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

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

Abstract

Gaze behaviors such as eye-contact or shared attention are important markers for diagnosing developmental disorders in children. While previous studies have looked at some of these elements, the analysis is usually performed on private datasets and is restricted to lab settings. Furthermore, all publicly available gaze target prediction benchmarks mostly contain instances of adults, which makes models trained on them less applicable to scenarios with young children. In this paper, we propose the first study for predicting the gaze target of children and interacting adults. To this end, we introduce the ChildPlay dataset: a curated collection of short video clips featuring children playing and interacting with adults in uncontrolled environments (e.g. kindergarten, therapy centers, preschools etc.), which we annotate with rich gaze information. We further propose a new model for gaze target prediction that is geometrically grounded by explicitly identifying the scene parts in the 3D field of view (3DFoV) of the person, leveraging recent geometry preserving depth inference methods. Our model achieves state of the art results on benchmark datasets and ChildPlay. Furthermore, results show that looking at faces prediction performance on children is much worse than on adults, and can be significantly improved by fine-tuning models using child gaze annotations. Our dataset is available at https://www.idiap.ch/en/dataset/childplay-gaze*. Code will be made available soon.*

1. Introduction

Gaze is a non-verbal cue that provides rich information about people. In particular, it plays fundamental roles in social interactions and human communication, like initiating interaction, showing attention, monitoring the floor, or regulating intimacy, and as such finds many applications in human interaction analysis. Hence, gaze extraction and analysis finds applications in human-human or human-robot interaction [60, 46], including psychological studies [44] and medical diagnoses.

* indicates equal contribution

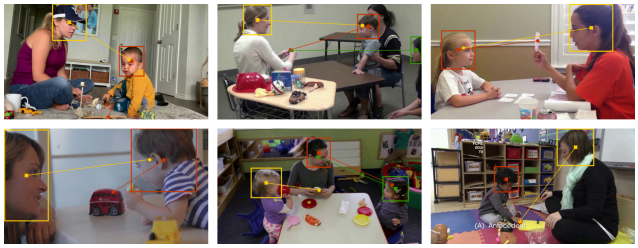

Figure 1. Sample images from the ChildPlay dataset with head bounding box and gaze point annotations. Such scenes strongly depart from existing gaze benchmarks (*e.g.* standing adults).

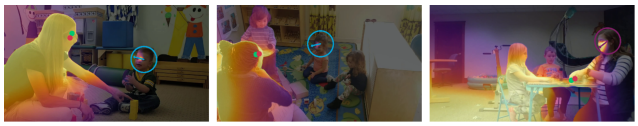

Figure 2. Qualitative results of our geometrically grounded model on ChildPlay. Our 3D Field of View (3DFoV) highlights potential gaze targets, excluding objects where the depth does not match. Gaze target predictions are given in green and GT ones in red.

In this regard, acquiring appropriate social gaze behavior skills is important in the cognitive development of children. For instance, gaze following has been shown to help children with language acquisition [5, 6], while gaze aversion can help children perform better in cognitively demanding tasks [48]. Furthermore, abnormal gaze patterns are known to correlate with several neuro-developmental disorders like Autism Spectrum Disorder (ASD) [59, 43]. This has led to the development of specific gold standard markers for the screening of these impairments, e.g. by measuring deficits in the initiation of joint attention or shared attention [13].

ChildPlay dataset. Due to this importance, several methods have been proposed to analyze children's gaze, especially in the context of autism [54, 8, 24, 56, 1, 61, 62, 9, 35]. However, they are all tested on private datasets [12] due to the sensitive nature of the data, hindering proper comparison across algorithms. Some of them may be accessible in anonymized formats (e.g. pose skeletons), but this makes them difficult to use for the study of gaze behavior. While other datasets have gaze related labels involving

autistic children [4, 53], they are usually recorded in a fixed lab setting and with coarse annotations. Alternatively, we could use standard public benchmarks like [11] for learning gaze prediction models, but children and the physical situations and task performed (seating on the floor, playing with objects) would be severely underrepresented. It has been shown that training models on datasets with mainly adults can lead to a significant drop in performance when tested on children, e.g. for body landmark prediction [58]. Given the importance of pose information in gaze prediction [23, 3] and the difference in gaze behavior between adults and children [17], there is a need for gaze annotated datasets featuring lower age groups in general settings.

To address this problem, we introduce the ChildPlay dataset: a set of videos featuring children in free-play environments interacting with their surrounding. The dataset is rich in unprompted social behaviors, communicative gestures and interactions and features high quality dense gaze annotations, including a gaze class to account for special scenarios that arise in 2D gaze following. Further, the 2D gaze information can be used to model other attention-related behaviors like shared attention, gaze shifts, eye contact and fixation points with minimal processing. To the best of our knowledge, we are the first to establish a more representative gaze dataset aiming to cover children.

3D Field of View (3DFoV) for gaze target prediction. Recasens et al. [52] introduced this task, also called gaze following. In contrast with previous gaze objectives which were mainly attempting at inferring the 3D gaze direction from head and eye images [18, 39], it aims to predict the image 2D gaze location of a person in the image for arbitrary and general scenes. Since then, many works have embraced this paradigm [37, 10, 69, 21, 45, 30], proposing new models exploiting temporal information [11], or exploiting further cues like depth [16, 2, 23]. Indeed, inferring and understanding depth is crucial, as it provides information about the scene structure enabling geometric reasoning and ruling out salient objects or people which fall along the 2D line of sight of a person, but are actually not visible to the person in the 3D space. In this context, as most datasets do not have depth information, some methods opted for pre-trained monocular depth (or disparity) estimators to extract scene depth cues [16, 2, 23]. However, such algorithms [50, 67] typically estimate the depth up to an unknown shift and scale factors which often result in stretched and distorted scenes unsuitable for proper 3D analysis.

In this paper we provide a more geometrically grounded approach leveraging a new algorithm [47] addressing these points, correcting shifts, and yielding geometry-preserving depth maps that can be used to derive a proper scene point cloud, and explicitly match the predicted 3D gaze vector with this point cloud to derive the 3DFoV of the person. In experiments, we show that this method generalizes well, providing better cross-dataset performance.

New gaze metric: looking at heads precision (P.Head). Standard performance metric for gaze following either have no physical interpretation (the Area Under Curve, AUC), or may not provide rich enough information about performance, as is the case of 2D distance metrics (how far is a 2D gaze prediction from the GT). In practice, one is more interested at semantics, e.g. how accurate is a model at predicting the category (person, body part, object) of image regions being looked at. As objects might be hard to annotate at scale, in this paper, leveraging highly accurate head detectors and since heads are one of the most important gaze category in many applications (ex. child looking at clinician for ASD diagnosis [13]), we propose to exploit looking at head precision (P.Head) metric for performance evaluation. We show that this measure greatly varies across datasets and can have rather low performance, that the distance and P.Head metrics may disagree and performance on children is rather different than on adults. On ChildPlay, while children exhibit a better distance performance, their P.Head metric is much worse.

