applied sciences

MDPI

Article

A Comprehensive Study on Deep Learning-Based 3D Hand Pose Estimation Methods

Theocharis Chatzis *,†, Andreas Stergioulas *,†, Dimitrios Konstantinidis, Kosmas Dimitropoulos and Petros Daras

Visual Computing Lab at Information Technologies Institute of Centre for Reseach and Technology Hellas, VCL of CERTH/ITI Hellas, 57001 Thessaloniki, Greece; dikonsta@iti.gr (D.K.); dimitrop@iti.gr (K.D.); daras@iti.gr (P.D.)

* Correspondence: hatzis@iti.gr (T.C.); andrster@iti.gr (A.S.)

† These authors contributed equally to this work.

Received: 8 September 2020; Accepted: 25 September 2020; Published: 30 September 2020

Abstract: The field of 3D hand pose estimation has been gaining a lot of attention recently, due to its significance in several applications that require human-computer interaction (HCI). The utilization of technological advances, such as cost-efficient depth cameras coupled with the explosive progress of Deep Neural Networks (DNNs), has led to a significant boost in the development of robust markerless 3D hand pose estimation methods. Nonetheless, finger occlusions and rapid motions still pose significant challenges to the accuracy of such methods. In this survey, we provide a comprehensive study of the most representative deep learning-based methods in literature and propose a new taxonomy heavily based on the input data modality, being RGB, depth, or multimodal information. Finally, we demonstrate results on the most popular RGB and depth-based datasets and discuss potential research directions in this rapidly growing field.

Keywords: computer vision; deep learning; neural networks; 3D hand pose estimation

1. Introduction

Markerless hand pose estimation can be defined as the task of predicting the position and orientation of the hand and fingers relative to some coordinate system, given an RGB image and/or volumetric data captured from a depth camera. It is an important research topic and even a fundamental one for the creation of immersive virtual and augmented reality (VR/AR) systems, efficient gesture-based human-computer interaction (HCI) systems and 3D humanoid reconstruction systems [1–4]. Accurate hand pose estimation can enhance user experience in VR systems by enabling the performance of plausible and realistic virtual hand movements, as well as contribute towards a better understanding of human actions in smart HCI systems [5–8], thus enabling a more intelligent interaction between users and smart systems. Apart from the aforementioned applications, hand pose estimation is crucial in a number of other tasks, such as gesture recognition [9,10], action recognition [11,12], support systems for motor impairments patients [13], sign language recognition [14–18], and representation [19]. In sign language recognition especially, accurate hand pose estimation is beneficial for promoting social inclusion and enhancing accessibility in the Deaf community.

Although there has been a rapid progress in hand pose estimation, it still remains a challenging task with unique characteristics and difficulties, due to hardware limitations and constraints manifesting from the physiology of hands. The most important challenges that are present when implementing a markerless hand pose estimation method are the following:

- High articulation: A human hand has more than 20 degrees of freedom (DOF) [20], therefore, a lot of parameters are needed in order to properly model the complexity of hand, as well as finger movements.
- Occlusions: During the performance of a gesture, fingers of the same hand may be strongly occluded by each other, other body parts, or objects. This could potentially result in hidden hand parts or different fingers mistakenly inferred in the same location.
- Low resolution: The size of a hand, let alone a finger, occupy a small area in an RGB image or a depth map, if it is not the focus of attention or there is some distance between the camera and the hand. Additionally, limitations in camera technologies (e.g., lens resolution, depth sensing capabilities) may lead to inaccurate data, further hindering the results of hand pose estimation methods.
- Annotated data: The creation of annotated 3D data is a costly and time-consuming task. In order to capture accurate 3D data, an expensive marker-based motion capture system or a massive multi-view camera setting is required [21].
- Rapid hand and finger movements: Usually, hand motion is highly non-monotonic and consists of very fast movements, sudden stops, and joint rotations that are not met in other body parts. Currently, most conventional depth and RGB cameras can support 30 to 60 frames per second (fps) but still fall short to capture the speed of hand motions, resulting in blurry frames or uncorrelated consecutive frames.

Recently, the inclusion of neural networks in computer vision tasks, such as image classification [22,23], object detection [24,25], semantic segmentation [26,27], human pose estimation [28,29], camera recalibration for mobile robotics, navigation and augmented reality [30], etc., has yielded state-of-the-art results in such tasks. Moreover, recent end-to-end Deep Neural Networks (DNNs), e.g., convolutional neural networks (CNNs) [31], recurrent neural networks (RNNs) [32], auto-encoders [33], and generative adversarial networks (GANs) [34], have proven to be capable of extracting much more meaningful features from the available data than previous handcrafted techniques. The employment of such networks has further raised the performance of methods that engage with complex data, such as RGB images. Consequently, hand pose estimation has also progressed rapidly by the use of neural networks [35–37]. For example, one of the fundamental challenges of hand pose estimation, namely hand localization, has been alleviated by the utilization of DNNs [38]. Additionally, DNNs also enabled the implementation of 3D hand pose estimation methods from RGB images and videos, and recent methods are one step further in being applied on real world scenarios, whereas older methods were limited on depth data and required sub-optimal rearrangements for each specific dataset.

Different from existing review papers that are solely focused on depth-based methods [39,40], in this paper, we conduct a comprehensive study on the recent DNN-based 3D hand pose estimation methods, taking into account depth-based, RGB-based, and multimodal methods. More specifically, our contributions can be summarized as follows:

- We provide a comprehensive overview of the modern deep learning methods on 3D hand pose estimation, along with a brief overview of earlier machine learning methods for context. For each presented work, we describe the proposed method and the network architecture.
- We propose a new taxonomy for a better categorization and presentation of deep-learning-based 3D hand pose estimation methods. More specifically, in this paper, the following three categories are used: RGB-based, depth-based, and multimodal approaches.
- We present the most significant datasets employed by the research community, along with state-of-the-art results on each one.
- We draw conclusions and suggest possible future directions in terms of datasets and model architectures.

In order to explore the literature, we utilized the following digital libraries: IEEEXplore, Springer(link), arxiv.org, and the Google scholar search engine. The screening criteria for the selection of the works presented in this paper are: (a) literature works that address 3D hand pose estimation (2D methods are excluded), (b) literature works based on single-view camera setups (multi-view camera setups are excluded), (c) recent deep-learning-based methods published from 2014 until July 2020, (d) methods utilizing standalone visual marker-less input (i.e., no specialized equipment, such as gloves, markers, or specialized sensing equipment), and e) works that achieve state-of-the-art results on well-known and highly experimented datasets, such as Imperial College Vision Lab (ICVL) [41], NYU Hand Pose Dataset [42], Rendered Handpose Dataset (RHD) [43], Stereo Hand Pose Tracking Benchmark (STB) [44], and HANDS 2017 [45].

The rest of this survey is organized as follows. In Section 2, previous categorization methods and early machine learning methods are described. DNN-based 3D hand pose estimation methods are classified and reviewed explicitly in Section 3. Since training data play an important role in DNN-based methods, the existing 3D hand pose estimation datasets, the utilized evaluation metrics, and state-of-the-art results in each dataset are summarized in Section 4. Finally, conclusions and potential future research directions are highlighted in Section 5.

2. Previous Categorizations and Early Machine Learning Methods

2.1. Previously Proposed Categorizations

Various taxonomies have been proposed in order to categorize 3D hand pose estimation methods. Erol et al. [20] considered two main categories: appearance-based and model-based approaches. Appearance-based methods [46–52] do not need a prior knowledge about the hand pose. Instead, they extract features from the input images and learn a direct mapping to the hand pose space, using either random forests or deep convolutional neural networks. The most distinguishing characteristic of the appearance-based methods is that they require extensive training, and they do not employ a hand model. Consequently, the performance of these algorithms depends heavily on the quality, as well as the quantity, of training data. Moreover, since the encoded image representations include all the necessary information that is needed to predict a hand pose, these methods are not computationally expensive. On the other hand, model-based algorithms [53–59] establish a prior 3D hand model to match the images into the predefined model. At each frame, an algorithm performs an exploration in order to acquire the pose and shape of the hand model that best matches the features extracted from the input image. These approaches require a good initialization for the first frame, while, for the subsequent ones, they can either utilize a prediction model or use the predicted pose from the previous frame. Afterwards a similarity function measures the discrepancy between the ground-truth pose and the selected one. Eventually, an optimization strategy is necessary for model convergence. The most popular optimization method is particle swarm optimization (PSO) [60] and iterative closest point (ICP) [61]. Model-based methods are better at handling self-occlusions and other factors, such as rotation and scaling, but this comes at a cost since they tend to be complex and computationally expensive.

Furthermore, 3D hand pose estimation methods can be divided into regression-based [62,63] and detection-based [43,64] methods, according to the algorithm they apply. The key objective of methods that fall into the first category is to directly predict joint locations, which are continuous numeric values corresponding to the input image. In general, regression-based algorithms predict the output values based on input features extracted from training data, modeling dependencies between target poses and input images. On the contrary, detection-based methods predict a probability density map of each joint, where the probability at each point denotes how likely it is that the specific keypoint is located at this position. Specifically, these methods use volumetric heatmaps to form the joint location predictions and, afterwards, utilize an argmax function in order to obtain the position of a keypoint from a computed probability map.

Another categorization can be defined based on the strategy of finger and palm pose estimation. Holistic methods [65,66] estimate the hand pose by receiving the whole hand region as input, which is more robust but not very accurate at estimating fingers. Hierarchical methods [35,63,67] estimate the finger pose conditioned on the estimated palm pose. As a result, they achieve more accurate predictions, but they rely on the noisy estimation of palm pose.

Oikonomidis et al. [56] differentiated between disjoint evidence methods and joint evidence methods, according to the way that partial evidence regarding the individual rigid parts of the hand contributes to the final prediction. Disjoint evidence methods [68,69] are computationally efficient, as they reduce the search space by considering individual parts in isolation, prior to evaluating them against observations. The disadvantage of these methods is that they need additional mechanisms to explicitly handle part interactions, such as collisions and occlusions. At the other end of the spectrum, joint-evidence methods [56,70–72] consider all parts in the context of full object hypotheses. Part interactions are easily modeled, but their computational requirements are rather high.

This paper focuses mainly on the most recent DNN-based methods due to their superior performance on the 3D hand pose estimation field, in comparison to early machine learning methods. Thus, we initially categorize these methods according to their input modality to RGB-based, depth-based, and multimodal methods. The intuition behind our taxonomy scheme derives from the fact that such categorization of methods further improves the comprehensibility of the current survey, and, additionally, it could potentially be very useful for researchers to attend previous works and explore new directions based on the available data.

However, for the sake of completeness and due to the fact that machine learning techniques were the staple for the initially proposed hand pose estimation methods, the following sub-section describes the most influential and fundamental early machine learning methods.

2.2. Early Machine Learning Methods

Early works on the field of 3D hand pose estimation employ traditional machine learning techniques side by side with specialized technologies, such as gloves [73], that allow for better optimization. Amongst them, the Random Forest [74] algorithm and its variations are the most effective and excessively utilized. The Random Forest method is a machine learning ensemble meta-algorithm for classification or regression that falls under the category of bootstrap aggregating (bagging). It consists of a multitude of decision trees and outputs the class with the most predictions or the mean prediction of the individual trees, depending on the task.

