M2S2: A Multi-Modal Sensor System for Remote Animal Motion Capture in the Wild*

Azraa Vally[1†], Gerald Maswoswere[1†], Nicholas Bowden[1], Stephen Paine[1], Paul Amayo[1], Andrew Markham[2], and Amir Patel[1,3], *Senior Member, IEEE*

.

Abstract—Capturing animal locomotion in the wild is far more challenging than in controlled laboratory settings. Wildlife subjects move unpredictably, and issues like scaling, occlusion, lighting changes, and the lack of ground truth data make motion capture difficult. Unlike human biomechanics, where machine learning thrives with annotated datasets, such resources are scarce for wildlife. Multi-modal sensing offers a solution by combining the strengths of various sensors, such as LiDAR and thermal cameras, to compensate for individual sensor limitations. Additionally, some sensors, like LiDAR, can provide training data for monocular pose estimation models. We introduce M2S2, a Multi-Modal Sensor System for capturing animal motion in the wild. M2S2 integrates RGB, depth, thermal, event, LiDAR, and acoustic sensors to overcome challenges like synchronization and calibration. We showcase its application with data from cheetahs, offering a new resource for advancing sensor fusion algorithms in wildlife motion capture.

Index Terms—Sensor systems, motion capture, pose estimation, sensor fusion

I. INTRODUCTION

Understanding animal motion in natural environments is essential for insights into ecology, evolutionary biology, and neuroscience. While controlled laboratory settings have been informative, the complexity of outdoor environments poses significant sensing challenges like occlusion and lighting variability. In human motion capture, deep learning has addressed these challenges but requires extensive ground truth data currently impractical for wildlife [1].

Multi-modal sensor data integration can overcome these obstacles by leveraging the strengths of different modalities [2]. This approach not only provides robust state estimation but also opens up new possibilities for generating ground truth data. However, a comprehensive system for capturing synchronized multi-modal animal data in the field is non-existent.

We present the M2S2 (Multi-Modal Sensor System), an open-source platform that captures synchronized high-speed RGB, depth camera, thermal, mmWave radar, event camera, acoustic, and LiDAR data from animals. This system addresses the challenges of time synchronization and extrinsic calibration of diverse sensors. We demonstrate the system's capabilities with data captured from cheetahs during high-speed pursuits and low-light conditions, providing a valuable resource for developing new fusion algorithms for wildlife pose estimation.

All code and data are available on our GitHub page: M2S2 GitHub.

*This work was supported by the National Research Foundation (NRF) of South Africa grant number 137762 as well as the Google Research Scholar Program.
Authors are with the [1]Department of Electrical Engineering at the University of Cape Town, South Africa, [2] Department of Computer Science, University of Oxford, UK [3] Department of Computer Science, University College London, UK. [†]Azraa Vally and Gerald Maswoswere are co-first authors: vllazr002@myuct.ac.za, mswger001@myuct.ac.za.

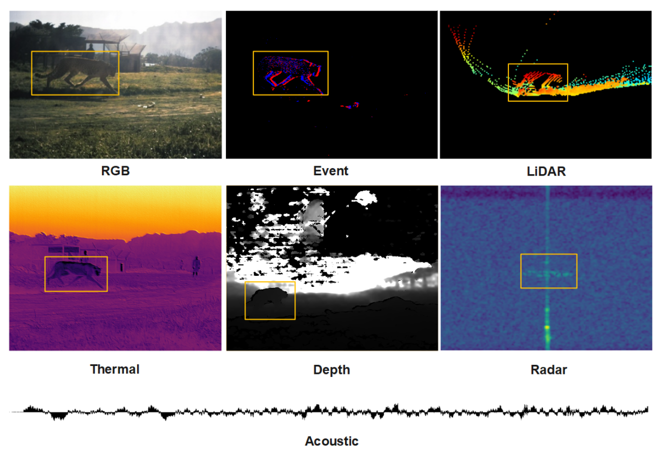

Fig. 1. Example of Multi-modal sensor data captured by M2S2 of cheetahs running in the wild.