Contributions. Our main contributions are:

- We introduce the ChildPlay dataset, a curated collection of clips recording children playing and interacting with adults in uncontrolled environments, annotated with rich gaze information;
- We propose a new model for gaze target prediction that relies on the explicit modeling of the 3DFoV by exploiting geometrically consistent inferred depth maps.
- We propose to use the Looking At Head Precision metric to characterize performance.

Extensive experiments on the GazeFollow, VideoAttentionTarget and ChildPlay datasets demonstrate that our approach produces the best or state-of-the-art results, motivating further studies on the topic. The dataset and models are publicly available.

2. Related Work

We discuss works related to gaze target prediction and highlight the methods that use depth information. We discuss datasets on gaze and children in Section 3.4.

Gaze Target Prediction. Traditional methods for gaze following rely on a 2-branch architecture consisting of a scene branch to identify salient regions in the image and a human-centric branch to infer the general gaze direction of the target person [51, 11, 37, 30, 16, 23]. Various ideas have since been proposed in the literature to boost this typical architecture, namely, inferring a 2D gaze cone [37, 23], using multimodal information [16, 23, 22], leveraging the temporal context [11], or improving computational efficiency for multi-person scenarios [65, 30]. There are also other related tasks that incorporate semantic information such as

detecting eye-contact [41, 3], inferring the gaze target object [64, 66], or shared attention behavior [14] to cite a few.

Gaze Target Prediction using Depth. Fang et al. [16] used a pre-trained monocular depth estimator to extract the scene depth. They split this map into three depth-based saliency maps depending on the depth of the target person, and used a pre-trained gaze estimation model to select the corresponding one for a coarse matching. Jin et al. [31] attach auxiliary branches during training to predict scene disparity and predict a 3D orientation vector. However, these are not used as input to the model and depth is implicitly encoded in the features. Hu et al [28] follow a similar strategy as ours but match a coarse predicted 3D gaze vector with the derived scene point cloud. Further, they mainly target the use of RGB-D images, and their derived point cloud when dealing with RGB images is not geometrically consistent due to their pre-trained depth estimator [67]. Bao et al. [2] also derive a scene point cloud and attempt to correct it by using humans as reference objects. However, they do not predict a 3D gaze vector and hence do not perform any explicit matching of predicted gaze and depth. Gupta et al [23] treat the depth map as an input and hence do not perform any matching of 3D gaze and depth.

3. ChildPlay Dataset

In the following, we describe several aspects of the ChildPlay dataset i.e. data collection strategy, annotation protocol, statistics and comparison to benchmarks.

3.1. Data collection strategy

Data selection. We relied on the YouTube video search engine with queries like "children playing toys", "childcare center", or "kids observation" to retrieve videos matching our aim [1]. To foster quality we only looked for videos with an aspect ratio of 16:9 and a resolution of 720p or 1080p. We downloaded the audio files as well, although in many cases they are not produced by the scene (e.g. commentary).

Clip selection. The scene context of our videos ranged from childcare facilities and schools to homes and therapy centers. As full videos contain many irrelevant parts, we selected clips with children and featuring interesting gaze movements and social interactions. We also made sure that clips contain no scene cuts, blurriness, large overlaid graphical items, heavy zooming or camera movement. Finally, to foster diversity, we limited the duration and number of clips taken from each video or Youtube channel.

Content. We obtain a dataset of 401 clips, mainly restricted to indoor environments, showing at least 1 child, but oftentimes include 1 or 2 adults and multiple children. The age group of the children varies from toddlers to pre-teenagers. The dominant activity of children is "playing with toys", but the dataset also includes a few clips containing other interactions such as behavioral therapy exercises.

[1] Around 10% of our data have the CC BY license, whereas the other have no license, i.e. they follow the default YouTube license.

3.2. Annotation protocol

We performed a dense annotation of gaze information[2]. In every clip we selected up to 3 people and for each of them, in every frame we annotated the head bounding box, a 2D gaze point, and a gaze label. We also provide the class label for adult vs child. Two main points were taken into account to ensure high quality and confidence in the 2D annotations: enforcing semantically consistent 2D gaze annotations, i.e. the annotated 2D location has to be on the object being looked at (cf Figure 3) in anticipation of a transition to semantically aware gaze evaluation metrics (as motivated by the P.Head metric), and the introduction of a gaze label.

The gaze label addresses an important limitation with existing datasets, in which annotating a 2D gaze point is enforced in every frame, with only a standard inside vs outside label to denote if the person looks within the frame or not. However, there are many situations where annotating 2D gaze points is highly challenging, if not impossible. For example, when a person shifts attention from one location to another, in the VideoAttentionTarget dataset we often observed that intermediate frames during the shift were annotated using the outside class which is inconsistent with the definition. To avoid this, our gaze label was defined to include 7 non-overlapping classes to properly account for special scenarios (precise definition in appendix):

- inside-frame;
- outside-frame;
- gaze-shift;
- occluded;
- eyes-closed;
- uncertain, and
- not-annotated.

In practice, inside-frame (85.3%) and outside-frame (5.4%) are the dominant classes, but all other ones where a confident annotation can not be made still represent 9.3% of the frames.

Finally, to evaluate the inter-annotator agreement in terms of 2D target localization we followed the usual practice. We had 2000 instances being double coded, and evaluated the performance of a human (as prediction) against the other one (used as GT). Evaluation is reported in Table 3 (under the "Human" baseline), demonstrating similar agreement as on other datasets (see 4).

[2] The collection and annotation of the dataset has been approved by our Data and Research and Ethics Committee. According to the national law, downloading Youtube data for training and annotation is allowed for research purposes. For distribution, similarly to other datasets [20], we will share the video links, scripts to extract frames, and corresponding annotations.

Figure 3. Example of two annotations close in L_2 distance but very different semantically (i.e. on different objects) and distant depth-wise. Green is correct, red is incorrect.

3.3. Statistics

Table 1 provides a summary of the main statistics of our ChildPlay dataset. It contains 401 short clips averaging 10s, extracted from 95 videos originating from 44 different YouTube channels. In total, there are around $258k$ annotation instances (i.e. 62% for children) distributed across $120,549$ frames and annotated by 7 people using the LabelBox platform [34]. Figure 1 shows sample images along with the corresponding annotations.

Figure 4 summarizes the distribution of various geometric quantities, highlighting major differences between children and adults. The distribution of head sizes covers a fairly wide range, and reflecting that people are located at various distances from the camera. Moreover, we can notice that adults mostly look down to observe children and their activities, whereas children mostly look down at their toys with few instances where they raise their heads to gaze back at the adults, an important behavior difference which is also corroborated by the high difference in probability of looking at faces (see Tab. 2). We can also observe that the distance from the head to the gaze point is typically between 10% and 60% of image side, with again children more focused on close targets than adults.

