Deep learning techniques for single person human pose estimation

Andrea Alfieri
TU Delft
a.alfieri-1@student.tudelft.nl

Abstract

Solving the human pose estimation (HPE) task with deep neural network solutions is an arduous problem which requires an accurate design of the systems, and still often needs a copious amount of computational power. With this work, I try to provide a report in which the ideas found in the most important papers of the field are discussed and compared. My analysis involves publications on *single-person* HPE from many years ago as well as the most recent state-of-the-art, and shows how each solution built upon previous work to improve its quality. Moreover, this survey also contains an extensive overview on the most common datasets and evaluation metrics used to measure the performances of these systems.

ACM Reference Format:
Andrea Alfieri. 2020. Deep learning techniques for single person human pose estimation. In *TU Delft: Literature Survey, November, 2020, Delft, NL.* ACM, New York, NY, USA, 12 pages. https://doi.org/00.0000/0000000.0000000

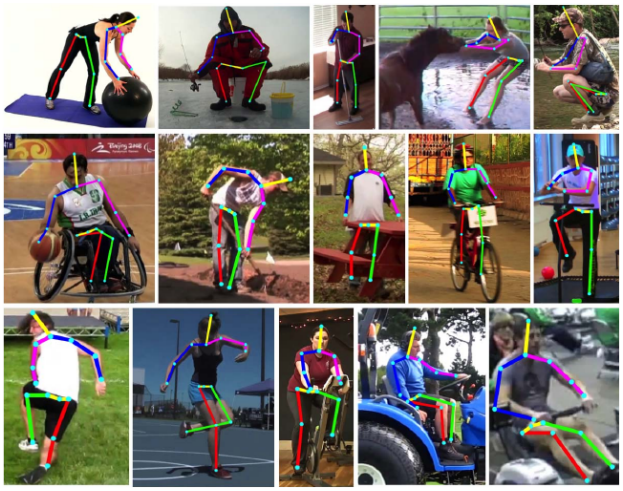

Figure 1. This figure represents an example of what single-person human pose estimation systems are able to produce. In this case, 15 different joints are detected for each image, and the system is able to distinguish between right and left limbs for a variety of different poses. [4]

1 Introduction

Human pose estimation is defined as the task of estimating the pose of one or more people in an image or video. Its goal is usually to detect where the body joints of each person are, as shown in the example of figure 1. In this case, only one person is pictured in each image and only 15 key points are detected, but many papers describe systems which are able to provide much more detailed outputs, such as a distinction between the right and left eye or between the wrist and the hand for multiple subjects in the same image.

Achieving good results on HPE and accurately understanding a person's posture and limb articulation can be extremely helpful for higher level tasks such as activity recognition, video surveillance and self-driving [9, 39]. However, HPE techniques also present some challenges that can heavily undermine the performances of the system. Occlusion, for example, happens when other people or parts of the environment cover joints that should otherwise be visible. The same concept applies when parts of the body cover other parts of the same person with respect to the camera, which is defined as self-occlusion. Different networks and novel techniques [27, 45] have been tested to solve the occlusion problem, but it still represents today one of the main factors that affect the power of the HPE systems. Pytel et al. [44] provide an interesting study that shows how much occlusion can influence the performance. Moreover, the same joint can appear differently in each image. For example, a good dataset for HPE should include an extended set of diverse clothing to allow the systems to generalize well.

Research on HPE has been conducted for quite a long time, but deep learning techniques have only been tested for less than a decade. This probably happened because it is around that time that computers and GPUs started to become more efficient at running systems which contained millions of floating point parameters. The main goal of this work is to provide the reader with a thorough overview of the main research papers related to this subject from the

TU Delft, November, 2020, Delft, NL

ACM ISBN XXX-X-XXXX-XXXX-X/00/00…$1.00
https://doi.org/00.0000/0000000.0000000

deep learning perspective, starting from the first attempts at moving from classical approaches to deep learning solutions around 2014, to the most recent architectures providing the current state of the art performances. Moreover, this survey also describes the most common evaluation metrics used to assess the performance of the different architectures and analyzes a set of different datasets that are more popular in the field.

Other recent surveys have been consulted, mainly Chen et al. [9] and Dang et al. [14], to allow me to understand what kind of information about the papers that are being discussed is important to include in a survey. Moreover, these surveys have been a valuable source for an initial set of papers to read and to find out which works are currently considered as state of the art by the community. However, while other surveys are usually just a large list of citations followed by one or two descriptive sentences, the objective of my work is to dig deeper into the details of each research that is being analyzed and provide the reader with enough information to fully understand the strengths and weaknesses of each work and their main differences.

The focus of my survey is **2D single-person** human pose estimation, which means that the goal of systems described in the following sections is to output the most accurate pose on a 2D plane when fed with an input image representing a single subject. A different and more complex task is *multi-person pose estimation*, where the input image pictures a group of subjects in different poses. In this case, the two main lines followed by the research community are the top-down and bottom-up approaches. In the bottom-up approach [6, 21, 28, 30], all body joints are detected in the image and are then grouped through various algorithms. On the other hand, the top-down approach [10, 19, 38] first draws bounding boxes around each person and then conducts pose estimation on each subject. *3D pose estimation* is a different task which tries to output a 3D model of the human body from a 2D image. The same distinction between single-person [3, 41, 47, 58] and multi-person [36, 37, 48] pose estimation can be applied here as well. An example of 3D HPE is shown in figure 2.

I decided to focus on single person HPE because it is still an incredibly active research field and a task which can be solved with simple to extremely complex solutions. Moreover, multi-person HPE systems often make use of the same techniques described here, which are occasionally also relevant for 3D HPE and other tasks such as object detection or action recognition.

Following the choice adopted by many authors [40, 51] to divide single-person HPE into two main branches, this work will also follow the same partitioning and first describe the most common *regression-based* methods, followed by an analysis of the *detection-based* methods.

2 Regression vs Detection

All existing approaches to human pose estimation can be assigned to two main categories: regression based and detection based. **Regression-based** approaches attempt at defining a direct mapping from the input image to the *coordinates* of the body joints. Since this is a highly non-linear problem [9], the direct approach is considered by many to be the most difficult one, and generally provides worse performances than detection because of its lack of robustness, unless some gimmicks are included in the implementation. However, these tricks usually generate systems than can no longer be strictly categorized as regression-based or detection-based, and are usually defined as hybrid solutions.

