Dynamic 3D Scene Reconstruction from Classroom Videos

Sebastian Janampa[1], Phuong Tran[1], Erik Guaderrama[1], Sylvia Celedón-Pattichis[2] , Marios S. Pattichis[1]

[1] Department of Electrical and Computer Engineering, The University of New Mexico, Albuquerque, NM 87131-0001, USA.
Email: {sebasjr1966, pnt204, eguaderrama, pattichi}@unm.edu
[2] Department of Curriculum and Instruction, College of Education, The University of Texas at Austin, Austin, TX 78712, USA.
Email: sylvia.celedon@austin.utexas.edu

2024 58th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-5405-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/IEEECONF60004.2024.10943077

***Abstract*—The paper describes the development of a system for estimating 3D speaker geometry from raw images of collaborative classroom videos. The proposed system integrates methods for 2D and 3D pose estimation with depth estimation and camera calibration to detect and reconstruct the 3D speaker geometry of a collaborative group of students. Results on the Human3.6M dataset show that the system can estimate 3D poses reasonably well without the need to pre-train on the Human3.6M dataset. Furthermore, for classroom videos, the proposed system outperformed a baseline approach trained on the Human3.6M dataset. The proposed system is used to provide the 3D speaker geometry to a new speaker diarization system that performs well in noisy classroom environments.**

***Index Terms*—scene reconstruction, 3D pose estimation, 2D pose estimation, speaker diarization.**

I. Introduction

Our current paper is focused on reconstructing the 3D student speaker geometry under occlusions. Our interest in dynamic 3D scene reconstruction from raw videos comes from the need to assess student learning in collaborative learning classrooms.

We present an example of our collaborative classroom video scene in fig. 1. Based on the raw image, we want to explore models for reconstructing the 3D student speaker geometry with respect to the table. Here, we note that a single microphone is located at the center of the table. Based on the location of the students with respect to the microphone, we can construct a 3D acoustic model for speaker diarization (determining who is talking and when) as described in [1]. We will demonstrate our approach for 3D pose estimation and for speaker diarization.

In our earlier research on classroom videos, we focused on video activity classification. In [2], we demonstrated a successful approach for dynamic participant tracking in long classroom videos. In [3], we demonstrated a fast and accurate method for detecting typing and writing activities. Similarly, we note related work in educational video analysis has been reported in [4], [5], [6].

The primary contribution of this paper is the development of a modular pipeline for reconstructing the 3D speaker geometry in collaborative classrooms. Our approach allows 3D scene reconstruction from a raw video image. Thus, it is suitable for dynamic 3D scene analysis from image samples selected from classroom videos. We demonstrate that our proposed approach enables speaker geometry based on the estimated 3D speaker geometry.

Fig. 1: Collaborative learning classroom example.

We provide extensive background for our approach in section II. We then describe our method in section III. We provide results in section IV and concluding remarks in section V.

II. Background

We summarize the background in four subsections. First, we review previous research on collaborative classroom video analysis focused on video activity recognition. Second, we describe the speaker diarization application that relies on the use of 3D speaker geometry. Third, we review methods for performing camera calibration from raw images. Fourth, we provide a summary of PoseNet, a method for estimating 3D poses from raw images.

A. Collaborative classroom video analysis

In this section, we summarize earlier research for analyzing collaborative classroom videos. We refer to [7] for a recent review of video activity recognition systems.

1) Long-Term Human Participation Assessment: In [2], the authors tackled the problem of tracking student participants in 90-minute videos. The paper introduced methods for dealing with pose variation, occlusion, and students entering and

leaving the scene dynamically. The paper used a video face recognition method described in [8]. The system demonstrated excellent results for tracking students under occlusion.

2) Long-term human video activity quantification of student participation: An effective system for recognizing student activities in classroom videos was described in [3]. In [3], the authors describe methods for recognizing typing and writing activities. The method relies on the use of a hand detection system described in [9]. The system achieved 80% accuracy, outperforming several popular methods while using 1,000 to 1,500 times fewer parameters.

