed within a consistent environment, an Apple M1 Pro CPU/GPU [210] operating on macOS Ventura 13.4.1 (c) [211] for the synthetic dataset generation, and Google Colaboratory environments with V100 GPUs and Ubuntu 18.04 [212] for network training. The synthetic generation programme was compiled and executed using Xcode Version 14.3.1 (14E300c) [213], which internally uses the Apple Clang compiler [214].

5.1 Synthetic data generation tool

C++ (C++20 dialect) [215] and OpenGL 4.1 [216] are the primary language and libraries used in the execution of this project. While alternatives like DirectX or Vulkan offer improved performance [217], opting for OpenGL was grounded in its flexibility, capabilities, and well-established graphics programming ecosystem. OpenGL and C++ provide a high level of control for developing customised, efficient code through its low-level capabilities. Furthermore, despite more contemporary solutions, OpenGL remains an industry standard with extensive troubleshooting resources, widely recognised and utilised. Lastly, OpenGL's cross-platform nature enhances the replicability of the resulting tool.

Alongside OpenGL, this implementation relies on Glew (version 2.2.0) [218] and GLFW (version 3.3.8) [219]. These libraries provide a more straightforward API for loading and interacting with OpenGL and its interface components. The selection of these two libraries over alternatives stems from their simplicity and superior performance in the selected system. OpenGL is a complex graphics API with some operating mechanisms that can be intricate for inexperienced developers. However, the internal workings of OpenGL

are not discussed in this document, as they are beyond the scope of this dissertation, and it does not offer any relevant novelty. Similarly, and for the same reason, trivial computer graphics concepts like geometric transformations are not to be addressed in this document either.

The core structure of the synthetic dataset generation tool comprises eleven classes implementing different tasks of the program. This codebase builds upon the code developed for various modules throughout the degree to which this dissertation pertains. Moreover, the base structure of the code, as well as the code employed for bone rigging, was constructed using guidance and example code provided by LearnOpenGL [220]. Although the code has been extensively rewritten, to the point of not sharing the same structure, remnants of the original code, such as functions and code fragments, persist within this implementation.

This code functions through a primary loop producing a single synthetic image on each iteration. Before entering the loop, the scene, objects, and variations-to-be are initialised. Within each iteration, variations are applied to the scene, key points are recomputed, the scene is rendered, and the resulting data is stored in the dataset directory.

The term scene refers to the arrangement of the elements that contribute to the final image that the pipeline renders. In this case, the scene encompasses five distinct elements: the hand model, the light source, the camera, and the background plane. The variations applied to these elements take values within finite ranges to produce images that match the expected results. Table 5.1 shows the different ranges that variations of each attribute can take by randomly sampling them. For example, the position of the hand and the light is determined by shifting it in X, Y, and Z coordinates randomly selected from a range of choices. Then, the hand is randomly rotated following the same approach. These limits guarantee that the hand appears correctly on the image, not getting clipped or being too small. It is worth mentioning that the hand model is normalised to a 1.0x1.0x1.0 bounding box centred in the centre of coordinates. This allows for transformations to be applied with ease. Likewise, joint rotation angles are sampled from their corresponding ranges. However, unlike arm rotation, the joint angle ranges limit movement to realistically provide the 23 degrees of freedom of the hand model depicted in Figure 2.3. Each joint has its corresponding limits and rotation axes. Last, skin tone, shininess, and background selection are also controlled by their own ranges.

The selected variations associated with each attribute are generated at the beginning of the execution. Then, every image takes a random instance among the pre-computed options of each attribute. This mechanic facilitates control over the dimensions of the variation scope. In other words, it enables to specify of a precise count of possible values a variation can take. For example, the tool can be configured to produce a myriad of different positions, orientations and tones, making each image utterly different from the other, but to constrain that only two different backgrounds can be used through the whole dataset. This proceeding allows the detailed assessment of the weight that each variation has in the performance and generalisation of the final model.

However, contrasting with the significant methodological level of detail required to optimise synthetic datasets, it is noteworthy that implementing most of these variations is technologically straightforward from the perspective of contemporary computer graphics. For instance, variating the position, orientation and or joint angles are well-known mechanics not requiring further explanation. Nevertheless, other options are subject to choices that do need to be documented for replicability. Namely, random backgrounds are implemented by applying images of the UASOL dataset [209] as textures of a skybox behind the scene, changing OpenGL's depth function to avoid clipping it. On the other hand, the colour of the

skin is altered using functions based on exposure adjusting. Lastly, the light variations are produced by changing the parameters of the light model: position, intensity, and ambient light.

The geometry transformations and alterations needed to apply these variations are applied to either vertices or colour within the shaders. These shaders, executed in the GPU to produce the final image, implement a modification of the widely known Blinn-Phong shading model [221], enriched with the incorporation of a normal map for detailing. This modification introduces a minimum light level to prevent extreme shadows that deviate from real-world situations. While not achieving full photorealism, the Blinn-Phong shading model produces an acceptable approximation for the standards of this project, especially considering that much of the criticism against it refers to aspects not relevant here—namely, intricate light interactions and material complexities, neither of which are required in this scene.

Most alterations discussed throughout this chapter are directly programmed in the 3D engine, independent of the geometry in use—i.e., geometric transformations and shader parameters. Conversely, applying other changes relies directly on data built into the 3D model. Skeletal rigging, for instance, needs to be included in the model to achieve realistic-looking poses. Thus, due to the scarcity of publicly available hand models possessing suitable characteristics, a commercial model was acquired from the exchange site CGtrader [222]. This model is highly photorealistic and contains an accurate bone structure that can be effectively employed for changing the hand's pose and computing the key points that serve as the annotations of the dataset, aligning their order with those featured in the FreiHand dataset. The model and its skeletal hierarchy are loaded into the scene using the Assimp library (version 5.2.5) [223].

Lastly, the tool developed does not feature an elaborate graphical user interface. Instead, it produces a window displaying the resulting images as they generate. Progress metrics are given to the user through the terminal, and the dataset's configuration is achieved by using constants, providing a structured and predefined approach to dataset customisation.

Attribute	**Type**	**Minimum**	**Maximum**
Finger flexion	Float	-5.0	90.0
Finger abduction	Float	-10.0	10.0
Thumb flexion	Float	-80.0	10
Thumb abduction	Float	-30.0	30.0
Wrist flexion	Float	-90.0	90.0
Wrist abduction	Float	-30.0	30.0
Wrist pronation	Float	-60.0	60.0
Position X	Float	-0.3	0.3
Position Y	Float	-0.3	0.5
Position Z	Float	-0.4	0.4
Rotation angles	Float	-90.0	90.0
Skin tone	Float	0.05	2.0
Light XYZ position	Float	-5.0	5.0
Light intensity	Float	5.0	40.0
Shininess	Float	1.0	50.0
Image	Integer	0	14042

Table 5.1: Type and range that the variations of each attribute can take in the synthetic data generation process.

5.2 Hand pose estimation

The evaluation step operates independently of the data synthesis process. Consequently, relying on available code bases for training the hand pose estimation models is sufficient, requiring minimal compatibility adjustments. The primary requirement is to structure the data in a format that can be seamlessly integrated into the system without extensive modifications.

Python (version 3.9.6) [224] is the language of choice for the training process, alongside the machine learning library PyTorch [225]. Python is currently the most prevalent choice of language for machine learning purposes. In addition, PyTorch stands out among contemporary deep-learning libraries due to its simple and transparent API. These characteristics significantly ease interaction with the model, making the training and inference less burdensome.

While adapted to accommodate the generated datasets, the most significant share of the code for training and inferring the Shallow UNet network has been obtained from the repository authored by Olga Chernytska [226]. This project, publicly available on GitHub, aligns with the choices outlined in the Experimental Design chapter of this dissertation. In particular, it uses a modified UNet model for training a neural network to estimate hand poses. The reasoning behind the training choices is comprehensively discussed in the author's Master's Thesis [227].

In this code, the training process employs the Intersection over Union Loss, which is particularly suited for tasks involving geometric differences as an accuracy criterion, aiming to minimise disparities. The chosen optimiser is the Stochastic Gradient Descent, employed to prevent training issues. To guarantee convergence despite SDG, the Learning Rate Scheduler is used. This scheduler continually evaluates the validation loss for plateaus throughout the training, indicating a slowdown in the model's progress. In those cases, the scheduler reduces the learning rate enabling the model to make more nuanced adjustments.

In addition, an analogous code is used to infer and evaluate the produced models. This program loads the trained models, feeds an unseen test set of the FreiHand dataset to the trained model, and compares the outputs with the ground truth data.

Chapter 6

Results

This chapter presents the resulting metrics and results of the experiments. The implications and significance of these findings are subsequently examined in the Discussion chapter. The numerical results showcased stem from thorough experimentation, precise data gathering, and meticulous analysis.

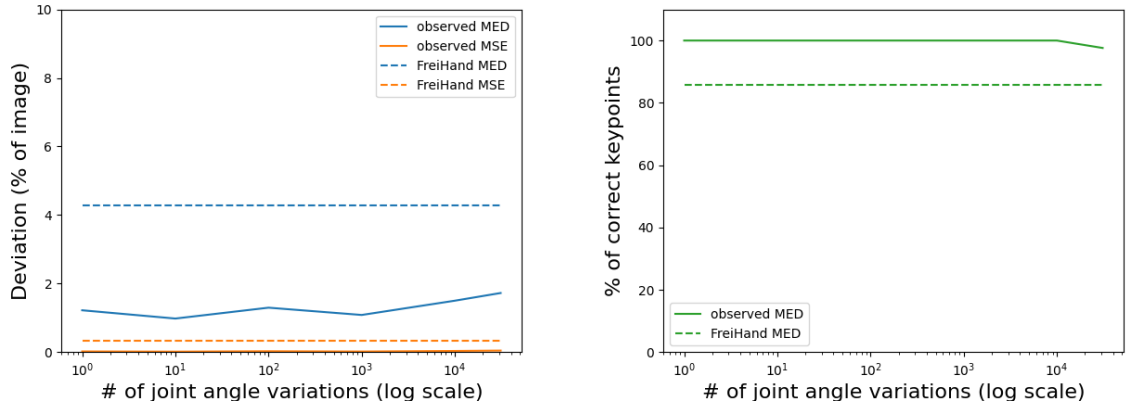
Since the desired output of hand pose estimation is the correct identification and location of the key points, the system needs to assess the validity and accuracy of its predictions by using point-to-point comparisons. Therefore, this chapter uses the Mean Euclidian Distance, Mean Square Error, and Percentage of Correct Keypoints as the result metrics. For this project, the criterion of keypoint correction has been established as a deviation of 5 pixels or less from the ground truth data.

The process of applying different amounts of variations to each attribute has produced a total of 41 different datasets, which have been used to infer synthetic and real-world hand images.

6.1 Joint angles

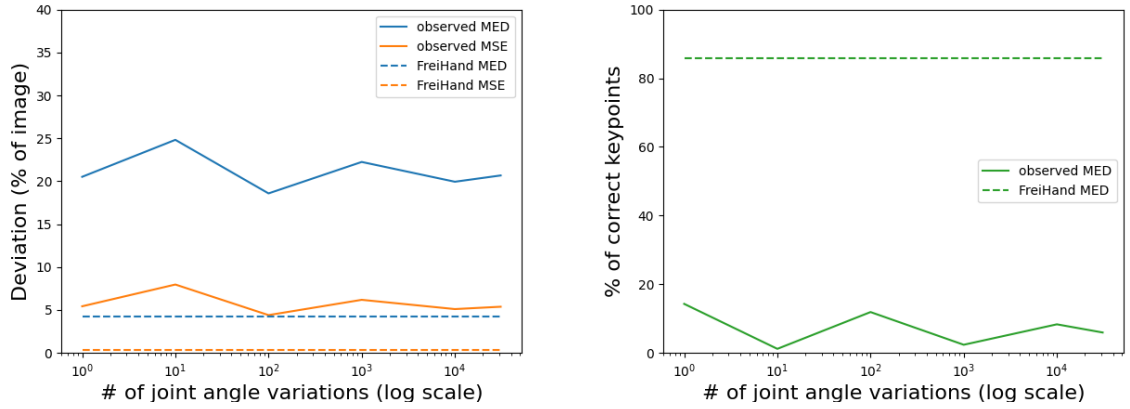
Figure 6.1 shows that the ability of the trained models to estimate different hand positions, without any other variation present, remains close to 100% across all cases, regardless of the number of introduced variations. Only a slight decrease in the maximum value of the Figure is observed, where all images have completely distinct positions from each other. Similarly, in these cases, the Mean Euclidian Distance and Mean Squared Error remain stable, well positively below the benchmark.

As shown in Figure 6.2, the results of generalising this model to infer real-world data are much less favourable. In this Figure, the three metrics also remain relatively stable, despite the various additional variations in hand joint rotations. However, the values remain negative in this case, with high error and low correct points. Fluctuation is observed in the plots, although no evident positive or negative trend exists.