On the other hand, **detection-based** approaches attempt at detecting the different body parts of the human body, with the two most common output representations being *image patches* and *heatmaps* of the joint locations. Therefore, this distinction is purely based on the type of output of the network, which then affects how the rest of it is designed.

2.1 Deep Convolutional Networks - AlexNet

We are going to begin our discussion with the paper published by Krizhevsky et al. [31] in 2012. Despite this work not being directly related to human pose estimation, my personal belief is for it to hold great importance for the research topic we are analyzing. In fact, the authors are able to prove how CNNs (i.e. Convolutional neural networks) can be extremely powerful on computer vision tasks, by developing a simple network which achieved significantly better results than the previous state of the art for object detection and classification on the ImageNet dataset[15]. ImageNet is a large collection of annotated images commonly used for many computer vision tasks, whose annotations were collected with the use of Amazon Mechanical Turk, and regularly used to compare different systems and their performances.

The neural network architecture consisted of 5 convolutional layers followed by max-pooling layers, and 3 fully connected layers, with a final 1000-way softmax defining the output of the system. Moreover, the authors included and tested the "dropout" technique, a regularization method which was introduced shortly before, and were able to show how it can greatly reduce overfitting in the fully connected layers.

I believe this paper to be an important milestone in the research on human pose estimation because different authors, such as Pfister et al. [43] and Toshev and Szegedy [56], defined their networks as an adaptation of the one proposed by Krizhevsky et al. [31], often by only modifying the output representation. This architecture is also known as AlexNet.

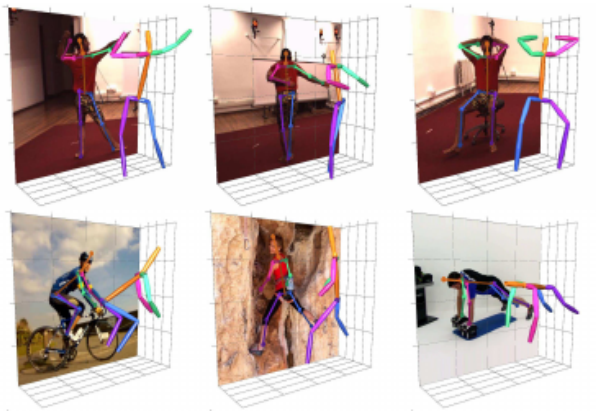

Figure 2. Sun et al. [51] used a bone representation to encode long distance relations between joints, a solution which is applicable to both 2D and 3D HPE. In this image, the most accurate predictions of the network for 3D HPE are pictured.

3 Regression based approaches

My discussion will now turn towards the regression based approach, which attempts at directly mapping the input image to the output joint coordinates.

3.1 Direct regression

The first important attempt at a full deep learning based solution came in 2014 with Toshev and Szegedy [56] and their **DeepPose** architecture. This work was the first one to implement a modified version of the AlexNet architecture as a baseline for the network, which was used to try to solve HPE as a regression problem towards body joints.

This modified implementation also involved 5 convolutional layers and 3 fully connected ones but also used a *cascade of regressors* which would first take the original image as input and then progressively refine the output by using higher resolution sub-images. By fixing the input size at 220x220, the original image would first be submitted at a much lower resolution, while the following stages of the cascade would focus only on a higher detailed crop centered on the previous prediction. Each stage followed the same network architecture but learned different parameters.

The prediction refinement technique is still applied by many recent state-of-the-art networks and is what allowed the system designed by Toshev and Szegedy [56] to have a manageable number of parameters (40 million), even though their number remains quite large if compared to solutions that came afterwards. According to the authors, the biggest advantage of the use of a full deep neural network architecture lies in the fact that it does not required an explicit body model to work, since this is implicitly learned during training.

3.2 Bone representation

Although what Toshev and Szegedy [56] created proved how the regression-based approaches can provide good results if designed correctly, they could not keep up with the results that detection-based was achieving for a few years. It was only in 2017, with the paper we are going to discuss now and published by Sun et al. [51], that regression regained the interest of the public.

According to the authors, «a central problem is that [regression based approaches] simply minimize the per-joint location errors independently but ignore the internal structures of the pose. In other words, joint dependence is not well exploited». Therefore, the designed solution is a structure-aware one which uses **bones** instead of joints to represent the pose of the subject. This approach, defined by the authors as *compositional pose regression*, finds in bones a more primitive, stable and easier to learn representation that simultaneously encodes long distance interactions between the bones. The reason why this paper was able to improve the previous state of the art results is because the preceding solutions were based on minimizing each per-joint location error independently from the others.

This is the novelty that this paper brought to the community: by only re-parametrizing the pose representation, which is the network output, and the loss function, the authors are able to bring an improvement to the previous state of the art results with a method that can also be applied without any change to 3D pose estimation, as shown in figure 2. For the first time, they have proven that directly mixed learning is effective.

Moreover, I genuinely appreciated how the authors validated their work through a comprehensive evaluation which included new metrics, an interesting ablation study, and a thorough comparison with state of the art results on both 2D and 3D benchmarks.

3.3 Multi-task learning

The biggest step forward for the regression based methods came, in my opinion, in 2017 with Luvizon et al. [35], which focused on the differentiability of the pipeline and used the *soft-argmax* function to convert feature maps directly to body joint coordinates, resulting in a fully differentiable framework. This final step was usually done with the use of the *argmax* function, which is not differentiable and was «breaking the learning chain of neural networks»[35].

