2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Active Human Pose Estimation via an Autonomous UAV Agent

Jingxi Chen, Botao He, Chahat Deep Singh, Cornelia Fermüller, Yiannis Aloimonos
Perception and Robotics Group, University of Maryland - College Park

2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3503-7770-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IROS58592.2024.10801780

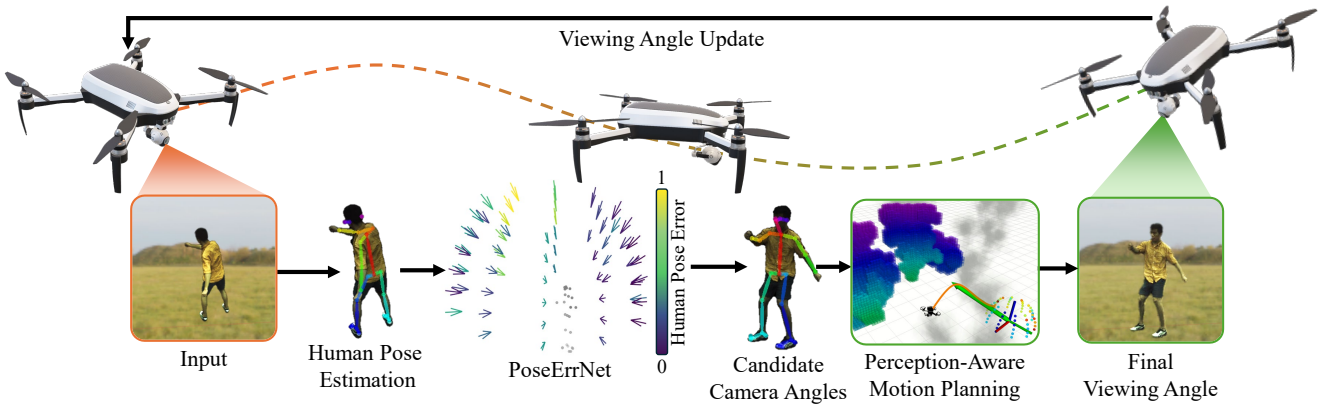


Fig. 1: The proposed pipeline for an autonomous drone to perform active human pose estimation follows a specific loop. Initially, the drone captures an input image, which is processed to perform 2D human pose estimation, yielding an imperfect 2D skeleton. Utilizing this skeleton, our designed PoseErrNet predicts the pose estimation error across a hemispherical space of camera viewing angles and suggests candidate camera viewing angles for the best next view. These candidate views are subsequently integrated into a planner focusing on perception-aware planning and navigation goal planning.

***Abstract*— One of the core activities of an active observer involves moving to secure a "better" view of the scene, where the definition of "better" is task-dependent. This paper focuses on the task of human pose estimation from videos capturing a person's activity. Self-occlusions within the scene can complicate or even prevent accurate human pose estimation. To address this, relocating the camera to a new vantage point is necessary to clarify the view, thereby improving 2D human pose estimation. This paper formalizes the process of achieving an improved viewpoint. Our proposed solution to this challenge comprises three main components: a NeRF-based Drone-View Data Generation Framework, an On-Drone Network for Camera View Error Estimation, and a Combined Planner for devising a feasible motion plan to reposition the camera based on the predicted errors for camera views. The Data Generation Framework utilizes NeRF-based methods to generate a comprehensive dataset of human poses and activities, enhancing the drone's adaptability in various scenarios. The Camera View Error Estimation Network is designed to evaluate the current human pose and identify the most promising next viewing angles for the drone, ensuring a reliable and precise pose estimation from those angles. Finally, the combined planner incorporates these angles while considering the drone's physical and environmental limitations, employing efficient algorithms to navigate safe and effective flight paths. This system represents a significant advancement in active 2D human pose estimation for an autonomous UAV agent, offering substantial potential for applications in aerial cinematography by improving the performance of autonomous human pose estimation and maintaining the operational safety and efficiency of UAVs.**

I. Introduction

Recent advances in aerial robotic technologies have significantly improved the use and abilities of aerial robots in many industries. [1]–[3]. A key aspect of modern aerial robots is their ability to be equipped with video cameras, transforming them into dynamic platforms for aerial videography [4]. The mobility and agility of aerial robots make them highly effective for aerial cinematography, allowing for versatile footage capture from optimal angles with minimal equipment. The growing demand in entertainment, industrial, and military sectors has shifted aerial cinematography's focus from static objects to dynamic human subjects, and the technical challenge is how to autonomously adjust a drone's viewing direction to best view human subjects during navigation.

The challenges associated with autonomously adjusting the viewing direction for UAV-based human pose estimation include determining the criteria for modifying the UAV's viewing direction during human inspection, and navigating the UAV in a way that balances perceptual guidance with navigation objectives, such as feasible motion plans and collision avoidance.

