ChildPlay: A New Benchmark for Understanding Children's Gaze Behaviour Supplementary Material

Samy Tafasca* Anshul Gupta* Jean-Marc Odobez
Idiap Research Institute, Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne, Switzerland
{stafasca, agupta, odobez}@idiap.ch

1. Supplementary

1.1. More information on ChildPlay

Gaze Classes. ChildPlay is annotated with 7 non-overlapping gaze classes to enable high quality gaze annotations. These are defined as follows:

- inside-frame: when the gaze target is located within the frame and is visible;
- outside-frame: the gaze target is outside the frame;
- gaze-shift: when the person shifts attention from one location to the next during at least two frames. In case of interest, shorter shifts (i.e. saccades) can be recovered by identifying sudden changes in gaze points that are annotated as inside-frame;
- occluded: the 2D gaze target is within the frame but is totally occluded (hence cannot be annotated);
- uncertain: the gaze target cannot be determined confidently (lack of salient elements in the gaze direction, several possible targets);
- eyes-closed: used in rare cases where a child closes their eyes (e.g. during hide-and-seek);
- not-annotated: none of the options above is applicable.

Semantics. We compare the semantics of the gaze targets for ChildPlay and VideoAttentionTarget in Table 1. Our ChildPlay dataset is far more balanced[1], while also having 50% more frames and twice as much scene variety.

Dataset	things-person	things-other	stuff	not-detected
VideoAttention [4]	80.85%	8.05%	3.60%	7.50%
ChildPlay	45.19%	18.66%	12.62%	23.53%

Table 1. Comparison of gaze target semantic class between ChildPlay and VideoAttentionTarget. Numbers were obtained by running a panoptic segmentation model [3] on images and retrieving the semantic class of each annotated gaze point.

1.2. More Children Datasets

One of the major motivations behind building datasets of children is the study of neurodevelopmental disorders exhibiting symptoms in humans from an early age. For this reason, many benchmarks studied in the literature cover topics such as motor control, brain imaging, emotions, speech, and social interactions. Nevertheless, most of them are ultimately never shared due to privacy considerations and ethics regulations [5]. We previously listed some of the children datasets directly related to autism behaviors, in this section, we cover a few publicly available ones that feature pose annotations. Since the body proportions of humans change significantly from birth to adulthood [16], it is important for younger age groups to be well represented in research benchmarks, particularly for applications targeting them. Table 2 summarizes the notable ones.

1.3. Point Cloud Comparison

Monocular Depth Estimation. Depth datasets can be put under three categories:

- Absolute Depth: These datasets provide the absolute depth of the scene. The data is recorded using sensors such as LiDARS, time of flight cameras etc. ex. KITTI [7]

* indicates equal contribution

[1] After manual inspection, we found that most of the `not-detected` instances in ChildPlay correspond to objects that were not detected by the segmentation, and which would fall into the `things-other` category.

Name	Type	Setting	Size	Annotations
Sciortino et al. [16]	Video	SSBD dataset + youtube	1176 images of 104 subjects	2D pose keypoints
DREAM [2]	Video	Interactions with robot No raw data, only extracted features and annotations	306 hours of therapy (102) subjects	3D pose keypoints
BabyPose [13]	Video Depth	Preterm Infant movement in NICUs	16000 frames · 16 depth videos · 16 patients	2D pose keypoints
SyRIP [10]	Image	Hybrid: real + synthetic YouTube and Google images	Real: 700 images (140+ subjects) Synthetic: 1000 images	2D pose keypoints
MINI-RGBD [9]	Video Depth	Synthetic: obtained by registering SMIL to real sequences of moving infants. Constrained environment	12000 frames · 12 sequences	2D and 3D keypoints

Table 2. Summary of selected pose estimation children datasets.

- Up to Scale (UTS) Depth: These datasets provide the depth of the scene up to an unknown scale C_1. The absolute depth $d*$ can be recovered from UTS depth d as $d*^{-1} = C_1.d^{-1}$. ex. Megadepth [12]
- Up to Shift and Scale (UTSS) Depth: These datasets provide the disparity of scene. They are obtained from stereo movies and photos by computing the optical flow. The absolute depth can be recovered from the disparity D as $d*^{-1} = C_1.(D + C_2)$. C_2, also known as shift, depends on the camera parameters and is crucial for reconstructing geometry preserving point clouds. However, the shift is typically unknown. ex. MiDaS [15]

Recent methods for monocular depth estimation [15][17] have leveraged UTSS depth data due to it's high diversity, and shown better generalization when tested on unseen datasets. However, they can only predict UTSS depth so the reconstructed point clouds are not geometry preserving. Hence, methods for gaze target prediction that use these algorithms rely on course matching [6] or attempt to correct the point cloud based on prior assumptions [1].

We study two recent methods for monocular depth estimation that aim to generate geometry-preserving point clouds while still leveraging UTSS data. Wei et al. [18] predict UTSS depth and use it to construct a (distorted) point cloud. A point cloud module then recovers the shift factor from the distorted point cloud. On the other hand, Patakin et al. [14] train on a mix of absolute, UTS and UTSS depth data. The absolute and UTS depth data provide supervision such that the algorithm predicts UTS depth.