In 2018, with Luvizon et al. [34], the same authors released a *multi-task* framework for jointly 2D and 3D pose estimation pictured in figure 3, which also made use of the soft-argmax function. Multi-task learning attempts at solving multiple tasks at the same time and with the same network, so that their differences and commonalities can be exploited for a more accurate solution. In fact, the best advantage of deep

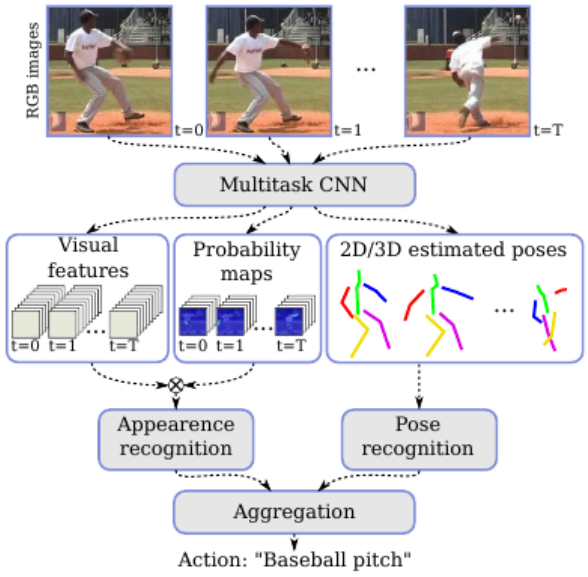

Figure 3. This image represents the *multi-task* framework designed by Luvizon et al. [34]. A single network takes T frames of a video as input and outputs the pose of the subject and an action category. The two tasks are able to influence each other for a greater performance.

learning is considered by many to be «its capability to perform end-to-end optimization» [34] and, as suggested by Kokkinos [29], this is even more true for multitask problems, where related tasks can benefit from one another.

As baseline network for the HPE task, the authors used the Inception-V4 architecture [52] followed by multiple prediction blocks that would iteratively refine the estimations by looking at three different image resolutions. The soft-argmax function then allowed them to stack the action recognition pipeline on top of pose estimation, which generated a multitask network that could be trained from end-to-end. This solution provided performances which are quite close to those achieved by the state of the art of the detection based approach, with a PCKh@0.5 score of 91.2 on the MPII dataset.

After 2018, the single-person 2D HPE research almost uniquely focused on the detection based approaches, but papers like Pavllo et al. [42] (2019) or Ci et al. [13] (2020) are still applying regression-based techniques for monocular 3D human pose estimation.

4 Detection based approaches

In contrast with regression-based approaches, detection-based ones attempt at detecting the different parts of the human body, which are usually represented through either a heatmap or multiple image patches. Different techniques can then be applied to determine the correct connection of the different detected parts.

4.1 First deep learning approach

In the case of detection-based, the first important step towards a full deep learning solution came in 2014 with Tompson et al. [55]. With this paper, the authors presented a hybrid solution for HPE which consists of two parts: a deep Convolutional network and a Markov Random Field. A **Markov Random Field** (**MRF**) can be considered similar to a Bayesian network, but is undirected and may also be acyclic. They have been used to solve many computer vision problems by posing them as energy minimization problems. [32]

The first stage of the system is the CNN, which is pictured in figure 4. The goal of this network is to detect the different parts of the body through a standard and quite straightforward architecture. The input of the network is the RGB image containing the subject and the output is a heat-map which «produces a per-pixel likelihood for key joint locations on the human skeleton» [55]. The singularity of this network is the use of *multi-resolution*, where two different sizes of sliding windows as used at the same time. These are pictured as the blue and the orange windows in Fig. 4.

The novelty brought by the work of Tompson et al. [55] lies in the second stage of the system, the Markov random field. In fact, simple body part detection which only makes use of local appearance is highly sensitive to body occlusion and complex backgrounds, and is therefore usually not sufficient to achieve good performances. This second part is defined as higher level *Spatial-model* formulated as a MRF-like model which is used to «constrain the joint inter-connectivity and enforce global pose consistency» [55].

As stated by the authors of the paper, the main expectation of this stage is not to increase the performance of detections that are already satisfactory, but to «remove the false positive outliers that are anatomically incorrect» [55]. Moreover, differently from other previous works [23], the Spatial-model also contains parameters which are learned, therefore creating a fully trainable solution.

4.2 Stacked Hourglass

A big step forward for the entire deep learning community came in 2016 with Newell et al. [39], which is one of the most cited papers this survey is analyzing. After reading about the research carried out by others, such as the previously discussed Tompson et al. [55] and Carreira et al. [7], the authors built upon their work and designed the well known **Stacked Hourglass Network**, an architecture capable of processing features across all scales through «successive steps of pooling and upsampling that are done to produce a final set of predictions»[39].

As shown in figure 5, a single hourglass module pools down the input image to a very low resolution of 4x4 pixels through multiple convolutional and max pooling layers, and successively begins the sequence of upsampling and the

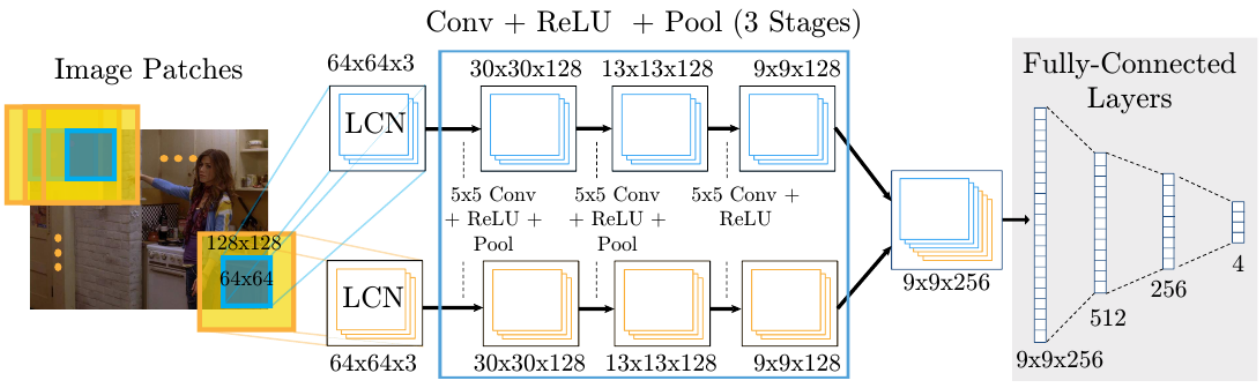

Figure 4. The first stage of the network designed by Tompson et al. [55]. The authors use two different window sizes to analyze the input image at different scales, which are denoted in blue and orange in this figure.

