An Improved Mixture Density Network for 3D Human Pose Estimation with Ordinal Ranking

Yiqi Wu 1,2,*, Shichao Ma 1, Dejun Zhang 1,2,*, Weilun Huang 3 and Yilin Chen 2,4

1 School of Computer Science, China University of Geosciences, Wuhan 430074, China; wuyq@cug.edu.cn (Y.W.); msc@cug.edu.cn (S.M.)
2 Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology), Wuhan 430205, China; yilinchen@wit.edu.cn
3 School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China; 201920632@cug.edu.cn
4 School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China

* Correspondence: zhangdejun@cug.edu.cn

Abstract: Estimating accurate 3D human poses from 2D images remains a challenge due to the lack of explicit depth information in 2D data. This paper proposes an improved mixture density network for 3D human pose estimation called the Locally Connected Mixture Density Network (LCMDN). Instead of conducting direct coordinate regression or providing unimodal estimates per joint, our approach predicts multiple possible hypotheses by the Mixture Density Network (MDN). Our network can be divided into two steps: the 2D joint points are estimated from the input images first; then, the information of human joints correlation is extracted by a feature extractor. After the human pose feature is extracted, multiple pose hypotheses are generated via the hypotheses generator. In addition, to make better use of the relationship between human joints, we introduce the Locally Connected Network (LCN) as a generic formulation to replace the traditional Fully Connected Network (FCN), which is applied to a feature extraction module. Finally, to select the most appropriate 3D pose result, a 3D pose selector based on the ordinal ranking of joints is adopted to score the predicted pose. The LCMDN improves the representation capability and robustness of the original MDN method notably. Experiments are conducted on the Human3.6M and MPII dataset. The average Mean Per Joint Position Error (MPJPE) of our proposed LCMDN reaches 50 mm on the Human3.6M dataset, which is on par or better than the state-of-the-art works. The qualitative results on the MPII dataset show that our network has a strong generalization ability.

Keywords: 3D human pose estimation; mixture density network; Gaussian mixture model; graphic convolutional network; ordinal ranking

1. Introduction

With the popularity of virtual reality devices, the task of 3D human pose estimation from a monocular image is gaining importance and becoming one of the challenging tasks in the field of computer vision. A general and effective method for 3D human pose estimation is a two-stage approach, in which the 2D joints are detected from a monocular input first [1], and then, the 3D pose is generated from the detected 2D joints [2–9]. The advantage of the two-stage approach is that it directly predicts a 3D pose from simple 2D joints, diminishing the interference of input RGB images, such as variations in the background scene, lighting, clothing color, etc. In addition, the two-stage approach can fully utilize the 2D joint labels in the human pose dataset and is adapted to both indoor and in the wild scenarios.

While deep learning brings remarkable progress in many computer vision tasks, 3D human pose estimation from a monocular image still faces the problem of depth ambiguity. To be specific, different 3D human poses may have the same or similar 2D projections...
under different conditions, such as lighting or viewing angle, making the task of estimating 3D human poses based on deep learning from a single RGB image uncertain. Some methods use various geometric constraints between joints, such as joint limits [2] and bone length [10], etc., to eliminate infeasible 3D poses. However, there may still be the problem of multiple 3D poses with similar 2D projections. Some recent approaches generate multiple hypotheses of geometrically feasible 3D human pose to solve the uncertainty problem. Jahangiri and Yuille [11] were the first to propose the idea of generating multiple 3D pose hypotheses.

In contrast to single estimation, it can effectively eliminate model over-fitting and enhance the generalization capability. Based on it, Li et al. [12] proposes a two-stage approach that adopts the Mixture Density Networks (MDN) [13] for the 3D human pose estimation task and achieves state-of-the-art results. The network first lifts the 2D joint into a latent space via a feature extractor only implemented by a simple Fully Connected Network (FCN); then, multiple hypotheses are generated by a hypothesis generator. However, the network has two obvious drawbacks due to its several linear layers. First, a number of learnable parameters that come from dense connections might result in over-fitting. Second, the robustness of this network is insufficient due to the limited feature extraction capability caused by an overly simplistic network structure. In addition, this method outputs multiple results, and the selection of a single result relies on the ground truth label. Hence, the actual estimation results are random and unstable in real-world applications where there is no 3D human pose ground truth label.

In order to solve the problems above, we propose the Locally Connected Mixture Density Network (LCMDN) for 3D human pose estimation. The LCMDN proposes improvements for the MDN [12] and is inspired by the idea of the Locally Connected Network (LCN) [14,15], which introduces a variant of GCN for the 2D to 3D task.