3.4. Comparison to other datasets

Gaze Datasets. Table 1 presents an overview of existing gaze prediction datasets that are publicly available. The two most related datasets are GazeFollow and VideoAttentionTarget. GazeFollow [51] is a large-scale image dataset featuring $130K$ independent instances, but it suffers from low resolution, average annotation quality and lack of temporal context. Nevertheless, given its rich diversity, it remains a good dataset to use for pre-training. VideoAttentionTarget [11] is a recent video dataset built from high resolution clips taken from popular TV shows. Since it was extracted from 50 shows, the diversity of the scenes remains limited. In this view, ChildPlay definitively offers complementary content to VAT: much more children and a large diversity of pose, behaviors and situations (looking up and down at children or adults) as opposed to dominantly standing, sitting, and talking people, and a very strong bias towards looking at people's faces, as shown in Table 2.

Our ChildPlay dataset is far more balanced. while also having 50% more frames and twice as much scene variety.

Other datasets differ significantly from ChildPlay in their scope (beyond having very little children). Several of them address attention related tasks (Co-attention [14], looking-at-each-other (LAEO) [41]), specific settings like retail [64], or are much smaller in size and diversity hence can mainly be used for evaluation, not training [37].

Children Datasets. The Multimodal Dyadic Behavior (MMDB) dataset [53], the Self-Stimulatory Behaviors dataset (SSBD) [49], DREAM [4] and 3D-AD [55] are all datasets meant to tackle different aspects of autism, be it stimming behaviors (arm flapping, head banging), speech and vocalizations, communicative gestures (e.g. pointing, reaching, etc.) or gaze patterns (e.g. shared attention, eye-contact). However, they are either anonymized, limited in terms of behaviors or restricted to lab environments (e.g. screening or therapy sessions). In contrast, ChildPlay boasts a higher diversity of scenes, people, gestures, viewing angles and lighting conditions. More information about other children datasets can be found in appendix.

4. Model Architecture

4.1. Approach overview

Our network architecture is illustrated in Figure 5. Similar to other methods [11, 16, 23], our architecture comprises two main pathways. On one hand, the gaze pathway (GP) aims at predicting the scene elements which are in the 3D Field of View (3DFoV) of the person, represented by the heatmap $\mathbf{V}$. To do so, it takes as input the image crop $\mathbf{I}_h$ of the person's head and predicts its 3D gaze direction $\mathbf{g_p}$ as well as a gaze embedding $\mathbf{e_g}$. The gaze $\mathbf{g_p}$ is then combined with the scene consistent 3D point cloud $\mathbf{P}$ inferred from the image to generate the 3DFoV heatmap.

On the other hand, the scene pathway combines the image with the GP information (the location of the head represented by the head mask $\mathbf{I}_m$, $\mathbf{V}$ and the gaze embedding $\mathbf{e_g}$) to infer the in-out label $\mathbf{o}_p$ (i.e. looking inside the frame or outside) and the attention heatmap $\mathbf{A}_p$. We details these elements below. However, given the importance of the scene structure representation, we first describe how the scene point cloud is obtained.

4.2. Point cloud generation

To obtain our point cloud in the camera coordinate system $\mathbf{P^c} = \{\mathbf{P}_i^\mathbf{c} = (X_i^\mathbf{c}, Y_i^\mathbf{c}, Z_i^\mathbf{c})\}$ associated to the 2D pixels defined in the image plane $\mathbf{p}_i = (x_i, y_i)$, we need to know the scene depth as well as the intrinsic parameters of the camera. As these are not available, we need to infer them from the data and make assumptions.

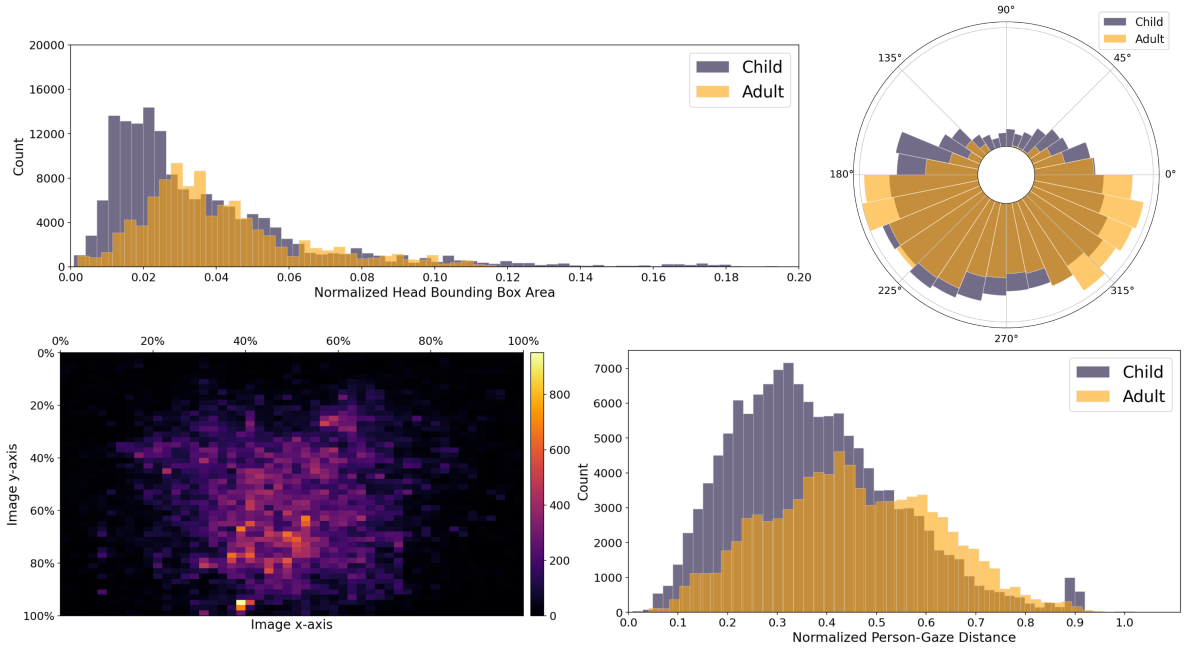

Figure 4. Geometric statistics of the ChildPlay annotations. [Top-Left] Distribution of head bounding box area normalized w.r.t image size. [Top-Right] Distribution of gaze angles in the image frame. [Bottom-Left] 2D histogram of gaze points. [Bottom-Right] Distribution of the distance between each person (i.e. center of the head) and their gaze point, normalized w.r.t the image sides.

