Exploring Aspects of Social Interaction using Machine Learning

Nele-Pauline Suffo[1], Pierre-Etienne Martin[2], Anam Zahra[2], Daniel Haun[2], & Manuel Bohn[1, 2]

[1] Institute of Psychology in Education, Leuphana University Lüneburg

[2] Max Planck Institute for Evolutionary Anthropology

Author Note

The authors made the following contributions. Nele-Pauline Suffo: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Manuel Bohn: Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to Nele-Pauline Suffo, Universitätsallee 1, 21335 Lüneburg. E-mail: nele.suffo@leuphana.de

Abstract

tbd

Exploring Aspects of Social Interaction using Machine Learning

Introduction

Methodology

The Quantex dataset includes

Dataset Description

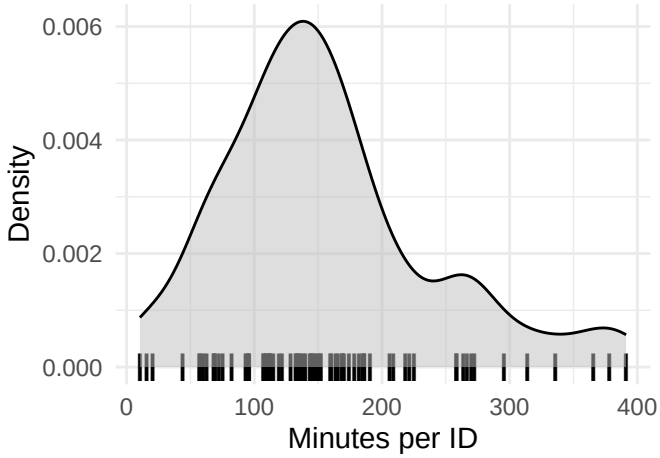

Figure 1. Video recording duration (in minutes) per Child in the Quantex Dataset.

Statistics.

Annotation Strategy. The dataset annotations cover four key elements: persons, faces, gaze direction, objects the child interacts with. Gaze information identifies whether a detected person's gaze is directed toward the child or not. For every detected person (or reflection of a person, such as in a mirror) and face, additional attributes like age and gender are collected. Faces are annotated even when occluded or blurry to ensure comprehensive coverage of interactions. Partially visible faces are also annotated if key facial features, such as the nose, eyes, or mouth, remain identifiable. Objects are categorized into six distinct groups: book, screen, animal, food, toy, and kitchenware, with

an additional category for other objects. The dataset focus is on detecting and labeling instances of (social) interaction and engagement through these key categories. The annotation strategy is displayed in Figure 2.