Figure 1 is the pipeline of our proposed LCMDN. Our network can be divided into two steps: the 2D joint information is first obtained from the input image, and then, a feature extractor is used to extract features, and the parameters of a Gaussian mixture model (GMM) is obtained by the hypotheses generator. Multiple predictions of 3D human poses can be generated from the GMM. Specifically, our network uses LCN instead of the simple dense connections in a feature extractor, reducing the challenge of over-fitting due to sparse joint connections between joints. Our model predicts multiple possible hypotheses by MDN so that its robustness is significantly improved. Moreover, we introduce a pose selector to select one pose that best matches the human joint correspondence to solve the problem of ground truth absence in real-world applications.

Figure 1. Pipeline for the proposed Locally Connected Mixture Density Network (LCMDN). Our network takes RGB images as input and outputs multiple predictions of 3D pose first. Then, a 3D pose selector is trained to select the best estimation result.

We test the proposed LCMDN on public datasets. Qualitative experiments show that our method can effectively obtain accurate 3D human poses from 2D images. Quantitative experiments show that our network achieves outstanding estimation accuracy. In addition, we conduct robustness tests and ablation studies for the proposed LCMDN for further analyses of the network’s ability. The main contributions of this paper are listed as follows:
We propose an LCN-based human pose estimation network that learns a Gaussian mixture model matching the distribution of human joints to output multiple hypotheses. LCN is applied to a 3D human pose estimation task with multiple pose outputs, which improves the accuracy of the estimation task by learning the structural relationships of human joints. A 3D pose selector is designed to select the best predicted 3D human pose. In the selector, an ordinal matrix containing joints relationship is learned from the input RGB images via an hourglass network. Our network achieves comparable or better results than the state-of-the-art in terms of accuracy and visualization with better robustness, and experimental results on the MPII human dataset validate the generalization ability of our method.

2. Related Work

2.1. Graph Convolutional Networks

Graph convolution network (GCN) is commonly visible in computer vision and achieves state-of-the-art performance by leveraging GCN to model the relations such as temporal sequences and visual objects. GCN can be divided into two categories: the spatial domain and the spectral domain. For the spatial domain, the convolution process is performed with the Fourier transform, while for the latter, the spectral domain is applied directly to the nodes of the graph and their neighbors. Our work belongs to the latter stream, where convolution operates in the spectral domain.

Currently, several works attempt to apply GCN to human pose estimation tasks. Zhao et al. proposed semantic graph convolution and integrated nonlocal layers into the network to expand the acceptance domain of the graph kernel. The LCN is proposed to overcome the limitations of the GCN by assigning dedicated rather than shared filters to different joints.

In this paper, we apply the idea of graph convolution to the MDN-based human pose estimation task to replace FCN. Our proposed LCMDN learns the semantic relationships between body joints by introducing a unique weight matrix for each joint point.

2.2. 3D Pose Estimation

Existing methods for estimating 3D human pose are divided into two categories according to their inferring process. One is an end-to-end approach based on deep convolutional neural networks (CNNs) where 3D human poses are directly generated from the input images. Zhou et al. represent the 3D pose as a sparse representation and predict the 3D pose using an expectation–maximization (EM) algorithm. Park et al. concatenate detected 2D poses and additional information on relative location among multiple joints to improve traditional CNNs. Pavlakos et al. come up with a volumetric representation to predict 3D heatmaps considered as a volumetric version of the stack hourglass network. Yang et al. proposes an adversarial network to separate the ground truth 3D labels from generated labels. Simultaneously, Zhou et al. propose a weakly supervised transfer learning approach by using mixed 2D and 3D annotations where 3D pose labels in indoor environments can be applied to in-the-wild inputs.

The other method first predicts 2D pose joints from a monocular image followed by lifting 2D to 3D human poses by fitting a probabilistic 3D pose model. The two-stage approach decouples a complicated problem into two easier processes. Akhter et al. estimate the 3D pose from detected 2D joints using a multi-stage approach providing an over-complete dictionary of poses. Bogo et al. optimize the error between the reprojected 3D pose and detected 2D joints. Martinez et al. directly regress a 3D pose from given 2D joints by a simple fully connected residual network. To solve the uncertainty problem from 2D to 3D joints, Jahangiri and Yuille et al. first propose an approach to generate multiple pose estimation hypotheses to solve this problem. The authors first learned from a set of 3D human poses uniformly sampled from a 3D human pose dataset and then learned a 3D GMM model using conditional sampling.
improved on this by combining the traditional neural network with a mixed density model to solve the depth in 3D human pose estimation blurring and occlusion problems.