Name	Type	Shows	Frames	Instances	Origin	Annotations
GazeFollow [51]	image	-	122, 143	130, 339	SUN, MS-COCO, ImageNet, ...	eye location · 2D gaze point · inside/outside
VideoAttentionTarget [11]	video	50 (606)	71, 666	164, 541	TV shows	gaze point · inside/outside
VideoCoAtt [14]	video	20 (400)	493, 242	138, 203	TV shows	shared gaze object bbox
DL Gaze [37]	video	4 (86)	95, 000	6, 348	Manual collection	gaze point
UCO-LAEO [41]	video	4 (129)	18, 000	36, 358	TV shows	LAEO class
AVA-LAEO [41]	video	298	$1.4M$	172, 330	Movies	LAEO class
VACATION [15]	video	50	96, 993	164, 365	TV Shows	gaze object bbox · gaze communication label
GOO [64]	image	-	201, 552	172, 330	Manual collection Synthetic	gaze point · gaze object · object bbox
ChildPlay	video	95 (401)	120, 549	257, 928	YouTube	gaze point · gaze class

Table 1. Summary of gaze estimation datasets. All datasets provide head bounding boxes (or pairs of them for LAEO).

Regarding depth, we leverage the pre-trained model of [47] to predict the depth $Z_i^{\mathbf{c}}$ of each pixel. As explained earlier, we chose this model as it generates geometrically consistent depth maps which are crucial for doing a proper 3D analysis of the scene. As to camera parameters, we make standard assumptions: square pixels, no skew, and the principal point at the image center. The more important parameter is the focal length, which is required to avoid scene stretching. In this paper, we estimate it using the pre-trained model of [68]. As a result, denoting by W and H the image width and height, we obtain the simplified projection equation:

$$\begin{bmatrix} X_i^{\mathbf{c}} \\ Y_i^{\mathbf{c}} \\ Z_i^{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} f & 0 & W/2 \\ 0 & f & H/2 \\ 0 & 0 & 1 \end{bmatrix}^{-1} . \begin{bmatrix} x_i.Z_i^{\mathbf{c}} \\ y_i.Z_i^{\mathbf{c}} \\ Z_i^{\mathbf{c}} \end{bmatrix} \quad (1)$$

enabling us to build our point cloud $\mathbf{P^c}$.

Note that $\mathbf{P^c}$ is defined in the camera coordinate system. However, as our aim is to evaluate the scene elements visible from the person's viewpoint, we transform it in the local eye coordinate system C^{eye} in which the gaze vector is predicted (see next section), resulting in $\mathbf{P^e}$. Following [33], the origin of C^{eye} is defined by the eye location $\mathbf{P}^{\mathbf{c}}_{eye}$, and the basis vectors (E_x, E_y, E_z) are such that E_z is the unit

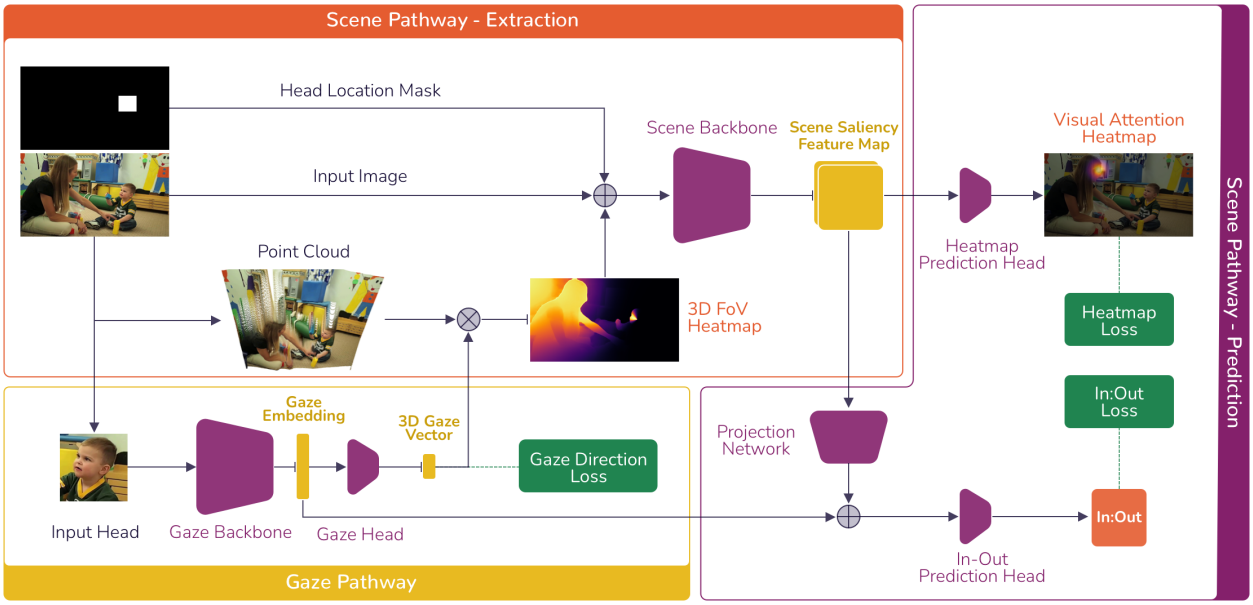

Figure 5. Overview of our proposed architecture. The Gaze Network processes the head crop to predict a 3D gaze vector, which is then used with inferred point cloud to generate a heatmap of the 3D Field-of-view. The Scene Pathway further combines this map with the image and a head location mask to predict a feature map highlighting salient items in the scene. This map is further used to predict on one hand the visual attention map $\mathbf{A}_p$, and on the other hand, with the gaze embedding $\mathbf{e_g}$, the in-vs-out gaze label.

vector from the camera to the eye, and E_x and E_y are in the plane perpendicular to E_z.

$$\mathbf{P^e} = \begin{bmatrix} E_x^{\mathrm{T}} \\ E_y^{\mathrm{T}} \\ E_z^{\mathrm{T}} \end{bmatrix} .(\mathbf{P^c} - \mathbf{P}^{\mathbf{c}}_{eye}) \tag{2}$$

We provide an analysis of the quality of the generated point clouds in the appendix.

4.3. Gaze pathway

This pathway comprises several steps to generate the 3DFoV heatmap $\mathbf{V}$, as described below.

