西北工业大学

数字图像处理–论文翻译

原论文标题: AdaFuse: Adaptive Multiview Fusion for Accurate Human Pose Estimation in the Wild

宋绍淇

计算机学院

人工智能

2024 年 11 月

学号: 2022303297

Northwestern Polytechnical University

Noname manuscript No.
(will be inserted by the editor)

AdaFuse: Adaptive Multiview Fusion for Accurate Human Pose Estimation in the Wild

Zhe Zhang[1†] · **Chunyu Wang**[2†] · **Weichao Qiu**[3] · **Wenhu Qin**[1*] · **Wenjun Zeng**[2]

Received: date / Accepted: date

Abstract Occlusion is probably the biggest challenge for human pose estimation in the wild. Typical solutions often rely on intrusive sensors such as IMUs to detect occluded joints. To make the task truly unconstrained, we present *AdaFuse*, an adaptive multiview fusion method, which can enhance the features in occluded views by leveraging those in visible views. The core of *AdaFuse* is to determine the point-point correspondence between two views which we solve effectively by exploring the sparsity of the heatmap representation. We also learn an adaptive fusion weight for each camera view to reflect its feature quality in order to reduce the chance that good features are undesirably corrupted by "bad" views. The fusion model is trained end-to-end with the pose estimation network, and can be directly applied to new camera configurations without additional adaptation. We extensively evaluate the approach on three public datasets including Human3.6M, Total Capture and CMU Panoptic. It outperforms the state-of-the-arts on all of them. We also create a large scale synthetic dataset *Occlusion-Person*, which allows us to perform numerical evaluation on the occluded joints, as it provides occlusion labels for every joint in the images. The dataset and code are released at https://github.com/zhezh/adafuse-3d-human-pose.

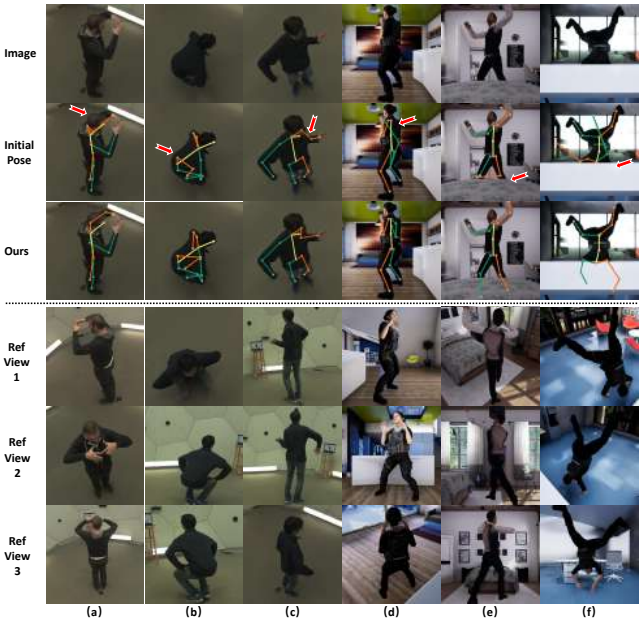

Fig. 1 Our approach accurately detects the poses even though they are occluded by leveraging the features in other views. The bottom three rows are images from other view angles of the scene for readers to better perceive the 3D poses of the actors.

Keywords Human pose estimation · Multiple camera fusion · Epipolar geometry

Zhe Zhang
E-mail: zhangzhecnjs@gmail.com

Chunyu Wang
E-mail: chnuwa@microsoft.com

Weichao Qiu
E-mail: qiuwc@gmail.com

Wenhu Qin
E-mail: qinwenhu@seu.edu.cn

Wenjun Zeng
E-mail: wezeng@microsoft.com

[1] Southeast University, Nanjing, China
[2] Microsoft Research Asia, Beijing, China
[3] The Johns Hopkins University, MD, USA
[*] Corresponding Author
[†] Zhe Zhang and Chunyu Wang have contributed equally. Work done when Zhe Zhang is an intern at Microsoft Research Asia

1 Introduction

Accurately estimating 3D human pose from multiple cameras has been a longstanding goal in computer vision (Liu et al., 2011; Bo and Sminchisescu, 2010; Gall et al., 2010; Rhodin et al., 2018; Amin et al., 2013; Burenius et al., 2013; Pavlakos et al., 2017; Belagiannis et al., 2014). The ultimate goal is to recover absolute 3D locations of the body joints in a world coordinate system from multiple cameras placed in natural environments. The task has attracted a lot of attention because it can benefit many applications such as augmented and virtual reality (Starner et al., 2003), human-computer-interaction and intelligent player analysis in sport videos (Bridgeman et al., 2019).

The task is often addressed by a simple two-step framework. In the first step, it tries to detect the 2D poses in all camera views, for example, by a convolutional neural network (Cao et al., 2017; Xiao et al., 2018). Then in the second step, it recovers the 3D pose from the multiview 2D poses either by analytical methods (Burenius et al., 2013; Pavlakos et al., 2017; Belagiannis et al., 2014; Qiu et al., 2019; Amin et al., 2013) or by discriminative models (Iskakov et al., 2019; Tu et al., 2020). The camera parameters are usually assumed known in these approaches. The development of powerful network architectures such as (Newell et al., 2016) has notably improved the 2D pose estimation quality, which in turn reduces the 3D error remarkably. For example, in (Qiu et al., 2019), the 3D error on Human3.6M (Ionescu et al., 2014) decreases significantly from 52mm to 26mm.

However, obtaining small errors on benchmark datasets does not imply that the task has been truly solved unless the challenges such as background clutter, human appearance variation and occlusion encountered in real world applications are well addressed. In fact, a growing amount of efforts (Zhou et al., 2017; Ci et al., 2019; Yang et al., 2018; Rogez and Schmid, 2016; Pavlakos et al., 2018; Ci et al., 2020) have been devoted to improving the pose estimation performance in challenging scenarios, for example, by augmenting the training dataset (Zhou et al., 2017; Yang et al., 2018; Varol et al., 2017) with more images or by using more robust sensors such as IMUs (Trumble et al., 2017). We will discuss about this type of work in more details in section 2.

In this work, we propose to solve the problem in a different way by multiview feature fusion. The approach is orthogonal to the previous efforts. As shown in Figure 1, our approach can accurately detect the joints even when they are occluded in certain views. The motivation behind our approach is that a joint occluded in one view may be visible in other views. So it is generally helpful to fuse the features at the corresponding locations in different views. To that end, we present a flexible multiview fusion approach termed *AdaFuse*. Figure 2 shows the pipeline. It first uses camera parameters to compute the point-line correspondence between a pair of views. Then it "finds" the matched point on the line by exploring the sparsity of the heatmap representation without performing the challenging point-point matching. Finally, the features of the matched points in different views are fused. The approach can effectively improve the feature quality in occluded views. In addition, for a new environment with different camera poses, we can directly use *AdaFuse* without re-training as long as the camera parameters are available. This improves the applicability of the approach in real applications.

The performance of *AdaFuse* is further boosted by learning an adaptive fusion weight for each view to reflect its feature quality. This weight is leveraged in fusion in order to reduce the impact of low-quality views. If a joint is occluded in one view, its features are also likely corrupted. In this case, we hope to give a small weight to this view when performing multiview fusion such that the high-quality features in the visible views are dominant, and are free from being corrupted by low-quality features. We add some simple layers to the pose estimation network to predict heatmap quality based on the heatmap distribution and cross view consistency. We observe in our experiments that the use of adaptive fusion notably improves the performance.

We evaluate our approach on three public datasets including Human3.6M (Ionescu et al., 2014), Total Capture (Trumble et al., 2017) and CMU Panoptic (Joo et al., 2019). It outperforms the state-of-the-arts demonstrating the effectiveness of our approach. In addition, we also compare it to a number of standard multiview fusion methods such as RANSAC in order to give more detailed insights. We evaluate the generalization capability of our approach by training and testing on different datasets. We also create a synthetic human pose dataset in which human are purposely occluded by objects. The dataset allows us to perform evaluation on the occluded joints.

The rest of the paper is organized as follows. In section 2, we discuss the related work on multiview 3D human pose estimation with special focus on the approaches that aim to improve the performance in challenging environments. Section 3 introduces the basics for multiview feature fusion to lay the groundwork for *AdaFuse*. Then we describe how we learn adaptive weight for each camera view to reflect the feature quality, as well as the details of *AdaFuse*. In sections 5 and 6, we introduce the experimental datasets and results, respectively. Section 7 concludes this work.

2 Related Work

We first review the related work on multiview 3D human pose estimation in section 2.1. Then section 2.2 summarizes the techniques that are used to improve the in-the-wild performance. Finally, in section 2.3, we discuss the approaches on consensus learning such as RANSAC. This is necessary

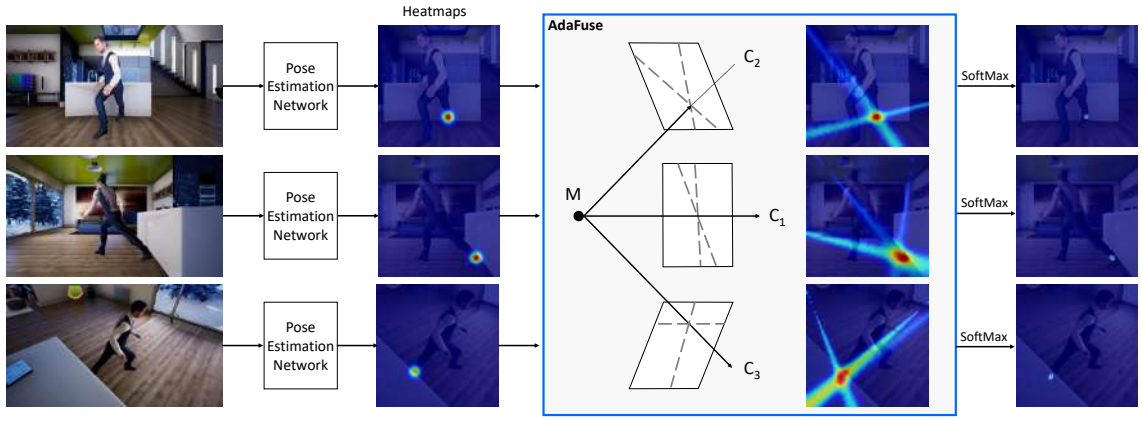

Fig. 2 Overview of *AdaFuse*. It takes multiview images as input and outputs 2D poses of all views jointly. It first uses a pose estimation network to obtain 2D heatmaps for each view. Then on top of epipolar geometry, the heatmaps from all camera views are fused. Finally, we apply the SoftMax operator to suppress the small noises introduced in fusion. Consequently, pose estimation in each view benefits from other views.

for multiple sensor fusion because the sensors could have contradictory predictions and the outliers should be removed to ensure the good fusion quality.

2.1 Multiview 3D Human Pose Estimation

We briefly classify the multiview 3D human pose estimation methods into two classes. The first class is model-based approaches which are also known as analysis-by-synthesis approaches (Liu et al., 2011; Gall et al., 2010; Moeslund et al., 2006; Sigal et al., 2010; Perez et al., 2004). They first model human body by simple primitives such as sticks and cylinders. Then the parameters of the model (*i.e.* poses) are continuously updated according to the observations in multiview images until the model can be explained by the image features. The resulted optimization problem is usually non-convex. So expensive sampling techniques are often used. The main difference among those approaches lies in the adopted image features and the optimization algorithms. We refer the interested readers to earlier survey papers such as (Moeslund et al., 2006).

The advantage of the model-based approaches lies in its capability to handle occlusion because of the inherent structure prior embedded in human model. These approaches aggregate the local features as evidence to infer the global model parameters with the inherent human body structure as constraints. So if a joint is occluded, it can still rely on other joints to guess the possible locations that are consistent with the prior. However, the model-based approaches get larger 3D errors than the model-free approaches due to the difficult optimization problems.

The second class is model-free approaches (Qiu et al., 2019; Iskakov et al., 2019; Burenius et al., 2013; Pavlakos et al., 2017; Dong et al., 2019; Amin et al., 2013; Belagiannis et al., 2014; Xie et al., 2020) which often follow a two-step framework. They first detect 2D poses in images of all camera views. Then with the aid of camera parameters, they recover the 3D pose using either triangulation (Amin et al., 2013; Iskakov et al., 2019) or pictorial structure models (Burenius et al., 2013; Pavlakos et al., 2017; Dong et al., 2019). Recursive pictorial structure model is introduced in (Qiu et al., 2019) to speed up the inference process. The authors in (Iskakov et al., 2019) also propose to use learnable triangulation (Hartley and Zisserman, 2003) for human pose estimation which is more robust to inaccurate 2D poses. If the 2D poses are accurate, the recovered 3D poses are guaranteed to be accurate without worrying about being trapped in local optimum as the model-based methods.

The development of more powerful network architectures (Newell et al., 2016; Sun et al., 2019) has dramatically improved the 2D pose estimation accuracy on benchmark datasets, which in turn also decreases the 3D pose estimation error. For example, on the most popular benchmark Human3.6M (Ionescu et al., 2014), the 3D MPJPE error has decreased to about 20mm which can meet the requirements of many real-life applications.

2.2 Improving "In the Wild" Performance

Sensors Occlusion is probably the biggest challenge for in-the-wild scenarios. One straightforward solution is to use additional sensors such as IMUs (Trumble et al., 2017) and radio signals (Zhao et al., 2019), which are not impacted by occlusion. For example, Roetenberg *et al*. (Roetenberg et al., 2009) place 17 IMUs at the rigid bones. If the measurements are accurate, the 3D pose is fully determined. In practice, however, the accuracy is limited by the drifting problem. To that end, some approaches (Trumble et al., 2017; von Marcard et al., 2018; Gilbert et al., 2019; Malleson et al., 2017; Zhang et al., 2020) propose to fuse images and IMUs to achieve more robust pose estimation. Some works (Zhao et al., 2019; Li et al., 2019; Zhao et al., 2018) leverage the

fact that wireless signals in the WiFi frequencies traverse walls and reflect off the human body, and propose a radio-based system that can estimate 2D poses even when persons are completely occluded by walls. However, these approaches also have their own problems. For example, how to effectively fuse visual and inertial signals for IMU-based approaches? Besides, wearing sensors on the body is intrusive, and is not acceptable in some scenarios such as football games. On the other hand, the WiFi-based solutions cannot deal with self-occlusion which is a big limitation.

Data Augmentation Collecting more images for model training is an effective approach to improve the generalization performance. For example in (Zhou et al., 2017; Qiu et al., 2019), the authors propose to use the MPII (Andriluka et al., 2014) and the COCO (Lin et al., 2014) datasets to help train the 2D module of the 3D pose estimators which effectively reduces the risk of over-fitting to simple training datasets. However, annotating a sufficiently large pose dataset is expensive and time consuming. So some approaches (Rogez and Schmid, 2016; Varol et al., 2017; Hoffmann et al., 2019; Chen et al., 2016; Lassner et al., 2017) propose to generate synthetic images. The main issue is to bridge the gap between the synthetic and real images such that the model trained on synthetic images can be applied to real images. To that end, some approaches such as (Peng et al., 2018) propose to use generative adversarial networks to generate realistic images.

Spatial-Temporal Context Models Some approaches propose to use spatial-temporal context models to jointly detect all joints in a video sequence such that each joint can benefit from other joints in the same or neighboring frames. Intuitively, if a body joint is occluded thus is difficult to be detected according to its own appearance, they can use the locations of other joints to guess the possible location. For example, in a previous work (Cao et al., 2017; Kreiss et al., 2019), the authors propose to detect body parts, *i.e*. the links connecting two joints, in addition to the individual joints. This provides a chance to mutually enhance the detection of the two linked joints. In (Cheng et al., 2019; Pavllo et al., 2019), temporal convolution is utilized to deal with occlusion in current frames. Some works such as (Qiu et al., 2019) propose to establish the spatial correspondence across multiple camera views, and leverage multi-view features for robust joint detection. Significant performance improvement has been achieved for the occluded joints on several benchmark datasets. The main drawback of the approach (Qiu et al., 2019) is the lack of flexibility in practice since it needs to train a separate fusion network for every possible camera placement. Our work differs from (Qiu et al., 2019) in that it can be applied to new environments with different numbers of cameras and different camera poses without additional adaptation. We will compare the two methods in the experiments.

2.3 Consensus Learning

A fundamental problem in multi-sensor fusion is to detect and remove outliers as the sensors may produce inconsistent measurements. RANSAC (Fischler and Bolles, 1981) is the most commonly used outlier detection method. The main assumption is that the dataset consists of inliers. It produces reasonable results only with a certain probability which increases as the number of inliers. In practice, when the number of sensors is small, the probability of detecting the real outliers is also small. For example, in multi-view human pose estimation, the number of cameras is only four to eight for most benchmark datasets (Ionescu et al., 2014; Trumble et al., 2017). For such cases, we observe that RANSAC may not be the best option.

In recent years, uncertainty learning (Kendall and Gal, 2017; Gal and Ghahramani, 2015; Lakshminarayanan et al., 2017; Zafar et al., 2019; Lakshminarayanan et al., 2017; Pleiss et al., 2017) has attracted a lot of attention which is particularly important for high-risk applications such as autonomous driving and medical diagnosis (Gal, 2016; Ghahramani, 2016). The main idea is that, when a model makes a prediction, it also outputs a score reflecting the confidence of the prediction. Consider an autonomous car that uses a neural network to detect people. If the network is not confident about the prediction, the car could probably rely on other sensors for making the correct decision. Uncertainty is introduced to computer vision in (Kendall and Gal, 2017; Kreiss et al., 2019; He et al., 2019; Ilg et al., 2018). Another branch of approaches such as (Guo et al., 2017; Pleiss et al., 2017) propose to learn uncertainty by calibration. They propose to train the model such that the probability associated with the predicted class label agrees with its ground truth correctness likelihood.

The concept of uncertainty can be leveraged to reduce the impact of outliers. For example, in (Iskakov et al., 2019), the authors propose to predict an uncertainty score for each joint in each view. The score is used to weigh each view when doing triangulation. This dramatically reduces the 3D pose estimation error. Inspired by the success of uncertainty learning in computer vision tasks, we propose to learn uncertainty for multiview feature fusion. The predicted uncertainty is used as a weight when fusing multiview features. We show this adaptive feature fusion could effectively improve the fusion quality.

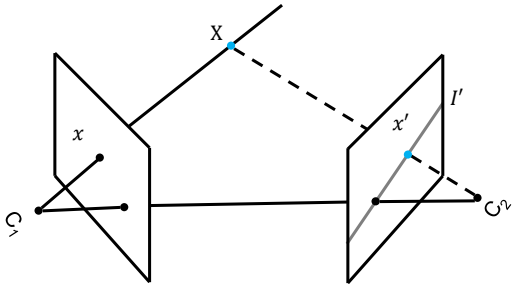

Fig. 3 Illustration of the point-line correspondence in two views. For an arbitrary point $\mathbf{x}$ in one view, the corresponding point $\mathbf{x}'$ in another view has to lie on the epipolar line $\mathbf{l}'$. This is the core of *AdaFuse* for finding corresponding points in other views.