Considering the special characteristics of human joints, most of the multi-pose estimation hypothesis methods do not exploit the structural relationships between joints. In view of this, we apply the LCN to the 3D human pose estimation task with multi-pose output and design a 3D human pose estimation network that can exploit the structural connectivity relationships of human joints effectively based on a step-by-step training model.

Moreover, the methods using a mixed density model [12,34] have the shortcomings that the selection of optimal results is unstable in practice, as 3D pose labels are not provided in the real-world applications. To solve this problem, we use the hourglass network to learn the position relationship between the joints from the images and use it as an indicator to select the 3D human pose that best matches the actual results.

3. Locally Connected Mixture Density Network

In this section, we first introduce the modeling of our proposed LCMDN. The architecture of LCMDN is shown in Figure 2. The LCMDN takes the 2D image as input and outputs a correct 3D human pose. The network consists of a 2D pose estimator, feature extractor, hypothesis generator, and 3D pose selector. Details of each module are introduced in the following subsections.

![Figure 2. Illustration of the proposed Locally Connected Mixture Density Networks. The network consists of a 2D pose estimator, a feature extractor, a 3D pose hypotheses generator, and a 3D pose selector. The hypotheses generator outputs multiple hypothetical 3D poses from the detected 2D joints. The ordinal matrix is used to generate the ordinal ranking of joints to select the correct human pose estimation. M denotes the numbers of Gaussian components.](image)

3.1. Model Representation

Different from directly training CNNs to estimate the 3D human poses from images or 2D poses, our network can estimate multiple diverse 3D pose hypotheses by learning the Gaussian distribution of human body poses. More specifically, the probability density of 3D pose joints \( Y \in \mathbb{R}^{3N} \) can be represented as a linear combination of Gaussian kernel functions when inputting the 2D joints \( x \in \mathbb{R}^{2N} \), which are denoted as

\[
p(Y|x, \Theta) = \sum_{m=1}^{M} \alpha_m(x, \Theta) \phi_m(y|x, \Theta),
\]

(1)
where $\Theta$ represents the learnable parameters of the network, $M$ is the number of Gaussian kernels, $a_m$ is the mixing coefficient of the $m$th Gaussian distribution, and $\phi_m$ represents the $m$th Gaussian distribution, which can be denoted as

$$\phi_m(y|x, \Theta) = \frac{1}{(2\pi)^{c/2}\sigma_m(x, \Theta)^c} \exp\left(-\frac{\|y - \mu_m(x, \Theta)\|^2}{2\sigma_m(x, \Theta)^2}\right), \quad (2)$$

where $\mu_m$ and $\sigma_m$ denote the mean and variance of the $m$th kernel, respectively.

The whole training process of our network can be described as follows: given the 2D joint detections, it outputs the parameters of the Gaussian mixture model. The multiple hypotheses for 3D human pose estimation can be generated from the Gaussian mixture model.

Furthermore, in our proposed LCMDN, to select the best 3D pose from the multiple hypotheses, a joints ordinal ranking matrix is generated from the input image, which indicates the depth position relationship between joints. The estimated pose that matches the relationship best is considered as the final estimated 3D human pose.

3.2. Two-Dimensional (2D) Pose Estimator and Feature Extractor

To extract 2D pose joints, the state-of-the-art stacked hourglass network [1] is adopted as the 2D pose estimation module. We use LCN layers to extract more valid information in the feature extractor.

For existing classical MDN-based 3D pose estimation methods [35], the fully connected layers are often adopted for feature extraction. The Laplace operator of FCN and LCN can be transformed into the product of a structure matrix and a weight matrix, which can be formalized as:

$$y = X(S \odot W), \quad (3)$$

where structure matrix $S$ is shared for all LCN layers and is constructed based on the specified joints dependence. Weight matrix $W$ is learned end-to-end, and it varies with different LCN layers.

Figure 3 reflects the difference in the structure matrix between FCN and LCN. Taking Joint2 in the figure as an example, FCN considers the relationship between it and all other joints, while LCN only considers associated joints, such as Joint1, Joint3, and Joint4.

Joint dependence is determined by ensuring whether the distance between two joints is less than a hyper-parameter $K$. For instance, the distance between Joint2 and Joint1 is 1 because they are directly connected. The distance from Joint2 to Joint4 is 2 because they are connected indirectly through Joint3.

Figure 3. The difference of structure matrix between LCN and GCN. For Joint2, the LCN considers only the relationships within the orange joints, while the FCN considers the relationships with all other joints.
Specifically, as shown in Figure 2, the feature extractor takes the information of the coordinates of the $J$ joints obtained from the RGB images as input, obtains the feature vectors containing the relationship of the joints through the LCN layers, and projects them to a high-dimensional vector. Each LCN layer also contains the batch normalization layer and activation layer (ReLU). Residual connection is introduced between every two LCN layers to ensure the validity of the features learned by deeper layers.