To address the challenges posed by dynamic videography using UAVs, we propose a sophisticated, integrated autonomous UAV videography system, as illustrated in Fig. 2. This system is engineered to intelligently interpret human poses and proactively reposition itself to capture optimal visual content. The architecture of this system can be divided into three primary components: 1) Drone-View Human Subject Data Generation Framework. This framework is designed to capture a wide range of human poses and actions under varying environmental perspectives. By utilizing advanced vision techniques (HumanNerf) [5], this framework

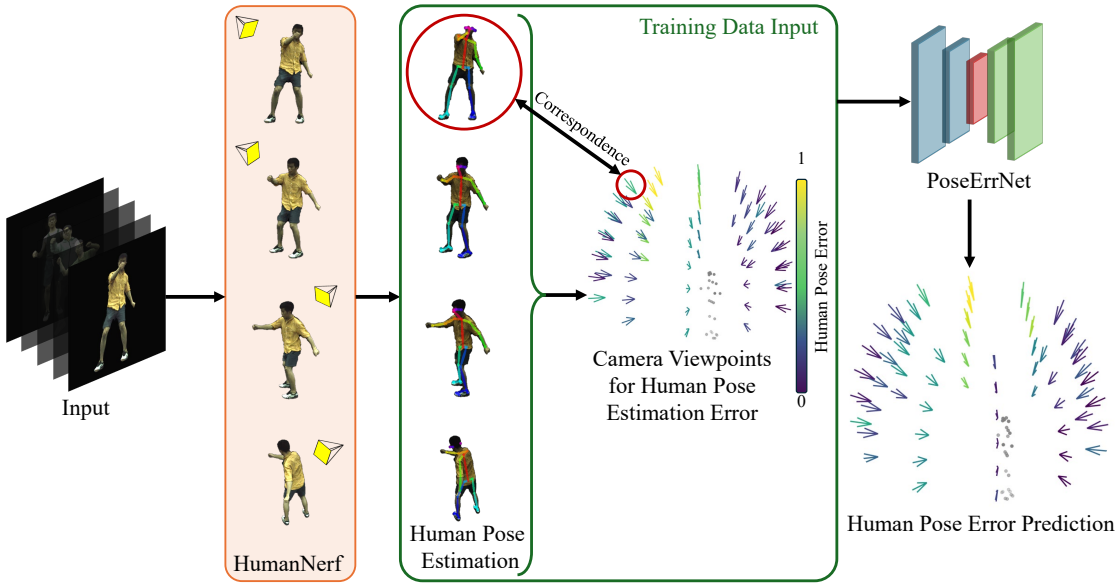


Fig. 2: Our proposed approach features an integrated system with three key components: 1) Drone-View Data Synthesis, which generates realistic drone perspectives of human subjects from various camera angles and human poses, alongside calculating the associated human pose estimation error for these views to serve as training data pairs. 2) PoseErrNet, a network trained on the generated drone-view data pairs, is capable of predicting a 3D perception guidance field for the selection of candidate viewing angles. 3) A comprehensive planner that integrates traditional navigation cost maps with the 3D perception guidance field derived from PoseErrNet. This integration enables effective motion planning, collision avoidance, and the execution of the next-best viewing angle selection for accurate human pose estimation.

will enable the UAV to have a profound understanding of human subjects in different scenarios, enhancing its ability to adapt to real-world videography tasks. 2) Robust and Efficient On-Drone Network for Viewing Angle Estimation. This network is tailored to analyze the current human pose and compute the best subset of the next viewing angles. It aims to process complex visual inputs in real-time, ensuring the UAV can react promptly and accurately to dynamic subjects. The efficiency of this network is crucial, as it directly impacts the UAV's ability to operate under computational and power constraints typically associated with autonomous drones. 3) Combined Planner for Feasible Motion Plan. Our proposed system is a sophisticated planning module that combines the network's viewing angle recommendations with the UAV's dynamic and environmental constraints. The planner employs advanced algorithms to chart a feasible motion plan that not only adheres to the suggested viewing angles but also respects the physical limitations of the UAV and the navigational challenges posed by the environment. Using this planner, the UAV can maneuver in complex environments with agility and precision, ensuring high-quality videography while maintaining safety and operational efficiency.

In Sec. III, we will delve into the specifics of generating human subject data from drone views. Following that, in Sec. IV, we explain our approach using PoseErrNet to transform an imperfect detection of 2D human keypoints, into an error vector for 2D human pose estimation (HPE) across all predefined hemispherical camera viewing angles. This process creates what we refer to as the 3D perception guidance field. Our goal was to design a lightweight network capable of learning the correlation between the optimal subset of next viewing angles and the current human pose estimation, utilizing a dataset we generated for this purpose. The robust estimation of the 3D perception guidance field is crucial as it provides candidate camera viewing angles for the ensuing motion planning phase. For the motion planning part, based on [6] and [7], we crafted a perception-aware motion planning framework. This framework not only incorporates the 3D perception guidance field but also is capable of generating a smooth flight trajectory, avoiding occlusions between the target and the UAV, and ensuring the safety of the flight.

Our contributions can be summarized as:

- A drone-view data generation framework for different human poses.
- A robust and efficient network running on the drone for estimating the best subset of the next viewing angles based on the current human pose estimation.
- A combined planner that combines the perspective-aware guidance from the network and traditional navigation constraints into a feasible motion plan for improving 2D HPE, a computer vision task.

These three interconnected components are seamlessly integrated into a system designed for application in real-world scenarios.