II. LITERATURE REVIEW

A. Multi-modal Sensor Systems

Multi-modal sensing refers to the integration of multiple sensory inputs, such as vision, audio, touch, and more, to gather a comprehensive understanding of an environment or object. This approach offers several advantages, including enhanced perception and context awareness, improved robustness in noisy or dynamic settings, and the ability to provide richer and more accurate information for applications such as autonomous vehicles [2] and pose estimation [3].

Multi-modal data is challenging to generate for two primary reasons. The first is that time-synchronization of sensors with varying data rates requires careful software design. The second reason is the extrinsic calibration (estimating the pose between various sensors) can be a challenge as it requires multiple calibration experiments (no single method works for all sensors). Nevertheless, we have summarized a collection of notable multi-modal sensor systems in Table 1. As can be seen, our proposed system is the first to cover such a multitude of sensors for any application.

TABLE 1. Comparison of Existing Multi-Modal Systems. The letters D, T, R, L, E and A stand for Depth, Thermal, Radar, LiDAR, Event and Acoustic, respectively.

	RGB	D	T	R	L	E	A
Lim et al. [4]	✓	✓		✓			
Chen et al.* [5]	✓			✓	✓		
Fang et al.* [6]	✓	✓		✓	✓		✓
Cheng et al.* [7]	✓	✓	✓	✓	✓		
Sizhe et al. [8]	✓	✓		✓			
Yang et al. [3]	✓	✓		✓	✓		
M2S2	✓	✓	✓	✓	✓	✓	✓

Systems marked with an * do not provide information regarding extrinsic calibration of the sensors used.

B. Animal Pose Datasets

Recent advancements have seen a proliferation of animal datasets aimed at various research objectives. Collections of 2D pose datasets have been developed for macaques [9], horses [10], and birds [11], along with large-scale general pose datasets like AP-36K [12]. Despite these developments, 3D animal pose datasets remain scarce, primarily due to the challenges of obtaining accurate ground truth, which often requires manual labeling from multi-view videos [13]. Marker-less motion capture systems have emerged as a promising solution for overcoming these challenges, particularly in studies involving wild animals in natural environments. Unlike traditional systems (e.g., optical markers and IMUs), which are intrusive and may alter animal behavior, marker-less systems are non-intrusive and adaptable, allowing for data collection in uncontrolled and dynamic settings [14], [15]. Noteworthy contributions in 3D pose estimation include datasets for monkeys [16], dogs [17], and various wild quadrupeds [18]. Among these, only Patel et al. [19] incorporate RGB, depth, and thermal data, albeit in a controlled setting.

III. SYSTEM DESIGN

A. Hardware Design

M2S2 combines the current state-of-the-art remote sensors into one modular system. Each of the sensors have been selected for their complementary strengths for pose estimation and are briefly described below with the specific sensors selected depicted in Fig. 2.

1) High-speed RGB camera: RGB video for pose estimation is advantageous due to its affordability as well as the density of its data. It has limitations, including difficulties in low-light, glare, or occluded scenarios and challenges in precise 3D pose estimation without additional sensors. It has successfully been employed in animal pose inference using both multi-view [20] and monocular camera data [21].

2) Depth camera: Depth cameras, utilizing structured light or time-of-flight to emit infrared patterns, measure distortion to calculate depth [22]. They offer accurate 3D data for pose estimation but are limited by strong ambient light and range. These technologies have been effectively used for tracking animals in controlled environments, like mice in laboratories [19] and domestic dogs, as demonstrated by the RGBD-Dog dataset provided by Kearney et al. [23]. However, their application has not yet extended to wild animals.

3) Thermal Camera: Advantages of thermal cameras are that they can operate in complete darkness and are insensitive to visible light conditions, making them suitable for nighttime and low-visibility scenarios which are typical for wildlife. They capture thermal radiation which is less affected by subject appearance [24]. However, they cannot capture color and fine texture details, limiting pose estimation precision. Animal pose estimation with thermal cameras is yet to be explored.