B. Speaker diarization using virtual microphone array and 3D scene analysis

We review our development of Physics-based models for identifying speakers from digital videos. Our approach is to combine 3D scene analysis with audio signal processing to identify the speakers. We begin by introducing the speaker diarization problem.

Speaker diarization refers to the problem of determining the speaker within a given time interval. For our real-life collaborative classroom videos, the problem is particularly challenging. Specifically, we need to identify who is talking from a group of 3 to 5 active speakers. At the same time, we have a single microphone recording the group against a background of over 20 active speakers. In addition, our speakers are often bilingual, speaking in both English and Spanish.

More recently, we introduced an effective method for speaker diarization based on the 3D speaker geometry [1]. The approach relied on a manual estimate of the 3D speaker geometry. In this paper, we will report results from using this method based on extracting the 3D speaker methods automatically (also see earlier efforts in [10]).

The system in [1] used the 3D speaker geometry to simulate audios from the different speakers. The system also simulated recording the simulated audio signals as they would have been recorded by a virtual microphone array designed to separate the speakers. By comparing features from the simulated speakers against features extracted from the actual speakers, the system was able to identify the actual speakers. The system in [1] significantly outperformed Amazon AWS and Google Cloud in identifying speakers from 2 to 5 speakers.

C. Camera calibration using a single raw video image

Camera calibration determines the camera parameters that map 3D world coordinates to the 2D pixel coordinates as given by:

$$\rho\tilde{\mathbf{x}} = \mu\mathbf{R}[\mathbf{K}|\mathbf{t_c}]\tilde{\mathbf{X}} \tag{1}$$

where ρ and μ denote scaling factors, $\tilde{\mathbf{x}}$ denotes the 2D coordinates in homogeneous form, $\tilde{\mathbf{X}}$ denotes the 3D world coordinates in homogeneous form, $\mathbf{R}$ denotes a rotation matrix, $\mathbf{K}$ denotes the internal parameters of the camera, and $\mathbf{t_c}$ represents the position of the camera in the 3D world. The rotation matrix $\mathbf{R}$ stores the rotation angles: the pitch, roll and yaw. We assume that the yaw angle is 0° since it is not possible to estimate it from single-view images without imposing additional constraints. For the internal camera parameters, we have:

$$\mathbf{K} = \begin{pmatrix} f_x & v & x_c \\ 0 & f_y & y_c \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

where f_x and f_y denote the focal length parameters along the x- and y- axes, the principal point (x_c, y_c) represents the center of the image in each axis, and v represents the skew parameter. We consider a simplified camera model with zero skew: $v = 0$, square pixels: $f_x = f_y = f$, and $\mathbf{t_c} = \mathbf{0}$. The assumptions reduce eq. (1) to

$$\rho\tilde{\mathbf{x}} = \mu\mathbf{R}\begin{pmatrix} f & 0 & x_c \\ 0 & f & y_c \\ 0 & 0 & 1 \end{pmatrix}\mathbf{X}. \tag{3}$$

Traditional camera calibration is performed using a sequence of images of known calibration patterns (e.g., chessboard images). Alternatively, camera calibration can be performed using geometric features such as line segments and vanishing points [11], [12] .

For our paper, we consider the use of deep convolutional neural networks that can estimate camera parameter based on image content (e.g., see [13]–[17]). We also note that it is possible to combine geometric priors (line segments) with image content to improve estimation as described in [18], [19]. Unfortunately, the use of line segments requires post-processing, and the use of image content is limited by the kernel size associated with the use of convolutional neural network models.

Some of the limitations of convolutional neural networks can be addressed with the use of end-to-end transformer-based models [20]–[22]. These models use an attention mechanism to model long-distance relations between image regions [23]. They also make use of line segment information without the need for post-processing.

For the current paper, we will use the multi-Scale defOrmable transFormer (SOFI) model for camera calibratIon with enhanced line queries as described in [22]. Beyond earlier models such as [20], [21], SOFI provided intra- and cross-scale interaction and gave state-of-the-art results for camera calibration. Based on SOFI, we first extract line segments using the linear-time line segment detector [24] (LSD) followed by passing the line-segments and the input image to the network.