Fig. 4 Epipolar geometry based heatmap fusion. For each location $\mathbf{x}$ in the first view, we first compute the corresponding epipolar lines in the other two views. Then we find the largest responses on the two lines, respectively and add them to the original response at $\mathbf{x}$.

3 The Basics for Multiview Fusion

We first introduce the basics for multiview fusion to lay the groundwork for *AdaFuse*. In particular, we discuss how to establish the point-point correspondence between two views such that the features correspond to the same 3D space point can be fused together. The narrow baseline correspondence can be solved efficiently by local feature matching. However, in the context of multiview human pose estimation where only a small number of cameras are placed far away from each other, the local features cannot be robustly detected and matched especially for texture-less human regions. This poses a serious challenge.

To solve the problem, we present a coarse-to-fine approach to find matched points. It first establishes the point-to-line correspondence between two views by epipolar geometry, and then implicitly determine the point-to-point correspondence by exploring the sparsity of the heatmap representations. The approach notably simplifies the task because it avoids the challenging step of finding the exact correspondence. We first introduce epipolar geometry in section 3.1 in order to determine the point-to-line correspondence. Then in section 3.2, we describe how we adapt epipolar geometry to perform multiview heatmap fusion. Finally, we discuss the side effect caused by the simplified fusion strategy and our solution in section 3.3.

3.1 Epipolar Geometry

Let us denote a point in 3D space as $\mathbf{X} \in \mathcal{R}^{4\times1}$ as shown in Figure 3. This could be the location of a body joint in the context of pose estimation. Note that homogeneous coordinate and column vector are used to represent a point. The 3D point is imaged in two camera views, at $\mathbf{x} = \mathbf{P}\mathbf{X}$ in the first, and $\mathbf{x}' = \mathbf{P}'\mathbf{X}$ in the second, where $\mathbf{x}$ and $\mathbf{x}' \in \mathcal{R}^{3\times1}$ represent 2D points in images, $\mathbf{P}$ and $\mathbf{P}' \in \mathcal{R}^{3\times4}$ are the projection matrix for each camera. Since the two 2D points correspond to the same 3D point and have the same semantic meanings, their features can be safely fused such that each view benefits from the other view.

The epipolar geometry (Hartley and Zisserman, 2003) between two views is essentially the geometry of the intersection of the image planes with the pencil of planes having the baseline as axis. The baseline is the line joining the camera centers C_1 and C_2. In particular, for each location $\mathbf{x}$ in the first view, it helps us to determine the location of the corresponding point $\mathbf{x}'$ in the second view without having to know $\mathbf{X}$.

We can see from Figure 3 that the image points $\mathbf{x}$ and $\mathbf{x}'$, the 3D point $\mathbf{X}$, and the camera centers $\mathbf{C_1}$ and $\mathbf{C_2}$ lie on the same plane π. The plane intersects with the two image planes at epipolar lines $\mathbf{l}$ and $\mathbf{l}'$, respectively. In particular,

$$\begin{aligned} \mathbf{l}' &= \mathbf{F}\mathbf{x} \\ \mathbf{l} &= \mathbf{F}^\top \mathbf{x}', \end{aligned} \tag{1}$$

where $\mathbf{F} \in \mathcal{R}^{3\times3}$ is fundamental matrix which can be derived from $\mathbf{P}$ and $\mathbf{P}'$. Readers can refer to (Hartley and Zisserman, 2003) for detail derivation. In addition, the rays back-projected from $\mathbf{x}$ and $\mathbf{x}'$ intersect at $\mathbf{X}$, and the rays are coplanar, lying in π. It is straightforward to derive that the location of $\mathbf{x}'$ which corresponds to $\mathbf{x}$ is guaranteed to lie on the epipolar line $\mathbf{l}'$. However, we have to leverage additional information such as appearance to determine the exact location of $\mathbf{x}'$ on $\mathbf{l}'$.

In the context of multiview feature fusion, for every image point $\mathbf{x}$, we need to find the corresponding point $\mathbf{x}'$ in the second view so that we can fuse the features at $\mathbf{x}$ with those at $\mathbf{x}'$ and obtain more robust pose estimations. Since we do not know the depth of $\mathbf{X}$, it could move freely on the line defined by the camera center $\mathbf{C_1}$ and image point $\mathbf{x}$. However, we know that $\mathbf{x}'$ cannot span the entire image plane but is restricted to the line $\mathbf{l}'$. In the following section 3.2, we will describe how we perform multiview feature fusion based on epipolar geometry.

Sampson Distance In practice, usually we have 2D measurements $\mathbf{x}$ and $\mathbf{x}'$ corresponding to the same 3D location $\mathbf{X}$ which is unknown. Due to measurement noise and errors,

the line $\mathbf{C_1}\mathbf{x}$ and $\mathbf{C_2}\mathbf{x}'$ might not intersect exactly at location $\mathbf{X}$. To obtain the optimal estimation for $\mathbf{X}$, we search for $\hat{\mathbf{X}}$ subject to

$$d^2_{Reproj} = \min_{\hat{\mathbf{X}}} d^2\left(\mathbf{x}, \mathbf{P}\hat{\mathbf{X}}\right) + d^2\left(\mathbf{x}', \mathbf{P}'\hat{\mathbf{X}}\right), \tag{2}$$

where $d(\cdot)$ denotes Euclidean distance, d_{Reproj} represents the reprojection distance between $\mathbf{x}$ and $\mathbf{x}'$. Since there is optimization process when obtaining d_{Reproj}, we adopt an one-step method which is its first-order approximation (Hartley and Zisserman, 2003). This approximation is also called Sampson distance as

$$d_{Sampson} = \frac{\mathbf{x}'^{\top}\mathbf{F}\mathbf{x}}{(\mathbf{F}\mathbf{x})_1^2 + (\mathbf{F}\mathbf{x})_2^2 + (\mathbf{F}^{\top}\mathbf{x}')_1^2 + (\mathbf{F}^{\top}\mathbf{x}')_2^2}, \tag{3}$$

where $\mathbf{F}$ is fundamental matrix, the subscript 1 or 2 denotes the first or second element of a vector. By using Sampson distance, we can directly obtain distance between a pair of locations without knowing the intermediate $\hat{\mathbf{X}}$. In *AdaFuse*, we use Sampson distance to represent to what extent a pair of 2D joint detections support each other.

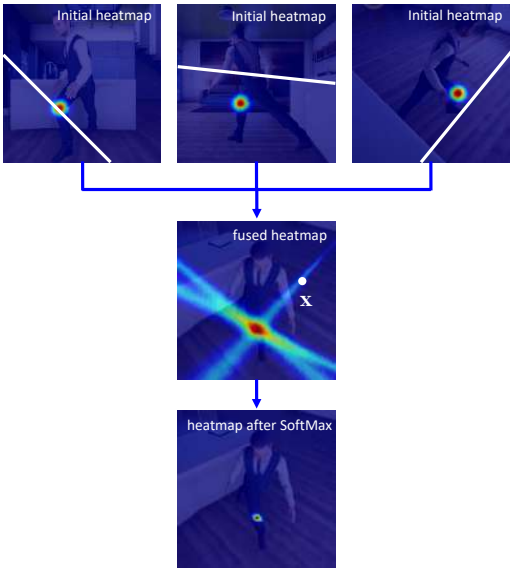

Fig. 5 The ambiguity problem in our simplified multiview fusion approach and our solution. We can see from the "fused heatmap" that the correct location has the largest response which is as expected. However, for an incorrect location $\mathbf{x}$, there is also a chance that the response is also enhanced by at most one view. Fortunately, the correct location will be enhanced more times (three times in this example) leading to the largest response. So we apply the SoftMax operator to the fused heatmap to reduce the responses at incorrect locations.

3.2 Heatmap Fusion

Multiview fusion is applied to heatmaps rather than intermediate features as shown in Figure 2. This is because heatmap has the nice property of sparsity which can simplify the point-point matching. A heatmap produces a per-pixel likelihood for joint locations in the image. Specifically, it is generated as a two-dimensional Gaussian distribution centered at the coordinate of the joint. So it has a small number of large responses near the joint location, and a large number of zeros at other locations. See Figure 4 (a) for an example heatmap of the right knee joint.

The sparse heatmaps allow us to safely skip the exact point-point matching because the features at the "zero" locations on the epipolar line are not contributing to the feature fusion. As a result, instead of trying to find the exact corresponding location in the other view, *we simply select the largest response on the epipolar line as the matched point*. This is a reasonable simplification because the corresponding point usually has the largest response. For example, in Figure 4, for each location $\mathbf{x}$, we first compute the corresponding epipolar lines in the other two camera views. Then we find the largest responses on the two epipolar lines, respectively and fuse them with the response at $\mathbf{x}$.

Let us denote the heatmap in view v as $\mathbf{H}^v$. The response at the location $\mathbf{x}$ of the heatmap is denoted as $\mathbf{H}^v(\mathbf{x})$. The corresponding epipolar line of $\mathbf{x}$ in view u is denoted as $\mathbf{I}^u(\mathbf{x})$ which consists of a number of discrete locations on the heatmap $\mathbf{H}^u$. The epipolar line can be analytically computed based on the camera parameters for every location $\mathbf{x}$. Then we formulate multiview fusion as

$$\hat{\mathbf{H}}^v(\mathbf{x}) = \lambda\mathbf{H}^v(\mathbf{x}) + \frac{1-\lambda}{N}\sum_{u=1}^{N}\max_{\mathbf{x}'\in\mathbf{I}^u(\mathbf{x})}\mathbf{H}^u(\mathbf{x}'), \tag{4}$$

where $\hat{\mathbf{H}}$ denotes the fused heatmap and N is the number of camera views which contribute to the fusion of current view. The parameter λ balances the responses in the current and other views.

3.3 Side Effect and Solution

One side effect caused by the simplified fusion model (*i.e.* Eq. (4)) is that some background locations may be enhanced undesirably. We visualize an example in the second row of Figure 5. We can see that many background pixels, for example $\mathbf{x}$, have non-zero responses which are caused by fusion. This phenomenon happens because multiple epipolar lines (in other views) may pass the ground truth joint location which has large responses, and some of the epipolar lines actually correspond to background pixels in the current view. This is explained in Figure 5. For a location $\mathbf{x}$ in the current view, the corresponding epipolar lines in the other three views are drawn in the first row. We can see that although $\mathbf{x}$ is not at a meaningful joint location, the epipolar

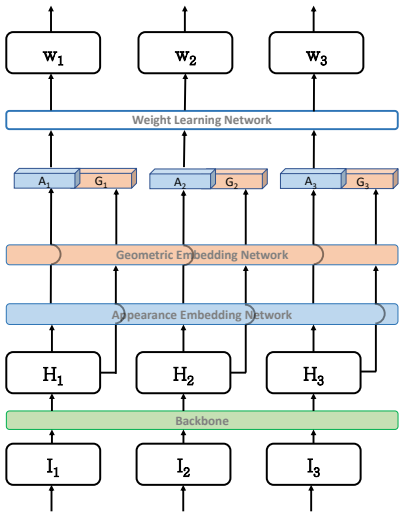

Fig. 6 Network for learning adaptive fusion weights. The backbone network for pose estimation is used to extract heatmaps $\mathbf{H}_v$ for each view $\mathbf{I}_v$. The heatmaps are fed to *appearance embedding network* and *geometry embedding network*, respectively, to extract features, which are concatenated and fed to a *weight learning network* to learn the fusion weights which reflect the heatmap quality in each view. The weights are used for multiview fusion.

line in the first view passes the ground truth knee joint and leads to a large unexpected response for $\mathbf{x}$.

Fortunately, there are patterns for the background pixels that could be undesirably impacted. In general, the pixels that are impacted by a high response location in another view are guaranteed to lie on the same line. More importantly, the lines that correspond to different views do not overlap. It means, for a location $\mathbf{x}$ in the background, its response can only be enhanced by at most one view. In contrast, the location which corresponds to meaningful body joints will be enhanced by multiple views. In other words, the correct location is guaranteed to have the largest response for general cases. So we take advantage of this observation and directly apply the SoftMax operator to remove the small responses. See the third row in Figure 5 for the effect. We can see that only the large responses around the joint location are preserved.

3.4 Implementation Details

It is worth noting that the above fusion method does not have learnable parameters. So we only need to train the backbone network such as SimpleBaseline (Xiao et al., 2018) to estimate pose heatmaps. The loss function for training the backbone network is defined as MSE loss between the estimated heatmaps and ground truth heatmaps. In the testing stage, given the heatmaps estimated by SimpleBaseline, we fuse them deterministically by our approach.

4 Adaptive Weight for Multiview Fusion

The fusion strategy introduced in the previous section treats all views evenly without considering the feature quality of each view. Note that the fusion weight is $\frac{1-\lambda}{N}$ for the N views in Eq. (4). However, the strategy is problematic in some cases where the heatmaps of some camera views are incorrect. This is because those features may undesirably mess up the features in good views, leading to a completely incorrect 2D pose estimation results.

To solve this problem, we present a weight learning network to learn an *adaptive weight* for each view to faithfully reflect its heatmap quality. It takes inputs of the heatmaps of N-views extracted by the pose estimation network, and regresses N weights ω^u. Then multiview fusion is rewritten to consider the weights as follows

$$\hat{\mathbf{H}}^v(\mathbf{x}) = \omega^v \mathbf{H}^v(\mathbf{x}) + \sum_{u=1}^{N} \omega^u \max_{\mathbf{x}' \in \mathbf{I}^u(\mathbf{x})} \mathbf{H}^u(\mathbf{x}'), \tag{5}$$

The prediction of the adaptive fusion weight ω is implemented by a lightweight neural network as shown in Figure 6. On top of the heatmaps $\mathbf{H}$ provided by the pose estimation network, we extract two types of information for making the prediction. The first is the appearance embedding which extracts information such as the distribution characteristics of the heatmaps. The second is the geometry embedding which considers the cross-view location consistency. The two terms are complementary to each other. The proposed weight learning network can be joined with the pose estimation network for end-to-end training without enforcing supervision on the weights.

4.1 The Appearance Embedding

The heatmap of each joint actually contains rich information to infer its heatmap quality. For example, if the predicted heatmap has a desired shape of Gaussian kernel, then in many cases, the heatmap quality is good. In contrast, if the predicted heatmap has random and small responses all over the space (for example, when the joint is occluded), then the quality is likely to be bad.

We propose a simple network to extract appearance embeddings for each joint in each camera view. Figure 7 shows the network structure. Starting from the heatmaps $\mathbf{H}_i$, we apply a convolutional layer to extract features. Then the features are down-sampled by average pooling and fed to a Fully Connected (FC) layer for extracting the appearance embeddings. Different joint types and camera views share the same weights. We only show the network for a single view and a single joint for simplicity. The appearance embedding network is jointly learned end-to-end with the pose estimation network.

Fig. 7 The appearance embedding network for predicting the fusion weight. i is the index of camera views. The parameters in the network are shared for all views and joints. See also Figure 6 for how the appearance embedding A_i is used for determining the fusion weight.

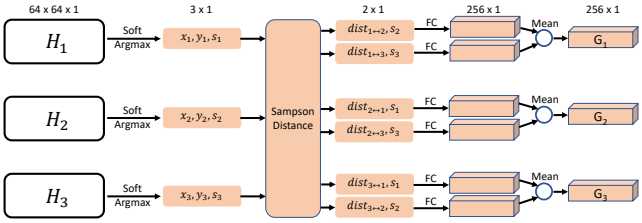

Fig. 8 The geometry embedding network for predicting the fusion weight. For each joint in each camera view (three views are shown in this example), it generates a 256-dimensional embedding to reflect the heatmap (pose) quality. Note that the FC is shared for all branches.

4.2 The Geometry Embedding

The appearance embedding alone is not sufficient for some challenging cases where the heatmaps have the desired shape of Gaussian kernel but at the wrong locations. One such example is when the left knee is detected at the location of right knee which is usually known as the "double counting" problem to the community. To solve this problem, we propose to leverage the location consistency information among all camera views. Our core motivation is that the predicted joint location in one camera view is more reliable if it agrees with the locations in other views.

We implement this idea by a geometry embedding network as shown in Figure 8. Starting from the heatmaps $\mathbf{H}$, we first apply the "soft-argmax" operator (Sun et al., 2018) to obtain the location (x, y) of the joint in each view. We also get the heatmap response value s in that location to reflect its confidence. Then we compute the Sampson distance (Hartley and Zisserman, 2003) $dist_{i\leftrightarrow j}$ between the current view and other views to measure the correspondence or consistency error. A small $dist_{i\leftrightarrow j}$ means the joint locations in the two views are consistent. Intuitively, the location that is consistent with most views is more reliable. Finally, we propose to use a FC layer to embed the Sampson distance into a feature vector. The feature vectors of all camera pairs are then averaged to obtain the final geometry embedding.

4.3 Weight Learning Network

We propose a simple network consisting of three FC layers to transform the concatenated appearance and geometric embeddings to regress the final weight. It is worth noting that we do not train the weight learning network independently. Instead, we join it with the pose estimation network to minimize the fused 2D heatmap loss without enforcing intermediate supervision on the fusion weights. The first column in Figure 9 shows some example weights predicted by our approach. We can see that when the joints are occluded, and are localized at incorrect locations, the corresponding fusion weights are indeed smaller than other joints.

5 Datasets and Metrics

We introduce the three datasets used for evaluation and the corresponding metrics. We also describe how we construct the synthetic person dataset *Occlusion-Person* which has a large amount of human-object occlusion.

Table 1 The statistics of the public multiview pose estimation datasets. Only the *Occlusion-Person* dataset provides occlusion labels.

Dataset	Frames	Cameras	Occluded Joints
Human3.6M	784k	4	-
Total Capture	236k	8	-
Panoptic	36k	31	-
Occlusion-Person	73k	8	20.3%

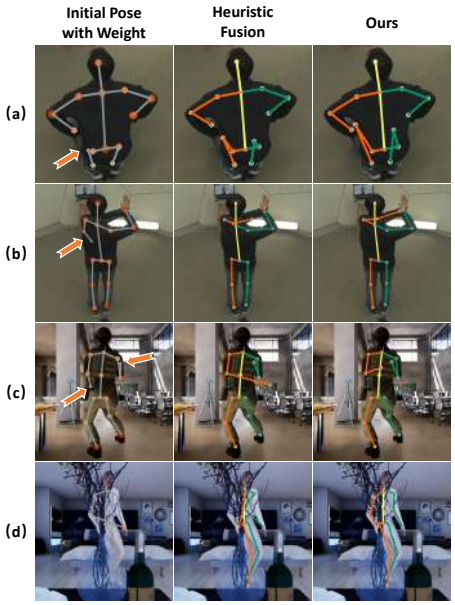

Fig. 9 We visualize the predicted fusion weights by the size of the markers in the first column. A large marker denotes a larger weight. The rest two columns show the poses estimated by *HeuristicFuse* and *AdaFuse*, respectively. Our *AdaFuse* has clearly better estimations due to the consideration of the feature quality in every view.