### 3.3. Hypotheses Generator

Unlike the regression of 3D human joint point coordinates directly from human pose features, the role of the hypotheses generator is to learn the Gaussian mixture distribution model of human 3D joint points from human pose features. As shown in Figure 2, the human pose feature is fed into three different linear connectivity layers and outputs the mixing coefficients $\alpha$, mean $\mu$ and variance $\sigma$ of the Gaussian mixture model, using three kinds of activation functions, where $M$ denotes the number of Gaussian kernels.

Specifically, suppose the human features learned in the feature generator are $x$. $F_\mu$, $F_\mu$, and $F_\sigma$ represent three different linear layers, respectively. The mean value $\mu$ reflects the average degree of 3D human joint point information, so the mean value can be calculated by one linear layer, which is denoted as:

$$\mu = F_\mu(x)$$

The mixing coefficients $\alpha$ reflect the weights of the single Gaussian model with a sum of 1, and each value satisfies the range from 0 to 1. Therefore, the softmax function is used as the activation function in the output, which is defined as:

$$\alpha = \frac{F_\alpha(x)}{\sum F_\alpha(x)}$$

To ensure the validity of the variance $\sigma$, the ELU function is chosen as the activation function to calculate the variance. In addition, in order to make the value of the variance always greater than 0, each term of the original formula definition is added by 1. The modified ELU function is defined as function 6, where $\gamma$ represents a scale for negative factors.

$$\sigma = \begin{cases} F_\sigma(x) + 1, & F_\sigma(x) \geq 0 \\ \gamma(\exp(F_\sigma(x)) - 1) + 1, & F_\sigma(x) < 0 \end{cases}$$

### 3.4. 3D Pose Selector

In recent work [36–39], ordinal relations have been used in 3D human estimation tasks with ordinal annotations to impose penalties for violations of ordinal depth constraints. The 3D pose selector is introduced to select the most realistic 3D pose from the generated multi-pose human hypotheses.

Similar to Sharma [36], we use the ordinal ranking to estimate the depth and reflectance of human joints. Our network uses human RGB images as input and a four-layer stacked hourglass network to obtain the ordinal matrix. As shown in Figure 2, the human joints are listed in rows and columns of the ordinal matrix. The matrix reflects the ordinal relations of each pair of human joints, in which $>,=,<$ represents the positional relationship of less, greater and equal, respectively. See more details of the ordinal matrix in [36].

The whole training process is divided into two parts. Firstly, multiple 3D poses are output from the Hypothesis Generator. Then, the pose selector is used to obtain the ordinal matrix from the input images. The obtained matrix is compared with the ordinal matrix generated from the predicted poses. The one that matches best is used for the backpropagation in the training phase.

During the testing phase, multiple human pose hypotheses are generated, and corresponding ordinal matrices are outputted. Then, the pose selector selects the most realistic 3D pose by comparing ordinal matrices and finding the best match.
4. Experiments

In this section, qualitative and quantitative experiments are conducted to evaluate our proposed LCMDN. At first, implementation details, datasets and metrics are presented. Then, we test the network on the Human3.6M [40] and MPII [41] dataset and make comparisons to the state-of-the-art methods. Some additional ablation studies are designed to demonstrate the capability of the network.

4.1. Training Details and Developing Environment

For the optimizer, we choose the Adam [42] with exponential decay and set the initial learning rate to 0.001. For the initialization of the network, the Kaiming initialization [43] is applied. The batch size is 64, and the network is trained for 50 epochs. The testing time of our proposed LCMDN from the RGB image to 2D pose is about 78 ms and from 2D pose to 3D pose is about 0.8 ms on average. The max-norm constraint is used to ensure the weight of each layer in [0, 1]. The value of \( a_i(x) \) and \( \sigma_i(x) \) are fixed to \([1 \times 10^{-8}, 1]\) and \([1 \times 10^{-15}, 1 \times 10^{15}]\), respectively, to maintain the training loss. As a two-stage estimation method, the stacked hourglass network [1] method for 2D joints detection is adopted. We use the MPII dataset for the pre-training of the 2D joints detection network and the Human3.6M dataset for the fine-tuning.

The hardware for experiments includes the following: CPU: Intel i5-10500, GPU: RTX 3060, RAM: 32 GB. The developing environment is Ubuntu16.04, CUDA 10.1, cuDNN 7.4, and TensorFlow.