(a) (b)

Figure 6.1: Comparison of accuracy metrics to infer unseen synthetic data on synthetic datasets with varying joint angle variations. a) Solid blue line: Mean Euclidian Distance; solid orange line: Mean Squared Error; dashed blue line: Benchmark MED (Real-world data model); dashed orange line: Benchmark MSE (Real-world data model). b) Solid green line: Percentage of Correct Keypoints; dashed blue line: Benchmark PCK (Real-world data model).



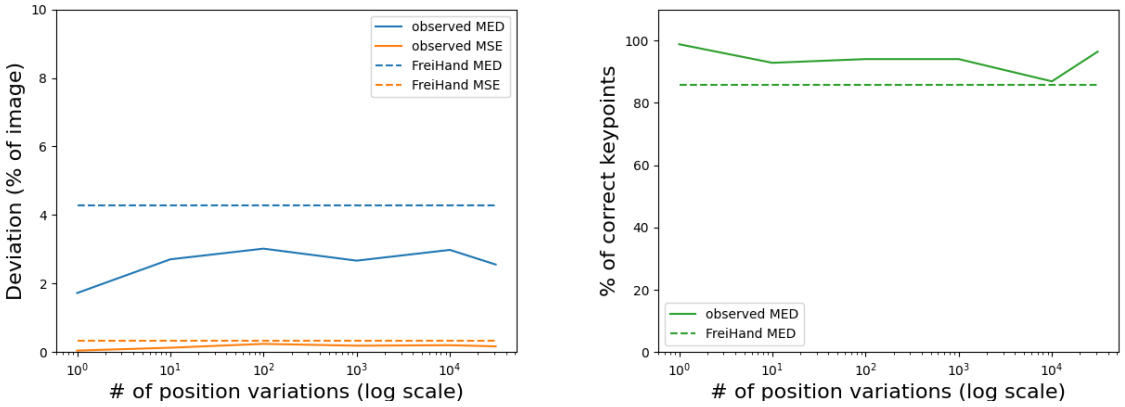
(a) (b)

Figure 6.2: Comparison of accuracy metrics to infer real-world data on synthetic datasets with varying joint angle variations. The meaning of a) and b) remain as seen in Figure 6.1.

6.2 Position

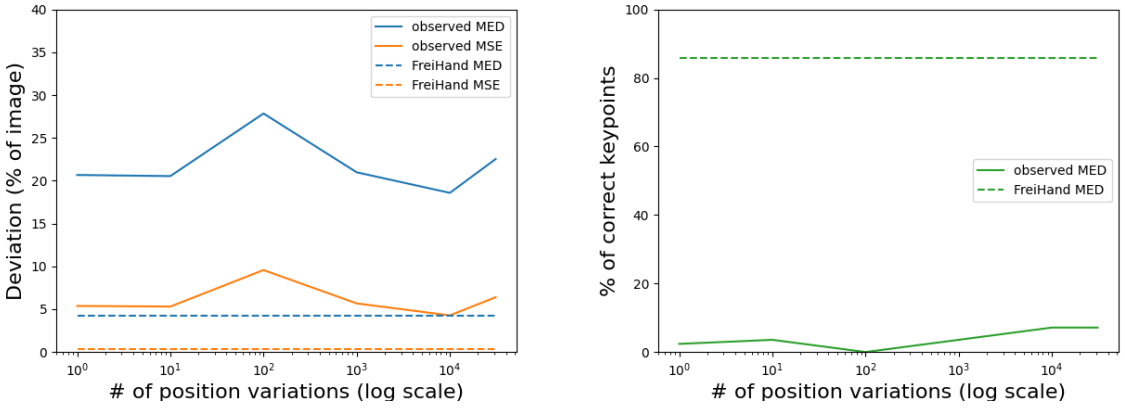
Once again, Figure 6.3 presents similar results to those shown in Figure 6.1, especially regarding the curve's stability. While this figure does not display results as extreme as in the case with no variations, the plot illustrates that the detection quality remains relatively steady, slightly hinting a downward trend.

As for the impact of positional variations on the detection quality using real images, a modest contrast regarding the previous Figures is tangible in Figure 6.4. In this case, albeit minor, an upward trend in hand pose estimation becomes apparent as more variations are introduced. Nevertheless, the attained data remains quite low. Such a small variation could be easily ignored, but being the magnitudes in that metric so low, they must be taken into account.



(a) (b)

Figure 6.3: Comparison of accuracy metrics to infer unseen synthetic data on synthetic datasets with varying hand positions. The meaning of a) and b) remain as seen in Figure 6.1.



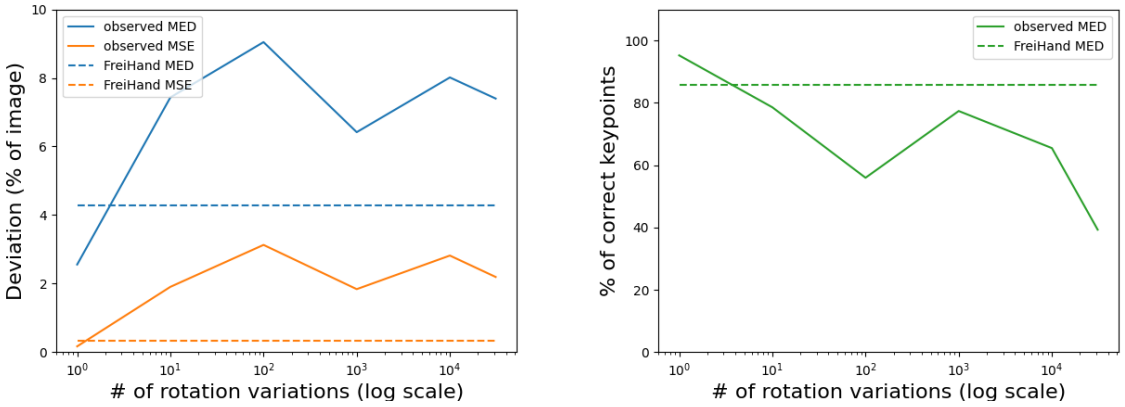
(a) (b)

Figure 6.4: Comparison of accuracy metrics to infer real-world data on synthetic datasets with varying hand positions. The meaning of a) and b) remain as seen in Figure 6.1.

6.3 Rotation

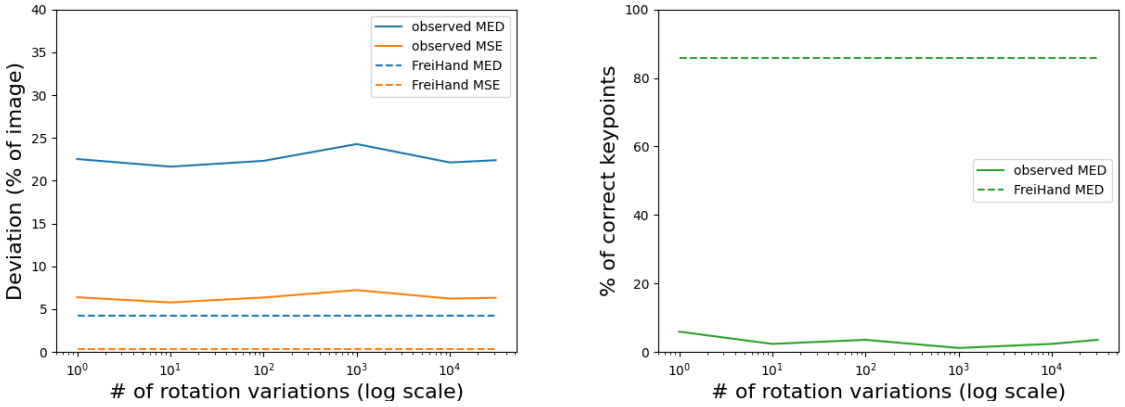
In contrast to the modest variations displayed by the previous plots, Figure 6.5 demonstrates a clear deterioration as additional rotations are introduced. The total difference observed amounts to approximately a 60% reduction in the percentage of correct key points. Furthermore, concerning the Mean Euclidean Distance and Mean Squared Error, they exhibit changes in the opposite direction that align perfectly with this decline in detection quality.

Conversely, these variations seem to have little to no impact on detecting real-world images, as the metric trends remain highly stable across the diagram in Figure 6.6. While Figure 6.5 showcased a notable decline in quality, this second plot shows values that persist as poor outcomes with little alteration.



(a) (b)

Figure 6.5: Comparison of accuracy metrics to infer unseen synthetic data on synthetic datasets with varying hand orientations. The meaning of a) and b) remain as seen in Figure 6.1.



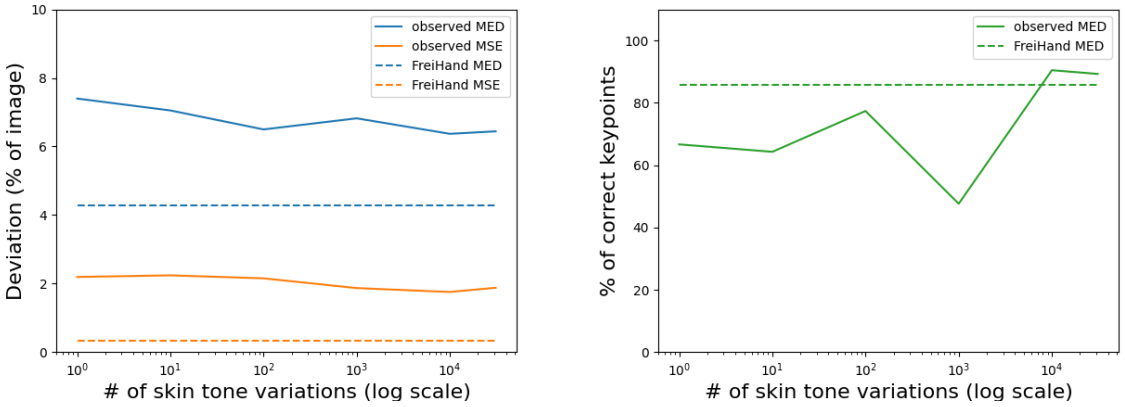
(a) (b)

Figure 6.6: Comparison of accuracy metrics to infer real-world data on synthetic datasets with varying hand orientations. The meaning of a) and b) remain as seen in Figure 6.1.

6.4 Skin tone

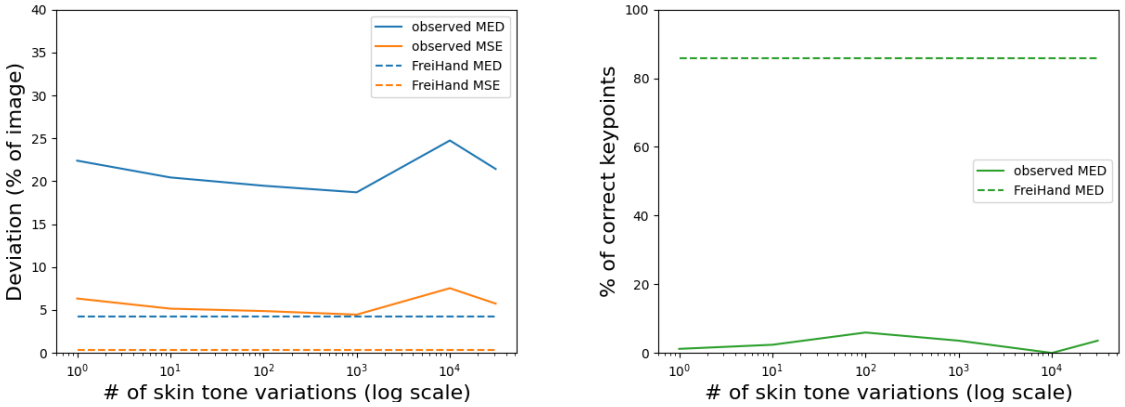
Figures 6.7 and 6.8 depict an interesting behaviour when considering the four plots together. Firstly, the Mean Euclidean Distance and Mean Squared Error show a consistent level of stability as skin tone variations are added to the dataset. However, the percentage of correct key points showcases a notable contrast, as a modest yet somewhat erratic improvement can be seen in the plot.

Despite this increase in the PCK when inferring synthetic images, the results do not translate into comparable outcomes in the context of real-world images. In Figure 6.8, a quite stable behaviour can be seen throughout the different amounts of orientations. This pattern echoes observations made in other analogous cases, further emphasising the distinct nature of real-world image detection challenges.



(a) (b)

Figure 6.7: Comparison of accuracy metrics to infer unseen synthetic data on synthetic datasets with varying skin tones. The meaning of a) and b) remain as seen in Figure 6.1.