4) Event Camera: An event camera is a specialized type of camera that captures changes in pixel intensity rather than traditional frames. Advantages include high temporal resolution for fast motion tracking and low power consumption. However, they have lower spatial resolution, limiting detailed pose estimation, and require unique algorithms for data processing. Animal pose estimation using event cameras has not been investigated but it has been shown to generate extremely high-speed motion capture in humans [25].

5) LiDAR: LiDAR sensors offer precise 3D (metric) spatial information, enabling accurate pose estimation even in different lighting conditions. However, disadvantages include their relatively high cost, sparsity of point cloud data and susceptibility to interference from weather conditions like fog or rain, which can impact their reliability in adverse environments [2]. LiDAR has successfully been combined with RGB images for human pose estimation [26] but has also not been explored for animal pose estimation.

6) mmWave Radar: FMCW (Frequency Modulated Continuous Wave) mmWave radar emits continuous radio waves and measures reflections to determine distance and velocity of objects, through the Doppler shift (the output is a 3D cube with range, Doppler and angle as the 3-axes). It works in various conditions, including darkness and adverse weather. Challenges include complex signal processing and poor angular resolution. Human pose estimation has been demonstrated with mmWave-only systems and with depth cameras [3].

7) Audio: Audio's potential to enhance motion capture is often overlooked, yet it offers a unique dimension to perception. Existing research shows promise, such as estimating upper torso motion of violin players [27] or refining pose estimation in galloping horses [28]. Given its untapped potential compared to other sensors, there is a pressing need to explore audio's role in enhancing motion capture accuracy.

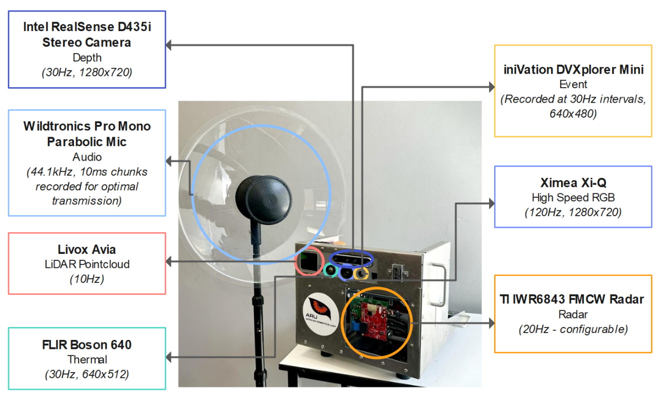

Fig. 2. Front view of M2S2 showing all sensors. Note: frame-rates for sensors were chosen to avoid memory overhead.

8) Mechanical Design: The system's mechanical design, modeled features a durable aluminum casing (365mm x 300mm) with 3D-printed sensor mounts, a tripod for the microphone, and aluminum handles for portability. All CAD files are available on our GitHub page.

B. Software Design

1) Architecture Overview: The platform integrates various sensors on a compact platform with ROS2 [29] running on an Intel NUC Mini PC (13th Gen, i7-1360P Processor, 16GB RAM) for effective data capture and synchronization. Each sensor is managed by ROS2 nodes for control and data logging, as shown in Fig. 3.

Sensor data is published to ROS2 topics and visualized using built-in tools. Unlike typical ROS2 bag storage, our setup uses eCAL [30] for data logging. eCAL's advantages over RTPS DDS middleware are discussed in Section III-B2.

2) eCAL RMW: eCAL RMW (eCAL ROS2 middleware wrapper) is based on the open-source eCAL framework [30], enabling ROS2 components to communicate using eCAL middleware. It replaces ROS2 RTPS DDS, which can be resource-intensive and introduce latency [29]. eCAL RMW focuses on simplicity and maximum performance, offering efficient data communication, reduced storage and bandwidth costs, and low-latency data transfer.