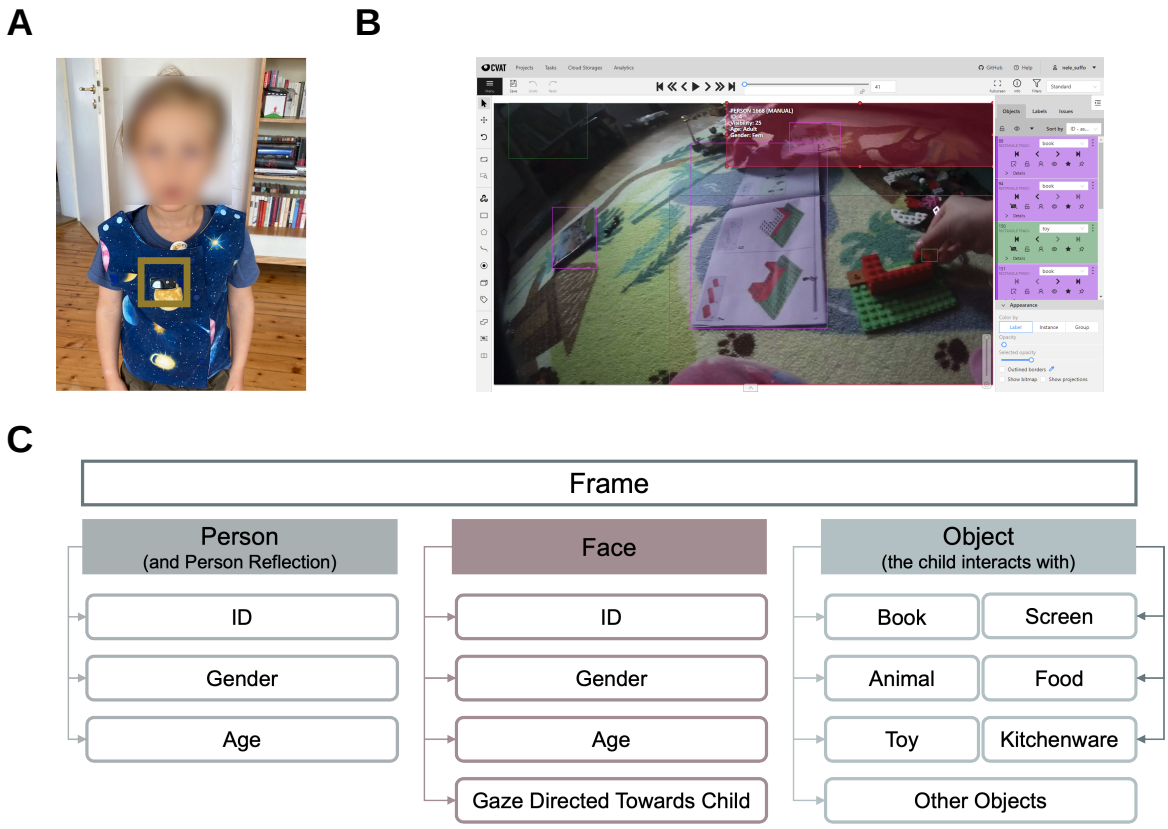

Figure 2. **A** – Vest with the embedded camera worn by the children, **B** – CVAT platform utilized for video annotation, **C** – Annotation Strategy in the Quantex dataset.

Data Collection

This study collected egocentric video recordings from 76 children, aged 3 to 5 years, over a span of 73 months. Participating families lived in a mid-sized city in Germany. To capture the children's everyday experiences, a wearable vest equipped with a camera was used, as shown in figure 2. The camera, a *PatrolEyes WiFi HD Infrared Police Body Camera*, provided high-definition video (1920x1080p at 30 fps) with a 140-degree

Table 1

Evaluation metrics for the YOLO11 face detection model trained on the Quantex dataset.

Dataset	Precision	Recall	F1-Score
Quantex	0.90	0.83	0.86

wide-angle lens and also recorded audio. Children were free to move around and engage in their usual activities at home without any interference or instructions given to their parents.

Data Preprocessing

For the video data, the annotation strategy required persons, faces, and objects to be labeled even when only partially visible, as long as key features such as facial landmarks (e.g., nose, eye, or mouth) or parts of a person or object were clearly visible. Frames that were too blurry due to movement were marked as "noise" and excluded from further analysis. Additionally, frames where the child was not wearing the camera, as well as any scenes containing nudity, were also labele d as noise and removed from the dataset. To prepare the video data for analysis, one frame per second was annotated, corresponding to every 30th frame in the video. Similarly, every 30th raw frame was extracted from the annotated video files. No preprocessing was applied to the audio data, which was used in its raw form for analysis.

Automated Analysis Pipeline

Person Detection.

Face Detection.

Gaze Classification.

Voice Detection and Classification.

Feature Extraction

Results

Presence of Aspects of Social Interaction

Presence of a Person.

Presence of a Face.

Presence of Gaze Directed at the Child.

Presence of Language.

Co-occurrence of Aspects of Social Interaction

General Discussion

References

Supplementary Material

Person Detection: Model Selection, Data Preprocessing, and Performance Evaluation

In this study, we utilized Ultralytics' YOLO11 (Jocher & Qiu, 2024), a state-of-the-art object detection model recognized for its efficiency and accuracy. Initially, we experimented with Multi-Task Cascaded Convolutional Networks (MTCNN) face detection algorithm (Zhang, Zhang, Li, & Qiao, 2016); however, it achieved a recall of less than 50% on our dataset. Furthermore, the MTCNN model demands very specific picture input for fine-tuning, which is time-consuming. In contrast, YOLO11, released in October 2024, introduces new features such as the C2PSA block, which enhances spatial attention within feature maps, allowing the model to focus more precisely on critical areas of an image. Additionally, YOLO11 incorporates the C3K2 block, designed to be faster and more efficient, enhancing the overall performance of the feature aggregation process (Khanam & Hussain, 2024). Moreover, we already had a data preparation pipeline established for YOLO11 due to its application in person detection within our project. Consequently, we selected YOLO11 for face detection and fine-tuned it on our egocentric dataset, captured using chest-mounted cameras, to adapt it to the unique characteristics of our data.

Dataset Splitting and Balancing. We started data preprocessing with a dataset comprising a total of 113799 images from 80 annotated videos.Prior to partitioning this dataset into training, validation, and testing subsets, we analyzed the proportion of images containing annotated faces versus those without. Our analysis revealed that 45.25% of the images included at least one annotated face. To maintain this inherent distribution across all subsets, we initially applied a stratified split, ensuring that each subset—training, validation, and testing—preserved the original 19% to 81% ratio of images with faces to images without faces.

