Aerial View 3D Human Pose Estimation using Double Vector Quantized-Variational AutoEncoders

Juheon Hwang
Yonsei University
Seoul, South Korea
consecrated.h@yonsei.ac.kr

Jiwoo Kang*
Sookmyung Women's University
Seoul, South Korea
jwkang@sookmyung.ac.kr

Abstract

This study introduces a novel methodology for the precise estimation of the three-dimensional (3D) pose of individuals based on images captured from aerial viewpoints, particularly from top-to-bottom viewpoints. A motion capture system utilized for surveillance purposes is frequently constrained in its ability to capture dynamic scenarios, primarily due to the limited field of view of a third-person-view camera. To address the problem at hand, various approaches employ aerial views to overcome limitations in spatial constraints. Nevertheless, when observing the unmanned aerial vehicle (UAV) from an aerial perspective, it is common for the lower body to appear diminished and obstructed by the upper body. This phenomenon results in pose estimation that is highly unreliable and inaccurate. To overcome the existing limitation, we present a novel approach that utilizes the Vector Quantized-Variational AutoEncoder (VQ-VAE) to accurately predict and optimize the 3D human pose from aerial images. Thus, we introduce a novel pipeline for pose estimation and optimization using the codebook by learning aerial image features and pose features from large human pose datasets with VQ-VAE. The proposed method with the vector quantizer of VQ-VAEs can help improve the generalization capabilities of 3D pose estimation from aerial top-to-bottom viewpoints. Through conducting comparative experiments, our method has demonstrated a substantial enhancement in performance compared to those of existing state-of-the-art methods.

1. Introduction

In recent years, many graphics applications, such as games, augmented reality (AR), and virtual reality (VR), are based on 3D human poses. As technology advances, not only graphics applications but also dangerous situations through action recognition of 3D human poses are being applied to various surveillance situations. Previously studied 3D human pose estimation methods are broadly divided into marker-based motion capture (MoCap) and markerless systems using only RGB cameras [6, 15, 20, 24]. The markerless system is used when the target is unable to wear active or passive markers, which is a standard method for 3D pose estimation in surveillance situations. Since these markerless systems use calibrated multi-view cameras, there is a strong physical requirement that the subject never leaves the fixed recording volume inside the laboratory. However, in various surveillance situations, the target does not stay within the fixed recording volume, rather the target moves wherever they want, thus the use of a markerless system with multi-view cameras is not appropriate.

For this reason, some methods utilize the view of an unmanned aerial vehicle (UAV) to avoid spatial constraints of capturing the target. The images captured from the top-to-bottom view on the UAV include the normal camera view, but also those captured from right above the target, resulting in significant occlusion. Since the situation of captur-

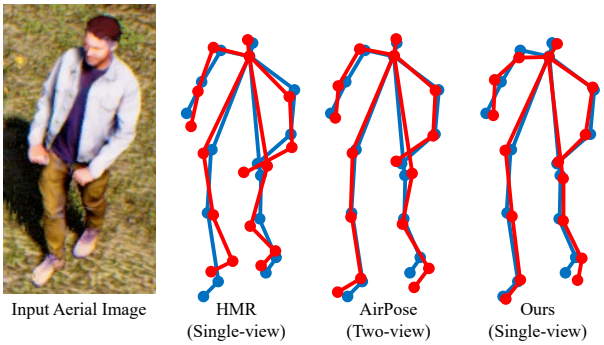

Figure 1. Predicted 3D poses (red) are overlaid on the ground truth (blue). Although our method uses a single-view as input, it is comparable to the results of AirPose [28] using multiple views, and shows significantly more accurate pose prediction performance than HMR [10] using the same single-view.

*Corresponding author.

