View Planning for High Fidelity 3D Reconstruction of a Moving Actor

Qingyuan Jiang[1], and Volkan Isler[1]

***Abstract*— Capturing and reconstructing a human actor's motion is important for filmmaking and gaming. Currently, motion capture systems with static cameras are used for pixel-level, high-fidelity reconstructions. Such setups are costly, require installation and calibration and, more importantly, confine the user to a predetermined area. In this work, we present a drone-based motion capture system that can alleviate these limitations. We present a complete system implementation and study view planning, which is critical for achieving high-quality reconstructions. The main challenge is that the reconstruction algorithms are computationally costly, but views need to be planned in real time. To address this challenge, we introduce Pixel-Per-Area (PPA) as a reconstruction quality proxy and plan views by maximizing the PPA of the faces of a simple geometric shape representing the actor. Through experiments in simulation, we show that PPA is highly correlated with reconstruction quality. We also conduct real-world experiments showing that our system can produce dynamic 3D reconstructions of good quality. We discuss how this view planning algorithm can be extended with predicted future human poses on human surface representation. We share our code for the simulation experiments in the link: https://github.com/Qingyuan-Jiang/view_planning_3dhuman.**

I. INTRODUCTION

Capturing a human actor's motion with fine details is essential due to its applications in virtual reality and related metaverse applications. However, obtaining such *high-fidelity* reconstructions [1], [2] remains a challenging problem due to the actor's motion. Recently developed multi-camera systems can generate reconstructions at the sub-pixel level of hand details or facial gestures [3], [4]. The primary limitation of these systems is that they rely on stationary, pre-calibrated cameras, which confine the user to a motion capture area. In this work, we tackle the challenge of acquiring images with drones for the high-fidelity reconstruction of a dynamic human actor. We focus on the view planning strategy to improve the reconstruction quality.

There are many technical challenges in designing such a view planning algorithm when the actor is dynamic: The planning space is heavily enlarged due to the uncertainty of the actor's movement. Also, the target surface keeps changing; in contrast, existing works on high-fidelity human 3D reconstruction are computationally costly and slow. There is no prior information on the human surface and no closed-form objective functions to plan in real time.

Therefore, we present a view planning strategy that addresses these challenges (Fig. 1). We 1) present a new objective function, Pixels-Per-Area (PPA), to measure the fidelity of 3D reconstruction for a single 3D patch. 2) We represent the human out-surface as a set of 3D patches and propose the view planning algorithm that optimizes PPA proxy based on the set. We investigate our view planning performance in different fidelity: from highest fidelity, where each patch is a triangle prism from the observed partial mesh, to the simplest, a cuboid with five faces.

Fig. 1. **Capturing images with a flying camera for the high-fidelity reconstruction of a dynamic actor.** We build a drone system to capture a dynamic actor's high-fidelity 3D reconstruction and study the view planning strategy for better reconstruction quality.

Our contributions can be summarized as follows.

- We propose to use the Pixels-Per-Area (PPA) function as a proxy for reconstruction quality. We formulate the view planning algorithm as the optimization problem of maximizing the PPA proxy.
- Through experiments, we show that PPA proxy is highly correlated with 3D reconstruction quality. Meanwhile, we show the impact of the human representation in different fidelity on the view planning algorithm performance.
- We built a drone system that can produce high-fidelity 3D reconstructions of a dynamic, moving human actor.

II. RELATED WORK

Reconstruction with flying cameras has received significant attention. Many works aim for a dynamic human target and a 3D skeleton pose reconstruction. Meanwhile, other works build high-fidelity reconstructions for large-scale static objects like buildings.

A. Human pose reconstruction with drones

There is an increasing amount of work on controlling drones to extract human skeleton poses actively, with the purpose of better pose reconstruction quality [5]–[7] or better artistic meaning [8], [9]. On the other hand, some studies

*This work is supported by the NSF NRI Grant #2022894.