Qualitative Results. We provide a qualitative comparison of point clouds generated using the depth maps from Ranftl et al. [15], Wei et al [18] and Patakin et al. [14] in Figure 1. We observe that the point clouds generated using the depth maps from Wei et al. and Patakin et al. generally have less distortion of scene elements, and better maintain the depth between objects. The point clouds from Patakin et al. in particular seem to preserve the geometry of the scene best.

Gaze Vector Stability. To quantitatively compare the methods of Wei et al. [18] and Patakin et al. [14], we investigate which algorithm generates more stable gaze vectors. This is crucial as we rely on their generated gaze vectors as ground truth. The test is based on the fact that the gaze vector for a person (camera coordinate system) should be the same irrespective of their distance from the camera. We perform the test as follows:

- We take 5 random crops of an image
- For each crop, we compute the depth (Wei et al. or Patakin et al.) and focal length
- We then reconstruct the point cloud $\mathbf{P^c}$ following the protocol defined in Section 4.2, and obtain the gaze vector for each crop as $\mathbf{g_{gt}}^{\mathbf{c}} = \frac{\mathbf{P}^{\mathbf{c}}_{gaze} - \mathbf{P}^{\mathbf{c}}_{eye}}{||\mathbf{P}^{\mathbf{c}}_{gaze} - \mathbf{P}^{\mathbf{c}}_{eye}||}$
- The stability is given by the standard deviation of the gaze vector across the crops

For a more robust estimate, we perform this procedure for the first frame of every clip in the ChildPlay training set, and compute the median standard deviation. The values for the method of Wei et al. are [0.041, 0.032, 0.095] while the values for the method of Patakin et al. are [0.026, 0.019, 0.075]. The median standard deviation for Patakin et al. is lower, especially for the z component, indicating that it generates more stable gaze vectors.

Figure 1. Comparison of point clouds generated using the depth maps from Ranftl et al. [15] (row 2), Wei et al. [18] (row 3) and Patakin et al. [14] (row 4) on ChildPlay images. The point clouds generated using Patakin et al. appear to best preserve the geometry of the scene.

1.4. Training Details

Head Bounding Boxes. The provided head box annotations for GazeFollow are not consistent and sometimes include the whole head, and at other times just the face of the person. Hence, we re-extract the head boxes using a pre-trained Yolov5 model [11] and use these for all our experiments.

Eye Location. For GazeFollow, we use the annotated eye location, and for the VideoAttentionTarget and ChildPlay datasets we use the center of the annotated head bounding box as the eye location.

Input Aspect Ratio. Previous methods [4][8] distort the scene and head images to the model input size. To avoid this, we expand the head bounding box to a square to match the Human-Centric module's input aspect ratio. We also carefully crop and pad scene images to the Scene-Centric module's input aspect ratio during training and validation so that there is no distortion. During the test phase, we do not perform any cropping/padding and instead scale the longer side of the scene image to the Scene-Centric module's input width.

References

[1] Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14126–14135, 2022. 2

[2] Erik Billing, Tony Belpaeme, Haibin Cai, Hoang-Long Cao, Anamaria Ciocan, Cristina Costescu, Daniel David, Robert Homewood, Daniel Hernandez Garcia, Pablo Gómez Esteban, et al. The dream dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy. *PloS one*, 15(8):e0236939, 2020. 2

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 1

[4] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020. 1, 3

[5] Ryan Anthony J de Belen, Tomasz Bednarz, Arcot Sowmya, and Dennis Del Favero. Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational psychiatry*, 10(1):1–20, 2020. 1

[6] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11390–11399, June 2021. 2

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[8] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction:

Application to privacy-sensitive settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5041–5050, 2022. 3

[9] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Raphael Weinberger, and A Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[10] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadabbas. Invariant representation learning for infant pose estimation with small data. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 2

[11] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Kalen Michael, Jiacong Fang, imyhxy, Lorna, Colin Wong, (Zeng Yifu), Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Max Strobel, Mrinal Jain, Lorenzo Mammana, and xylieong. ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations, Aug. 2022. 3

[12] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2

[13] Lucia Migliorelli, Sara Moccia, Rocco Pietrini, Virgilio Paolo Carnielli, and Emanuele Frontoni. The babypose dataset. *Data in brief*, 33:106329, 2020. 2

[14] Nikolay Patakin, Anna Vorontsova, Mikhail Artemyev, and Anton Konushin. Single-stage 3d geometry-preserving depth estimation model training on dataset mixtures with uncalibrated stereo data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1705–1714, 2022. 2, 3

[15] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2, 3

[16] Giuseppa Sciortino, Giovanni Maria Farinella, Sebastiano Battiato, Marco Leo, and Cosimo Distante. On the estimation of children's poses. In *International conference on image analysis and processing*, pages 410–421. Springer, 2017. 1, 2

[17] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 2

[18] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3