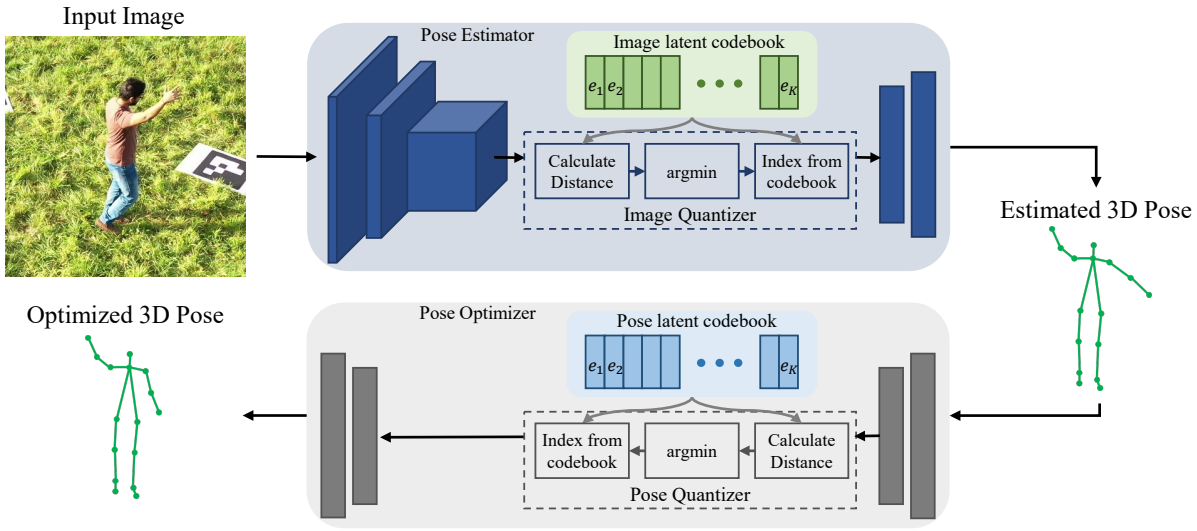

Figure 2. An architecture overview of the proposed method. The pose estimator takes a single-view aerial image as input and extracts a 2D joint heatmap from the image. The features are then extracted from the heatmap to obtain the closest latent vector in the codebook, and this is used to estimate the 3D pose. The pose optimizer takes as input an estimated 3D pose with increased channels for coordinates via positional encoding to extract features. From these features, the closest latent vector in the codebook is then retrieved and used to predict the optimized 3D pose. These consecutive processes for 3D pose estimation increase the generalization performance for unseen or highly occulded data, allowing accurate predictions of 3D human poses for top-to-bottom viewpoint images.

ing the target from the top viewpoint is not common, 3D pose estimation for such images using conventional methods [10] (HMR) is challenging, as shown in Fig. 1, where AirPose [28] uses multi-view UAV images for 3D pose estimation. However, this method has a strong restriction of multi-view and, in practical situations, it is common to use only one UAV. Therefore, we restrict ourselves to accurately predicting the 3D pose from a top-to-bottom view of the target, from images captured by a single UAV.

Images taken from the top-to-bottom view often have more occlusion than normal, similar to egocentric images. Thus, several efforts have been made to address the problem in 3D pose estimation from the top-to-bottom view images. Xu *et al*. [41], predicted the 3D pose by predicting the depth of each joint based on the heatmap of the 2D joints predicted from the input image. Tome *et al*. [34] predicted the 3D pose by regressing the location of 3D joints on a heatmap. However, these methods have the limitation that the 3D pose determined by self-occlusion is not accurate. Wang et al. [37] proposed a method in which 3D poses are estimated against existing methods [34, 41]. Then they trained a 3D pose optimizer using a structure of Variational AutoEncoder (VAE) pre-trained on a large-scale human 3D pose dataset (*i.e*., AMASS [17]) to produce an accurate 3D pose from the top-to-bottom viewpoint.

Despite these efforts, arm or leg joint positions are often incorrectly predicted by 3D human pose estimation methods. In order to overcome these problems, we employed Vector Quantized-Variational AutoEncoder (VQ-VAE) to predict and optimize human position from photographs captured from aerial viewpoints. Here, we introduce a novel pipeline for pose estimation and optimization using the codebook by learning image features and pose features from aerial-view images with VQ-VAE. The utilization of the vector quantizer in VQ-VAE has been proposed as a method to enhance the generalization capabilities of 3D pose estimation when dealing with top-to-bottom viewpoint images. Our method shows a significant performance gain for 3D pose estimation thanks to the vector quantizer of VQ-VAEs. An example of pose estimation comparisons between the proposed and current state-of-the-art methods from aerial viewpoint image is illustrated in Fig. 1. We used an image in AirPose [28] synthetic data, where human images with the corresponding poses are constructed from the view of an unmanned aerial vehicle (UAV). The outperformed performance demonstrates the generalization ability in 3D pose estimation over current state-of-the-art pose estimation methods.

2. Related work

2.1. 3D Human Pose Estimation

Markerless human pose estimation is a well-established and extensively studied field within the domain of computer vision. Among the various approaches for 3D human pose estimation, two prominent methods are those utilizing multi-view cameras and those employing monocular cameras. Methods that employ multi-view cameras [5, 6, 39, 42]

use calibrated information from cameras to transform the 2D human pose, predicted from images captured by each camera, into a 3D pose using bundle adjustment [35]. In recent years, there has been significant progress in the field of predicting 3D pose using features extracted from multiple multi-view images. In particular, the methods in [39, 44] have conducted research on this topic, employing deep learning techniques, such as transformers. These studies have demonstrated the effectiveness of utilizing deep learning methods for predicting 3D pose based on multi-view image features.

Furthermore, research has been conducted on the development of techniques utilizing monocular cameras to progress from merely predicting a 2D pose on a 2D image to predicting a 3D pose, aligning with the advancements in deep learning. Methods for predicting the 3D pose directly from an image have been extensively investigated in two main directions. The first direction involves predicting the depth of a predicted 2D human pose [21, 41, 45]. The second direction focuses on predicting the 3D pose of an image through regression [18, 31, 34]. An alternative methodology involves the extraction of a 2D pose from an image, followed by lifting the 2D pose to a 3D pose through the utilization of a new neural network. These approaches have used fully connected networks [4, 19], temporal convolutional networks [2, 26], graph convolutional networks [1, 3, 38], and transformer networks [13, 30]. Finally, a method has been developed to accurately align a human parametric mesh with a 2D image by utilizing the SMPL model [16] . This approach enables the estimation of a human's 3D pose and mesh [10]. To tackle the challenge of achieving human 3D pose estimation in challenging scenarios, such as capturing a target from an aerial viewpoint in a top-to-bottom view, we employed the following methods [34, 37, 41], as comparison methods, which focus on egocentric view 3D human pose estimation, have the most similar domain to the top-to-bottom view. We conducted a comparative analysis of our method with AirPose [28] as a reference method, as well as HMR [10], which serves as the baseline of that method, although it does not use a single view image for effective comparisons.

