



Review

A Survey on GAN-Based Data Augmentation for Hand Pose Estimation Problem

Farnaz Farahanipad ^{*,†} , Mohammad Rezaei [†], Mohammad Sadegh Nasr , Farhad Kamangar and Vassilis Athitsos

Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX 76019, USA; mohammad.rezaei@mavs.uta.edu (M.R.); mohammadsadegh.nasr@mavs.uta.edu (M.S.N.); kamangar@uta.edu (F.K.); athitsos@uta.edu (V.A.)

* Correspondence: farnaz.farahanipad@mavs.uta.edu

† These authors contributed equally to this work.

Abstract: Deep learning solutions for hand pose estimation are now very reliant on comprehensive datasets covering diverse camera perspectives, lighting conditions, shapes, and pose variations. While acquiring such datasets is a challenging task, several studies circumvent this problem by exploiting synthetic data, but this does not guarantee that they will work well in real situations mainly due to the gap between the distribution of synthetic and real data. One recent popular solution to the domain shift problem is learning the mapping function between different domains through generative adversarial networks. In this study, we present a comprehensive study on effective hand pose estimation approaches, which are comprised of the leveraged generative adversarial network (GAN), providing a comprehensive training dataset with different modalities. Benefiting from GAN, these algorithms can augment data to a variety of hand shapes and poses where data manipulation is intuitively controlled and greatly realistic. Next, we present related hand pose datasets and performance comparison of some of these methods for the hand pose estimation problem. The quantitative and qualitative results indicate that the state-of-the-art hand pose estimators can be greatly improved with the aid of the training data generated by these GAN-based data augmentation methods. These methods are able to beat the baseline approaches with better visual quality and higher values in most of the metrics (PCK and ME) on both the STB and NYU datasets. Finally, in conclusion, the limitation of the current methods and future directions are discussed.

Keywords: generative adversarial networks; hand pose estimation; data augmentation; domain translation; semi-supervised learning; weakly supervised learning



Citation: Farahanipad, F.; Rezaei, M.; Nasr, M.S.; Kamangar, F.; Athitsos, V. A Survey on GAN-Based Data Augmentation for Hand Pose Estimation Problem. *Technologies* **2022**, *10*, 43. <https://doi.org/10.3390/technologies10020043>

Academic Editors: Abdellah Chehri and Pietro Zanuttigh

Received: 31 December 2021

Accepted: 4 February 2022

Published: 21 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Hand pose estimation, which is a problem of predicting the 2D/3D position of hand joints, given an RGB/depth input, is receiving a lot of attention in the computer vision field. It has been applied in many applications, such as human–computer interaction (HCI) [1], gesture recognition [2–4], sign language recognition [5–8], interactive games [9–11], user interface controls [12], computer-aided design (CAD) [13], etc. In recent years, by the advancements in deep learning algorithms, data-driven approaches have become more advantageous and have led to significant improvements in 2D/3D hand pose estimation, as a large number of annotated datasets have become available [14–16]. However, acquiring accurate 3D labeled data requires an expensive marker-based motion capture system or a massive multi-view camera setting. Therefore, to avoid annotating such large datasets, which is costly, time consuming and labor intensive, researchers are trying to find alternative approaches that can leverage them. One upcoming solution is to use synthetic data for training, where data are automatically annotated and convenient for generating a large scale of data with accurate ground truth. Although image synthesis can be generated using a physical renderer, there are usually a few differences between real and synthetic data,

without consideration of depth sensor noise in a realistic way. Therefore, models trained on the synthetic data suffers from the domain shift problem, and they fail to perform well on real datasets, due to the domain gap between the real and synthetic datasets.

The most promising approach is to use generative models that learn to discover the essence of data and find a best distribution to represent it. Generative adversarial networks [17], or GANs in short, are a class of generative models, where two neural networks, generator and discriminator, contest with each other in a zero-sum game, where one agent's gain is another agent's loss. Given a training set, the generator learns to generate new data with the same statistics as the training set, while the discriminator's goal is to distinguish between real and generated samples. GANs have the ability to translate source synthetic images into realistic target-like images for training purposes. This is known as domain transfer learning. Several state-of-the-art transfer learning research works used GANs to enforce the alignment of the latent feature space. The conditional generative adversarial networks (CGANs) [18], which is an extension of GAN, has the ability to train synthetic models to generate images based on auxiliary information. Due to the popularity of the framework, it has become the foundation for many successful architectures, such as CycleGAN [19], StyleGAN [20], PixelRNN [21], DiscoGAN [22], etc.

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

Despite the large body of works that have been conducted on hand pose estimation using generative adversarial networks, no recent all-round survey has been conducted on it. As far as we know, this is the first survey among current publications which focused on GAN-based data augmentation for hand pose estimation problem. Moreover, different from existing review papers on the hand pose estimation problem which mainly discuss depth-based methods [23,24], in this paper, we present a comprehensive study on the most recent GAN-based methods based on input data modality, i.e., RGB, depth, or multi-modal information. Another point of motivation of our work is that researchers do attach much importance to semi/unsupervised learning using GANs.