[1]Qingyuan Jiang and Volkan Isler are with Department of Computer Science, University of Minnesota, Twin Cities, Minneapolis, MN, 55455 {jian0345, isler}@umn.edu

extend this setting to multiple drones [10]–[15]. Compared to these works, our system is designed for high-fidelity reconstructions beyond skeleton poses by proposing a proxy for reconstruction quality and using geometric primitives as intermediate actor models.

B. View planning for high-fidelity 3D reconstruction of static objects

Researchers also use drones to reconstruct objects in high fidelity on a large scale, such as static buildings [16]–[19], landslide [20], orchards [21]–[23], or urban areas [24], [25]. While such UAV systems can produce 3D reconstruction results for static targets, reconstructing a dynamic object is intractable because the target's surface changes across time, and the Next Best View algorithms do not work in such cases. In our work, we model the moving human as a set of observed patches and update them in real time.

III. PROBLEM FORMULATION

In this section, we introduce the representation of the human surface and formulate the view planning on such representation as an optimization problem (Fig. 2).

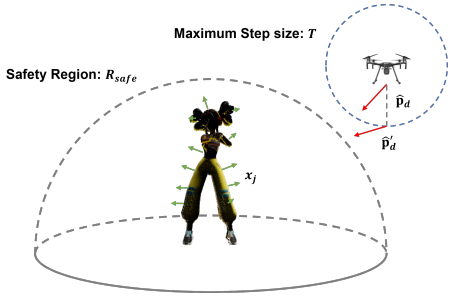

Fig. 2. **Formulation**. We model the actor as a set of geometric primitives with surface normals. We would like to find a viewpoint that minimizes the reconstruction error while maintaining a minimum distance to ensure the actor's safety.

A. Representation

We model the actor as a set of patches. A patch is defined as a planar surface of bounded size, represented with i) a point at its centroid and ii) a normal vector indicating the orientation. For example, a triangle from a mesh can be treated as a patch in the highest resolution.

Suppose we have $\mathbf{m}$-patches from the actor to visit in a 3-dimensional space. Mathematically, we use the geometric center as its representation point. We use $\mathbf{p}_j \in \mathbb{R}^3$ to denote the position of the point with index j and use $\mathbf{n}_j \in \mathbb{R}^3$ to denote the normal vector of the patch. The pose of the j-th patch is denoted by $\mathbf{x}_j \in \mathbb{R}^6$, where $\mathbf{x}_j = (\mathbf{p}_j, \mathbf{n}_j)$. The actor is modeled as the set of patches $\mathcal{X} = \{\mathbf{x}_j\}$. Meanwhile, we use $\mathbf{x}_a = (\mathbf{p}_a, \mathbf{n}_a)$ to denote the actor's pose as a whole. Note that all normal vectors has normalized length, i.e. $\|\mathbf{n}_j\| = 1$, $\|\mathbf{n}_a\| = 1$.

We denote the localization estimation of a drone's pose as $\hat{\mathbf{x}}_d = (\hat{\mathbf{p}}_d, \hat{\mathbf{n}}_d)$. From the current drone's pose $\hat{\mathbf{x}}_d$, we estimate the actor as a set of patches. We denote it as $\mathcal{X}_{est} = \{\hat{\mathbf{x}}_j\}$. Similarly, we have $\|\mathbf{n}_d\| = \|\hat{\mathbf{n}}_d\| = \|\hat{\mathbf{n}}_j\| = 1$.