II. Related Works

1) Autonomous Aerial Human Inspection: Existing research on autonomous aerial inspection of human subjects primarily aims at achieving planning autonomy but often lacks objective guidance on subsequent movements. As a result, high-level guidance for the inspection tasks is typically expected to come from human operators, as seen

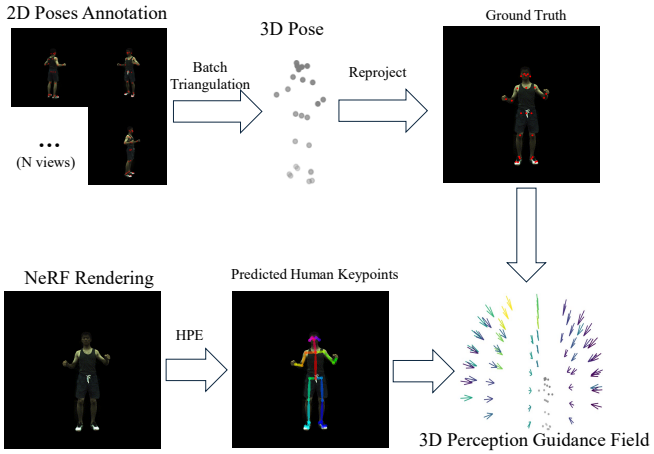


Fig. 3: The process for generating NeRF-based drone-view images of human subjects and 3D perception guidance field data involves using 2D annotations to conduct batch triangulation, resulting in a 3D skeleton for a given human pose. We then render the synthesized image for "drone views", reproject the ground truth 3D skeleton onto NeRF poses to obtain ground truth 2D keypoints, and employ an arbitrary HPE network to predict these keypoints for computing the per camera view HPE error. Through this method, we successfully acquire paired data comprising 2D observations and the corresponding 3D perception guidance field.

in various studies [8]–[14]. While the autonomous planning for UAVs to follow and robustly track mobile objects along optimized trajectories is well-documented [15]–[18], these works generally do not address the capability of UAVs to autonomously execute perception-aware objectives, such as human pose estimation, without human operator inputs.

2) 2D Human Pose Estimation: Human Pose Estimation, both in 2D and 3D, has received attention for two decades because of its many applications, including activity understanding, healthcare, AR/VR, and robotics [19]–[21]. Advancements in deep learning techniques have significantly enhanced the performance of 2D Human Pose Estimation (HPE), leading to robust and efficient solutions for both single and multiple individuals. [22]–[27]. However, the near-perfect performance of 2D Human Pose Estimation (HPE) often relies on ideal input images of humans without any occlusion of body parts. This assumption becomes challenging in the context of autonomous UAV inspections of humans. As the UAV moves, the camera's view of human subjects can easily be obstructed by environmental elements or self-occlusion of human body parts.

3) Neural Radiance Field: NeRF [28] and its extensions [29]–[35] enable high-quality and continuous rendering of static 3D scenes. A natural progression is to expand the neural radiance field approach to encompass dynamic scene representation. [36]–[41]. In the context of dynamic scene representations, our work is most closely related to the neural representation of dynamic human subjects [5], [42], [43].

III. Drone-View Data Acquisition

As illustrated in Fig. 2, the drone views of a human subject can be represented in a hemispherical space. This type of data can be acquired through one of three methods: 1) Drone Capture, which involves using UAVs to obtain images of human subjects in specific poses from multiple angles. 2) Camera Array, which entails setting up an array of cameras to cover the hemispherical space, with a focus on achieving time synchronization among the cameras. 3) Utilization of simulation software like Blender [44] or Unity [45] to create projected image views from human models. Each method comes with its own set of challenges and practical considerations. The pros and cons of these methods for generating drone-view data are detailed in Table I, highlighting economic and engineering costs. For instance, deploying multiple drones for data capture or creating a camera array incurs significant economic costs due to the hardware required and demands considerable engineering effort for calibration and synchronization. Conversely, simulation software offers a low economic cost option, though achieving a high-quality simulation presents substantial engineering challenges. The table also compares the accuracy of the desired viewing angles, the resolution, or the detail level at which capture angles are set, and the realism of capturing human subjects and poses. Drone capture and camera arrays provide realism in both appearance and pose since they employ real-world methods. In contrast, achieving realistic captures of both appearance and human poses proves difficult with simulation software.

Our research leverages the innovative free-viewpoint rendering method, HumanNeRF [5], which is designed for rendering images with complex human poses, perfectly meeting our data acquisition needs. This technique allows for the free-view synthesis of a human image in a specific pose. We configure the camera poses and viewing angles to match the desired hemispherical drone camera pose for the captured image and then render drone-view images for various human poses in ZJU-MoCap Dataset [46], [47].

After rendering the drone-view images, we follow the approach depicted in Fig. 3 to compute the human pose estimation error for each camera view. This process enables us to gather the desired training data pairs, linking each 2D human skeleton estimation to a 3D perception guidance field.