However, this stratified split resulted in a significant class imbalance within the

training set, which could adversely affect the model's learning process. In imbalanced datasets, models tend to be biased toward the majority class, often predicting it more frequently while misclassifying or overlooking minority class instances. This can lead to poor recall for the minority class (Hasanin, Khoshgoftaar, Leevy, & Bauder, 2019). Additionally, Yolo11's gradient-based learning algorithm struggles to adjust decision boundaries effectively when trained on imbalanced data, potentially causing slow or unstable convergence and requiring extensive hyperparameter tuning (Kaur, Pannu, & Malhi, 2020).

To mitigate this issue, we employed an undersampling technique on the training data. Specifically, we identified the number of images containing faces in the training set (41192 frames) and randomly sampled an equal number of images from the non-face category. This approach balanced the training dataset to consist of 50% images with faces and 50% images without faces, thereby addressing the class imbalance and facilitating more effective model training.

Consequently, the final data distribution was as follows: the balanced training dataset comprised 82384 frames, while the validation and test datasets contained 11379 and 11382 frames, respectively. Notably, the validation and test sets retained the original 19% face presence, ensuring that the model's performance evaluation remained representative of the real-world data distribution.

Training and Convergence. Model training was conducted on a Linux server equipped with an Intel(R) Xeon(R) Silver 4214Y CPU @ 2.20GHz with 48 cores, a Quadro RTX 8000 GPU and 188 GB of RAM. The model was trained for a total of 86 epochs. The training process utilized YOLO11's built-in data augmentation, a batch size of 16, a cosine annealing learning rate scheduler, and early stopping after 10 epochs without improvement, with a maximum of 200 epochs.

The loss function of the YOLOv11 model comprises three main components: Box

Loss, Classification Loss, and Distribution Focal Loss (DFL). Box Loss estimates the difference between predicted bounding boxes and ground truth boxes to assess the model's localization accuracy. Classification Loss measures the model's ability to properly identify detected objects; however, in our study, this component is less relevant due to our focus on a single class—faces. DFL improves the model's ability to detect challenging objects by prioritizing difficult-to-detect instances.

During the training process, we observed that all three loss components decreased over time, indicating effective learning and improved performance, as visible in in figure 6. A steady decrease in Box Loss indicates that the model is becoming increasingly accurate in localizing faces within images. This is consistent with the rapid convergence of Classification Loss, revealing the model's ability to reliably recognize faces. The decrease in DFL over time indicates that the model is getting better at focusing on and correctly identifying difficult-to-detect faces, which improves its overall detection capabilities. These data collectively represent the model's gradual improvement in both localization and identification tasks during the training period.

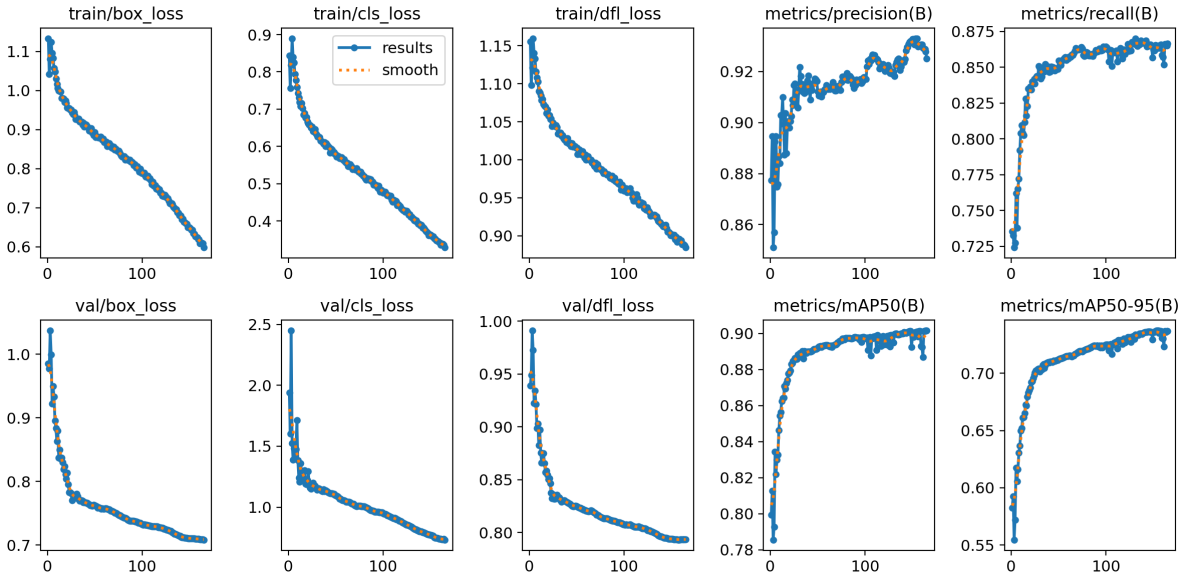

Figure 3. Training and Validation Loss Curves for the YOLO11 face detection model.