B. Pixels-Per-Area (PPA) as reconstruction quality proxy

Because we do not have the ground truth of human patches $\mathbf{x}_j$, we define Pixels-Per-Area (PPA) in this part as a proxy for the reconstruction quality. We define the Pixels-Per-Area (PPA) as the projection area in the image plane of a 3D patch (Fig. 3), which is a function of the drone's pose $\mathbf{x}_d$ and the pose of a patch $\mathbf{x}_j$:

$$\mathbf{ppa}\left(\mathbf{x}_d, \mathbf{x}_j\right) = \frac{\cos(\alpha(\mathbf{n}_d, \mathbf{n}_j))}{d(\mathbf{p}_d, \mathbf{p}_j)} \tag{1}$$

$\alpha(\mathbf{n}_d, \mathbf{n}_j)$ defines the acute angle between $\mathbf{n}_j$ and $\mathbf{n}_d$. $d(\mathbf{p}_d, \mathbf{p}_j)$ defines the Euclidean distance between the drone and the patch given as: $\cos\left(\alpha(\mathbf{n}_d, \mathbf{n}_j)\right) = \frac{\mathbf{n}_d \cdot \mathbf{n}_j}{\|\mathbf{n}_d\| \cdot \|\mathbf{n}_j\|}$, $d(\mathbf{p}_d, \mathbf{p}_j) = \|\mathbf{p}_d - \mathbf{p}_j\|$. As its name implies, the PPA function describes the pixels occupied in the image plane by a 3D area. We use PPA as a proxy of the reconstruction quality, and we will show later in Sec. VI-A the correlation between PPA values and the reconstruction quality.

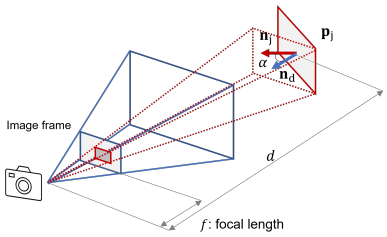

Fig. 3. **Geometric meaning of the PPA value**. We define the PPA value of a patch as the ratio between the projection area in the image plane (colored in blue) and the patch's original area. (colored in red). f: the focal length of a camera.

C. Formulation

Given an estimation of the drone's pose $\hat{\mathbf{x}}_d$ and actor observation $\mathcal{X}_{est}$, we would like to compute a new view point $\hat{\mathbf{x}}'_d = \left(\hat{\mathbf{p}}'_d, \hat{\mathbf{n}}'_d\right)$ in a local area with better reconstruction quality. Before that, we need to define the safety region and the maximum step size.

1) Safety regions: To ensure the actor's safety, we define a hemispherical space around the actor as the safe region. We use $R_{\text{safe}} > 0$ to denote the radius of the hemispherical space. The distance between the updated viewpoint and the actor's position should always be greater than R_{safe}.

2) Maximum step size: To constraint the new viewpoint close to the current estimation, we define a maximum step size T, such that $\|\hat{\mathbf{p}}'_d - \hat{\mathbf{p}}_d\| \leq T$.

3) Formulation: Now that we are ready to formulate the problem. Given an estimation of camera pose $\hat{\mathbf{x}}_d$ and an estimation of the actor model from reconstruction $\mathcal{X}_{est}$, we would like to find a new viewpoint $\hat{\mathbf{x}}'_d$ within step size T and out of safety region, such that the PPA value is maximized.

$$\begin{aligned} \hat{\mathbf{x}}'_d = \arg\max_{\hat{\mathbf{x}}'_d} \quad & \sum_j \mathbf{ppa}\left(\hat{\mathbf{x}}'_d, \hat{\mathbf{x}}_j\right) \\ \text{s.t.} \quad & \|\hat{\mathbf{p}}'_d - \hat{\mathbf{p}}_d\| \leq T, \quad \|\hat{\mathbf{p}}'_d - \mathbf{p}_a\| \geq R_{\text{safe}} \end{aligned} \tag{2}$$

We will solve the formulated problem above with our view planning module described in Sec. IV.