Keskin et al. [68], inspired by the huge success of Reference [69], applied the revolutionary—at the time—approach of intermediate representations on the hand pose estimation field, instead of directly inferring the hand pose in a high-dimensional space. Initially, a single depth image was segmented into a dense probabilistic hand part labeling. Subsequently, the authors employed a computationally efficient depth comparison and assigned at each pixel a hand part label with Random Forests, trained on a large amount of synthetic depth images. Lastly, the mean shift algorithm was employed to estimate the probability of each class label with weighted Gaussian kernels placed on each sample. The same authors proposed two novel multi-layered Random Forest networks, namely the Global Experts Network (GEN) and the Local Experts Network (LEN) [75], as an extension to their previous work. They trained Random Forests on clusters of the training set to classify each pixel into hand parts and introduced a Random Forest variation, called Shape Classification Tree (SCT) to assign a cluster label to each pixel. Then the input image was directed to the correct expert: either pose label for the whole image (GEN) or local clusters for individual pixels (LEN).

However, since the above approaches required a large amount of per-pixel labeled samples, the use of synthetic depth images led to performance deterioration when applied in real-world data. Tang et al. [76] addressed this drawback using a unified single-layered Semi-supervised Transductive Regression (STR) forest, which was trained on both a realistic and a synthetic dataset, in order to learn the underlying relationship between these sets. Yet, considering that the STR did not contain any structural information about the joints, a data-driven pseudo-kinematic technique was introduced to refine the joint predictions. Liang et al. [77] proposed a hand parsing scheme to extract a high-level description of the hand from depth images, which was robust to complex hand configurations. They developed a depth-context feature and introduced a distance-adaptive selection method to further enhance its performance. Furthermore, they utilized a novel Superpixel Markov Random Field (SMRF) framework to model the spatial and temporal constraints, resulting in less misclassified regions, in comparison to pixel-level filtering methods.

Tang et al. [41] presented a novel Latent Regression Forest (LRF), a divide-and-conquer search method guided by a latent tree model that reflected the hierarchical structure of the hand. This structured coarse-to-fine search recursively divided the hand region until all joints were located, combining the benefits of both holistic and pixel-based methods, while the hand topological model enforced kinematic constraints.

Methods that are based on pixel classification are less successful in the hand pose estimation field compared to human pose estimation. The major differences are the larger variations in finger articulation, occlusions, and camera viewpoints that are apparent in hand motions and the lack of structural information that these methods provide, leading to uncertain pixel classification, as Sun et al. [63] pointed out. To overcome these issues, the authors extended a previous work on human pose estimation [78], using a sequence of weak regressors to progressively estimate the hand pose. They proposed a cascaded hierarchical approach to regress the pose of different parts sequentially, based on the order of their articulation complexity, since, as they pointed out, different object parts did not share the same amount of variations and DOFs.

3. Deep Learning Methods

In this paper, we propose a new taxonomy that groups each work based on the modality of the input data. As such, there are three distinct categories: (a) depth-based, (b) RGB-based, and (c) multimodal approaches.

Each category is further divided into more sub-categories. The depth-based approaches are grouped based on whether they directly use 2D depth maps or process them to extract 3D data representations, such as point clouds or 3D voxel representations. The RGB-based approaches are categorized based on whether they directly regress the 3D hand joint locations or a heatmap of the joints or employ a predefined hand model. The last category describing multimodal approaches is divided according to the input modality, required during evaluation, to unimodal and multimodal evaluation methods. A visualization of the proposed taxonomy can be observed in Figure 1.

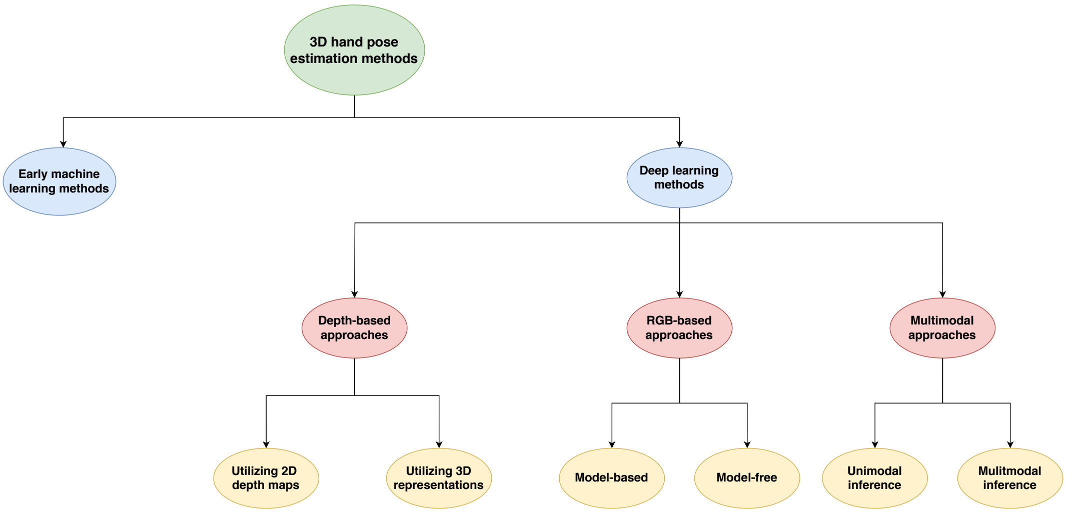

Figure 1. Proposed taxonomy. Each work is grouped by the modality of the input data, i.e., RGB, depth, and multimodal data, and then by the per-category distinct characteristics that best describe them.

In order to improve the comprehensibility of this work, an introduction of hand models is deemed necessary. In general, a hand model imposes a geometrical prior of feasible poses and possible joint rotations. Usually, hand models are employed to further refine the predicted pose and consequently constrain the neural network's predictions to humanly possible hand poses. Such models may be developed explicitly by the authors or instead use an already existing hand model.

Many hand models have been proposed that use different ways to represent a hand, such as assembled geometric primitives [56], sum of Gaussians [79], spheres [80], or triangulated meshes [81].

Recently, a differentiable triangulated hand model has been introduced, namely the MANO hand model [82], and is gaining a lot of traction. The hand is generated by a differentiable function $M(\beta, \theta)$, where β are the shape parameters and θ the pose parameters. The complete function is defined as:

$$M(\beta, \theta) = W(T(\beta, \theta), J(\beta), \theta, w), \tag{1}$$

where W is a Linear Blend Skinning function (LBS) which is applied to a rigged template hand mesh T with 16 joints J. The parameter w denotes the blend weights. By default, the MANO hand model's differentiable nature makes it very convenient to use alongside DNN-based methods; thus, most recent deep learning 3D hand pose estimation methods that employ a hand model utilize the MANO model.

3.1. Depth-Based Approaches

As was previously mentioned, depth-based approaches are classified based on the way that they utilize the depth map, meaning that the input to these methods can be either 2D depth maps or 3D representations, such as hand point clouds or 3D voxels. In the following sub-sections, we describe recent works with respect to the aforementioned categorization of Depth-based hand pose estimation approaches.

3.1.1. 2D Depth Map Utilization

Historically, the estimation of 3D hand joint locations was performed by leveraging depth images. Early deep learning methods did not deviate in that matter and employed depth data, as well, in the form of 2D depth images. In one of the earliest works, Tompson et al. [42] proposed the creation of a labeled ground-truth dataset of hand gestures by automatically annotating depth maps with the use of a Linear-Blend-Skinning (LBS) hand model, thus forming the NYU dataset. This dataset was then employed to train a 2-stage CNN network that outputs probability heatmaps for each joint.

Oberweger et al. [35] investigated different CNN network structures in order to find the most efficient one and introduced a pose prior module denoted as DeepPrior, which is effectively a bottleneck linear layer with significantly less neurons than the total number of joints ($\ll$3 J). The weights of the bottleneck layer were initialized by a PCA applied to the 3D hand pose data. The same authors, in DeepPrior++ [83], extended DeepPrior in three ways: (a) they incorporated the ResNet architecture into the network, (b) they introduced additional data augmentations during training, and (c) they improved the hand detection sub-task by introducing a CNN detection refinement network.

Both DeepPrior and DeepPrior++ tried to physically constrain the predicted hand pose by the use of the bottleneck layer. Zhou et al. [84] proposed to additionally take into consideration the functionality and importance of each finger. They designed a deep CNN network, named as Hand Branch Ensemble (HBE), that had three branches: one for the thumb, one for the index, and the last one to express the rest of the fingers. The features of each branch were fused together and projected to a lower dimensional representation by a similar bottleneck layer and further processed in order to directly regress the 3D joint positions.

Rad et al. [85] proposed an approach that applied domain transfer between synthetic and real depth images. More specifically, the authors rendered the 3D pose over the real images in order to create the corresponding synthetic ones and used a network to map the extracted features of the real images to the feature space of the synthetic ones. Accordingly, they trained a network to minimize the distance between the mapped and the synthetic features and a decoder trained on the synthetic images to infer the 3D pose.

Another way to constrain the final predicted hand pose is the use of intermediate heatmaps to guide the regression of the 3D hand joints. Du et al. [86] decomposed the hand pose estimation task into two sub-tasks, palm pose estimation and finger pose estimation, as well as the integrated heatmaps as constraints during the feature extraction phase from the input 2D depth map.

Differently, Ren et al. [87] proposed a re-parameterization layer to estimate intermediate representations, such as 3D heatmaps and 3D unit vector fields from joint positions. Their network design consisted of stacked joint regression modules with re-parameterization layers, except from the final regression module that directly outputted the final hand pose prediction.

Some works chose to employ a hand model in order to geometrically constrain and ensure the validity of the final hand pose predictions. Zhou et al. [88] used a method similar to Reference [35] CNN architecture to estimate the pose parameters of a hand model in joint-angle format. The hand model was from libhand [89] and has 23 joints with 26 DOFs. A differentiable forward kinematics function was employed to map pose parameters to 3D joint positions. Malik et al. [90] extended this work in order to generalize over different hand shapes, by jointly estimating 3D hand pose and bone-lengths of the hand skeleton. Initially, a CNN was employed to predict and bone scale and pose parameters, and, subsequently, a forward kinematics layer processed them to infer the 3D joint positions. The same authors [91] further imposed structural constraints, considering length and inter-distance of fingers, as well as kinematic distance, to better capture the hand shape. They regressed the center of the hand using a CNN, and the cropped hand region was then fed to another CNN, namely PoseCNN, to estimate the hand pose coupled with the above constraints. Sinha et al. [62] designed a method that took a cropped 2D hand depth map as input and outputted an activation feature that synchronized with other features in a population database, using a matrix completion method. Based on the global pose initialization estimated from the matrix completion method, the parameters of a hand model developed by the authors with 18 joints and 21 DOFs were estimated in order to infer the final hand pose.