Table 2

Evaluation metrics for the YOLO11 face detection model trained on the Quantex dataset. False Positive Rate and False Negative Rate are given in percentages.

Dataset	Precision	Recall	F1-Score	False Positive Rate	False Negative Rate
Quantex	0.90	0.83	0.86	2.1	14.0

A **B**

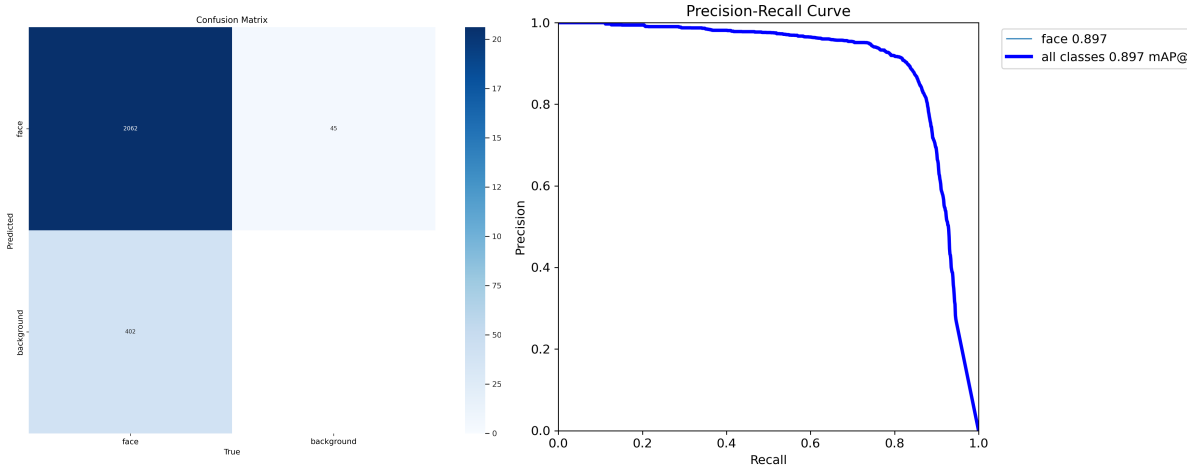

Figure 4. **A** - Confusion Matrix for the YOLO11 face detection model trained on the Quantex dataset. **B** - Precision-Recall Curve for the YOLO11 face detection model.

Model Evaluation Metrics. The YOLO11 model achieved a precision of 0.92 and a recall of 0.87 on the testing set, resulting in an F1-score of 0.90. These metrics, summarized in figure **??** indicate the model's strong performance in accurately identifying faces while minimizing errors. The precision-recall curve, displayed in figure 4, further illustrates this performance, with the curve remaining close to the top-left corner. This positioning signifies that the model maintains high precision and recall across various thresholds, underscoring its effectiveness in detecting faces with confidence.

Analysis of the confusion matrix reveals that 86% of all faces are correctly identified by the model, corresponding to 1905 true positives, while 234 faces were missed (false negatives). False negatives predominantly occurred in scenarios where faces were in the background, blurred due to motion, or occluded by the child's body. In such instances, adjacent frames often provided clearer views, aiding in more accurate classification. The model exhibited a false positive rate of approximately 2.72%, with 251 images incorrectly classified as containing faces when none were present. These false positives were often attributed to objects or toys resembling facial features. Given the relatively low false positive rate, this small number should not raise significant concerns. In face detection systems, a balance between false positives and false negatives is often necessary, and a 2.1% false positive rate is generally considered acceptable.

To provide a comprehensive understanding of the model's performance, we have included visual examples of true positives, false positives, and false negatives in figure 5. These images highlight the model's strengths and areas where challenges persist, offering insights into specific scenarios that influence detection accuracy.

Overall, the YOLO11 model demonstrates robust performance in face detection tasks. However, challenges remain in dynamic scenarios, particularly with partially visible, rotated, or side-view faces. These findings underscore the complexities inherent in analyzing egocentric video data, where movement and varying perspectives introduce

additional challenges.

