




Article

Say What You Are Looking At: An Attention-Based Interactive System for Autistic Children

Furong Deng ^{1,†}, Yu Zhou ^{2,†}, Sifan Song ³, Zijian Jiang ⁴, Lifu Chen ⁵, Jionglong Su ^{6,*} , Zhenglong Sun ^{7,*} 
and Jiaming Zhang ^{4,8,*} 

- ¹ School of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China; furongdeng@outlook.com
 - ² School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China; joeyz1118@gmail.com
 - ³ Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China; Sifan.Song19@student.xjtlu.edu.cn
 - ⁴ Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518129, China; zijianjiang@cuhk.edu.cn
 - ⁵ DoGoodly International Education Center (Shenzhen) Co., Ltd., Smart Children Education Center, Shenzhen 518000, China; ertongxinli@126.com
 - ⁶ School of AI and Advanced Computing, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
 - ⁷ School of Science and Engineering, The Chinese University of Hong (Shenzhen), Shenzhen 518172, China
 - ⁸ Institute of Robotics and Intelligent Manufacturing, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China
- * Correspondence: Jionglong.Su@xjtlu.edu.cn (J.S.); sunzhenglong@cuhk.edu.cn (Z.S.); zhangjiaming@cuhk.edu.cn (J.Z.)
- † Both authors contributed equally to this work.



Citation: Deng, F.; Zhou, Y.; Song, S.; Jiang, Z.; Chen, L.; Su, J.; Sun, Z.; Zhang, J. Say What You Are Looking At: An Attention-Based Interactive System for Autistic Children. *Appl. Sci.* **2021**, *11*, 7426. <https://doi.org/10.3390/app11167426>

Academic Editor: Takashi Kuremoto

Received: 29 June 2021

Accepted: 4 August 2021

Published: 12 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Gaze-following is an effective way for intention understanding in human–robot interaction, which aims to follow the gaze of humans to estimate what object is being observed. Most of the existing methods require people and objects to appear in the same image. Due to the limitation in the view of the camera, these methods are not applicable in practice. To address this problem, we propose a method of gaze following that utilizes a geometric map for better estimation. With the help of the map, this method is competitive for cross-frame estimation. On the basis of this method, we propose a novel gaze-based image caption system, which has been studied for the first time. Our experiments demonstrate that the system follows the gaze and describes objects accurately. We believe that this system is competent for autistic children's rehabilitation training, pension service robots, and other applications.

Keywords: human–robot interaction; image caption; simultaneous localization and mapping; visual attention

1. Introduction

Humans are very good at understanding the intentions of others by following the gaze. We can infer that the child is interested in a ball if they keep staring at it. We also can find crucial clues through the suspect's attention at the scene of the crime. This ability leads us to obtain obscure but essential information. If robots also have the capability of gaze-following, they would be competent for many human–robot interaction tasks, including helping doctors with rehabilitation training for autism [1,2]. This is the goal that we set to achieve. For autistic children, they are usually interested in some abnormal objects, such as bottle caps or door handles, instead of toys that non-autistic children like.

Autistic children may pay attention to the object they are interested in for a long time and ignore the doctor's instructions in rehabilitation training. In order to understand what the children are interested in, the robot also needs to talk to gain the children's attention and guide them to participate in the doctor's task. This is challenging for a robot. As

shown in Figure 1, the general image caption focuses on describing what the robot sees, while the gaze-based image caption focuses on describing what the child is looking at. Obviously, the latter is more suitable for interaction scenarios. In this paper, we propose a gaze-based image caption system. We equipped a depth camera on the robot's chest to collect images; the type of camera we used is RealSense D435. This camera can obtain a color image and the corresponding depth image. The depth map measures the distance of every pixel, which is used to reconstruct geometric relations of the gaze and the map. Given a video sequence of depth images and the corresponding color images, the camera builds a point cloud map for locating what the children are interested in and generates a caption to interact with them.