D. PoseNet

In [25], the authors proposed a promising method for 3D pose estimation. Their approach consisted of three models: (1) DetectNet for human pose detection, (2) PoseNet for 2D keypoint estimation with relative depth, and (3) RootNet for absolute depth estimation. We will be using PoseNet in our dynamic scene analysis method. Unfortunately, [25] is a multi-stage network that can result in slow inference for estimating poses for multiple humans in the scene.

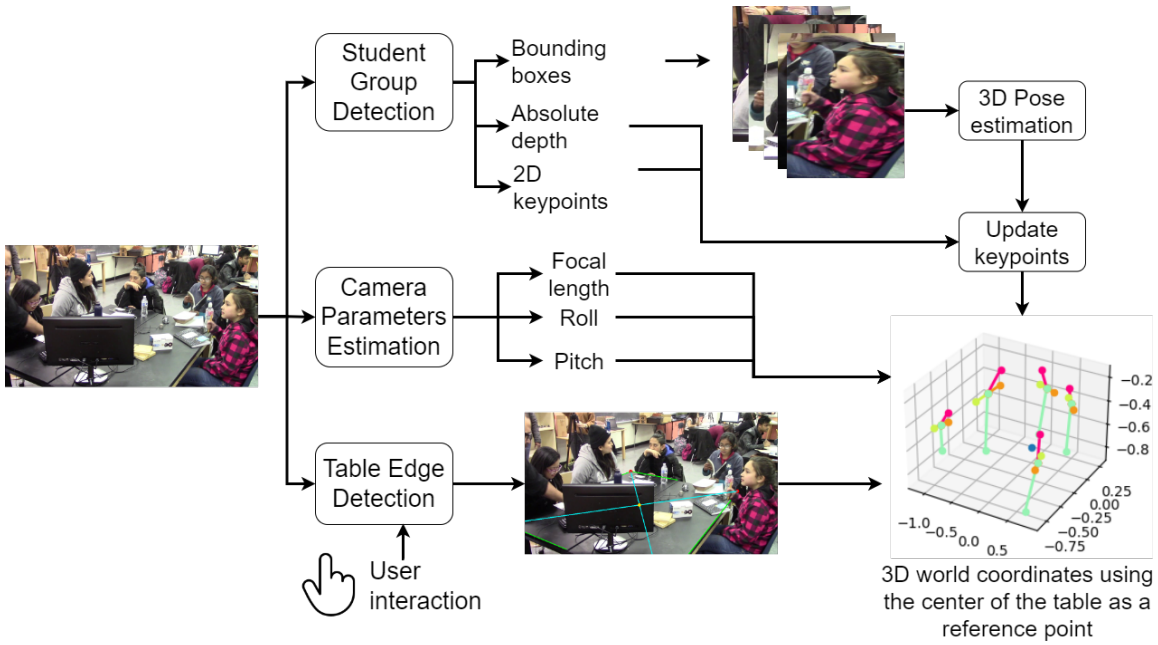

Fig. 2: An overview of our 3D scene reconstruction system based on real-life collaborative learning videos.

III. Methodology

We present an overview of the proposed system in fig. 2. Our system processes raw input videos to compute student group detection, camera parameters estimation, and table edge detection. We combine results from the three modules to reconstruct the 3D locations of each detected student. In what follows, we provide descriptions of the primary components of the system.

A. Student group detection

We provide a detailed diagram of our student group detection module in fig. 3. In parallel, our student group detection module processes a select number of raw input images to: (1) segment objects and humans in the scene, (2) estimate the depth map using Depth Anything V2 [26], and (3) estimate 2D poses for all detected persons using the YOLOv11 backbone [27].

The results of the segmentation step are further processed to remove all objects and all humans for which we get low confidence (< 0.5) For each detected person, we use the estimated depth map to select the student group that is closest to the camera. We apply bipartite matching to associate the 2D poses with the detected student groups. We then remove students for whom the head keypoint is occluded to produce the final results.