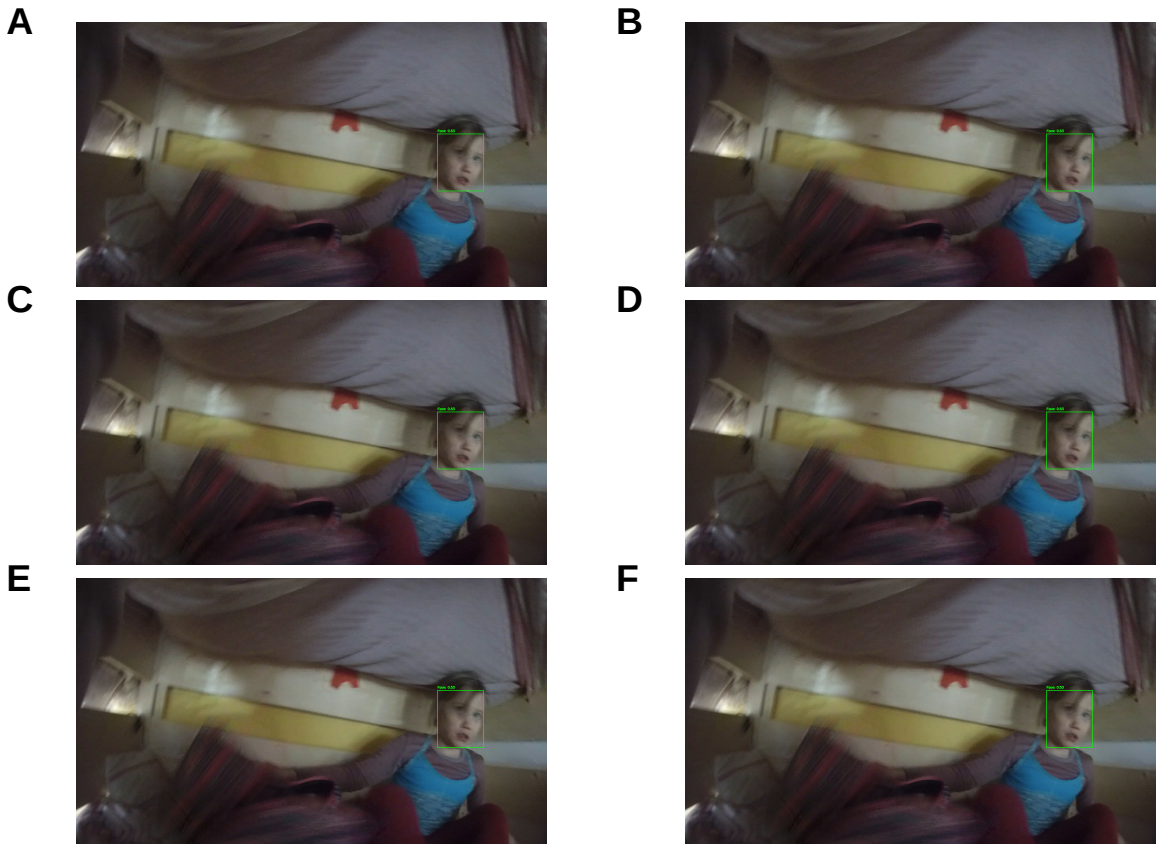

Figure 5. **A**, **B** - Examples of True Positives, **C**, **D** – Examples of False Negatives, **E**, **F** – Examples of False Positives in the YOLO11 face detection model.

Face Detection: Model Selection, Data Preprocessing, and Performance Evaluation

In this study, we utilized Ultralytics' YOLO11 (Jocher & Qiu, 2024), a state-of-the-art object detection model recognized for its efficiency and accuracy. Initially, we experimented with Multi-Task Cascaded Convolutional Networks (MTCNN) face detection algorithm (Zhang et al., 2016); however, it achieved a recall of less than 50% on our dataset. Furthermore, the MTCNN model demands very specific picture input for fine-tuning, which is time-consuming. In contrast, YOLO11, released in October 2024,

introduces new features such as the C2PSA block, which enhances spatial attention within feature maps, allowing the model to focus more precisely on critical areas of an image. Additionally, YOLO11 incorporates the C3K2 block, designed to be faster and more efficient, enhancing the overall performance of the feature aggregation process (Khanam & Hussain, 2024). Moreover, we already had a data preparation pipeline established for YOLO11 due to its application in person detection within our project. Consequently, we selected YOLO11 for face detection and fine-tuned it on our egocentric dataset, captured using chest-mounted cameras, to adapt it to the unique characteristics of our data.