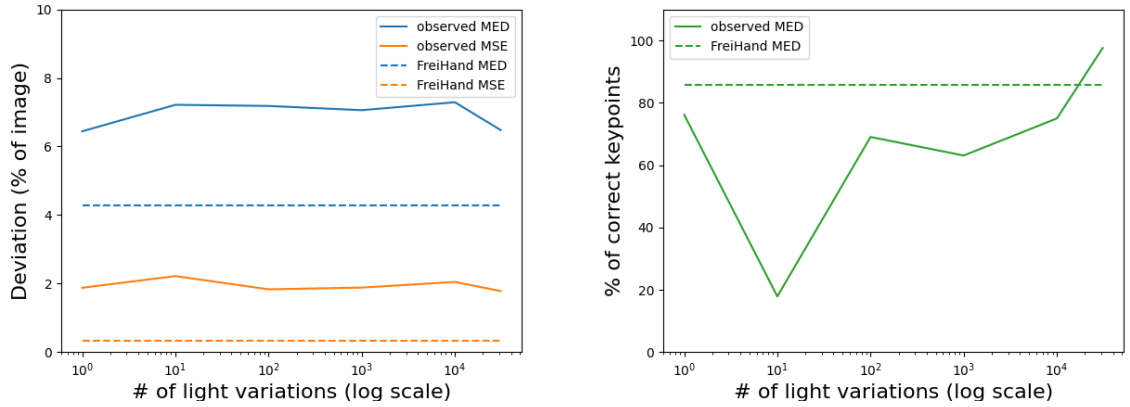
(a) (b)

Figure 6.8: Comparison of accuracy metrics to infer real-world data on synthetic datasets with varying skin tones. The meaning of a) and b) remain as seen in Figure 6.1.

6.5 Light

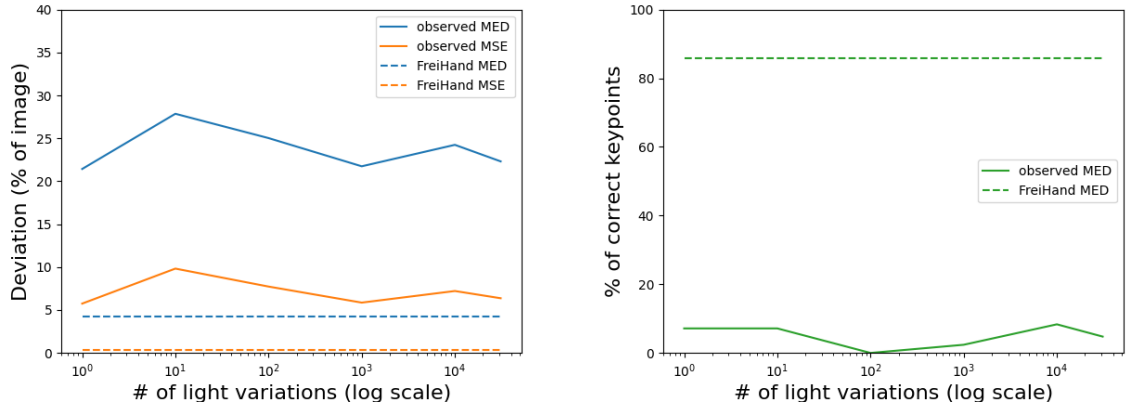
Figure 6.9 portrays intriguing behaviour, whereas while Figure 6.9a exhibits minimal variations, Figure 6.9b displays highly erratic changes, albeit with a positive trend. However, this trend requires a more comprehensive discussion to make sense of it in the Discussion chapter.

In this instance, Figure 6.10 does exhibit a more direct correlation with Figure 6.9. While there is no distinct positive trend in this case, we can indeed discern the erratic behaviour seen in the previous graph. However, this behaviour occurs on a much smaller scale due to its considerably lower overall values.



(a) (b)

Figure 6.9: Comparison of accuracy metrics to infer unseen synthetic data on synthetic datasets with varying light settings. The meaning of a) and b) remain as seen in Figure 6.1.



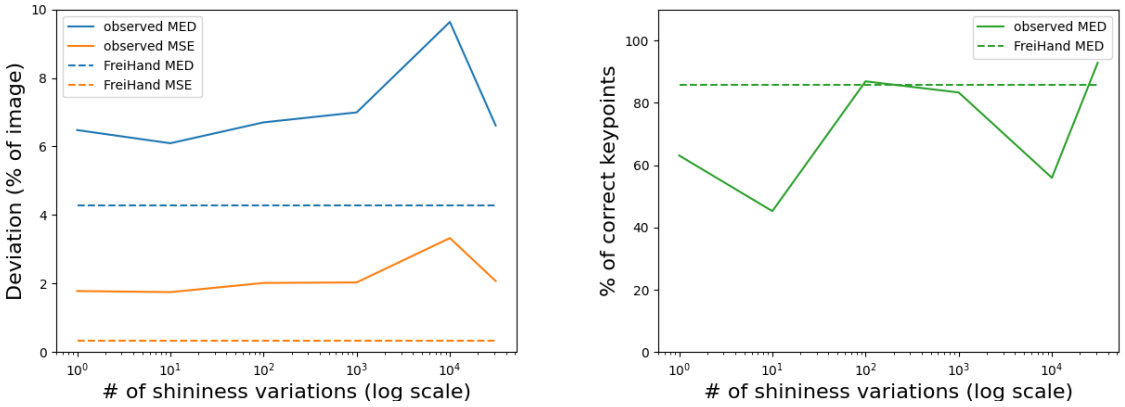
(a) (b)

Figure 6.10: Comparison of accuracy metrics to infer real-world data on synthetic datasets with varying light settings. The meaning of a) and b) remain as seen in Figure 6.1.

6.6 Shininess

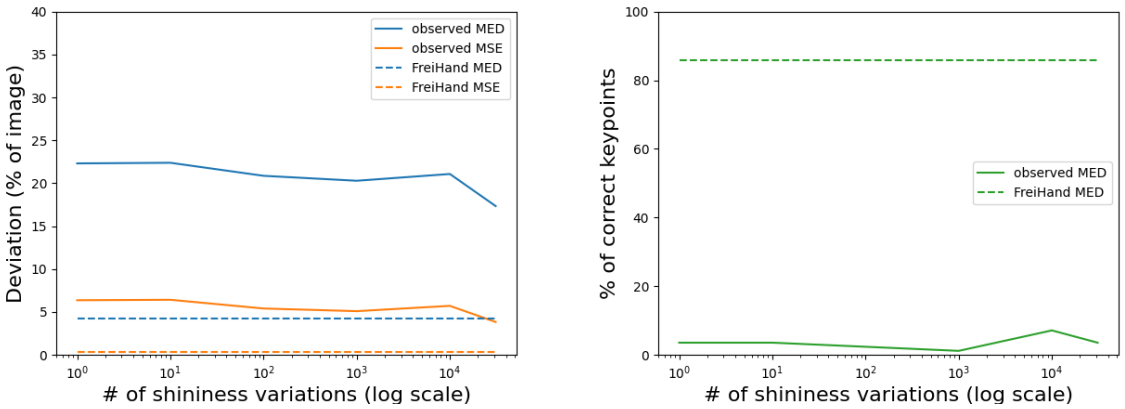
Figure 6.11 again illustrates erratic behaviour, albeit with a positive trend regarding PCK. On the other hand, the error fluctuates around a consistent value, demonstrating a pattern that suggests a growing trend, although the final value decreases once more. This behaviour gives rise to uncertainty as to whether it indicates a rising or stable trend.

Figure 6.12, however, aligns with the trends observed in previous figures about real-world image inference. Once more, the values remain almost unchanged, despite alterations in the synthetic dataset variations.



(a) (b)

Figure 6.11: Comparison of accuracy metrics to infer unseen synthetic data on synthetic datasets with varying shininess levels. The meaning of a) and b) remain as seen in Figure 6.1.



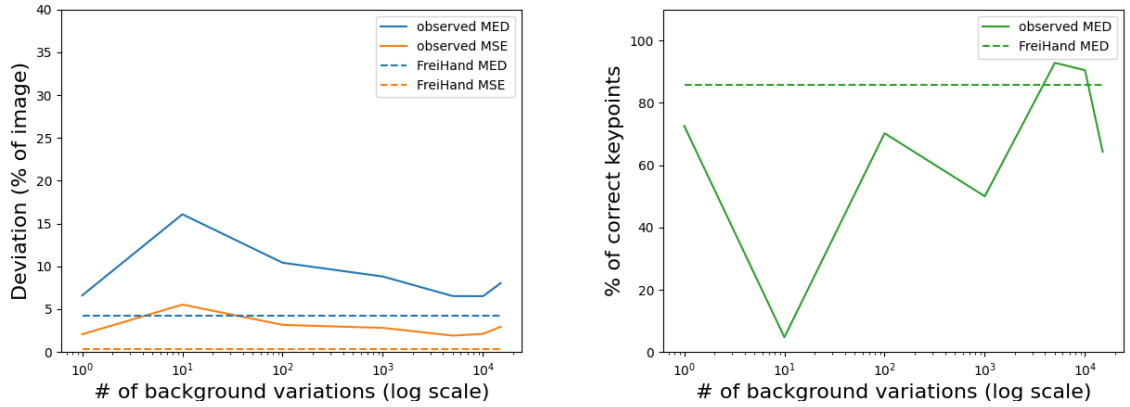
(a) (b)

Figure 6.12: Comparison of accuracy metrics to infer real-world data on synthetic datasets with varying shininess levels. The meaning of a) and b) remain as seen in Figure 6.1.

6.7 Background

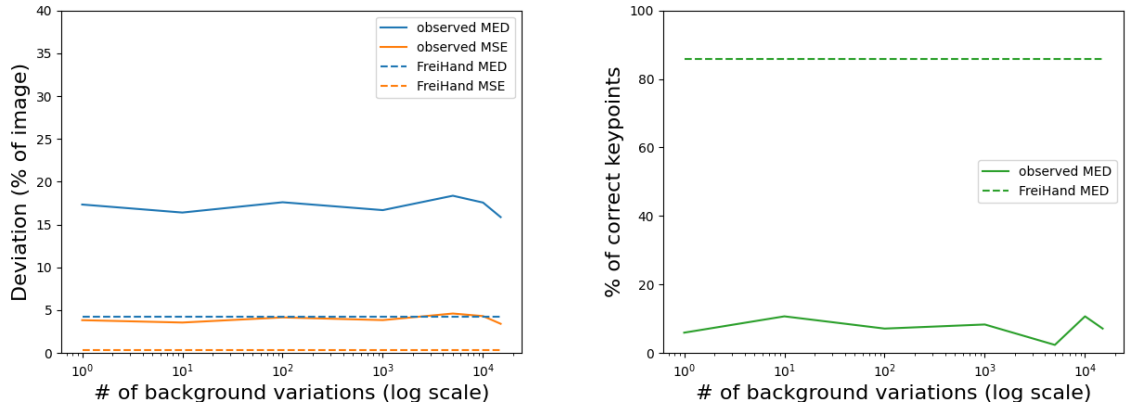
Figure 6.13, concerning the introduction of random backgrounds in synthetic images, undeniably exhibits the most erratic behavior among all variables depicted in graphs within this section. With the inclusion of random backgrounds, the error experiences a noticeable surge, although gradually receding as more backgrounds are added to the image. Similarly, the PCK demonstrates a corresponding behavior, albeit in reverse.

Figure 6.14 once more presents a stable behavior, with minimal alterations upon the introduction of random backgrounds, both in terms of errors and the percentage of correct key points.



(a) (b)

Figure 6.13: Comparison of accuracy metrics to infer unseen synthetic data on synthetic datasets with varying background variations. The meaning of a) and b) remain as seen in Figure 6.1.



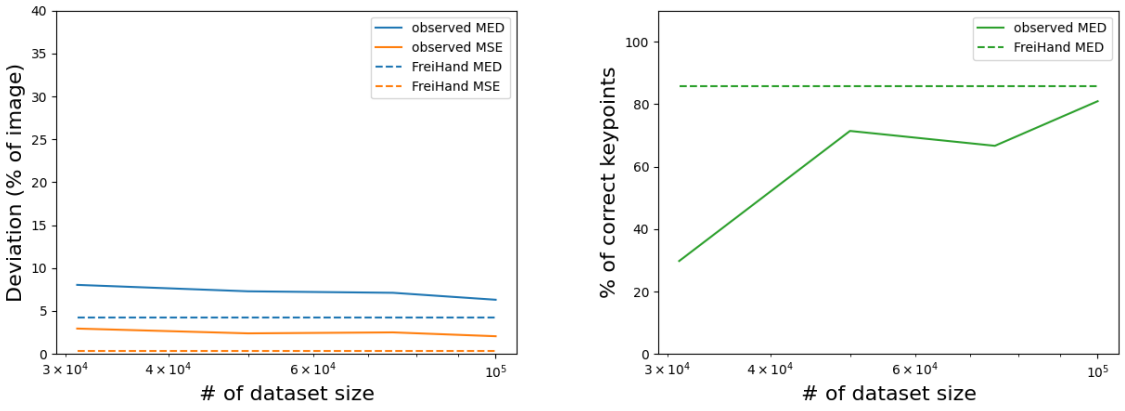
(a) (b)

Figure 6.14: Comparison of accuracy metrics to infer real-world data on synthetic datasets with varying background variations. The meaning of a) and b) remain as seen in Figure 6.1.