In what follows, in Section 2, we discuss the challenge followed by a comprehensive study of the most representative GAN-based data augmentation studies in solving the hand pose estimation problem in Section 3. Additionally, the existing hand pose datasets, the evaluation metrics, and the state-of-the-art results on two common datasets are summarized in Section 4.

Finally, potential research directions in this rapidly growing field and conclusions are highlighted in Sections 5 and 6, respectively.

2. Challenge Analysis

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

- **Annotation difficulties:** Existing learning-based methods require a large number of labeled data to accurately estimate hand poses. However, acquiring precise labels is costly and labor intensive.
- **Lack of various modalities:** Most of the existing hand pose datasets only contain RGB images, depth frames or infrared images instead of paired modalities.
- **Requirement for variety and diversity:** The real datasets are limited in a quantity and coverage, mainly due to the difficulty of annotations, annotation accuracy, hand shape and viewpoint variations, and articulation coverage.
- **Occlusions:** Due to the high degree of freedom (DoF), the fingers can be heavily articulated. In particular, hand-object and hand-hand interaction scenarios are still a

big challenge, due to object occlusion and the lack of a large annotated dataset. Severe occlusion might lead to loose information on some hand parts or different fingers mistakenly. To handle occlusion, several studies resorted to a multi-camera setup from different viewpoints; however, it is expensive and complex to set up a synchronous and calibrated system with multiple sensors.

- **Rapid hand and finger movements:** Most conventional RGB/depth cameras cannot capture the speed of the hand motions and, thus, cause blurry frames or uncorrelated consecutive frames, which directly affect the hand pose estimation results.

Although many existing methods try to address these challenges with powerful learning-based approaches, as the effectiveness of generative deep learning aroused, many researchers try to address these through generative adversarial networks. Such methods dominate the benchmarks on large public datasets, such as NYU [25], ICVL [26], and FreiHAND [27]. In what follows, we first explain GANs, then we discuss studies on hand pose estimation, focusing on addressing the above challenges through data augmentation using GANs.

3. GAN-Based Hand Pose Data Augmentation

The generative adversarial network (GAN) is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data such that the model can be used to generate new examples as similarly as possible to the original dataset. GAN consists of two networks called the generator and discriminator; Figure 1a. The generator takes a simple random variable and generates new examples, and the discriminator tries to distinguish real samples from the generated ones. The two models are trained together in a zero-sum game—adversarial—until the discriminator model is fooled about half of the time, meaning that the generator model generates plausible examples. Although the original framework [17] has no control of what is to be generated and it is only dependent on random noise, in a later study [18], the authors introduced conditional-GAN, where they add the conditional input vector c concatenated with noise vector z and feed the resulting vector into the generator. This conditional GAN can be used to generate examples from a domain of a given type. This allows for some of the more impressive applications of GANs, such as image-to-image translation, style transfer, photo colorization, and so on.

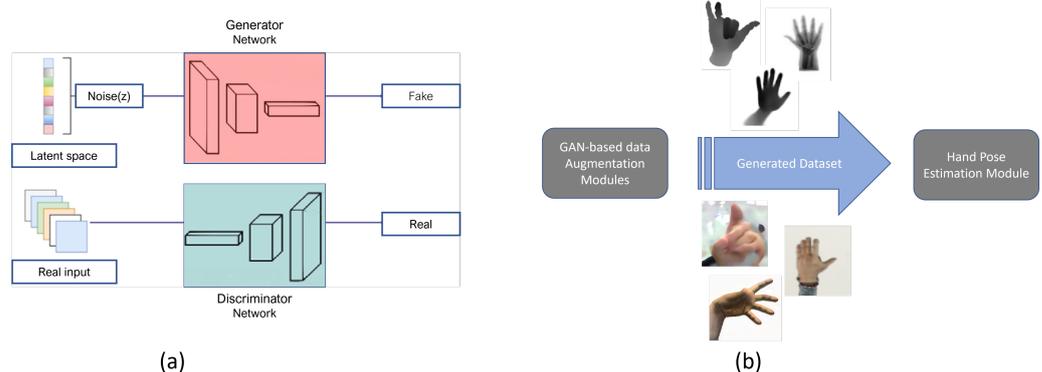


Figure 1. GAN-based hand pose data augmentation. (a) Overview on generative adversarial network, (b) procedure illustration of using generated data in HPE problem.

GANs are perhaps best known for their contributions to realistic image synthesis and model patterns of motion in video. GANs are able to enhance synthetic datasets such that the statistical distribution resembles a real-world dataset. Many approaches explore how to better manipulate images by applying GAN models [19,28,29]. Although image synthesis can be generated using a physical renderer, the difference between real and synthetic data is not considered in the image synthesis process. Moreover, GANs' successful ability to model high-dimensional data, handle missing data, and the capacity of GANs to provide

multi-modal outputs or multiple plausible answers made researchers more ambitious to the extent that they use GANs for the hand pose estimation problem either by generating data in new modalities or by realistic image synthesis through eliminating the domain gap between the synthetic and real data (Figure 1b). Despite the large body of work that has been conducted on hand pose estimation through GANs, to the best of our knowledge, this is the first review paper on data augmentation for hand pose estimation using GANs. Moreover, unlike other studies that focus on a single modality, such as depth or RGB, in this survey we cover various modalities. Below is a comprehensive survey on GAN-based hand pose data augmentation based on GANs' application.