IV. VIEW PLANNING METHODOLOGY

We make a one-step update on the drone's pose estimation by maximizing the summation of PPA values throughout patches with Eq. 2. To do so, we calculate the gradient from the Jacobian vector with respect to the drone's pose estimation as described in Eq. 3, whose components are given in Eq. 4 and Eq. 5.

$$\frac{\partial}{\partial \hat{\mathbf{x}}_d} \sum_j \mathbf{ppa}\left(\hat{\mathbf{x}}_d^{'}, \hat{\mathbf{x}}_j\right) = \sum_j \begin{bmatrix} \frac{\partial}{\partial \hat{\mathbf{p}}_d} \mathbf{ppa}\left(\hat{\mathbf{x}}_d^{'}, \hat{\mathbf{x}}_j\right) \\ \frac{\partial}{\partial \hat{\mathbf{n}}_d} \mathbf{ppa}\left(\hat{\mathbf{x}}_d^{'}, \hat{\mathbf{x}}_j\right) \end{bmatrix} \tag{3}$$

$$\frac{\partial}{\partial \hat{\mathbf{p}}_d} \sum_j \mathbf{ppa}\left(\hat{\mathbf{x}}_d^{'}, \hat{\mathbf{x}}_j\right) = -\sum_j \frac{\hat{\mathbf{n}}_d \cdot \hat{\mathbf{n}}_j}{\|\hat{\mathbf{p}}_d - \hat{\mathbf{p}}_j\|^3} \cdot (\hat{\mathbf{p}}_d - \hat{\mathbf{p}}_j) \tag{4}$$

$$\frac{\partial}{\partial \hat{\mathbf{n}}_d} \sum_j \mathbf{ppa}\left(\hat{\mathbf{x}}_d^{'}, \hat{\mathbf{x}}_j\right) = \sum_j \frac{\hat{\mathbf{n}}_j - (\hat{\mathbf{n}}_d \cdot \hat{\mathbf{n}}_j) \cdot \hat{\mathbf{n}}_d}{\|\hat{\mathbf{p}}_d - \hat{\mathbf{p}}_j\|} \tag{5}$$

By calculating the Jacobian vector, we obtain the gradient of PPA values with respect to the drone's poses and update our drone's pose by following its direction and moving by a step size ΔT in such a way that the constraints described in Eq. 2 are satisfied.

$$\begin{aligned} &\hat{\mathbf{x}}_d^{'} = \hat{\mathbf{x}}_d + \frac{\partial}{\partial \hat{\mathbf{x}}_d} \sum_j \mathbf{ppa}\left(\hat{\mathbf{x}}_d^{'}, \hat{\mathbf{x}}_j\right) \cdot \Delta T \\ &\text{s.t.} \quad \|\hat{\mathbf{p}}_d^{'} - \hat{\mathbf{p}}_d\| \leq T, \quad \|\hat{\mathbf{p}}_d^{'} - \mathbf{p}_a\| \geq R_{\text{safe}} \end{aligned} \tag{6}$$

V. SYSTEM DESIGN

We built the drone system to validate our view planning in a real system. Besides our online view planning algorithm for high fidelity reconstruction in Sec. IV, the system also includes the capabilities of i) Online actor localization and heading direction estimation, ii) offline reconstruction with the Iterative Closest Point (ICP) method. We assume the actor is on the ground and localize its 2D pose following [26] with a multiple-layer perception (MLP) as the heading direction estimator. Then, we build human patches based on the 2D pose and plan views as described in Sec. IV. We use RGB-D images from onboard cameras to produce high-fidelity reconstruction and calibrated views with the Iterative Closest Point method. We use DJI M100 as our working drone, with Jetson TX1 as our computing unit and Realsense D435 (15Hz) as our onboard camera.

VI. EXPERIMENTS

We study the effectiveness of our system through the following questions:

1) How does the PPA function perform as a proxy for the reconstruction quality?
2) How is the reconstruction quality improved by optimizing the PPA metric?

We conduct both simulation and real-world experiments using Microsoft Airsim [27], Unreal Engine [28], and Mixamo animations [29] for evaluating reconstructions.

A. PPA as the reconstruction quality proxy

In this part, we show the correlation between PPA values, reconstruction quality, and validity using geometry primitives to represent a human actor.