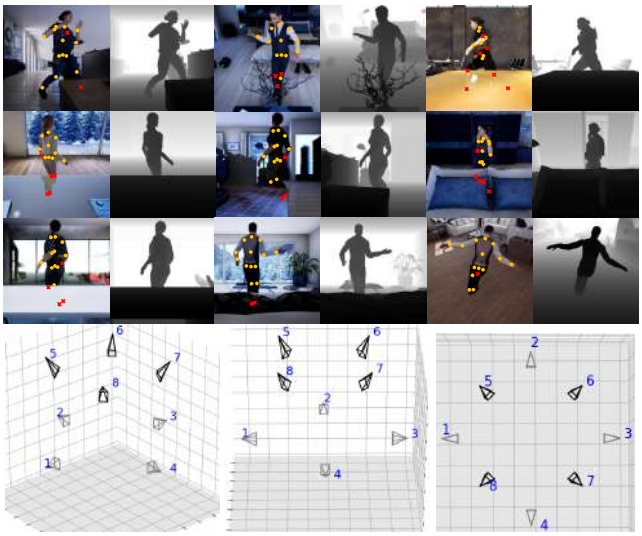

Fig. 10 We show some typical images, ground-truth 2D joint locations and the depth maps from the *Occlusion-Person* dataset. The joint represented by red "x" means it is occluded. The **bottom** row shows spacial configuration of the eight cameras used in the dataset from different view angles.

5.1 Datasets

The Human3.6M Dataset (Ionescu et al., 2014) It provides synchronized images captured by four cameras. There are seven subjects performing daily actions. We use a cross-subject evaluation scheme where subjects $1, 5, 6, 7, 8$ are used for training and $9, 11$ for testing. We also use the MPII dataset (Andriluka et al., 2014) to augment the training data to avoid over-fitting to the simple background. Since the MPII dataset provides only monocular images, we only train the backbone network before multiview fusion.

The Total Capture Dataset (Trumble et al., 2017) It provides synchronized person images captured by eight cameras. Following the dataset convention, the training set consists of "ROM1,2,3", "Freestyle1,2", "Walking1,3", "Acting1,2" and "Running1" on subjects 1, 2 and 3. The testing set consists of "Freestyle3 (**FS3**)", "Acting3 (**A3**)" and "Walking2 (**W2**)" on subjects 1,2,3,4 and 5.

The CMU Panoptic Dataset (Joo et al., 2019) This recently introduced dataset provides images captured by dozens of cameras. We uniformly select six cameras to evaluate the impact of the number of cameras on 3D pose estimation. In particular, the cameras 1, 2, and 10 are firstly selected to construct a 3-view experiment setting. Then the cameras 13, 3 and 23 are sequentially added to the previous three cameras to construct a four, five and six view experiment setting, respectively. We follow the practice of the previous work (Xiang et al., 2019) to select the training and testing sequences which consist of only one person. Since few works have reported numerical results on this dataset, we only compare our approach to the baselines.

The Occlusion-Person Dataset The previous benchmarks do not provide occlusion labels for the joints in images which prevents us from performing numerical evaluation on the occluded joints. In addition, the amount of occlusion in the benchmarks is limited. To address the limitations, we propose to construct this synthetic dataset *Occlusion-Person*. We adopt UnrealCV (Qiu et al., 2017) to render multiview images and depth maps from 3D models. In particular, thirteen human models of different clothes are put into nine different scenes such as living rooms, bedrooms and offices. The human models are driven by the poses selected from the CMU Motion Capture database. We purposely use objects such as sofas and desks to occlude some body joints. Eight cameras are placed in each scene to render the multiview images and the depth maps. The eight cameras are placed evenly every 45 degree on a circle of two meters radius at about 0.9 and 2.3 meters high, respectively. We provide the 3D locations of 15 joints as ground truth. Figure 10 shows some sample images from the dataset and spacial configuration of the cameras.

The occlusion label for each joint in an image is obtained by comparing its depth value (available in the depth map), to the depth of the 3D joint in the camera coordinate system. If the difference between the two depth values is smaller than 30cm, then the joint is not occluded. Otherwise, it is occluded. Table 1 compares this dataset to the existing benchmarks. In particular, about 20% of the body joints are occluded in our dataset. We use 75% of the dataset for training and 25% for validation.

5.2 Metrics

2D Metrics The Percentage of Correct Keypoints (PCK) metric introduced in Andriluka et al. (2014) is commonly used for 2D pose evaluation. PCKh@t measures the percentage of the estimated joints whose distance from the ground-truth joints is smaller than t times of the head length. Following the previous works, we report results when t is $\frac{1}{2}$. Since the head length is not provided in the used three benchmarks, we approximately set it to be 2.5% of the human bounding box width for all benchmarks.

3D Metrics The 3D pose estimation accuracy is measured by Mean Per Joint Position Error (MPJPE) between a ground truth 3D pose $y = [p_1^3, \cdots, p_M^3]$ and an estimated 3D pose $\bar{y} = [\bar{p_1^3}, \cdots, \bar{p_M^3}]$: MPJPE $= \frac{1}{M}\sum_{i=1}^{M} \|p_i^3 - \bar{p_i^3}\|_2$ where M is the number of joints in a pose. We do not align the estimated 3D poses to the ground truth by Procrustes. This is referred to as protocol 1 in some works (Martinez et al., 2017; Tome et al., 2018)

Table 2 The 2D pose estimation accuracy (PCKh@t) of the baseline methods and our approach on the Human3.6M dataset. We report results for each individual joint and the average over all joints.

Methods	Root	Belly	Neck	Nose	Head	Hip	Knee	Ankle	Shlder	Elbow	Wrist	*Mean*
NoFuse	95.8	77.1	60.4	86.4	86.2	79.3	81.5	58.6	65.1	78.3	70.1	74.8
HeuristicFuse	96.0	79.3	60.7	**88.4**	**86.8**	83.1	84.5	60.0	**66.9**	82.1	75.2	77.3
ScoreFuse	96.2	79.3	61.6	88.3	86.2	83.3	84.3	60.5	66.6	83.1	77.4	77.8
AdaFuse (Ours)	**96.2**	**79.3**	**61.6**	88.3	86.3	**83.5**	**86.4**	**61.1**	66.7	**86.0**	**80.1**	**78.8**

Table 3 The 3D pose estimation error (mm) of the baseline methods and our approach on the Human3.6M dataset.

Methods	Belly	Neck	Nose	Head	Hip	Knee	Ankle	Shlder	Elbow	Wrist	Mean
NoFuse	21.6	16.8	15.7	11.3	**17.8**	25.8	35.8	22.0	26.8	34.1	22.9
HeuristicFuse	21.6	16.8	15.7	11.0	17.9	23.0	32.7	21.9	25.0	25.7	21.0
ScoreFuse	21.4	16.7	15.8	10.9	18.3	21.3	30.8	21.8	23.3	23.2	20.1
RANSAC	21.6	16.8	**15.7**	11.2	17.9	23.9	34.6	22.0	25.8	28.2	21.8
AdaFuse (Ours)	**21.3**	**16.7**	15.8	**10.9**	18.3	**20.6**	**30.2**	**21.8**	**21.3**	**21.1**	**19.5**

6 Experimental Results

We compare our approach to four baselines. The first is *NoFuse* which estimates 2D poses independently for each view without multiview fusion. The second is *HeuristicFuse* which assigns a fixed fusion weight for each view according to Eq. (4). The parameter λ is set to be 0.5 by cross-validation. The third baseline is *ScoreFuse* which uses the same formulation as AdaFuse, i.e. Eq. (5), for feature fusion. It differs from AdaFuse only in the way we compute ω. In particular, ScoreFuse computes ω as the maximum value of the heatmap $\mathbf{H}$. Our approach is denoted as *AdaFuse* which uses the predicted weight for fusion as in Eq. (5). All of the four methods use triangulation (Hartley and Zisserman, 2003) to estimate 3D pose from the multiview 2D poses. We also compare to a baseline *RANSAC* which does not perform multiview fusion, but uses RANSAC to remove the outliers in triangulation.

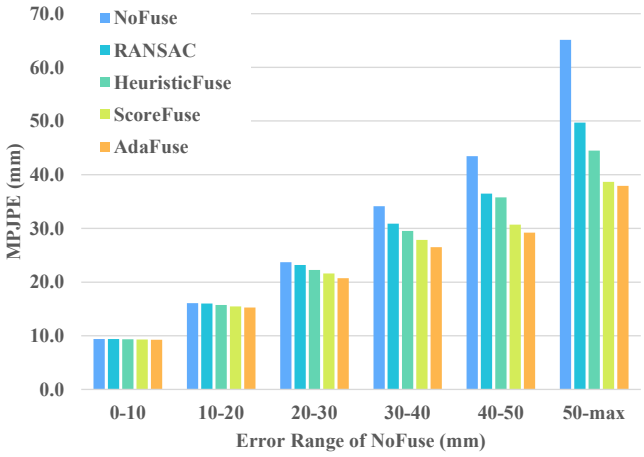

Fig. 11 We divide the test set of Human3.6M into to six groups according to the error of *NoFuse*. We compute the average error for every baseline and every group, respectively.

6.1 Results on Human3.6M

2D Pose Estimation Results The 2D pose estimation results are presented in Table 2. All multiview fusion methods remarkably outperform *NoFuse*. The improvement is most significant for the Elbow and Wrist joints because they are frequently occluded by human body. The results demonstrate that multiview fusion is an effective strategy to handle occlusion. *AdaFuse* achieves the highest average accuracy among all fusion methods validating that learning appropriate fusion weights can effectively reduce the negative impact caused by the features of low-quality views.

3D Pose Estimation Results Table 3 shows the 3D pose estimation errors of the baselines and our approach. We can see that *NoFuse* gets an average error of 22.9mm. This is a very strong baseline whose error is only slightly larger than the state-of-the-arts (see Table 4). On top of this strong baseline, we observe that adding multiview fusion can further reduce the 3D pose estimation errors.

HeuristicFuse gets a smaller error than *NoFuse* which is consistent with the 2D results in Table 2. The mean error only decreases by 1.9mm because most examples are relatively easy leaving little space for improvement. However, significant improvement is achieved for the challenging joints such as Wrist. The *ScoreFuse* gets a smaller error than *HeuristicFuse*. It means assigning small weights to low-quality views helps improve the quality of the fused heatmaps. Finally, our approach *AdaFuse*, which determines the fusion weight by considering both appearance cues and geometry consistency, notably decreases the average error to 19.5mm. Considering the baseline is already very strong, the improvement is significant. We notice that AdaFuse achieves slightly worse results on a small number of joints such as hip and head. This is mainly because these joints are rarely oc-

cluded in the datasets so the 2D pose estimator can obtain very accurate estimations for them. Further applying cross view fusion will introduce small noise to heatmaps leading to slightly worse 2D pose estimation accuracy. But when occlusion occurs which is often the case in practice, the benefit brought by cross view fusion will be much more significant than the harm caused by the small noise.

RANSAC is the de facto standard for solving robust estimation problems. As shown in Table 3, it outperforms *NoFuse* by removing some outlier 2D poses in triangulation. However, it is not as effective as the multiview fusion methods because the latter also attempt to refine, in addition to removing, the outlier poses. Another reason is that the number of cameras in this task is small which reduces the chance of finding the true outliers. In addition, we find that *RANSAC* is very sensitive to the threshold used for determining whether a data point is inlier or outlier. In our experiments, we set the threshold by cross validation.

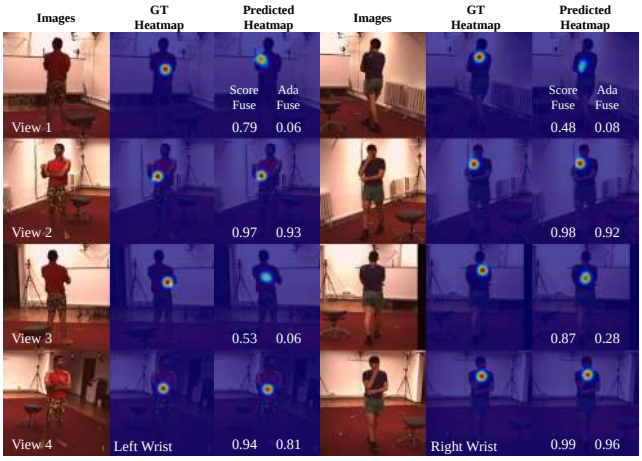

Fig. 12 We visualize the weights predicted by the *ScoreFuse* and *AdaFuse*, respectively. For example, in the first example (left sub-figure), the pose estimation network generates a high response at the wrong location for the first view. Consequently, *ScoreFuse* undesirably gives a large weight. In contrast, *AdaFuse* gives a small weight by identifying that its location are inconsistent with other views.

To better understand the improvement brought by *AdaFuse*, we divide the testing samples of the Human3.6M dataset into six groups according to the 3D errors of *NoFuse*. Then we compute the average error for each group. Figure 11 shows the results of various baselines. We can see that *AdaFuse* achieves the most significant improvement when the original error of *NoFuse* is large. However, even when the pose estimations of *NoFuse* are already accurate, *AdaFuse* can still reduce the error slightly.

Ablation Study on Fusion Weights One typical situation where *ScoreFuse* fails is when the pose estimation network generates large scores at *inaccurate* locations. In this case, *AdaFuse* can outperform *ScoreFuse* by leveraging the multiview geometry consistency. To support this conjecture, we visualize some typical heatmaps and the corresponding fusion weights predicted by the two methods, respectively, in Figure 12. We find that the heatmap responses are large for the four views although the locations are inaccurate for the first and third view. *ScoreFuse* gives large weights for all views which finally leads to a corrupted heatmap. In contrast, *AdaFuse* identifies that the predicted locations in the first and third view are inconsistent with the other two views in spite of their large scores. So it decreases the weights to ensure the good quality of the fused heatmap.

In addition, we also conduct ablation study on *AdaFuse* by using only one of two embedding networks. When we only use either the *appearance embedding* or *geometry embedding*, the 3D errors increase to 20.3mm and 19.9mm, respectively. Note that the improvement is actually much larger on those challenging examples. The results validate that the two embeddings are complementary.

Comparison to the State-of-the-arts Table 4 compares our approach to the state-of-the-arts. We can see that our approach outperforms all of them. Note that two approaches, *i.e. Triangulation* and *Volumetric*, are used in (Iskakov et al., 2019) to lift 2D poses to 3D. The *Triangulation* approach is more comparable to ours. Our approach *AdaFuse* decreases the error of (Iskakov et al., 2019) by about $13\%(=\frac{22.6-19.5}{22.6})$. The improvement is significant considering that the error of the state-of-the-art is already very small.

6.2 Results on Panoptic

We evaluate the impact of the number of cameras on this dataset. Figure 13 shows the mean 3D errors when three to six cameras are used, respectively. In general, the error decreases when more cameras are used for most baselines. However, we observe that the error of *NoFuse* actually becomes larger when the camera number increases from three to four. This undesirable phenomenon happens because the new camera view is very challenging thus the 2D pose estimation results are inaccurate. However, for our approach *AdaFuse*, the negative impact of low-quality heatmaps in individual views is limited due to the adaptive multiview fusion. We can see that the error of *AdaFuse* consistently decreases when the number of cameras increases. Since there is not a commonly adopted evaluation protocol and very few works have reported results on this new dataset, we do not compare our approach to the other approaches.

Table 4 The 3D pose estimation errors (mm) of the state-of-the-arts and our approach on the Human3.6M dataset. We report results for each of the 15 actions individually and also the average error over all actions. T- Iskakov et al. (2019) means triangulation is used. V- Iskakov et al. (2019) means volumetric method is used.

Methods	Direct	Disc.	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	*MPJPE*
Trumble et al. (2017)	92.7	85.9	72.3	93.2	86.2	101.2	75.1	78.0	83.5	94.8	85.8	82.0	114.6	94.9	79.7	87.3
Pavlakos et al. (2017)	41.2	49.2	42.8	43.4	55.6	46.9	40.3	63.7	97.6	119.0	52.1	42.7	51.9	41.8	39.4	56.9
Tome et al. (2018)	43.3	49.6	42.0	48.8	51.1	64.3	40.3	43.3	66.0	95.2	50.2	52.2	51.1	43.9	45.3	52.8
Qiu et al. (2019)	24.0	26.7	23.2	24.3	24.8	22.8	24.1	28.6	32.1	26.9	30.9	25.6	25.0	28.0	24.4	26.2
T- Iskakov et al. (2019)	20.4	22.6	20.5	19.7	22.1	20.6	19.5	23.0	25.8	33.0	23.0	21.6	20.7	23.7	21.3	22.6
V- Iskakov et al. (2019)	18.8	20.0	19.3	**18.7**	20.2	19.3	**18.7**	22.3	**23.3**	29.1	21.2	**20.3**	19.3	21.6	19.8	20.8
NoFuse	20.1	22.2	20.2	22.2	23.9	18.2	20.6	25.9	37.0	24.6	22.4	22.5	18.2	22.8	18.5	22.9
AdaFuse (Ours)	**17.8**	**19.5**	**17.6**	20.7	**19.3**	**16.8**	18.9	**20.2**	25.7	**20.1**	**19.2**	20.5	**17.2**	**20.5**	**17.3**	**19.5**

Table 5 The 2D pose estimation accuracy (PCKh@t) of the baselines and our approach for the **occluded joints** on the *Occlusion-Person* dataset. We report results for each joint type individually, and also the average accuracy over all joint types.

Methods	Hip	Knee	Ankle	Shlder	Elbow	Wrist	Avg.
NoFuse	63.4	21.5	17.0	29.5	14.6	12.4	30.9
HeuristicFuse	76.9	59.0	73.4	63.5	49.0	54.8	65.0
ScoreFuse	90.9	88.6	88.1	86.0	93.2	86.8	89.8
AdaFuse	**96.5**	**96.0**	**92.5**	**94.1**	**98.3**	**93.2**	**95.5**

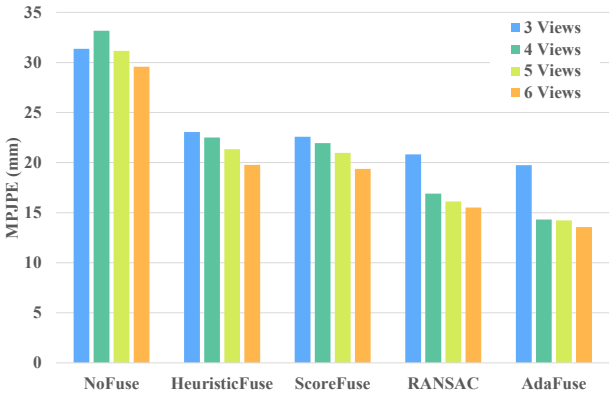

Fig. 13 The 3D pose estimation errors on the Panoptic dataset when different numbers of cameras are used.

6.3 Results on Occlusion-Person

2D Pose Estimation Results Table 5 shows the results on the *occluded* joints. Only about 30.9% of the occluded joints can be accurately detected by *NoFuse*. The result is reasonable because the features of the occluded joints are severely corrupted. All of the three multiview fusion methods remarkably improve the accuracy. In particular, more than 90% of the occluded joints are correctly detected by *AdaFuse*. The results demonstrate the advantages of our strategy for learning the fusion weights.

3D Pose Estimation Results We show the 3D pose estimation error (mm) for each joint type in Table 6. *NoFuse* results in a large error of 48.1mm. By improving the 2D pose estimation results on the occluded joints, the 3D errors are also significantly reduced, especially for the joints on the limbs such as Ankles and Wrists. In particular, our approach decreases the 3D error significantly to 12.6mm.