3.1. Image Style Transfer and Data Augmentation

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

In [30], they proposed a data-driven approach to generate depth hand images given ground-truth hand poses using a generative model. The style transfer is applied to generate the image with the style equivalent to the style image and the content from the content image. The style and content are defined based on the loss functions by measuring how far away the synthesized images are from the perfect style transfer. The proposed style-transfer network aims to transform the smooth synthetic images to become depth hand images more similar to the real ones. Figure 2 shows the architectural structure of the developed method. It contains three parts: a generator to transform the 3D hand pose into a deep hand image, and a discriminator which determines the authenticity of the generated image and the style-transfer network. At the end, they performed 3D hand pose regression on generated depth hand images based on the residual convolutional neural network. Their approach was evaluated and analyzed on three publicly available datasets—NYU [25], ICVL [26], and MSRA gesture [31] datasets—and it was shown to boost hand pose estimation performance when used as training images.

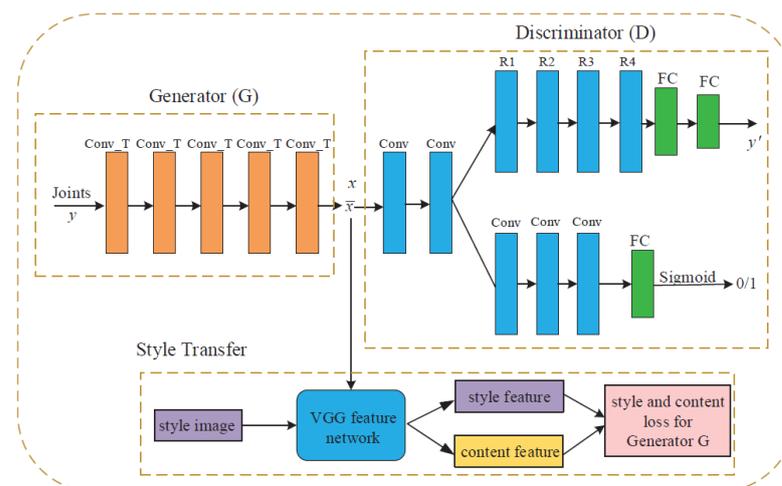


Figure 2. Flowchart of the proposed method in [30], covering the generator, the discriminator, and style-transfer networks in detail. Originally used in [30].

To increase the amount of training data, Shrivastava et al. [32] proposed a framework which uses simulated and unsupervised learning to fit a model that uses unlabeled real data to improve the realism of a simulator's rendered data. They performed an experiment using real hand depth maps from the NYU [25] hand pose dataset in an extended version

of SimGAN [32], and successfully added realistic noise to synthetic frames to better imitate imperfect real frames that are captured by depth cameras. Figure 3 gives an overview of the proposed model. Once the synthetic data are generated by a black box simulator, they are refined using a neural network called the ‘refiner network’. The refiner network is trained using adversarial loss from [17] to enforce the refined images similar to the real ones.

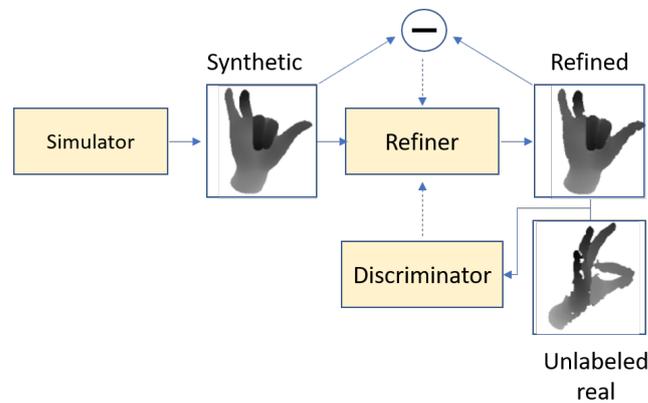


Figure 3. Overview of SimGAN; the self-regularization term minimizes the image difference between the synthetic and the refined images. Adapted from [32].

3.2. Domain Translation

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

Image-to-Image Translation

Image-to-image translation can be considered a type of image synthesis which maps an image from one domain to a corresponding image in another domain. It can be viewed as a generalization of style transfer since it not only transfers the style but also manipulates the attributes of the objects. Pix2Pix [29] and CycleGAN [19] are the most popular ones in supervised and unsupervised image-to-image translation. Pix2pix makes the assumption that paired data are available for the image translation problem that is being solved. In Pix2pix, model G was trained to translate images from domain X to domain Y. CycleGAN does the same, but additionally, it also trains a model F that translates images in the opposite direction—from domain Y to domain X. CycleGAN was created in order to support working with unpaired data since having paired data available is actually rather rare, and collecting such data can require a large amount of resources.