4.2. Dataset and Metric

**Human3.6M dataset [40]:** The Human3.6M dataset is a large public human pose dataset with about 3.6 million indoor images. These images are captured to show 15 different human actions performed by seven people. The dataset offers both 2D and 3D human pose ground truth by labels of human joints’ positions. Different subjects are selected for training (1,5,6,7,8) and testing (9,11). We follow [5] for the standard normalization of both 2D and 3D joints.

**MPII dataset [41]:** The MPII dataset is a small 2D human pose dataset with image data mainly from major video sites, such as YouTube, and it provides before and after unannotated frames. The MPII dataset includes about 25,000 images with human behavioral actions, including over 40,000 annotated human joints, covering 410 human activities, including walking, sitting, skiing, and hiking. It is one of the mainstream human pose datasets. Since the MPII dataset does not provide 3D pose information, we only report qualitative visualization results for the MPII dataset.

**Evaluation Metric:** For Human3.6M, there are two common evaluation protocols. One standard protocol computes the Mean Per Joint Position Error (MPJPE) in millimeters between the ground truth and predictions after being aligned with the root joint. This metric is called Protocol #1. The other protocol further aligns the predictions with the ground truth, leveraging a rigid transformation. This protocol is called Protocol #2. We use the more challenging Protocol #1 for evaluation in all the experiments.

4.3. Results on Human3.6M Dataset

The experiment results of our method and several latest methods for 3D pose estimation on the Human3.6M dataset are shown in Table 1. From the table, the LCMDN outperforms other methods in most human actions. It achieves an average error of 50 mm in 15 human behavior actions, which is a decrease of 5.4% compared to the previous best average error of 52.7 mm. For a relatively simple action “Photo”, the error reaches 58.1 mm, which decreases about 5.8% compared to the previous best result of 61.5 mm. Even for some more complex actions, such as “Eating”, the error is reduced from 44.7 to 44.5 mm.
From the results, although the LCMDN is without any multi-view or video information, it still outperforms most other spatial or temporal constraints-based methods due to the extraction of human joints correlation, which indicates the improvement of our model.

### Table 1. The numerical results under protocol # 1 and comparisons on the Human3.6M dataset (mm). (Best result in bold.)

| Method            | Direct. | Discuss. | Eating | Greet | Phone | Photo | Pose | Purch. | Sitting | SittingD. | Smoke | Wait | WalkD. | Walk | WalkT. | Avg. |
|-------------------|---------|----------|--------|-------|-------|-------|------|--------|---------|-----------|-------|------|--------|------|-------|------|
| Lin et al. [40]   | 132.7   | 183.6    | 132.3  | 164.4 | 162.1 | 205.9 | 150.6 | 171.3  | 151.6   | 243.0    | 162.1 | 170.7| 177.1  | 162.1|
| Du et al. [44]    | 85.1    | 112.7    | 104.9  | 122.1 | 139.1 | 135.9 | 105.8 | 166.2  | 117.5   | 226.9    | 120.0 | 117.7| 137.4  | 106.5|
| Zhou et al. [28]  | 87.4    | 109.3    | 87.1   | 103.2 | 116.2 | 143.3 | 106.9 | 99.8   | 124.5   | 199.2    | 107.4 | 118.1| 114.2  | 97.7 | 113.0|
| Pavlakos et al. [26] | 67.4    | 71.9    | 66.7   | 69.1  | 72.0  | 77.0  | 65.0  | 68.3   | 83.7    | 96.5     | 71.7  | 65.8 | 74.9   | 59.1 | 63.2 | 71.9|
| Jahangiri et al. [11] | 63.1    | 55.9    | 58.1   | 64.5  | 68.7  | 61.3  | 55.6  | 86.1   | 117.6   | 71.0     | 71.2  | 66.3 | 57.1   | 62.5 | 61.0 | 68.0|
| Zhou et al. [10]  | 54.8    | 60.7    | 58.2   | 71.4  | 62.0  | 65.5  | 53.8  | 55.6   | 75.2    | 111.6    | 64.1  | 66.0 | 51.4   | 63.2 | 55.3 | 64.9|
| Martinez et al. [5]   | 51.8    | 56.2    | 58.1   | 59.0  | 69.5  | 78.4  | 55.2  | 58.1   | 74.0    | 94.6     | 62.3  | 59.1 | 65.1   | 49.5 | 52.4 | 62.9|
| Lee et al. [31]    | 43.8    | 51.7    | 48.8   | 53.1  | 52.2  | 74.9  | 52.7  | 44.6   | 56.9    | 74.3     | 56.7 | 66.4 | 47.5   | 68.4 | 45.6 | 55.8|
| Li et al. [12]     | 43.8    | 48.6    | 49.1   | 49.8  | 57.6  | 61.5  | 45.9  | 48.3   | 62.0    | 73.4     | 54.8  | 50.6 | 56.0   | 43.4 | 45.5 | 52.7|
| Ci et al. [14]     | 46.8    | 52.3    | 44.7   | 50.4  | 52.9  | 68.9  | 49.6  | 46.4   | 60.2    | 78.9     | 51.2  | 50.0 | 54.8   | 40.4 | 43.3 | 52.7|
| LCMDN              | 42.0    | 47.1    | 44.5   | 48.2  | 54.5  | 58.1  | 44.0  | 45.8   | 57.9    | 71.4     | 52.0  | 48.7 | 52.7   | 41.3 | 42.3 | 50.0|