6.8 Dataset size

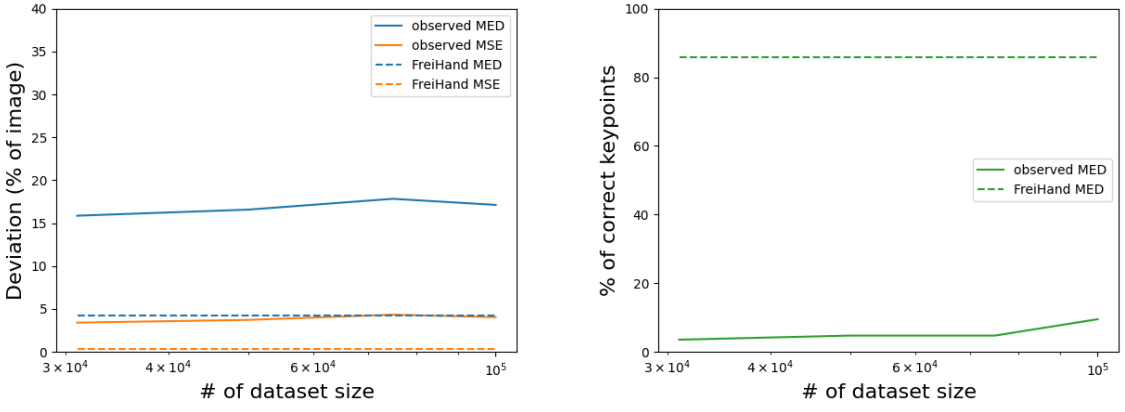
Figure 6.15, undeniably exhibits the most notorious performance increase among all variations, taking the PCK metric from a 35% to an outstanding 78%.

Figure 6.16 shows an increasing tendency as dataset size increases, although the scale of the improvement is very limited.



(a) (b)

Figure 6.15: Comparison of accuracy metrics to infer unseen synthetic data on synthetic datasets with varying dataset sizes. The meaning of a) and b) remain as seen in Figure 6.1.



(a) (b)

Figure 6.16: Comparison of accuracy metrics to infer real-world data on synthetic datasets with varying dataset sizes. The meaning of a) and b) remain as seen in Figure 6.1.

Chapter 7

Discussion

This section discusses the implications of the outcomes presented in the Results section. Since the results have two distinct trends, separate discussions about hand pose estimation from synthetic and real-world images are conducted. This disparity aligns with the hypotheses put forth by some works discussed in the literature review, which stress the criticality of the synthetic-to-real gap.

However, despite this distinction, each section also includes cross-observations, contributing context and noteworthy insights for effectively interpreting the data and understanding the complete topic.

7.1 Hand pose estimation from synthetic inputs

The results obtained from the experiments provide a deeper understanding of the intricacies of building a synthetic data generation system and how the produced data must correlate with an equivalent dataset acquired from the real world. In certain instances, the results either confirm or refute hypotheses formulated early in this document. In other cases, the data introduce new questions and reflections, enabling to draw of unexpected conclusions from the experiments.

Figure 6.1 serves as an example of this phenomenon. This plot shows an almost impeccable ability to estimate the pose of synthetically generated hands from images accurately. This unexpected result imparts insightful information in a manner not initially envisaged. It is essential to acknowledge that the primary intent behind these initial datasets was precisely to showcase overfitting scenarios to improve the model's performance beyond that point. However, the resultant outcome diverges from the expected result. While it holds true that these datasets exhibit minimal variability, rendering them exceptionally susceptible to overfitting, these models show an ability to infer unseen data that goes beyond expectations. This behaviour suggests that the neural network possesses surplus capacities to discern patterns within these synthetically generated images. Despite its non-apparent nature, this conclusion is promising. It prompts the inclination to believe these synthetic images inherently encapsulate some requisites to function as viable inputs eventually.

Figure 6.3 further prompts thinking in that direction, as by introducing numerous additional variations, a consistently higher accuracy than that of the real-world dataset is maintained, even for previously unseen

images. Similarly, although depicting very different results, Figure 6.5 suggests interesting information from which conclusions can be drawn. Examining the FreiHand dataset reveals a deliberate and evident absence of images featuring substantial occlusions. The tool proposed in this dissertation for synthetic dataset generation does not incorporate any such restrictions, resulting in the inclusion of exceptionally challenging or even impossible images to estimate. This contributes insignificantly to training and introduces deceptive images into the test set.

The approach of consciously removing specific characteristics from the dataset suggests something quite valuable when generating datasets and could be a consideration for potential system enhancements in the future. Inputs lacking pertinent information required for generating output are not desirable within the dataset. While it might seem counterintuitive, as they represent real-world examples a neural network could encounter, on the other hand, if there is no discernible pattern from which to learn, the only likely outcome is the potential confusion of the neural network.

In addition, Figure 6.7 presents an intriguing question: Why does the introduction of variations in skin tone lead to improved detection of synthetic inputs? It is not apparent that augmenting variability within other aspects of the dataset could enhance precision on top of making it more generalisable, but such an outcome is actually possible.

However, this circumstance gains more clarity by observing Figure 6.9. While they may not possess the same form, these two figures have a certain parallelism. In both cases, they display a scenario of high volatility in detection quality metrics, yet they share an overall increasing trend: as new variations are introduced, detection quality slowly improves. The reason behind this occurrence likely resides in colour variability within the hand, as both lighting and skin tone variations fundamentally involve colour differences. A similar rationale presumably underlies the other two graphs as well. More significant variability compels the neural network to identify common patterns, potentially pushing it to find a better-fitting minimum during the gradient descent process. In easier problems with less variability, a different, less optimal, relative minimum might have been reached.

All the figures show an evident instability in the metrics obtained by inferring synthetic datasets in models trained on themselves. These outcomes indicate that the system employs a logic for generating variations that are not optimal for achieving the best detection. The craft of data selection is a widely recognised concept in Machine Learning, involving data manipulation to attain an optimally suited set for our system, despite it not necessarily being a flawless reflection of the real world it aims to model.

In this case, certain variations have been introduced, which might not be optimal to include: consider specific combinations of rotations and positions that obscure parts of the hand, specific lighting parameters that hinder accurate element differentiation, or variations that render the images less realistic and, consequently, less readily extrapolated to the real-world domain.

The observation regarding the dataset size increase illustrated in Figure 6.15 depicts the relationship between dataset size and performance, and signifies a remarkable breakthrough in this dissertation. As the dataset size expands, the dataset's ability to produce better models improves as well, substantiating the notion that the path of scaling the solution could potentially lead to achieving genuine generalisation.

7.2 Hand pose estimation from real-world inputs

This project's main objective is to transfer the data and patterns from the synthetic domain to the real-world domain, enabling the use of models trained with synthetic data for real-world problems. Within the process of assessing the feasibility of such a tool, there are specific expected and unexpected outcomes that the experiments' results have pointed out.

The behaviour observed in Figures 6.2 and 6.4 is precisely the expected behaviour for those datasets, as the synthetic data intentionally incorporates minimal variations, making it challenging to infer real-world hand position data in any way.

However, an intriguing observation is that Figure 6.2 achieves better results than Figures 6.4 and 6.6. This behaviour can likely be attributed to an uneven spatial distribution of data in the FreiHand dataset. The synthetic datasets in Figure 6.2 consist of data with variations limited solely to join angle settings. In other words, the hand is always centred on the image and consistently facing away from the camera.

A model trained on such data exhibits a distinct tendency to produce key points with a similar pattern due to the overfitting induced by the high data repetition. Therefore, the fact that Figure 6.2 displays better precision metrics than Figures 6.4 and 6.6 likely indicates that images close to the original hand position in the synthetic dataset are much more prevalent than others. While not directly leading to immediate system functionality, these types of observations can contribute to adapting the generated datasets to desired domains to refine their outcomes.

However, the one genuinely conclusive conclusion that can be drawn from the results obtained is that models trained on these artificial datasets cannot at all bridge the synthetic-to-real gap. All resulting graphs from inferring real-world images with models trained on synthetic images exhibit inferior results, resulting in an absolute inability to detect hand positions in real-world images. This possibility had already been anticipated in the literature review, as many authors suggested that this was the major challenge encountered in synthetic dataset creation. The obtained data corroborate these hypotheses.

Nevertheless, the Figures show that the different numbers of variations barely impact performance metrics, which consistently remain at minimal levels. The case most notably inducing a change is in Figure 6.14, where it observed that random backgrounds seem to impact the obtained results. However, these remain at levels far from being considered a successful system for hand pose detection in real-world images.

The mention of the dataset size increase in Figure 6.16 suggests that as the dataset size grows, there is potential for promising results to be achieved. This could imply that larger datasets might lead to better performance, more accurate predictions, and improved outcomes in the context of generalisation to the real-world domain. However, it is noted that despite the potential benefits, the exploration of this path has been limited due to resource and time constraints.

Chapter 8

Conclusions

The advancement of computer vision systems has brought about a revolution driven by the formidable capabilities of deep neural networks. However, as systems grow in complexity, the demand for extensive datasets becomes urgent. Creating these large datasets with the required characteristics presents significant labour and resource challenges.

Consequently, the exploration of procedural generation techniques has gained relevance as a means to overcome this hurdle. Controlled variations impact the quality of detection and the reliability of the resulting system, although often proving insufficient for bridging the domain gap.

This dissertation delves into these issues within the context of hand pose estimation. To assess the efficacy of the generated datasets, an advanced computer vision system is trained to detect key points in hand images, leveraging both the procedurally generated dataset and traditionally annotated datasets. Comparative analyses then scrutinise the system's performance on real-world data, examining the influence of procedural variations on accuracy, robustness, and generalisation capabilities.

The findings in this dissertation can be used as a starting point towards building a generic pipeline capable of generating synthetic datasets from automated variations and realistic 3D models. However, further work needs to improve in bridging the synthetic-to-real gap to create usable datasets. This conclusion echoes discussions in the literature and is reinforced by the literature review within this dissertation: domain transfer remains the major challenge in synthetic dataset generation, a claim supported by the results presented here.

These conclusions pave the way for future research to build upon this approach, specifically focusing on minimising this gap and achieving effective generalisation from synthetic to real-world data. This research has provided valuable insights into the intricacies of constructing such systems and an in-depth understanding of the history of the field and current approaches.

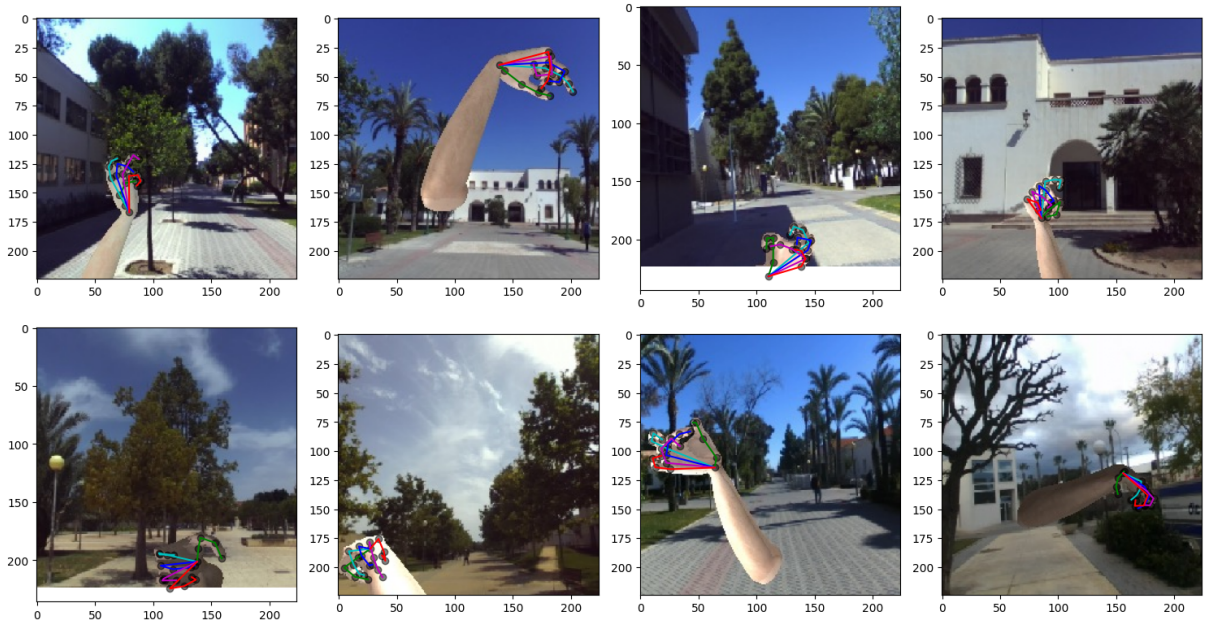


Figure 8.1: Detail of a resulting dataset with random backgrounds.