Gaze Prediction Network. Its aim is to predict the gaze direction $\mathbf{g_p}$ defined in the local coordinate system C^{eye} associated with the head crop $\mathbf{I}_h$ [33]. C^{eye} is used rather than the camera coordinate system, as the gaze depends mainly on appearance (head pose and eyes) and not on the head location within the image. This network is composed of a gaze prediction backbone $\mathcal{G}_b$ and of a gaze prediction head $\mathcal{G}_h$. The first one, $\mathcal{G}_b$, is a ResNet-18 [25] network that predicts the gaze embedding $\mathbf{e_g}$ from the head crop $\mathbf{I}_h$, while the second is an MLP with 2 layers followed by tanh activation which transforms this gaze embedding into the unit 3D gaze vector prediction $\mathbf{g_p}$:

$$\mathbf{e_g} = \mathcal{G}_b(\mathbf{I}_h) \text{ and } \mathbf{g_p} = \mathcal{G}_h(\mathbf{e_g}) \tag{3}$$

3DFoV heatmap V generation. Its goal is to highlight the scene parts lying in the gaze direction of the person. To do so, given the point cloud $\mathbf{P^e}$ and the gaze prediction $\mathbf{g_p}$, we simply compute the cosine similarity $\mathbf{c}$ between $\mathbf{g_p}$ and every point $\mathbf{P}^{\mathbf{e}}_i$ in $\mathbf{P^e}$, and further apply an exponential decay function for values with lower similarity to enhance the scene parts which are more in the gaze focus:

$$\mathbf{V}_i = \begin{cases} \mathbf{c}_i, & \text{if } \mathbf{c}_i > 0.9 \\ 0.9 \times \frac{\exp(5 \times \mathbf{c}_i)}{\exp(5 \times 0.9)}, & \text{otherwise.} \end{cases} \tag{4}$$

and $\mathbf{c}_i = \mathbf{g_p} . \frac{\mathbf{P}^{\mathbf{e}}_i}{||\mathbf{P}^{\mathbf{e}}_i||}$. Note that this formulation of the 3DFoV is differentiable, allowing end-to-end training.

4.4. Scene Pathway

The scene pathway combines the scene information (the image $\mathbf{I}$) with the 3DFoV heatmap $\mathbf{V}$ of the person (to which we add the head location mask $\mathbf{I}_m$ to better characterize the location and scale of the person in the scene) and the gaze embedding $\mathbf{e_g}$ to infer his attention (in-out indicator $\mathbf{o}_p$ and visual attention heatmap $\mathbf{A}_p$), according to:

$$\mathbf{F} = \mathcal{F}([\mathbf{I}, \mathbf{V}, \mathbf{I}_m]) \tag{5}$$

$$\mathbf{A}_p = \mathcal{R}(\mathbf{F}) \text{ and } \mathbf{o}_p = \mathcal{O}([\mathbf{e_s}, \mathbf{e_g}]) \text{ with } \mathbf{e_s} = \mathcal{C}(\mathbf{F}) \tag{6}$$

which we explain below.

Saliency feature extraction. The scene backbone network $\mathcal{F}$ is an encoder-decoder architecture producing a set $\mathbf{F}$ of

gaze saliency feature maps. The encoder is an EfficientNet-B1 [63] network while the decoder is a Feature Pyramid Network (FPN) [38]. The FPN contains skip connections that help retain high resolution spatial information which will improve gaze target localization. The concatenation of inputs in Eq. 5 can be considered as early fusion of the scene and gaze information and has been shown to give better performance compared to late fusion schemes, e.g. [23].

Attention prediction. It is summarized in Eq. 6, and comprises two parts. The main one is the attention prediction head network $\mathcal{R}$ which process the feature maps $\mathbf{F}$ to predict the gaze target heatmap $\mathbf{A}_p$, whose maximum gives us the gaze target location. It is a CNN block with 6 layers of dilated convolutions, and a 1x1 conv regression layer.

The second one is the in-out prediction head $\mathcal{O}$ deciding whether the gaze target is within the frame. It is an MLP with 2 layers followed by a sigmoid activation that fuses the gaze embedding $\mathbf{e_g}$ with the scene embedding $\mathbf{e_s}$ to reach a decision. The embedding $\mathbf{e_s}$ is derived from the gaze saliency feature maps $\mathbf{F}$ through the compression network $\mathcal{C}$ (a CNN block with 3 strided convolution layers followed by max pooling).

4.5. Ground truth and loss definition

Heatmap GT. The gaze target location is encoded in a standard way [11] as a GT heatmap $\mathbf{A}_{gt}$ with a 2D isotropic gaussian centered at the annotated gaze target location, and with a standard deviation σ defined in proportion to the heatmap size according to $\sigma = \frac{(W_{hm}+H_{hm})}{2} \cdot \frac{3}{64}$, where (W_{hm}, H_{hm}) are the heatmap dimensions. This results in $\sigma = 3$ pixels for a heatmap of size $(64, 64)$ which corresponds to the value used in other methods [11].

3D Gaze Vector pseudo GT. While other methods only use the 2D information to drive the gaze pathway estimation, we propose to use our geometrically consistent point cloud to define a pseudo 3D gaze direction ground truth. Given the 2D gaze target GT, we obtain the corresponding 3D gaze point $\mathbf{P}^{\mathbf{e}}_{gaze}$ in the point cloud defined in the eye coordinate system C^{eye} (see Section 4.2), and simply derive the 3D gaze unit vector accordingly as $\mathbf{g_{gt}} = \frac{\mathbf{P}^{\mathbf{e}}_{gaze}}{||\mathbf{P}^{\mathbf{e}}_{gaze}||}$

Loss definitions. Learning is driven by three losses:

$$\mathcal{L} = \lambda_{hm}\mathcal{L}_{hm} + \lambda_{dir}\mathcal{L}_{dir} + \lambda_{io}\mathcal{L}_{io} \quad (7)$$

The first loss is the visual attention heatmap loss, defined as in other works as the L2 loss between the predicted and GT heatmaps: $\mathcal{L}_{hm} = \|\mathbf{A}_p - \mathbf{A}_{gt}\|_2^2$. The second loss is the in-out loss $\mathcal{L}_{io}$, and is classically defined as the binary cross entropy between the predicted $\mathbf{o}_p$ and ground truth $\mathbf{o}_{gt}$ in-vs-out of frame label.

Finally, to the contrary of previous works which relied only on 2D (in plane) gaze losses, we propose in this work to introduce a 3D direction loss to drive the gaze pathway network. It is defined so as to maximizes the cosine similarity between the prediction and the GT 3D gaze $\mathbf{g_{gt}}$:

$$\mathcal{L}_{dir} = 1 - < \mathbf{g_p}, \mathbf{g_{gt}} > \quad (8)$$

where $< a, b >$ denotes the inner product between a and b.

%Head	GazeFollow [52]	VAT [11]	ChildPlay *children*	ChildPlay *adults*
All	23.0	69.0	15.7	44.4
Multi	30.6	71.0	16.9	44.6

Table 2. GT percentage of looking at head instances. Statistics for all images (1st row) or images with at least 2 persons (2nd row).

5. Experiments

5.1. Experimental Protocol.

Implementation Details. The gaze network head $\mathcal{G}_h$ is pre-trained on the Gaze360 dataset [33] and processes the head crop at a resolution of 224×224. The Scene Pathway encoder is pre-trained on the ImageNet dataset [57] and processes the scene image at a resolution of 512×288. During the test phase, we maintain the original aspect ratio of the scene image and scale the longer side to 512.

Datasets. We train our models on 3 datasets - GazeFollow, ChildPlay and VideoAttentionTarget. More details about these datasets are provided in Section 3.4.