Dataset Splitting and Balancing. We started data preprocessing with a dataset comprising a total of 91706 images from 64 annotated videos.Prior to partitioning this dataset into training, validation, and testing subsets, we analyzed the proportion of images containing annotated faces versus those without. Our analysis revealed that 19% of the images included at least one annotated face. To maintain this inherent distribution across all subsets, we initially applied a stratified split, ensuring that each subset—training, validation, and testing—preserved the original 19% to 81% ratio of images with faces to images without faces.

However, this stratified split resulted in a significant class imbalance within the training set, which could adversely affect the model's learning process. In imbalanced datasets, models tend to be biased toward the majority class, often predicting it more frequently while misclassifying or overlooking minority class instances. This can lead to poor recall for the minority class (Hasanin et al., 2019). Additionally, Yolo11's gradient-based learning algorithm struggles to adjust decision boundaries effectively when trained on imbalanced data, potentially causing slow or unstable convergence and requiring extensive hyperparameter tuning (Kaur et al., 2020).

To mitigate this issue, we employed an undersampling technique on the training data. Specifically, we identified the number of images containing faces in the training set (13708 frames) and randomly sampled an equal number of images from the non-face category.

This approach balanced the training dataset to consist of 50% images with faces and 50% images without faces, thereby addressing the class imbalance and facilitating more effective model training.

Consequently, the final data distribution was as follows: the balanced training dataset comprised 27416 frames, while the validation and test datasets contained 9169 and 9173 frames, respectively. Notably, the validation and test sets retained the original 19% face presence, ensuring that the model's performance evaluation remained representative of the real-world data distribution.

Training and Convergence. Model training was conducted on a Linux server equipped with an Intel(R) Xeon(R) Silver 4214Y CPU @ 2.20GHz with 48 cores, a Quadro RTX 8000 GPU and 188 GB of RAM. The model was trained for a total of 86 epochs. The training process utilized YOLO11's built-in data augmentation, a batch size of 16, a cosine annealing learning rate scheduler, and early stopping after 10 epochs without improvement, with a maximum of 200 epochs.

The loss function of the YOLOv11 model comprises three main components: Box Loss, Classification Loss, and Distribution Focal Loss (DFL). Box Loss estimates the difference between predicted bounding boxes and ground truth boxes to assess the model's localization accuracy. Classification Loss measures the model's ability to properly identify detected objects; however, in our study, this component is less relevant due to our focus on a single class—faces. DFL improves the model's ability to detect challenging objects by prioritizing difficult-to-detect instances.

During the training process, we observed that all three loss components decreased over time, indicating effective learning and improved performance, as visible in in figure 6. A steady decrease in Box Loss indicates that the model is becoming increasingly accurate in localizing faces within images. This is consistent with the rapid convergence of Classification Loss, revealing the model's ability to reliably recognize faces. The decrease in

DFL over time indicates that the model is getting better at focusing on and correctly identifying difficult-to-detect faces, which improves its overall detection capabilities. These data collectively represent the model's gradual improvement in both localization and identification tasks during the training period.