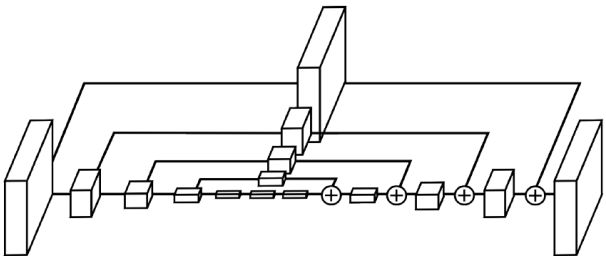

Figure 5. A single *Hourglass* module. Note how the resolution is progressively reduced to a small size and then upscaled again to the original resolution. [39]

combination of features across scales. While the system designed by Tompson et al. [55] was running different pipelines in parallel to process the same image at only two different resolutions, which were then combined for the final prediction, Newell et al. [39] use a single pipeline with skip layers and combines the information of two adjacent resolutions through «nearest neighbor upsampling of the lower resolution followed by an elementwise addition of the two sets of features»[39].

Moreover, after Carreira et al. [7] provided a hint of the potential of multiple iterative stages and intermediate supervision, Newell et al. [39] decided to stack multiple hourglass modules end-to-end by feeding the input of the previous module directly into the following one. This technique gives the network a procedure for repeated bottom-up and top-down inference which allows for a reevaluation of the initial estimates, and they proved how this method, in conjunction with intermediate supervision, is critical to improve the performance of the network. In the final design eight modules are used, which do not share the weights.

This solution provided state of the art results on human pose estimation for their time, but most importantly defined a new powerful architecture which other researchers could build upon.

4.3 Multi-context attention

The first important improvement on the Stacked Hourglass architecture proposed by Newell et al. [39] came only one year later with Chu et al. [12], which were able to improve the performance of the Stacked Hourglass network by replacing different modules with more advanced ones.

The idea was to design an architecture which could combine multiple contextual information in a more precise manner, a technique which had been proved effective on many computer vision tasks such as pedestrian detection [60].

In particular, Chu et al. [12] adopted the stacked hourglass architecture to generate different attention maps *at multiple resolutions* focusing on the human body, which could help recover occluded body parts and distinguish the body from cluttered backgrounds. First, the authors designed a novel attention model based on **Conditional Random Fields** (CRFs) which replaced the widely used spacial Softmax normalization. This solution showed important improvements because of the ability of CRFs to model the correlation among neighbouring regions of the attention map, which is a crucial task for detecting the correct pose of the human body in the input image.

Moreover, the residual unit of the original architecture was replaced by a novel "*Hourglass Residual Unit*", which was able to increase the receptive field of the network. «With this architecture, [they enrich] the information received by the output of each building block, which makes the whole framework more robust to scale change.»[12]

In conclusion, I consider this paper to be an important one because of the ability of the authors to thoroughly comprehend the Stacked Hourglass architecture and improve it to provide performances which today are still close to those achieved by state of the art solutions.

4.4 Hourglass for the state of the art

Finally, we turn our discussion towards two papers based on the Hourglass architecture which can be currently considered as state of the art for the single person HPE task: Tang et al. [54] (2018) and Tang and Wu [53] (2019), which respectively achieved a PCKh@0.5 score on the MPII dataset of 92.3 and 92.7.

The novelty brought by Tang et al. [54] lies in the use of *compositional models*, a representation of patterns with hierarchies of meaningful parts and subparts. In particular, compositional models are able to exploit the compositionality of the human body by representing it as a hierarchy which satisfies some articulation constraints.

With this paper, the authors are able to exploit CNNs, which had been used by the vast majority of papers on HPE, to **learn** the compositional model of human bodies and to resolve low-level ambiguities in high-level pose predictions. Specifically, this model is able to improve the performance of the system because it implicitly learns the relationship among the different body parts in a new and more effective way. This helps in detecting body parts that are hard to detect because of overlapping parts, other people in the same image or a cluttered background. Therefore, it is once more clear how previous state of the art performances can only be improved by directly addressing the weaknesses of the best architectures in the most detailed way. In this case, occlusion.

Moreover, the authors propose a novel part representation similar to the one previously discussed and introduced by Sun et al. [51], which is based on *bones* instead of joints.

On the other hand, Tang and Wu [53] try to create a network which can further comprehend how different body parts are related to each other, by proposing a *part-based branching network* (PBN) to learn specific representations to each part group, instead of predicting all joint heatmaps from one single branch.

4.5 Improving the computational cost

After the achievement by Tang and Wu [53] of such performance, research started focusing more on how to make these networks run on systems with a constrained computational power, such as a small laptop or a mobile phone. Zhang et al. [61] in 2019, tried to design a lightweight Hourglass network by examining the «redundancy degree of existing state-of-the-art pose CNN architecture designs». Moreover, they applied the concept of *knowledge distillation*, where they attempted at transferring knowledge from a pre-trained larger teacher network to a tiny target pose network, which is deployed at test time. They called their solution FastPose.

Isack et al. [22] in 2020, also achieve state of the art performances with a more efficient and lightweight model called RePose, by explicitly incorporating geometric priors about the human body into the framework and by applying the prediction refinement technique which was also used by different papers we discussed before.

While Tang and Wu [53] was using more than 25 million parameters in its implementation, Zhang et al. [61] and Isack et al. [22] only used 3 million and 4 million respectively. Their PCKh@0.5 scores on the MPII dataset were 91.1 and 90.29 respectively, which are comparable to the performance of the original Stacked Hourglass architecture presented by Newell et al. [39], but in a much lighter fashion. Other interesting papers which try to increase the time efficiency of HPE are Hwang et al. [20], Jiang et al. [24] and Zhang et al. [61].