It supports tools like eCAL recorder, monitor, and player, simplifying data recording and playback. Various tests revealed eCAL consistently recorded

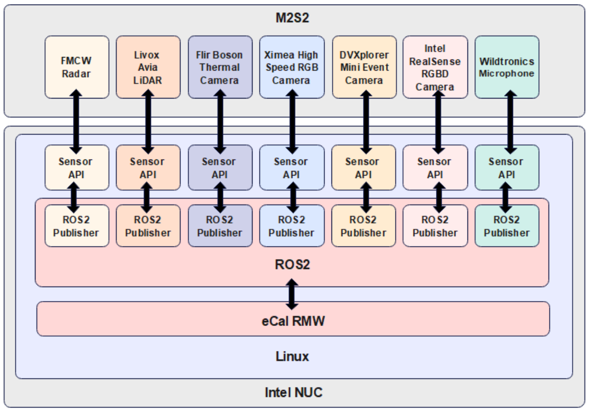

Fig. 3. Overview of M2S2 Software Architecture

high-payload data without dropping frames, outperforming the the alternative (ROS2 bags), which showed significant frame loss under increased sensor loads (approx. 30% of frames were lost with a simple two sensor setup). eCAL stores data in the accessible HDF5 format, ensuring seamless deserialization and dataset integration.

3) Synchronization: Accurate synchronization is essential for comprehensive multi-modal animal datasets. While LiDAR, Intel RealSense, and Ximea sensors support hardware synchronization, others (radar, event, thermal, microphone) rely on software alignment. We evaluated synchronization using a GPS-synchronized pulsing LED captured by RGB, depth, and event cameras, achieving 30 ms alignment (mean 10.99 ms, std. 10.28 ms). This approach suffices for many applications but may introduce errors for rapid maneuvers. Future work should incorporate advanced hardware synchronization and select sensors with built-in capabilities to enhance multi-modal data integration.

IV. CALIBRATION

Extrinsic calibration defines spatial relationships between sensors, crucial for sensor fusion in systems like M2S2 [2]. Given that no single algorithm is capable of calibrating all sensors in M2S2, a modular strategy with the Intel RealSense camera as the reference (Fig. 4) was employed. Unlike microphone arrays, the single parabolic microphone used (which has a beam width of approximately 43° at 1 kHz) is not designed to capture spatial information; therefore, spatial calibration with respect to the cameras is neither applicable nor useful for our application.

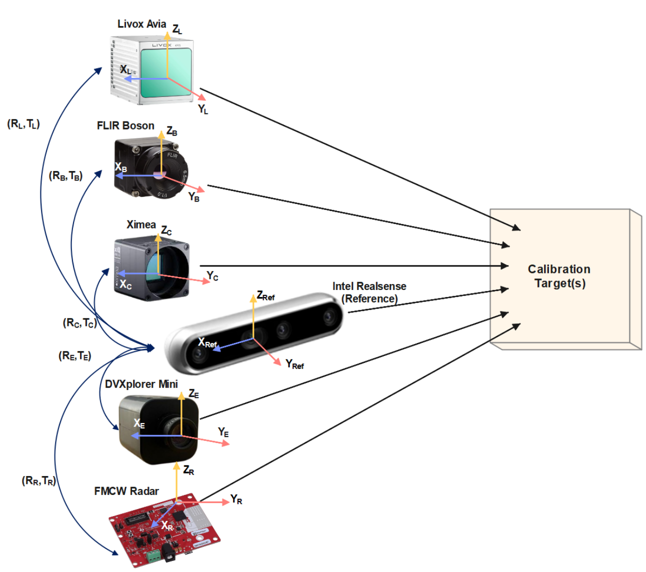

Fig. 4. Calibration Overview for M2S2, with the Intel RealSense as the reference frame.