B. Keypoints updates

Given a cropped image that contains a single human, PoseNet estimates the relative depth D_r of each human keypoint with respect to the pelvis (root) keypoint. For our application, due to occlusion, we keep the keypoints associated with the nose, the shoulder thorax, and the pelvis. Then, we use the 2D keypoints J_{2D}, the absolute D_a and relative depths of each detected human and eq. (1) to estimate the 3D keypoints J_{3D} as given by:

$$J_{3D} = \rho \bar{J}_{3D} = \rho (RK)^{-1} \tilde{J}_{2D}, \tag{4}$$

where $\tilde{J}_2 D = [J_{2D}^T | 1]^T$ is the homogeneous representation of J_{2D} and ρ is a scalar which satisfies:

$$\rho \sqrt{\bar{J}_{3D}^T \bar{J}_{3D}} = D_a + D_r. \tag{5}$$

PoseNet does not perform well when humans are severely occluded. Thus, although PoseNet computes 2D keypoints, we use YOLOV11 for 2D keypoint estimation. We address PoseNet drawback by using the 2D keypoints estimation from YOLOV11.

C. Table Edge Detection

Robust table edge detection is key for reliable 3D scene reconstruction. Although we experimented with multiple versions of YOLO [27]–[30], RT-DETR [31], and YOLO-World [32], we were not able to obtain robust detections. Grounding DINO [33] gave reasonable results with a high latency of 10 seconds per image. To address the issue of object detectors, we ask the users to verify the table edges.

Once the table has been verified, we estimate the 2D center of the table as the intersection of its two diagonals as shown in fig. 2. We then use the depth map to estimate the depth corresponding to the table's center. We get the final 3D coordinates by applying eq. (4) followed by (5) (assuming $D_r = 0$).

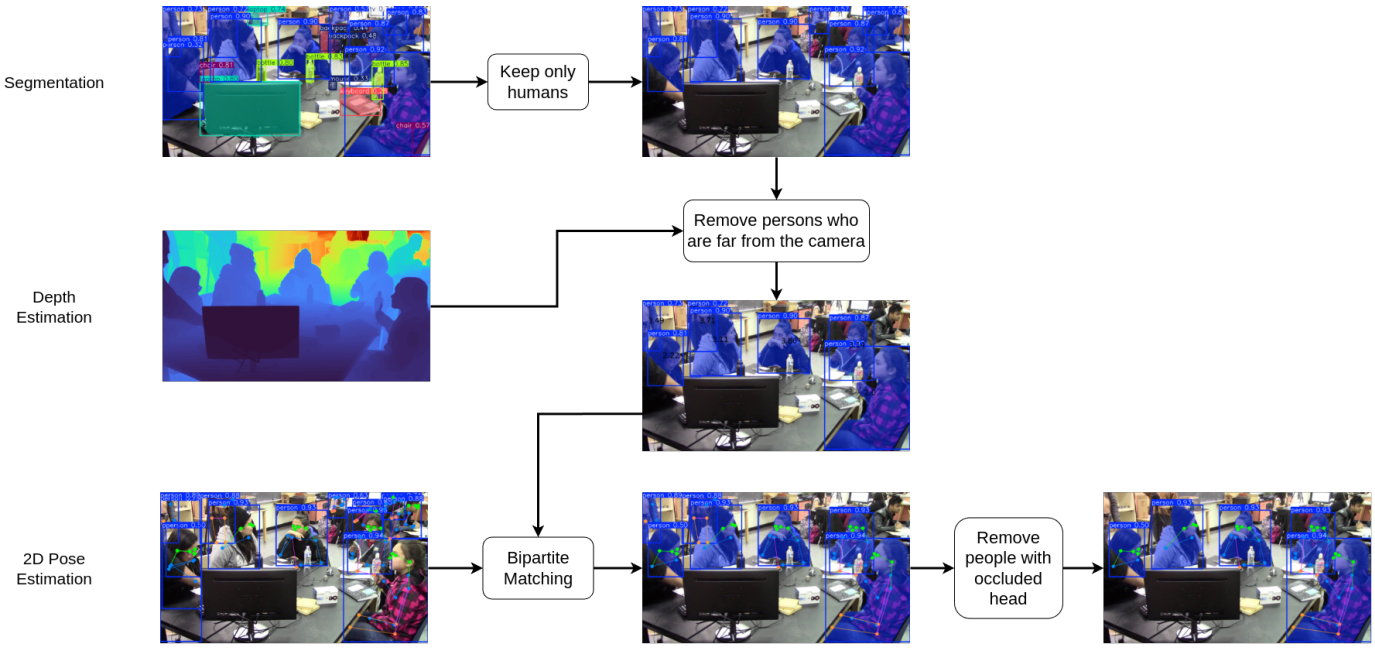

Fig. 3: Student group detection module.