Malik et al. [92] designed a CNN network that was better fitted to leverage their developed skinned hand model, which had 26 DOFs and 22 joints. Given a 2D depth image, the proposed network predicted the pose parameters θ, the bones scale α and the shape parameters β. A Hand Pose and Shape Layer (HPSL) took the estimated θ, α, and β parameters as input and predicted, using a forward kinematics function, the 3D joint positions P, as well as the vertices V. Both joint loss and vertex loss were taken into consideration to train the network. Later on, the same authors [93] introduced a weakly-supervised framework, incorporating intermediate bone representations. More specifically, a CNN network was initially employed to regress 3D bone vectors B from a single depth image. Accordingly, the B vectors were processed by a bone-to-joint layer to estimate 3D joint locations, preserving the hand structure and then a decoder was used to predict dense mesh vectors. Additionally, the authors employed a depth synthesizer during training to reconstruct the 2D depth image from the 3D hand mesh prediction, providing weak supervision.

Wan et al. [94] proposed a self-supervised method that fitted a sphere hand model to a depth image and then rendered the estimated hand pose to a depth image, for comparison with the original one. The hand model consisted of $N = 41$ spheres, 11 for the palm and 6 for each finger. For extra supervision, additional loss terms were added, more specifically a multi-view projection consistency loss, a kinematic constraints prior that penalized infeasible joint positions, a bone length loss, and a collision loss.

As the effectiveness of generative deep learning networks increased, several works tried to estimate hand poses by modeling the statistical relationship of the 3D pose space and the corresponding space of the input data using generative networks. GANs and Variational AutoEncoders (VAEs) have been proven pretty effective in capturing the statistics of the underlying feature space. However, there exist some early works that did not utilize such powerful networks but, rather, attempted to estimate 3D hand pose using a generative framework. For instance, Oberweger et al. [95] proposed a deterministic generative approach that was not based on sampling from a distribution. They designed a CNN network consisting of 3 sub-modules: a pose prediction network, a depth map synthesizer network, and a pose updater network. The goal was to directly predict the 3D joint locations from the initial depth map, use the predicted 3D pose to synthesize a depth map, and feed it into the updater CNN to further refine the predicted pose.

Wan et al. [65] designed a network (named Crossing Nets) that combined GANs and VAEs with the purpose of mapping the latent space of depth images to the latent space of 3D poses, assuming a one-to-one mapping between depth image and pose. The VAE part of Crossing Nets was responsible for the generation of the hand poses latent space, creating the distribution Z_y. The GAN part, sampled from the distribution of the latent space of depth images Z_x, in order to synthesize a depth map close to the original one. A fully connected layer was used to map the two spaces together. By mapping the two latent spaces, the network combined the generalization capabilities of the GAN part and the implicit pose constraints learned by the VAE and improved on the discriminative pose estimation.

Abdi et al. [51] proposed a method to make use of the abundance of synthetic samples datasets in order to accurately estimate the hand pose of real samples. Therefore, they created a shared latent space between three different domains: synthetic depth images, real depth images, and poses. The pose latent space was created by a VAE, whereas the real and synthetic images domain by a VAE-GAN hybrid network. The cycle consistency and weight sharing constraints [96] were put to use in order to learn the shared latent space of real and synthetic depth images.

Baek et al. [52], in contrast to the above mentioned generative works, proposed to augment the latent space of training examples by adding variations in viewpoints and shapes in the 3D pose domain. Afterwards, a GAN was trained to synthesize depth maps for the newly created 3D poses, as there were no corresponding depth map entries. The hand pose estimator and the hand pose generator were trained in a unified manner, enforcing the cycle consistency constraint.

A method that did not employ any of the constraints that were used in the aforementioned works is the one proposed by Wan et al. [97]. The authors proposed a dense pixel-wise estimation method, where a 2D CNN stacked hourglass network [98] estimated 2D heatmaps, 3D heatmaps, and dense 3D unit vectors, from a given 2D depth image. The 3D heatmaps coupled with the 3D unit vectors effectively compose a 3D offset vector between the 2D depth image pixels and the hand joints. The authors argued that working with offset vectors made the estimations translation-invariant and boosted the method's generalization capabilities to different combinations of finger poses.

Pixel-wise estimations, in general, are time-consuming, and estimations for background pixels may distract the neural network from learning effective features in the hand region [99]. Xiong et al. [100] proposed a method that did not densely regressed pixel-wise predictions, instead predicted 3D joint positions by aggregating the estimations of multiple anchor points. The main idea was to densely position anchors on the depth image and estimate the 3D joint positions by aggregating the estimations of multiple anchor points. Thus, the anchor points can be viewed as the local regressors with different viewpoints to the regressed joint. The network consisted of 3 branches, one for depth estimation, one for weighting the anchors of a joint (Anchor proposal branch), and an in-plain offset branch which computed the offset from an anchor point towards a joint. Afterwards, this method used the anchor proposals to locally regress the 3D location of a joint.

3.1.2. 3D Representation Utilization

There are two 3D representations that a 2D depth map can be converted to: (a) point clouds and (b) 3D voxels. Most works that leveraged 3D data utilized point clouds, while 3D voxels were not explored as much. Ge et al. [67] converted the 2D depth images to point clouds using the camera's intrinsic parameters (i.e., principal point location and focal length) and projected it to x, y, and z axis in order to get multiple views of the same hand pose. Three identical CNNs were employed to extract heatmaps from each projected view, which were later fused together in order to estimate the 3D hand pose.

Even though Ge et al. utilized the resulting point cloud of a depth image, they did not work directly with 3D points, since the designed network still required as input a 2D depth map. The development of PointNet [101], a network that works directly on point clouds, motivated works that utilized point clouds at input level. Later on, Ge et al. [36] proposed a hierarchical PointNet, the Hand PointNet, that takes N points as input and extracts the hand features in a hierarchical way, which enabled better generalization. Additionally, a fingertip refinement network processed the neighboring points of the estimated fingertips and further refined the hand pose. The global hand orientation was estimated by an OBB (oriented bounding box) by performing PCA on the 3D coordinates of the points that the OBB surrounded. The final 3D hand joint locations were estimated by direct regression. In a later work, Ge et al. [99] used a stacked PointNet architecture to directly take N sampled and normalized points from the point cloud as input and produce a set of heatmaps, as well as unit vector fields. These outputs were then put to use, in order to infer the offset of each sampled point from the hand joints and estimate the 3D coordinates of the latter.

On the other hand, Chen et al. [102] proposed a new network structure and used a spatial transformer to further improve the performance of the network. The designed framework consisted of a semantic segmentation network and a pose regression network. Two transformation matrices prediction networks were also implemented. The semantic segmentation network predicted for each point its corresponding semantic label. The pose regression network, using the semantic segmentation network's output together with the original point cloud, estimated the hand joint locations. The transformation matrices prediction networks estimated the necessary transformation matrices to normalize the input points and transform the 3D joints to the original space. Both the semantic segmentation and the regression network utilized the PointNet architecture.

Zhu et al. [103] utilized a modified version of the PointNet++ [104] architecture, as well as an ensemble strategy, where they proposed n possible poses for a hand point cloud and weighted them, in order to get the final predicted 3D joint locations. They also presented a data augmentation method that divided the hand point cloud into 16 parts based on Euler distance and then bent the fingers according to kinematics constraints, thus creating a new gesture.

In order to overcome PointNet's need for a meticulous preprocessing of the input point cloud, Li et al. [105] proposed a residual Permutation Equivariant Layer (PEL) [106] deep network to compute point-wise features and feed them to a point-to-pose voting module. The point-to-pose module consisted of 2 fully connected layers, one responsible for the computation of the importance of each point to the final pose estimation and the other for the production of 1-dimensional heat vectors or direct 3D location values, depending on the selection between detection and regression. In their work, the regression-based variation slightly outperformed the detection-based variation.

Chen et al. [107], in order to utilize depth images without 3D annotation, proposed a hand pose encoder-decoder network, where the encoder hierarchically extracted a vector representation from the point cloud and the decoder reconstructed the point cloud from the extracted vector. The encoder architecture was based on the SO-Net [108], which built a self-organizing map (SOM) [109] and performed a hierarchical feature extraction of a point cloud and SOM nodes. Additionally, a hand pose regression module made use of the encoder's output by concatenating the extracted vector with the latent point cloud features, to directly estimate the output hand pose. The initial input were N sampled points from the point cloud, with $N = 1024$.

Different from the above, Moon et al. [64] cast the objective of 3D hand pose estimation to a voxel-to-voxel likelihood for each joint. The 2D depth map was projected into a 3D voxel representation and then fed into a 3D CNN encoder-decoder network that was based on the hourglass network [98] and directly predicted the per-voxel likelihood for each keypoint. The voxel with the highest likelihood was considered to be the joint location and as such, it was wrapped to the real world coordinates. The authors suggested that, by working with voxels instead of a 2D depth map, they eliminated the perspective distortion that may occur when there is a direct regression from 2D to 3D. Similarly, Huang et al. [110] used a 3D hourglass model to directly regress 3D joint heatmaps. In order to add extra hand skeleton constraints, heatmaps for each bone were also estimated by their network, which worked as intermediate supervision. The number of bones varied depending on the dataset, for example, in the MSRA dataset, where there are 21 joints, the number of bones was set to 20.

Wu et al. [111] proposed a hand pose estimation network where intermediate supervision in the form of dense feature maps constrained the feature space of the joint predictions. The dense feature maps were explicitly designed to incorporate the geometric and spatial information of hand joints. Even through they also experimented with using 3D voxels as input, the best results were acquired by employing the projection technique of Reference [67] in combination with the outputs of Reference [64] as supervision feature maps.

Ge et al. [112] proposed a 3D CNN-based hand pose estimation method that could directly regress 3D joint locations from 3D voxels. More specifically, they represented the 3D hand as a 3D volume of projective Directional Truncated Signed Distance Function (D-TSDF) [113] values, which were fed into three 3D CNN layers and three fully connected layers. The authors suggest that a 3D CNN with 3D volumes as input, is easy to train end-to-end. In Reference [50], the same authors expended their previous work. The output of their 3D CNN network was a low dimensional representation of 3D hand joint relative locations in the 3D volume. By performing PCA reconstruction and coordinate transformations, they obtained the 3D hand joint locations in the camera's coordinate system. Additionally, they took into consideration the complete hand surface as intermediate supervision. A 3D U-Net [114] was applied on the input voxels in order to generate a truncated distance function (TDF) volume of complete hand surface points which was later concatenated with the original input voxel volume. A 3D DenseNet further processed the concatenated volumes and predicted the desired 3D hand pose. Both 3D U-Net and 3D DenseNet are pretrained separately.