Impact of Number of Occluded Views We also evaluate the impact of the number of occluded views on this dataset. In particular, we classify each joint into one of five groups according to the number of occluded views, and report the average joint error for each group, respectively. The results are shown in Table 7. We can see that when the joints are visible in all views, the simple baseline *NoFuse* also achieves a very small error of 13.0mm. However, the error increases dramatically to 82.6mm when four views are occluded. Recall that there are eight views in total for this dataset. In contrast, the multiview fusion methods, especially our *AdaFuse*, achieves consistently smaller errors than *NoFuse*. More importantly, the error increase is much slower than *NoFuse* when more camera views are occluded which validates the robustness of our approach to occlusion.

Generalization Power The only learnable parameters in our fusion approach are in the appearance embedding and geometry embedding networks. In this section, we evaluate whether the *AdaFuse* weight prediction network learned on *Occlusion-Person* can be directly applied to the other datasets. In particular, we append the *AdaFuse* weight prediction network learned on *Occlusion-Person* to the 2D pose estimators trained on each dataset itself as the final model for evaluation. Table 8 shows the 3D pose estimation results on various datasets. We find that the fusion network learned on the synthetic *Occlusion-Person* dataset achieves similar performance on the three realistic datasets compared to the networks learned on each of the target dataset, respectively. The promising results validate that the fusion model has strong generalization power. It is also worth noting that our approach can naturally handle different numbers of cameras for two reasons. First, the parameters in the appearance embedding network and the geometry embedding network are shared for all camera views. Second, the "Mean" operator in the geometry embedding network makes it independent of the number of views as shown in Figure 7 and Figure 8. In summary, *AdaFuse* is ready to be deployed in new environ-

Table 6 The 3D pose estimation error (*mm*) of the baselines and our approach on the *Occlusion-Person* dataset. We report the result on each joint individually and also the average over all joints. The second row shows the percentage of the joints that are occluded for each joint type.

	Root	Belly	Neck	Hip	Knee	Ankle	Shlder	Elbow	Wrist	Mean
Occluded (%)	14.3%	13.7%	7.6%	23.0%	25.0%	23.5%	16.8%	25.3%	21.7%	
NoFuse	10.0	12.2	12.5	16.8	61.1	113.9	28.0	63.7	60.3	48.1
HeuristicFuse	8.8	10.7	**11.5**	14.2	21.1	19.2	17.5	23.6	24.1	18.0
ScoreFuse	8.4	12.6	12.6	14.7	17.5	17.1	16.1	13.2	16.9	15.0
RANSAC	8.6	11.2	11.7	12.9	18.8	17.9	17.1	14.5	19.7	15.5
AdaFuse (Ours)	**7.2**	**10.6**	**11.6**	**11.7**	**13.8**	**15.7**	**14.2**	**9.9**	**14.4**	**12.6**

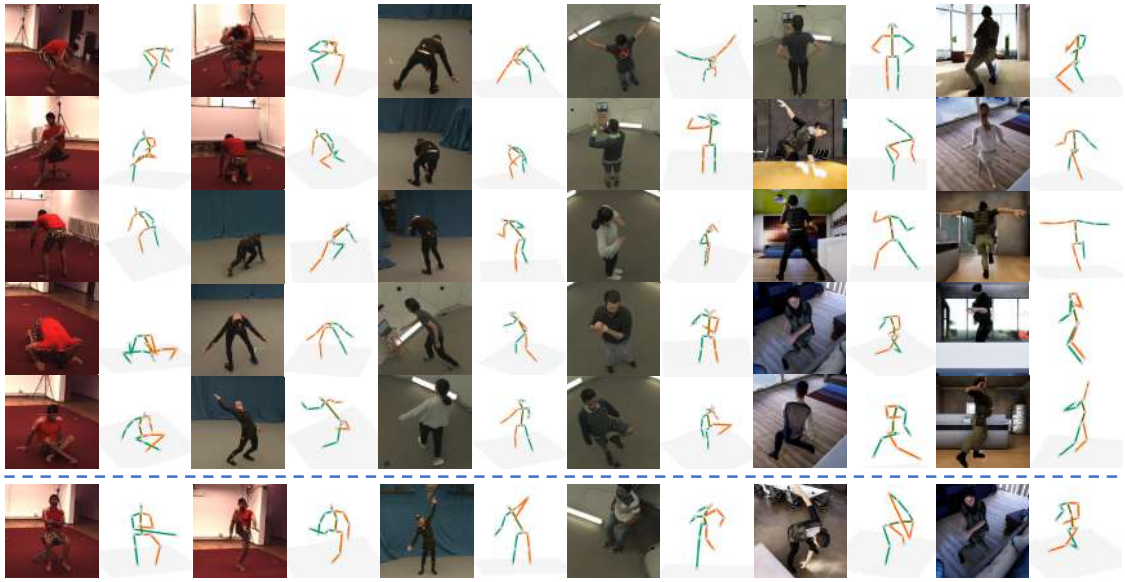

Fig. 14 We demonstrate some 3D pose estimation examples obtained by *AdaFuse*. The last row shows some failure cases.

Table 7 The 3D pose estimation error (mm) of the baseline methods and our approach on the Occlusion-Person dataset. We group the the 3D joints by number of occluded views (8 views in all). We show each group's joint number percentage in the second row.

Occluded Views	4	3	2	1	0
Percentage	2%	15%	38%	35%	10%
NoFuse	82.6	70.2	59.7	33.7	13.0
HeuristicFuse	30.5	19.9	15.9	13.5	11.1
ScoreFuse	25.0	18.1	15.2	13.4	12.6
RANSAC	36.5	24.5	19.4	14.3	11.7
AdaFuse (Ours)	**21.7**	**14.8**	**12.5**	**11.5**	**10.8**

ments of different camera poses without additional adaptation.

6.4 Results on Total Capture

We report the 3D pose estimation results on the Total Capture dataset in Table 9. It is worth noting that some methods also use IMUs in addition to the multiview images. We can see that our approach outperforms all of the previous methods. We notice that the error of our approach is slightly larger than LSTM-AE (Trumble et al., 2018) for the "W2 (walking)" action of S4,5. We tend to think it is because LSTM can get significant benefits when it is applied to periodic actions such as "walking". This is also observed independently in another work (Gilbert et al., 2019).

We show some 3D pose estimation examples in Figure 14. In most cases, our approach can accurately estimate the 3D poses. One typical situation where the approach fails is when 2D pose estimation results are inaccurate for many camera views. For example in the Panoptic dataset, when human begin to enter the dome, they may be occluded in multiple views. In this case, the heatmaps in each view are of low-quality. Therefore the fused heatmaps will also have degraded quality, leading to inaccurate 2D pose estimations.

7 Summary and Future Work

We present a multiview fusion approach *AdaFuse* to handle the occlusion problem in human pose estimation. *AdaFuse* has practical values in that it is very simple and can be flexibly applied to new environments without additional adaptation. In addition, it can be combined with any 2D pose estimation networks. We extensively evaluate the effectiveness of the approach on three benchmark datasets. The approach

Table 8 The 3D pose estimation errors MPJPE (mm) when *AdaFuse* weight prediction network is trained on *Occlusion-Person* or directly trained on the Evaluation dataset, respectively. The 2D pose estimators for generating the initial heatmaps are trained on each Evaluation dataset separately.

Evaluation Dataset	AdaFuse Trained on		NoFuse	HeuristicFuse	ScoreFuse	RANSAC
	Evaluation Dataset	Occlusion-Person				
Human3.6M	19.5	19.4	22.9	21.0	20.1	21.8
Panoptic 4 views	14.7	14.6	33.2	22.5	21.9	16.9
Panoptic 6 views	13.6	13.9	29.6	19.8	19.4	15.5
Total Capture	19.2	20.1	29.4	20.0	20.5	20.5

Table 9 The 3D pose estimation errors MPJPE (mm) of different methods on the Total Capture dataset.

Methods	IMUs	Temporal	Subjects(S1,2,3)			Subjects(S4,5)			Mean
			W2	A3	FS3	W2	A3	FS3	
(Trumble et al., 2017)	✓	✓	48.3	94.3	122.3	84.3	154.5	168.5	107.3
(Wei et al., 2016)			79.0	106.5	112.1	79.0	73.7	149.3	99.8
(Gilbert et al., 2019)	✓		19.2	42.3	48.8	24.7	58.8	61.8	42.6
(Trumble et al., 2018)		✓	13.0	23.0	47.0	**21.8**	40.9	68.5	34.1
(Qiu et al., 2019)			19	21	28	32	33	54	29
NoFuse			15.9	18.5	29.9	33.9	33.8	60.0	29.4
HeuristicFuse			7.8	11.6	19.6	23.3	26.9	44.8	20.0
ScoreFuse			9.7	13.1	19.9	23.9	27.2	41.4	20.5
RANSAC			8.4	11.6	20.5	23.3	27.2	45.7	20.5
AdaFuse (Ours)			**7.2**	**10.8**	**18.5**	22.8	**26.6**	**42.9**	**19.2**

outperforms the state-of-the-arts remarkably. We also construct a large scale human dataset which has severe occlusion to promote more research along this direction. Our next step of work is to leverage temporal information to further improve the pose estimation accuracy.

References

Amin S, Andriluka M, Rohrbach M, Schiele B (2013) Multi-view pictorial structures for 3D human pose estimation. In: BMVC

Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2D human pose estimation: New benchmark and state of the art analysis. In: CVPR, pp 3686–3693

Belagiannis V, Amin S, Andriluka M, Schiele B, Navab N, Ilic S (2014) 3d pictorial structures for multiple human pose estimation. In: CVPR, pp 1669–1676

Bo L, Sminchisescu C (2010) Twin gaussian processes for structured prediction. IJCV 87(1-2):28

Bridgeman L, Volino M, Guillemaut JY, Hilton A (2019) Multi-person 3d pose estimation and tracking in sports. In: CVPRW, pp 0–0

Burenius M, Sullivan J, Carlsson S (2013) 3D pictorial structures for multiple view articulated pose estimation. In: CVPR, pp 3618–3625

Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR, pp 7291–7299

Chen W, Wang H, Li Y, Su H, Wang Z, Tu C, Lischinski D, Cohen-Or D, Chen B (2016) Synthesizing training images for boosting human 3d pose estimation. In: 3DV, IEEE, pp 479–488

Cheng Y, Yang B, Wang B, Yan W, Tan RT (2019) Occlusion-aware networks for 3d human pose estimation in video. In: ICCV, pp 723–732

Ci H, Wang C, Ma X, Wang Y (2019) Optimizing network structure for 3d human pose estimation. In: ICCV, pp 915–922

Ci H, Ma X, Wang C, Wang Y (2020) Locally connected network for monocular 3d human pose estimation. In: T-PAMI

Dong J, Jiang W, Huang Q, Bao H, Zhou X (2019) Fast and robust multi-person 3d pose estimation from multiple views. In: CVPR, pp 7792–7801

Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6):381–395

Gal Y (2016) Uncertainty in deep learning. PhD thesis, PhD thesis, University of Cambridge

Gal Y, Ghahramani Z (2015) Dropout as a bayesian approximation: Insights and applications. In: Deep Learning Workshop, ICML, vol 1, p 2

Gall J, Rosenhahn B, Brox T, Seidel HP (2010) Optimization and filtering for human motion capture. IJCV 87(1-2):75

Ghahramani Z (2016) A history of bayesian neural networks. In: NIPS Workshop on Bayesian Deep Learning

Gilbert A, Trumble M, Malleson C, Hilton A, Collomosse J (2019) Fusing visual and inertial sensors with semantics for 3d human pose estimation. IJCV 127(4):381–397

Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: ICML, JMLR. org, pp 1321–1330

Hartley R, Zisserman A (2003) Multiple view geometry in computer vision. Cambridge university press

He Y, Zhu C, Wang J, Savvides M, Zhang X (2019) Bounding box regression with uncertainty for accurate object detection. In: CVPR, pp 2888–2897

Hoffmann DT, Tzionas D, Black MJ, Tang S (2019) Learning to train with synthetic humans. In: German Conference on Pattern Recognition, Springer, pp 609–623

Ilg E, Cicek O, Galesso S, Klein A, Makansi O, Hutter F, Brox T (2018) Uncertainty estimates and multi-hypotheses networks for optical flow. In: ECCV, pp 652–667

Ionescu C, Papava D, Olaru V, Sminchisescu C (2014) Human3. 6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. T-PAMI pp 1325–1339

Iskakov K, Burkov E, Lempitsky V, Malkov Y (2019) Learnable triangulation of human pose. arXiv preprint arXiv:190505754

Joo H, Simon T, Li X, Liu H, Tan L, Gui L, Banerjee S, Godisart T, Nabbe B, Matthews I, et al. (2019) Panoptic studio: A massively multiview system for social interaction capture. T-PAMI 41(1):190–204

Kendall A, Gal Y (2017) What uncertainties do we need in bayesian deep learning for computer vision? In: NIPS, pp 5574–5584

Kreiss S, Bertoni L, Alahi A (2019) Pifpaf: Composite fields for human pose estimation. In: CVPR, pp 11977–11986

Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: NIPS, pp 6402–6413

Lassner C, Romero J, Kiefel M, Bogo F, Black MJ, Gehler PV (2017) Unite the people: Closing the loop between 3d and 2d human representations. In: CVPR, pp 6050–6059

Li T, Fan L, Zhao M, Liu Y, Katabi D (2019) Making the invisible visible: Action recognition through walls and occlusions. In: ICCV, pp 872–881

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: ECCV, Springer, pp 740–755

Liu Y, Stoll C, Gall J, Seidel HP, Theobalt C (2011) Markerless motion capture of interacting characters using multi-view image segmentation. In: CVPR, IEEE, pp 1249–1256

Malleson C, Gilbert A, Trumble M, Collomosse J, Hilton A, Volino M (2017) Real-time full-body motion capture from video and imus. In: 3DV, IEEE, pp 449–457

von Marcard T, Henschel R, Black MJ, Rosenhahn B, Pons-Moll G (2018) Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV, pp 601–617

Martinez J, Hossain R, Romero J, Little JJ (2017) A simple yet effective baseline for 3D human pose estimation. In: ICCV, p 5

Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Computer vision and image understanding 104(2-3):90–126

Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: ECCV, Springer, pp 483–499

Pavlakos G, Zhou X, Derpanis KG, Daniilidis K (2017) Harvesting multiple views for marker-less 3D human pose annotations. In: CVPR, pp 1253–1262

Pavlakos G, Zhou X, Daniilidis K (2018) Ordinal depth supervision for 3d human pose estimation. In: CVPR, pp 7307–7316

Pavllo D, Feichtenhofer C, Grangier D, Auli M (2019) 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR, pp 7753–7762

Peng X, Tang Z, Yang F, Feris RS, Metaxas D (2018) Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In: CVPR, pp 2226–2234

Perez P, Vermaak J, Blake A (2004) Data fusion for visual tracking with particles. Proceedings of the IEEE 92(3):495–513

Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ (2017) On fairness and calibration. In: NIPS, pp 5680–5689

Qiu H, Wang C, Wang J, Wang N, Zeng W (2019) Cross view fusion for 3d human pose estimation. In: ICCV, pp 4342–4351

Qiu W, Zhong F, Zhang Y, Qiao S, Xiao Z, Kim TS, Wang Y (2017) Unrealcv: Virtual worlds for computer vision. In: Proceedings of the 25th ACM international conference on Multimedia, ACM, pp 1221–1224

Rhodin H, Spörri J, Katircioglu I, Constantin V, Meyer F, Müller E, Salzmann M, Fua P (2018) Learning monocular 3d human pose estimation from multi-view images. In: CVPR, pp 8437–8446

Roetenberg D, Luinge H, Slycke P (2009) Xsens mvn: full 6dof human motion tracking using miniature inertial sensors. Xsens Motion Technologies BV, Tech Rep 1

Rogez G, Schmid C (2016) Mocap-guided data augmentation for 3d pose estimation in the wild. In: NIPS, pp 3108–

3116
Sigal L, Balan AO, Black MJ (2010) Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV 87(1-2):4
Starner T, Leibe B, Minnen D, Westyn T, Hurst A, Weeks J (2003) The perceptive workbench: Computer-vision-based gesture tracking, object tracking, and 3d reconstruction for augmented desks. Machine Vision and Applications 14(1):59–71
Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: CVPR, pp 5693–5703
Sun X, Xiao B, Wei F, Liang S, Wei Y (2018) Integral human pose regression. In: ECCV, pp 529–545
Tome D, Toso M, Agapito L, Russell C (2018) Rethinking pose in 3D: Multi-stage refinement and recovery for markerless motion capture. In: 3DV, pp 474–483
Trumble M, Gilbert A, Malleson C, Hilton A, Collomosse J (2017) Total capture: 3D human pose estimation fusing video and inertial sensors. In: BMVC, pp 1–13
Trumble M, Gilbert A, Hilton A, Collomosse J (2018) Deep autoencoder for combined human pose estimation and body model upscaling. In: ECCV, pp 784–800
Tu H, Wang C, Zeng W (2020) Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: ECCV, pp 1–16
Varol G, Romero J, Martin X, Mahmood N, Black MJ, Laptev I, Schmid C (2017) Learning from synthetic humans. In: CVPR, pp 109–117
Wei SE, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: CVPR, pp 4724–4732
Xiang D, Joo H, Sheikh Y (2019) Monocular total capture: Posing face, body, and hands in the wild. In: CVPR
Xiao B, Wu H, Wei Y (2018) Simple baselines for human pose estimation and tracking. In: ECCV, pp 466–481
Xie R, Wang C, Wang C (2020) Metafuse: A pre-trained fusion model for human pose estimation. In: CVPR
Yang W, Ouyang W, Wang X, Ren J, Li H, Wang X (2018) 3d human pose estimation in the wild by adversarial learning. In: CVPR, pp 5255–5264
Zafar U, Ghafoor M, Zia T, Ahmed G, Latif A, Malik KR, Sharif AM (2019) Face recognition with bayesian convolutional networks for robust surveillance systems. EURASIP Journal on Image and Video Processing 2019(1):10
Zhang Z, Wang C, Qin W, Zeng W (2020) Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In: CVPR, pp 2200–2209
Zhao M, Li T, Abu Alsheikh M, Tian Y, Zhao H, Torralba A, Katabi D (2018) Through-wall human pose estimation using radio signals. In: CVPR, pp 7356–7365
Zhao M, Liu Y, Raghu A, Li T, Zhao H, Torralba A, Katabi D (2019) Through-wall human mesh recovery using radio signals. In: ICCV, pp 10113–10122
Zhou X, Huang Q, Sun X, Xue X, Wei Y (2017) Towards 3D human pose estimation in the wild: a weakly-supervised approach. In: ICCV, pp 398–407

Noname manuscript No.
(will be inserted by the editor)

AdaFuse：自适应多视图融合用于野外准确人体姿态估计

Zhe Zhang[1†] · **Chunyu Wang**[2†] · **Weichao Qiu**[3] · **Wenhu Qin**[1*] · **Wenjun Zeng**[2]

Received: date / Accepted: date

摘要 遮挡可能是野外人体姿态估计中最大的问题。典型的解决方案通常依赖于侵入性传感器，如惯性测量单元（IMU），来检测被遮挡的关节。为了使任务真正无约束，我们提出了 *AdaFuse*，一种自适应多视图融合方法，能够通过利用可见视角中的特征来增强被遮挡视角中的特征。*AdaFuse* 的核心是确定两个视角之间的点-点对应关系，我们通过探索热图表示的稀疏性有效地解决了这个问题。我们还为每个摄像头视角学习自适应融合权重，以反映其特征质量，从而减少“坏”视角不良特征对“好”特征的破坏。融合模型与姿态估计网络端到端训练，可以直接应用于新的摄像头配置，无需额外适配。我们在三个公共数据集上广泛评估了该方法，包括 Human3.6M、Total Capture 和 CMU Panoptic，并在所有数据集上均超越了现有的最先进方法。我们还创建了一个大规模合成数据集 *Occlusion-Person*，该数据集允许我们对被遮挡的关节进行数值评估，因为它为图像中的每个关节提供了遮挡标签。数据集和代码已发布在https://github.com/zhezh/adafuse-3d-human-pose.