8.1 Limitations

This section acknowledges the existing limitations of the research due to scope, resources, and time constraints. These limitations have influenced the extent of the research and may affect the implications of the findings.

First, this project has inevitably limited the range of variations explored within the study due to the limited resources and time available. Generating synthetic datasets and training neural networks are resource-intensive processes. The computational demands and time required to generate and process these datasets have posed significant challenges. Therefore, simplifications had to be applied to some seemingly minor aspects of the project, and variation ranges were limited to reduce the number of examples to compute. This limitation could potentially have impacted the robustness of the trained models.

Besides, as discussed, using one 3D hand model in this project introduces inherent bias. Although several variations may have reduced these biases, the selected 3D model cannot capture the full diversity of real-world hand variations.

Finally, combining all the elements in the pipeline and producing the final image may also introduce limitations to the system by adding visual bias, cues and artefacts. On the one hand, all simplified shading models introduce unrealism to some degree. Secondly, the images produced are small JPG images (224x224 pixels) to match the original dataset and reduce execution times. However, small image rendering alongside JPG compression produces artefacts that can affect detection.

Chapter 9

Future work

The central finding of this dissertation underscores that closing the gap between synthetic and real domains is the path to achieving the collective objectives within this research domain. While numerous research lines are currently open, there is space for potential additions to the methodological framework presented in this dissertation.

Firstly, several improvements have already been introduced across sections of this dissertation. The main improvement would be to leverage the apparent benefit of increasing the dataset size to obtain more generalisable datasets. To train these datasets, important processing resources are needed.

Other options include the incorporation of more realistic shaders, introducing additional variations, integrating more 3D models to mitigate bias and enhance generalisation, addressing the challenge of inefficient data discarding, and more.

Moreover, this research can be continued by employing statistical methodologies to find the intricate interrelationships and collective impacts of diverse variations on generalisation capabilities. A deeper comprehension can be obtained by treating the research as a singular multi-dimensional space.

Another strategic step involves subjecting the synthetic data generator to evaluation using diverse neural network architectures. This investigation could uncover better performance of these architectures to the specific problem.

Given the importance of background selection in the literature, a novel direction could research the generation of randomized backgrounds as integral components of the scenes instead of selecting pre-existing backgrounds. Similar techniques, successfully employed in texture generation, have demonstrated their potential to train systems to recognise shapes.

Finally, strategically incorporating Generative Adversarial Networks (GANs) could be pivotal. By employing GANs, the images generated could undergo domain transfer, aligning them more closely with the patterns exhibited by real-world data.

Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[2] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020, number: 2 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[3] M. R. Villarreal, "English: Main division on the (right) human hand." Jan. 2007. [Online]. Available: https://commons.wikimedia.org/wiki/File:Scheme_human_hand_bones-en.svg

[4] M. Oliveira, H. Chatbri, S. Little, Y. Ferstl, N. E. O'Connor, and A. Sutherland, "Irish Sign Language Recognition Using Principal Component Analysis and Convolutional Neural Networks," in *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov. 2017, pp. 1–8.

[5] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," Feb. 2015, arXiv:1405.0312 [cs]. [Online]. Available: http://arxiv.org/abs/1405.0312

[6] K. von Neumann-Cosel, E. Roth, D. Lehmann, J. Speth, and A. Knoll, "Testing of Image Processing Algorithms on Synthetic Data," in *2009 Fourth International Conference on Software Engineering Advances*, Sep. 2009, pp. 169–172.

[7] P. S. Rajpura, H. Bojinov, and R. S. Hegde, "Object Detection Using Deep CNNs Trained on Synthetic Images," Sep. 2017, arXiv:1706.06782 [cs]. [Online]. Available: http://arxiv.org/abs/1706.06782

[8] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5908–5916, iSSN: 2380-7504.

[9] "GANerated Hands Dataset." [Online]. Available: https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/GANeratedDataset.htm

[10] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning Optical Flow with Convolutional Networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 2758–2766, iSSN: 2380-7504.

[11] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla, "Model-based hand tracking using a hierarchical Bayesian filter," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1372–1384, Sep. 2006, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[12] 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," Sep. 2019, arXiv:1909.04349 [cs]. [Online]. Available: http://arxiv.org/abs/1909.04349

[13] Z. Ma, H. Ling, Y.-Z. Song, T. Hospedales, W. Jia, Y. Peng, and A. Han, "IEEE Access Special Section Editorial: Recent Advantages of Computer Vision," *IEEE Access*, vol. 6, pp. 31 481–31 485, 2018, conference Name: IEEE Access.

[14] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023, conference Name: Proceedings of the IEEE.

[15] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," *Computational Intelligence and Neuroscience*, vol. 2018, p. e7068349, Feb. 2018, publisher: Hindawi. [Online]. Available: https://www.hindawi.com/journals/cin/2018/7068349/

[16] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, "Survey on Deep Neural Networks in Speech and Vision Systems," *Neurocomputing*, vol. 417, pp. 302–321, Dec. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220311619

[17] A. Lapedes and R. Farber, "How Neural Nets Work," in *Evolution, Learning and Cognition*. WORLD SCIENTIFIC, Jan. 1989, pp. 331–346. [Online]. Available: https://www.worldscientific.com/doi/10.1142/9789814434102_0012

[18] R. Uhrig, "Introduction to artificial neural networks," in *Proceedings of IECON '95 - 21st Annual Conference on IEEE Industrial Electronics*, vol. 1, Nov. 1995, pp. 33–37 vol.1.

[19] M. Bianchini and F. Scarselli, "On the Complexity of Neural Network Classifiers: A Comparison Between Shallow and Deep Architectures," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1553–1565, Aug. 2014, conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[20] Z. Buczolich, "Solution to the gradient problem of C. E. Weil," *Revista Matemática Iberoamericana*, pp. 889–910, 2005. [Online]. Available: https://ems.press/doi/10.4171/rmi/439

[21] G. Philipp, D. Song, and J. G. Carbonell, "The exploding gradient problem demystified - definition, prevalence, impact, origin, tradeoffs, and solutions," Apr. 2018, arXiv:1712.05577 [cs]. [Online]. Available: http://arxiv.org/abs/1712.05577

[22] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," Feb. 2017, arXiv:1611.03530 [cs]. [Online]. Available: http://arxiv.org/abs/1611.03530

[23] Z. Du, X. Li, and J. Wu, "Accelerating the Training of HTK on GPU with CUDA," in *2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum*, May 2012, pp. 1907–1914.

[24] S. Pal, E. Ebrahimi, A. Zulfiqar, Y. Fu, V. Zhang, S. Migacz, D. Nellans, and P. Gupta, "Optimizing Multi-GPU Parallelization Strategies for Deep Learning Training," *IEEE Micro*, vol. 39, no. 5, pp. 91–101, Sep. 2019, conference Name: IEEE Micro.

[25] G. Albuquerque, T. Lowe, and M. Magnor, "Synthetic Generation of High-Dimensional Datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2317–2324, Dec. 2011, conference Name: IEEE Transactions on Visualization and Computer Graphics.

[26] S. Mehta, M. Dawande, G. Janakiraman, and V. Mookerjee, "How to Sell a Dataset? Pricing Policies for Data Monetization," Rochester, NY, Aug. 2019. [Online]. Available: https://papers.ssrn.com/abstract=3333296

[27] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in Deploying Machine Learning: A Survey of Case Studies," *ACM Computing Surveys*, vol. 55, no. 6, pp. 114:1–114:29, Dec. 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3533378

[28] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and Information Systems*, vol. 34, no. 3, pp. 483–519, Mar. 2013. [Online]. Available: https://doi.org/10.1007/s10115-012-0487-8

[29] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: A narrative review," *PLOS Digital Health*, vol. 2, no. 1, p. e0000082, Jan. 2023, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082

[30] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Data augmentation using synthetic data for time series classification with deep residual networks," Aug. 2018, arXiv:1808.02455 [cs]. [Online]. Available: http://arxiv.org/abs/1808.02455

[31] W. Jiang, K. Zhang, N. Wang, and M. Yu, "MeshCut data augmentation for deep learning in computer vision," *PLOS ONE*, vol. 15, no. 12, p. e0243613, Dec. 2020, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0243613

[32] P. Kaur, B. S. Khehra, and E. B. S. Mavi, "Data Augmentation for Object Detection: A Review," in *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, Aug. 2021, pp. 537–543, iSSN: 1558-3899.

[33] A. Kortylewski, A. Schneider, T. Gerig, B. Egger, A. Morel-Forster, and T. Vetter, "Training Deep Face Recognition Systems with Synthetic Data," Feb. 2018, arXiv:1802.05891 [cs]. [Online]. Available: http://arxiv.org/abs/1802.05891

[34] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, and G. Epelde, "Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing," *JMIR Medical Informatics*, vol. 8, no. 7, p. e18910, Jul. 2020, company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://medinform.jmir.org/2020/7/e18910

[35] Q. D. Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. L. Saux, and D. Filliat, *3D Hand Gesture Recognition Using a Depth and Skeletal Dataset*. The Eurographics Association, 2017, accepted: 2017-04-22T17:17:41Z ISSN: 1997-0471. [Online]. Available: https://diglib.eg.org:443/xmlui/handle/10.2312/3dor20171049

[36] Y. Cai, L. Ge, J. Cai, N. M. Thalmann, and J. Yuan, "3D Hand Pose Estimation Using Synthetic Data and Weakly Labeled RGB Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3739–3753, Nov. 2021, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[37] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, "3D Hand Shape and Pose Estimation from a Single RGB Image," Apr. 2019, arXiv:1903.00812 [cs]. [Online]. Available: http://arxiv.org/abs/1903.00812

[38] F. Gomez-Donoso, M. Cazorla, A. Garcia-Garcia, and J. Garcia-Rodriguez, "Automatic Schaeffer's gestures recognition system," *Expert Systems*, vol. 33, no. 5, pp. 480–488, 2016, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.12160. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12160

[39] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, publisher: American Association for the Advancement of Science. [Online]. Available: https://www.science.org/doi/10.1126/science.aaa8415

[40] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, Jul. 1959, conference Name: IBM Journal of Research and Development.

[41] O. Firschein, "R63-70 Experiments on Machine Learning to Recognize," *IEEE Transactions on Electronic Computers*, vol. EC-12, no. 4, pp. 420–421, Aug. 1963, conference Name: IEEE Transactions on Electronic Computers.

[42] D. Michie, "Experiments on the Mechanization of Game-Learning Part I. Characterization of the Model and its parameters," *The Computer Journal*, vol. 6, no. 3, pp. 232–236, Nov. 1963. [Online]. Available: https://doi.org/10.1093/comjnl/6.3.232

[43] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, conference Name: IEEE Transactions on Information Theory.

[44] A. K. Griffith, "A comparison and evaluation of three machine learning procedures as applied to the game of checkers," *Artificial Intelligence*, vol. 5, no. 2, pp. 137–148, Jun. 1974. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0004370274900277

[45] J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, "1 - AN OVERVIEW OF MACHINE LEARNING," in *Machine Learning*, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Eds. San Francisco (CA): Morgan Kaufmann, Jan. 1983, pp. 3–23. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080510545500054

[46] J. D. Schaffer, "Some experiments in machine learning using vector evaluated genetic algorithms," Vanderbilt Univ.,Nashville, TN, Tech. Rep., Jan. 1985. [Online]. Available: https://www.osti.gov/biblio/5673304

[47] S. Athey, "The Impact of Machine Learning on Economics," in *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, Jan. 2018, pp. 507–547. [Online]. Available: https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/impact-machine-learning-economics

[48] G. Choy, O. Khalilzadeh, M. Michalski, S. Do, A. E. Samir, O. S. Pianykh, J. R. Geis, P. V. Pandharipande, J. A. Brink, and K. J. Dreyer, "Current Applications and Future Impact of Machine Learning in Radiology," *Radiology*, vol. 288, no. 2, pp. 318–328, Aug. 2018, publisher: Radiological Society of North America. [Online]. Available: https://pubs.rsna.org/doi/abs/10.1148/radiol.2018171820

[49] H. Pallathadka, M. Mustafa, D. T. Sanchez, G. Sekhar Sajja, S. Gour, and M. Naved, "IMPACT OF MACHINE learning ON Management, healthcare AND AGRICULTURE," *Materials Today: Proceedings*, vol. 80, pp. 2803–2806, Jan. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221478532104894X

[50] K. P. Murphy, *Machine learning: a probabilistic perspective*, ser. Adaptive computation and machine learning series. Cambridge, MA: MIT Press, 2012.