2.2. Flying Motion Capture Systems

As the utilization of aerial robots continues to gain popularity, the graphics community has recently put forth a range of tools and algorithms aimed at planning physically realistic quadrotor camera trajectories for aerial videography. These tools facilitate the strategic design of aerial shots within a 3D virtual environment, employing offline optimization techniques to consider both aesthetic objectives and constraints related to robot modeling. The methods presented in [9] and [7] generate quadrotor trajectories based on user-defined space-time keyframes. On the other hand, the method proposed in [27] takes a physically infeasible trajectories by adjusting the velocity according to a non-linear quadrotor model, resulting in the computation of the closest feasible trajectory. The work in [7] examines the various factors that influence the recognition of aerial images and presents an optimization scheme derived from the findings. All of the aforementioned methods are offline in nature, rendering them incapable of generating control inputs suitable for application in a dynamic environment. Utilizing model predictive control (MPC) formulations, the work in [22] optimizes cinematic constraints, such as visibility and position on the screen, considering the robot constraints for a single quadrotor. The work of [22] extends this work to multiple drones to facilitate actor-centric tracking along geometric paths. In the field of robotics, researchers have proposed various methods for reconstructing 3D trajectories of individuals in motion using a camera mounted on a micro-automated vehicle (MAV) while simultaneously mapping the surrounding environment [12, 14]. In contrast, the objective of this paper is to reconstruct the full 3D body pose of a subject in motion, while simultaneously planning the trajectory of the MAV in such a way that markers become visible. To successfully complete this task, it is necessary to employ multiple quadrotors and accurately estimate their positions, as well as the positions of the skeletal joints.

Recently, researchers have been utilizing a single UAV equipped with a camera to develop two novel technologies: Flycons [23] and Drocap [46]. Flycon necessitates the utilization of LED markers on subjects, which employ the advanced infrared-based MoCap algorithm. The Drawcap is a markerless approach that employs a low-latency fitting-based method to calculate the 3D skeleton of the subject. Additionally, the UAV remains in a fixed position throughout the entire sequence. Flycap [40] uses RGB-D cameras mounted on multiple UAVs within an indoor setting to generate a sequence of 3D point clouds for reconstruction purposes over a period of time. AirCap [32] proposed a method for autonomous UAV formation to capture multi-view imagery and optimize 3D poses and geometry for offline [29] using onboard GPS-based self-localization for accurate positioning. In the present study, we introduce a novel methodology for accurately estimating 3D poses using a single aerial view image.

3. Method

3.1. Architecture

The schematic representation of the proposed methodology for human 3D pose estimation is illustrated in Fig. 2. Our methodology endeavors to accurately predict human 3D pose of humans from an aerial view image captured by a UAV mounted camera. To achieve our objective, our approach involves utilizing a pose estimator that predicts a 3D

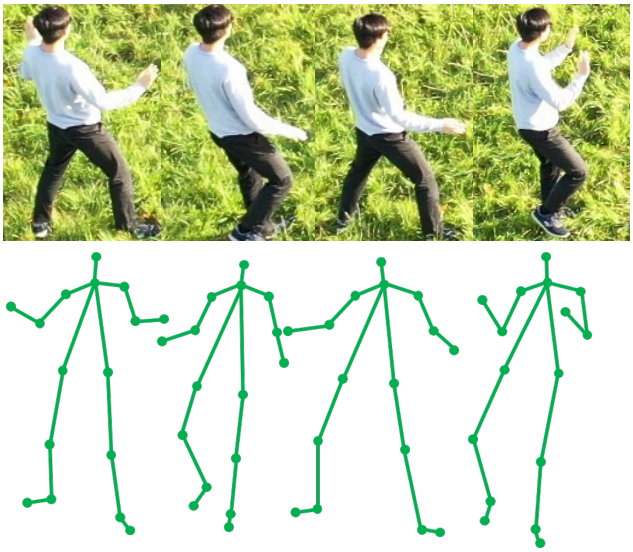

Figure 3. 3D human pose estimation results on AirPose real data. Our method shows accurate 3D pose estimation performance even in situations where some parts of the target's body are partially occluded.

pose from an input aerial image. Additionally, we employ a pose optimizer to generate a 3D pose that closely resembles the one predicted by the pose estimator as input. The pose estimator is trained by utilizing a dataset consisting of aerial images and corresponding human 3D pose pairs, while the pose optimizer is trained on a varied and comprehensive collection of human 3D poses. The estimator and optimizer employ the VQ-VAE framework to learn a codebook that captures the overall distribution of features derived from aerial viewpoint images and human 3D poses. Our methodology enables the prediction and optimization of the pose for a new image or pose by utilizing the latent code that is most similar to it. Thus, this approach effectively provides valuable insight into the distribution of the dataset through the codebook.

3.1.1 Pose Estimator Architecture

The pose estimator estimates a human 3D pose from an aerial image. The initial stage of the pose estimator involves the utilization of a variation of ResNet [8] to predict the heatmap. Additionally, a convolutional neural network (CNN) and fully connected layers (FC) are used to extract the features of the aerial image, with the predicted heatmap serves as input. In the intermediate stage, the convolutional neural network (CNN) and fully connected layers (FC) are used to extract features from aerial images. These features are then utilized to train a codebook for aerial image features using the Vector Quantization Variational Autoencoder (VQ-VAE) method. Subsequently, the latent code that closely resembles the features extracted from the codebook is retrieved. In the final stage, the human 3D pose is predicted by regressing it on the most closely related latent code from the VQ-VAE. We adopt the network architectures proposed in [34] for pose estimation and one in [36] for VQ-VAE.