Zhe Zhang
E-mail: zhangzhecnjs@gmail.com

Chunyu Wang
E-mail: chnuwa@microsoft.com

Weichao Qiu
E-mail: qiuwc@gmail.com

Wenhu Qin
E-mail: qinwenhu@seu.edu.cn

Wenjun Zeng
E-mail: wezeng@microsoft.com

[1] Southeast University, Nanjing, China
[2] Microsoft Research Asia, Beijing, China
[3] The Johns Hopkins University, MD, USA
[*] Corresponding Author
[†] Zhe Zhang and Chunyu Wang have contributed equally. Work done when Zhe Zhang is an intern at Microsoft Research Asia

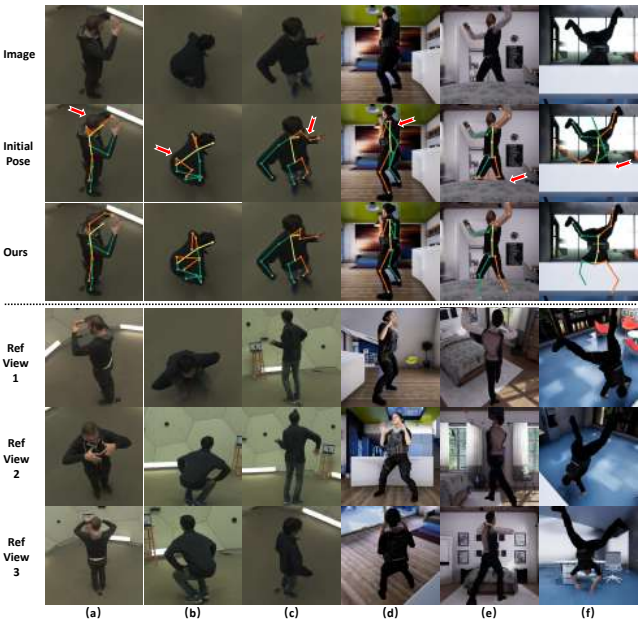

图 1 我们的方法通过利用其他视角中的特征，准确地检测到即使被遮挡的姿态。底部三行是来自场景其他视角的图像，帮助读者更好地感知演员的 3D 姿态。

Keywords 人体姿态估计 · 多摄像头融合 · 极几何

1 引言

从多个摄像头准确估计 3D 人体姿态一直是计算机视觉中的一个长期目标 (Liu et al., 2011; Bo and Sminchisescu, 2010; Gall et al., 2010; Rhodin et al., 2018; Amin et al., 2013; Burenius et al., 2013; Pavlakos et al., 2017; Belagiannis et al., 2014)。最终目标是从自然环境中布置的多个摄像头恢复身体关节在世界坐标系中的绝对 3D 位置。由于这一任务能够惠及许多应用领域，如增强现实和虚拟现实 (Starner et al., 2003)、人机交互以及体育视频中的智能运动员分析 (Bridgeman et al., 2019)，因此引起了广泛关注。

该任务通常通过一个简单的两步框架来解决。在第一步中，它尝试检测所有摄像头视角中的 2D 姿态，例如通过卷积神经网络 (Cao et al., 2017; Xiao et al., 2018)。然后，在第二步中，利用多视角的 2D 姿态，通过解析方法 (Burenius et al., 2013; Pavlakos et al., 2017; Belagiannis et al., 2014; Qiu et al., 2019; Amin et al., 2013) 或判别模型 (Iskakov et al., 2019; Tu et al., 2020) 来恢复 3D 姿态。在这些方法中，通常假设摄像头参数是已知的。强大的网络架构（如 (Newell et al., 2016)）的开发显著提高了 2D 姿态估计的质量，从而显著降低了 3D 误差。例如，在 (Qiu et al., 2019) 中，Human3.6M (Ionescu et al., 2014) 上的 3D 误差从 52mm 显著下降到 26mm。

然而，在基准数据集上获得较小的误差并不意味着该任务已经真正解决，除非现实应用中遇到的挑战，如背景杂乱、人类外观变化和遮挡问题，得到了很好的处理。事实上，越来越多的研究工作 (Zhou et al., 2017; Ci et al., 2019; Yang et al., 2018; Rogez and Schmid, 2016; Pavlakos et al., 2018; Ci et al., 2020) 致力于提高在复杂场景下的姿态估计性能，例如，通过增强训练数据集 (Zhou et al., 2017; Yang et al., 2018; Varol et al., 2017)（增加更多图像）或使用更为稳健的传感器，如惯性测量单元（IMU）(Trumble et al., 2017)。我们将在2节中更详细地讨论这一类工作。

在本研究中，我们提出通过多视图特征融合来以不同方式解决这一问题。该方法与之前的努力是正交的。如图1所示，我们的方法即使在某些视角中关节被遮挡时，仍能准确检测到关节。我们方法背后的动机是，某个视角中被遮挡的关节在其他视角中可能是可见的。因此，将不同视角中对应位置的特征进行融合通常是有帮助的。为此，我们提出了一种灵活的多视图融合方法，命名为 *AdaFuse*。图2展示了该方法的流程。首先，利用摄像头参数计算一对视角之间的点-线对应关系。然后，通过探索热图表示的稀疏性，而非进行具有挑战性的点-点匹配，来“找到”线上匹配的点。最后，融合不同视角中匹配点的特征。该方法可以有效提高遮挡视角中特征的质量。此外，对于具有不同摄像头姿态的新环境，只要摄像头参数可用，我们就可以直接使用 *AdaFuse*，无需重新训练。这提高了该方法在实际应用中的适用性。

AdaFuse 的性能进一步通过为每个视角学习自适应融合权重来提升，以反映其特征质量。该权重在融合过程中被利用，从而减少低质量视角的影响。如果某个关节在一个视角中被遮挡，则其特征也可能被损坏。在这种情况下，我们希望在进行多视图融合时给这个视角一个较小的权重，使得可见视角中的高质量特征占主导地位，并且不受低质量特征的干扰。我们向姿态估计网络中添加了一些简单的层，用于基于热图分布和视角一致性预测热图质量。在我们的实验中，我们观察到使用自适应融合显著提高了性能。

我们在三个公共数据集上评估了我们的方法，包括 Human3.6M (Ionescu et al., 2014)、Total Capture (Trumble et al., 2017) 和 CMU Panoptic (Joo et al., 2019)。结果表明，该方法超越了现有的最先进方法，验证了我们方法的有效性。此外，我们还将其与若干标准的多视图融合方法（如 RANSAC）进行了对比，以提供更详细的见解。我们通过在不同数据集上进行训练和测试，评估了我们方法的泛化能力。我们还创建了一个合成的人体姿态数据集，在该数据集中人体被物体故意遮挡。该数据集使我们能够对被遮挡的关节进行评估。

本文的其余部分结构如下。第2节讨论了与多视图 3D 人体姿态估计相关的工作，特别关注那些旨在提高复杂环境中性能的方法。第3节介绍了多视图特征融合的基础知识，为 *AdaFuse* 奠定基础。然后，我们描述了如何为每个摄像头视角学习自适应权重以反映特征质量，并详细阐述 *AdaFuse* 的方法。第5节

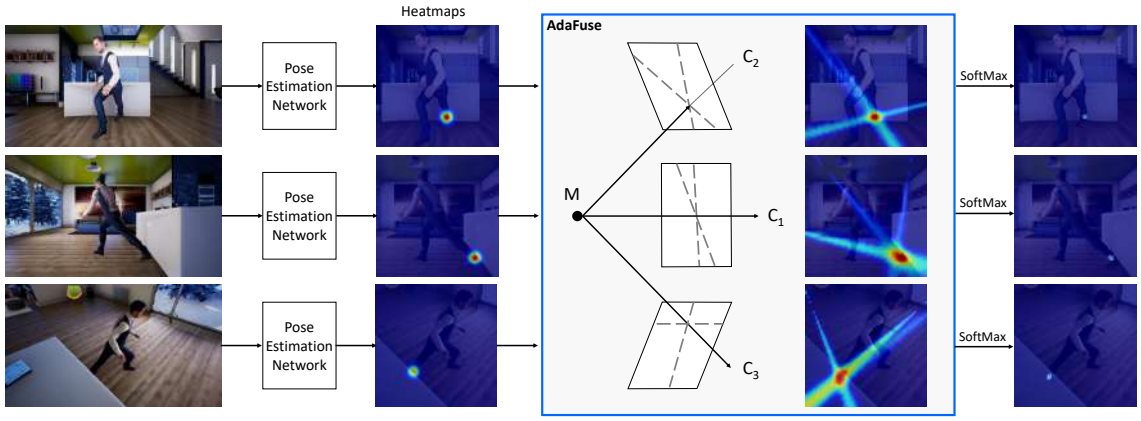

图 2 *AdaFuse* 的概述。它以多视图图像作为输入，并联合输出所有视角的 2D 姿态。首先，使用姿态估计网络为每个视角获取 2D 热图。然后，在极几何的基础上，将所有摄像头视角的热图进行融合。最后，应用 SoftMax 操作来抑制融合过程中引入的小噪声。因此，每个视角中的姿态估计都能从其他视角中受益。

和第6节分别介绍了实验数据集和实验结果。第7节总结了本文工作。

2 相关工作

我们首先在第2.1节回顾多视图 3D 人体姿态估计的相关工作。接着，第2.2节总结了用于提高“野外”环境中性能的技术。最后，在第2.3节中，我们讨论了共识学习方法，如 RANSAC。这对于多传感器融合是必要的，因为传感器可能会有矛盾的预测，且应去除离群值以确保良好的融合质量。

2.1 多视图 3D 人体姿态估计

我们将多视图 3D 人体姿态估计方法简要分为两类。第一类是基于模型的方法，也称为分析-合成方法 (Liu et al., 2011; Gall et al., 2010; Moeslund et al., 2006; Sigal et al., 2010; Perez et al., 2004)。这些方法首先通过简单的原始模型（如棍棒和圆柱体）来建模人体。然后，根据多视角图像中的观测，持续更新模型参数（即姿态），直到模型能够通过图像特征进行解释。所得到的优化问题通常是非凸的，因此常常需要使用昂贵的采样技术。这些方法之间的主要区别在于采用的图像特征和优化算法。对于感兴趣的读者，可以参考早期的综述文章，如 (Moeslund et al., 2006)。

基于模型的方法的优点在于其能够处理遮挡问题，因为人体模型中嵌入了固有的结构先验。这些方法将局部特征聚合为证据，以推断全局模型参数，并利用固有的人体结构作为约束。因此，如果某个关节被遮挡，它仍然可以依赖其他关节来猜测与先验一致的可能位置。然而，由于优化问题的复杂性，基于模型的方法比基于模型的方法产生更大的 3D 误差。

第二类是基于模型的方法 (Qiu et al., 2019; Iskakov et al., 2019; Burenius et al., 2013; Pavlakos et al., 2017; Dong et al., 2019; Amin et al., 2013; Belagiannis et al., 2014; Xie et al., 2020)，通常采用两步框架。它们首先在所有摄像头视角的图像中检测 2D 姿态。然后，在摄像头参数的帮助下，使用三角测量 (Amin et al., 2013; Iskakov et al., 2019) 或图形结构模型 (Burenius et al., 2013; Pavlakos et al., 2017; Dong et al., 2019) 恢复 3D 姿态。在 (Qiu et al., 2019) 中引入了递归图形结构模型，以加速推理过程。(Iskakov et al., 2019) 中的作者还提出使用可学习三角测量 (Hartley and Zisserman, 2003) 进行人体姿态估计，这对不准确的 2D 姿态具有更强的鲁棒性。如果 2D 姿态是准确的，那么恢复的 3D 姿态保证是准确的，而不必担心像基于模型的方法那样陷入局部最优。

更强大的网络架构的发展 (Newell et al., 2016; Sun et al., 2019) 显著提高了基准数据集上的 2D 姿态估计精度，从而也降低了 3D 姿态估计误差。例如，在最流行的基准数据集 Human3.6M (Ionescu et al.,

2014) 上，3D MPJPE 误差已降至约 20mm，满足了许多实际应用的要求。

2.2 提升“野外”性能

传感器 遮挡可能是“野外”场景中的最大挑战。一个直接的解决方案是使用额外的传感器，如 IMU (Trumble et al., 2017) 和无线电信号 (Zhao et al., 2019)，这些传感器不受遮挡的影响。例如，Roetenberg 等人 (Roetenberg et al., 2009) 在刚性骨骼上放置了 17 个 IMU。如果测量是准确的，则 3D 姿态可以完全确定。然而，在实际中，精度受到漂移问题的限制。为此，一些方法 (Trumble et al., 2017; von Marcard et al., 2018; Gilbert et al., 2019; Malleson et al., 2017; Zhang et al., 2020) 提出融合图像和 IMU，以实现更稳健的姿态估计。还有一些研究 (Zhao et al., 2019; Li et al., 2019; Zhao et al., 2018) 利用 WiFi 频率的无线信号可以穿透墙壁并从人体反射的事实，提出了一种基于无线电的系统，即使人在墙壁后完全被遮挡，也能估计 2D 姿态。然而，这些方法也有其自身的问题。例如，如何有效地融合视觉和惯性信号以用于基于 IMU 的方法？此外，在身体上佩戴传感器具有侵入性，在一些场景中（如足球比赛）是不可接受的。另一方面，基于 WiFi 的方法无法处理自遮挡问题，这是其一大限制。

数据增强 收集更多图像用于模型训练是提高泛化性能的有效方法。例如在 (Zhou et al., 2017; Qiu et al., 2019) 中，作者建议使用 MPII (Andriluka et al., 2014) 和 COCO (Lin et al., 2014) 数据集来帮助训练 3D 姿态估计器的 2D 模块，从而有效降低过拟合于简单训练数据集的风险。然而，标注一个足够大的姿态数据集既昂贵又耗时。因此，一些方法 (Rogez and Schmid, 2016; Varol et al., 2017; Hoffmann et al., 2019; Chen et al., 2016; Lassner et al., 2017) 提出生成合成图像。主要问题在于弥合合成图像与真实图像之间的差距，使得在合成图像上训练的模型可以应用于真实图像。为此，一些方法如 (Peng et al., 2018) 提出使用生成对抗网络（GAN）来生成逼真的图像。

时空上下文模型 一些方法提出使用时空上下文模型来联合检测视频序列中的所有关节，使每个关节都能从同一帧或相邻帧中的其他关节中受益。直观上，如果某个关节被遮挡，因此难以根据其自身外观被检测到，可以利用其他关节的位置来推测其可能的位置。例如，在之前的工作 (Cao et al., 2017; Kreiss et al., 2019) 中，作者提出除了检测单独的关节外，还检测身体部件（即连接两个关节的连杆）。这为相连关节的检测提供了相互增强的机会。在 (Cheng et al., 2019; Pavllo et al., 2019) 中，采用了时间卷积来处理当前帧中的遮挡问题。一些研究 (Qiu et al., 2019) 提出在多个摄像头视角之间建立空间对应关系，并利用多视图特征进行稳健的关节检测。在多个基准数据集上，对于被遮挡关节的检测性能取得了显著提升。(Qiu et al., 2019) 方法的主要缺点是在实际应用中缺乏灵活性，因为它需要为每一种可能的摄像头布局训练一个单独的融合网络。我们的工作与 (Qiu et al., 2019) 不同之处在于，我们的方法可以应用于具有不同摄像头数量和不同摄像头姿态的新环境，而无需额外的适配。在实验中，我们将对这两种方法进行比较。

2.3 共识学习

在多传感器融合中，一个基本问题是检测和去除离群值，因为传感器可能会产生不一致的测量结果。RANSAC (Fischler and Bolles, 1981) 是最常用的离群值检测方法。其主要假设是数据集中包含内点。RANSAC 仅在一定概率下生成合理的结果，而这一概率随着内点数量的增加而提高。在实际应用中，当传感器数量较少时，检测真实离群值的概率也较低。例如，在多视图人体姿态估计中，对于大多数基准数据集 (Ionescu et al., 2014; Trumble et al., 2017)，摄像头的数量通常只有四到八个。在这种情况下，我们观察到 RANSAC 可能不是最佳选择。

近年来，不确定性学习 (Kendall and Gal, 2017; Gal and Ghahramani, 2015; Lakshminarayanan et al., 2017; Zafar et al., 2019; Pleiss et al., 2017) 吸引了大量关注，尤其在自动驾驶和医学诊断等高风险应用中尤为重要 (Gal, 2016; Ghahramani, 2016)。其主要思想是，当模型进行预测时，它还会输出一个反映

预测置信度的分数。例如，考虑一辆使用神经网络检测行人的自动驾驶汽车。如果网络对其预测不够自信，汽车可能会依赖其他传感器来做出正确的决策。不确定性被引入到计算机视觉中 (Kendall and Gal, 2017; Kreiss et al., 2019; He et al., 2019; Ilg et al., 2018)。另一类方法如 (Guo et al., 2017; Pleiss et al., 2017) 通过校准来学习不确定性，它们提出训练模型，使得与预测类别标签相关的概率与其真实正确性的可能性相一致。

不确定性的概念可以用来减少离群值的影响。例如，在 (Iskakov et al., 2019) 中，作者提出为每个视角中的每个关节预测一个不确定性得分。该得分在进行三角测量时用于加权各个视角，从而显著减少 3D 姿态估计误差。受不确定性学习在计算机视觉任务中成功的启发，我们提出在多视图特征融合中学习不确定性。预测的不确定性在融合多视图特征时作为权重使用。我们展示了这种自适应特征融合能够有效提高融合质量。

3 多视图融合基础

我们首先介绍多视图融合的基础知识，为 *AdaFuse* 奠定基础。具体而言，我们讨论如何在两个视角之间建立点-点对应关系，使得对应于相同 3D 空间点的特征可以融合在一起。窄基线的对应关系可以通过局部特征匹配高效解决。然而，在多视图人体姿态估计的背景下，仅有少量的摄像头相互之间距离较远，局部特征尤其在人体无纹理区域中无法稳健地检测和匹配，这带来了严峻的挑战。