1) Experiment setup in the simulation.: We obtain ground truth from Airsim using a mesh representation. Triangles and surface normals are used as patches $\mathbf{x}_j$, with viewpoints sampled in spherical coordinates. PPA values are computed per view. We also calculate the corresponding PPA values if we model the actor as a cuboid, i.e., five faces as patches. We use the following reconstruction quality.

i) Coverage of triangles. We project each pixel from the sampled view into 3D space with depth information. We define a triangle as visible if any point from the view falls within its prism volume (height fixed at 1cm). We define coverage as the visible triangle percentage. **ii) Average Pixels-Per-Triangle**. We also use the average number of pixels on a mesh triangle as our evaluation metric.

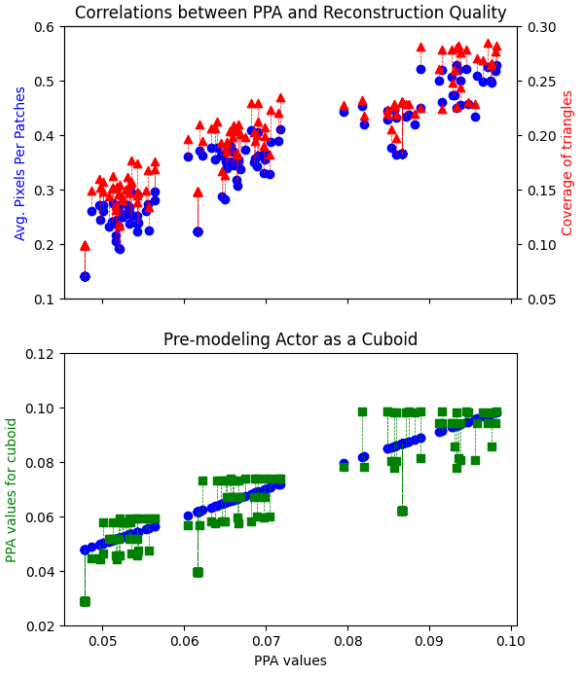

Fig. 4. **Correlation between PPA values and the reconstruction quality.** We show the correlation between PPA values and reconstruction quality and validity to use geometry primitives as the actor model. X-axis for both figures: PPA values on each human mesh triangle. Y-axis (top): metrics for reconstruction quality, average pixels per patch in blue dots, coverage ratio in red triangle. Y-axis (bottom): PPA values based on the cuboid model. Blue dots are the PPA values on each mesh triangle and hence on $y = x$ diagonal. The corresponding PPA values on the cuboid faces are connected to the green dots.

2) Results: Quantitative results are shown in Fig 4. Each dot is a sampled view. In the top figure of Fig. 4, we

TABLE I

RECONSTRUCTION RESULTS FROM VIEW PLANNING ALGORITHMS.

Modules	No Plan	Greedy	PPA Cuboid	PPA Mesh	Enum. Coverage	Enum. CD.
Coverage [%]↑	12.6	12.8	12.9	**13.7**	14.2	13.8
CD [mm] ↓	45.36	44.72	44.65	**44.24**	43.65	43.10
Coverage (noise) [%]↑	12.6	12.9	13.0	**13.6**	14.1	13.6
CD (noise) [mm] ↓	45.36	45.25	45.08	**44.46**	43.80	43.13

show a monotonic relationship between PPA and both quality metrics. The bottom figure indicates that cuboid and mesh-based PPA values are comparable, especially in areas with a planar surface (e.g., chest/back), making cuboids a reasonable simplification for planning.

B. PPA as the view planning strategy

In this part, we compare reconstruction quality using PPA-optimized views against baseline planning strategies.

1) Metric: We use triangle coverage and Chamfer Distance (CD) [30] as our evaluation metrics. Chamfer distance in the forward direction computes the accuracy of the reconstruction, whereas, in the backward direction, it models the coverage of the ground truth point cloud by the reconstruction. We report the mean of the two directions as the total reconstruction error. All numbers are reported in millimeters.