Considering some joints are often occluded or missing due to various kinds of interference in realistic scenarios, we simulate the situation of missing joins by randomly dropping one or two limb joints to test our network. The results are reported in Table 2 together with the performance of some existing methods which generate single hypotheses or use CNN-based networks. From the table, the performance of our LCMDN has significantly improved compared to the baseline methods, which further indicates the robustness of our method.

### Table 2. Results of missing joints and comparisons. The first three rows show the results of one missing joint, and the last three rows show the results of two missing joints. (Best result in bold.)

| Method            | Direct. | Discuss. | Eating | Greet | Phone | Photo | Pose | Purch. | Sitting | SittingD. | Smoke | Wait | WalkD. | Walk | WalkT. | Avg. |
|-------------------|---------|----------|--------|-------|-------|-------|------|--------|---------|-----------|-------|------|--------|------|-------|------|
| Jahangiri et al. [11] | 108.6   | 105.9   | 105.6  | 109.0 | 105.5 | 109.9 | 102.0 | 111.3  | 119.6   | 107.8    | 107.1 | 111.3| 108.4  | 107.0|
| Martinez et al. [5]   | 57.4    | 64.6    | 64.3   | 65.6  | 73.3  | 85.5  | 61.0  | 62.1   | 84.0    | 101.1    | 68.2  | 66.7 | 70.8   | 55.6 |
| Li et al. [12]       | 48.9    | 53.9    | 54.5   | 55.5  | 62.6  | 70.4  | 51.3  | 52.0   | 69.7    | 83.9     | 60.7  | 57.2 | 62.4   | 48.3 | 50.8 | 58.8|
| LCMDN              | 46.4    | 50.9    | 50.8   | 51.9  | 58.2  | 64.6  | 47.7  | 48.7   | 64.2    | 77.6     | 56.8  | 53.6 | 57.7   | 45.0 | 54.4 | 54.7|

To demonstrate the effectiveness of our method more intuitively, we give some visualization results in Figure 4. As can be seen from rows 1 to 3 of the figure, for simple poses with low ambiguity, such as standing, each Gaussian kernel produces almost identical hypotheses, which indicates that a single Gaussian distribution is effective for predicting simple human poses. However, for human poses with more occlusions and ambiguities, such as the sitting pose, different human pose hypotheses are generated, such as the obvious difference between column 5 and column 7 in row 5, which further demonstrates that the multi-pose estimation model can alleviate the uncertainty problem in the single-pose prediction model by generating multiple human pose hypotheses.
Figure 4. Visualization results on Human3.6M dataset. The first column shows the 2D poses estimated by the stacked hourglass network. The second column shows the ground truth offered by the dataset. The following 5 columns are the estimated hypothetical results generated by the hypotheses generator of the LCMDN, and the poses surrounded by orange frames are the final estimated poses selected by the LCMDN.

4.4. Ablation Study

Different step of neighbors. To verify the effect of the neighborhood distance $K$ (in Section 3.2) on the experimental results, we investigate the value of $K$ from 1 to 4. Results and comparisons with LCN [14] on the Human3.6M dataset are shown in Table 3. It indicates that a small value of $K$ leads to poorer representation capability of our model. However, too large values also bring the decay of performance due to the redundant features. Therefore, $K$ is set to 2 for our network, which ensures that the network can learn valid human joint structure features while reducing the learning of redundant information.

| Method  | 1   | 2   | 3   | 4   |
|---------|-----|-----|-----|-----|
| LCN [14]| 58.77 | 57.73 | 57.56 | 58.37 |
| LCMDN   | 51.28 | 50.02 | 50.42 | 50.45 |
Different number of kernels. There are $M$ Gaussian kernels in Equation (1), which represent $M$ sets of different results generated by the hypotheses generator of our LCMDN. We train four different models with $M$ set to 1, 3, 5, and 8, respectively. The results and the comparisons with [12] are shown on the Human3.6M dataset in Table 4. The results show that multiple Gaussian has better performance than single Gaussian. However, when the number of Gaussian kernels is greater than 5, there is only a tiny improvement. In order to balance the computational burden and capability of the network, the number of Gaussian kernels $M$ is set to 5 for our network.