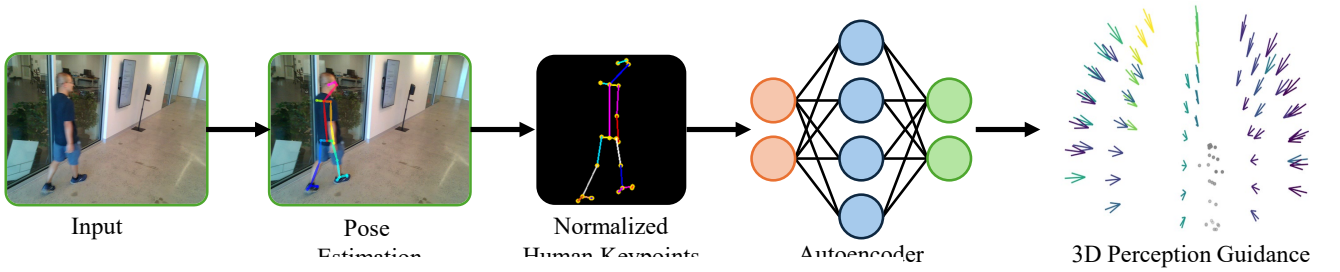
IV. Mapping 2D Observations to 3D Perception Guidance Fields

After acquiring our training data pairs for drone-view human images and 3D perception guidance field. A simplified auto-encoder network is employed for visual guidance, as depicted in Fig. 4. This network architecture is characterized by a minimal number of weights, enhancing its efficiency. For dealing with the sim-to-real gap, we proposed a process to normalize the input drone-view data, this normalization process includes first converting the input image into the HPE resulting keypoints and then normalizing the detected keypoints to account for translation, rotation, and scale variance in the input keypoint due to different human, drone-to-human distance or instability during the flight of the drone.

The proposed normalization process for input HPE keypoints is straightforward yet effective. We begin by identifying the human spine, typically the line between

Method	Cost		Viewing Angle		Realistic Capture	
	Ecnomoic	Engineering	Accuracy	Resolution	Appearance	Human Pose
Drone Capture	Medium	High	Low	Low	✓	✓
Camera Array	High	High	High	High	✓	✓
Simulation Software	Low	High	High	High	–	–
Ours	Low	Low	High	High	✓	✓

TABLE I: Comparison of different drone-view data acquisition methods. Cost, related to economic and engineering cost of the capture method (*Density*). Viewing Angle, consists of viewing angle accuracy and resolution of the capture method (*Viewing Angle*). Realistic Capture, is related to whether the capture method can capture realistic appearance and human pose (*Realistic Caprure).*



deal with the sim-to-real gap and with scale,
ork to map from normalized 2D observations

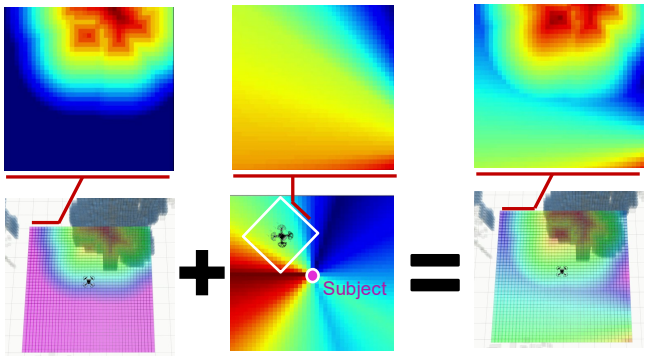


Fig. 5: Illustration for the calculating P-ESDF.

the neck and the midpoint of the hips. Once the spine is determined, we translate all keypoints to align the spine's midpoint to a consistent coordinate, addressing the translation variance of the input keypoints. Furthermore, we rotate all keypoints to orient the spine's direction upwards and to the right, countering the rotation variance of the input keypoints. Finally, we scale all keypoints to ensure the spine length remains constant. An example of normalized human keypoints is illustrated in Fig. 4.

Our normalization process for the HPE keypoints enhances robustness against variances in translation, rotation, and scale in the inputs. By using the normalized coordinates of the keypoints, represented as a vector, as input to our network, we simplify the design of the network architecture. The proposed PoseErrNet is an autoencoder with a minimal number of layers, benefiting greatly from the simplicity and normalization of its inputs.

V. Perception-Aware Motion Planning

We introduce our proposed perception-aware motion planning framework to bridge the gap between perception and motion planning. In the proposed framework, the perception loss is added to the motion planning cost function as one of the costs. To achieve that, we built a differentiable distance field called Pose-enhanced Euclidean Distance Field (P-ESDF), noted as $\mathcal{P}$, each of its element is then represented as $\{p_i \in \mathcal{P} | i \in \mathbb{Z}^+\}$. The construction of the field is described below.

The output of the PoseErrNet is a 2-D map $\mathcal{E} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$, where m and n are dimensions of the proposed Perception Guidance Field. The first step is to transform $\mathcal{E}$ to the subject frame in $\mathbb{R}^3$ with the subject in the center, as shown in the bottom sub-figure of Fig. 5(b). Then, we project the transformed map $\mathcal{E}^{sub}$ from the subject frame to the drone frame to get the $\mathcal{E}^{drone}$, as shown in the upper sub-figure of Fig. 5(b). To simutanrously perform obstacle avoidance and viewpoint targeting, the $\mathcal{E}^{drone}$ need to be merged with the standard ESDF $\mathbf{E}$ to get the final $\mathcal{P}$, as illustrated in Fig. 5. The merge process can be expressed as:

$$\mathcal{P} = \sum_i \lambda \mathcal{E}_i^{drone} + (1 - \lambda)\mathbf{E}_i. \tag{1}$$