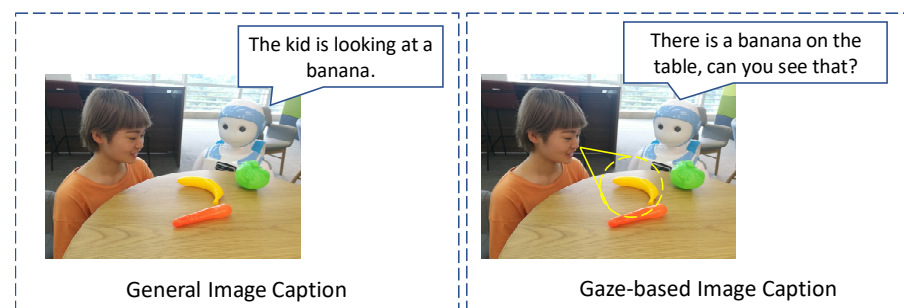


Figure 1. The difference between general image caption and gaze-based image caption.

1.1. Gaze-Following

Gaze-following is a problem of predicting where people are looking at in a given image or video. Many previous works have used wearable devices such as glasses with a camera to track the gaze direction and locate the target from a 3D map [3,4]. However, the wearable devices are usually unavailable in practical applications. A more promising way is to predict the gaze from images directly. Fathi et al. [5] and Marin-Jimenez et al. [6] address this problem by assuming that people are only looking at each other. They used the posture and position of the head as clues to locate subjects. Inspired by the study of electric fields, Park et al. [7] used “social charges” to represent latent quantities that drive the primary gaze behavior of members of a social group. Recasens et al. [8] make a great contribution on first publishing a gaze-following dataset and proposing a two-pathway model (a gaze pathway and a saliency pathway). However, Lian et al. [9] and Parks et al. [10] consider an internal connection between gaze and saliency, rather than complete independence. For example, the gaze point is always located at a salient place along the gaze direction. Drawing on this idea, they proposed new models with better performance. Chong et al. [11,12] propose a new deep structure that models the gaze over time. It directly learns the gaze-relevant scene regions by face feature instead of giving head posture. Unfortunately, the above methods require that the people and the objects appear in the same image. For example, a child is watching TV, but the TV is not in the view of camera. These methods cannot keep following the gaze when the camera turns to the side of the TV. Recasens et al. [13] name this problem as cross-frame gaze-following. They proposed a model adding to a transformation pathway based on their previous work [8]. Given two images from different views, this model keeps following the gaze by estimating the transformation between these two frames. However, the deep learning method for geometric estimation is inaccurate and increases the uncertainty of the result. Therefore, we propose a different gaze-following method. We select a sequence of depth and color images as the input and built a three-dimensional (3D) map for predicting the gaze point. Our method provides a new solution for cross-frame gaze-following. The main differences between the above methods and ours are that (a) we build and use 3D maps online for cross-frame gaze-following, and (b) compared with other geometric methods, we propose an occlusion detection mechanism that minimizes error prediction.