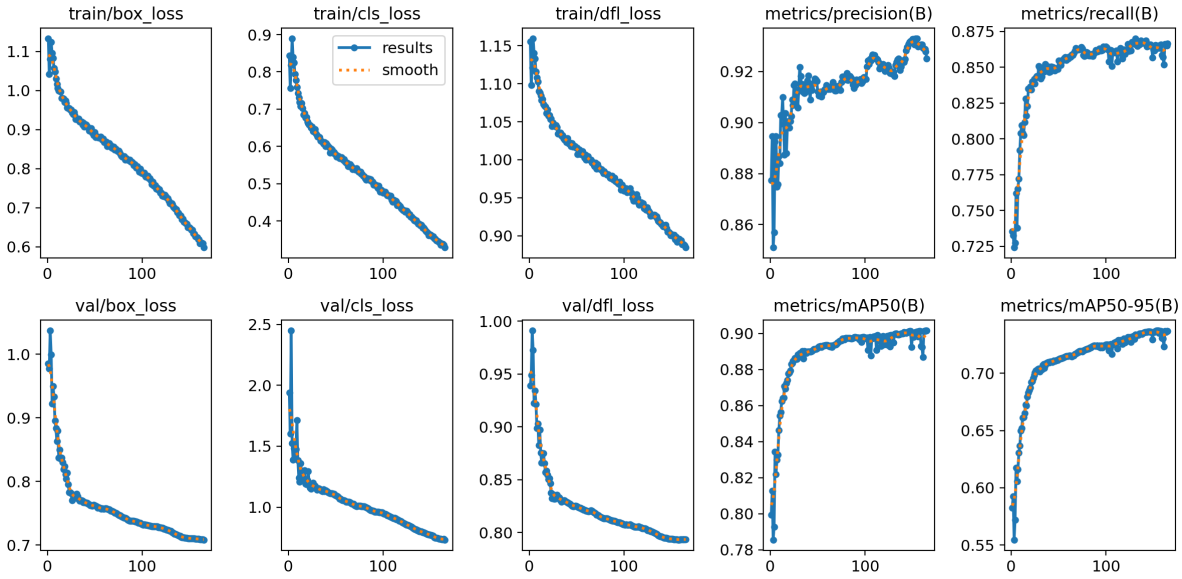

Figure 6. Training and Validation Loss Curves for the YOLO11 face detection model.

Model Evaluation Metrics. The YOLO11 model achieved a precision of 0.92 and a recall of 0.87 on the testing set, resulting in an F1-score of 0.90. These metrics, summarized in figure **??** indicate the model's strong performance in accurately identifying faces while minimizing errors. The precision-recall curve, displayed in figure 7, further illustrates this performance, with the curve remaining close to the top-left corner. This positioning signifies that the model maintains high precision and recall across various thresholds, underscoring its effectiveness in detecting faces with confidence.

Analysis of the confusion matrix reveals that 86% of all faces are correctly identified by the model, corresponding to 1905 true positives, while 234 faces were missed (false negatives). False negatives predominantly occurred in scenarios where faces were in the background, blurred due to motion, or occluded by the child's body. In such instances,

Table 3

Evaluation metrics for the YOLO11 face detection model trained on the Quantex dataset. False Positive Rate and False Negative Rate are given in percentages.

Dataset	Precision	Recall	F1-Score	False Positive Rate	False Negative Rate
Quantex	0.90	0.83	0.86	2.1	14.0

A **B**

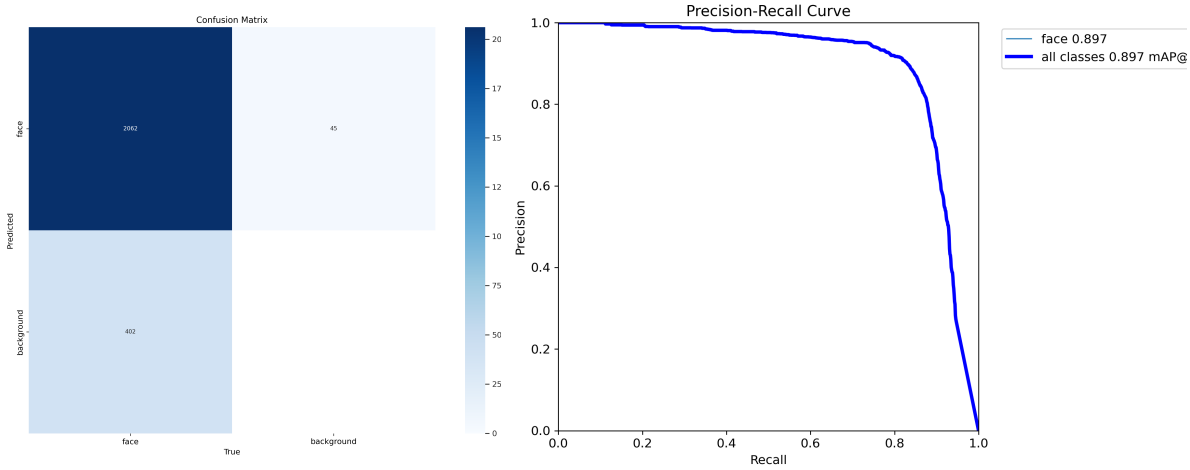

Figure 7. **A** - Confusion Matrix for the YOLO11 face detection model trained on the Quantex dataset. **B** - Precision-Recall Curve for the YOLO11 face detection model.

adjacent frames often provided clearer views, aiding in more accurate classification. The model exhibited a false positive rate of approximately 3.57%, with 251 images incorrectly classified as containing faces when none were present. These false positives were often attributed to objects or toys resembling facial features. Given the relatively low false positive rate, this small number should not raise significant concerns. In face detection systems, a balance between false positives and false negatives is often necessary, and a 2.1% false positive rate is generally considered acceptable.