5 GANs

First introduced by Goodfellow et al. [18] in 2014, GANs gained great interest from the public in the past few years because of their ability to allow unsupervised training of generative model, and have performed well on generating natural images such as human faces and indoor scenes. While finally being able to avoid the blur effect of variational autoencoders, they have also been considered unstable and difficult to train in their early life. However, this problem became smaller and smaller as researchers introduced more complex networks. In particular, «two neural networks [i.e. a **generator** and a **discriminator**] contest with each other in a game (in the form of a zero-sum game, where one agent's gain is another agent's loss). [...] The core idea of a GANs is based on the "indirect" training through the discriminator, which itself is also being updated dynamically»[1]

Because of the success of GANs, people also attempted at using these architectures in the field of supervised learning. This is the case of the papers we are going to discuss next.

5.1 Self adversarial training for HPE

After the increase of interest towards GANs and the success of the Stacked Hourglass architecture [39], Chou et al. [11] attempted at employing a generative adversarial network as their learning paradigm for human pose estimation.

In particular, the generator and the discriminator were designed as two networks with the same architecture based on the Stacked Hourglass. While the generator was designed to map each input image to keypoint heatmaps, the objective of the discriminator was to distinguish real heatmaps from generated ones. The framework of this adversarial network is shown in figure 6.

In theory, if the training phase is successful, the generator should finally be able to produce heatmaps that the discriminator in no longer capable of distinguishing from real ones. Therefore, the discriminator is discarded and the generator is used as the predictor for the testing phase.

The idea behind the design proposed by Chou et al. [11] comes from the fact that HPE can become a tricky task when occlusion and fuzzy backgrounds are present, a problem

[1] https://en.wikipedia.org/wiki/Generative_adversarial_network

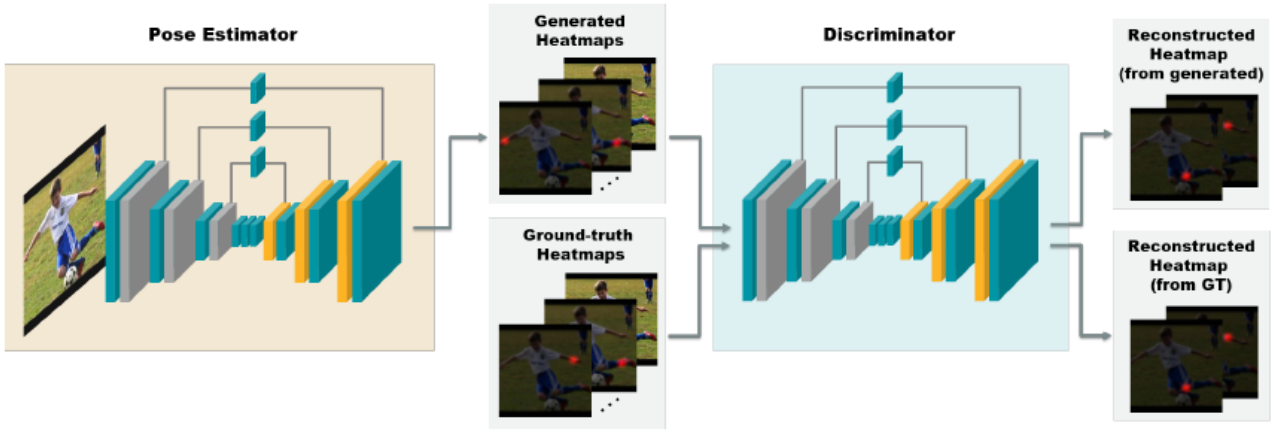

Figure 6. GAN architecture for Human Pose Estimation, as proposed by Chou et al. [11]. The discriminator is fed either with a generated heatmap or a ground-truth one but is not told which one is which. Through backpropagation, both networks are able to learn at the same time, even if they have different goals.

which many addressed by including a human body model in their architecture. In this solution, this can be considered the advantage of including a discriminator: that it should be able to distinguish implausible poses, which implicitly makes the generator learn a human body model that can be used in parallel with regular detection by the network.

5.2 Adding more discriminators

A similar solution to the one proposed by Chou et al. [11] was introduced by Chen et al. [8] in 2017. Again, the training style of GANs is employed, but this time with an important difference: the generator is not only used to determine the pose of the human bodies, but simultaneously outputs an occlusion heatmap, which are both then fed to the discriminator.

Moreover, two discriminators are used. The first one is the standard one which is simply employed for classification of real vs fake pose, while the second one has to make a decision on the confidence of the predicted pose heatmaps. This architecture was able to achieve results close to the state of the art with a PCKh score of 91.9.

5.3 Self attention and Multiple-Instance Learning

Traditional GANs are great at generating high-resolution details as a function of only spatially local points in lower-resolution feature maps. [62] With the advent of **Self Attention GANs** (SAGAN), proposed by Zhang et al. [62], GANs became able to also model attention-driven, long-range dependencies for the task of image generation. In fact, the self-attention module is able is calculate the «response at a position as a weighted sum of the features at all positions, where the weights - or attention vectors - are calculated with only a small computational cost»[62]. Even though a similar effect could also be achieved by increasing the size of the convolutional kernels, that would also generate a loss in the computational and statistical efficiency of the network.

While GANs had been proven efficient for the HPE task too, they were also able to only learn local body joints structural constraints. In 2019, Wang et al. [57] used self-attention to design a network which could finally learn long-range body joint dependencies and generate more consistent, realistic and accurate body structures, struggling less against occlusion and crowded backgrounds. Both the network generator and discriminator used a Hourglass architecture as their backbone with self-attention architectures. The generator used a 4-stack hourglass network where the attention modules were integrated in the last three. This solution was able to achieve a PCKh@0.5 score of 92.3 on the MPII dataset, a result which comes close to the current SOTA in HPE.

On the other hand, Shamsolmoali et al. [50] in 2020 used two residual Multiple-instance learning (MIL) models with identical architectures as generator and discriminator of their system. **Multiple-instance learning** is a technique that can be applied to supervised learning where the learner is fed with a labeled set of instances, instead of individually labeled ones, called *bag*. For example, in the case of binary classification, the bag could be labeled as negative if all instances are negative and positive if at least one instance is positive. Then the learner is required to label either single instances or entire bags correctly. Shamsolmoali et al. [50] decided to use this technique because of how it previously improved the accuracy for the action recognition task [1]. In their case, the input bag contains different poses for the same image. The different poses are generated by adding noise to the ground truth pose. The bag is then labeled with the ground truth pose. Shamsolmoali et al. [50] also achieved a PCKh@0.5

score of 92.3 on the MPII dataset, a close performance to the state-of-the-art.