Then, to guide the drone with the proposed P-ESDF, we design pose penalty J_{pose} as the function of p. Assume the path is constructed by a series of waypoints $\{\mathbf{p_k} \subset \mathcal{R}^3 | k \in \mathbb{Z}^+\}$, and define $\Xi(\mathbf{p_k})$ as the value of $\mathcal{P}$ at the position of $\mathbf{p_k}$. The J_{pose} can be expressed as

$$J_{pose} = \lambda_p \sum_{i=0}^{M} c(p_i)\mathbf{p_k}' \tag{2}$$

where $\mathbf{p}_k' = \frac{\partial \Xi(\mathbf{p}_k)}{\partial \mathbf{p}_k}$ can be efficiently acquired from P-ESDF, and $c(\mathbf{p}_k)$ can be expressed as:

$$c(\mathbf{p}_k) = \begin{cases} \frac{1}{2\rho}(\Xi(\mathbf{p}_k) - \rho)^2, \Xi(p_k) \leq \rho \\ 0 \qquad\qquad\quad , \Xi(\mathbf{p}_k) > \rho \end{cases} \tag{3}$$

The result is shown in Fig. 5.

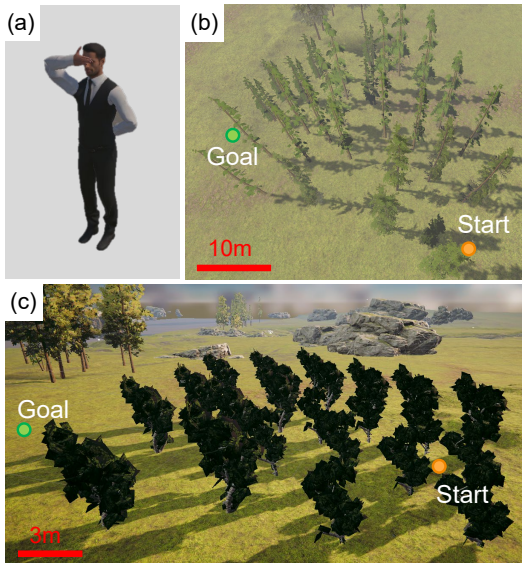


Fig. 6: The testing environment for the proposed motion planning framework.

VI. Experiments

To demonstrate the effectiveness of our system from different dimensions, we conduct 3 tasks in simulated environments with varied scales and complexity, as shown in Fig. 6. The first task is to estimate a static challenging pose, as shown in Fig. 6 (a). The later two tasks are to estimate the human pose online during walking in the forests, as shown in Fig. 6 (b-c). The drone needs to simutaneously track the person, choose the best view-point, avoid occlusion and ensure the safety.

A. Implementation Details

The implementation of PoseErrNet is an autoencoder with 4 layers in the encoder layers and 4 layers in the decoder layers, with total 0.011M parameters. It runs on NVIDIA RTX 3070Ti GPU with 5000 HZ inference rate. The TRT-Pose is utilized for 2-D human pose estimation. It runs on NVIDIA RTX 3070Ti GPU with 15 HZ update rate. The simulated vehicle is equipped with an Intel D435 depth camera, which is used both as the range sensor for navigation planning and camera sensor for RGB images. The onboard autonomy system of the UAV integrates several key navigation modules from the development environment of [48]. These include kinodynamic path search, mapping module and GUI. These components serve as fundamental navigation modules. The proposed perception-aware planner is on the top of the navigation system. The framework runs on a laptop with i7-12700H CPU. We configure the navigation system to update at 15Hz and perform trajectory optimization at each sensor update. The spatial resolution is set as $0.2m$. The P-ESDF is set a $10 \times 10m$ area with the vehicle in the center.

B. Evaluation of 3D Perception Guidance Field Generation

To demonstrate the performance of the proposed 3D perception guidance field generation, we showcase an example in challenging scenarios as shown in Fig. 7.

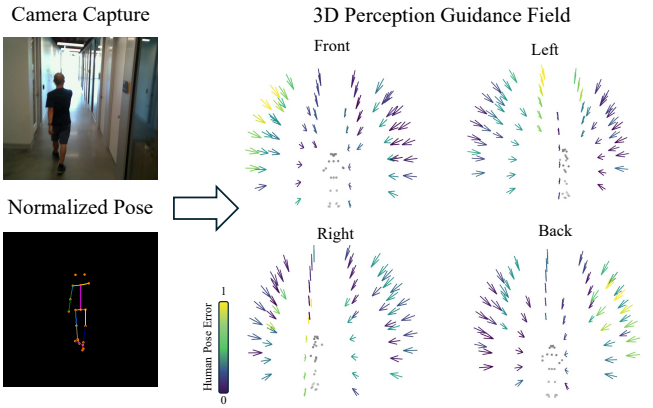


Fig. 7: The result for 3D perception guidance field prediction on the example real-world data. Each arrow within the 3D perception guidance field represents a camera viewing angle. Colors are utilized to denote the error of HPE at a camera viewing angle.