为了解决这一问题，我们提出了一种由粗到细的方法来寻找匹配点。首先，通过极几何在两个视角之间建立点-线对应关系，然后通过探索热图表示的稀疏性，隐式地确定点-点对应关系。该方法显著简化了任务，因为它避免了寻找精确对应关系的困难步骤。我们首先在第3.1节介绍极几何，以确定点-线对应关系。接着在第3.2节中，描述如何调整极几何来执行多视图热图融合。最后，我们在第3.3节讨论简化融合策略所引发的副作用以及我们的解决方案。

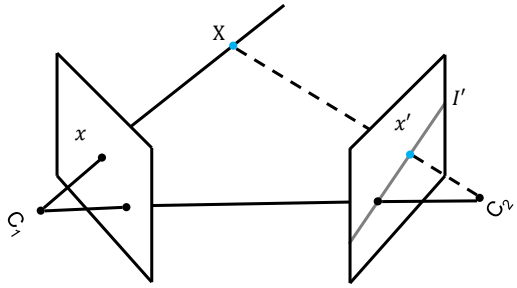

图 3 两视角中点-线对应关系的示意图。对于一个视角中的任意点 $\mathbf{x}$，其在另一个视角中的对应点 $\mathbf{x}'$ 必须位于极线 $\mathbf{l}'$ 上。这是 *AdaFuse* 在其他视角中寻找对应点的核心原理。

3.1 极几何

我们用 $\mathbf{X} \in \mathcal{R}^{4\times1}$ 表示 3D 空间中的一个点，如图 3 所示。在姿态估计的背景下，这可以表示一个身体关节的位置。注意，这里使用齐次坐标和列向量来表示一个点。3D 点分别在两个摄像机视角中成像，在第一个视角的成像为 $\mathbf{x} = \mathbf{P}\mathbf{X}$，在第二个视角的成像为 $\mathbf{x}' = \mathbf{P}'\mathbf{X}$，其中 $\mathbf{x}$ 和 $\mathbf{x}' \in \mathcal{R}^{3\times1}$ 表示图像中的 2D 点，$\mathbf{P}$ 和 $\mathbf{P}' \in \mathcal{R}^{3\times4}$ 是每个摄像头的投影矩阵。由于这两个 2D 点对应相同的 3D 点，并具有相同的语义意义，因此可以安全地融合它们的特征，使每个视角都能从其他视角中获益。

两个视角之间的极几何 (Hartley and Zisserman, 2003) 本质上是图像平面与以基线为轴的平面族的交集几何。基线是连接摄像头中心 C_1 和 C_2 的直线。特别地，对于第一个视角中的任意位置 $\mathbf{x}$，极几何帮助我们确定第二个视角中对应点 $\mathbf{x}'$ 的位置，而无需知道 $\mathbf{X}$。

从图 3 可以看出，图像点 $\mathbf{x}$ 和 $\mathbf{x}'$、3D 点 $\mathbf{X}$ 以及摄像头中心 $\mathbf{C_1}$ 和 $\mathbf{C_2}$ 位于同一平面 π 上。该平面与两个图像平面分别相交于极线 $\mathbf{l}$ 和 $\mathbf{l}'$。具体来说，

$$\begin{aligned} \mathbf{l}' &= \mathbf{F}\mathbf{x} \\ \mathbf{l} &= \mathbf{F}^\top\mathbf{x}', \end{aligned} \tag{1}$$

其中 $\mathbf{F} \in \mathcal{R}^{3\times3}$ 是基础矩阵（fundamental matrix），可由 $\mathbf{P}$ 和 $\mathbf{P}'$ 推导得出。详细的推导过程请参考 (Hartley and Zisserman, 2003)。

此外，从 $\mathbf{x}$ 和 $\mathbf{x}'$ 反向投影的射线相交于 $\mathbf{X}$，并且这些射线共面，即位于平面 π 上。由此可以直接推导出，对应于 $\mathbf{x}$ 的 $\mathbf{x}'$ 的位置必定位于极线 $\mathbf{l}'$ 上。

(a) (b) (c)

图 4 基于极几何的热图融合。对于第一个视角中的每个位置 $\mathbf{x}$，我们首先计算其在其他两个视角中的对应极线。然后分别在这两条线上找到响应最大的点，并将其响应值加到 $\mathbf{x}$ 的原始响应上。

(a) (b) (c)

图 5 基于极几何的热图融合。对于第一个视角中的每个位置 $\mathbf{x}$，我们首先计算其在其他两个视角中的对应极线。然后分别在这两条线上找到响应最大的点，并将其响应值加到 $\mathbf{x}$ 的原始响应上。

然而，我们还需要利用额外的信息（如外观特征）来确定 $\mathbf{x}'$ 在极线 $\mathbf{I}'$ 上的确切位置。

在多视图特征融合的背景下，对于每个图像点 $\mathbf{x}$，我们需要找到第二视角中的对应点 $\mathbf{x}'$，以便将 $\mathbf{x}$ 处的特征与 $\mathbf{x}'$ 处的特征融合，从而获得更稳健的姿态估计。由于我们不知道 $\mathbf{X}$ 的深度，因此 $\mathbf{X}$ 可以在由摄像头中心 $\mathbf{C}_1$ 和图像点 $\mathbf{x}$ 定义的直线上自由移动。然而，我们知道 $\mathbf{x}'$ 并不能遍布整个图像平面，而是被限制在极线 $\mathbf{I}'$ 上。在接下来的第 3.2 节中，我们将描述如何基于极几何进行多视图特征融合。

在多视图特征融合的背景下，对于每个图像点 $\mathbf{x}$，我们需要找到第二视角中的对应点 $\mathbf{x}'$，以便将 $\mathbf{x}$ 处的特征与 $\mathbf{x}'$ 处的特征融合，从而获得更稳健的姿态估计。由于我们不知道 $\mathbf{X}$ 的深度，因此 $\mathbf{X}$ 可以在由摄像头中心 $\mathbf{C}_1$ 和图像点 $\mathbf{x}$ 定义的直线上自由移动。然而，我们知道 $\mathbf{x}'$ 并不能遍布整个图像平面，而是被限制在极线 $\mathbf{I}'$ 上。在接下来的第 3.2 节中，我们将描述如何基于极几何进行多视图特征融合。

Sampson 距离 在实际应用中，通常我们有两个 2D 测量值 $\mathbf{x}$ 和 $\mathbf{x}'$，它们对应相同的 3D 位置 $\mathbf{X}$，但 $\mathbf{X}$ 是未知的。由于测量噪声和误差，直线 $\mathbf{C_1x}$ 和 $\mathbf{C_2x}'$ 可能不会精确地在位置 $\mathbf{X}$ 相交。为了获得 $\mathbf{X}$ 的最佳估计，我们需要搜索 $\widehat{\mathbf{X}}$，使得满足以下条件：

$$d^2_{Reproj} = \min_{\hat{\mathbf{x}}} d^2\left(\mathbf{x}, \mathbf{P}\widehat{\mathbf{X}}\right) + d^2\left(\mathbf{x}', \mathbf{P}'\widehat{\mathbf{X}}\right), \tag{2}$$

其中，$d(\cdot)$ 表示欧几里得距离，d_{Reproj} 表示 $\mathbf{x}$ 和 $\mathbf{x}'$ 之间的重投影误差。

由于计算 d_{Reproj} 时涉及优化过程，我们采用了一种一步法，即它的一阶近似 (Hartley and Zisserman, 2003)。这种近似也称为 Sampson 距离，其公式为：

$$d_{Sampson} = \frac{\mathbf{x}'^{\top}\mathbf{F}\mathbf{x}}{(\mathbf{F}\mathbf{x})_1^2 + (\mathbf{F}\mathbf{x})_2^2 + (\mathbf{F}^{\top}\mathbf{x}')_1^2 + (\mathbf{F}^{\top}\mathbf{x}')_2^2}, \tag{3}$$

其中，$\mathbf{F}$ 是基本矩阵，子标记 1 或 2 表示向量的第一个或第二个元素。通过使用 Sampson 距离，我们可以直接获得一对位置之间的距离，而不需要知道中间的 $\widehat{\mathbf{X}}$。在 *AdaFuse* 中，我们使用 Sampson 距离来表示一对 2D 关节点检测在多大程度上相互支持。

3.2 热图融合

多视图融合应用于热图而非中间特征，如图 2 所示。这是因为热图具有稀疏性的优点，可以简化点对点的匹配过程。热图为图像中的关节位置生成每个像素的可能性。具体来说，热图是以关节坐标为中心的二维高斯分布。因此，它在关节位置附近产生少量的大响应，而在其他位置则产生大量的零响应。图 5 (a) 展示了右膝关节的热图示例。

稀疏热图使我们可以安全地跳过精确的点对点匹配，因为在视差线上的“零”位置的特征不会对特征融合产生影响。因此，我们只需选择视差线上的最大响应作为匹配点。这种简化是合理的，因为相应的点通常会有最大的响应。例如，在图 5 中，对于每个位置 $\mathbf{x}$，我们首先计算其他两个相机视角中的对应视差线。然后，我们分别在这两条视差线上找到最大响应，并将它们与 $\mathbf{x}$ 位置的响应进行融合。

我们将视图 v 中的热图表示为 $\mathbf{H}^v$，热图在位置 $\mathbf{x}$ 处的响应记为 $\mathbf{H}^v(\mathbf{x})$。位置 $\mathbf{x}$ 在视图 u 中的对应视差线表示为 $\mathbf{I}^u(\mathbf{x})$，它由热图 $\mathbf{H}^u$ 上的一系列离散位

置组成。视差线可以根据相机参数为每个位置 $\mathbf{x}$ 进行解析计算。然后我们将多视图融合表示为：

$$\widehat{\mathbf{H}}^v(\mathbf{x}) = \lambda \mathbf{H}^v(\mathbf{x}) + \frac{1-\lambda}{N} \sum_{u=1}^{N} \max_{\mathbf{x}' \in \mathbf{I}^u(\mathbf{x})} \mathbf{H}^u(\mathbf{x}'), \tag{4}$$

其中，$\widehat{\mathbf{H}}$ 表示融合后的热图，N 是参与当前视图融合的相机视角数。参数 λ 用于平衡当前视图与其他视图的响应。

3.3 副作用与解决方案

简化的融合模型（即公式 (4)）引发的一个副作用是，一些背景位置可能会被不当地增强。我们在图 **??** 的第二行中可视化了一个示例。可以看到，许多背景像素，例如 $\mathbf{x}$，具有非零响应，这些响应是由融合过程引起的。这个现象发生的原因是，多个视差线（在其他视图中）可能经过真实的关节位置，并产生大的响应，而这些视差线实际上在当前视图中对应的是背景像素。如图 **??** 所解释，对于当前视图中的位置 $\mathbf{x}$，其他三个视图中的对应视差线在第一行中绘制出来。我们可以看到，尽管 $\mathbf{x}$ 不是一个有意义的关节位置，但第一个视图中的视差线经过真实的膝关节位置，并导致 $\mathbf{x}$ 产生了一个大的非预期响应。

幸运的是，背景像素受不希望的影响存在一些模式。通常，由另一个视图中的高响应位置影响的像素是保证位于同一条线上的。更重要的是，来自不同视图的视差线不会重叠。这意味着，对于背景中的某个位置 $\mathbf{x}$，它的响应最多只能被一个视图增强。相比之下，与有意义的身体关节对应的位置将会受到多个视图的增强。换句话说，正确的位置在一般情况下是保证具有最大响应的。因此，我们利用这个观察结果，直接应用 SoftMax 操作来去除小的响应。图 **??** 的第三行展示了效果。可以看到，只有位于关节位置周围的大响应被保留下来。

3.4 实现细节

值得注意的是，以上融合方法没有可学习的参数。因此，我们只需要训练骨干网络，如 SimpleBaseline (Xiao et al., 2018)，以估计姿态热图。训练骨干网络的损失函数定义为估计热图与真实热图之间的均方误差（MSE）。在测试阶段，给定由 SimpleBaseline 估计的热图，我们通过我们的方法确定性地进行融合。

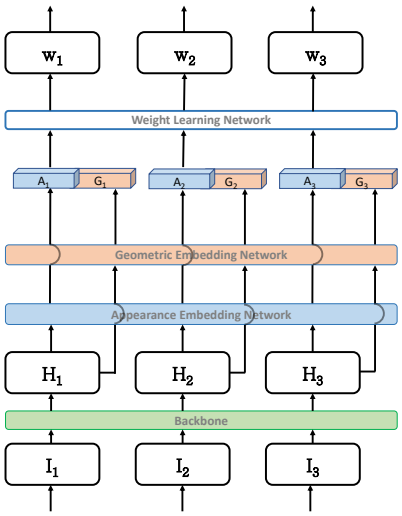

图 6 用于学习自适应融合权重的网络。姿态估计的骨干网络用于提取每个视角 $\mathbf{I}_v$ 的热图 $\mathbf{H}_v$。这些热图分别输入到 外观嵌入网络和 几何嵌入网络中提取特征，提取的特征被拼接后输入到 权重学习网络中，学习反映每个视角热图质量的融合权重。该权重用于多视角融合。

4 多视角融合的自适应权重

上一节中介绍的融合策略对所有视角一视同仁，并未考虑每个视角的特征质量。注意到在公式 (4) 中，融合权重是 $\frac{1-\lambda}{N}$，对于 N 个视角。然而，该策略在某些情况下存在问题，尤其是当某些相机视角的热图不正确时。因为这些特征可能会不良地干扰正确视角中的特征，导致完全错误的 2D 姿态估计结果。

为了解决这个问题，我们提出了一个权重学习网络，学习每个视角的 自适应权重，以真实地反映其热图质量。该网络以姿态估计网络提取的 N 个视角的热图作为输入，回归出 N 个权重 ω^u。然后，多视角融合重新写作，考虑这些权重如下所示。

$$\widehat{\mathbf{H}}^v(\mathbf{x}) = \omega^v \mathbf{H}^v(\mathbf{x}) + \sum_{u=1}^{N} \omega^u \max_{\mathbf{x}' \in \mathbf{I}^u(\mathbf{x})} \mathbf{H}^u(\mathbf{x}'), \tag{5}$$

自适应融合权重 ω 的预测是通过一个轻量级神经网络实现的，如图 6 所示。在姿态估计网络提供的热图 $\mathbf{H}$ 上，我们提取两种类型的信息来进行预测。

图 7 用于预测融合权重的外观嵌入网络。i 是相机视角的索引。网络中的参数对所有视角和关节共享。另请参见图 6，了解外观嵌入 A_i 如何用于确定融合权重。

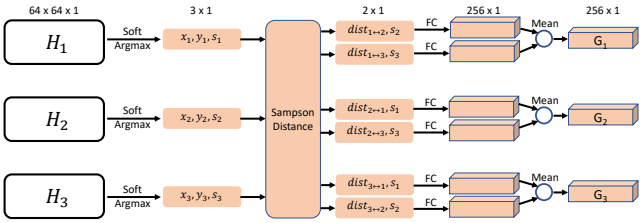

图 8 用于预测融合权重的几何嵌入网络。对于每个相机视角中的每个关节（此示例中显示了三个视角），它生成一个 256 维的嵌入，反映热图（姿态）的质量。注意，所有分支共享同一个全连接（FC）层。

第一种是外观嵌入，它提取诸如热图分布特征等信息。第二种是几何嵌入，它考虑了跨视角位置的一致性。这两个部分是互补的。所提出的权重学习网络可以与姿态估计网络联合进行端到端训练，无需对权重进行监督。

4.1 外观嵌入

每个关节的热图实际上包含了丰富的信息，可以用来推测其热图质量。例如，如果预测的热图具有理想的高斯核形状，则在许多情况下，热图质量较好。相反，如果预测的热图在空间中具有随机且小的响应（例如当关节被遮挡时），则质量很可能较差。

我们提出了一个简单的网络，用于提取每个相机视角中每个关节的外观嵌入。图 7 显示了该网络结构。从热图 $\mathbf{H}_i$ 开始，我们应用卷积层提取特征。然后通过平均池化对特征进行下采样，并将其输入到全连接（FC）层中提取外观嵌入。不同的关节类型和相机视角共享相同的权重。为了简化展示，我们只展示了单视角和单关节的网络结构。外观嵌入网络与姿态估计网络共同进行端到端学习。

4.2 几何嵌入

仅仅依赖外观嵌入不足以处理一些具有挑战性的情况，这些情况中热图虽然具有理想的高斯核形状，但位置不正确。一个这样的例子是，当左膝被检测到在右膝的位置时，这通常被称为“重复计数”问题。为了解决这个问题，我们提出利用所有相机视角之间的位置一致性信息。我们的核心动机是，如果一个相机视角中的预测关节位置与其他视角中的位置一致，那么该位置的可靠性会更高。

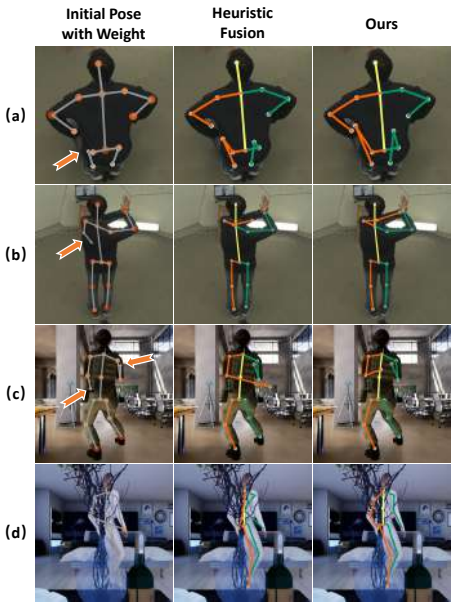

图 9 我们通过第一列中标记的大小来可视化预测的融合权重。较大的标记表示更大的权重。其余两列分别显示了由 *HeuristicFuse* 和 *AdaFuse* 估计的姿态。由于考虑了每个视角的特征质量，我们的 *AdaFuse* 显示了明显更好的估计结果。

我们通过一个几何嵌入网络实现了这一思想，如图 8 所示。从热图 $\mathbf{H}$ 开始，我们首先应用“软-最大”操作 (Sun et al., 2018) 来获得每个视角中关节的位置 (x, y)。我们还获取该位置的热图响应值 s，以反映其置信度。接下来，我们计算当前视角与其他视角之间的 Sampson 距离 (Hartley and Zisserman, 2003) $dist_{i\leftrightarrow j}$，用于衡量对应关系或一致性误差。较小的 $dist_{i\leftrightarrow j}$ 表示两个视角中的关节位置是一致的。直观地讲，与大多数视角一致的位置更可靠。最后，我们提出使用一个全连接（FC）层将 Sampson 距离