In conclusion, my opinion is for GANs to be a solution which has been interesting the community because of their ability to automatically focus on the most difficult aspects of the task they are given through the competition of the discriminator and the generator. However, for now they are only able to reach state of the art performances at the cost of being still quite difficult to design and train correctly, meaning that a few more years of research are probably necessary for their use in supervised learning to be comparable to more standard solutions.

6 Common Evaluation Metrics

To compare the performances of different systems, one would need two things: a shared evaluation metric and a shared set of images to test the networks on. While the second aspect will be covered in section 7, I will use this section to list and describe the most common metrics used in the field.

6.1 Percentage of Correct Parts

PCP is probably the most simple metric one could think of, and has been used by many researchers for quite a long time. Ferrari et al. [17], for example, uses it in 2008. Its goal is to measure the percentage of limbs that are correctly detected. After defining a threshold, «a limb is correctly localized if its two endpoints are within the threshold from the corresponding ground truth endpoints»[9]. Many papers often provide the mean PCP of all limbs as well as the single PCP of more important parts such as the torso or the head.

6.2 Percentage of Correct Keypoints

Similarly to PCP, PCK detects the number of correctly detected joints. Again, a joint is considered correct if it is found within a threshold distance from the ground truth. The threshold can be defined in different ways: Yang and Ramanan [59] determine it as a fraction of the person bounding box size while Andriluka et al. [2] use 50% of the head length, which is also called PCKh@0.5 and is used in Table 1 to compare the different systems discussed in this survey.

6.3 Average Precision

Once we have a way of determining if a prediction is correct or not, therefore being able to count the number true and false positives (TP/TN) and true and false negatives (TN/FN), a further analysis can be accomplished through the use of Average Precision, which is widely use in the object detection task. Precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

When considering each detection one by one, we are able to generate a precision-recall curve $p(r)$. The Average Precision is defined as the area under such curve or

$$AP = \int_0^1 p(r)dr \tag{3}$$

which is always $0 \leq AP \leq 1$ since precision and recall are also within 0 and 1. Finally, the mAP (mean average precision) is defined as the average precision of each body joint for the HPE task.

6.4 FPS and Giga FLOPs

Since the final goal of HPE is to have the most accurate prediction in the least amount of time, the speed of the network must also be assessed through an evaluation metric. The most used ones are FPS and GFLOPs. FPS, or frames per second are used to measure the processing speed of the input images and show how many inputs can be processed in a given time frame. Giga FLOPs, on the other hand, are a measure of the network itself and show how computationally heavy it is based on the number of weights and how difficult it's operations are for a GPU.

7 Datasets

Similarly to many other computer vision tasks, datasets hold great importance for training and testing networks designed for HPE. For human pose estimation, having a large set of labeled images is crucial if we want our model to output realistic poses. Moreover, the dataset we choose should be able to show the network many different poses in many different contexts, and should be complex enough to highlight where the model is lacking. These datasets should include both simple and complex backgrounds, different viewpoints, occluded parts, different body types and as many different poses as possible.

With this section, I am trying to show the reader all the most famous datasets among the community and compare their strengths and weaknesses, as well as describing which techniques were used to collect and label the data.

7.1 FLIC

The first important dataset specifically designed to train and test for the HPE task was presented in 2013 by Sapp and Taskar [49], called FLIC (Frames Labeled in Cinema). Before this, most datasets were containing less than 1000 images, and those with more (e.g. H3D [5], PASCAL VOC [16]) had insufficient resolution or images were severely non-frontal or occluded.

The final version of FLIC contained 5003 images which were derived from popular Hollywood movies by a multi-step process: first, a person detector was run on every tenth frame of 30 movies. Then, the most confident predictions were sent to Amazon Mechanical Turk, a popular crowd-sourcing platform which they used to label each image with

Figure 7. An example from the Leeds Sports Pose dataset. Note how these people are all wearing sport outfits and how the legs are all in a different spatial positions with respect to the head. [25]

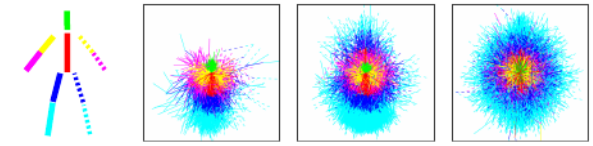

Figure 8. A comparison of the distribution of poses in the different datasets, realized by centering the image around the neck and plotting the position of other body parts. Image (d) represents the extended version of the LSP dataset, which is clearly the most balanced one. [26]

10 upper body joints by paying multiple workers $0.01 each for every labeled image and successively aggregating their answers. Finally, 20% of the images were set aside for the test set.

I believe this paper to be an important milestone for two main reasons. On one hand, the authors presented a large and well designed dataset which could be used by other researchers for testing their systems with more accuracy. On the other hand, they showed how crowdsourcing could provide excellent results if designed correctly, also for image labeling, a task which can be tricky due to cheating or lazy workers.

7.2 LSP

The Leeds Sports Pose dataset came in 2010, presented by Johnson and Everingham [25]. The main goal of this new dataset was to also include different constrained poses and have a larger dataset than the one previously used by the majority, the Iterative Image Parsing (IIP) dataset [46], which only contained approximately 300 images.

For this reason, the authors collected and labeled 2,000 images from Flickr, by searching through sport tags such as "parkour" or "baseball", which contain many highly challenging poses. A few images taken from this dataset can be see in figure 7. One year later, the same authors released a much bigger version of the same dataset [26] containing 10,000 images, whose goal was to have a good balance of all possible poses, as shown in figure 8. This work required an accurate design of a crowdsourcing task which was performed on Amazon Mechanical Turk and collected roughly 400 annotations per hour, which were added to the standard LSP dataset. All images contained labels for 14 different body joints, but some consider this dataset not reliable because the system often did not discard labels from bad crowdworkers, which were not pre-selected. Moreover, since the LSP dataset is limited to sport scenes, most people are wearing tight sport outfits, which is a great limit of this dataset.