[51] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017, arXiv:1412.6980 [cs]. [Online]. Available: http://arxiv.org/abs/1412.6980

[52] M. Takeda and J. W. Goodman, "Neural networks for computation: number representations and programming complexity," *Applied Optics*, vol. 25, no. 18, pp. 3033–3046, Sep. 1986, publisher: Optica Publishing Group. [Online]. Available: https://opg.optica.org/ao/abstract.cfm?uri=ao-25-18-3033

[53] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943. [Online]. Available: https://doi.org/10.1007/BF02478259

[54] P. Sharma and A. Singh, "Era of deep neural networks: A review," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2017, pp. 1–5.

[55] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep Learning vs. Traditional Computer Vision," in *Advances in Computer Vision*, ser. Advances in Intelligent Systems and Computing, K. Arai and S. Kapoor, Eds. Cham: Springer International Publishing, 2020, pp. 128–144.

[56] S. Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 06, no. 02, pp. 107–116, Apr. 1998, publisher: World Scientific Publishing Co. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0218488598000094

[57] J. Su, J. Faraone, J. Liu, Y. Zhao, D. B. Thomas, P. H. W. Leong, and P. Y. K. Cheung, "Redundancy-Reduced MobileNet Acceleration on Reconfigurable Logic for ImageNet Classification," in *Applied Reconfigurable Computing. Architectures, Tools, and Applications*, ser. Lecture Notes in Computer Science, N. Voros, M. Huebner, G. Keramidas, D. Goehringer, C. Antonopoulos, and P. C. Diniz, Eds. Cham: Springer International Publishing, 2018, pp. 16–28.

[58] S. De, A. Mukherjee, and E. Ullah, "Convergence Guarantees for RMSProp and ADAM in Non-Convex Optimization and an Empirical Comparison to Nesterov Acceleration," *arXiv: Learning*, Jul. 2018. [Online]. Available: https://www.semanticscholar.org/paper/Convergence-Guarantees-for-RMSProp-and-ADAM-in-and-De-Mukherjee/f0a8159948d0b5d5035980c97b88038d444a1454

[59] S. Messelodi, C. M. Modena, M. Zanin, F. G. B. De Natale, F. Granelli, E. Betterle, and A. Guarise, "Intelligent extended floating car data collection," *Expert Systems with Applications*, vol. 36, no. 3, Part 1, pp. 4213–4227, Apr. 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417408002042

[60] V. Sevetlidis, G. Pavlidis, S. Mouroutsos, and A. Gasteratos, "Tackling Dataset Bias With an Automated Collection of Real-World Samples," *IEEE Access*, vol. 10, pp. 126 832–126 844, 2022, conference Name: IEEE Access.

[61] R. Pandey, H. Purohit, C. Castillo, and V. L. Shalin, "Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning," *International Journal of Human-Computer Studies*, vol. 160, p. 102772, Apr. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581922000015

[62] S. I. Nikolenko, "Introduction: The Data Problem," *Springer Optimization and Its Applications*, pp. 1–17, 2021, publisher: Springer. [Online]. Available: https://ideas.repec.org//h/spr/spochp/978-3-030-75178-4_1.html

[63] G. Paulin and M. Ivasic-Kos, "Review and analysis of synthetic dataset generation methods and techniques for application in computer vision," *Artificial Intelligence Review*, vol. 56, no. 9, pp. 9221–9265, Sep. 2023. [Online]. Available: https://doi.org/10.1007/s10462-022-10358-3

[64] N. Takemoto, L. Araújo, T. Coimbra, M. Tygel, S. Avila, and E. Borin, "Enriching synthetic data with real noise using Neural Style Transfer," Aug. 2019.

[65] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic Data - A Privacy Mirage," *ArXiv*, Nov. 2020. [Online]. Available: https://www.semanticscholar.org/paper/Synthetic-Data-A-Privacy-Mirage-Stadler-Oprisanu/c059d356f70560a955085a7d4625e49929657029

[66] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Empirically Analyzing the Effect of Dataset Biases on Deep Face Recognition Systems," Apr. 2018, arXiv:1712.01619 [cs]. [Online]. Available: http://arxiv.org/abs/1712.01619

[67] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Deep Domain-Adversarial Image Generation for Domain Generalisation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 13 025–13 032, Apr. 2020, number: 07. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/7003

[68] N. Gadipudi, I. Elamvazuthi, M. Sanmugam, L. I. Izhar, T. Prasetyo, R. Jegadeeshwaran, and S. S. A. Ali, "Synthetic to Real Gap Estimation of Autonomous Driving Datasets using Feature Embedding," in *2022 IEEE 5th International Symposium in Robotics and Manufacturing Automation (ROMA)*, Aug. 2022, pp. 1–5.

[69] E. Barsoum, "Articulated Hand Pose Estimation Review," *ArXiv*, Apr. 2016. [Online]. Available: https://www.semanticscholar.org/paper/Articulated-Hand-Pose-Estimation-Review-Barsoum/fe49526fef68e26217022fc56e043b278aee8446

[70] C. T. Wadsworth, "Clinical Anatomy and Mechanics of the Wrist and Hand," *Journal of Orthopaedic & Sports Physical Therapy*, vol. 4, no. 4, pp. 206–216, Apr. 1983, publisher: Journal of Orthopaedic & Sports Physical Therapy. [Online]. Available: https://www.jospt.org/doi/10.2519/jospt.1983.4.4.206

[71] V. Pavlovic, R. Sharma, and T. Huang, "Visual interpretation of hand gestures for human-computer interaction: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 677–695, Jul. 1997, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[72] S. Han, B. Liu, R. Cabezas, C. D. Twigg, P. Zhang, J. Petkau, T.-H. Yu, C.-J. Tai, M. Akbay, Z. Wang, A. Nitzan, G. Dong, Y. Ye, L. Tao, C. Wan, and R. Wang, "MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality," *ACM Transactions on Graphics*, vol. 39, no. 4, pp. 87:87:1–87:87:13, Aug. 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3386569.3392452

[73] M. Melyani, Y. Heryadi, A. Trisetyarso, B. S. Abbas, W. Suparta, and F. L. Gaol, "Framework of Mobile Game Design as an Assistive Technology for Children with Motor Disabilities," *Journal of Games, Game Art, and Gamification*, vol. 4, no. 1, 2019, number: 1. [Online]. Available: https://journal.binus.ac.id/index.php/jggag/article/view/7460

[74] M. De MEULDER, "The Legal Recognition of Sign Languages," *Sign Language Studies*, vol. 15, no. 4, pp. 498–506, 2015, publisher: Gallaudet University Press. [Online]. Available: https://www.jstor.org/stable/26191000

[75] I. Papastratis, C. Chatzikonstantinou, D. Konstantinidis, K. Dimitropoulos, and P. Daras, "Artificial Intelligence Technologies for Sign Language," *Sensors (Basel, Switzerland)*, vol. 21, no. 17, p. 5843, Aug. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8434597/

[76] W. Sandler, "The uniformity and diversity of language: Evidence from sign language," *Lingua*, vol. 120, no. 12, pp. 2727–2732, Dec. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0024384110000768

[77] B. Schwarz, "Mapping the world in 3D," *Nature Photonics*, vol. 4, no. 7, pp. 429–430, Jul. 2010, number: 7 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nphoton.2010.148

[78] L. Cruz, D. Lucio, and L. Velho, "Kinect and RGBD Images: Challenges and Applications," in *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials*, Aug. 2012, pp. 36–49.

[79] A. Thippur, C. H. Ek, and H. Kjellström, "Inferring hand pose: A comparative study of visual shape features," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Apr. 2013, pp. 1–8.

[80] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks," *ACM Transactions on Graphics*, vol. 33, no. 5, pp. 169:1–169:10, Sep. 2014. [Online]. Available: https://dl.acm.org/doi/10.1145/2629500

[81] "NYU Hand Pose Dataset." [Online]. Available: https://jonathantompson.github.io/NYU_Hand_Pose_Dataset.htm

[82] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and Robust Hand Tracking from Depth," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, Jun. 2014, pp. 1106–1113. [Online]. Available: https://ieeexplore.ieee.org/document/6909541

[83] "Homepage of Xiao Sun." [Online]. Available: https://jimmysuen.github.io/

[84] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim, "BigHand2.2M Benchmark: Hand Pose Dataset and State of the Art Analysis," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2605–2613, iSSN: 1063-6919.

[85] "FreiHAND Dataset." [Online]. Available: https://lmb.informatik.uni-freiburg.de/projects/freihand/

[86] F. Gomez-Donoso, S. Orts-Escolano, and M. Cazorla, "Large-scale Multiview 3D Hand Pose Dataset," Jul. 2017, arXiv:1707.03742 [cs]. [Online]. Available: http://arxiv.org/abs/1707.03742

[87] "Multiview Hand Pose Dataset." [Online]. Available: http://www.rovit.ua.es/dataset/mhpdataset/

[88] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, Jun. 2018, pp. 49–59. [Online]. Available: https://ieeexplore.ieee.org/document/8578111/

[89] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010. [Online]. Available: https://doi.org/10.1007/s11263-009-0275-4

[90] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255, iSSN: 1063-6919.

[91] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *Computer Vision – ECCV 2012*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 7576, pp. 746–760, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-642-33715-4_54

[92] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, Jun. 2015, pp. 567–576. [Online]. Available: http://ieeexplore.ieee.org/document/7298655/

[93] S. I. Nikolenko, "Synthetic Data for Basic Computer Vision Problems," in *Synthetic Data for Deep Learning*, ser. Springer Optimization and Its Applications, S. I. Nikolenko, Ed. Cham: Springer International Publishing, 2021, pp. 161–194. [Online]. Available: https://doi.org/10.1007/978-3-030-75178-4_6

[94] K. Man and J. Chahl, "A Review of Synthetic Image Data and Its Use in Computer Vision," *Journal of Imaging*, vol. 8, no. 11, p. 310, Nov. 2022, number: 11 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2313-433X/8/11/310

[95] J. Little and A. Verri, "Analysis of differential and matching methods for optical flow," in *Workshop on Visual Motion [1989] Proceedings*, Mar. 1989, pp. 173–180.

[96] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, Feb. 1994. [Online]. Available: https://doi.org/10.1007/BF01420984

[97] C. Stewart, "MINPRAN: a new robust estimator for computer vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 925–938, Oct. 1995, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[98] Y. Leedan and P. Meer, "Estimation with bilinear constraints in computer vision," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, Jan. 1998, pp. 733–738.

[99] Y. Genc, J. Ponce, Y. Leedan, and P. Meer, "Parameterized image varieties and estimation with bilinear constraints," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2, Jun. 1999, pp. 67–72 Vol. 2, iSSN: 1063-6919.

[100] Y. Leedan and P. Meer, "Heteroscedastic Regression in Computer Vision: Problems with Bilinear Constraint," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 127–150, Jun. 2000. [Online]. Available: https://doi.org/10.1023/A:1008185619375

[101] W. Freeman and E. Pasztor, "Learning low-level vision," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, Sep. 1999, pp. 1182–1189 vol.2.

[102] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981. [Online]. Available: https://dl.acm.org/doi/10.1145/358669.358692

[103] H. Wang and D. Suter, "Robust adaptive-scale parametric model estimation for computer vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1459–1474, Nov. 2004, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[104] J. Matas and O. Chum, "Randomized RANSAC with Td,d test," *Image and Vision Computing*, vol. 22, no. 10, pp. 837–842, Sep. 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885604000514

[105] J.-G. Wang, E. Sung, and R. Venkateswarlu, "Estimating the eye gaze from one eye," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 83–103, Apr. 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314204001134

[106] T. Vaudrey, C. Rabe, R. Klette, and J. Milburn, "Differences between stereo and motion behaviour on synthetic and real-world stereo sequences," in *2008 23rd International Conference Image and Vision Computing New Zealand*, Nov. 2008, pp. 1–6, iSSN: 2151-2205.

[107] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A Database and Evaluation Methodology for Optical Flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, Mar. 2011. [Online]. Available: https://doi.org/10.1007/s11263-010-0390-2

[108] M. Peris, S. Martull, A. Maki, Y. Ohkawa, and K. Fukui, "Towards a simulation driven stereo vision system," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov. 2012, pp. 1038–1042, iSSN: 1051-4651.

[109] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A Naturalistic Open Source Movie for Optical Flow Evaluation," in *Computer Vision – ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer, 2012, pp. 611–625.

[110] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 4040–4048, arXiv:1512.02134 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1512.02134

[111] N. Mayer, E. Ilg, P. Fischer, C. Hazirbas, D. Cremers, A. Dosovitskiy, and T. Brox, "What Makes Good Synthetic Training Data for Learning Disparity and Optical Flow Estimation?" *International Journal of Computer Vision*, vol. 126, no. 9, pp. 942–960, Sep. 2018. [Online]. Available: https://doi.org/10.1007/s11263-018-1082-6

[112] X. Peng, B. Sun, K. Ali, and K. Saenko, "Learning Deep Object Detectors from 3D Models," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1278–1286, iSSN: 2380-7504.