1.2. Attention-Based Image Caption

Computers are expected to describe the world from a human perspective. An image contains a great quantity of information. How do computers choose valuable information? The answer is the attention mechanism. Most of the existing works focus on describing the whole image as human as possible. Xu et al. [14] and Lu et al. [15] generated the contextual attention by a recurrent neural network (RNN) [16] and adjusted the weight of the context vector to generate a sentence. Cornia et al. [17,18] considered attention as visual salient information that is prominent and simple to be noticed, such as high-contrast objects. Moreover, Liu et al. [19] and Sugano et al. [20] modeled the attention as gaze data of the subjects' eye movements while watching the video. Some extended methods of image caption were studied to describe the details of the image. Johnson et al. [21] used the region proposal network (RPN) [22] to generate multiple local bounding boxes to extract features. Each region feature is used to generate a sentence. Subsequently, Yang et al. [23] improved the model using joint inference and context fusion. Wang et al. [24] proposed a multilayer dense attention model to minimize the interference due to non-salient information. However, all these attention measurement methods are quite subjective, and the results vary with people's personalities, age, and emotions. These methods may not be suitable for some people, such as autistic children. To obtain a more objective result, we ensured that the attention is reflected only by the gaze rather than other factors. We propose a novel method that attaches weights to the regions where people are looking at. We note that our gaze data are measured in the video, while the gaze data of methods [19,20] are collected by the people who are watching the video. In this paper, a gaze-based image caption system is proposed on the basis of the work of [21]. We utilized the result of gaze-following to select the description candidates and to describe the attention area. Similar to [25], our system was mainly designed for autistic children, but it can also be applied to general people. We have three main contributions in this paper: (1) A novel gaze-following method is proposed on the basis of spatial geometry. It predicts attention regions by the spatial relationship between the map and the sightline. (2) An image caption method guided by the gaze is proposed. It describes the region concentrated by users according to attention heatmaps. (3) For the first time, we studied the problem of describing the region where people are looking at and combine image caption with gaze-following.

2. Methods

We propose an object description system on the basis of third-person visual attention. The system predicts the interested area of a person and describes this area through an object description algorithm. Figure 2 shows the framework of the whole system. It consists of visual attention prediction and object description. With the support of SLAM (simultaneous localization and mapping) and gaze tracking techniques, we constructed a 3D map of the environment online and track the gaze of the person in real time. We utilized the geometric relationship between the map and the sightline to assign weights in the map for indicating the degree of interest. Finally, the object description algorithm was used to preferentially describe objects with high attention. There are four main challenges in this system: (1) object occlusion detection: some occluded objects in the map are invisible, and the algorithm has to determine the obscured point clouds in the sparse map; (2) map update: since moving objects are significantly destructive to the structure of the map, the point clouds of these objects should be updated in real time to avoid residual traces of previous ones; (3) sightline error: the results of gaze tracking are often affected by a number of factors, resulting in violent shaking, and it is therefore essential to improve the stability of gaze tracking; (4) describing specific areas according to the actual attention of the human eye. We describe our algorithm in detail along with the solution to these challenges in Sections 2.1 and 2.2, respectively.

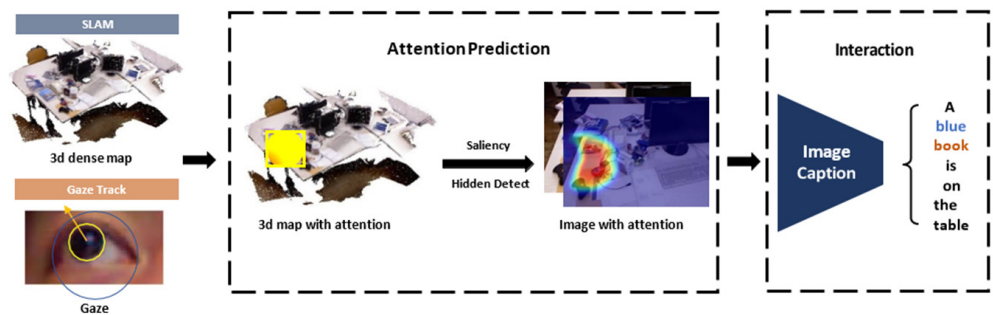


Figure 2. The system takes a video as input. It builds a map and estimates the gaze simultaneously. Then, an attention heat map is predicted. Finally, the system selects the attention region to generate a sentence from the image.