7.3 MPII

The "Max Planck Institute for Informatics" Human Pose dataset was released in 2014 by Andriluka et al. [2] and is still considered by many to be the go-to dataset for assessing the performance of networks designed for the HPE task. In fact, it makes a significant advance in terms of diversity and difficulty, as well as the mere size of the collection, when compared to previously used datasets. By collecting images representing over 800 human activities, the authors tried to create something which was not limited in its scope and variability. In total, more than 40,000 images were collected and extensively labeled with 16 «positions of body joints, full 3D torso and head orientation, occlusion labels for joints and body parts, and activity labels» [2]. The variety of this collection is not only represented by the different human activities, but also from the presence of indoor and outdoor scenes and both amateur and professional recordings, as well as many interactions of people with various objects and environments.

To collect the data, the authors queried the video platform YouTube based on activity descriptions, which resulted in a large set of almost 4,000 videos. Then, frames of these video were manually picked with the restriction of them being at least 5 seconds apart from each other, and labels were put around the people present in these frames, which generated more than 40,000 images. Finally, these cropped images were sent to Amazon Mechanical Turk and labeled by crowdworkers. However, this time a careful pre-selection of the workers was performed and the collected labels were closely inspected to maintain high data quality.

While already being a complete dataset for single-person HPE, MPII can also be used for training and testing multi-person HPE networks.

7.4 COCO

COCO[2] Common Objects in Context is a huge large dataset that was presented in 2014 by Lin et al. [33] and is currently used every year for different challenges, including Detection,

[2] https://cocodataset.org/

Paper	*Summary*	*PCKh@0.5 (%)*
Krizhevsky et al. [31] (2012)	Proves how CNNs can be extremely powerful on computer vision tasks. Introduces AlexNet (5 convolutional layers + 3 fully connected)	n/a
Regression Based		
Toshev and Szegedy [56] (2014)	Uses AlexNet for HPE and introduces the *cascade of regressors* to progressively refine the output	–
Sun et al. [51] (2017)	Revives the interest in regression based methods. Uses bones instead of joints because of a more primitive and stable representation, which also encodes long distance interaction of body parts.	86.4
Luvizon et al. [34] (2018)	Merges HPE and action recognition in a single multi-task fully differentiable pipeline to achieve close to state of the art performances on 2D/3D HPE.	91.2
Detection Based		
Tompson et al. [55] (2014)	First full deep neural net solution. Introduces the DeepPose architecture, which makes use of a double scale CNN and a Markov Random Field.	79.6
Newell et al. [39] (2016)	Introduces the Stacked Hourglass architecture, which is able to process the input image at many different scales in a single sequential pipeline	90.9
Chu et al. [12] (2017)	The authors improved the stacked hourglass architecture by replacing the residual unit and improved its performance by generating first different attention maps with the use of CRFs	91.5
Tang et al. [54] (2018)	Uses compositional models to reach state of the art results by performing extremely well on overlapping parts and cluttered backgrounds	92.3
Tang and Wu [53] (2019)	Further explores the idea of the hierarchy of the human body by proposing a part-based branching network. Currently the best performance on single-person HPE	<u>92.7</u>
Zhang et al. [61] (2019)	They successfully distilled the Stacked Hourglass network with minimal drop in performance	91.1
Isack et al. [22] (2020)	Designed a lightweight network by incorporating geometric priors and using a stacked architecture for prediction refinement	90.3
GANs		
Chou et al. [11] (2017)	First one to obtain promising results with GANs	–
Chen et al. [8] (2017)	Uses two discriminators to discriminate fake poses from real ones and for generating body joint heatmaps	91.9
Wang et al. [57] (2019)	Exploits the work of [62] and uses self-attention GANs, which can better model long-range relations in a single image	92.3
Shamsolmoali et al. [50] (2020)	Use the Multiple instance learning technique with GANs and feed the network with a bag of different noisy poses which is labeled with the ground truth pose	92.3

Table 1. This table provides a summary of the most important papers discussed in this survey, as well as a comparison of their performances based on their PCKh@0.5 score on the MPII dataset. Some of them do not have a PCKh score because it was not reported or were tested on a different dataset. The current best performance is underlined.

Keypoint[3] and DensePose. For the detection challenge the train, test and validation sets contain more than 200,000 images in total, with 250,000 person instances labeled with keypoints.

This well known collection was originally designed for object detection and segmentation with the majority of instances pictured in their natural environment. The number of images in the set is large because they have been collected from the most popular search engines, i.e. Google, Bing and Flickr and annotated, as usual, on Amazon Mechanical Turk.

While being a much richer dataset than MPII, it might not be a preferable choice because it also contains imprecise labels or images without any label. However, this might also come as an advantage for unsupervised and semi supervised learning techniques. In conclusion, the COCO keypoint 2017 dataset represents each person with 17 joints, similarly to MPII.

[3] https://cocodataset.org/#keypoints-2020

8 Conclusion

I would like to conclude this survey with a table that summarizes the most important aspects of the works described in this paper, hoping for this to be a helpful resource for a reader that is approaching and wants to learn more about this research field the same way as I was doing so weeks ago. Table 1 contains a short overview of each work, as well as a comparison of their PCKh@0.5 score on the MPII dataset.

Moreover, I hope for my analysis of datasets and common evaluation metrics to provide an effective starting point for researchers wanting to concretely build and test a supervised learning solution for this task.

In conclusion, this work tried to present a rigorous and exhaustive overview of how deep learning solutions for human pose estimation evolved throughout the years. The main aspect I would like the reader to notice is how each presented solution could achieve greater performances only by directly addressing and building upon the weaknesses of the previous state of the art, a method which all researchers should follow. As future work, I would like to expand this survey by analyzing more papers that I did not include in my work and also consider solutions for multi-person and 3D HPE.

References

[1] Saad Ali and Mubarak Shah. 2008. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence* 32, 2 (2008), 288–303.

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*. 3686–3693.

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. 2019. Exploiting temporal context for 3D human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3395–3404.

[4] Vasileios Belagiannis and Andrew Zisserman. 2017. Recurrent human pose estimation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 468–475.

[5] Lubomir Bourdev and Jitendra Malik. 2009. Poselets: Body part detectors trained using 3d human pose annotations. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 1365–1372.