Training. We train for 40 epochs on GazeFollow. Following the protocol of [11], we fine-tune the model trained on GazeFollow for 20 epochs on VideoAttentionTarget. We adopt the same protocol for ChildPlay, and fine-tune the model trained on GazeFollow for 20 epochs. We use the AdamW optimizer [40] and set the learning rate as $2.5e-4$ for training on GazeFollow, and as $2.5e-5$ for fine-tuning on VideoAttentionTarget and ChildPlay. The loss coefficients are set as $\lambda_{hm} = 100$, $\lambda_{dir} = 0.1$ and $\lambda_{io} = 1$.

Validation. As GazeFollow and VideoAttentionTarget do not propose any validation set, we split a portion of the training set and use it for validation. Our GazeFollow validation split contains 4499 instances, and our VideoAttentionTarget validation split contains 6726 instances from 3 shows. The epoch with the best distance score on the validation set is used for testing.

5.2. Metrics

For evaluation, we use standard metric (AUC, Distance, AP) that we complement with a more semantic one: The precision of looking at heads (P.Head), as described below.

AUC. The predicted gaze heatmap is compared against the binarized GT gaze heatmap. This is used to plot the TPR vs FPR curve. AUC is the area under this curve.

Distance. The predicted gaze location is obtained from the

	Children				Adults				Full data			
Model	**AUC↑**	**Dist↓**	**AP↑**	**P.Head↑**	**AUC↑**	**Dist↓**	**AP↑**	**P.Head↑**	**AUC↑**	**Dist↓**	**AP↑**	**P.Head↑**
Gupta [23]†	0.926	0.136	-	0.435	0.919	0.151	-	0.621	0.923	0.142	-	0.518
Ours - 2D cone†	0.929	0.125	-	0.472	0.934	0.131	-	0.664	0.931	0.127	-	0.567
Ours†	0.934	0.112	-	0.509	0.930	0.119	-	0.681	0.932	0.115	-	0.602
Gupta [23]	0.923	0.106	0.980	0.648	0.914	0.123	0.987	0.731	0.919	0.113	0.983	0.694
Ours - 2D cone	0.925	0.118	0.937	0.564	0.927	0.125	0.955	0.717	0.926	0.121	0.944	0.644
Ours	0.939	0.098	0.989	0.604	0.928	0.121	0.983	0.704	0.935	0.107	0.986	0.663
Human	-	-	-	-	-	-	-	-	0.911	0.048	0.993	-

Table 3. Results on the ChildPlay dataset. The best results are given in red and the second best results are given in blue. † indicates that the model was not fine-tuned on ChildPlay.

A: Model	**AUC↑**	**Avg.Dist↓**	**Min.Dist↓**	**P.Head↑**
Fang [16]	0.922	0.124	0.067	-
Hu [28]	-	0.135	0.075	-
Bao [2]	0.928	0.122	-	-
Jin [31]	0.920	0.118	0.063	-
Chong [11]	0.921	0.137	0.077	0.708
Gupta [23]	0.933	0.134	0.071	0.750
Ours - 2D cone*	0.939	0.122	0.062	0.762
Ours*	0.936	0.125	0.064	0.760
Human	0.924	0.096	0.040	-

B: Model	**AUC↑**	**Dist↓**	**AP↑**	**P.Head↑**
Gupta [23]†	0.907	0.137	-	0.887
Ours - 2D cone†	0.915	0.128	-	0.894
Ours†	0.911	0.123	-	0.900
Fang [16]	0.878	0.124	0.872	-
Bao [2]	0.885	0.120	0.869	-
Jin [31]	0.898	0.109	0.897	-
Chong [11]	0.854	0.147	0.848	-
Gupta [23]*	0.897	0.134	0.864	0.903
Ours - 2D cone*	0.909	0.120	0.856	0.892
Ours*	0.914	0.109	0.834	0.902
Human	0.921	0.051	0.925	-

Table 4. Results on GazeFollow (A) and VideoAttentionTarget (B) with the best results in red and second best results in blue. * indicates that the model follows a proper protocol, using a validation split to select the model. † indicates that the model was not fine-tuned on VAT.

arg max of the predicted gaze heatmap. The Distance is the L2 distance between the predicted and GT gaze location on a 1×1 image. When multiple annotations are available (ex. GazeFollow) we can compute the minimum and average distance statistics.

Average Precision (AP). It is used to compute the performance for in vs out of frame gaze classification.

Looking at Heads (P.Head). The GT to compute this metric was obtained as follows. We first run a robust and powerful pre-trained Yolo-v5 [32] based head detector on images to get the head bounding boxes of everyone in the scene, and apply tracking [26]. We verified and further validated the obtained tracks. To obtain the GT, we then check if the annotated gaze target of a person falls inside a detected head box. As the GazeFollow test set contains multiple annotation, we check that at least two annotations fall inside the same detected head box. We provide the GT statistics for our datasets in Table 2.

At evaluation time, we perform the same procedure for each prediction to decide whether it is a gaze on a face. Finally, we compare the results with the GT, and compute the precision score.

5.3. Tested Models

ChildPlay. Other than our proposed model, we train the Image model of Gupta et al. [23] on our ChildPlay dataset. We also show results without any fine-tuning on ChildPlay, i.e. the models are only trained on GazeFollow.

GazeFollow and VideoAttentionTarget. We train our proposed model on the GazeFollow and VideoAttentionTarget benchmarks. We also re-train the Image model of Gupta et al. [23] on VideoAttentionTarget following our new training and validation splits. For state of the art, we compare results with the static model of Chong et al. [11], as well as models using depth information and using the same head crop input: Fang et al. [16], Hu et al. [28], Gupta et al. [23], Bao et al. [2] and Jin et al. [31].

Ablation: 3DFoV vs 2D cone. To see the benefit of using an explicit 3DFoV, we compare our approach to using a standard 2D gaze cone (similar to [23]) on the three datasets. Here the 2D gaze cone is derived by computing the similarity of the projected 3D gaze vector and the 2D scene locations (no decay factor).

5.4. Results

GazeFollow and VideoAttentionTarget (VAT). Our results on GazeFollow and VideoAttentionTarget are given in Table 4. As can be seen, our model achieves high results. Compared to the state-of-the-art, it is in par with the best method [31], which also used depth but without modeling an explicit 3DFoV, and may not use a validation set for evaluation. Indeed, on GazeFollow, although the Avg.Dist is

slightly worse for our approach, the Min.Dist metric is the same, and on VAT dataset both methods perform equally for Dist (0.109). Compared to [23] which follows the same protocol, our method performs much better.

Looking at the P.Head metric, we can notice that performance is in general quite high, esp. on the VAT dataset that has a large bias towards looking at heads (Table 2), so the true positives dominate the false positives.