[113] E. Bochinski, V. Eiselein, and T. Sikora, "Training a convolutional neural network for multi-class object detection using solely virtual world data," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug. 2016, pp. 278–285.

[114] S. Hinterstoisser, V. Lepetit, P. Wohlhart, and K. Konolige, "On Pre-Trained Image Features and Synthetic Images for Deep Learning," Nov. 2017, arXiv:1710.10710 [cs]. [Online]. Available: http://arxiv.org/abs/1710.10710

[115] F. E. Nowruzi, P. Kapoor, D. Kolhatkar, F. A. Hassanat, R. Laganiere, and J. Rebut, "How much real data do we actually need: Analyzing object detection performance using synthetic and real data," Jul. 2019, arXiv:1907.07061 [cs]. [Online]. Available: http://arxiv.org/abs/1907.07061

[116] S. Hinterstoisser, O. Pauly, H. Heibel, M. Marek, and M. Bokeloh, "An Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Instance Detection," Feb. 2019, arXiv:1902.09967 [cs]. [Online]. Available: http://arxiv.org/abs/1902.09967

[117] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An Information-Rich 3D Model Repository," Dec. 2015, arXiv:1512.03012 [cs]. [Online]. Available: http://arxiv.org/abs/1512.03012

[118] L. Yi, V. G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, and L. Guibas, "A scalable active framework for region annotation in 3D shape collections," *ACM Transactions on Graphics*, vol. 35, no. 6, pp. 210:1–210:12, Dec. 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2980179.2980238

[119] L. Yi, L. Guibas, A. Hertzmann, V. G. Kim, H. Su, and E. Yumer, "Learning hierarchical shape segmentation and labeling from online repositories," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 70:1–70:12, Jul. 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3072959.3073652

[120] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A Large-scale Benchmark for Fine-grained and Hierarchical Part-level 3D Object Understanding," Dec. 2018, arXiv:1812.02713 [cs]. [Online]. Available: http://arxiv.org/abs/1812.02713

[121] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, "SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-training on Indoor Segmentation?" in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2697–2706, iSSN: 2380-7504.

[122] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, "SceneNet: Understanding Real World Indoor Scenes With Synthetic Data," Nov. 2015, arXiv:1511.07041 [cs]. [Online]. Available: http://arxiv.org/abs/1511.07041

[123] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, and J. M. Alvarez, "Effective Use of Synthetic Data for Urban Scene Semantic Segmentation," Jul. 2018, arXiv:1807.06132 [cs]. [Online]. Available: http://arxiv.org/abs/1807.06132

[124] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," Jan. 2018, arXiv:1703.06870 [cs]. [Online]. Available: http://arxiv.org/abs/1703.06870

[125] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," Jun. 2016, arXiv:1412.7062 [cs] version: 4. [Online]. Available: http://arxiv.org/abs/1412.7062

[126] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic, "Seeing 3D Chairs: Exemplar Part-Based 2D-3D Alignment Using a Large Dataset of CAD Models," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 3762–3769, iSSN: 1063-6919.

[127] F. Liu, S. Wang, D. Ding, Q. Yuan, Z. Yao, Z. Pan, and H. Li, "Retrieving indoor objects: 2D-3D alignment using single image and interactive ROI-based refinement," *Computers & Graphics*, vol. 70, pp. 108–117, Feb. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S009784931730122X

[128] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik, "Aligning 3D models to RGB-D images of cluttered scenes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 4731–4740, iSSN: 1063-6919.

[129] T. Hodan, P. Haluza, S. Obdrzalek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects," Jan. 2017, arXiv:1701.05498 [cs]. [Online]. Available: http://arxiv.org/abs/1701.05498

[130] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects," Sep. 2018, arXiv:1809.10790 [cs]. [Online]. Available: http://arxiv.org/abs/1809.10790

[131] E. Richardson, M. Sela, and R. Kimmel, "3D Face Reconstruction by Learning from Synthetic Data," Sep. 2016, arXiv:1609.04387 [cs]. [Online]. Available: http://arxiv.org/abs/1609.04387

[132] H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, and C. Rother, "Augmented Reality Meets Computer Vision : Efficient Data Generation for Urban Driving Scenes," Aug. 2017, arXiv:1708.01566 [cs]. [Online]. Available: http://arxiv.org/abs/1708.01566

[133] 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*, vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[134] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," Feb. 2019, arXiv:1809.11096 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1809.11096

[135] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-Shot Text-to-Image Generation," Feb. 2021, arXiv:2102.12092 [cs]. [Online]. Available: http://arxiv.org/abs/2102.12092

[136] M. John and S. Santhanalakshmi, "Image augmentation using GAN models in Computer Vision," in *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, Oct. 2021, pp. 1194–1201.

[137] Z. Wang, Q. She, and T. E. Ward, "Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy," *ACM Computing Surveys*, vol. 54, no. 2, pp. 37:1–37:38, Feb. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3439723

[138] M. Kang, J. Shin, and J. Park, "StudioGAN: A Taxonomy and Benchmark of GANs for Image Synthesis," Jul. 2022, arXiv:2206.09479 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2206.09479

[139] A. Borji, "Pros and Cons of GAN Evaluation Measures: New Developments," Oct. 2021, arXiv:2103.09396 [cs]. [Online]. Available: http://arxiv.org/abs/2103.09396

[140] A. Hindupur, "The GAN Zoo," Aug. 2023, original-date: 2017-04-14T16:45:24Z. [Online]. Available: https://github.com/hindupuravinash/the-gan-zoo

[141] V. Allken, N. O. Handegard, S. Rosen, T. Schreyeck, T. Mahiout, and K. Malde, "Fish species identification using a convolutional neural network trained on synthetic data," *ICES Journal of Marine Science*, vol. 76, no. 1, pp. 342–349, Jan. 2019. [Online]. Available: https://doi.org/10.1093/icesjms/fsy147

[142] T. Dahmen, P. Trampert, F. Boughorbel, J. Sprenger, M. Klusch, K. Fischer, C. Kübel, and P. Slusallek, "Digital reality: a model-based approach to supervised learning from synthetic data," *AI Perspectives*, vol. 1, no. 1, p. 2, Sep. 2019. [Online]. Available: https://doi.org/10.1186/s42467-019-0002-0

[143] X. Qi, Q. Chen, J. Jia, and V. Koltun, "Semi-parametric Image Synthesis," Apr. 2018, arXiv:1804.10992 [cs]. [Online]. Available: http://arxiv.org/abs/1804.10992

[144] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '99. USA: ACM Press/Addison-Wesley Publishing Co., Jul. 1999, pp. 187–194. [Online]. Available: https://dl.acm.org/doi/10.1145/311535.311556

[145] ——, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[146] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3D Morphable Model Learnt From 10,000 Faces," 2016, pp. 5543–5552. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Booth_A_3D_Morphable_CVPR_2016_paper.html

[147] B. Egger, W. A. P. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter, "3D Morphable Face Models – Past, Present and Future," Apr. 2020, arXiv:1909.01815 [cs]. [Online]. Available: http://arxiv.org/abs/1909.01815

[148] Z. An, W. Deng, T. Yuan, and J. Hu, "Deep Transfer Network with 3D Morphable Models for Face Recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, May 2018, pp. 416–422.

[149] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data," 2019, pp. 0–0. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/html/BEFA/Kortylewski_Analyzing_and_Reducing_the_Damage_of_Dataset_Bias_to_Face_CVPRW_2019_paper.html

[150] S. I. Nikolenko, "Synthetic Simulated Environments," in *Synthetic Data for Deep Learning*, ser. Springer Optimization and Its Applications, S. I. Nikolenko, Ed. Cham: Springer International Publishing, 2021, pp. 195–215. [Online]. Available: https://doi.org/10.1007/978-3-030-75178-4_7

[151] M. Müller, V. Casser, J. Lahoud, N. Smith, and B. Ghanem, "Sim4CV: A Photo-Realistic Simulator for Computer Vision Applications," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 902–919, Sep. 2018. [Online]. Available: https://doi.org/10.1007/s11263-018-1073-7

[152] N. Jaipuria, X. Zhang, R. Bhasin, M. Arafa, P. Chakravarty, S. Shrivastava, S. Manglani, and V. N. Murali, "Deflating Dataset Bias Using Synthetic Data Augmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2020, pp. 3344–3353, iSSN: 2160-7516.

[153] G. Riegler, M. Urschler, M. Rüther, H. Bischof, and D. Stern, "Anatomical Landmark Detection in Medical Applications Driven by Synthetic Data," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec. 2015, pp. 85–89.

[154] W. Qiu and A. Yuille, "UnrealCV: Connecting Computer Vision to Unreal Engine," Sep. 2016, arXiv:1609.01326 [cs]. [Online]. Available: http://arxiv.org/abs/1609.01326

[155] W. Qiu, F. Zhong, Y. Zhang, S. Qiao, Z. Xiao, T. S. Kim, and Y. Wang, "UnrealCV: Virtual Worlds for Computer Vision," in *Proceedings of the 25th ACM international conference on Multimedia*. Mountain View California USA: ACM, Oct. 2017, pp. 1221–1224. [Online]. Available: https://dl.acm.org/doi/10.1145/3123266.3129396

[156] S. Borkman, A. Crespi, S. Dhakad, S. Ganguly, J. Hogins, Y.-C. Jhang, M. Kamalzadeh, B. Li, S. Leal, P. Parisi, C. Romero, W. Smith, A. Thaman, S. Warren, and N. Yadav, "Unity Perception: Generate Synthetic Data for Computer Vision," Jul. 2021, arXiv:2107.04259 [cs]. [Online]. Available: http://arxiv.org/abs/2107.04259

[157] H. Tang and K. Jia, "A New Benchmark: On the Utility of Synthetic Data With Blender for Bare Supervised Learning and Downstream Domain Adaptation," 2023, pp. 15 954–15 964. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Tang_A_New_Benchmark_On_the_Utility_of_Synthetic_Data_With_CVPR_2023_paper.html

[158] E. Hatay, J. Ma, H. Sun, J. Fang, Z. Gao, and H. Yu, "Learning to Detect Phone-related Pedestrian Distracted Behaviors with Synthetic Data," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 2975–2983. [Online]. Available: https://ieeexplore.ieee.org/document/9522912/

[159] J. Shermeyer, T. Hossler, A. Van Etten, D. Hogan, R. Lewis, and D. Kim, "RarePlanes: Synthetic Data Takes Flight," Nov. 2020, arXiv:2006.02963 [cs]. [Online]. Available: http://arxiv.org/abs/2006.02963

[160] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, "PAMTRI: Pose-Aware Multi-Task Learning for Vehicle Re-Identification Using Highly Randomized Synthetic Data," May 2020, arXiv:2005.00673 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2005.00673

[161] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for Data: Ground Truth from Computer Games," Aug. 2016, arXiv:1608.02192 [cs]. [Online]. Available: http://arxiv.org/abs/1608.02192

[162] S. R. Richter, Z. Hayder, and V. Koltun, "Playing for Benchmarks," Sep. 2017, arXiv:1709.07322 [cs]. [Online]. Available: http://arxiv.org/abs/1709.07322

[163] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun, "Real time hand pose estimation using depth sensors," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov. 2011, pp. 1228–1234.

[164] ——, "Hand Pose Estimation and Hand Shape Classification Using Multi-layered Randomized Decision Forests," in *Computer Vision – ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer, 2012, pp. 852–863.

[165] D. Tang, T.-H. Yu, and T.-K. Kim, "Real-Time Articulated Hand Pose Estimation Using Semi-supervised Transductive Regression Forests," 2013, pp. 3224–3231. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2013/html/Tang_Real-Time_Articulated_Hand_2013_ICCV_paper.html

[166] J. Molina, J. A. Pajuelo, M. Escudero-Viñolo, J. Bescós, and J. M. Martínez, "A natural and synthetic corpus for benchmarking of hand gesture recognition systems," *Machine Vision and Applications*, vol. 25, no. 4, pp. 943–954, May 2014. [Online]. Available: https://doi.org/10.1007/s00138-013-0576-z

[167] X. Deng, S. Yang, Y. Zhang, P. Tan, L. Chang, and H. Wang, "Hand3D: Hand Pose Estimation using 3D Neural Network," Apr. 2017, arXiv:1704.02224 [cs]. [Online]. Available: http://arxiv.org/abs/1704.02224

[168] M. Madadi, S. Escalera, A. Carruesco, C. Andujar, X. Baró, and J. Gonzàlez, "Top-down model fitting for hand pose recovery in sequences of depth images," *Image and Vision Computing*, vol. 79, pp. 63–75, Nov. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885618301513

[169] C. Zimmermann and T. Brox, "Learning to Estimate 3D Hand Pose from Single RGB Images," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017, pp. 4913–4921. [Online]. Available: http://ieeexplore.ieee.org/document/8237787/

[170] 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)*, Sep. 2018, pp. 110–119, iSSN: 2475-7888.