More recently, Malik et al. [115] introduced a novel 3D CNN architecture to estimate 3D shape and 3D hand pose from voxelized depth maps. Initially, they employed a voxel-to-voxel pose regression network to estimate 3D joint heatmaps. The predicted 3D heatmaps coupled with the voxelized depth map were then processed by two sub-networks in order to obtain the voxelized shape and the hand surface. The first one was a voxel-to-voxel shape regression network that established a one-to-one mapping between the voxelized depth map and the voxelized shape. As for the second one, it was a voxel-to-surface regression network aimed to capture the hand topology, which the voxelized shape could not preserve. During training, a depth synthesizer was attached after each of these sub-networks to reconstruct voxelized depth maps, thus providing weak supervision. Furthermore, the authors utilized a shape registration module to fit the predicted hand surface to the predicted voxelized shape. Lastly, they proposed a 3D data augmentation scheme, applying 3D rotations in voxelized depth maps and predicted 3D heatmaps that led to enhanced network accuracy.

3.2. RGB-Based Approaches

The RGB-based methods are classified into two sub-categories, model-free and model-based methods. The model-free category includes methods that directly predict 3D joint positions or heatmaps, whereas the model-based category describes methods that predict parameters of a hand model, such as pose and shape.

3.2.1. Model-Free Approaches

One of the first deep learning methods that directly estimated 3D hand pose from RGB images is the work of Zimmermann et al. [43]. In their paper, the authors designed a network consisting of three sub-networks: a hand segmentation net that detected the hand and according to the predicted hand mask segmented the image, a 2D keypoint detection network that produced score maps for each joint and a 3D pose regression network that took the score maps and predicted the canonical 3D joint locations and a transformation matrix to the world coordinates. Additionally, the authors also developed a large synthetic dataset that facilitated the training process.

Iqbal et al. [116] proposed a 2.5D pose representation from single RGB images that was then reconstructed to 3D. For the purpose of obtaining 2.5D pose representations, both 2D joint locations, as well as depth values, are needed, so the authors designed a new heatmap representation, referred as 2.5D heatmap that consisted of a 2D heatmap and a depth map for each joint. However, no depth data were required during training, since the network learned to produce depth maps intrinsically. The network architecture consisted of an hourglass network which produced a latent 2.5D representation of the input RGB image, containing the latent 2D heatmaps and depth maps.

Spurr et al. [117] proposed a variant of the above mentioned network, since they considered 2.5D representations to be better suited with the set of loss terms that they developed. Those terms implicitly enforced biomechanical constraints on the 3D joint predictions, even for networks trained only with 2D supervision. Such biomechanical constraints supervise the range of valid bone lengths, the range of valid palm structure, and the range of valid joint angles of the thumb and fingers, and, as a result, the final predictions were constrained to anatomically plausible poses.

To differentiate from their previous work that relied on carefully designed loss terms to constrain the predicted poses to plausible ones, Spurr et al. [49] proposed a method that involved latent space alignment between the 3D pose space and the input data space. The authors considered as input data 2D keypoints, RGB images and to evaluate the generability of their method, 2D depth images. However, their main focus was on RGB images. The network consisted of two VAE networks, one for building the pose latent space and one for creating the input data space. Each VAE encoded and decoded the data to their respective modality, i.e., $RGB \rightarrow RGB$ and $3D \rightarrow 3D$. The latent space alignment was achieved by using the trained decoder of the $3D \rightarrow 3D$ VAE and the trained encoder of the $RGB \rightarrow RGB$ VAE.

In order to align the latent space of 3D hand poses and RGB images, Theodoridis et al. [118] proposed a novel variational framework that was trained in two distinct phases. During the first training stage, two VAEs were independently trained to create the cross-modal latent space of $RGB \rightarrow 3D$ and the single-modal latent space of $3D \rightarrow 3D$. Subsequently, in the next training phase, the above branches remained frozen and a variational module (mapper) was introduced with the purpose to map the cross-modal latent space to the well-structured and more informative single-modal one. Despite the fact that this framework was generic and not explicitly designed for the 3D hand pose estimation task, since they presented experiments on food image analysis datasets too, it yielded state-of-the-art results on the task of RGB-based 3D hand pose estimation among model-free approaches.

Yang et al. [119] proposed a method that constructed a refined latent space by disentangling the representations of 3D hand poses and hand images, called disentangled VAE (dVAE). Each latent space was decomposed to its meaningful components; the 3D pose space was decomposed to the canonical coordinates and the viewpoint factor spaces and the RGB space to the world coordinates 3D hand pose space and background. During training, they considered a decoder for $RGB \rightarrow RGB$, a decoder for $RGB \rightarrow 3D_{canonical}$, and one for $RGB \rightarrow Viewpoint$.

More recently, Gu et al. [120] proposed a latent space disentangling method, as well. However they chose to decompose the latent space to 3D hand pose space and modality specific characteristics. Two VAEs were designed to extract the latent space of each modality to 3D pose and modality information, with an extra discriminator connecting the two latent spaces by aligning the 3D hand space. An extra fully connected network was tasked with the translation of the RGB specific information to 3D hand pose modality information and an additional 3D pose modality adversarial discriminator was deployed in order to regulate the translation.

3.2.2. Model-Based Approaches

One of the first deep learning RGB-based methods that made use of a hand model was the work of Panteleris et al. [121]. The authors presented a method that employed the detection network YOLO v2 [122] for hand detection, a pretrained OpenPose [28] network for 2D joint positions prediction and an assembled geometric primitives hand model with 26 DOFs and 21 joints. The hand model was fitted by minimising the distances of the projections of its joints to the 2D joints predictions acquired from OpenPose, using non-linear least-squares minimization.

In Reference [123], Mueller et al. proposed the creation of a synthetic dataset, with an additional GAN for the translation of synthetic hand images to realistic hand images. The translation GAN was trained using the cycle consistency with a geometric consistency loss that preserved the hand pose during translation. A regression CNN was trained on the synthetic data, in order to predict 2D joint heatmaps, as well as 3D positions relative to the root joint. The 2D/3D predictions were utilized during the kinematic hand model fitting to ensure the anatomically correctness of the resulting hand poses.

In Reference [124], Zhang et al. incorporated the MANO hand model in their 3D hand mesh recovery method, called Hand Mesh Recovery (HAMR) framework. The mesh representations were extracted by regressing the mesh and camera parameters of MANO. The network was trained with five loss terms: a 2D heatmap loss, a camera perspective loss, a 2D keypoint projection loss, a 3D joint position loss, and a geometric constraints loss. The 3D joint locations were estimated by linear interpolations between the vertices of the mesh. The MANO model was also employed by Boukhayma et al. [125] to predict hand pose and shape from images in the wild. A ResNet-50 was used as the encoder network and generated the pose, shape and view parameters that the decoder, consisting of the MANO model and a reprojection module, processed in order to regress the final hand pose and shape estimation. By estimating the view parameters, the authors were able to project the estimated hand pose to 2D image coordinates and leverage 2D weak annotations as supervision. The encoder network was pretrained on a synthetic dataset created by the authors, in which hand images were paired with ground-truth camera and hand parameters. The training procedure was supervised by a combination of four loss terms: a 2D joint projection loss, a 3D joint loss, a hand mask loss, and a regularization loss that was applied on the encoder's hand model parameters outputs to reduce their magnitude and allow only physically plausible hand reconstructions.

Baek et al. [126] proposed a method that leverages the MANO hand model with 3D supervision, as well as hand segmentation masks and 2D supervision, to train a hand pose estimation CNN. Their network took an RGB image as input, estimated the 2D joint positions, and used them along with the image feature vector to predict the camera and the hand model parameters. At testing time, the initial 3D mesh estimation was further refined using gradient descent. The gradients were computed by projecting the predicted 3D mesh to 2D and comparing them with the predicted 2D joint locations. Differently from previous approaches, the 2D joints and segmentation masks were not kept fixed during the refinement step; instead, they recursively improved the 2D predictions.

While previous model-based methods employed a hand model to constrain the predicted poses to be humanly feasible, the structure of hands has not been exploited thoroughly. He et al. [127] presented an adversarial method where a generator predicted an initial prior pose using the MANO hand model and then a Graph Convolutional Network (GCN) refinement module was employed to further refine the prior pose. In particular, the GCN refinement module used the prior pose and the image features as inputs and predicted a deformation that was added to the features of the initial prior pose. The developed discriminator supervised over three input sources: 3D hand poses, hand bones computed from the 3D poses, and the input image.

3.3. Multimodal Approaches

Recently, a few works have leveraged auxiliary types of data in order to obtain better pose estimations and alleviate depth ambiguity and finger occlusions. Information, such as depth maps, heatmaps, point clouds, or even motion capture (MoCap) data, is better suited to capture the structure of the hand pose than RGB images. There are two ways of incorporating additional modalities into the framework: either as supplementary information available only during training in order to boost the RGB inference (unimodal inference) or as accessible information throughout the whole procedure (multimodal inference).

3.3.1. Unimodal Inference

Yuan et al. [128] were among the first to utilize depth data during training, by employing a two-staged training strategy to estimate poses from each modality with two CNNs. Initially, they regressed 3D joint locations from the depth-based network, while, on the second stage, they froze its parameters and used the RGB-based network in order to train paired images. Two losses were utilized, one for the pose regression task and one for mid-level representations. In addition, they embedded hand masks extracted from depth images into the mid-layers of the RGB-based CNN to alleviate the noisy background features.

To differentiate from the vast majority of 3D hand pose estimation methods, Cai et al. [129] proposed a weakly supervised approach to dismiss the necessity of 3D annotations. They adopted a framework that consists of an encoder-decoder architecture to estimate heatmaps and 2D joint locations, followed by a regression network to infer the 3D pose from the estimated heatmaps and a depth regularizer that was responsible for rendering a depth image from the predicted 3D hand pose. In this fashion, the depth regularizer coupled with the L1 loss between the generated depth image and the corresponding ground-truth can be seen as weak supervision for 3D hand pose regression. In the same spirit, Dibra et al. [130] introduced an unsupervised framework to address the necessity of real-world training data, on condition that paired depth maps are provided alongside RGB images. Initially, the authors used a network to localize the hand region, which was then processed by a CNN, namely SynthNet, trained explicitly on synthetic data to infer joint rotation angles. In order to allow for unlabeled real-world RGB images, they extended SynthNet with a depth loss component, using the ground-truth paired depth map and the predicted angles coupled with a point cloud sampled from their hand model.

On the other hand, not all datasets include paired RGB-depth images. To address this issue, Chen et al. [131] introduced a variant of GAN, denoted as depth-image Guided GAN (DGGAN), that comprised two sub-networks: a depth map reconstruction sub-network and a hand pose estimation sub-network. The first sub-network processed the RGB image and generated a depth map that the hand pose estimation sub-network leveraged in addition with the corresponding RGB image, in order to estimate a 2D hand pose, heatmaps, and 3D joint locations, as in Reference [129]. The key contribution of this work was that the reconstructed depth image, produced by the depth regularizer, was guided by the synthesized depth map.

Interestingly, Zhao et al. [132] proposed a framework to transfer knowledge from a dataset containing more than one source, e.g., RGB and depth maps, to a single source target dataset, using a teacher-student scheme. The authors modeled cross-modal knowledge as priors on the parameters of the student network; thus, it can be seen as a regularizer that can be learned by meta-learning.