In [33], Chen et al. suggested blending a synthetic hand poses generated by an augmented reality (AR) simulator with real background images to generate more realistic hand images, which later served as training data. Inspired by the pix2pix [29] which leverages the shape map to constrain the output image, they proposed a tonality-alignment

GAN (TAGAN) to take the color distribution and shape features into account. Evaluation on multiple hand pose datasets indicates that their proposed approach outperforms state-of-the-art methods in both 2D and 3D hand pose estimation. Figure 4 gives an overview of the proposed method.

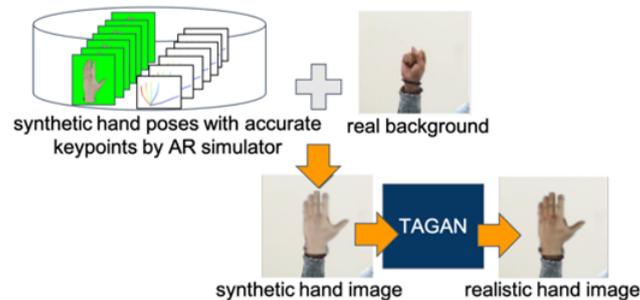


Figure 4. Overview of the TAGAN method for realistic hand image synthesis [33]. Synthetic pose by an AR simulator is blended with real background to yield a synthetic hand image, which is then fed to the proposed TAGAN to produce a more realistic hand image. Originally used in [33].

In another study by Wu et al., they proposed to directly generate realistic hand images from 3D pose and synthetic depth maps. However, unlike pose-guided person image generation, pose-guided hand generation is more challenging due to self-similarity and self-occlusion. To address these difficulties, they proposed a four-module model, MM-Hand, which contains 3D pose embedding, multi-modality encoding, progressive transfer, and image modality decoding [34]. They aimed to convert 3D hand poses to depth maps using a depth map generator. More specifically, in the 3D pose embedding module, they project the 3D hand pose onto a 2D image, given the projection matrix, which is followed by connecting the keypoints on each finger with an ellipsoid, using different colors. Then, a palm surrogate is formed through connecting a polygon from the base of each finger and wrist. Then, the depth map generator, which is a pix2pix-based model, is trained to synthesize depth maps based on any given 3D pose. Their experimental results show that the augmented hand images by their proposed approach significantly improved the 3D hand pose estimation results, even with reduced training data. The synthesized hand images using the proposed MM-Hand on the two benchmark datasets STB and RHP are shown in Figure 5.

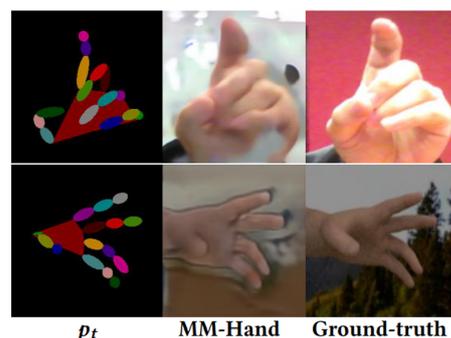


Figure 5. Qualitative results under MM-Hand model originally reported in [34]. Synthesized hand images using MM-Hand on two datasets, STB and RHP. From top to bottom: the STB dataset and the RHP dataset.

Moreover, to address the lack of various modalities problem, the authors in [35] presented a depth-image guided GAN model named DGGAN, which includes two sub-networks: a depth-map reconstruction module and a hand pose estimation module. Once the depth-map reconstruction module is trained using the GAN loss, it is able to generate a depth map of a hand based on the RGB input image. The second module trained using

the task loss estimates hand poses from the input RGB and the GAN-reconstructed depth images. They aim to reconstruct the depth map from RGB hand images in the absence of paired RGB and depth training data. Once the depth maps are constructed from the RGB images, the hand pose estimation modules takes both RGB and depth images to estimate the 3D hand pose first by estimating the 2D hand keypoints on the RGB images followed by regressing the 3D hand poses from the inferred 2D keypoints. Next, exploiting the reconstructed depth map, it regularizes the inferred 3D hand poses. Experimental results on multiple benchmark datasets show that the synthesized depth maps produced by DGGAN are quite effective, yielding new state-of-the-art results in estimation accuracy by notably reducing the mean 3D end-point errors (EPE).

In another study [36], to generate new modalities, Haiderbhai et al. introduced a novel architecture based on the pix2pix model. They proposed a method of synthetic X-ray generation using conditional generative adversarial networks and created triplets for X-ray, pose, and RGB images of natural hand poses sampled from the NYU hand pose dataset. As a result, they introduced a two-module network. The first one aims to generate a 2D image of the pose, given the RGB input. Next, the output is stacked with the original RGB, which is used as input for the second module, which is identical to the pix2pix architecture. Their proposed model, pix2ray, has the advantages of creating X-ray simulations in situations where only the 2D input is available and generating more clear results, especially in occluded cases.

In [37], to improve hand pose estimation on weakly blurred infrared (IR) images under fast hand motion, the authors proposed a method based on domain transfer learning. The proposed model consists of a hand image generator (HIG), hand image discriminator (HID) and three hand pose estimators (HPE). The HIG synthesizes a depth image given input IR images. To train the HIG network, adapted by [29], they used the pair of unblurred depth and IR images with slow hand movement. The HID classifies whether the generated depth map conforms to the human hand depth map. The HPEs estimate the hand skeleton given an input depth image from the actual depth sensor, synthesized depth map, and IR image. It is worth mentioning that collecting depth and IR images from a single sensor eliminates the additional effort for depth image labeling. Moreover, since consistency loss is back propagated from the results of HPE, given the real depth image, the training is self-supervised. The proposed model is able to effectively improve hand pose estimation results in infrared images by generating un-blurred depth images as shown in Figure 6.