2) Baselines: We evaluate reconstruction quality from four methods as shown in Fig. 5: 1) no planning; 2) greedy planning (closest next view); 3) planning on PPA with a cuboid human model; 4) planning on PPA with observed mesh; 5) We also enumerate the viewing quality metrics in the local area and use them as our upper bound, which can not be reached during actual flights since we can not obtain the mesh ground truth of the actor surface. In practice, we set the safe radius $R_{safe} = 8$m and step size $T = 1.0$m.

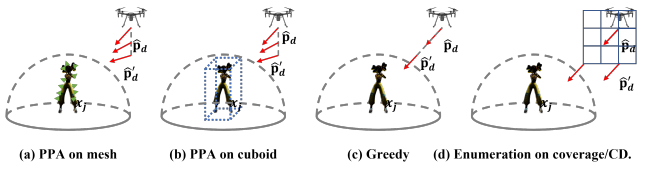

Fig. 5. **Methods and baselines** We compare our planning strategy with other baselines on the reconstruction results. From left to right, we show the view planning methods, including PPA based on the mesh, PPA based on a cuboid, greedy method, and enumeration on viewing quality.

3) Analysis: From the results in Table I, we show that optimizing our PPA optimization improves both coverage and accuracy. Meanwhile, mesh-based PPA gives better results than cuboid-based, though cuboid is more computationally efficient. Besides, we test the sensitivity of the view planning algorithm when the error of human 2D pose is included. With added Gaussian noise ($\mu = 0$m, $\sigma = 0.5$m) to actor pose, PPA-based methods remain robust, as shown in Table I.

4) Real-world Results: We also show qualitative results from the real-world experiments in Fig. 6, where we visualize the reconstructed dynamics of the actor walking in the 3D world. Real-world reconstructions confirm the effectiveness of PPA-based view planning. More results are included in the supplementary video.

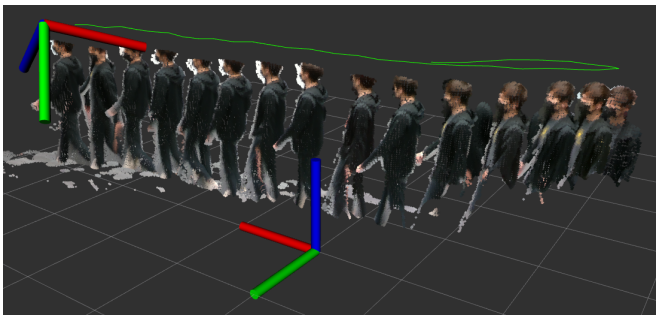

Fig. 6. **Reconstruction results from real drone setup.** We show our reconstructed human actor from the real drone in ROS Rviz. Green lines are the calibrated camera path. Results are shown as a point cloud in the middle. We plot the world and camera frame on the mid-bottom and top-left of the image.

VII. HUMAN MOTION PREDICTION EXTENSION

In the section above, we solve the view planning problem by maximizing PPA values over a set of patches representing the human surface, observed in real time. One limitation of the algorithm is that it 'responds' to human motion and, therefore, tracks the human passively. Our recent work [31] predicts long-term human skeleton poses from past poses and the surrounding environment. As an extension of this view planning algorithm, we would like to model human patches considering future poses. Our minimization process will remain similar but have a time dimension and can be solved by dynamic programming for an optimal solution or by other methods for a sub-optimal solution. This way, we can actively plan future views with predictions as prior knowledge.

VIII. CONCLUSION

This paper presented a view planning method for capturing high-fidelity 3D reconstructions of a dynamic actor based on a Pixels-Per-Area (PPA) function, used as a proxy for reconstruction quality. We modeled the actor surface as a set of 3D patches and experimented with different geometry representations. Experiments in simulation validated the correlation between PPA and reconstruction quality, showing improved reconstruction results with our proposed view-planning method. We also built a real-world drone system and demonstrated successful reconstructions in a real setting. In future work, we aim to extend the view planning algorithm by leveraging human motion prediction results and considering future surfaces in the optimization process.