3.1.2 Pose Optimizer Architecture

The pose optimizer is employed to address the issue of inaccuracies arising from monocular aerial images and self-occluded regions. The pose optimizer utilizes a process to extract the features of the estimated human 3D pose. These features are then transferred to the pose latent space, where the optimizer retrieves the most similar latent code from the codebook to recover the human 3D pose. Here, the VQ-VAE utilized by the pose optimizer is not trained on an aerial image dataset, but rather on a comprehensive real-world human 3D pose dataset. This enables the model to learn the codebook for various human motions. The proposed method facilitates the restoration of a natural pose based on a 3D human pose in the real world. Instead of directly using the locations of the 3D joints as input, we used positional encoding (PE) [33] to enhance the channel of the location. We adopt the architectural framework of the pose optimizer as described in [36].

3.2. Losses

Our method uses two loss functions for pose estimation [34] and pose optimization [43].

3.2.1 Pose Estimation Loss

First, the pose estimator is trained using the following loss function as

$$L_{pose} = L_{2D} + \lambda_{AE} L_{AE} + \lambda_{vq} L_{vq} \tag{1}$$

where L_{2D} is the 2D pose detection loss, L_{AE} is the autoencoding loss, and L_{vq} is the vector quantization loss [36]. The objective of the 2D pose detection loss L_{2D} is to facilitate the training of the ResNet component responsible for predicting the heatmap of the aerial image. The autoencoding loss L_{AE} helps the pose estimator by aiding in the prediction of the pose from the heatmap.

3.2.2 Pose Optimizer Loss

We use the vector quantization loss L_{vq} in (1) for pose estimation. In other words, we use the vector quantization loss L_{vq} in (2) only in the training process of the pose estimator. In detail, the vector quantization loss L_{vq} is composed of three loss terms as

$$L_{vq} = L_{re} + \lambda_{embed} L_{embed} + \lambda_{commit} L_{commit} \tag{2}$$

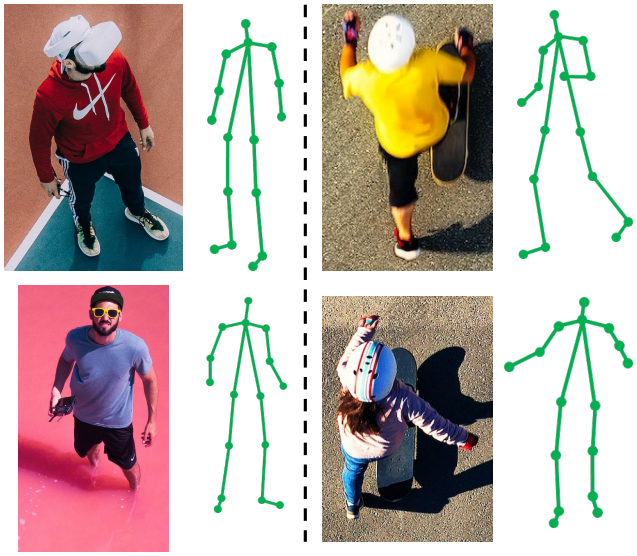

Figure 4. 3D human pose estimation results on real-world image. Our method predicts accurate 3D poses from a variety of real-world images. This demonstrates the high generalization performance of our proposed method. © Martin Sanchez (left column) and Red Zeppelin (right column) uploaded in Unsplash.

where L_{re} is the reconstruction loss, L_{embed} is the embedding loss, and L_{commit} is the commitment loss. The reconstruction loss L_{re} optimizes the VQ-VAE encoder and decoder of the VQ-VAE model, aiming to minimize the discrepancy between input and output. The embedding loss L_{embed} makes the learning of feature vector values extracted by the encoder within the codebook's latent code. The commitment loss L_{commit} ensures that the features extracted by the encoder closely align with the latent code in the codebook.

4. Experiments

4.1. Datasets

We used the AirPose dataset [28] and the AMASS dataset [17] for the pose estimator and pose optimizer, respectively. The AirPose dataset [28] was constructed for multi-view aerial 3D human pose estimation and consists of synthetic and real-world datasets. The synthetic dataset was generated using Unreal Engine (UE) to render realistic human scans [25] from the viewpoints of two UAVs. The synthetic dataset consists of a total of 60,000 images (30,000 images per UAV) and is accompanied by SMPL-X fittings for the scans, from which 3D poses can be generated. The real-world dataset consists of two real data sequences acquired by two DJI Mavic UAVs equipped with RGB cameras. One of the UAVs hovers in place, and the other one is manually flown around the target. We used the AirPose synthetic dataset for training.

The AMASS dataset [17] is a comprehensive collection that integrates 15 distinct motion capture datasets. In the AMASS dataset, a thorough manual inspection was carried out to correct and integrate all data, specifically focusing on identifying and rectifying any instances of swapped or mislabeled joint markers in human 3D poses. The dataset comprises a total of 344 subjects, encompassing 11,265 motions and spanning a duration of 40 hours of recordings.

4.2. Implementation Details

We used 15 of the 23 SMPL joints for our experiments, excluding hands, collars, spines, and pelvis for simplicity. This simplification helps to obtain more results from self-occluded images.