ChildPlay. Our results on ChildPlay are given in Table 3. As it can be seen, our model shows much better cross-dataset generalization performance compared to our model with a 2D gaze cone and the model of Gupta et al. [23]. We see a general improvement in performance for all models after fine-tuning, with Gupta et al. [23] benefiting more from it and potentially slightly overfiting the dataset statistics and priors. Nevertheless, although the benefit is lower for our model, it is still the best on all metrics but the P.Head metric. Ultimately and interestingly, the gap in performance compared to human performance suggests a large potential for improvement.

Children vs Adults. Looking more in details, ChildPlay results show that the distance scores are slightly better for children compared to adults. However, this can mainly be attributed to the fact that gaze targets are on average closer to the child than to the adult (see Fig 4), a point also reported by Tu et al. [65] who showed that gaze target prediction models have lower performance (using a distance metric) for targets further away.

This contrasts with the P.Head metric, which shows that in this case, the performance is significantly lower for children than for adults (18.7% lower). This validates our hypothesis that different performance metrics are needed to fully assess models, and that models trained on existing datasets suffer when tested on children. This last point is corroborated by the fact that after fine-tuning, all models have a much larger improvement for children compared to adults whether for the distance or P.Head metrics, highlighting the importance of training the model with children data.

3DFoV vs 2D cone saliency. Results show that our model (Ours) with an explicit 3DFoV performs on par or much better than a model relying only a 2D saliency cone (Ours-2D cone). On GazeFollow, results are very slightly worse, which can be due to the fact that GazeFollow contains relatively simple scenes where depth is less important, and with a bias towards people being in the foreground (in the vast majority of a images, only the gaze of the person with the largest face in the scene is annotated). However, on VideoAttentionTarget and ChildPlay, our model with 3DFoV demonstrates much better cross-dataset generalization, and remains significantly better after fine-tuning. All this highlights the importance of depth information and the interest of using a geometrically consistent 3DFoV method.

6. Conclusion

In this paper we proposed a new dataset of children playing and interacting with adults. Our dataset has rich gaze annotations which is of interest for analysis of child gaze behaviour and gaze target prediction generally. We also proposed a new model that uses depth information to construct a geometrically grounded 3D field of view of a person. Our models achieve state of the art results on public benchmarks and ChildPlay. In particular, experiments indicate that training on ChildPlay can yield performance improvements for child gaze prediction, and that using semantic metrics (looking at faces) is useful to further characterize gaze models. In the future we will supplement ChildPlay with other layers of annotations (e.g. human-human-object interaction labels) and we encourage the research community to do the same.

Acknowledgement. This research has been supported by the AI4Autism project (Digital Phenotyping of Autism Spectrum Disorders in children, grant agreement no. CRSII5_202235 / 1) of the the Sinergia interdisciplinary program of the SNSF.

References

[1] Salvatore Maria Anzalone, Jean Xavier, Sofiane Boucenna, Lucia Billeci, Antonio Narzisi, Filippo Muratori, David Cohen, and Mohamed Chetouani. Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment. *Pattern Recognition Letters*, 118:42–50, 2019. 1

[2] 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, 3, 8, 13

[3] Younes Belkada, Lorenzo Bertoni, Romain Caristan, Taylor Mordan, and Alexandre Alahi. Do pedestrians pay attention? eye contact detection in the wild, 2021. 2, 3

[4] 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, 4, 14

[5] Rechele Brooks and Andrew N Meltzoff. The development of gaze following and its relation to language. *Developmental science*, 8(6):535–543, 2005. 1

[6] Rechele Brooks and Andrew N Meltzoff. Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of child language*, 35(1):207–220, 2008. 1

[7] 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. 13

[8] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M. Jones, Agata Rozga, and James M. Rehg. Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20, 2017. 1

[9] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M Jones, Agata Rozga, and James M Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20, 2017. 1

[10] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 383–398, 2018. 2

[11] 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. 2, 4, 5, 7, 8, 13, 14

[12] 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, 13

[13] Fith Edition. Diagnostic and statistical manual of mental disorders. *American Psychiatric Association*, 21:591–643, 2013. 1, 2

[14] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6460–6468, 2018. 3, 4, 5

[15] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5724–5733, 2019. 5

[16] 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, 3, 4, 8, 13

[17] John M Franchak, David J Heeger, Uri Hasson, and Karen E Adolph. Free viewing gaze behavior in infants and adults. *Infancy*, 21(3):262–287, 2016. 2

[18] K. Funes and J.-M. Odobez. Gaze estimation from multimodal kinect data. In *CVPR Workshop on Face and Gesture and Kinect demonstration competition (Best Student Paper Award), Providence*, june 2012. 2

[19] 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. 13

[20] Chunhui Gu, Chen Sun, David A Ross, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of CVPR*, 2018. 3

[21] Jian Guan, Liming Yin, Jianguo Sun, Shuhan Qi, Xuan Wang, and Qing Liao. Enhanced gaze following via object detection and human pose estimation. In *International Conference on Multimedia Modeling*, pages 502–513. Springer, 2020. 2

[22] Jian Guan, Liming Yin, Jianguo Sun, Shuhan Qi, Xuan Wang, and Qing Liao. Enhanced gaze following via object detection and human pose estimation. In *International Conference on Multimedia Modeling*, pages 502–513. Springer, 2020. 2

[23] 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. 2, 3, 4, 7, 8, 9, 14

[24] Jordan Hashemi, Geraldine Dawson, Kimberly L.H. Carpenter, Kathleen Campbell, Qiang Qiu, Steven Espinosa, Samuel Marsan, Jeffery P. Baker, Helen L. Egger, and Guillermo Sapiro. Computer Vision Analysis for Quantification of Autism Risk Behaviors. *IEEE Transactions on Affective Computing*, 3045(AUGUST):1–12, 2018. 1

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[26] Alexandre Heili, Adolfo Lopez-Mendez, and Jean Marc Odobez. Exploiting long-term connectivity and visual motion in crf-based multi-person tracking. *IEEE Transactions on Image Processing*, 23(7):3040–3056, 2014. 8

[27] 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. 14

[28] Zhengxi Hu, Dingye Yang, Shilei Cheng, Lei Zhou, Shichao Wu, and Jingtai Liu. We know where they are looking at from the rgb-d camera: Gaze following in 3d. *IEEE Transactions on Instrumentation and Measurement*, 2022. 3, 8

[29] 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. 14

[30] Tianlei Jin, Zheyuan Lin, Shiqiang Zhu, Wen Wang, and Shunda Hu. Multi-person gaze-following with numerical coordinate regression. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021. 2

[31] Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924, 2022. 3, 8

[32] 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. 8, 14

[33] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019. 5, 6, 7

[34] Labelbox. Labelbox, online, 2022. 4

[35] Beibin Li, Quan Wang, Erin Barney, Logan Hart, Carla Wall, Katarzyna Chawarska, Irati Saez de Urabain, Timothy J Smith, and Frederick Shic. Modified dbscan algorithm on oculomotor fixation identification. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 337–338, 2016. 1