REFERENCES

[1] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, H. Farid, and T. Kanade, "Virtual space teleconferencing using a sea of cameras," in *Proc. First International Conference on Medical Robotics and Computer Assisted Surgery*, vol. 26, 1994.

[2] J. Lanier, "Virtually there," *Scientific American*, vol. 284, no. 4, pp. 66–75, 2001, publisher: JSTOR.

[3] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic Studio: A Massively Multiview System for Social Motion Capture," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015.

[4] Z. Yu, J. S. Yoon, I. K. Lee, P. Venkatesh, J. Park, J. Yu, and H. S. Park, "HUMBI: A Large Multiview Dataset of Human Body Expressions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] X. Zhou, S. Liu, G. Pavlakos, V. Kumar, and K. Daniilidis, "Human Motion Capture Using a Drone," *arXiv:1804.06112 [cs]*, Apr. 2018, arXiv: 1804.06112. [Online]. Available: http://arxiv.org/abs/1804.06112

[6] A. Pirinen, E. Gärtner, and C. Sminchisescu, "Domes to drones: Self-supervised active triangulation for 3d human pose reconstruction," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[7] S. Kiciroglu, H. Rhodin, S. N. Sinha, M. Salzmann, and P. Fua, "ActiveMoCap: Optimized Viewpoint Selection for Active Human Motion Capture," *arXiv:1912.08568 [cs]*, June 2020, arXiv: 1912.08568. [Online]. Available: http://arxiv.org/abs/1912.08568

[8] R. Bonatti, W. Wang, C. Ho, A. Ahuja, M. Gschwindt, E. Camci, E. Kayacan, S. Choudhury, and S. Scherer, "Autonomous aerial cinematography in unstructured environments with learned artistic decision-making," *Journal of Field Robotics*, vol. 37, no. 4, pp. 606–641, June 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/rob.21931

[9] R. Bonatti, A. Bucker, S. Scherer, M. Mukadam, and J. Hodgins, "Batteries, camera, action! Learning a semantic control space for expressive robot cinematography," *arXiv:2011.10118 [cs]*, Mar. 2021, arXiv: 2011.10118. [Online]. Available: http://arxiv.org/abs/2011.10118

[10] A. Alcántara, J. Capitán, R. Cunha, and A. Ollero, "Optimal Trajectory Planning for Cinematography with Multiple Unmanned Aerial Vehicles," *Robotics and Autonomous Systems*, vol. 140, p. 103778, June 2021, arXiv: 2009.04234. [Online]. Available: http://arxiv.org/abs/2009.04234

[11] R. Tallamraju, N. Saini, E. Bonetto, M. Pabst, Y. T. Liu, M. J. Black, and A. Ahmad, "AirCapRL: Autonomous Aerial Human Motion Capture using Deep Reinforcement Learning," *arXiv:2007.06343 [cs]*, Aug. 2020, arXiv: 2007.06343. [Online]. Available: http://arxiv.org/abs/2007.06343

[12] R. Tallamraju, E. Price, R. Ludwig, K. Karlapalem, H. H. Bulthoff, M. J. Black, and A. Ahmad, "Active Perception Based Formation Control for Multiple Aerial Vehicles," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4491–4498, Oct. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8784232/

[13] N. Saini, E. Price, R. Tallamraju, R. Enficiaud, R. Ludwig, I. Martinovic, A. Ahmad, and M. J. Black, "Markerless outdoor human motion capture using multiple autonomous micro aerial vehicles," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 823–832.

[14] A. Ahmad, E. Price, R. Tallamraju, N. Saini, G. Lawless, R. Ludwig, I. Martinovic, H. H. Bülthoff, and M. J. Black, "AirCap – Aerial Outdoor Motion Capture," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2019), Workshop on Aerial Swarms*, Nov. 2019.