Fig. 4: Robust 3D Pose estimation using YOLO and PoseNet. The left image shows the results for PosNet alone. Right image shows the results when using PoseNet with YOLO.

Human Detector	Pose Est.	Depth Est.	MPJPE (mm)
DetecNet	PoseNet	RootNet	53.70
YOLOV11	PoseNet	RootNet	54.24
YOLOV11	YOLOV11	RootNet	64.71
YOLOV11	PoseNet	Depth-AnythingV2	75.24
YOLOV11	YoloV11	Depth-AnythingV2	80.79

TABLE I: Out of distribution results for 3D pose estimation on the Human3.6M dataset.

IV. Results

We summarize our results in three subsections. First, we present results for 3D pose estimation on the Human3.6M dataset. Second, we present an example that demonstrates how our proposed approach outperforms RootNet for collaborative classroom videos. Third, we present results for speaker diarization using our estimated 3D scene geometry.

A. Out of distribution results on 3D pose estimation

We report results for out-of-distribution results for the Human3.6M dataset [34]. Human3.6M provides 3.6 million annotated poses with significant diversity based for various human activities. We summarize our results and report how each one of our components affects the estimation of the 3D joints in table I. We recall that neither YOLOV11 nor DepthAnythingV2 were trained on the Human3.6M dataset. Surprisingly, our proposed pipeline, has an increased error of just 27 mm beyond the baseline results.

At the same time, our proposed approach performs much better than the baseline method in our classroom video examples. We present an example in fig. 4. PoseNet gave poor keypoint predictions for the students on the left and right of the image. For the same example, YOLOV11 gave excellent results. Thus, the performance of the baseline method is significantly degraded for out-of-distribution datasets.

B. 3D Pose Estimation Example

In fig. 5, we present an example of our proposed method and compare it against RootNet [25]. We note that RootNet performs poorly because of the occlusion of the pelvis (root keypoint). This root keypoint is used to estimate the absolute depth. In contrast, our pipeline uses DepthAnythingV2 [26] to produce better results. Unfortunately, a limitation of our use of DepthAnythingV2 is that it requires 5 seconds to estimate a single depth map (high latency).

C. Speaker Diarization Using 3D Speaker Geometry

We next present results for speaker diarization using the 3D speaker geometry that was estimated using our proposed method. The 3D scene is described in terms of the audio sources, room size (fixed), the microphone located in the center of the table, and the table itself (with its associated absorption coefficient). Here, we note that the audio sources require an accurate estimation of the mouths of the student