[36] 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. 13

[37] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018. 2, 4, 5

[38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 7

[39] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation with calibration. In *British Machine Vision Conference 2018, BMVC 2018*, 2018. 2

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 7

[41] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019. 3, 4, 5

[42] Lucia Migliorelli, Sara Moccia, Rocco Pietrini, Virgilio Paolo Carnielli, and Emanuele Frontoni. The babypose dataset. *Data in brief*, 33:106329, 2020. 14

[43] Peter Mundy, Marian Sigman, and Connie Kasari. A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and developmental Disorders*, 20(1):115–128, 1990. 1

[44] Skanda Muralidhar, Rémy Siegfried, Jean-Marc Odobez, and Daniel Gatica-Perez. Facing employers and customers: What do gaze and expressions tell about soft skills? In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia, MUM 2018, Cairo, Egypt, November 25-28, 2018*, pages 121–126, 2018. 1

[45] Zhixiong Nan, Jingjing Jiang, Xiaofeng Gao, Sanping Zhou, Weiliang Zuo, Ping Wei, and Nanning Zheng. Predicting task-driven attention via integrating bottom-up stimulus and top-down guidance. *IEEE Transactions on Image Processing*, 30:8293–8305, 2021. 2

[46] Catharine Oertel, Patrik Jonell, Dimosthenis Kontogiorgos, Kenneth Funes Mora, Jean-Marc Odobez, and Joakim Gustafson. Towards an engagement-aware attentive artificial listener for multi-party interactions. *Frontiers in Robotics and AI*, 8:189, 2021. 1

[47] 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, 5, 13, 14, 15

[48] Fiona G Phelps, Gwyneth Doherty-Sneddon, and Hannah Warnock. Helping children think: Gaze aversion and teaching. *British journal of developmental psychology*, 24(3):577–588, 2006. 1

[49] Shyam Rajagopalan, Abhinav Dhall, and Roland Goecke. Self-stimulatory behaviours in the wild for autism diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 755–761, 2013. 4

[50] 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, 13, 15

[51] Adria Recasens*, Aditya Khosla*, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. * indicates equal contribution. 2, 4, 5

[52] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443, 2017. 2, 7

[53] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. Decoding children's social behavior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3414–3421, 2013. 2, 4

[54] J M Rehg, G D Abowd, A Rozga, M Romero, M A Clements, S Sclaroff, I Essa, O Y Ousley, Yin Li, Chanho Kim, H Rao, J C Kim, L L Presti, Jianming Zhang, D Lantsman, J Bidwell, and Zhefan Ye. Decoding Children's Social Behavior. In *Computer Vision and Pattern Recognition (CVPR)*, jun 2013. 1

[55] Omar Rihawi, Djamal Merad, and Jean-Luc Damoiseaux. 3d-ad: 3d-autism dataset for repetitive behaviours with kinect sensor. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017. 4

[56] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W. Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19):1–12, 2018. 1

[57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 7

[58] 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. 2, 13, 14

[59] Atsushi Senju and Mark H Johnson. Atypical eye contact in autism: models, mechanisms and development. *Neuroscience & Biobehavioral Reviews*, 33(8):1204–1214, 2009. 1

[60] S. Sheikhi and J.M. Odobez. Combining dynamic head pose and gaze mapping with the robot conversational state for attention recognition in human-robot interactions. *Pattern Recognition Letters*, 66:81–90, Nov. 2015. 1

[61] Frederick Shic, Jessica Bradshaw, Ami Klin, Brian Scassellati, and Katarzyna Chawarska. Limited activity monitoring in toddlers with autism spectrum disorder. *Brain research*, 1380:246–254, 2011. 1

[62] Uzma Haque Syeda, Ziaul Zafar, Zishan Zahidul Islam, Syed Mahir Tazwar, Miftahul Jannat Rasna, Koichi Kise, and Md Atiqur Rahman Ahad. Visual face scanning and emotion perception analysis between autistic and typically developing children. In *Proceedings of the 2017 acm international joint conference on pervasive and ubiquitous computing and proceedings of the 2017 acm international symposium on wearable computers*, pages 844–853, 2017. 1

[63] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 7

[64] Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3125–3133, 2021. 3, 4, 5

[65] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-end human-gaze-target detection with transformers. *arXiv preprint arXiv:2203.10433*, 2022. 2, 9

[66] Binglu Wang, Tao Hu, Baoshan Li, Xiaojuan Chen, and Zhijie Zhang. Gatector: A unified framework for gaze object prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19588–19597, 2022. 3

[67] 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, 3, 13

[68] 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. 5, 13, 14, 15

[69] Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. Learning to draw sight lines. *International Journal of Computer Vision*, 128(5):1076–1100, 2020. 2

7. Supplementary

7.1. 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 5. Our ChildPlay dataset is far more balanced[3], while also having 50% more frames and twice as much scene variety.

7.2. 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 [12]. 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 [58], it is important for younger age groups to be well represented in research benchmarks, particularly for applications targeting them. Table 6 summarizes the notable ones.

[3]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.

Dataset	things-person	things-other	stuff	not-detected
VideoAttention [11]	80.85%	8.05%	3.60%	7.50%
ChildPlay	45.19%	18.66%	12.62%	23.53%

Table 5. Comparison of gaze target semantic class between ChildPlay and VideoAttentionTarget. Numbers were obtained by running a panoptic segmentation model [7] on images and retrieving the semantic class of each annotated gaze point.

7.3. 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 [19]
- 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 [36]
- 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 [50]

Recent methods for monocular depth estimation [50][67] 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 [16] or attempt to correct the point cloud based on prior assumptions [2].

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. [68] 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. [47] 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

Name	Type	Setting	Size	Annotations
Sciortino et al. [58]	Video	SSBD dataset + youtube	1176 images of 104 subjects	2D pose keypoints
DREAM [4]	Video	Interactions with robot No raw data, only extracted features and annotations	306 hours of therapy (102) subjects	3D pose keypoints
BabyPose [42]	Video Depth	Preterm Infant movement in NICUs	16000 frames · 16 depth videos · 16 patients	2D pose keypoints
SyRIP [29]	Image	Hybrid: real + synthetic YouTube and Google images	Real: 700 images (140+ subjects) Synthetic: 1000 images	2D pose keypoints
MINI-RGBD [27]	Video Depth	Synthetic: obtained by registering SMIL to real sequences of moving infants. Constrained environment	12000 frames · 12 sequences	2D and 3D keypoints

Table 6. Summary of selected pose estimation children datasets.

et al. [50], Wei et al [68] and Patakin et al. [47] in Figure 6. 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. [68] and Patakin et al. [47], 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.

7.4. 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 [32] 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 [11][23] 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.

Figure 6. Comparison of point clouds generated using the depth maps from Ranftl et al. [50] (row 2), Wei et al. [68] (row 3) and Patakin et al. [47] (row 4) on ChildPlay images. The point clouds generated using Patakin et al. appear to best preserve the geometry of the scene.