Yang et al. [133] proposed to learn joint latent representation employing a VAE-based framework. More specifically, they encoded RGB, point clouds, and heatmaps, producing their respective latent spaces, and, afterwards, they aligned these manifolds using either KL divergence, reducing their divergence or product of Gaussian experts and creating an additional joint space. Subsequently, they decoded each latent space to every available modality. They obtained the best results with the latest method, in which the encoders of different modalities can be trained independently. It should be noted that the utilization of shared decoders had a great impact in aligning the latent spaces.

Ge et al. [134] employed a Graph CNN to infer the hand mesh from the estimated heatmaps and extracted RGB features. The first training stage also included the 3D hand pose inference from the predicted mesh in a fully supervised manner with heatmap, mesh, and 3D hand pose guidance. The second training stage was a fine-tuning on a real-world dataset in a weakly supervised manner, introducing a pseudo-ground-truth mess and depth loss. The major drawback of this method is that, in order to acquire ground-truth paired mesh information, a special synthetic dataset is compulsory, which is either rare to find or hard to create.

Zhou et al. [135] divided the problem into regressing root-relative scale-normalized depth and 2D coordinates following Reference [116] and, additionally, predicted root-relative normalized 3D coordinates to tackle depth-scale ambiguity. Initially, they introduced a joint detection network, called DetNet, in order to extract features from a real-world dataset using 2D joint supervision, while estimating 3D joint positions can be guided from a synthetic dataset. The authors also proposed IKNet, an inverse kinematics network that regressed the joint rotations from the predicted 3D pose in a single feed-forward pass. This framework was also the first work to use hand MoCap data to guide joint rotations during training.

3.3.2. Multimodal Inference

Mueller et al. [136] created a colored depth map from each paired RGB-depth image which was processed by a CNN to produce 2D hand position heatmaps and consequently localized the hand region. A second CNN was then employed to regress 3D joint locations, as well as 2D joint heatmaps, from the cropped colored depth map. Subsequently, they estimated the joint angles of a kinematic pose tracking framework from the predicted 3D joint positions and 2D joint heatmaps to refine the pose and achieve temporal smoothness.

Kazakos et al. [137] proposed a two-stream architecture to incorporate depth information into RGB images, investigating various fusion strategies and rules. In particular, they considered fusing at (a) input-level, which refers to placing the fusion block after the input layer, (b) feature-level, which employs the fusion block after any convolution, pooling or fully connected layer and (c) score-level that fuses the predictions of each stream. As for the rules, they considered max, sum, concatenation, and convolutional fusion for the feature-level fusion and concatenation and convolutional fusion for the input-level one. On the contrary, the authors introduced a learnable function for the last fusion technique, called locally connected fusion function. Intriguingly, they concluded that, regardless of the utilized strategy or rule, the recognition performance of the network was not significantly affected, and, furthermore, the two-stream framework barely outperformed the corresponding depth-based single-stream framework.

4. Datasets & Metrics

Datasets play an important role in DNN-based 3D hand pose estimation, since the performance of DNN methods is tied to the quality and quantity of training data. The existing hand pose datasets contain either synthesized hand images or real-world examples. Synthetic datasets are more cost-efficient to produce than real-world datasets, since the exact 3D annotations for computer generated hands can be easily extracted; however, the hand images are not realistic enough and hinder the performance in real-life scenarios. This section presents the most significant 3D hand pose estimation datasets for DNN-based methods, as summarized in Table 1, the evaluation metrics that are used to gauze the performance of a method, and results of recent state-of-the-art 3D hand pose estimation methods.

Table 1. A representative catalogue of 3D hand pose estimation datasets.

Dataset	Modality	Year	View	Type	# of Subjects	# of Joints	# of Frames
ICVL [41]	D	2014	3rd	Real	10	16	332.5 K
NYU [42]	D	2014	3rd	Real	2	36	81 K
MSRA15 [63]	D	2015	3rd	Real	9	21	76.5 K
HandNet [138]	D	2015	3rd	Real	10	6	212 K
BigHand2.2M [139]	D	2017	ego/3rd	Real	10	21	2.2 M
HANDS 2017 [45]	D	2017	ego/3rd	Real	20	21	1.2 M
SynHand5M [92]	D	2018	3rd	Synthetic	-	22	5 M
FreiHAND [140]	RGB	2019	3rd	Real	32	21	134 K
Dexter1 [79]	RGB+D	2013	3rd	Real	1	6	2 K
Dexter+Object [141]	RGB+D	2016	3rd	Real	2	5	3 K
RHD [43]	RGB+D	2017	3rd	Synthetic	20	21	44 K
STB [44]	RGB+D	2017	3rd	Real	1	21	18 K
EgoDexter [136]	RGB+D	2017	ego	Real	4	5	3190
SynthHands [136]	RGB+D	2017	ego	Synthetic	2	21	63.5 K

4.1. Datasets

In recent years, a few datasets suitable for 3D hand pose estimation have been proposed. Earlier datasets were composed only of depth data; however, since the development of robust methods that leverage RGB images, more datasets that contain both RGB and depth images have been introduced. In the following paragraphs, we have compiled and described the most frequently used datasets, omitting the ones that are rarely employed by the research community. Standard rules for dataset creation do not exist; therefore, the type of data (i.e., synthetic or real data), the number of joints and the resolution of the images varies between datasets.

Imperial College Vision Lab Hand Posture Dataset (ICVL) [41] contains a collection of sequences from 10 different subjects, resulting in a total of 332.5 K depth images, 331 K for training and 1.5 K for testing, with 320 × 240 resolution. The depth images have high quality and sharp outlines with little noise, but, as Reference [35] pointed out, a large number of samples are not correctly annotated.

NYU Hand Pose Dataset [42] includes (RGB-D) frames from 3 Kinect cameras, providing a frontal and two side views. The training set consists of 72 K samples for training performed by one subject, while the test set contains 8.2 K samples from two subjects, with a resolution of 640 × 480. Although it is composed of a wide variety of unique poses, its shortcoming is the fact that it provides only one hand shape in training set. In total, 36 joint are annotated but most works evaluate their performance on a subset of 14 joints. This dataset is considered challenging since depth maps exhibit typical artifacts of structured light sensors: noisy outlines and missing depth values.

BigHand2.2M [139] is a large-scale real-world dataset that was recorded using six magnetic sensors attached on the hand, thus providing quite accurate 6D measurements. It consists of 2.2 M depth maps, obtained from 10 subjects with 640 × 480 resolution. Locations of 21 joints were obtained by applying inverse kinematics on a hand model with 31 DOF with kinematic constraints.

MSRA15 [63] comprises about 76K RGB images with 21 annotated joints with 320 × 240 resolution, captured using a time-of-flight camera. It contains 9 subjects performing 17 gestures covering a large number of viewpoints. For evaluation the most common schema is the leave-one-subject-out cross-validation; the training is conducted on 8 different subjects and the evaluation on the remaining one. The major drawbacks of this dataset is its limited size and the high error rate as far as annotations are concerned, e.g., missing finger annotations [139].

Handnet [138] includes 202 K training and 10 K testing depth images, as well as more than 2.7 K samples for validation, with a resolution of 320 × 240. Samples were recorded with a RealSense RGB-D camera and depict the 6D postures of the hand, as well as the position and orientation of each fingertip.

HANDS 2017 [45] consists of about 1.1 M depth images at 640 × 480 resolution, 957 K for training and 295K for testing, sampled from BigHand2.2M [139] and First-Person Hand Action [142], combining the large diversity of hand configurations, hand shapes, and viewpoints of Reference [139] and occlusions contained in Reference [142]. The training set consists of 5 subjects, while the test set comprises additional 5 unseen subjects. The pose annotations for each image are the positions of 21 hand joints.

SynHand5M [92] is a million-scale synthetic dataset containing 5 M depth images, 4.5 M for training and 500 K for testing, with a resolution of 320 × 240. It provides annotations of 22 3D joint positions defined on 26 DOFs and 1193 3D hand mesh vertices, as well as joint angles and segmentation masks.

FreiHand [140] is a multi-view RGB dataset containing hand-object interactions. In total, it encompasses 134 K samples at 224 × 224 resolution, 130 K for training and 4 K for evaluation. Ground-truth 3D locations for 21 keypoints of the hand, as well as the 3D hand shape, the hand mask, and the intrinsic camera matrix, are provided.

Rendered Handpose Dataset (RHD) [43] consists of 43,986 rendered hand images, 41,258 for training and 2728 for evaluation. Totally, there exist 39 actions performed by 20 characters, with 320 × 320 resolution. For each image depth map, segmentation mask and 3D and 2D keypoint annotations are provided. This dataset is considered to be exceedingly challenging due to the fact that it contains large variations in viewpoints and hand shapes, as well as occluded fingers, large visual diversity, and noise.

Stereo Hand Pose Tracking Benchmark (STB) [44] comprises one subject performing 12 sequences with 6 different backgrounds. Altogether it includes 18 K frames, 15 K for training and 3 K for testing, with a resolution of 640 × 480. Since 3D keypoint annotations are provided, 2D keypoint locations can be easily obtained using the intrinsic parameters of the camera.

EgoDexter [136] includes RGB-D images recorded with a body-mounted camera from egocentric viewpoints of hand interactions with objects in real cluttered scenes, complex hand-object interactions, and natural lighting. It is composed of 4 test sequences, with 3190 frames at 640 × 480 resolution, performed by 4 subjects. EgoDexter is employed for evaluation purposes, as no training set is provided.

Dexter+Object [141] provides six test sequences performed by two subjects, one male and one female. Similarly to EgoDexter, it is only utilized for evaluation purposes. The recordings were conducted by a static camera, containing interactions of a single hand with a cuboid object. This dataset is composed of about 3K samples, with a resolution of 640 × 320 providing RGB images, depth maps, and annotations for fingertips and cuboid corners.

Dexter1 [79] contains 7 sequences of a single subject's right hand performing slow and fast hand motions. Roughly the first 250 frames in each sequence correspond to slow motions, while the rest of them are fast motions. It includes about 2 K frames with a resolution of 320 × 240. As Reference [143] indicated, this dataset suffers from pretty inaccurate calibration parameters and synchronization issues.

SynthHands [136] contains 63.5 K RGB and depth synthetic images from an egocentric viewpoint of two subjects, captured using the Intel RealSense SR300 with 640 × 480 resolution. Real object textures and background images were used, capturing the variations in skin color, shape, background clutter, camera viewpoint, and hand-object interactions. In total, there exist interactions with 7 different virtual objects. Thus, the subject was able to see the rendered scene in real time leading to the above mentioned interactions with objects. As for the annotated data, it provides accurate locations of 21 joints.

4.2. Metrics

The two most common metrics utilized to quantitatively evaluate 3D hand pose estimation methods is mean End-Point-Error (EPE) and Area Under the Curve (AUC) on the Percentage of Correct Keypoints (PCK). Mean EPE is the average 3D Euclidean distance between predicted and ground-truth joints. On the other hand, PCK measures the mean percentage of predicted joint locations that fall within certain error thresholds in comparison to correct poses. Table 2 demonstrates the top results in the most commonly utilized datasets and the employed metric(s), as well as the input data modality, leveraged during training of the method responsible for the state-of-the-art results. It should be noted that, since not all works evaluate their performance using both metrics, Table 2 includes only the reported results on each method, measured in mm as far as mEPE is concerned and within 20–50 mm error threshold for AUC.