Similarly to AirPose [28], the original was not used directly for the input image, but the bounding box of the target human was cropped and the image was resized to 224×224 to ensure that the proportions of the human did not change. AirPose uses the SMPL parameters predicted by HMR [10] with the corresponding images as initial input. Still, our method first uses only human images as input to predict the heatmap for the 2D joint location. Thus, we generated the 2D joint heatmap from the dataloader for training. Similarly, our method directly predicts 3D human pose, and thus, unlike AirPose, it extracts and uses 3D joint locations from SMPL parameters.

4.3. Evaluation Metrics

Similar to previous studies [37], we conducted a comparative analysis of the performance of our method using two different metrics: Procrustes Analysis-Mean Per Joint Position Error (PA-MPJPE) and Bone length Aligned-Mean Per Joint Position Error (BA-MPJPE). PA-MPJPE aligns the estimated pose and the ground truth pose using Procrustes analysis [11] and subsequently computes the average error between the joint positions. BA-MPJPE calculates the average discrepancy in joint positions between a resized estimated pose and a ground truth pose. This alignment is achieved through the Procrustes analysis, which scales the poses to match the bone length of a standard skeleton. Therefore, PA-MPJPE assesses the accuracy of the pose in its actual form, whereas BA-MPJPE evaluates the normalized accuracy of the pose after removing the body scale.

4.4. Qualitative Results

In Fig. 3 and Fig. 4, it can be seen that our proposed model accurately predicts the 3D human pose from real aerial top-to-bottom view images. In particular, Fig. 3 shows the qualitative results for selected frames extracted from the AirPose real dataset. From this set of results, it can be seen that the estimated 3D human pose is accurately predicted for the input human image, even for images not used

in training. Since our method includes a pose optimizer, we can see that even in this case of severe occlusion, the pose is accurately predicted for the unseen parts. Additionally, qualitative results for aerial top-to-bottom view images obtained from the image-sharing site, Unsplash, are shown in Fig. 4. In particular, the person wearing yellow in the second row of captured images has a highly elevated aerial viewpoint from the ground. These results confirm that our proposed method accurately predicts even very challenging poses, such as the human pose of moving forward on a skateboard. These results confirm that the estimated 3D human pose is accurately represented for the input human image, even for images not used in training.

Fig. 5 shows a qualitative comparison of the different methods trained on the AirPose synthetic dataset. Since aerial top-to-bottom view images are challenging with a large self-occluded region similar to egocentric images, we used Mo^2Cap^2 [41] and xR-egopose [34] as comparison methods. We also used the Wang et al. method [37] as comparison methods, which added a VAE optimizer to each of these methods. Since these methods do not predict SMPL parameters but directly predict 3D human pose, these methods were trained with approaches described in the implementation details along with our method. Furthermore, we also compared the HMR [10] used as a baseline in AirPose. In addition, we compared AirPose [28] with multi-view images, which allows us to make a qualitative comparison of the difference between our method and the method with multi-view images.

Other methods, except for AirPose [28], use single-view images as input, so we can see that they are generally not accurate in predicting the 3D pose in situations with a high degree of self-occlusion. In particular, it can be seen that the Wang et al. method [37], which optimizes the pose via VAE, predicts the pose more accurately than using vanilla Mo^2Cap^2 [41] and xR-egopose [34]. However, since the method of Wang et al. optimizes the previously predicted 3D pose, when the estimated 3D pose used as input is significantly different from the ground truth, the 3D pose after optimization shows a significantly different appearance from the ground truth. Furthermore, since HMR [10] uses the method of fitting an SMPL mesh to a 2D image, it can be seen that the pose prediction of that part is not accurate due to the lack of information in the part where self-occlusion occurs. In contrast, our method estimates the pose more accurately than other methods using the VQ-VAE pose estimator and optimizer. Finally, we can see that the performance of our method is comparable to the results of AirPose [28], which uses multi-view images as input, when compared to other methods. This demonstrates that our method has achieved high generalization performance on datasets trained with VQ-VAE and that it is robust to self-occlusion.

Table 1. Experimental results on AirPose synthetic test dataset. AirPose [28] uses two-view images as input, and the other methods, including ours, use single-view images as input. The proposed method exhibits the highest level of accuracy compared to other single-view based methods, thereby highlighting the significant advantage of the proposed approach.

	PA-MPJPE (↓)	BA-MPJPE (↓)
Mo^2Cap^2 [41]	107.24	80.69
Mo^2Cap^2+Wang [37]	95.43	71.18
xR-Egopose [34]	102.12	76.49
xR-Egopose+Wang [37]	91.24	68.76
HMR [10]	92.35	69.57
Ours	**78.60**	**62.01**
AirPose [28]	75.88	59.61

4.5. Quantitative Results

We used the AirPose synthetic test dataset [28] for quantitative results comparison. We extracted joints from the SMPL parameter provided by the AirPose synthetic test to compare PA-MPJPE and BA-MPJPE for the 15 joints used for training. For quantitative comparisons with state-of-the-art methods, we used the same comparison methods used in qualitative comparison, as described in Fig. 5. The results of the quantitative measurements in terms of PA-MPJPE and BA-MPJPE are summarized in Table 1. AirPose [28] uses multi-view images as input, thus it is listed separately at the bottom of the Table 1. It can be seen that the method [37] of Wang et al. with VAE-based pose optimizer outperforms Mo^2Cap^2 [41] and xR-egopose [34] without pose optimizer for each case. However, despite using a pose optimizer, the method of Wang et al. performs similarly to HMR [10]. This is because HMR predicts the pose based on the SMPL parameters and these parameters also act as strong constraints on the pose. We can verify that our method predicts 3D pose most accurately among single-view methods by the results showing that ourperforms better than Wang et al.'s method and HMR. The results quantitatively show that our method has increased the accuracy of pose estimation over current state-of-the-art methods based on the better generalization performance that is achieved through VQ-VAE. Furthermore, it can be seen that the MPJPE of AirPose [28] and the MPJPE of our method are not significantly different, demonstrating that our method predicts the 3D pose in single-view situations as accurately as multi-view due to the significant generalization performance of VQ-VAE.