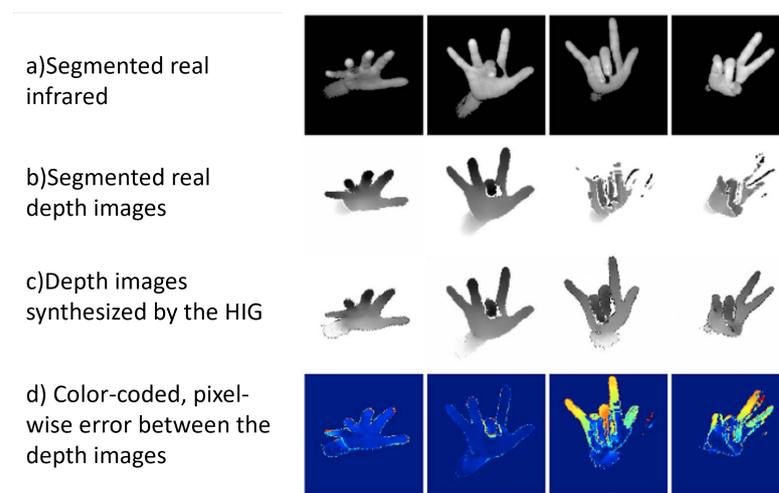


Figure 6. The HIG synthesizes a depth map from an infrared map. In the case of slow motion (the first and second column), the largest discrepancy is shown near the outline of the hand due to sensor noise. In the case of fast motion (the third and fourth column), the largest discrepancy is shown in blurry fingers. Originally reported in [37].

Since acquiring a large paired dataset can be difficult and expensive, inspired by CyclicGAN, Mueller et al. applied cycleGAN for realistic appearances of generated synthetic samples to reduce the synthetic-real domain gap [38]. They proposed a translation model, named GANerated, based on cycle-consistent adversarial networks (CycleGAN) to transfer the synthetic images to “real” ones so as to reduce the domain shift between them. Mueller et al. controlled the process through these two objectives: first converting synthesized image to real and calculating synth2real loss, and again converting the result to synthesized image and calculating real2synth loss. To make the images even more realistic, they also randomly put some background behind the hands. To simulate the occlusion, they artificially put some objects in front of the hand.

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

In another study [39], inspired by cycleGAN [19], the authors applied a generative adversarial network to estimate hand poses through one-to-one relation between the disparity maps and 3D hand pose models. They aimed to enrich the existing dataset by augmenting them. Unlike other studies, they synthesized data in the skeleton space (instead of depth-map space), where data manipulation is intuitively controlled and simplified and, thereafter, automatically transfers them to realistic depth maps. Their proposed model consists of three networks: hand pose generator (HPG), hand pose discriminator (HPD), and hand pose estimator (HPE). The job of HPG is to generate a hand based on the 3D representation of joints while the HPD tries to determine how real or fake the generated samples are. The HPE is responsible for estimating the 3D hand pose based on the input depth map. During the training, these three networks are optimized to reduce the error of HPE. In the inference time, the algorithm refines the 3D model, which is guided by HPG to generate the most realistic depth map. More detailed architecture can be found in [39].

Although the recent studies try to solve an issue of lacking reliable RGB/depth datasets through generations of hand images, most of these works have focused on the generation of realistic appearances of hands without considering the temporal information. In [40], leveraged temporal information, they presented an unsupervised domain adaptation strategy based on CycleGAN for 3D hand-object joint reconstruction. Exploited by 3D geometric constraints and cycle consistency, their approach is able to effectively transfer annotation from the synthetic source images to an unlabeled real target domain. Moreover, by embedding short-term and long-term temporal consistency loss, the proposed model leverages unlabeled video to fine tune the model and outperforms the state-of-the-art models on benchmark datasets.

4. Results and Discussion

4.1. Benchmark Datasets

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

- **NYU Hand Pose Dataset** It has 72,000 images as training and 8000 as testing data. Data are collected by 3 Microsoft Kinect cameras from 3 different views with 36 3D annotations. It is the most commonly used dataset in the hand pose estimation problem since it covers a variety of poses in RGB and depth modalities.
- **Imperial College Vision Lab Hand Posture Dataset (ICVL)** The ICVL contains 300,000 training and 1600 images as testing images. All depth images are captured by Intel RealSense and, in total, 16 hand joints are initialized by the output of the camera and manually refined.