Table 2. Results of top 3D hand pose estimation methods in terms of accuracy, categorized by dataset and our proposed taxonomy. Results in mEPE are in mm and for Area Under the Curve (AUC) within 20–50 mm error threshold.

Datasets	Input Modality	Sub-Category	Method	Results
ICVL [41]	Depth-based	2D depth map	Zhou et al. [103]	6.24 mEPE
			Ren et al. [87]	6.26 mEPE
			Xiong et al. [100]	6.46 mEPE
		3D representations	Moon et al. [64]	6.28 mEPE
			Ge et al. [99]	6.30 mEPE
			Ge et al. [50]	6.70 mEPE
NYU [42]	Depth-based	2D depth map	Rad et al. [85]	7.40 mEPE
			Ren et al. [87]	7.78 mEPE
			Xiong et al. [100]	8.61 mepe
		3D representations	Moon et al. [64]	8.42mEPE
			Malik et al. [115]	8.72 mEPE
			Huang et al. [110]	8.9 mEPE
MSRA15 [63]	Depth-based	2D depth map	Ren et al. [87]	7.16 mEPE
			Du et al. [86]	7.20 mEPE
			Wan et al. [97]	7.20 mepe
		3D representations	Moon et al. [64]	7.49 mEPE
			Huang et al. [110]	7.40 mEPE
			Ge et al. [99]	7.70 mEPE
HANDS 2017 [45]	Depth-based	2D depth map	Zhou et al. [84]	5.26 mEPE
			Wu et al. [111]	5.90 mEPE
			Ren et al. [87]	8.33 mEPE
		3D representations	Li et al. [105]	9.82 mepe
			Moon et al. [64]	9.95mEPE
SynHand5M [92]	Depth-based	2D depth map	Malik et al. [93]	4.32 mEPE
			Malik et al. [93]	6.30 mEPE
		3D representations	Malik et al. [115]	3.75 mEPE
			Moon et al. [64]	3.81 mEPE
Dexter+Object [141]	RGB-based	Model-based	Zhang et al. [124]	0.825 AUC
			Boukhayma et al. [125]	25.53 mEPE/0.763 AUC
			Baek et al. [126]	0.610 AUC
		Model-free	Spurr et al. [117]	0.820 AUC
			Iqbal [116]	0.710 AUC
	Multimodal	Unimodal Inference	Zhou et al. [135]	0.948 AUC
RHD [43]	RGB-based	Model-based	He et al. [127]	12.40 mEPE
			Baek et al. [126]	0.926 AUC
			Zhang et al. [124]	0.901 AUC
		Model-free	Iqbal et al. [116]	13.41mEPE/0.940 AUC
			Theodoridis et al. [118]	15.61 mEPE/0.907 AUC
			Spurr et al. [49]	19.73 mEPE/0.849 AUC
	Multimodal	Unimodal inference	Yang et al. [133]	13.14 mEPE/0.943 AUC
			Ge et al. [134]	0.920 AUC
			Cai et al. [129]	0.887 AUC
STB [44]	RGB-based	Model-based	He et al. [127]	3.96 mEPE
			Baek et al. [126]	0.995 AUC
			Zhang et al. [124]	0.995 AUC
		Model-free	Theodoridis et al. [118]	6.93 mEPE/0.997AUC
			Iqbal et al. [116]	0.994 AUC
			Spurr et al. [49]	8.56 mEPE/0.983 AUC
	Multimodal	Unimodal Inference	Ge et al. [134]	6.37 mEPE/0.998 AUC
			Yang et al. [133]	7.05 mEPE/0.996 AUC
			Cai et al. [129]	0.994 AUC
EgoDexter [136]	RGB-based	Model-based	Boukhayma et al. [125]	45.33mEPE/0.674AUC
		Model-free	Iqbal et al. [116]	0.580AUC
	Multimodal	Unimodal Inference	Zhou et al. [135]	0.811 AUC

5. Conclusions and Future Directions

Three-dimensional hand pose estimation is a very popular research field and, thanks to the emergence of deep learning, recent approaches have seen a significant breakthrough. Earlier methods, due to hardware limitations and smaller amount of available training data, required a very careful design, targeted on the specific dataset that they were employed. Recent DNN-based methods are more powerful, efficient, and have better generalization capabilities than their predecessors.

Despite that, there are still some unresolved challenges that create a gap between research and practical 3D hand pose estimation applications. Most of the existing works assume that: (a) the hand is positioned in front of a clean background making it easy to segment, (b) only one hand is considered to be present in a scene, and (c) hand-to-hand or hand-to-object interactions are not well explored. Three-dimensional hand pose detection in the wild is far from perfect, unlike the results achieved in existing datasets. Additionally, even though most of the methods presented in this survey are capable of real-time performance on top-end PCs, applications for mobile devices still fall short, due to their limited processing capabilities.

Moreover, selecting an optimal method for a task at hand depends highly on the requirements and constraints imposed by the task (e.g., cost, accuracy, speed, etc.). For instance, highly accurate 3D hand representation is currently achieved by gloves, although deep networks show promising results. On the other hand, real-time applications or low-specification systems require deep networks with minimum trainable parameters, usually sacrificing accuracy for speed. However, elaborate network architectures with high memory and processing footprint achieve very robust results. Furthermore, the state-of-the-art in the hand pose estimation field changes so rapidly that novel deep networks achieving state-of-the-art results are continuously emerging, rendering older networks obsolete.

Furthermore, existing works have validated that training strategies, such as intermediate supervision or physiological constraints and network designs, such as residual feature aggregation, stacked network architectures, and hand model-based priors, improve the recognition performance. Figure 2a draws a comparison between various DNN architectures in terms of network's performance (Accuracy), number of research papers (Volume of works), scope for improvement based on the performance gains and the volume of works (Future potential), computational complexity during training (Training complexity), and time complexity of inference (Inference speed). As can be observed, the research community has extensively utilized 2D CNN architectures in the past, and now the interest has shifted towards other architectures, such as 3D CNN and generative models. Since the frameworks based on 2D CNNs contain a limited number of parameters compared to the rest of the methods, their inference speed is low, coming at a performance cost. Additionally, they do not introduce, in general, any extra component during training, e.g., discriminator, maintaining a relatively low training complexity.

On the other hand, it is deduced that there is much more room for improvement on works that employ generative networks, such as VAEs and GANs, since they have shown promising results, even though they are not exhaustively explored. The construction of latent spaces, where similar input poses are embedded close to each other, proves that these networks can learn a manifold of feasible hand poses, further boosting 3D hand pose estimation results. Moreover, in order to create humanly interpretable and tunable latent representations, methods producing disentangled latent sub-spaces should be investigated thoroughly. However, the fact that they require more delicate training strategies, especially applied to GANs due to the adversarial training, consists a minor drawback leading to a rather high training complexity. To this end, we believe that there should be an extensive research focus on fully exploiting the representation capabilities of such methods and thus fulfilling their potentials.

Figure 2b demonstrates some attributes for each of the three major proposed categories. Similarly, to Figure 2a, Volume of work, Accuracy, and Future potential are used to give a brief overview of the methods, alongside 3D hand pose datasets on respective modality (Existing datasets) and computational cost of each modality (Hardware requirements). It can be seen that most existing works, as well as datasets, deal with depth data, whereas RGB approaches have bloomed more recently, and multimodal ones are still not thoroughly explored. We argue that these approaches should be explored more intensively, since leveraging all available data is not as straightforward as it seems.

Lastly, more diverse and well-annotated datasets may help to alleviate the assumptions about hand positioning, improving the networks' ability to handle cluttered scenes and irregular poses. The creation of datasets specifically focused on challenging scenarios, such as sign language, is an option; however, high-quality annotated 3D data are not easy to acquire. Synthetic technology is capable of generating an abundance of training data, but networks trained on synthetic data have been observed to under-perform when applied to real data due to the domain gap.

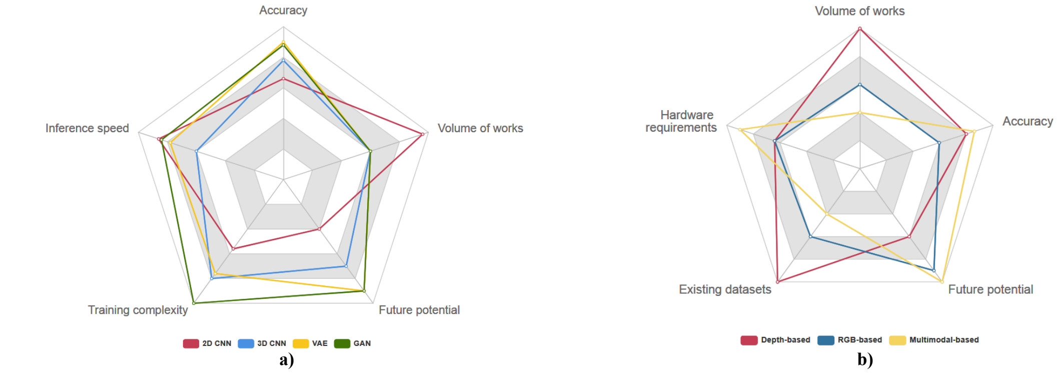

Figure 2. Radar charts showcasing the findings of this survey for (**a**) different Deep Neural Network (DNN) architectures and (**b**) methods based on our proposed taxonomy.

Author Contributions: Conceptualization, T.C., A.S., D.K., K.D., P.D.; Formal analysis, T.C., A.S., D.K., K.D., P.D.; Writing—original draft preparation, T.C., A.S., D.K.; Writing—review and editing, K.D., P.D.; Supervision, K.D., P.D.; Project administration, P.D.; Funding acquisition, P.D. All authors have read and agreed to the published version of the manuscript.

Funding: The work leading to these results received funding from the Greek General Secretariat of Research and Technology under contract T1EΔK-02469 EPIKOINONO.