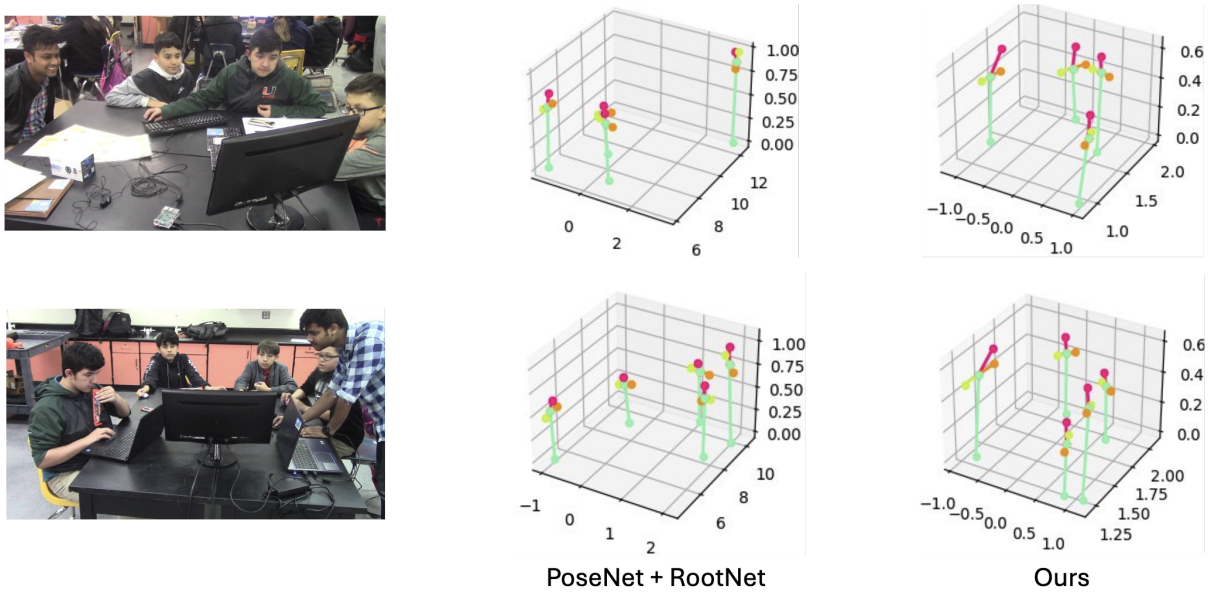

Fig. 5: Comparison of 3D human keypoint estimation between RootNet and our pipeline.

speakers derived from our model. An acoustic model is constructed that tries to identify the active speaker by comparing the actual audio recording against hypothesized speakers (see [1] for details). In what follows, we summarize the dataset and provide final results for the system.

For this experiment, we apply our dynamic scene method to estimate the 3D speaker geometry based on the first frame of an 8-minute video. Within the 8-minute video, we assume that the 3D speaker geometry will not change. For the example, we have four primary speakers that we are interested in and over 20 background speakers that we need to ignore. For the example, we assume that the four primary speakers do not talk over each other. We labeled 175 audio samples associated with one of the primary speakers. We then split our audio samples to use 80% for training, 10% for validation, and 10% for final testing. Our goal here is to use the final testing set to determine who is talking (speaker diarization).

We summarize the results in table II. Overall, the performance of the 3D scene estimate rivals the performance of setting up the parameters manually. Based on accuracy, the automated system exhibits a 4% drop from 70.58% to 66.70%.

TABLE II: Speaker diarization results. The MaxOrder parameter refers to the number of reflections used in the acoustic simulation model. The SnR parameter refers to the signal to noise ratio used by the Wiener filter simulation of the estimated room impulse response for each speaker. The optimized parameters were computed using the validation set.

Method	**MaxOrder**	**SnR**	**Acc**
Manual estimation	3	9	70.58
3D estimations	7	7	66.70

Nevertheless, the automated system can produce results based on a single video frame. Hence, the proposed system can detect changes in 3D speaker geometry resulting from routine student movements.

V. Conclusion & Future Work

Our paper summarizes our proposed approach for reconstructing 3D classroom scenes from raw images. Our approach combines the results from two different pose estimation models. Then, without training on the Human3.6M dataset, we have found that our proposed method gives acceptable results for 3D pose estimation. We have also shown that our system enables accurate speaker diarization by providing the 3D speaker geometry to the method described in [1]. We are currently developing a computer-assisted system to enable educational researchers to assess learning in collaborative learning videos.

VI. Acknowledgements

Some of the material is based upon work supported by the National Science Foundation under Grant No. 1949230 and Grant No. 1613637. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] A. Gomez, M. S. Pattichis, and S. Celedón-Pattichis, "Speaker diarization and identification from single channel classroom audio recordings using virtual microphones," *IEEE Access*, vol. 10, pp. 56256–56266, 2022.

[2] W. Shi, P. Tran, S. Celedón-Pattichis, and M. S. Pattichis, "Long-term human participation assessment in collaborative learning environments using dynamic scene analysis," *IEEE Access*, 2024.

[3] V. Jatla, S. Teeparthi, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Long-term human video activity quantification of student participation," in *2021 55th Asilomar Conference on Signals, Systems, and Computers*, pp. 1132–1135, IEEE, 2021.