A. Thermal Camera

For geometric calibration of thermal cameras, distinctive features visible in the IR spectrum are required, achievable by leveraging the thermal conductivity and emissivity of materials. Following the approach of ElSheikh et al. [24], we utilized a laser-cut MDF checkerboard pattern overlaid onto an aluminum surface, heated with halogen lamps for visibility in both RGB and IR cameras. A mean reprojection error of 0.45 pixels was achieved.

B. LiDAR

Geometric LiDAR-camera calibration is commonly performed by extracting the checkerboard edges from the camera images, and detecting the rectangular checkerboard plane in the LiDAR data, according to the board dimensions computed [31]. We used the MATLAB LiDAR Camera Calibrator app [32] and achieved a mean translational error of 0.027 m and a mean rotational error of 4.02°.

C. Event Camera

Event cameras detect changes in pixel intensity, with pixels acting independently. We used the E2Calib toolbox [33], which integrates event data into constant time periods and reconstructs visible images. These images were used in MATLAB Stereo Camera Calibration App [34] for extrinsic calibration. A checkerboard calibration target, as recommended in [34], was employed for this process, achieving a mean reprojection error of 0.33 pixels.

D. mmWave Radar

Radar-camera calibration was performed by determining points of correspondence (using a metal corner reflector) between the radar and camera to determine the rotation and translation of the radar as done in [35]. A RMSE of 0.18 m for the Y position and 1.82 m for X position was obtained. The large error is due to the comparatively poor angle resolution of the radar. 2 shows the radar parameters as configured for calibration.

TABLE 2. Radar Parameters

Radar Parameter	Value
Maximum Range [m]	22.5
Range Resolution [m]	0.23
Maximum Velocity [m/s]	±20
Velocity Resolution [m/s]	1.5
Azimuth Resolution [°]	15
Elevation Resolution [°]	58

E. Ximea High-speed Camera

In addition we also performed extrinsic calibration from the Ximea to the Realsense using the Matlab Stereo Camera calibrator app [34]. A checkerboard calibration target, as recommended in [34], was employed for this process. A mean reprojection error of 0.2 pixels was achieved.

V. MULTI-MODAL ANIMAL MOTION DATA

To validate the system we collected data of cheetahs during feedings and lure-chasing exercises at a wildlife sanctuary (Cheetah Outreach, South Africa). A supplementary video illustrating the system's capabilities form part of the supplementary material of the paper.

In Fig. 5 (left), we showcase integrating thermal imaging with LiDAR technology for nighttime depth perception, enabling accurate 3D pose estimation in low-light conditions. This fusion technique opens new avenues for training neural networks to convert monocular 2D thermal images into 3D pose estimations. Additionally (right), event cameras detect minute movements often missed by RGB cameras. Through calibrated integration of event and RGB cameras, we combine sparse event data with detailed RGB imagery, enriching scene understanding and facilitating denser tracking.

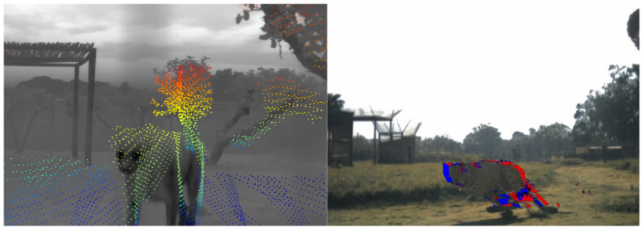

Fig. 5. *Left:* Fusion of LiDAR point cloud and thermal data of a cheetah observed at twilight. *Right:* Fusion of events and RGB images during a cheetah run.

VI. CONCLUSIONS AND FUTURE WORK

We presented M2S2, a system capturing multi-modal, time-synchronized sensor data from wildlife. We present the first multi-modal remote sensor data captured from animals movingly freely in the wild. We believe this system will enhance future 3D pose estimation of wildlife by creating novel data for sensor fusion algorithms which could make wildlife conservation more comprehensive.