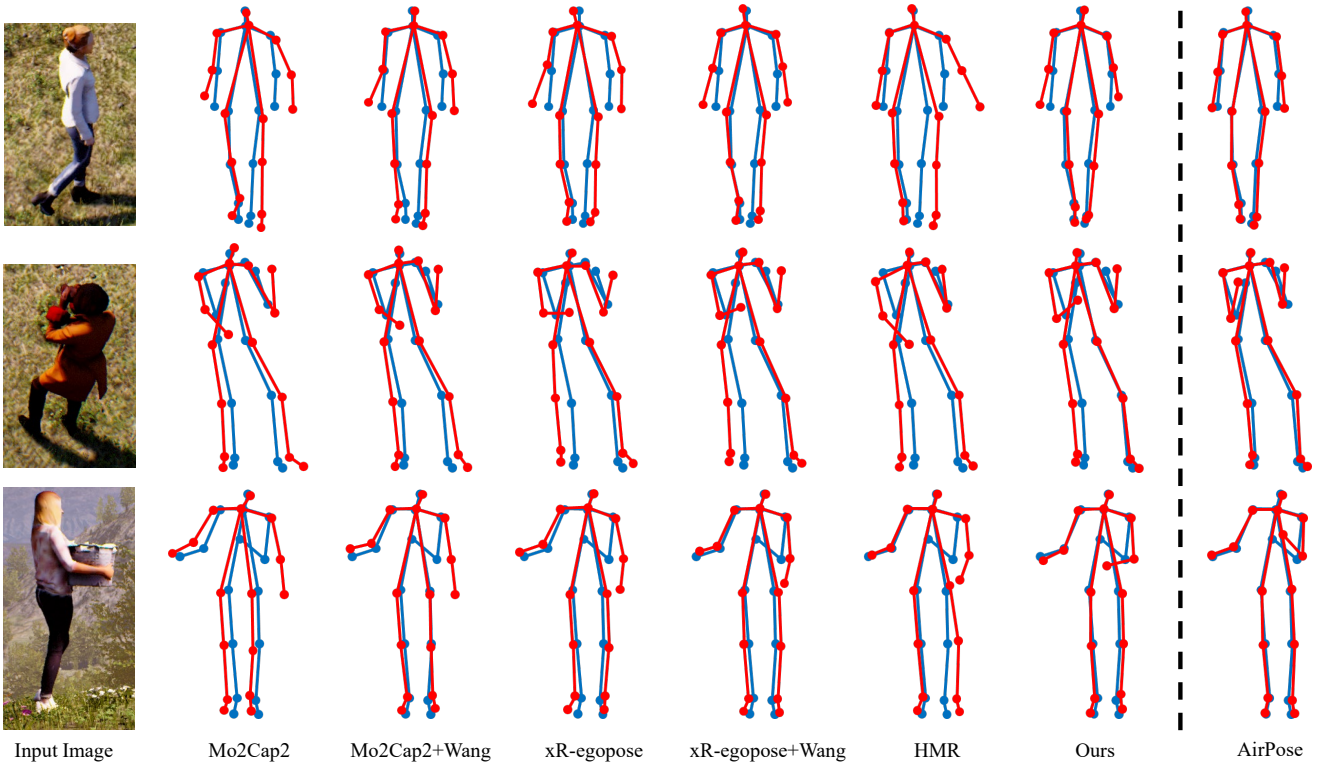

Figure 5. Qualitative comparisons of aerial 3D pose estimation on AirPose synthetic dataset. Predicted 3D poses (red) are overlaid on the ground truth (blue). Here, AirPose [28] uses two-view images as input, and the other methods, including ours, use the corresponding single-view input image illustrated in the figure. The proposed method demonstrates superior accuracy in estimating the 3D pose compared to other single-view pose estimation methods. Additionally, the accuracy of the 3D poses obtained from our methodoloy is on par with that of the multi-view pose estimation technique.

4.6. Ablation Study

In the ablation study, a comparison was made between our proposed method and two other approaches: without VQ-VAE in the pose estimator (referred to as "w/o VQ-VAE in Estimator"), and without pose optimizer (referred to as "w/o Pose Optimizer"). The evaluation was performed on the AirPose test dataset, and the results are presented in Table 2. w/o VQ-VAE in Estimator has a similar structure to vanilla xR-egopose [34] but with a VQ-VAE based pose optimizer added to the pose estimator. In w/o VQ-VAE in Estimator, the pose optimizer is used, but the performance of 3D pose prediction is lower than our complete method because VQ-VAE is removed from the features extracted by the estimator. However, it performs better than the method of Wang et al. [37] in Table 1, which confirm that the VQ-VAE based pose optimizer has better optimization performance than the VAE-based pose optimizer. In w/o Pose Optimizer, the performance of pose estimator alone without pose optimizer is worse than w/o VQ-VAE in Estimator. However, the w/o Pose Optimizer shows better performance than the results of Mo^2Cap^2 [41] and xR-egopose [34] results in Table 1. Moreover, w/o Pose Optimizer shows higher pose estimation performance than the methods of Wang et al. [37] and HMR [10], since the VQ-VAE based pose estimator can predict more accurate 3D pose than other methods even in single-view with the generalization performance of VQ-VAE in predicting 3D pose from an image.