2.1. Visual Attention Prediction

Before carrying out visual attention prediction, we need to process the results of gaze tracking. Since gaze estimation is greatly affected by the noise, we adopt head posture estimation that adjusts weights and generates more stable and accurate results according to the confidence level. We can see how to calculate the sightline from Equations (1) and (2).

$$p = \frac{p_l + p_r}{2} \tag{1}$$

$$d = \alpha_e \frac{d_l + d_r}{2} + \alpha_h d_h \tag{2}$$

Parameters p_l and p_r represent the left and right eye positions, respectively. The gaze direction of the left and right eyes are d_l and d_r respectively. The parameter d_h represents the head direction, and α_e, α_h are standardizing confidence levels. Equation (1) determines the equation of the sightline in 3D space, where p is the starting point of the sightline and d is the gaze direction. We then transform the linear equation from the camera coordinate to the world coordinate and represent the sight line with Equation (3). Notably, the sightline has a direction, and therefore the value of parameter t is greater than 0.

$$p' = p + td \tag{3}$$

We then obtain the visible point cloud within the map, which is marked as S , as shown in Figure 3. First, we construct a mathematical cone whose main axis has an inclination angle of ϵ . We specify that the points within this round table to be the visible point cloud S that reflects potential regions focused by people. For a point in the map, we first obtain the perpendicular foot to the sightline. After that, we calculate the distance from this point to the perpendicular foot and the distance from the perpendicular foot to the eye. By Equation (4), we determine whether this point is in the visible point cloud S .

$$p_i \in S, \text{ if } \tan(\epsilon) > \frac{\max(\text{dist}(p_i, p_f) - r, 0)}{\text{dist}(p, p_f)} \tag{4}$$

where p_i is a point in the map, p_f is the perpendicular foot of p_i to the sightline, ϵ is the inclination angle, r is the radius of the circle above the frustum, and $\text{dist}(*, *)$ is the distance between two points. The visible point cloud S determines the area where a person focuses on. However, some of the points in S are occluded. These points are not visible. Imprecise assignment of weights may distract the attention, and therefore in this paper, we propose a method to address this problem. For each visible point, we construct a cone. The direction

of the central axis of the cone is identical to the direction of the sightline. We specify that the point cloud within the cone is the occluded point cloud R , as shown in Equation (5).

$$p_j \in R, \text{ if } \tan(\theta) > \frac{\text{dist}(p_i, p_f)}{\text{dist}(p_f, p_j)} \quad (5)$$

where p_i is a point in visible point cloud S , p_f is the perpendicular foot of p_i , p_j is a point that needs to be checked, and θ is the inclination angle. We eliminated all the points in R so that the remaining point cloud reflects the actual attention region of the person. We assigned weights to the visible point cloud by constructing a Gaussian model, which is guided by the distance from points to sightline. The weight of each point cloud is inversely correlated with the distance, which implies that closer points allocated bigger weights. For convenience, we specify that the weight is in an interval from 0 to 1. On the basis of the 3D map with weight, we use a camera model to project this map onto an image plane. Subsequently, the discrete points with weight can be transformed into a continuous probability map using image processing techniques, such as dilation and filtering. We combine the results of the saliency detection and multiplied the two probability maps to obtain the final attention predictions. Algorithm 1 shows the detailed algorithm flow. When the person and the target do not appear in the same field of vision, we first observe the face and estimate the sightline of that person, then move the camera along the sightline until it reaches the map and finds out the visible point cloud. When the face moves out of view, its gaze direction is assumed to be unchanged, which corresponds to human behavior. The whole process is the same as that which we discuss above, except changing the field of vision and tracking the sightline.

Algorithm 1 Finding Visible Point Cloud

Input:

- 1: (a) Point cloud map M ;
- 2: (b) Eye position p ;
- 3: (c) Direction d ;
- 4: (d) Cone angle ϵ , θ ;
- 5: (e) radius r .