[171] Y. Cai, L. Ge, J. Cai, and J. Yuan, "Weakly-supervised 3D Hand Pose Estimation from Monocular RGB Images," 2018, pp. 666–682. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/html/Yujun_Cai_Weakly-supervised_3D_Hand_ECCV_2018_paper.html

[172] L. Yang, S. Chen, and A. Yao, "SemiHand: Semi-Supervised Hand Pose Estimation With Consistency," 2021, pp. 11 364–11 373. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Yang_SemiHand_Semi-Supervised_Hand_Pose_Estimation_With_Consistency_ICCV_2021_paper.html

[173] H. Park, D. Kim, S. Yim, T. Kwon, J. Jeong, W. Lee, J. Lee, B. Yoo, and G. Lee, "Generating Hand Posture and Motion Dataset for Hand Pose Estimation in Egocentric View," in *Virtual, Augmented and Mixed Reality: Design and Development*, ser. Lecture Notes in Computer Science, J. Y. C. Chen and G. Fragomeni, Eds. Cham: Springer International Publishing, 2022, pp. 329–337.

[174] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, "ViHASi: Virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods," in *2008 Second ACM/IEEE International Conference on Distributed Smart Cameras*, Sep. 2008, pp. 1–10.

[175] M. Khodabandeh, H. R. V. Joze, I. Zharkov, and V. Pradeep, "DIY Human Action Dataset Generation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, pp. 1529–152 910, iSSN: 2160-7516.

[176] S. E. Ebadi, Y.-C. Jhang, A. Zook, S. Dhakad, A. Crespi, P. Parisi, S. Borkman, J. Hogins, and S. Ganguly, "PeopleSansPeople: A Synthetic Data Generator for Human-Centric Computer Vision," Jul. 2022, arXiv:2112.09290 [cs]. [Online]. Available: http://arxiv.org/abs/2112.09290

[177] S. Meister and D. Kondermann, "Real versus realistically rendered scenes for optical flow evaluation," in *2011 14th ITG Conference on Electronic Media Technology*, Mar. 2011, pp. 1–6.

[178] Y. Movshovitz-Attias, T. Kanade, and Y. Sheikh, "How useful is photo-realistic rendering for visual learning?" Sep. 2016, arXiv:1603.08152 [cs]. [Online]. Available: http://arxiv.org/abs/1603.08152

[179] A. Tsirikoglou, J. Kronander, M. Wrenninge, and J. Unger, "Procedural Modeling and Physically Based Rendering for Synthetic Data Generation in Automotive Applications," Oct. 2017, arXiv:1710.06270 [cs]. [Online]. Available: http://arxiv.org/abs/1710.06270

[180] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2017, pp. 23–30, iSSN: 2153-0866.

[181] G. Georgakis, A. Mousavian, A. C. Berg, and J. Kosecka, "Synthesizing Training Data for Object Detection in Indoor Scenes," Sep. 2017, arXiv:1702.07836 [cs]. [Online]. Available: http://arxiv.org/abs/1702.07836

[182] J. Rehg and T. Kanade, "DigitEyes: vision-based hand tracking for human-computer interaction," in *Proceedings of 1994 IEEE Workshop on Motion of Non-rigid and Articulated Objects*, Nov. 1994, pp. 16–22.

[183] S. Lu, D. Metaxas, D. Samaras, and J. Oliensis, "Using multiple cues for hand tracking and model refinement," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 2, Jun. 2003, pp. II–443, iSSN: 1063-6919.

[184] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints," in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 2088–2095, iSSN: 2380-7504.

[185] 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*, ser. UIST '11. New York, NY, USA: Association for Computing Machinery, Oct. 2011, pp. 549–558. [Online]. Available: https://dl.acm.org/doi/10.1145/2047196.2047269

[186] I. Oikonomidis, N. Kyriazis, and A. Argyros, "Efficient model-based 3D tracking of hand articulations using Kinect," in *Procedings of the British Machine Vision Conference 2011*. Dundee: British Machine Vision Association, 2011, pp. 101.1–101.11. [Online]. Available: http://www.bmva.org/bmvc/2011/proceedings/paper101/index.html

[187] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall, "Capturing Hands in Action Using Discriminative Salient Points and Physics Simulation," *International Journal of Computer Vision*, vol. 118, no. 2, pp. 172–193, Jun. 2016. [Online]. Available: https://doi.org/10.1007/s11263-016-0895-4

[188] H. Liang, J. Yuan, D. Thalmann, and Z. Zhang, "Model-based hand pose estimation via spatial-temporal hand parsing and 3D fingertip localization," *The Visual Computer*, vol. 29, no. 6, pp. 837–848, Jun. 2013. [Online]. Available: https://doi.org/10.1007/s00371-013-0822-4

[189] C. Xu and L. Cheng, "Efficient Hand Pose Estimation from a Single Depth Image," in *2013 IEEE International Conference on Computer Vision*, Dec. 2013, pp. 3456–3462, iSSN: 2380-7504.

[190] D. Tang, J. Taylor, P. Kohli, C. Keskin, T.-K. Kim, and J. Shotton, "Opening the Black Box: Hierarchical Sampling Optimization for Estimating Human Hand Pose," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 3325–3333, iSSN: 2380-7504.

[191] J. Taylor, L. Bordeaux, T. Cashman, B. Corish, C. Keskin, T. Sharp, E. Soto, D. Sweeney, J. Valentin, B. Luff, A. Topalian, E. Wood, S. Khamis, P. Kohli, S. Izadi, R. Banks, A. Fitzgibbon, and J. Shotton, "Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 143:1–143:12, Jul. 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2897824.2925965

[192] U. Iqbal, P. Molchanov, T. Breuel, J. Gall, and J. Kautz, "Hand Pose Estimation via Latent 2.5D Heatmap Regression," in *Computer Vision – ECCV 2018*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 125–143.

[193] 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)*, Mar. 2018, pp. 436–445.

[194] B. Tekin, F. Bogo, and M. Pollefeys, "H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 4506–4515, iSSN: 2575-7075.

[195] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal Deep Variational Hand Pose Estimation," Mar. 2018, arXiv:1803.11404 [cs]. [Online]. Available: http://arxiv.org/abs/1803.11404

[196] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning Joint Reconstruction of Hands and Manipulated Objects," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 11 799–11 808, iSSN: 2575-7075.

[197] S. Baek, K. I. Kim, and T.-K. Kim, "Pushing the Envelope for RGB-Based Dense 3D Hand Pose Estimation via Neural Rendering," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 1067–1076, iSSN: 2575-7075.

[198] D. Xiang, H. Joo, and Y. Sheikh, "Monocular Total Capture: Posing Face, Body, and Hands in the Wild," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 10 957–10 966, iSSN: 2575-7075.

[199] Z. Zheng, Z. Hu, H. Qin, and J. Liu, "Stacked graph bone region U-net with bone representation for hand pose estimation and semi-supervised training," *Image and Vision Computing*, vol. 134, p. 104673, Jun. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885623000471

[200] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015, arXiv:1505.04597 [cs]. [Online]. Available: http://arxiv.org/abs/1505.04597

[201] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3D Hand Pose Estimation in Single Depth Images: From Single-View CNN to Multi-View CNNs," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 3593–3601, iSSN: 1063-6919.

[202] ——, "3D Convolutional Neural Networks for Efficient and Robust Hand Pose Estimation from Single Depth Images," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 5679–5688, iSSN: 1063-6919.

[203] L. Ge, Y. Cai, J. Weng, and J. Yuan, "Hand PointNet: 3D Hand Pose Estimation Using Point Sets," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 8417–8426, iSSN: 2575-7075.

[204] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Real-Time 3D Hand Pose Estimation with 3D Convolutional Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 956–970, Apr. 2019, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[205] M. Rezaei, R. Rastgoo, and V. Athitsos, "TriHorn-Net: A Model for Accurate Depth-Based 3D Hand Pose Estimation," Jun. 2022, arXiv:2206.07117 [cs] version: 2. [Online]. Available: http://arxiv.org/abs/2206.07117

[206] J. Cheng, Y. Wan, D. Zuo, C. Ma, J. Gu, P. Tan, H. Wang, X. Deng, and Y. Zhang, "Efficient Virtual View Selection for 3D Hand Pose Estimation," Mar. 2022, arXiv:2203.15458 [cs] version: 1. [Online]. Available: http://arxiv.org/abs/2203.15458

[207] C. Wan, T. Probst, L. Van Gool, and A. Yao, "Crossing Nets: Combining GANs and VAEs with a Shared Latent Space for Hand Pose Estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1196–1205, iSSN: 1063-6919.

[208] F. Mueller, M. Davis, F. Bernard, O. Sotnychenko, M. Verschoor, M. A. Otaduy, D. Casas, and C. Theobalt, "Real-time pose and shape reconstruction of two interacting hands with a single depth camera," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 49:1–49:13, Jul. 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3306346.3322958

[209] Z. Bauer, F. Gomez-Donoso, E. Cruz, S. Orts-Escolano, and M. Cazorla, "UASOL, a large-scale high-resolution outdoor stereo dataset," *Scientific Data*, vol. 6, no. 1, p. 162, Aug. 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41597-019-0168-5

[210] "MacBook Pro (14-inch, 2021) - Technical Specifications." [Online]. Available: https://support.apple.com/kb/SP854?viewlocale=en_US&locale=en_IE

[211] "About the security content of Rapid Security Responses for macOS Ventura 13.4.1," Jul. 2023. [Online]. Available: https://support.apple.com/en-us/HT213825

[212] "Tesla V100 GPUs are now generally available." [Online]. Available: https://cloud.google.com/blog/products/compute/tesla-v100-gpus-are-now-generally-available

[213] "Xcode 14.3.1 Release Notes." [Online]. Available: https://developer.apple.com/documentation/xcode-release-notes/xcode-14_3_1-release-notes

[214] A. Inc, "C++ Language Support - Xcode." [Online]. Available: https://developer.apple.com/xcode/cpp/

[215] "C++20 - cppreference.com." [Online]. Available: https://en.cppreference.com/w/cpp/20

[216] "OpenGL Capabilities Tables," 2017.

[217] M. Lujan, M. McCrary, B. W. Ford, and Z. Zong, "Vulkan vs OpenGL ES: Performance and Energy Efficiency Comparison on the big.LITTLE Architecture," in *2021 IEEE International Conference on Networking, Architecture and Storage (NAS)*, Oct. 2021, pp. 1–8.

[218] N. Stewart, "GLEW - The OpenGL Extension Wrangler Library," Aug. 2023, original-date: 2013-03-19T03:29:13Z. [Online]. Available: https://github.com/nigels-com/glew

[219] "Releases · glfw/glfw." [Online]. Available: https://github.com/glfw/glfw/releases

[220] "Learn OpenGL, extensive tutorial resource for learning Modern OpenGL." [Online]. Available: https://learnopengl.com/

[221] J. F. Blinn, "Models of light reflection for computer synthesized pictures," in *Proceedings of the 4th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '77. New York, NY, USA: Association for Computing Machinery, Jul. 1977, pp. 192–198. [Online]. Available: https://dl.acm.org/doi/10.1145/563858.563893

[222] slyusarrr, "Rigged Hand model Photorealistic with Nails | 3D model." [Online]. Available: https://www.cgtrader.com/3d-models/character/man/rigged-hand-model-photorealistic-with-nails

[223] "Releases · assimp/assimp." [Online]. Available: https://github.com/assimp/assimp/releases

[224] "Python Release Python 3.9.6." [Online]. Available: https://www.python.org/downloads/release/python-396/

[225] "Releases · pytorch/pytorch." [Online]. Available: https://github.com/pytorch/pytorch/releases

[226] O. Chernytska, "2D Hand Pose Estimation from RGB Image," Aug. 2023, original-date: 2021-04-08T16:52:06Z. [Online]. Available: https://github.com/OlgaChernytska/2D-Hand-Pose-Estimation-RGB

[227] ——, "3D Hand Pose Estimation from Single RGB Camera," Master's thesis, Ukranian Catholic University, Lviv, 2019, accepted: 2019-02-18T13:43:12Z. [Online]. Available: https://er.ucu.edu.ua/handle/1/1327

Appendix A

Use of AI tools

The AI-based tools Grammarly and ChatGPT have been employed, at varying degrees, to assist the production of this document, as specified below these lines.

Grammarly has been employed as the primary grammar check tool throughout the document to avoid grammar and orthography errors.

ChatGPT has been employed in specific document fragments to clarify the confusing text. It has been used in fragments containing intricate information and complex ideas requiring clear expression. This tool has been used to explore alternative, easier-to-read constructions in cases when the writer was not satisfied with the clarity or style of a fragment.