In Fig. 7, the algorithm begins by normalizing the detected imperfect 2D human keypoints, as outlined in the methodology section (Sec. IV). The robustness of our PoseErrNet's output against variations in scale, translation, and rotation of the input keypoint detection is enhanced by this normalization process. We captured a real-world video with a hand-held camera following a human subject and selected 146 representative frames. To these frames, we applied three levels of perturbation: T1 involves uniform random translation ranging from 0 to 5 pixels, uniform random rotation from 0 to 5 degrees, and uniform random scaling from 1.0 to 1.05; T2 includes uniform random translation from 0 to 10 pixels, uniform random rotation from 0 to 10 degrees, and uniform random scaling from 1.0 to 1.10; T3 consists of uniform random translation from 0 to 20 pixels, uniform random rotation from 0 to 20 degrees, and uniform random scaling from 1.0 to 1.15. As illustrated in Table II, we quantized the camera viewing angle error into 21 bins within its range and reported the percentage of bin changes from the results without perturbation for all camera viewing angle error predictions from PoseErrNet. Due to the input normalization, our PoseErrNet's output demonstrates robustness under various levels of input keypoint perturbations.

This normalized keypoints detection data is then input into PoseErrNet to predict the 3D perception guidance field. As an example in Fig. 7 from our real-world collected video frames, where a person moves forward while raising their right arm, obstructing their face. Viewpoints directly in front of the individual are assigned lower costs due to their superior visibility. In contrast, viewpoints from behind are usually associated with higher costs, as they are more prone to occluding important features like the face and arms. The error increases on the right-hand side due to occlusion caused by the raised arm. The left to the front side also with high error because the person's left leg moving forward creates self-occlusion of the right leg. Meanwhile, the area from the left to the back side exhibits the lowest error, indicating the best candidate camera viewing angles for this scenario.

Perturbation	Translation (%)	Rotation (%)	Scale (%)	All (%)
T1	5	8	12	17
T2	11	12	16	25
T3	19	18	21	34

TABLE II: Evaluation of Robustness of Pose Normalization for 3 perturbation levels T1, T2, and T3. The translation-only results (*Translation*), the rotation-only results (*Roataion*), the scale-only results (*Scale*) and combing all translation, rotation, scale results (*All*). Here we show after input keypoint normalization how many percentages of PoseErrNet output change due to input keypoint perturbation.

C. Perception-aware Motion Planning Experiment

To evaluate the efficacy of the On-drone Perception-aware Motion Planning method, we conducted an experimental implementation in complex, cluttered environments characterized by numerous obstacles, as depicted in Fig. 6.

As shown in Fig. 9 and Fig. 8, in obstacle-free environments, the UAV consistently adheres to the optimal viewpoint. However, upon encountering obstacles, the UAV demonstrates the capability to maintain this optimal viewpoint while simultaneously navigating around the obstructions. In scenarios where obstacles are proximate or densely situated, the planning framework proactively shifts to the second-best viewpoint. This adjustment is crucial for mitigating occlusion issues between the target and the UAV, and for maintaining safety by avoiding obstacles. Upon successful navigation past obstacles, and making sure that occlusions do not obstruct the view of the target, the UAV seamlessly plans and executes a trajectory to return to the primary, most advantageous viewpoint.

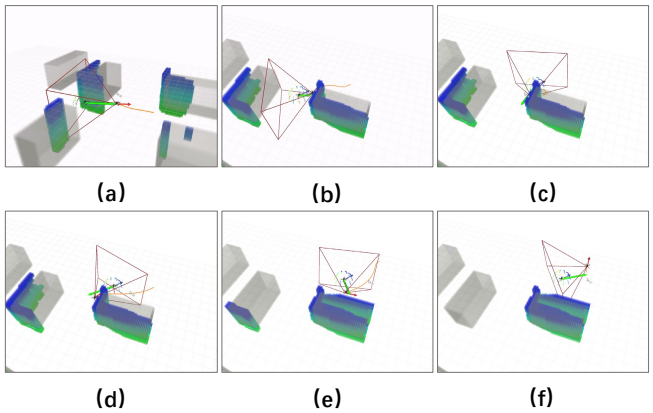


Fig. 8: Perception-aware motion planning experiment. **(a):** The UAV follows the best viewpoint (green arrow). **(b):** The UAV can still follow the best viewpoint and while avoiding the obstacle. **(c-d):** To avoid the occlusion between the target and the UAV, and also to avoid the obstacles for safety, the proposed planning framework automatically switches to the second best viewpoint. **(e-f):** After avoiding the obstacle, and if there is no occlusion between the target and the UAV, the UAV plans a smooth trajectory to return to the best viewpoint.

D. System-level Comparison

This experiment was designed to validate the robustness and overall performance of our integrated system. The perception aspects were simulated within Unity, while the control and motion planning components were simulated in ROS [49] using Rviz [50] for visual representation. We benchmark our method with Auto-Filmer [48] with different pre-setted view-angles. Auto-Filmer is a planner for autonomous tracking and videogrphy in unknown circumstance, it can track human with given viewpoint while keeping the drone safe. In this experiment, we set front, side and back views for Auto-Filmer to track. Two standard metrics are introduced to evaluate the performance: Percentage of Correct Key-points (PCK) and Mean Squared Error.