Output:

- 6: (a) Visible point cloud S ;
 - 7: (b) Hidden point cloud R .
 - 8: **for** each $p_i \in M$ **do**
 - 9: Find the foot point p_j of p_i along the sightline
 - 10: Calculate the distance between p_f and p_i
 - 11: Calculate the distance between p_f and p
 - 12: **if** p_i is satisfied with Equation (4) **then**
 - 13: Let p_i belongs to S
 - 14: **end if**
 - 15: **end for**
 - 16: **for** each $p_i \in S$ **do**
 - 17: Generate a line l with a slope of d starting at p_i
 - 18: **for** each $p_j \in S$ **do**
 - 19: Find the foot point p_f of p_j along the line l
 - 20: Calculate the distance between p_f and p_i
 - 21: Calculate the distance between p_f and p_j
 - 22: **if** p_j is satisfied with Equation (5) **then**
 - 23: Remove the p_j from S
 - 24: Let p_j belongs to R
 - 25: **end if**
 - 26: **end for**
 - 27: **end for**
-

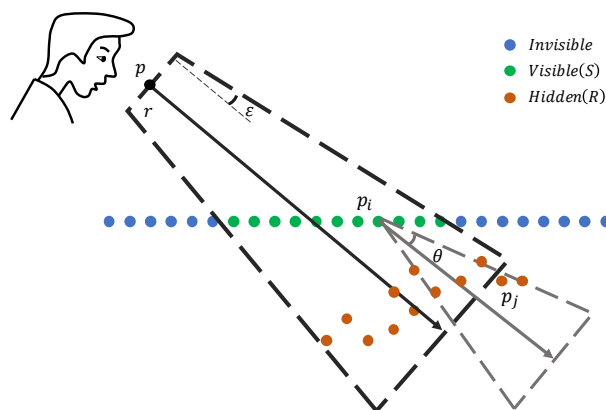


Figure 3. Occlusion detection and visual region estimation.

Since the scene often contains moving objects, e.g., people and animals whose motion traces are recorded by SLAM in the map, they often disrupt the original structure of the map and result in a large error of visual attention prediction. Therefore, real-time map updating is required. We employed a simple and effective strategy to ensure the real-time capabilities of the system. Before loading a new frame of the point cloud, we erase the camera-observable point cloud to ensure that the map stores the latest frame.

The camera-observable point cloud is defined as

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x/z \\ y/z \\ 1 \end{bmatrix} \tag{6}$$

where $0 \leq u \leq W$, $0 \leq v \leq H$, x, y, z is the position in the world of each point; u, v is the position in image; f_x, f_y, u_0, v_0 is the intrinsic parameters from camera; W is image width; and H is image height. We specify that a point can be updated when its projection is within the image range.

2.2. Object Description

A dense captioning task first proposed by Johnson et al. [21] and Yang et al. [23] introduces a method by adding joint inference and visual context based on [21] for performance improvement, and [25] also proposes to apply this task to rehabilitation robots. We trained the model using Visual Genome [26]. Figure 4 shows the overall network architecture. Given an image and its probability map created by gaze prediction, we first utilize CNN (convolutional neural network) [27] and region proposal network to generate a series of region features, then select the one with the highest weight density to output a sentence by the captioning model. Specifically, the captioning model consists of a recognition network and LSTM (long short-term memory) [28]. The recognition network is a fully connected network that takes the selected region feature as input and produces a string of visual text code. The RNN (recurrent neural network) utilizes LSTM to propagate the hidden state and recurrently sample the most likely next words. The framework of the whole network is similar to DenseCap [21], except that we add a box selecting module to focus on the attention region. The optimization objective of the dense captioning model is to minimize its loss function $L(I, S)$, which is given in Equation (7).

$$L(I, S) = - \sum_{t=1}^N \log p_t(S) \tag{7}$$

where I is the input image, and $S = (s_1, \dots, s_N)$ is a true sentence describing this image. We initialize the weights of the CNN with a pre-trained model on ImageNet [29] and add a regular term when the gradient is updated to avoid overfitting, as shown in Equation (8).

This helps to compare to training with an uninitialized model and without adding regular terms.

$$\omega_{i+1} = \frac{\omega_i - \partial L(I, S)}{\partial \omega_i} \quad (8)$$

where w_i is the weights in the network, and w_{i+1} is the weight to be updated.