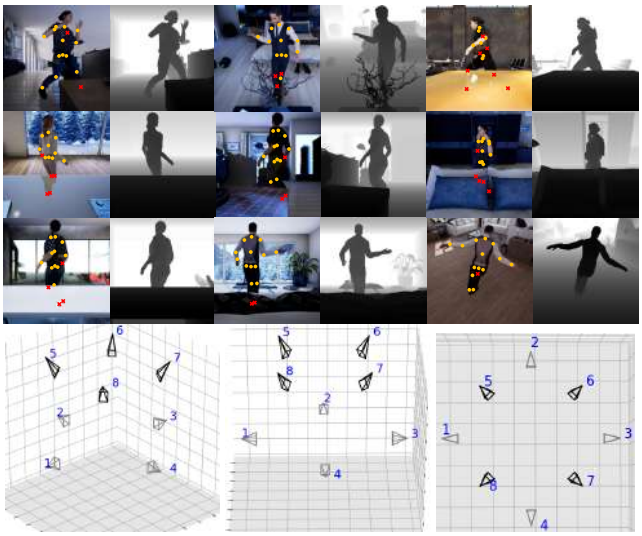

图 10 我们展示了来自 *Occlusion-Person* 数据集的典型图像、真实的 2D 关节位置以及深度图。红色“x”表示该关节被遮挡。下排显示了数据集中使用的八个相机从不同视角拍摄的空间配置。

嵌入到特征向量中。所有相机对的特征向量随后被平均，得到最终的几何嵌入。

4.3 权重学习网络

我们提出了一个简单的网络，由三层全连接（FC）层组成，用于将拼接后的外观和几何嵌入转化为回归最终权重。值得注意的是，我们并没有独立训练权重学习网络，而是将其与姿态估计网络结合，通过最小化融合后的 2D 热图损失来训练，而不对融合权重施加中间监督。图 9 中的第一列展示了我们方法预测的一些典型权重。我们可以看到，当关节被遮挡并且定位在错误位置时，相应的融合权重确实比其他关节要小。

5 数据集与评估指标

我们介绍了用于评估的三个数据集及其相应的评估指标。我们还描述了如何构建合成的 *Occlusion-Person* 数据集，该数据集具有大量的人的遮挡情况。

表 1 公开多视角姿态估计数据集的统计信息。只有 *Occlusion-Person* 数据集提供遮挡标签。

数据集	帧数	相机数量	遮挡关节
Human3.6M	784k	4	-
Total Capture	236k	8	-
Panoptic	36k	31	-
Occlusion-Person	73k	8	20.3%

5.1 数据集

Human3.6M 数据集 (Ionescu et al., 2014) 该数据集提供了由四台相机同步捕捉的图像，共有七个受试者执行日常动作。我们使用交叉受试者评估方案，其中受试者 1、5、6、7、8 用于训练，受试者 9 和 11 用于测试。为了避免过拟合简单背景，我们还使用 MPII 数据集 (Andriluka et al., 2014) 来扩充训练数据。由于 MPII 数据集仅提供单目图像，我们只训练骨干网络，而不进行多视角融合。

Total Capture 数据集 (Trumble et al., 2017) 该数据集提供了由八台相机同步捕捉的人的图像。按照数据集的惯例，训练集包括受试者 1、2 和 3 的“ROM1,2,3”、“Freestyle1,2”、“Walking1,3”、“Acting1,2” 和 “Running1”。测试集包括受试者 1、2、3、4 和 5 的 “Freestyle3（**FS3**）”、“Acting3（**A3**）” 和 “Walking2（**W2**）”。

CMU Panoptic 数据集 (Joo et al., 2019) 这是一个最近引入的数据集，提供了由多个相机捕捉的图像。我们统一选择了六个相机来评估相机数量对 3D 姿态估计的影响。具体来说，首先选择相机 1、2 和 10 来构建一个三视角的实验设置。然后，依次将相机 13、3 和 23 添加到之前的三个相机中，分别构建四视角、五视角和六视角的实验设置。我们遵循先前工作 (Xiang et al., 2019) 的做法，选择了仅包含一个人的训练和测试序列。由于很少有工作报告该数据集上的数值结果，因此我们只将我们的方法与基准进行比较。

Occlusion-Person 数据集 先前的基准数据集未提供图像中关节的遮挡标签，这使得我们无法对被遮挡的关节进行数值评估。此外，这些基准数据集中的遮

表 2 基准方法和我们方法在 Human3.6M 数据集上的 2D 姿态估计准确率（PCKh@t）。我们报告每个单独关节的结果以及所有关节的平均值。

方法	根部	腹部	颈部	鼻部	头部	臀部	膝部	脚踝	肩膀	肘部	手腕	平均
NoFuse	95.8	77.1	60.4	86.4	86.2	79.3	81.5	58.6	65.1	78.3	70.1	74.8
HeuristicFuse	96.0	79.3	60.7	**88.4**	**86.8**	83.1	84.5	60.0	**66.9**	82.1	75.2	77.3
ScoreFuse	96.2	79.3	61.6	88.3	86.2	83.3	84.3	60.5	66.6	83.1	77.4	77.8
AdaFuse (我们的)	**96.2**	**79.3**	**61.6**	88.3	86.3	**83.5**	**86.4**	**61.1**	66.7	**86.0**	**80.1**	**78.8**

表 3 基准方法和我们方法在 Human3.6M 数据集上的 3D 姿态估计误差（单位：毫米）。

方法	腹部	颈部	鼻部	头部	臀部	膝部	脚踝	肩膀	肘部	手腕	平均
NoFuse	21.6	16.8	15.7	11.3	**17.8**	25.8	35.8	22.0	26.8	34.1	22.9
HeuristicFuse	21.6	16.8	15.7	11.0	17.9	23.0	32.7	21.9	25.0	25.7	21.0
ScoreFuse	21.4	16.7	15.8	10.9	18.3	21.3	30.8	21.8	23.3	23.2	20.1
RANSAC	21.6	16.8	**15.7**	11.2	17.9	23.9	34.6	22.0	25.8	28.2	21.8
AdaFuse (我们的方法)	**21.3**	**16.7**	15.8	**10.9**	18.3	**20.6**	**30.2**	**21.8**	**21.3**	**21.1**	**19.5**

挡量也有限。为了解决这些限制，我们提出构建了合成数据集 *Occlusion-Person*。我们采用 UnrealCV (Qiu et al., 2017) 从 3D 模型渲染多视角图像和深度图。具体来说，13 个人体模型（穿着不同的衣物）被放置在 9 个不同的场景中，如客厅、卧室和办公室。人体模型的姿态从 CMU 动作捕捉数据库中选择。我们有意使用沙发、桌子等物体遮挡一些身体关节。每个场景中放置了 8 台相机来渲染多视角图像和深度图。这 8 台相机均匀地放置在一个半径为两米的圆圈上，每台相机的高度分别约为 0.9 米和 2.3 米。我们提供了 15 个关节的 3D 位置信息作为真实标签。图 10 显示了数据集中的一些示例图像以及相机的空间配置。

每个图像中关节的遮挡标签是通过比较其深度值（可在深度图中获得）与相机坐标系中 3D 关节的深度值来获取的。如果两个深度值之间的差异小于 30cm，则该关节未被遮挡。否则，该关节被认为是被遮挡的。表 1 将该数据集与现有基准数据集进行了比较。特别地，在我们的数据集中，大约 20% 的身体关节是被遮挡的。我们使用数据集的 75% 进行训练，25% 进行验证。

5.2 评价指标

2D 指标 在 Andriluka et al. (2014) 中引入的正确关键点百分比（PCK）是常用于 2D 姿态评估的指标。PCKh@t 衡量的是估计关节与真实关节之间的距离小于头部长度 t 倍的关节的百分比。根据以往的工作，我们报告 $t = \frac{1}{2}$ 时的结果。由于所用的三个基准数据集中没有提供头部长度，因此我们将其大致设置为所有基准数据集中人体边界框宽度的 2.5%。

3D 指标 3D 姿态估计准确度通过平均每关节位置误差（MPJPE）来衡量，该误差是地面真实 3D 姿态 $y = [p_1^3, \cdots, p_M^3]$ 和估计的 3D 姿态 $\bar{y} = [\bar{p_1^3}, \cdots, \bar{p_M^3}]$ 之间的欧氏距离：$\text{MPJPE} = \frac{1}{M}\sum_{i=1}^{M} \|p_i^3 - \bar{p_i^3}\|_2$，其中 M 为姿态中的关节数量。我们在计算误差时不使用 Procrustes 对估计的 3D 姿态进行对齐。这在一些工作中被称为协议 1 (Martinez et al., 2017; Tome et al., 2018)。

6 实验结果

我们将我们的方法与四个基准方法进行比较。第一个基准方法是 *NoFuse*，它独立地估计每个视角的 2D 姿态，而不进行多视角融合。第二个基准方法是 *HeuristicFuse*，它根据公式 (4) 为每个视角分配一个固定的融合权重，参数 λ 通过交叉验证设置为 0.5。第

三个基准方法是 *ScoreFuse*，它与 AdaFuse 使用相同的特征融合公式，即公式 (5)。它与 AdaFuse 的区别仅在于计算权重 ω 的方式。具体来说，ScoreFuse 将 ω 计算为热图 $\mathbf{H}$ 的最大值。我们的方法称为 *AdaFuse*，它使用预测的权重进行融合，如公式 (5) 所示。所有四种方法都使用三角测量 (Hartley and Zisserman, 2003) 从多视角的 2D 姿态估计 3D 姿态。我们还将 *RANSAC* 作为基准方法进行比较，RANSAC 不执行多视角融合，但使用 RANSAC 从三角测量中移除离群点。

6.1 Human3.6M 数据集上的结果

2D 姿态估计结果 2D 姿态估计结果见表 2。所有多视角融合方法均明显优于 *NoFuse*。其中，肘部和手腕关节的提升最为显著，因为它们经常被人体其他部分遮挡。结果表明，多视角融合是一种有效的策略来处理遮挡问题。*AdaFuse* 在所有融合方法中获得了最高的平均 准确率，验证了学习合适的融合权重可以有效减少低质量视角特征所带来的负面影响。

3D 姿态估计结果 表 3 显示了基准方法和我们方法的 3D 姿态估计误差。可以看到，*NoFuse* 的平均误差为 22.9mm，这是一个非常强的基准方法，其误差仅比当前最先进的方法稍大（见表 4）。在这个强基准的基础上，我们观察到，添加多视角融合能够进一步减少 3D 姿态估计的误差。

HeuristicFuse 的误差比 *NoFuse* 小，这与表 2 中的 2D 结果一致。平均误差仅减少了 1.9mm，因为大多数示例相对容易，改善空间较小。然而，对于一些具有挑战性的关节（如手腕），取得了显著的改善。*ScoreFuse* 的误差比 *HeuristicFuse* 更小。这意味着为低质量视角分配较小的权重有助于提高融合热图的质量。最后，我们的方法 *AdaFuse* 通过考虑外观线索和几何一致性来确定融合权重，显著地将平均误差降至 19.5mm。考虑到基准方法已经非常强大，这一改进是显著的。

我们注意到，AdaFuse 在一些关节（如臀部和头部）上的表现略有下降。这主要是因为这些关节在数据集中很少被遮挡，因此 2D 姿态估计器可以非常准确地估计这些关节的位置。进一步应用跨视角融合会给热图引入小的噪声，导致 2D 姿态估计准确度略微下降。但是，当发生遮挡时，这种情况在实际应用中非常常见，跨视角融合带来的好处远大于小噪声所带来的负面影响。

RANSAC 是解决稳健估计问题的标准方法。如表 3 所示，它通过去除三角测量中的一些离群点，优于 *NoFuse*。然而，它不如多视角融合方法有效，因为后者除了去除离群点外，还会进行精细化处理。另一个原因是，这个任务中的相机数量较少，减少了找到真实离群点的机会。此外，我们发现 *RANSAC* 对用于判断数据点是否为离群点的阈值非常敏感。在我们的实验中，阈值是通过交叉验证设置的。

Ablation Study on Fusion Weights 为了更好地理解 *AdaFuse* 带来的改进，我们将 Human3.6M 数据集的测试样本按照 *NoFuse* 的 3D 误差分成六组，然后计算每组的平均误差。图 **??** 显示了各种基准方法的结果。我们可以看到，当 *NoFuse* 的原始误差较大时，*AdaFuse* 获得了最显著的改进。然而，即使 *NoFuse* 的姿态估计已经很准确，*AdaFuse* 也能略微减少误差。

关于融合权重的消融实验 在 *ScoreFuse* 失败的典型情况下，姿态估计网络在不准确的位置产生了较大的响应。在这种情况下，*AdaFuse* 可以通过利用多视角几何一致性，超过 *ScoreFuse*。为了验证这一假设，我们在图 **??** 中分别可视化了两种方法预测的热图和对应的融合权重。我们发现，虽然第一和第三视角的位置不准确，但四个视角的热图响应都很大。*ScoreFuse* 对所有视角赋予较大的权重，最终导致了错误的热图。相比之下，*AdaFuse* 识别出第一和第三视角的预测位置与其他两个视角不一致，尽管它们的响应很大，因此降低了这些视角的权重，以确保融合热图的质量。

此外，我们还进行了 *AdaFuse* 的消融实验，单独使用两个嵌入网络中的一个。仅使用 *appearance embedding* 或 *geometry embedding* 时，3D 误差分别增加到 20.3mm 和 19.9mm。值得注意的是，改进在具有挑战性的示例中表现得尤为显著。实验结果验证了这两个嵌入是互补的。

表 4 在 Human3.6M 数据集上，现有方法和我们方法的 3D 姿态估计误差（单位：mm）。我们分别报告了 15 个动作的结果，并且计算了所有动作的平均误差。T- Iskakov et al. (2019) 表示使用了三角测量方法。V- Iskakov et al. (2019) 表示使用了体积方法。

方法	Direct	Disc.	Eat	Greet	Phone	Photo	Pose	Purch	Sit	SitD	Smoke	Wait	WalkD	Walk	WalkT	*MPJ*
Trumble et al. (2017)	92.7	85.9	72.3	93.2	86.2	101.2	75.1	78.0	83.5	94.8	85.8	82.0	114.6	94.9	79.7	87.3
Pavlakos et al. (2017)	41.2	49.2	42.8	43.4	55.6	46.9	40.3	63.7	97.6	119.0	52.1	42.7	51.9	41.8	39.4	56.9
Tome et al. (2018)	43.3	49.6	42.0	48.8	51.1	64.3	40.3	43.3	66.0	95.2	50.2	52.2	51.1	43.9	45.3	52.8
Qiu et al. (2019)	24.0	26.7	23.2	24.3	24.8	22.8	24.1	28.6	32.1	26.9	30.9	25.6	25.0	28.0	24.4	26.2
T- Iskakov et al. (2019)	20.4	22.6	20.5	19.7	22.1	20.6	19.5	23.0	25.8	33.0	23.0	21.6	20.7	23.7	21.3	22.6
V- Iskakov et al. (2019)	18.8	20.0	19.3	**18.7**	20.2	19.3	**18.7**	22.3	**23.3**	29.1	21.2	**20.3**	19.3	21.6	19.8	20.8
NoFuse	20.1	22.2	20.2	22.2	23.9	18.2	20.6	25.9	37.0	24.6	22.4	22.5	18.2	22.8	18.5	22.9
AdaFuse (我们的)	**17.8**	**19.5**	**17.6**	20.7	**19.3**	**16.8**	18.9	**20.2**	25.7	**20.1**	**19.2**	20.5	**17.2**	**20.5**	**17.3**	**19.5**

表 5 在 *Occlusion-Person* 数据集上，基线方法和我们方法对于**遮挡关节**的 2D 姿态估计准确度（PCKh@t）。我们分别报告每个关节类型的结果，并且计算所有关节类型的平均准确度。

方法	髋部	膝部	脚踝	肩部	肘部	手腕	平均
NoFuse	63.4	21.5	17.0	29.5	14.6	12.4	30.9
HeuristicFuse	76.9	59.0	73.4	63.5	49.0	54.8	65.0
ScoreFuse	90.9	88.6	88.1	86.0	93.2	86.8	89.8
AdaFuse	**96.5**	**96.0**	**92.5**	**94.1**	**98.3**	**93.2**	**95.5**

与最先进方法的比较 表 4 比较了我们的方法与最先进的技术。我们可以看到，*AdaFuse* 优于所有其他方法。需要注意的是，文献 (Iskakov et al., 2019) 中的两个方法 *Triangulation* 和 *Volumetric* 用于将 2D 姿态提升到 3D。*Triangulation* 方法与我们的方法更为接近。我们的 *AdaFuse* 将 (Iskakov et al., 2019) 的误差减少了约 13%（$\frac{22.6-19.5}{22.6}$）。考虑到最先进方法的误差已经非常小，这一改进是显著的。

6.2 Panoptic 数据集上的结果

我们评估了不同摄像头数量对该数据集的影响。图11展示了分别使用三到六个摄像头时的平均 3D 误差。通常情况下，当更多摄像头被使用时，大多数基线方法的误差都会减少。然而，我们观察到，当摄像头数量从三个增加到四个时，*NoFuse* 的误差反而变大。这一不理想的现象发生的原因是新增的摄像头视角非常具有挑战性，因此 2D 姿态估计结果不准确。然而，对于我们的方法 *AdaFuse*，由于自适应多视图融合的作用，低质量热图对单个视图的负面影响得到了有效限制。我们可以看到，当摄像头数量增加时，*AdaFuse* 的误差持续减小。由于目前没有广泛采用的评估协议，并且很少有工作报告了该新数据集上的结果，我们没有将我们的结果与其他方法进行比较。

6.3 Occlusion-Person 数据集上的结果

2D 姿态估计结果 表5展示了对于遮挡关节的结果。仅约 30.9% 的遮挡关节能够被 *NoFuse* 准确检测到。这个结果是合理的，因为遮挡关节的特征受到了严重的损坏。三种多视图融合方法显著提高了准确度。特别地，*AdaFuse* 正确检测了超过 90% 的遮挡关节。结果表明，我们的学习融合权重策略具有优势。

3D 姿态估计结果 我们在表6中展示了每个关节类型的 3D 姿态估计误差（单位：mm）。*NoFuse* 导致了 48.1mm 的较大误差。通过提高对遮挡关节的 2D 姿态估计结果，3D 误差显著减小，特别是对于肢体上的关节，如脚踝和手腕。特别地，我们的方法将 3D 误差显著降低到 12.6mm。

遮挡视角数量的影响 我们还评估了遮挡视角数量对该数据集的影响。具体地，我们根据每个关节的遮挡视角数量将其分类为五组，并分别报告每组的平均关节误差。结果如表7所示。当关节在所有视角中可见时，简单的基线方法 *NoFuse* 也能实现 13.0mm 的较小误差。然而，当四个视角被遮挡时，误差急剧增加到 82.6mm。回想一下，该数据集总共有八个

表 6 在 *Occlusion-Person* 数据集上，基线方法和我们方法的 3D 姿态估计误差（单位：*mm*）。我们分别报告每个关节的结果，并且计算所有关节的平均误差。第二行展示了每种关节类型的遮挡百分比。

	根部	腹部	颈部	髋部	膝部	脚踝	肩部	肘部	手腕	平均
遮挡（%）	14.3%	13.7%	7.6%	23.0%	25.0%	23.5%	16.8%	25.3%	21.7%	
NoFuse	10.0	12.2	12.5	16.8	61.1	113.9	28.0	63.7	60.3	48.1
HeuristicFuse	8.8	10.7	**11.5**	14.2	21.1	19.2	17.5	23.6	24.1	18.0
ScoreFuse	8.4	12.6	12.6	14.7	17.5	17.1	16.1	13.2	16.9	15.0
RANSAC	8.6	11.2	11.7	12.9	18.8	17.9	17.1	14.5	19.7	15.5
AdaFuse (我们的)	**7.2**	**10.6**	**11.6**	**11.7**	**13.8**	**15.7**	**14.2**	**9.9**	**14.4**	**12.6**