ACKNOWLEDGMENTS

The authors would like to thank Cheetah Outreach for access to their cheetahs and Google for their support. Additionally the authors ould like to thank Lindelani Mbatha for his work on initial development.

REFERENCES

[1] J. Smith, P. Loncomilla, and J. Ruiz-Del-Solar, "Human pose estimation using thermal images," *IEEE Access*, 2023.

[2] Y. Almalioglu, M. Turan, N. Trigoni, and A. Markham, "Deep learning-based robust positioning for all-weather autonomous driving," *Nature Machine Intelligence*, vol. 4, no. 9, pp. 749–760, 2022.

[3] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "Mm-fi: Multi-modal non-intrusive 4D human dataset for versatile wireless sensing," *arXiv preprint arXiv:2305.10345*, 2023.

[4] T.-Y. Lim, S. A. Markowitz, and M. N. Do, "Radical: A synchronized fmcw radar, depth, imu and rgb camera data dataset with low-level fmcw radar signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 941–953, 2021.

[5] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3D detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 172–181.

[6] S. Fang, A. Sarker, Z. Wang, M. Srivastava, B. Marlin, and D. Ganesan, "Design and deployment of a multi-modal multi-node sensor data collection platform," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 1041–1046.

[7] P. Cheng, Z. Xiong, Y. Bao, P. Zhuang, Y. Zhang, E. Blasch, and G. Chen, "A deep learning-enhanced multi-modal sensing platform for robust human object detection and tracking in challenging environments," *Electronics*, vol. 12, no. 16, p. 3423, 2023.

[8] S. An, Y. Li, and U. Ogras, "Mri: Multi-modal 3D human pose estimation dataset using mmwave, rgb-d, and inertial sensors," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 414–27 426, 2022.

[9] R. Labuguen, J. Matsumoto, S. B. Negrete, H. Nishimaru, H. Nishijo, M. Takada, Y. Go, K.-i. Inoue, and T. Shibata, "Macaquepose: A novel "In the Wild" macaque monkey pose dataset for markerless motion capture," *Frontiers in Behavioral Neuroscience*, vol. 14, p. 581154, 2021.

[10] A. Mathis, T. Biasi, S. Schneider, M. Yuksekgonul, B. Rogers, M. Bethge, and M. W. Mathis, "Pretraining boosts out-of-domain robustness for pose estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1859–1868.

[11] M. Badger, Y. Wang, A. Modh, A. Perkes, N. Kolotouros, B. G. Pfrommer, M. F. Schmidt, and K. Daniilidis, "3D bird reconstruction: A dataset, model, and shape recovery from a single view," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–17.

[12] Y. Yang, J. Yang, Y. Xu, J. Zhang, L. Lan, and D. Tao, "Apt-36k: A large-scale benchmark for animal pose estimation and tracking," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 301–17 313, 2022.

[13] T. W. Dunn, J. D. Marshall, K. S. Severson, D. E. Aldarondo, D. G. Hildebrand, S. N. Chettih, W. L. Wang, A. J. Gellis, D. E. Carlson, D. Aronov *et al.*, "Geometric deep learning enables 3D kinematic profiling across species and environments," *Nature Methods*, vol. 18, no. 5, pp. 564–573, 2021.

[14] I. Arent, F. P. Schmidt, M. Botsch, and V. Dürr, "Marker-less motion capture of insect locomotion with deep neural networks pre-trained on synthetic videos," *Frontiers in Behavioral Neuroscience*, vol. 15, p. 637806, 2021.

[15] E. Gambaretto and S. Corazza, "Markerless motion capture: The challenge of accuracy in capturing animal motions through model-based approaches," in *Videometrics, Range Imaging, and Applications X*, vol. 7447. SPIE, 2009, pp. 115–121.

[16] P. C. Bala, B. R. Eisenreich, S. B. M. Yoo, B. Y. Hayden, H. S. Park, and J. Zimmermann, "Automated markerless pose estimation in freely moving macaques with openmonkeystudio," *Nature Communications*, vol. 11, no. 1, p. 4560, 2020.