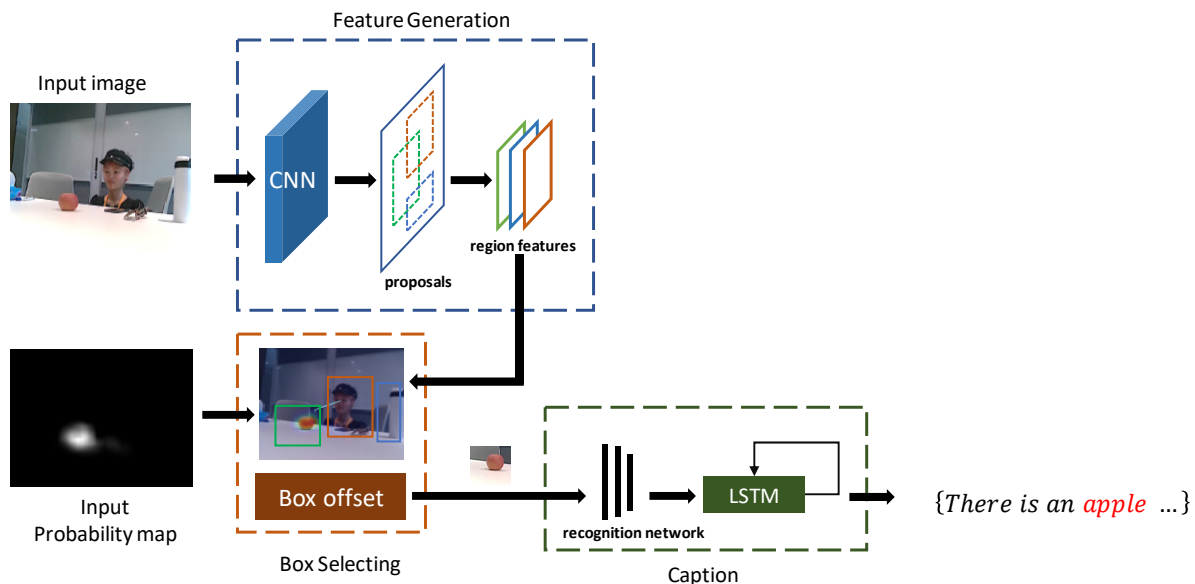


Figure 4. The framework of attention-based object description method.

3. Experimental Evaluation

In this section, we describe the attention prediction dataset for a rehabilitation training scenario for children with autism and evaluate our algorithm in this scenario, both in terms of visual attention prediction and object description. As advanced attention prediction algorithms, refs. [9] and [12] achieve accurate results in predicting the attention of people in the picture. They use the head posture as a clue to track people's attention by a neural network, while we focus on the geometric information brought by the 3D point cloud map and predict attention through the geometric relationship between sightline and the map. This also helps to achieve cross-frame prediction. We compare our method with [9] and [12] in visual attention prediction. Additionally, in object description, we compare with advanced general image caption methods [21,30,31] to show that our gazed-based method is more suitable for human–robot interaction.

3.1. Dataset

We used the Visual Genome dataset [26] for object description model training. It provides region proposals for each image, which may contain important information. Before training, we remove repetitive descriptions and use YoloV3 object detector [32] to crop images, because we focus on the specific objects that attract people. There is no need to involve all objects in the scene. Finally, we used 103,521 images including 688,143 regions for training and 4556 images with more than 70 descriptions including 40,989 regions for testing. For evaluation, we simulated rehabilitation scenes and capture 24 RGB-D videos from the perspective of the rehabilitation robot. Five of them were use to evaluate the performance of the cross-frame gaze estimation method. In actual scenes, the autistic child sits in front of a table and is instructed by a trainer to identify objects on the table, including daily-life tools, fruits, and animals. We require subjects to imitate the autistic child gazing at objects or playing with them and then provide a few sentences based on the objects as evaluation templates. We run our system on TianXP. During the test, the frame rate of visual attention prediction reach about 15 fps, and that of the object description is about 1 fps.