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

References

1. Piumsomboon, T.; Clark, A.; Billinghurst, M.; Cockburn, A. User-defined gestures for augmented reality. In Proceedings of the 14th IFIP TC 13 International Conference on Human-Computer Interaction, Cape Town, South Africa, 2–6 September 2013; Springer: Berlin, Germany, 2013; pp. 282–299.
2. Lee, T.; Hollerer, T. Multithreaded hybrid feature tracking for markerless augmented reality. *IEEE Trans. Vis. Comput. Graph.* **2009**, *15*, 355–368.
3. Jang, Y.; Noh, S.T.; Chang, H.J.; Kim, T.K.; Woo, W. 3d finger cape: Clicking action and position estimation under self-occlusions in egocentric viewpoint. *IEEE Trans. Vis. Comput. Graph.* **2015**, *21*, 501–510.
4. Kordelas, G.; Agapito, J.P.M.; Hernandez, J.V.; Daras, P. State-of-the-art algorithms for complete 3d model reconstruction. In Proceedings of the Summer School ENGAGE-Immersive and Engaging Interaction with VH on Internet, Zermatt, Switzerland, 13–15 September 2010; Volume 1315, p. 115.
5. Alexiadis, D.S.; Daras, P. Quaternionic signal processing techniques for automatic evaluation of dance performances from MoCap data. *IEEE Trans. Multimed.* **2014**, *16*, 1391–1406.
6. Alivizatou-Barakou, M.; Kitsikidis, A.; Tsalakanidou, F.; Dimitropoulos, K.; Giannis, C.; Nikolopoulos, S.; Al Kork, S.; Denby, B.; Buchman, L.; Adda-Decker, M.; et al. Intangible cultural heritage and new technologies: challenges and opportunities for cultural preservation and development. In *Mixed Reality and Gamification for Cultural Heritage*; Springer: New York, NY, USA, 2017; pp. 129–158.
7. Dimitropoulos, K.; Tsalakanidou, F.; Nikolopoulos, S.; Kompatsiaris, I.; Grammalidis, N.; Manitsaris, S.; Denby, B.; Crevier-Buchman, L.; Dupont, S.; Charisis, V.; et al. A multimodal approach for the safeguarding and transmission of intangible cultural heritage: The case of i-Treasures. *IEEE Intell. Syst.* **2018**, *33*, 3–16.
8. Caggianese, G.; Capece, N.; Erra, U.; Gallo, L.; Rinaldi, M. Freehand-Steering Locomotion Techniques for Immersive Virtual Environments: A Comparative Evaluation. *Int. J. Hum.–Comput. Interact.* **2020**, *36*, 1734–1755.
9. Kopuklu, O.; Kose, N.; Rigoll, G. Motion fused frames: Data level fusion strategy for hand gesture recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2103–2111.

10. Abavisani, M.; Joze, H.R.V.; Patel, V.M. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1165–1174.
11. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*; MIT Press: Montreal, QC, Canada, 2014; pp. 568–576.
12. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
13. Baulig, G.; Gulde, T.; Curio, C. Adapting egocentric visual hand pose estimation towards a robot-controlled exoskeleton. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
14. Papastratis, I.; Dimitropoulos, K.; Konstantinidis, D.; Daras, P. Continuous Sign Language Recognition Through Cross-Modal Alignment of Video and Text Embeddings in a Joint-Latent Space. *IEEE Access* **2020**, *8*, 91170–91180.
15. Koller, O.; Camgoz, C.; Ney, H.; Bowden, R. Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2306–2320.
16. Adaloglou, N.; Chatzis, T.; Papastratis, I.; Stergioulas, A.; Papadopoulos, G.T.; Zacharopoulou, V.; Xydopoulos, G.J.; Atzakas, K.; Papazachariou, D.; Daras, P. A Comprehensive Study on Sign Language Recognition Methods. *arXiv* **2020**, arXiv:2007.12530.
17. Konstantinidis, D.; Dimitropoulos, K.; Daras, P. A deep learning approach for analyzing video and skeletal features in sign language recognition. In Proceedings of the 2018 IEEE International Conference on Imaging Systems and Techniques (IST), Krakow, Poland, 16–18 October 2018; IEEE: New York, NY, USA, 2018; pp. 1–6.
18. Konstantinidis, D.; Dimitropoulos, K.; Daras, P. Sign language recognition based on hand and body skeletal data. In Proceedings of the 2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), Helsinki, Finland, 3–5 June 2018; IEEE: New York, NY, USA, 2018; pp. 1–4.
19. Stefanidis, K.; Konstantinidis, D.; Kalvourtzis, A.; Dimitropoulos, K.; Daras, P. 3D Technologies and Applications in Sign Language. In *Recent Advances in 3D Imaging, Modeling, and Reconstruction*; IGI Global: Hersey, PA, USA, 2020; pp. 50–78.
20. Erol, A.; Bebis, G.; Nicolescu, M.; Boyle, R.D.; Twombly, X. Vision-based hand pose estimation: A review. *Comput. Vis. Image Underst.* **2007**, *108*, 52–73.
21. Joo, H.; Simon, T.; Li, X.; Liu, H.; Tan, L.; Gui, L.; Banerjee, S.; Godisart, T.; Nabbe, B.; Matthews, I.; et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 190–204.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Communications of the ACM: New York, NY, USA, 2012; pp. 1097–1105.
23. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
24. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; Communications of the ACM: New York, NY, USA, 2015; pp. 91–99.
25. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
26. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
27. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]

28. Cao, Z.; Simon, T.; Wei, S.; Sheikh, Y.A. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
29. Xiang, D.; Joo, H.; Sheikh, Y. Monocular total capture: Posing face, body, and hands in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10965–10974.
30. Kendall, A.; Grimes, M.; Cipolla, R. Posenet: A convolutional network for real-time 6-dof camera relocalization. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2938–2946.
31. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks*; MIT Press: Cambridge, MA, USA, 1995; Volume 3361, p. 1995.
32. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
33. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
34. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems, Proceedings of the 2001 Neural Information Processing Systems (NIPS) Conference, Vancouver, BC, Canada, 3–8 December 2001*; A Bradford Book: Cambridge, MA, USA, 2014; pp. 2672–2680.
35. Oberweger, M.; Wohlhart, P.; Lepetit, V. Hands deep in deep learning for hand pose estimation. *arXiv* **2015**, arXiv:1502.06807.
36. Ge, L.; Cai, Y.; Weng, J.; Yuan, J. Hand pointnet: 3d hand pose estimation using point sets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8417–8426.
37. Wang, Y.; Peng, C.; Liu, Y. Mask-pose cascaded cnn for 2d hand pose estimation from single color image. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3258–3268. [CrossRef]
38. Gao, Q.; Liu, J.; Ju, Z. Robust real-time hand detection and localization for space human–robot interaction based on deep learning. *Neurocomputing* **2019**, *390*, 198–206. [CrossRef]
39. Yuan, S.; Garcia-Hernando, G.; Stenger, B.; Moon, G.; Yong Chang, J.; Mu Lee, K.; 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.
40. 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.
41. 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.
42. 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*, 1–10. [CrossRef]
43. Zimmermann, C.; Brox, T. Learning to estimate 3d hand pose from single rgb images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4903–4911.
44. Zhang, J.; Jiao, J.; Chen, M.; Qu, L.; Xu, X.; Yang, Q. A hand pose tracking benchmark from stereo matching. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: New York, NY, USA, 2017; pp. 982–986.
45. Yuan, S.; Ye, Q.; Garcia-Hernando, G.; Kim, T.K. The 2017 hands in the million challenge on 3d hand pose estimation. *arXiv* **2017**, arXiv:1707.02237.
46. Otberdout, N.; Ballihi, L.; Aboutajdine, D. Hand pose estimation based on deep learning depth map for hand gesture recognition. In Proceedings of the 2017 Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 17–19 April 2017; IEEE: New York, NY, USA, 2017; pp. 1–8.

47. 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 February 2016; pp. 87–94.
48. 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]
49. Spurr, A.; Song, J.; Park, S.; Hilliges, O. Cross-modal deep variational 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. 89–98.
50. Ge, L.; Liang, H.; Yuan, J.; Thalmann, D. Real-time 3D hand pose estimation with 3D convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 956–970. [CrossRef]
51. Abdi, M.; Abbasnejad, E.; Lim, C.P.; Nahavandi, S. 3d hand pose estimation using simulation and partial-supervision with a shared latent space. *arXiv* **2018**, arXiv:1807.05380.
52. Baek, S.; In Kim, K.; 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.
53. Liang, H.; Yuan, J.; Thalmann, D.; Zhang, Z. Model-based hand pose estimation via spatial-temporal hand parsing and 3D fingertip localization. *Vis. Comput.* **2013**, *29*, 837–848. [CrossRef]
54. Taylor, J.; Stebbing, R.; Ramakrishna, V.; Keskin, C.; Shotton, J.; Izadi, S.; Hertzmann, A.; Fitzgibbon, A. User-specific hand modeling from monocular depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 644–651.
55. Melax, S.; Keselman, L.; Orsten, S. Dynamics based 3D skeletal hand tracking. In Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, Orlando, FL, USA, 21 March 2013; p. 184.
56. Oikonomidis, I.; Kyriazis, N.; Argyros, A.A. Efficient model-based 3D tracking of hand articulations using Kinect. In Proceedings of the British Machine Vision Conference, Dundee, UK, 29 August–2 September 2011; Volume 1, p. 3.
57. Oikonomidis, I.; Kyriazis, N.; Argyros, A.A. Tracking the articulated motion of two strongly interacting hands. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: New Yorky, NY, USA, 2012; pp. 1862–1869.
58. Oikonomidis, I.; Lourakis, M.I.; Argyros, A.A. Evolutionary quasi-random search for hand articulations tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 3422–3429.
59. Roditakis, K.; Makris, A.; Argyros, A.A. Generative 3D Hand Tracking with Spatially Constrained Pose Sampling. In Proceedings of the British Machine Vision Conference, London, UK, 7–10 September 2017; Volume 1, p. 2.
60. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95-International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; IEEE: New York, NY, USA, 1995; Volume 4, pp. 1942–1948.
61. Tagliasacchi, A.; Schröder, M.; Tkach, A.; Bouaziz, S.; Botsch, M.; Pauly, M. Robust articulated-ICP for real-time hand tracking. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2015; Volume 34, pp. 101–114.
62. Sinha, A.; Choi, C.; Ramani, K. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4150–4158.
63. 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.
64. Moon, G.; Yong Chang, J.; Mu Lee, K. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5079–5088.
65. Wan, C.; Probst, T.; Van Gool, L.; Yao, A. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 680–689.