[15] C. Ho, A. Jong, H. Freeman, R. Rao, R. Bonatti, and S. Scherer, "3D Human Reconstruction in the Wild with Collaborative Aerial Cameras," *arXiv:2108.03936 [cs]*, Aug. 2021, arXiv: 2108.03936. [Online]. Available: http://arxiv.org/abs/2108.03936

[16] C. Wefelscheid, R. Hänsch, and O. Hellwich, "Three-Dimensional Building Reconstruction Using Images Obtained by Unmanned Aerial Vehicles," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXVIII-1/C22, pp. 183–188, Sept. 2012. [Online]. Available: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XXXVIII-1-C22/183/2011/

[17] S. Daftry, C. Hoppe, and H. Bischof, "Building with drones: Accurate 3D facade reconstruction using MAVs," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 3487–3494, iSSN: 1050-4729.

[18] T. Li, S. Hailes, S. Julier, and M. Liu, "UAV-based SLAM and 3D reconstruction system," in *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec. 2017, pp. 2496–2501.

[19] T. Koch, M. Körner, and F. Fraundorfer, "Automatic and Semantically-Aware 3D UAV Flight Planning for Image-Based 3D Reconstruction," *Remote Sensing*, vol. 11, no. 13, p. 1550, June 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/13/1550

[20] S. K. Gupta and D. P. Shukla, "Application of drone for landslide mapping, dimension estimation and its 3D reconstruction," *Journal of the Indian Society of Remote Sensing*, vol. 46, no. 6, pp. 903–914, June 2018. [Online]. Available: http://link.springer.com/10.1007/s12524-017-0727-1

[21] C. Peng and V. Isler, "Adaptive View Planning for Aerial 3D Reconstruction," *arXiv:1805.00506 [cs]*, Sept. 2019, arXiv: 1805.00506. [Online]. Available: http://arxiv.org/abs/1805.00506

[22] ——, "View Selection with Geometric Uncertainty Modeling," *arXiv:1704.00085 [cs]*, Feb. 2018, arXiv: 1704.00085. [Online]. Available: http://arxiv.org/abs/1704.00085

[23] ——, "Visual Coverage Path Planning for Urban Environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5961–5968, Oct. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9145689/

[24] H. Zhang, Y. Yao, K. Xie, C.-W. Fu, H. Zhang, and H. Huang, "Continuous aerial path planning for 3D urban scene reconstruction," *ACM Transactions on Graphics*, vol. 40, no. 6, pp. 1–15, Dec. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3478513.3480483

[25] W. Wahbeh, G. Müller, M. Ammann, and S. Nebiker, "Automatic Image-Based 3D Reconstruction Strategies for High-Fidelity Urban models-Comparison And Fusion of UAV And Mobile Mapping Imagery for Urban Design Studies," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B2-2022, pp. 461–468, May 2022. [Online]. Available: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B2-2022/461/2022/

[26] W. Wang, A. Ahuja, Y. Zhang, R. Bonatti, and S. Scherer, "Improved Generalization of Heading Direction Estimation for Aerial Filming Using Semi-Supervised Regression," in *2019 International Conference on Robotics and Automation (ICRA)*. Montreal, QC, Canada: IEEE, May 2019, pp. 5901–5907. [Online]. Available: https://ieeexplore.ieee.org/document/8793994/

[27] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles," in *Field and Service Robotics*, 2017, _eprint: arXiv:1705.05065. [Online]. Available: https://arxiv.org/abs/1705.05065

[28] Epic Games, "Unreal Engine," Apr. 2019. [Online]. Available: https://www.unrealengine.com

[29] Adobe Systems Inc., "Mixamo." [Online]. Available: https://www.mixamo.com

[30] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.

[31] Q. Jiang, B. Susam, J.-J. Chao, and V. Isler, "Map-Aware Human Pose Prediction for Robot Follow-Ahead," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 13 031–13 038.