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008* (2018).

[7] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. 2016. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4733–4742.

[8] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. 2017. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1212–1221.

[9] Yucheng Chen, Yingli Tian, and Mingyi He. 2020. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding* 192 (2020), 102897.

[10] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7103–7112.

[11] Chia-Jung Chou, Jui-Ting Chien, and Hwann-Tzong Chen. 2018. Self adversarial training for human pose estimation. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 17–30.

[12] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. 2017. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1831–1840.

[13] Hai Ci, Xiaoxuan Ma, Chunyu Wang, and Yizhou Wang. 2020. Locally Connected Network for Monocular 3D Human Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[14] Qi Dang, Jianqin Yin, Bin Wang, and Wenqing Zheng. 2019. Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology* 24, 6 (2019), 663–676.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2007. The PASCAL visual object classes challenge 2007 (VOC2007) results. (2007).

[17] Vittorio Ferrari, Manuel Marin-Jimenez, and Andrew Zisserman. 2008. Progressive search space reduction for human pose estimation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[19] Shaoli Huang, Mingming Gong, and Dacheng Tao. 2017. A coarse-fine network for keypoint localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 3028–3037.

[20] Dong-Hyun Hwang, Suntae Kim, Nicolas Monet, Hideki Koike, and Soonmin Bae. 2020. Lightweight 3D Human Pose Estimation Network Training Using Teacher-Student Learning. In *The IEEE Winter Conference on Applications of Computer Vision*. 479–488.

[21] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*. Springer, 34–50.

[22] Hossam Isack, Christian Haene, Cem Keskin, Sofien Bouaziz, Yuri Boykov, Shahram Izadi, and Sameh Khamis. 2020. RePose: Learning deep kinematic priors for fast human pose estimation. *arXiv preprint arXiv:2002.03933* (2020).

[23] Arjun Jain, Jonathan Tompson, Mykhaylo Andriluka, Graham W Taylor, and Christoph Bregler. 2013. Learning human pose estimation features with convolutional networks. *arXiv preprint arXiv:1312.7302* (2013).

[24] Chenru Jiang, Kaizhu Huang, Shufei Zhang, Xinheng Wang, and Jimin Xiao. 2020. Pay Attention Selectively and Comprehensively: Pyramid Gating Network for Human Pose Estimation without Pre-training. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2364–2371.

[25] Sam Johnson and Mark Everingham. 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation.. In *bmvc*, Vol. 2. Citeseer, 5.

[26] Sam Johnson and Mark Everingham. 2011. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*. IEEE, 1465–1472.

[27] Perla Sai Raj Kishore, Sudip Das, Partha Sarathi Mukherjee, and Ujjwal Bhattacharya. 2019. ClueNet: A Deep Framework for Occluded Pedestrian Pose Estimation.. In *BMVC*. 245.

[28] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. 2018. Multiposenet: Fast multi-person pose estimation using pose residual network. In *Proceedings of the European conference on computer vision (ECCV)*. 417–433.

[29] Iasonas Kokkinos. 2017. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6129–6138.

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

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[32] Stan Z Li. 1994. Markov random field models in computer vision. In *European conference on computer vision*. Springer, 361–370.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[34] Diogo C Luvizon, David Picard, and Hedi Tabia. 2018. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5137–5146.

[35] Diogo C Luvizon, Hedi Tabia, and David Picard. 2017. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics* 85 (2017), 15–22.

[36] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2019. Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. *arXiv preprint arXiv:1907.00837* (2019).

[37] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. 2018. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*. IEEE, 120–130.

[38] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. 2019. Posefix: Model-agnostic general human pose refinement network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7773–7781.

[39] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*. Springer, 483–499.

[40] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. 2018. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372* (2018).

[41] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. 2018. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7307–7316.

[42] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7753–7762.

[43] Tomas Pfister, Karen Simonyan, James Charles, and Andrew Zisserman. 2014. Deep convolutional neural networks for efficient pose estimation in gesture videos. In *Asian Conference on Computer Vision*. Springer, 538–552.

[44] Rafal Pytel, Osman Semih Kayhan, and Jan C van Gemert. 2020. Tilting at windmills: Data augmentation for deep pose estimation does not help with occlusions. *arXiv preprint arXiv:2010.10451* (2020).

[45] Umer Rafi, Juergen Gall, and Bastian Leibe. 2015. A semantic occlusion model for human pose estimation from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 67–74.

[46] Deva Ramanan. 2006. Learning to parse images of articulated objects. NIPS.

[47] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. 2018. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 750–767.

[48] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. 2017. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3433–3441.

[49] Ben Sapp and Ben Taskar. 2013. Modec: Multimodal decomposable models for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3674–3681.

[50] Pourya Shamsolmoali, Masoumeh Zareapoor, Huiyu Zhou, and Jie Yang. 2020. AMIL: Adversarial Multi-instance Learning for Human Pose Estimation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 1s (2020), 1–23.

[51] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. 2017. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*. 2602–2611.

[52] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261* (2016).

[53] Wei Tang and Ying Wu. 2019. Does Learning Specific Features for Related Parts Help Human Pose Estimation?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1107–1116.

[54] Wei Tang, Pei Yu, and Ying Wu. 2018. Deeply learned compositional models for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 190–206.

[55] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*. 1799–1807.

[56] Alexander Toshev and Christian Szegedy. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1653–1660.

[57] Xiangyang Wang, Zhongzheng Cao, Rui Wang, Zhi Liu, and Xiaoqiang Zhu. 2019. Improving human pose estimation with self-attention generative adversarial networks. *IEEE Access* 7 (2019), 119668–119680.

[58] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 2018. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5255–5264.

[59] Yi Yang and Deva Ramanan. 2012. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence* 35, 12 (2012), 2878–2890.

[60] Xingyu Zeng, Wanli Ouyang, and Xiaogang Wang. 2013. Multi-stage contextual deep learning for pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 121–128.

[61] Feng Zhang, Xiatian Zhu, and Mao Ye. 2019. Fast human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3517–3526.

[62] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *International Conference on Machine Learning*. PMLR, 7354–7363.