Table 4. Comparison between different numbers of kernels. Numbers in the first row represent the number of kernels.

| Method     | 1  | 3  | 5  | 8  |
|------------|----|----|----|----|
| Li et al. [12] | 62.9 | 55.2 | 52.7 | 52.6 |
| LCMDN      | 58.8 | 52.4 | 50.0 | 49.8 |

4.5. Three-Dimensional (3D) Human Pose Estimation on MPII Dataset

More qualitative experiments are conducted on MPII to testify the generalization ability of our LCMDN. The MPII dataset does not offer ground truth 3D labels, so the step of 2D to 3D estimation cannot be trained on it, and the numeric estimation accuracy cannot be evaluated. In the experiments, the network settings are the same as experiments in Section 4.3, and we also use the Human3.6M dataset for training.

Visualization results on the MPII dataset are shown in Figure 5. According to the observation, although the model is trained by the indoor dataset, it also achieves excellent generalization performance to in the wild scenes. Moreover, for actions with serious occlusion problems, such as rock climbing and horseback riding, our LCMDN can also estimate complete and correct 3D human pose. As shown in Figure 5, in the photo of rock climbing (column 5, row 1), although half of the human joints in the image are occluded, the LCMDN can still infer an accurate 3D human pose, which also proves the robustness of the LCMDN.

Figure 5. Qualitative results on the MPII dataset. Columns 1, 3, and 5 are the input images and columns 2, 4, and 6 are corresponding 3D poses generated by our network.

5. Conclusions

In this work, we present a novel network, namely LCMDN (Locally Connected Mixture Density Network), for 3D human pose estimation. Our network introduces the Locally Connected Network to the Mixture Density Network to integrate the representation and optimization ability of these two networks.

The LCMDN generates multiple 3D pose hypotheses at first. Then, the most realistic pose is selected based on the positional relationship of joints via the pose selector of the
LCMDN. It solves the result selection problem when there is no ground truth label to evaluate the best result of multiple hypotheses.

Experimental results show that the average MPJPE of our proposed network reaches 50 mm on the Human3.6M dataset, which is on par or better than the state-of-the-art works. Moreover, it is robust in the occlusion or undetected joints scenario and generalizes well for the in the wild scenario. In the future work, we will optimize our network \[45,46\] for better time performance and apply the idea of LCMDN into more extended fields, such as visual tracking \[47\], multi-view and multi-person pose estimation \[46,48\], or 3D human hand pose estimation \[49\].

**Author Contributions:** Y.W. and S.M. conceived of and designed the algorithm and the experiments. S.M. analyzed the data. Y.W. and S.M. wrote the manuscript. D.Z. supervised the research. D.Z. and W.H. provided suggestions for the proposed method and its evaluation and assisted in the preparation of the manuscript. W.H. and Y.C. collected and sorted out the literature. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the National Science Foundation of China (grant No. 61802355 and 61702350) and Hubei Key Laboratory of Intelligent Robot (HBIR 202105).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

**References**

1. 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; pp. 483–499.
2. Akhter, I.; Black, M.J. Pose-conditioned joint angle limits for 3D human pose reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1446–1455.
3. Zhou, X.; Zhu, M.; Leonardos, S.; Derpanis, K.G.; Daniilidis, K. Sparseness meets deepness: 3D human pose estimation from monocular video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4966–4975.
4. Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; Black, M.J. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 561–578.
5. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A simple yet effective baseline for 3D human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2640–2649.
6. Rayat Imitiaz Hossain, M.; Little, J.J. Exploiting temporal information for 3D human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 4–8 September 2018; pp. 68–84.
7. Chen, W.; Wang, H.; Li, Y.; Su, H.; Wang, Z.; Tu, C.; Lischinski, D.; Cohen-Or, D.; Chen, B. Synthesizing training images for boosting human 3D pose estimation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 479–488.
8. Yasin, H.; Iqbal, U.; Kruger, B.; Weber, A.; Gall, J. A dual-source approach for 3D pose estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4948–4956.
9. Moreno-Noguer, F. 3D human pose estimation from a single image via distance matrix regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2823–2832.
10. Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; Wei, Y. Towards 3D human pose estimation in the wild: A weakly-supervised approach. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 398–407.
11. Jahangiri, E.; Yuille, A.L. Generating multiple diverse hypotheses for human 3D pose consistent with 2d joint detections. In Proceedings of the IEEE Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 805–814.
12. Li, C.; Lee, G.H. Generating multiple hypotheses for 3D human pose estimation with mixture density network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9887–9895.
13. Bishop, C.M. *Mixture Density Networks*; Aston University: Birmingham, UK, 1994.
14. Ci, H.; Wang, C.; Ma, X.; Wang, Y. Optimizing Network Structure for 3D Human Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 2262–2271.
15. Ci, H.; Ma, X.; Wang, C.; Wang, Y. Locally connected network for monocular 3D human pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2022, 44, 1429–1442. [CrossRef] [PubMed]

16. Wang, X.; Gupta, A. Videos as space-time region graphs. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 399–417.

17. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LO, USA, 2–7 February 2018.

18. Yang, J.; Lu, J.; Lee, S.; Batra, D.; Parikh, D. Graph r-cnn for scene graph generation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 670–685.

19. Yao, T.; Fan, Y.; Li, Y.; Mei, T. Exploring visual relationship for image captioning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 684–699.

20. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* 2016, arXiv:1609.02907.

21. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3844–3852.

22. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3425–3435.

23. Hammond, D.K.; Vanderheyst, P.; Gribonval, R. Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.* 2011, 30, 129–150. [CrossRef]

24. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3D human pose estimation in the wild using improved cnn supervision. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 506–516.

25. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* 2017, arXiv:1710.10903.

26. Zou, L.; Peng, X.; Tian, Y.; Kapadia, M.; Metaxas, D.N. Semantic graph convolutional networks for 3D human pose regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7025–7034.

27. Lee, K.; Lee, I.; Lee, S. Propagating lstm: 3D pose estimation based on joint interdependency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 119–135.

28. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Deep kinematic pose regression. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 186–201.

29. Park, S.; Hwang, J.; Kwak, N. 3D human pose estimation using convolutional neural networks with 2d pose information. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 156–169.

30. Yang, W.; Ouyang, W.; Wang, X.; Ren, J.; Li, H.; Wang, X. 3D human pose estimation in the wild by adversarial learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5255–5264.

31. Lee, S.; Lee, I.; Lee, S. Propagating lstm: 3D pose estimation based on joint interdependency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 119–135.

32. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Integral human pose regression. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 529–545.

33. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.

34. Zou, L.; Huang, Z.; Gu, N.; Wang, F.; Yang, Z.; Wang, G. GMDN: A lightweight graph-based mixture density network for 3D human pose regression. *Comput. Graph.* 2021, 95, 115–122. [CrossRef]

35. Guillames, A.B. Mixture Density Networks for Distribution and Uncertainty Estimation. Ph.D. Thesis, Universitat Politècnica de Catalunya, Facultat d’Informàtica de Barcelona, Barcelona, Spain, 2017.

36. Sharma, S.; Varigonda, P.T.; Bindal, P.; Sharma, A.; Jain, A. Monocular 3D human pose estimation by generation and ordinal ranking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–November 2019; pp. 2325–2334.

37. Ronchi, M.R.; Mac Aodha, O.; Eng, R.; Perona, P. It’s all relative: Monocular 3D human pose estimation from weakly supervised data. *arXiv* 2018, arXiv:1805.06880.

38. Pons-Moll, G.; Fleet, D.J.; Rosenhahn, B. Posebits for monocular human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2337–2344.

39. Pavlakos, G.; Zhou, X.; Daniilidis, K. Ordinal depth supervision for 3D human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7307–7316.

40. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* 2013, 36, 1325–1339. [CrossRef] [PubMed]

41. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.

42. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* 2014, arXiv:1412.6980.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.

44. Du, Y.; Wong, Y.; Liu, Y.; Han, F.; Gui, Y.; Wang, Z.; Kankanhalli, M.; Geng, W. Marker-less 3D human motion capture with monocular image sequence and height-maps. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 20–36.

45. Zhang, D.; He, L.; Luo, M.; Xu, Z.; He, F. Weight asynchronous update: Improving the diversity of filters in a deep convolutional network. Comput. Vis. Media 2020, 6, 455–466. [CrossRef]

46. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. 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.

47. Zhang, D.; Zhang, Z.; Zou, L.; Xie, Z.; He, F.; Wu, Y.; Tu, Z. Part-based visual tracking with spatially regularized correlation filters. Vis. Comput. 2020, 36, 509–527. [CrossRef]

48. Zhang, D.; Wu, Y.; Guo, M.; Chen, Y. Deep Learning Methods for 3D Human Pose Estimation under Different Supervision Paradigms: A Survey. Electronics 2021, 10, 2267. [CrossRef]

49. Wu, Y.; Ma, S.; Zhang, D.; Sun, J. 3D Capsule Hand Pose Estimation Network Based on Structural Relationship Information. Symmetry 2020, 12, 1636. [CrossRef]