To provide a comprehensive understanding of the model's performance, we have included visual examples of true positives, false positives, and false negatives in figure 8. These images highlight the model's strengths and areas where challenges persist, offering insights into specific scenarios that influence detection accuracy.

Overall, the YOLO11 model demonstrates robust performance in face detection tasks. However, challenges remain in dynamic scenarios, particularly with partially visible, rotated, or side-view faces. These findings underscore the complexities inherent in analyzing egocentric video data, where movement and varying perspectives introduce additional challenges.

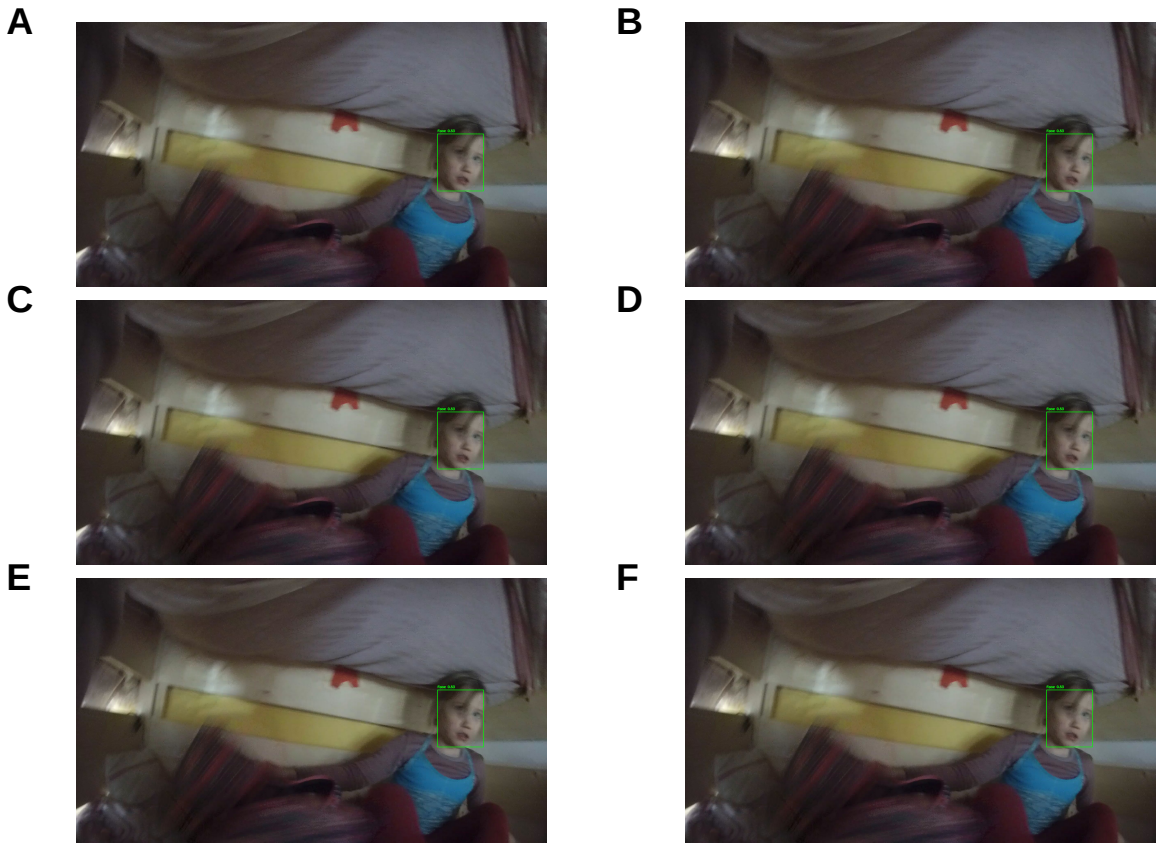

Figure 8. **A**, **B** - Examples of True Positives, **C**, **D** – Examples of False Negatives, **E**, **F** – Examples of False Positives in the YOLO11 face detection model.

References

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2019). Severely imbalanced Big Data challenges: Investigating data sampling approaches. *Journal of Big Data*, *6*(1), 107. https://doi.org/10.1186/s40537-019-0274-4

Jocher, G., & Qiu, J. (2024). *Ultralytics YOLO11*. Retrieved from https://github.com/ultralytics/ultralytics

Kaur, H., Pannu, H. S., & Malhi, A. K. (2020). A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Computing Surveys*, *52*(4), 1–36. https://doi.org/10.1145/3343440

Khanam, R., & Hussain, M. (2024, October 23). YOLOv11: An Overview of the Key

Architectural Enhancements. https://doi.org/10.48550/arXiv.2410.17725

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

Appendix