Fig. 6 shows the qualitative results of the ablation study. In the case of w/o Pose Optimizer, it shows the problem of inaccurate pose of occluded body parts in aerial images with significant self-occlusion. The reason for this problem is that the pose optimizer learns the pose prior of the body joints to compensate for the 3D pose, whereas the pose estimator predicts the 3D pose based on the 2D heatmap predicted from the input image, making it difficult to correct for an erroneous 3D pose. On the contrary, in the case of w/o VQ-VAE in Estimator, it can be seen that the VQ-VAE based pose optimizer corrects the 3D pose predicted by the vanilla pose estimator well and produces a comparatively accurate 3D pose.

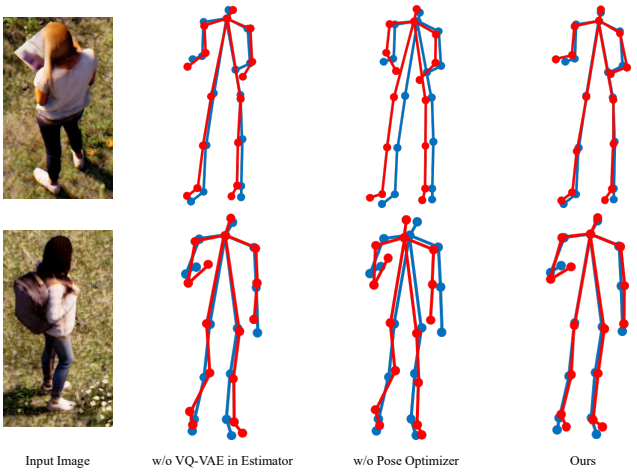

Figure 6. Ablation comparisons of aerial 3D pose estimation on AirPose synthetic dataset. Predicted 3D poses (red) are overlaid on the ground truth (blue). Without the pose optimizer, it can be seen that inaccurate 3D poses are predicted for body joints with strong self-occlusion, illustrating the effectiveness of the pose optimizer.

Table 2. Quantitative results of ablation study. The results indicate that VQ-VAE and pose estimator play a significant role in improving 3D pose estimation accuracy. It can be seen that the performance of w/o VQ-VAE in Estimator and w/o Pose Optimizer is comparable, which indicates that it is difficult to optimize the ground truth pose when an inaccurate 3D pose from the human pose model is input to the pose estimator.

	PA-MPJPE (↓)	BA-MPJPE (↓)
w/o VQ-VAE in Estimator	86.27	66.31
w/o Pose Optimizer	88.49	67.04
All	**78.60**	**62.01**

4.7. General Discussion

Our proposed method uses regression to predict 3D human pose based on 2D joints heatmap obtained from input aerial image in pose estimator. Then, the estimated 3D pose is corrected by a pose optimizer that has learned a 3D human pose prior through a large human 3D pose dataset. Our proposed method has shown significantly more accurate 3D pose estimation results than other state-of-the-art methods [10, 34, 37, 41], and comparable results to AirPose using two-view [28]. This means that the VQ-VAE based pose estimator predicts an precise 3D pose and the optimizer corrects the pose accurately in the part where self-exclusion occurs, resulting in a comparable performance to the method using two-view. The results show that our method significantly improves the generalization performance compared to the comparison method, which is very robust to 3D pose detection for unseen images and challenging images from top-to-down viewpoints, demonstrating the advantages of our method.

5. Conclusion

This study presents a novel approach for accurately estimating the 3D poses of humans using images obtained from aerial perspectives, specifically from top-to-bottom angles. The proposed method has addressed limitations in its capacity to capture dynamic scenarios, mainly attributed to the restricted field of view of a third-person-view camera. When conducting an aerial observation from top-to-bottom viewpoints, it is frequently observed that the lower body appears to be diminished and obstructed by the upper body. This phenomenon leads to pose estimation that is characterized by a high degree of unreliability and inaccuracy. To address the current constraints, we introduce a structure that incorporates VQ-VAE into the pose estimator and pose optimizer for accurate 3D pose prediction. In particular, we propose an innovative pipeline for pose estimation and optimization. Our approach involves using a codebook to learn aerial image features and pose features from extensive human pose datasets with the aid of VQ-VAE. The proposed method with the vector quantizer of VQ-VAEs presents a promising approach to enhance the generalization abilities of 3D pose estimation from aerial top-to-bottom viewpoints. The performance improvement of the VQ-VAE-based pose estimator and pose optimizer is demonstrated through comparison experiments with state-of-the-art methods and an ablation study, which confirms that VQ-VAE has good generalization performance in the pose estimation task. Furthermore, it is shown that the accuracy of 3D pose estimation from single-view images using VQ-VAE can be improved by using multi-view images.

The proposed technique is aimed at local 3D human pose estimation using images from single UAV mounted RGB camera, or images from top-to-bottom viewpoints, but we expect that it can be extended to perform global 3D human pose estimation using non-rigid structure-from-motion or simultaneous localization and mapping in the future.