- **MSRA15** This includes 9 subjects with 17 different gestures. In total, it has 76,000 depth images with 320×240 resolution, collected by Intel's Creative Interactive Camera, with 21 annotated joints.
- **BigHand2.2M** It contains 2.2 million real depth maps collected from 10 subjects. Since it is collected by six magnetic sensors, it has precisely 6D annotations.
- **Stereo Hand Pose Tracking Benchmark (STB)** STB includes 18,000 frames, 15,000 for training and 3000 for testing with 640×480 resolution. The 2D keypoint locations are obtained using the intrinsic parameters of the camera.
- **Rendered Hand pose Dataset (RHD)** It has 43,986 rendered hand images from 39 actions performed by 20 characters. Each depth image comes with segmentation mask, 3D and 2D keypoint annotations.

Modality, the type of data (i.e., synthetic or real data), the number of joints and the number of frames, are summarized in Table 1.

Table 1. Summary of hand pose estimation datasets commonly used in data augmentation using GANs.

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

4.2. Evaluation Protocol

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

4.3. Quantitative and Qualitative Results

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

For the NYU hand dataset, we choose refs. [30,39] since the other studies with NYU do not provide the quantitative results and only compare the quality of synthesized images. In Figure 7, the results are illustrated with and without the use of synthetic images for training on the NYU dataset. As it is reported in [30], the developed method obtains 0.4 mm reduction of the average 3D joint error, compared with the current best performance by Pose-REN [41]. Moreover, the results from ref. [39] also indicate the 3.2 mm reduction in mean error due to the increase in training samples from the proposed GAN-based data augmentation model. Furthermore, the developed methods are compared by the percentage of frames at different maximum error thresholds in Figure 7b. It has shown that both studies [39] and [30] achieved higher accuracy compared to the HPE base lines, [42] and [41], respectively.

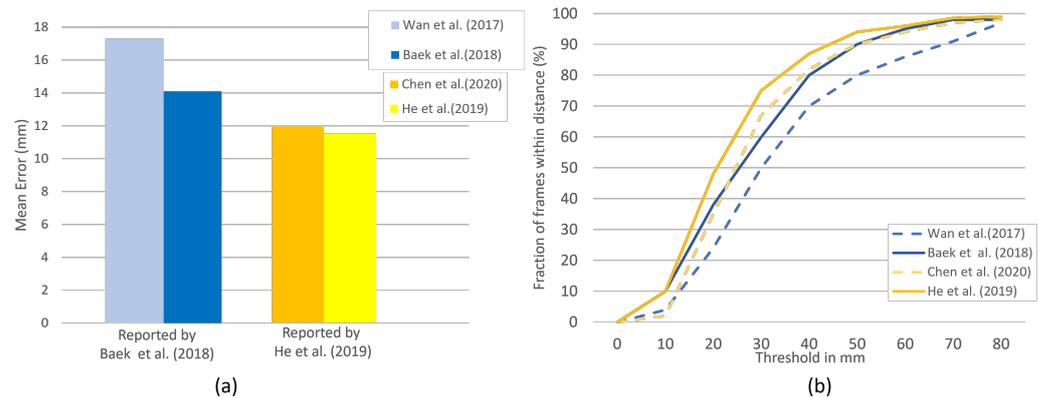


Figure 7. On NYU dataset, the contribution of the [30,39] methods to the accuracy are compared. (a) Mean error. (b) The fraction of frames over different maximum Euclidean distance error threshold. The larger the area under each curve, the better. (Best viewed on screen).

For the STB dataset, we compare DGGAN [35], GANerated [38], TAGAN [33], and MM-Hand [34] based on the reported PCK value in Figure 8. As it is mentioned, the larger the area under the curve, the higher the represented accuracy. The GANerated [38] has the lowest value of 0.965, compared to the others.

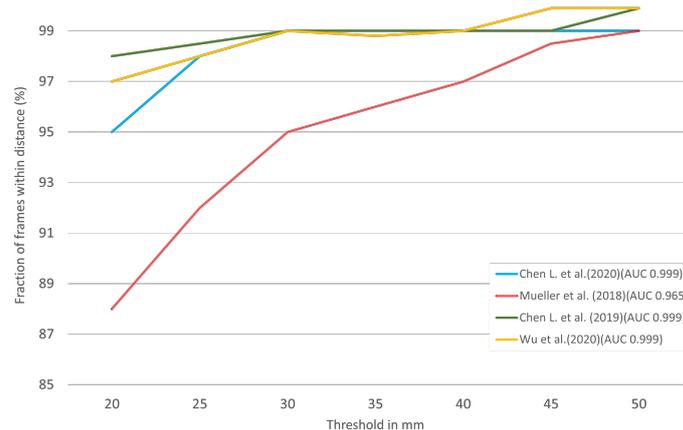


Figure 8. Comparison of [33–35,38] approaches for 3D pose estimation on the STB dataset. The fraction of frames over different maximum Euclidean distance error threshold. The larger area under the curve (AUC) represents better results. (Best viewed on screen).

5. Discussions and Future Directions

The explosion of interest in GANs is driven not only by their potential to learn deep, highly nonlinear mappings from a latent space into a data space and back, but also by their potential to make use of the vast quantities of unlabeled image data that remain closed to deep representation learning. While GAN has achieved great success due to its ability to generate realistic samples, GANs are still hard to train due to several common problematic unstable training and convergence behaviors, such as mode collapse, non-convergence and oscillatory behavior. To address the GAN challenges, recent studies are categorized in three main groups: proper architecture, loss function and optimization techniques. Therefore, a combination of careful balance during the adversarial optimization, finding new objective functions and the proper architecture can prove to outperform the state-of-the-art methods and can be a future research direction to explore. Moreover, due to the lack of robust and consistent metrics, coming up with good evaluation metric is still an open challenge to compare different GAN variants based on the visual assessment of the generated images.