66. Yang, H.; Zhang, J. Hand pose regression via a classification-guided approach. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Springer: Cham, Switzerland, 2016; pp. 452–466.
67. Ge, L.; Liang, H.; Yuan, J.; Thalmann, D. Robust 3d hand pose estimation in single depth images: From single-view cnn to multi-view cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3593–3601.
68. Keskin, C.; Kıraç, F.; Kara, Y.E.; Akarun, L. Real time hand pose estimation using depth sensors. In *Consumer Depth Cameras for Computer Vision*; Springer: New York, NY, USA, 2013; pp. 119–137.
69. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time human pose recognition in parts from single depth images. In Proceedings of the CVPR 2011, Providence, RI, USA, 20–25 June 2011; IEEE: New York, NY, USA, 2011; pp. 1297–1304.
70. Athitsos, V.; Sclaroff, S. Estimating 3D hand pose from a cluttered image. In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 18–20 June 2003; IEEE: New York, NY, USA, 2003; Volume 2, pp. II–432.
71. Oikonomidis, I.; Kyriazis, N.; Argyros, A.A. Markerless and efficient 26-dof hand pose recovery. In Proceedings of the 10th Asian Conference on Computer Vision, Queenstown, New Zealand, 8–12 November 2010; Springer: Berlin, Germany, 2010; pp. 744–757.
72. Oikonomidis, I.; Kyriazis, N.; Argyros, A.A. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; IEEE: New York, NY, USA, 2011; pp. 2088–2095.
73. Wang, R.Y.; Popović, J. Real-time hand-tracking with a color glove. *ACM Trans. Graph. (TOG)* **2009**, *28*, 1–8.
74. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
75. Keskin, C.; Kıraç, F.; Kara, Y.E.; Akarun, L. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin, Germany, 2012; pp. 852–863.
76. Tang, D.; Yu, T.H.; Kim, T.K. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 3224–3231.
77. Liang, H.; Yuan, J.; Thalmann, D. Parsing the hand in depth images. *IEEE Trans. Multimed.* **2014**, *16*, 1241–1253. [CrossRef]
78. Dollár, P.; Welinder, P.; Perona, P. Cascaded pose regression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; IEEE: New York, NY, USA, 2010; pp. 1078–1085.
79. Sridhar, S.; Oulasvirta, A.; Theobalt, C. Interactive markerless articulated hand motion tracking using RGB and depth data. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2456–2463.
80. Tkach, A.; Pauly, M.; Tagliasacchi, A. Sphere-meshes for real-time hand modeling and tracking. *ACM Trans. Graph. (ToG)* **2016**, *35*, 1–11. [CrossRef]
81. Tzionas, D.; Ballan, L.; Srikantha, A.; Aponte, P.; Pollefeys, M.; Gall, J. Capturing hands in action using discriminative salient points and physics simulation. *Int. J. Comput. Vis.* **2016**, *118*, 172–193. [CrossRef]
82. Romero, J.; Tzionas, D.; Black, M.J. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph. (ToG)* **2017**, *36*, 245. [CrossRef]
83. Oberweger, M.; Lepetit, V. Deepprior++: Improving fast and accurate 3d hand pose estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 585–594.
84. Zhou, Y.; Lu, J.; Du, K.; Lin, X.; Sun, Y.; Ma, X. Hbe: Hand branch ensemble network for real-time 3d hand pose estimation. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 501–516.
85. Rad, M.; Oberweger, M.; Lepetit, V. Feature mapping for learning fast and accurate 3d pose inference from synthetic images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4663–4672.

86. Du, K.; Lin, X.; Sun, Y.; Ma, X. Crossinfonet: Multi-task information sharing based hand pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9896–9905.
87. Ren, P.; Sun, H.; Qi, Q.; Wang, J.; Huang, W. SRN: Stacked Regression Network for Real-time 3D Hand Pose Estimation. In Proceedings of the 30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, 9–12 September 2019; p. 112.
88. Zhou, X.; Wan, Q.; Zhang, W.; Xue, X.; Wei, Y. Model-based Deep Hand Pose Estimation. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016.
89. Šarić, M. LibHand: A Library for Hand Articulation; Version 0.9. Available online: http://www.libhand.org/ (accessed on 31 July 2020).
90. Malik, J.; Elhayek, A.; Stricker, D. Simultaneous hand pose and skeleton bone-lengths estimation from a single depth image. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; IEEE: New York, NY, USA, 2017; pp. 557–565.
91. Malik, J.; Elhayek, A.; Stricker, D. Structure-aware 3d hand pose regression from a single depth image. In Proceedings of the International Conference on Virtual Reality and Augmented Reality, London, UK, 22–23 October 2018; Springer: Cham, Switzerland, 2018; pp. 3–17.
92. Malik, J.; Elhayek, A.; Nunnari, F.; Varanasi, K.; Tamaddon, K.; Heloir, A.; Stricker, D. Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; IEEE: New York, NY, USA, 2018; pp. 110–119.
93. Malik, J.; Elhayek, A.; Stricker, D. WHSP-Net: A Weakly-Supervised Approach for 3D Hand Shape and Pose Recovery from a Single Depth Image. *Sensors* **2019**, *19*, 3784. [CrossRef]
94. Wan, C.; Probst, T.; Gool, L.V.; Yao, A. Self-supervised 3d hand pose estimation through training by fitting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10853–10862.
95. Oberweger, M.; Wohlhart, P.; Lepetit, V. Training a feedback loop for hand pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 Decemebr 2015; pp. 3316–3324.
96. Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 700–708.
97. Wan, C.; Probst, T.; Van Gool, L.; Yao, A. Dense 3d regression for 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. 5147–5156.
98. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 483–499.
99. Ge, L.; Ren, Z.; Yuan, J. Point-to-point regression pointnet for 3d hand pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 475–491.
100. Xiong, F.; Zhang, B.; Xiao, Y.; Cao, Z.; Yu, T.; Zhou, J.T.; Yuan, J. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 793–802.
101. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
102. Chen, X.; Wang, G.; Zhang, C.; Kim, T.K.; Ji, X. Shpr-net: Deep semantic hand pose regression from point clouds. *IEEE Access* **2018**, *6*, 43425–43439. [CrossRef]
103. Zhu, T.; Sun, Y.; Ma, X.; Lin, X. Hand Pose Ensemble Learning Based on Grouping Features of Hand Point Sets. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
104. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5099–5108.

105. Li, S.; Lee, D. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11927–11936.
106. Ravanbakhsh, S.; Schneider, J.; Póczos, B. Deep Learning with Sets and Point Clouds. *arXiv* **2017**, arXiv:1611.04500.
107. Chen, Y.; Tu, Z.; Ge, L.; Zhang, D.; Chen, R.; Yuan, J. So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning. In Proceedings of the IEEE International Conference on Computer Vision, Long Beach, CA, USA, 16–20 June 2019; pp. 6961–6970.
108. Li, J.; Chen, B.M.; Hee Lee, G. So-net: Self-organizing network for point cloud analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9397–9406.
109. Kohonen, T.; Honkela, T. Kohonen network. *Scholarpedia* **2007**, *2*, 1568.
110. Huang, F.; Zeng, A.; Liu, M.; Qin, J.; Xu, Q. Structure-aware 3d hourglass network for hand pose estimation from single depth image. *arXiv* **2018**, arXiv:1812.10320.
111. Wu, X.; Finnegan, D.; O'Neill, E.; Yang, Y.L. Handmap: Robust hand pose estimation via intermediate dense guidance map supervision. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 237–253.
112. 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.
113. Song, S.; Xiao, J. Deep sliding shapes for amodal 3d object detection in rgb-d images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 808–816.
114. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland: 2015; pp. 234–241.
115. Malik, J.; Abdelaziz, I.; Elhayek, A.; Shimada, S.; Ali, S.A.; Golyanik, V.; Theobalt, C.; Stricker, D. HandVoxNet: Deep Voxel-Based Network for 3D Hand Shape and Pose Estimation from a Single Depth Map. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020; pp. 7113–7122.
116. Iqbal, U.; Molchanov, P.; Breuel Juergen Gall, T.; Kautz, J. Hand pose estimation via latent 2.5 d heatmap regression. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 118–134.
117. Spurr, A.; Iqbal, U.; Molchanov, P.; Hilliges, O.; Kautz, J. Weakly Supervised 3D Hand Pose Estimation via Biomechanical Constraints. *arXiv* **2020**, arXiv:2003.09282.
118. Theodoridis, T.; Chatzis, T.; Solachidis, V.; Dimitropoulos, K.; Daras, P. Cross-Modal Variational Alignment of Latent Spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 960–961.
119. Yang, L.; Yao, A. Disentangling latent hands for image synthesis and pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9877–9886.
120. Gu, J.; Wang, Z.; Ouyang, W.; Zhang, W.; Li, J.; Zhuo, L. 3D Hand Pose Estimation with Disentangled Cross-Modal Latent Space. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 391–400.
121. Panteleris, P.; Oikonomidis, I.; Argyros, A. Using a single rgb frame for real time 3d hand pose estimation in the wild. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; IEEE: New York, NY, USA, 2018; pp. 436–445.
122. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
123. Mueller, F.; Bernard, F.; Sotnychenko, O.; Mehta, D.; Sridhar, S.; Casas, D.; Theobalt, C. Ganerated 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.

124. Zhang, X.; Li, Q.; Mo, H.; Zhang, W.; Zheng, W. End-to-end hand mesh recovery from a monocular rgb image. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 2354–2364.
125. Boukhayma, A.; Bem, R.D.; Torr, P.H. 3d hand shape and pose from images in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10843–10852.
126. Baek, S.; Kim, K.I.; Kim, T.K. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1067–1076.
127. He, Y.; Hu, W.; Yang, S.; Qu, X.; Wan, P.; Guo, Z. 3D Hand Pose Estimation in the Wild via Graph Refinement under Adversarial Learning. *arXiv* **2019**, arXiv:1811.07376.
128. Yuan, S.; Stenger, B.; Kim, T. 3D Hand Pose Estimation from RGB Using Privileged Learning with Depth Data. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 2866–2873.
129. Cai, Y.; Ge, L.; Cai, J.; Yuan, J. Weakly-supervised 3d hand pose estimation from monocular rgb images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 666–682.
130. Dibra, E.; Melchior, S.; Balkis, A.; Wolf, T.; Oztireli, C.; Gross, M. Monocular RGB hand pose inference from unsupervised refinable nets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1075–1085.
131. Chen, L.; Lin, S.Y.; Xie, Y.; Lin, Y.Y.; Fan, W.; Xie, X. DGGAN: Depth-image Guided Generative Adversarial Networks forDisentangling RGB and Depth Images in 3D Hand Pose Estimation. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 411–419.
132. Zhao, L.; Peng, X.; Chen, Y.; Kapadia, M.; Metaxas, D.N. Knowledge as Priors: Cross-Modal Knowledge Generalization for Datasets without Superior Knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6528–6537.
133. Yang, L.; Li, S.; Lee, D.; Yao, A. Aligning latent spaces for 3d hand pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 2335–2343.
134. Ge, L.; Ren, Z.; Li, Y.; Xue, Z.; Wang, Y.; Cai, J.; Yuan, J. 3d hand shape and pose estimation from a single rgb image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10833–10842.
135. Zhou, Y.; Habermann, M.; Xu, W.; Habibie, I.; Theobalt, C.; Xu, F. Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020; pp. 5346–5355.
136. Mueller, F.; Mehta, D.; Sotnychenko, O.; Sridhar, S.; Casas, D.; Theobalt, C. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 1284–1293.
137. Kazakos, E.; Nikou, C.; Kakadiaris, I.A. On the Fusion of RGB and Depth Information for Hand Pose Estimation. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; IEEE: New York, NY, USA, 2018; pp. 868–872.
138. Wetzler, A.; Slossberg, R.; Kimmel, R. Rule of thumb: Deep derotation for improved fingertip detection. *arXiv* **2015**, arXiv:1507.05726.
139. Yuan, S.; Ye, Q.; Stenger, B.; Jain, S.; Kim, T.K. Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4866–4874.
140. 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 International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 813–822.
141. Sridhar, S.; Mueller, F.; Zollhöfer, M.; Casas, D.; Oulasvirta, A.; Theobalt, C. Real-time joint tracking of a hand manipulating an object from rgb-d input. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 294–310.

142. Garcia-Hernando, G.; Yuan, S.; Baek, S.; Kim, T.K. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018, pp. 409–419.

143. Simon, T.; Joo, H.; Matthews, I.; Sheikh, Y. Hand keypoint detection in single images using multiview bootstrapping. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1145–1153.