[17] B. Biggs, O. Boyne, J. Charles, A. Fitzgibbon, and R. Cipolla, "Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*. Springer, 2020, pp. 195–211.

[18] J. Xu, Y. Zhang, J. Peng, W. Ma, A. Jesslen, P. Ji, Q. Hu, J. Zhang, Q. Liu, J. Wang *et al.*, "Animal3d: A comprehensive dataset of 3D animal pose and shape," *arXiv preprint arXiv:2308.11737*, 2023.

[19] M. Patel, Y. Gu, L. C. Carstensen, M. E. Hasselmo, and M. Betke, "Animal pose tracking: 3D multimodal dataset and token-based pose optimization," *International Journal of Computer Vision*, vol. 131, no. 2, pp. 514–530, 2023.

[20] D. Joska, L. Clark, N. Muramatsu, R. Jericevich, F. Nicolls, A. Mathis, M. W. Mathis, and A. Patel, "Acinoset: A 3D pose estimation dataset and baseline models for cheetahs in the wild," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 901–13 908.

[21] S. Zuffi, A. Kanazawa, T. Berger-Wolf, and M. J. Black, "Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild"," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5359–5368.

[22] A. Kadambi, A. Bhandari, and R. Raskar, "3D depth cameras in vision: Benefits and limitations of the hardware: With an emphasis on the first- and second-generation kinect models," *Computer Vision and Machine Learning with RGB-D Sensors*, pp. 3–26, 2014.

[23] S. Kearney, W. Li, M. Parsons, K. I. Kim, and D. Cosker, "Rgbd-dog: Predicting canine pose from RGBD sensors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8336–8345.

[24] A. ElSheikh, B. A. Abu-Nabah, M. O. Hamdan, and G.-Y. Tian, "Infrared camera geometric calibration: A review and a precise thermal radiation checkerboard target," *Sensors*, vol. 23, no. 7, p. 3479, 2023.

[25] L. Xu, W. Xu, V. Golyanik, M. Habermann, L. Fang, and C. Theobalt, "Eventcap: Monocular 3D capture of high-speed human motions using an event camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4968–4978.

[26] A. Zanfir, M. Zanfir, A. Gorban, J. Ji, Y. Zhou, D. Anguelov, and C. Sminchisescu, "Hum3dil: Semi-supervised multi-modal 3D human pose estimation for autonomous driving," in *Conference on Robot Learning*. PMLR, 2023, pp. 1114–1124.

[27] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman, "Audio to body dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7574–7583.

[28] L. Ci, E. Hernlund, H. Kjellstrom, and S. Zuffi, "The sound of motion: Multimodal horse motion estimation from video and audio," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

[29] "ROS 2: Robot Operating System 2," https://index.ros.org/doc/ros2/, 2023, accessed: 13-09-2023.

[30] "eCAL: Efficient Communication Abstraction Layer," https://ecal.io/, 2023, accessed: 13-09-2023.

[31] L. Zhou, Z. Li, and M. Kaess, "Automatic extrinsic calibration of a camera and a 3D lidar using line and plane correspondences," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5562–5569.

[32] MathWorks, "Lidar camera calibrator app," 2024, accessed: 2024-12-05. [Online]. Available: https://www.mathworks.com/help/lidar/ref/lidarcameracalibrator-app.html

[33] M. Muglikar, M. Gehrig, D. Gehrig, and D. Scaramuzza, "How to calibrate your event camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1403–1409.

[34] A. Fetić, D. Jurić, and D. Osmanković, "The procedure of a camera calibration using camera calibration toolbox for matlab," in *2012 Proceedings of the 35th International Convention MIPRO*, 2012, pp. 1752–1757.

[35] S. A. M. Teck-Yian Lim and M. N. Do, "Radical: A synchronized FMCW radar, depth, IMU and RGB camera data dataset with low-level FMCW radar signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 941–953, 2021.