As indicated in Fig. 9 and Table. III, the optimal viewpoint (represented by a green arrow) dynamically adjusts in response to changes in the human subject's pose. The UAV is programmed to track and align with this best viewpoint in real-time, showcasing its responsiveness to the target's movements. In scenarios where both the target and obstacles are present within the UAV's operational environment, the system intelligently opts for the second-best viewpoint. This strategic choice is critical for avoiding visual occlusion between the UAV and the target, and for ensuring safety by steering clear of obstacles. This approach effectively demonstrates the system's capability to adapt to varying environmental conditions while maintaining high-quality perception and safe navigation.

VII. Future Work

In our work, the computation of the 3D perception guidance field is based solely on the current frame, without accounting for changes in human poses over time due to motion. This can result in latency in perception guidance and viewing-angle adjustment planning. In future work, we plan to address this limitation by incorporating a sequence-based network that considers past human poses and predicts future changes in the 3D perception guidance field, accommodating motion-induced pose changes.

VIII. Conclusion

The innovative approach detailed in this paper signifies a considerable leap forward in the domain of active 2D human pose estimation through the use of autonomous Unmanned Aerial Vehicles (UAVs). By weaving together a NeRF-based Drone-View Data Generation Framework, an On-Drone Network for Camera View Error Estimation, and a Combined Planner for strategic camera repositioning, our methodology effectively tackles the issue of self-occlusions in videos capturing human activities. This integrated system not only enhances the accuracy and reliability of human pose estimation from optimized camera viewing angles but also guarantees the adaptability and operational safety of drones across varied environments. The proposed method highlights the critical role of dynamic viewpoint

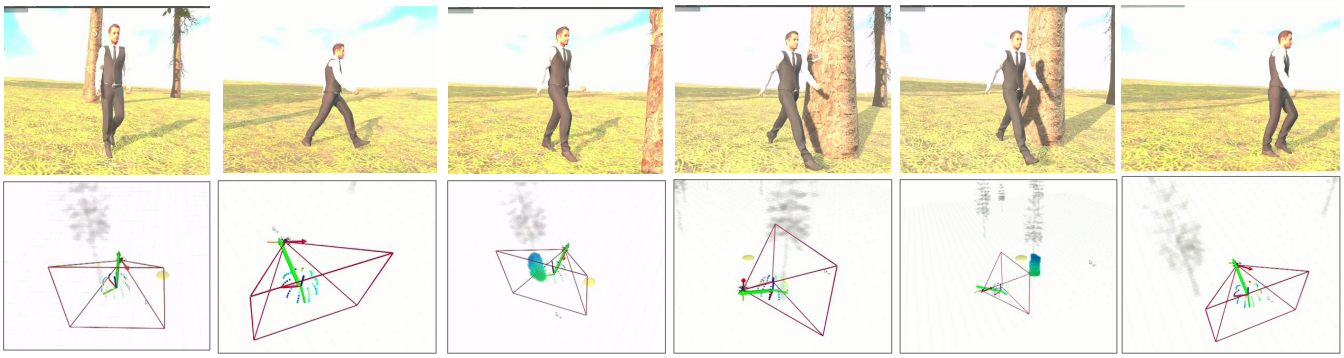


Fig. 9: Demonstration of the perception-aware motion planning. The upper row in each figure is an image captured using the onboard camera, and the bottom row illustrates the third-person view of the experiment. The 3D perception guidance field is represented as a hemispherical field around the drone with different colors indicating the quality of the view point (from green to red indicating good to poor). The green arrow means the best viewpoint evaluated altogether from the pose estimation accuracy, occlusion, and flight safety. **(a - b):** The best viewpoint (green arrow) changes with the human pose, and the UAV follows the best view point in real-time. **(c - f):** When both target and obstacles are observed in the environment, the UAV automatically chooses the second best viewpoint to avoid occlusions and ensure safety.

Method	Challenging Pose		Large-Scale		Dense		All	
	PCK	MSE	PCK	MSE	PCK	MSE	PCK	MSE
Auto-Filmer - Front	0.87	31.46	**0.91**	**24.10**	0.60	42.41	0.79	32.66
Auto-Filmer - Side	0.93	30.73	0.78	25.59	0.80	37.89	0.84	31.40
Auto-Filmer - Back	0.80	29.20	0.67	34.31	0.67	37.51	0.71	33.67
Ours	**1.0**	**22.60**	0.86	24.47	**0.92**	**23.61**	**0.92**	**23.56**

TABLE III: PCK and MSE Evaluation for System-level Experiments

optimization in elevating the quality of pose estimation, thereby paving new pathways for applications in sectors like aerial cinematography and surveillance. Experimental results from both simulation and real-world-captured data prove the efficacy of each component within our system and, more importantly, demonstrate the enhanced task-level performance of the integrated system.

IX. Acknowledgements

The support of NSF under awards OISE 2020624 and BCS 2318255 is greatly acknowledged.

References

[1] Claudia Stöcker, Rohan Bennett, Francesco Nex, Markus Gerke, and Jaap Zevenbergen. Review of the current state of uav regulations. *Remote sensing*, 9(5):459, 2017.

[2] Francesco Nex and Fabio Remondino. Uav for 3d mapping applications: a review. *Applied geomatics*, 6:1–15, 2014.

[3] Valerio Baiocchi, D Dominici, Martina Mormile, et al. Uav application in post-seismic environment. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40:21–26, 2013.