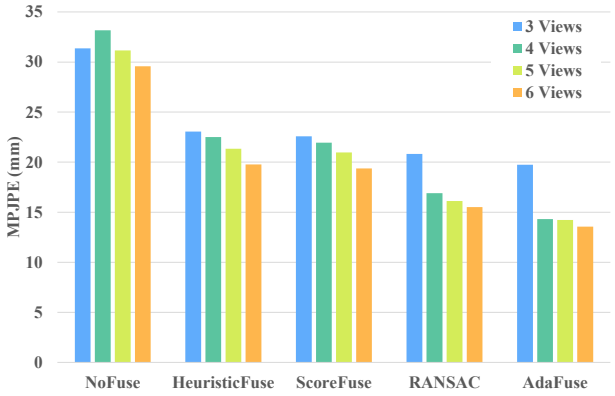

图 11 使用不同数量摄像头时，Panoptic 数据集上的 3D 姿态估计误差。

表 7 在 Occlusion-Person 数据集上，基线方法和我们方法的 3D 姿态估计误差（单位：mm）。我们根据遮挡视角的数量（总共有 8 个视角）对 3D 关节进行了分组。第二行显示了每组的关节数量百分比。

遮挡视角数量	4	3	2	1	0
百分比	2%	15%	38%	35%	10%
NoFuse	82.6	70.2	59.7	33.7	13.0
HeuristicFuse	30.5	19.9	15.9	13.5	11.1
ScoreFuse	25.0	18.1	15.2	13.4	12.6
RANSAC	36.5	24.5	19.4	14.3	11.7
AdaFuse (我们的)	**21.7**	**14.8**	**12.5**	**11.5**	**10.8**

视角。相比之下，多视图融合方法，特别是我们的 *AdaFuse*，在更多视角被遮挡时，始终保持比 *NoFuse* 更小的误差。更重要的是，当更多的摄像头视角被遮挡时，误差增长的速度明显比 *NoFuse* 慢，这验证了我们方法在遮挡情况下的鲁棒性。

泛化能力 我们融合方法中唯一可学习的参数是外观嵌入和几何嵌入网络中的参数。在本节中，我们评估了在 *Occlusion-Person* 数据集上训练的 *AdaFuse* 权重预测网络是否可以直接应用到其他数据集上。特别地，我们将 *AdaFuse* 在 *Occlusion-Person* 上学到的权重预测网络附加到分别在每个数据集上训练的 2D 姿态估计器上，作为最终模型进行评估。表8展示了在不同数据集上的 3D 姿态估计结果。我们发现，基于合成的 *Occlusion-Person* 数据集训练的融合网络在三个现实数据集上的表现与分别在每个目标数据集上训练的网络相比，取得了类似的性能。这些令人鼓舞的结果验证了我们的融合模型具有很强的泛化能力。值得注意的是，我们的方法可以自然地处理不同数量的摄像头视角，原因有二：首先，外观嵌入网络和几何嵌入网络中的参数对于所有视角都是共享的；其次，几何嵌入网络中的“平均”操作使得该网络不依赖于视角的数量，如图7和图8所示。总之，*AdaFuse* 能够在不同的摄像头姿态环境中无须额外适应地部署。

6.4 Total Capture 数据集的结果

我们在表9中报告了 Total Capture 数据集上的 3D 姿态估计结果。值得注意的是，一些方法除了多视角图像外，还使用了 IMU（惯性测量单元）。我们可以看到，我们的方法超越了所有之前的方法。我们注意到，对于 S4,5 中的“W2（行走）”动作，我们的方法的误差略大于 LSTM-AE (Trumble et al., 2018)。我们倾向于认为，这是因为 LSTM 在应用于“行走”等周期性动作时能够显著受益，这一点也在另一项工作中独立观察到 (Gilbert et al., 2019)。

我们在图12中展示了一些 3D 姿态估计示例。在大多数情况下，我们的方法能够准确估计 3D 姿态。

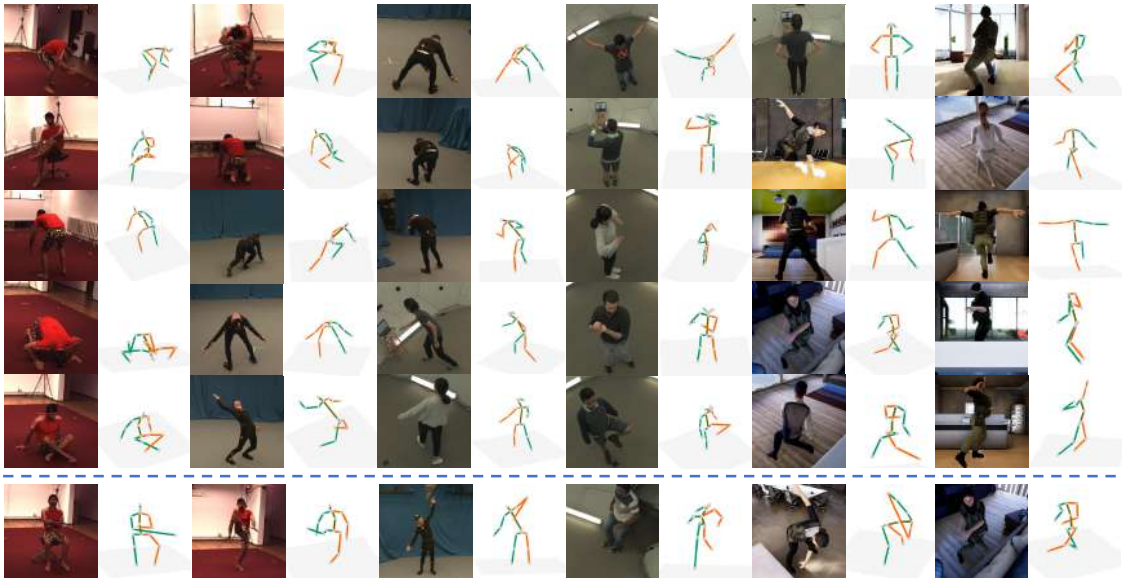

图 12 我们展示了通过 *AdaFuse* 获得的一些 3D 姿态估计示例。最后一行展示了一些失败的案例。

表 8 在不同数据集上训练 *AdaFuse* 权重预测网络时，3D 姿态估计误差 MPJPE（单位：mm）。*AdaFuse* 在 *Occlusion-Person* 或直接在 Evaluation 数据集上训练时的结果。用于生成初始热图的 2D 姿态估计器分别在每个 Evaluation 数据集上单独训练。

Evaluation 数据集	AdaFuse 训练于		NoFuse	HeuristicFuse	ScoreFuse	RANSAC
	Evaluation 数据集	Occlusion-Person				
Human3.6M	19.5	19.4	22.9	21.0	20.1	21.8
Panoptic 4 视角	14.7	14.6	33.2	22.5	21.9	16.9
Panoptic 6 视角	13.6	13.9	29.6	19.8	19.4	15.5
Total Capture	19.2	20.1	29.4	20.0	20.5	20.5

一个典型的失败情况是，当许多摄像头视角的 2D 姿态估计结果不准确时。例如，在 Panoptic 数据集中，当人类开始进入圆顶时，他们可能在多个视角中被遮挡。在这种情况下，每个视角中的热图质量较低。因此，融合后的热图质量也会降低，导致不准确的 2D 姿态估计。

7 总结与未来工作

我们提出了一种多视角融合方法 *AdaFuse*，用于解决人体姿态估计中的遮挡问题。*AdaFuse* 具有很高的实用价值，因为它非常简单且能够灵活地应用于新的环境，无需额外的适应性调整。此外，它可以与任何 2D 姿态估计网络结合使用。我们在三个基准数据集上广泛评估了该方法的有效性，结果表明该方法显著超越了现有的最先进方法。我们还构建了一个大规模的人体数据集，其中包含严重的遮挡情况，以促进该方向的进一步研究。我们的下一步工作是利用时间序列信息，进一步提升姿态估计的准确性。

参考文献

Amin S, Andriluka M, Rohrbach M, Schiele B (2013) Multi-view pictorial structures for 3D human pose estimation. In: BMVC

Andriluka M, Pishchulin L, Gehler P, Schiele B (2014) 2D human pose estimation: New benchmark and state of the art analysis. In: CVPR, pp 3686–3693

表 9 在 Total Capture 数据集上，不同方法的 3D 姿态估计误差 MPJPE（单位：mm）。

方法	IMUs	时间序列	受试者 (S1,2,3)			受试者 (S4,5)			平均
			W2	A3	FS3	W2	A3	FS3	
(Trumble et al., 2017)	✓	✓	48.3	94.3	122.3	84.3	154.5	168.5	107.3
(Wei et al., 2016)			79.0	106.5	112.1	79.0	73.7	149.3	99.8
(Gilbert et al., 2019)	✓		19.2	42.3	48.8	24.7	58.8	61.8	42.6
(Trumble et al., 2018)		✓	13.0	23.0	47.0	**21.8**	40.9	68.5	34.1
(Qiu et al., 2019)			19	21	28	32	33	54	29
NoFuse			15.9	18.5	29.9	33.9	33.8	60.0	29.4
HeuristicFuse			7.8	11.6	19.6	23.3	26.9	44.8	20.0
ScoreFuse			9.7	13.1	19.9	23.9	27.2	41.4	20.5
RANSAC			8.4	11.6	20.5	23.3	27.2	45.7	20.5
AdaFuse (我们的方法)			**7.2**	**10.8**	**18.5**	22.8	**26.6**	**42.9**	**19.2**

Belagiannis V, Amin S, Andriluka M, Schiele B, Navab N, Ilic S (2014) 3d pictorial structures for multiple human pose estimation. In: CVPR, pp 1669–1676

Bo L, Sminchisescu C (2010) Twin gaussian processes for structured prediction. IJCV 87(1-2):28

Bridgeman L, Volino M, Guillemaut JY, Hilton A (2019) Multi-person 3d pose estimation and tracking in sports. In: CVPRW, pp 0–0

Burenius M, Sullivan J, Carlsson S (2013) 3D pictorial structures for multiple view articulated pose estimation. In: CVPR, pp 3618–3625

Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR, pp 7291–7299

Chen W, Wang H, Li Y, Su H, Wang Z, Tu C, Lischinski D, Cohen-Or D, Chen B (2016) Synthesizing training images for boosting human 3d pose estimation. In: 3DV, IEEE, pp 479–488

Cheng Y, Yang B, Wang B, Yan W, Tan RT (2019) Occlusion-aware networks for 3d human pose estimation in video. In: ICCV, pp 723–732

Ci H, Wang C, Ma X, Wang Y (2019) Optimizing network structure for 3d human pose estimation. In: ICCV, pp 915–922

Ci H, Ma X, Wang C, Wang Y (2020) Locally connected network for monocular 3d human pose estimation. In: T-PAMI

Dong J, Jiang W, Huang Q, Bao H, Zhou X (2019) Fast and robust multi-person 3d pose estimation from multiple views. In: CVPR, pp 7792–7801

Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6):381–395

Gal Y (2016) Uncertainty in deep learning. PhD thesis, PhD thesis, University of Cambridge

Gal Y, Ghahramani Z (2015) Dropout as a bayesian approximation: Insights and applications. In: Deep Learning Workshop, ICML, vol 1, p 2

Gall J, Rosenhahn B, Brox T, Seidel HP (2010) Optimization and filtering for human motion capture. IJCV 87(1-2):75

Ghahramani Z (2016) A history of bayesian neural networks. In: NIPS Workshop on Bayesian Deep Learning

Gilbert A, Trumble M, Malleson C, Hilton A, Collomosse J (2019) Fusing visual and inertial sensors with semantics for 3d human pose estimation. IJCV 127(4):381–397

Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: ICML, JMLR. org, pp 1321–1330

Hartley R, Zisserman A (2003) Multiple view geometry in computer vision. Cambridge university press

He Y, Zhu C, Wang J, Savvides M, Zhang X (2019) Bounding box regression with uncertainty for accurate object detection. In: CVPR, pp 2888–2897

Hoffmann DT, Tzionas D, Black MJ, Tang S (2019) Learning to train with synthetic humans. In: German Conference on Pattern Recognition, Springer, pp 609–623

Ilg E, Cicek O, Galesso S, Klein A, Makansi O, Hutter F, Brox T (2018) Uncertainty estimates and multi-hypotheses networks for optical flow. In: ECCV, pp 652–667

Ionescu C, Papava D, Olaru V, Sminchisescu C (2014) Human3. 6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. T-PAMI pp 1325–1339

Iskakov K, Burkov E, Lempitsky V, Malkov Y (2019) Learnable triangulation of human pose. arXiv preprint arXiv:190505754

Joo H, Simon T, Li X, Liu H, Tan L, Gui L, Banerjee S, Godisart T, Nabbe B, Matthews I, et al. (2019) Panoptic studio: A massively multiview system for social interaction capture. T-PAMI 41(1):190–204

Kendall A, Gal Y (2017) What uncertainties do we need in bayesian deep learning for computer vision? In: NIPS, pp 5574–5584

Kreiss S, Bertoni L, Alahi A (2019) Pifpaf: Composite fields for human pose estimation. In: CVPR, pp 11977–11986

Lakshminarayanan B, Pritzel A, Blundell C (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. In: NIPS, pp 6402–6413

Lassner C, Romero J, Kiefel M, Bogo F, Black MJ, Gehler PV (2017) Unite the people: Closing the loop between 3d and 2d human representations. In: CVPR, pp 6050–6059

Li T, Fan L, Zhao M, Liu Y, Katabi D (2019) Making the invisible visible: Action recognition through walls and occlusions. In: ICCV, pp 872–881

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: ECCV, Springer, pp 740–755

Liu Y, Stoll C, Gall J, Seidel HP, Theobalt C (2011) Markerless motion capture of interacting characters using multi-view image segmentation. In: CVPR, IEEE, pp 1249–1256

Malleson C, Gilbert A, Trumble M, Collomosse J, Hilton A, Volino M (2017) Real-time full-body motion capture from video and imus. In: 3DV, IEEE, pp 449–457

von Marcard T, Henschel R, Black MJ, Rosenhahn B, Pons-Moll G (2018) Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV, pp 601–617

Martinez J, Hossain R, Romero J, Little JJ (2017) A simple yet effective baseline for 3D human pose estimation. In: ICCV, p 5

Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Computer vision and image understanding 104(2-3):90–126

Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: ECCV, Springer, pp 483–499

Pavlakos G, Zhou X, Derpanis KG, Daniilidis K (2017) Harvesting multiple views for marker-less 3D human pose annotations. In: CVPR, pp 1253–1262

Pavlakos G, Zhou X, Daniilidis K (2018) Ordinal depth supervision for 3d human pose estimation. In: CVPR, pp 7307–7316

Pavllo D, Feichtenhofer C, Grangier D, Auli M (2019) 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR, pp 7753–7762

Peng X, Tang Z, Yang F, Feris RS, Metaxas D (2018) Jointly optimize data augmentation and network training: Adversarial data augmentation in human

pose estimation. In: CVPR, pp 2226–2234

Perez P, Vermaak J, Blake A (2004) Data fusion for visual tracking with particles. Proceedings of the IEEE 92(3):495–513

Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ (2017) On fairness and calibration. In: NIPS, pp 5680–5689

Qiu H, Wang C, Wang J, Wang N, Zeng W (2019) Cross view fusion for 3d human pose estimation. In: ICCV, pp 4342–4351

Qiu W, Zhong F, Zhang Y, Qiao S, Xiao Z, Kim TS, Wang Y (2017) Unrealcv: Virtual worlds for computer vision. In: Proceedings of the 25th ACM international conference on Multimedia, ACM, pp 1221–1224

Rhodin H, Spörri J, Katircioglu I, Constantin V, Meyer F, Müller E, Salzmann M, Fua P (2018) Learning monocular 3d human pose estimation from multi-view images. In: CVPR, pp 8437–8446

Roetenberg D, Luinge H, Slycke P (2009) Xsens mvn: full 6dof human motion tracking using miniature inertial sensors. Xsens Motion Technologies BV, Tech Rep 1

Rogez G, Schmid C (2016) Mocap-guided data augmentation for 3d pose estimation in the wild. In: NIPS, pp 3108–3116

Sigal L, Balan AO, Black MJ (2010) Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV 87(1-2):4

Starner T, Leibe B, Minnen D, Westyn T, Hurst A, Weeks J (2003) The perceptive workbench: Computer-vision-based gesture tracking, object tracking, and 3d reconstruction for augmented desks. Machine Vision and Applications 14(1):59–71

Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: CVPR, pp 5693–5703

Sun X, Xiao B, Wei F, Liang S, Wei Y (2018) Integral human pose regression. In: ECCV, pp 529–545

Tome D, Toso M, Agapito L, Russell C (2018) Rethinking pose in 3D: Multi-stage refinement and recovery for markerless motion capture. In: 3DV, pp 474–483

Trumble M, Gilbert A, Malleson C, Hilton A, Collomosse J (2017) Total capture: 3D human pose estimation fusing video and inertial sensors. In: BMVC, pp 1–13

Trumble M, Gilbert A, Hilton A, Collomosse J (2018) Deep autoencoder for combined human pose estimation and body model upscaling. In: ECCV, pp 784–800

Tu H, Wang C, Zeng W (2020) Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: ECCV, pp 1–16

Varol G, Romero J, Martin X, Mahmood N, Black MJ, Laptev I, Schmid C (2017) Learning from synthetic humans. In: CVPR, pp 109–117

Wei SE, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: CVPR, pp 4724–4732

Xiang D, Joo H, Sheikh Y (2019) Monocular total capture: Posing face, body, and hands in the wild. In: CVPR

Xiao B, Wu H, Wei Y (2018) Simple baselines for human pose estimation and tracking. In: ECCV, pp 466–481

Xie R, Wang C, Wang C (2020) Metafuse: A pre-trained fusion model for human pose estimation. In: CVPR

Yang W, Ouyang W, Wang X, Ren J, Li H, Wang X (2018) 3d human pose estimation in the wild by adversarial learning. In: CVPR, pp 5255–5264

Zafar U, Ghafoor M, Zia T, Ahmed G, Latif A, Malik KR, Sharif AM (2019) Face recognition with bayesian convolutional networks for robust surveillance systems. EURASIP Journal on Image and Video Processing 2019(1):10

Zhang Z, Wang C, Qin W, Zeng W (2020) Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In: CVPR, pp 2200–2209

Zhao M, Li T, Abu Alsheikh M, Tian Y, Zhao H, Torralba A, Katabi D (2018) Through-wall human pose estimation using radio signals. In: CVPR, pp 7356–7365

Zhao M, Liu Y, Raghu A, Li T, Zhao H, Torralba A, Katabi D (2019) Through-wall human mesh recovery using radio signals. In: ICCV, pp 10113–10122

Zhou X, Huang Q, Sun X, Xue X, Wei Y (2017) Towards 3D human pose estimation in the wild: a weakly-supervised approach. In: ICCV, pp 398–407