On the other hand, despite the great performance of the current methods on hand pose estimation using GANs, still there remain difficulties in generalizing them to multi-hand

interaction. Furthermore, when it comes to evaluating GANs, there are many proposals but little consensus. Therefore, another future direction to study would be exploring good evaluation metrics in this field. Moreover, because of the interest of big technology companies in this field, perhaps in the near future, we can acquire much bigger and more generalized datasets generated by GAN and, therefore, very well-performing models on different modalities.

6. Conclusions

In this study, we reviewed the most recent state-of-the-art methods in data augmentation for hand pose estimation problem using GANs. Since most of the top-performed methods in HPE required large-scale training datasets, the current lack of large-scale training datasets that are accurate and diverse causes such methods to overfit. Moreover, manual hand-keypoint annotation is expensive, labor intensive, and still error-prone, often not being accurate enough either due to measurement errors or due to human errors. To address the quantitative and qualitative issues of hand pose training data and to enrich the hand pose dataset in modality and quantity, recent studies focus on using GANs to acquire comprehensive datasets in terms of quantity and modalities.

The main goal of this paper is to provide an overview of the methods used in hand pose estimation leveraged by GANs and point out the strengths and drawbacks of these methods. We classify these studies based on the use of GANs' application. In other words, we provide a detailed discussion of the most recent studies on image synthesis and image-to-image translation in HPE, where they aim to alleviate the burden of the costly 3D annotations in a real-world dataset. We aim to provide a simple guideline for those who want to apply GAN to the hand pose estimation problem and help further research in weakly/self-supervised learning.

Author Contributions: Conceptualization, F.F., M.S.N. and M.R.; introduction, F.F. and M.R.; challenge analysis, F.F.; GAN-based hand pose data augmentation, F.F. and M.S.N.; results and discussion, F.F. and M.R.; future direction, M.S.N.; writing—original draft preparation, F.F., M.R. and M.S.N.; writing—review and editing, F.F., M.R. and M.S.N.; visualization, F.F., M.R. and M.S.N.; supervision, F.K. and V.A.; project administration, F.K. and V.A.; funding acquisition, F.K. and V.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Krejov, P.; Gilbert, A.; Bowden, R. Guided optimisation through classification and regression for hand pose estimation. *Comput. Vis. Image Underst.* **2017**, *155*, 124–138. [[CrossRef](#)]
2. Zhou, Y.; Jiang, G.; Lin, Y. A novel finger and hand pose estimation technique for real-time hand gesture recognition. *Pattern Recognit.* **2016**, *49*, 102–114. [[CrossRef](#)]
3. Murugeswari, M.; Veluchamy, S. Hand gesture recognition system for real-time application. In Proceedings of the IEEE International Conference on Advanced Communications, Control and Computing Technologies, Ramanathapuram, India, 8–10 May 2014; pp. 1220–1225.
4. Carley, C.; Tomasi, C. Single-Frame Indexing for 3D Hand Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 101–109.
5. Isaacs, J.; Foo, S. Optimized wavelet hand pose estimation for American sign language recognition. In Proceedings of the 2004 Congress on Evolutionary Computation, Portland, OR, USA, 19–23 June 2004; pp. 797–802.
6. Jiang, S.; Sun, B.; Wang, L.; Bai, Y.; Li, K.; Fu, Y. Skeleton aware multi-modal sign language recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3413–3423.
7. Bilal, S.; Akmeliawati, R.; El Salami, M.J.; Shafie, A.A. Vision-based hand posture detection and recognition for Sign Language. In Proceedings of the 2011 4th International Conference on Mechatronics, Kuala Lumpur, Malaysia, 17–19 May 2011; pp. 1–16.