[4] Xin Zhou, Xiangyong Wen, Zhepei Wang, Yuman Gao, Haojia Li, Qianhao Wang, Tiankai Yang, Haojian Lu, Yanjun Cao, Chao Xu, et al. Swarm of micro flying robots in the wild. *Science Robotics*, 7(66):eabm5954, 2022.

[5] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022.

[6] Xiaoxia Zhou, Zhepei Wang, Chao Xu, and Fei Gao. Ego-planner: An esdf-free gradient-based local planner for quadrotors. *IEEE Robotics and Automation Letters*, 6:478–485, 2020.

[7] Qianhao Wang, Botao He, Zhiren Xun, Chao Xu, and Fei Gao. Gpa-teleoperation: Gaze enhanced perception-aware safe assistive aerial teleoperation. *IEEE Robotics and Automation Letters*, 7(2):5631–5638, 2022.

[8] Zhiwei Zhang, Yuhang Zhong, Junlong Guo, Qianhao Wang, Chao Xu, and Fei Gao. Auto filmer: Autonomous aerial videography under human interaction. *IEEE Robotics and Automation Letters*, 8(2):784–791, 2022.

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

[10] Niels Joubert, Dan B Goldman, Floraine Berthouzoz, Mike Roberts, James A Landay, Pat Hanrahan, et al. Towards a drone cinematographer: Guiding quadrotor cameras using visual composition principles. *arXiv preprint arXiv:1610.01691*, 2016.

[11] 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 the 2016 chi conference on human factors in computing systems*, pages 2508–2519, 2016.

[12] Christoph Gebhardt, Stefan Stevšić, and Otmar Hilliges. Optimizing for aesthetically pleasing quadrotor camera motion. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018.

[13] Ziquan Lan, Mohit Shridhar, David Hsu, and Shengdong Zhao. Xpose: Reinventing user interaction with flying cameras. In *Robotics: Science and Systems*, pages 1–9, 2017.

[14] Hao Kang, Haoxiang Li, Jianming Zhang, Xin Lu, and Bedrich Benes. Flycam: Multitouch gesture controlled drone gimbal photography. *IEEE Robotics and Automation Letters*, 3(4):3717–3724, 2018.

[15] Boseong Felipe Jeon and H Jin Kim. Online trajectory generation of a mav for chasing a moving target in 3d dense environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1115–1121. IEEE, 2019.

[16] Jialin Ji, Neng Pan, Chao Xu, and Fei Gao. Elastic tracker: A spatio-temporal trajectory planner for flexible aerial tracking. In *2022 International Conference on Robotics and Automation (ICRA)*, page 47–53. IEEE Press, 2022.

[17] Boseong Jeon, Yunwoo Lee, and H Jin Kim. Integrated motion planner for real-time aerial videography with a drone in a dense environment. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1243–1249. IEEE, 2020.

[18] Qianhao Wang, Yuman Gao, Jialin Ji, Chao Xu, and Fei Gao. Visibility-aware trajectory optimization with application to aerial tracking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5249–5256. IEEE, 2021.

[19] Gutemberg Guerra-Filho, Cornelia Fermüller, and Yiannis Aloimonos. Discovering a language for human activity. In *Proceedings of the AAAI 2005 Fall Symposium on anticipatory cognitive embodied systems*, 2005.

[20] Yi Li, Cornelia Fermüller, Yiannis Aloimonos, and Hui Ji. Learning shift-invariant sparse representation of actions. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2630–2637. IEEE, 2010.

[21] Snehesh Shrestha, Cornelia Fermüller, Tianyu Huang, Pyone Thant Win, Adam Zukerman, Chethan M Parameshwara, and Yiannis Aloimonos. Aimusicguru: Music assisted human pose correction. *arXiv preprint arXiv:2203.12829*, 2022.

[22] Xuecheng Nie, Jiashi Feng, Yiming Zuo, and Shuicheng Yan. Human pose estimation with parsing induced learner. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2100–2108, 2018.

[23] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1831–1840, 2017.

[24] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016.

[25] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2226–2234, 2018.

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

[27] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[29] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.

[30] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021.

[31] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

[32] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021.

[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. *Advances in neural information processing systems*, 33:7537–7547, 2020.

[34] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.

[35] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (ToG)*, 40(6):1–18, 2021.

[36] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021.

[37] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.

[38] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[39] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[40] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021.

[41] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021.

[42] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021.

[43] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021.

[44] Blender. Blender. https://www.blender.org/, 2024. Accessed: 2024-03-05.

[45] Unity. Unity. https://unity.com/, 2024. Accessed: 2024-03-05.

[46] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.

[47] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *CVPR*, 2021.

[48] Zhiwei Zhang, Yuhang Zhong, Junlong Guo, Qianhao Wang, Chao Xu, and Fei Gao. Auto filmer: Autonomous aerial videography under human interaction. *IEEE Robotics and Automation Letters*, 8(2):784–791, 2023.

[49] Morgan Quigley. Ros: an open-source robot operating system. In *IEEE International Conference on Robotics and Automation*, 2009.

[50] Hyeong Ryeol Kam, Sung Ho Lee, Taejung Park, and Chang Hun Kim. Rviz: a toolkit for real domain data visualization. *Telecommunication Systems*, 60(2):337–345, October 2015.