[4] M. Korban, P. Youngs, and S. T. Acton, "A multi-modal transformer network for action detection," *Pattern Recognition*, vol. 142, p. 109713, 2023.

[5] J. K. Foster, M. Korban, P. Youngs, G. S. Watson, and S. T. Acton, "Automatic classification of activities in classroom videos," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100207, 2024.

[6] M. Korban, P. Youngs, and S. T. Acton, "A semantic and motion-aware spatiotemporal transformer network for action detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 6055–6069, 2024.

[7] M. S. Pattichis, V. Jatla, and A. E. ulloa Cerna, "A review of machine learning methods applied to video analysis systems," in *2023 57th Asilomar Conference on Signals, Systems, and Computers*, pp. 1161–1165, IEEE, 2023.

[8] P. Tran, M. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Facial recognition in collaborative learning videos," in *International Conference on Computer Analysis of Images and Patterns*, pp. 252–261, Springer, 2021.

[9] S. Teeparthi, V. Jatla, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Fast hand detection in collaborative learning environments," in *International Conference on Computer Analysis of Images and Patterns*, pp. 445–454, Springer, 2021.

[10] L. Sanchez Tapia, A. Gomez, M. Esparza, V. Jatla, M. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Bilingual speech recognition by estimating speaker geometry from video data," in *International Conference on Computer Analysis of Images and Patterns*, pp. 79–89, Springer, 2021.

[11] F. Schaffalitzky and A. Zisserman, "Planar grouping for automatic detection of vanishing lines and points," *Image and Vision Computing*, vol. 18, no. 9, pp. 647–658, 2000.

[12] E. Tretyak, O. Barinova, P. Kohli, and V. Lempitsky, "Geometric image parsing in man-made environments," *International Journal of Computer Vision*, vol. 97, pp. 305–321, 2012.

[13] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs, "Deepfocal: A method for direct focal length estimation," in *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 1369–1373, 2015.

[14] S. Workman, M. Zhai, and N. Jacobs, "Horizon lines in the wild," in *British Machine Vision Conference (BMVC)*, 2016.

[15] O. Bogdan, V. Eckstein, F. Rameau, and J.-C. Bazin, "Deepcalib: a deep learning approach for automatic intrinsic calibration of wide field-of-view cameras," in *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, 2018.

[16] H. Lee, E. Shechtman, J. Wang, and S. Lee, "Automatic upright adjustment of photographs with robust camera calibration," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 833–844, 2013.

[17] W. Xian, Z. Li, M. Fisher, J. Eisenmann, E. Shechtman, and N. Snavely, "Uprightnet: geometry-aware camera orientation estimation from single images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9974–9983, 2019.

[18] M. Zhai, S. Workman, and N. Jacobs, "Detecting vanishing points using global image context in a non-manhattan world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5657–5665, 2016.

[19] J. Lee, M. Sung, H. Lee, and J. Kim, "Neural geometric parser for single image camera calibration," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 541–557, Springer, 2020.

[20] J. Lee, H. Go, H. Lee, S. Cho, M. Sung, and J. Kim, "Ctrl-c: Camera calibration transformer with line-classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16228–16237, 2021.

[21] X. Song, H. Kang, A. Moteki, G. Suzuki, Y. Kobayashi, and Z. Tan, "Mscc: Multi-scale transformers for camera calibration," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3262–3271, January 2024.

[22] S. Janampa and M. Pattichis, "Sofi: Multi-scale deformable transformer for camera calibration with enhanced line queries," *arXiv preprint arXiv:2409.15553*, 2024.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[24] R. von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "LSD: A fast line segment detector with a false detection control," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 722–732, apr 2010.

[25] G. Moon, J. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image," in *The IEEE Conference on International Conference on Computer Vision (ICCV)*, 2019.

[26] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv preprint arXiv:2406.09414*, 2024.

[27] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024.

[28] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023.

[29] C.-Y. Wang and H.-Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," 2024.

[30] L. L. e. a. Ao Wang, Hui Chen, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[31] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," 2023.

[32] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," *arXiv preprint arXiv:*, 2024.

[33] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[34] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.