8. Kirac, F.; Kara, Y.E.; Akarun, L. Hierarchically constrained 3D hand pose estimation using regression forests from single frame depth data. *Pattern Recognit. Lett.* **2014**, *50*, 91–100. [[CrossRef](#)]
9. Taylor, J.; Bordeaux, L.; Cashman, T.; Corish, B.; Keskin, C.; Sharp, T.; Soto, E.; Sweeney, D.; Valentin, J.; Luff, B.; et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Trans. Graph. (TOG)* **2016**, *35*, 143. [[CrossRef](#)]
10. Liang, H.; Wang, J.; Sun, Q.; Liu, Y.J.; Yuan, J.; Luo, J.; He, Y. Barehanded music: Real-time hand interaction for virtual piano. In Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, Redmond, WA, USA, 27–28 February 2016; pp. 87–94.
11. Zhang, Y.; Meruvia-Pastor, O. Operating virtual panels with hand gestures in immersive vr games. In Proceedings of the International Conference on Augmented Reality, Virtual Reality and Computer Graphics, Ugento, Italy, 12–15 June 2017; pp. 299–308.
12. Liang, H.; Yuan, J.; Lee, J.; Ge, L.; Thalmann, D. Hough forest with optimized leaves for global hand pose estimation with arbitrary postures. *IEEE Trans. Cybern.* **2017**, *49*, 527–541. [[CrossRef](#)] [[PubMed](#)]
13. Wang, R.; Paris, S.; Popović, J. 6D hands: Markerless hand-tracking for computer aided design. In Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, 16–19 October 2011; pp. 549–558.
14. Ge, L.; Liang, H.; Yuan, J.; Thalmann, D. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1991–2000.
15. Doosti, B.; Naha, S.; Mirbagheri, M.; Crandall, D.J. Hope-net: A graph-based model for hand-object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6608–6617.
16. Hasson, Y.; Tekin, B.; Bogo, F.; Laptev, I.; Pollefeys, M.; Schmid, C. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 571–580.
17. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
18. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
19. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
20. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
21. Van Oord, A.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel recurrent neural networks. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 1747–1756.
22. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1857–1865.
23. Yuan, S.; Garcia-Hernando, G.; Stenger, B.; Moon, G.; Chang, J.Y.; Lee, K.M.; Molchanov, P.; Kautz, J.; Honari, S.; Ge, L.; et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2636–2645.
24. Supancic, J.S.; Rogez, G.; Yang, Y.; Shotton, J.; Ramanan, D. Depth-based hand pose estimation: Data, methods, and challenges. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1868–1876.
25. Tompson, J.; Stein, M.; Lecun, Y.; Perlin, K. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Trans. Graph. (ToG)* **2014**, *33*, 169. [[CrossRef](#)]
26. Tang, D.; Jin Chang, H.; Tejani, A.; Kim, T.K. Latent regression forest: Structured estimation of 3d articulated hand posture. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3786–3793.
27. Zimmermann, C.; Ceylan, D.; Yang, J.; Russell, B.; Argus, M.; Brox, T. FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape from Single RGB Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 813–822.
28. Chen, Y.C.; Lin, Y.Y.; Yang, M.H.; Huang, J.B. Crdoco: Pixel-level domain transfer with cross-domain consistency. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1791–1800.
29. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
30. He, W.; Xie, Z.; Li, Y.; Wang, X.; Cai, W. Synthesizing depth hand images with GANs and style transfer for hand pose estimation. *Sensors* **2019**, *19*, 2919. [[CrossRef](#)] [[PubMed](#)]
31. Sun, X.; Wei, Y.; Liang, S.; Tang, X.; Sun, J. Cascaded hand pose regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 824–832.
32. Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; Webb, R. Learning from simulated and unsupervised images through adversarial training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2107–2116.

33. Chen, L.; Lin, S.Y.; Xie, Y.; Tang, H.; Xue, Y.; Lin, Y.Y.; Xie, X.; Fan, W. TAGAN: Tonality aligned generative adversarial networks for realistic hand pose synthesis. In Proceedings of the 30th British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019.
34. Wu, Z.; Hoang, D.; Lin, S.Y.; Xie, Y.; Chen, L.; Lin, Y.Y.; Wang, Z.; Fan, W. Mm-hand: 3d-aware multi-modal guided hand generative network for 3d hand pose synthesis. *arXiv* **2020**, arXiv:2010.01158.
35. Chen, L.; Lin, S.Y.; Xie, Y.; Lin, Y.Y.; Fan, W.; Xie, X. DGGAN: Depth-image guided generative adversarial networks for disentangling RGB and depth images in 3D hand pose estimation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Village, CO, USA, 1–5 March 2020; pp. 411–419.
36. Haiderbhai, M.; Ledesma, S.; Lee, S.C.; Seibold, M.; Fürnstahl, P.; Navab, N.; Fallavollita, P. Pix2xray: Converting RGB images into X-rays using generative adversarial networks. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 973–980. [[CrossRef](#)] [[PubMed](#)]
37. Park, G.; Kim, T.K.; Woo, W. 3D Hand Pose Estimation with a Single Infrared Camera via Domain Transfer Learning. In Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Porto de Galinhas, Brazil, 9–13 November 2020; pp. 588–599.
38. Mueller, F.; Bernard, F.; Sotnychenko, O.; Mehta, D.; Sridhar, S.; Casas, D.; Theobalt, C. Generated hands for real-time 3d hand tracking from monocular rgb. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 49–59.
39. Baek, S.; Kim, K.I.; Kim, T.K. Augmented skeleton space transfer for depth-based hand pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8330–8339.
40. Qi, M.; Remelli, E.; Salzmann, M.; Fua, P. Unsupervised Domain Adaptation with Temporal-Consistent Self-Training for 3D Hand-Object Joint Reconstruction. *arXiv* **2020**, arXiv:2012.11260.
41. Chen, X.; Wang, G.; Guo, H.; Zhang, C. Pose guided structured region ensemble network for cascaded hand pose estimation. *Neurocomputing* **2020**, *395*, 138–149. [[CrossRef](#)]
42. Wan, C.; Probst, T.; Van Gool, L.; Yao, A. Crossing nets: Dual generative models with a shared latent space for hand pose estimation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.