References

[1] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2272–2281, 2019. 3

[2] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 342–359. Springer, 2022. 3

[3] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3D human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2262–2271, 2019. 3

[4] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3D human pose prior with gradient fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4800–4810, 2023. 3

[5] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6981–6992, 2021. 2

[6] Juergen Gall, Bodo Rosenhahn, Thomas Brox, and Hans-Peter Seidel. Optimization and filtering for human motion capture: A multi-layer framework. *International Journal of Computer Vision*, 87:75–92, 2010. 1, 2

[7] Christoph Gebhardt, Benjamin Hepp, Tobias Nägeli, Stefan Stevšić, and Otmar Hilliges. Airways: Optimization-based planning of quadrotor trajectories according to high-level user goals. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, pages 2508–2519, 2016. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[9] Niels Joubert, Mike Roberts, Anh Truong, Floraine Berthouzoz, and Pat Hanrahan. An interactive tool for designing quadrotor camera shots. *ACM Transactions on Graphics*, 34(6):1–11, 2015. 3

[10] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1, 2, 3, 5, 6, 7, 8

[11] David G Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2):87–99, 1989. 5

[12] Rui Li, Minjian Pang, Cong Zhao, Guyue Zhou, and Lu Fang. Monocular long-term target following on UAVs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 29–37, 2016. 3

[13] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13147–13156, 2022. 3

[14] Hyon Lim and Sudipta N Sinha. Monocular localization of a moving person onboard a quadrotor MAV. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2182–2189. IEEE, 2015. 3

[15] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2720–2735, 2013. 1

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. 3

[17] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5442–5451, 2019. 2, 5

[18] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 72–88. Springer, 2022. 3

[19] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 3

[20] Thomas B Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001. 1

[21] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10133–10142, 2019. 3

[22] Tobias Nägeli, Javier Alonso-Mora, Alexander Domahidi, Daniela Rus, and Otmar Hilliges. Real-time motion planning for aerial videography with dynamic obstacle avoidance and viewpoint optimization. *IEEE Robotics and Automation Letters*, 2(3):1696–1703, 2017. 3

[23] Tobias Nägeli, Samuel Oberholzer, Silvan Plüss, Javier Alonso-Mora, and Otmar Hilliges. Flycon: Real-time environment-independent multi-view human pose estimation with aerial vehicles. *ACM Transactions on Graphics*, 37(6):1–14, 2018. 3

[24] Nobuyasu Nakano, Tetsuro Sakura, Kazuhiro Ueda, Leon Omura, Arata Kimura, Yoichi Iino, Senshi Fukashiro, and Shinsuke Yoshioka. Evaluation of 3D markerless motion capture accuracy using openpose with multiple video cameras. *Frontiers in Sports and Active Living*, 2:50, 2020. 1

[25] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 5

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

[27] Mike Roberts and Pat Hanrahan. Generating dynamically feasible trajectories for quadrotor cameras. *ACM Transactions on Graphics*, 35(4):1–11, 2016. 3

[28] Nitin Saini, Elia Bonetto, Eric Price, Aamir Ahmad, and Michael J Black. Airpose: Multi-view fusion network for aerial 3D human pose and shape estimation. *IEEE Robotics and Automation Letters*, 7(2):4805–4812, 2022. 1, 2, 3, 5, 6, 7, 8

[29] Nitin Saini, Eric Price, Rahul Tallamraju, Raffi Enficiaud, Roman Ludwig, Igor Martinovic, Aamir Ahmad, and Michael J Black. Markerless outdoor human motion capture using multiple autonomous micro aerial vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 823–832, 2019. 3

[30] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3D human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 461–478. Springer, 2022. 3

[31] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision*, pages 529–545, 2018. 3

[32] Rahul Tallamraju, Eric Price, Roman Ludwig, Kamalakar Karlapalem, Heinrich H Bülthoff, Michael J Black, and Aamir Ahmad. Active perception based formation control for multiple aerial vehicles. *IEEE Robotics and Automation Letters*, 4(4):4491–4498, 2019. 3

[33] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, 2020. 4

[34] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando De la Torre. Selfpose: 3D egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 3, 4, 6, 7, 8

[35] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International Workshop on Vision Algorithms Corfu*, pages 298–372. Springer, 2000. 3

[36] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 30, 2017. 4

[37] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3d human pose in global space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11500–11509, 2021. 2, 3, 5, 6, 7, 8

[38] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3D pose estimation from videos. In *Proceedings of the European Conference on Computer Vision*, pages 764–780. Springer, 2020. 3

[39] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct multi-view multi-person 3d human pose estimation. *Advances in Neural Information Processing Systems*, 2021. 2, 3

[40] Lan Xu, Yebin Liu, Wei Cheng, Kaiwen Guo, Guyue Zhou, Qionghai Dai, and Lu Fang. Flycap: Markerless motion capture using multiple autonomous flying cameras. *IEEE Transactions on Visualization and Computer Graphics*, 24(8):2284–2297, 2017. 3

[41] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo^2Cap^2: Real-time mobile 3D motion capture with a cap-mounted fisheye camera. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2093–2101, 2019. 2, 3, 6, 7, 8

[42] Angela Yao, Juergen Gall, and Luc Van Gool. Coupled action recognition and pose estimation from multiple views. *International journal of computer vision*, 100:16–37, 2012. 2

[43] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14730–14740, 2023. 4

[44] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023. 3

[45] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3D human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2344–2353, 2019. 3

[46] Xiaowei Zhou, Sikang Liu, Georgios Pavlakos, Vijay Kumar, and Kostas Daniilidis. Human motion capture using a drone. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2027–2033. IEEE, 2018. 3