3.2. Visual Attention Evaluation

The mapping, gaze-following, and saliency detection in our system are implemented by OpenFace [33], ORB-SLAM2 [34], and OpenCV module [35], respectively. We used our dataset to evaluate visual attention prediction and compare it with [9] and [12]. In order to keep the input of each algorithm the same, we were provided with both the face bounding box input of [9] and [12] methods by OpenFace. Moreover, we divided our visual attention prediction into three schemes: Gaze, Gaze + Hiddendetct, and Gaze + Hiddendetct + Saliency for showing the effect of occlusion detection and saliency detection. To measure visual attention prediction, we employed the widely used matrices, ROC (receiver operator characteristics), PR (precision recall), and AUC (area under the curve). The ROC curve effectively reflects the relationship between the true-positive rate and the false-positive rate. The PR curve is to visualize the accuracy of the model. Therefore, the large AUC of the two curves demonstrates the superiority of the model.

The test results are provided in Figure 5 and Table 1. The PR plot shows that the accuracy of this algorithm is higher than the others. The ROC curve of our algorithm is below that of [9]. However, the accuracy of our method is better than any of them. The prediction of [9] contains the best coverage area, but it also contains a relatively high level of false-positive cases. As for [12], it is highly accurate in some cases but is susceptible to the complex background. The prediction of [12] is more biased towards the region near the hand, and thus the overall performance is poor. Moreover, in the three experiments of our method, gaze had larger errors and more false positives. The reason is that the algorithm incorrectly assigns more weights to occluded objects. Gaze + HiddenDetect shows that occlusion detection significantly improves the accuracy of our system. On this basis, Gaze + HiddenDetect + Saliency further optimizes the details of attention regions, and Figure 6 shows its predictions. When the participant is looking at the green pepper on the left, only Gaze + HiddenDetect + Saliency accurately locate it. Figure 7 illustrates the effect of our attention prediction algorithm on cross-frame gaze estimation experiments. The target is not in the camera field at first. On the basis of the 3D point cloud map, the algorithm estimates the person's sightline and follows its direction until the point cloud falls into the cone of sight. Finally, we assign weight to these points and projected them into the image. This procedure shows that 3D point cloud map plays an important role in cross-frame estimation.

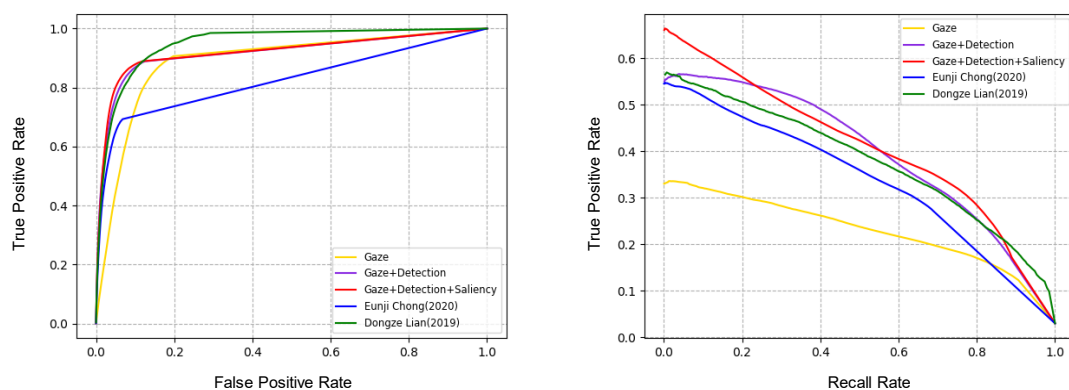


Figure 5. ROC curves (left), PR curves (right).

Table 1. PR and ROC values between different methods.

Method	ROC_{AUC}	PR_{AUC}
Gaze	0.889	0.230
Gaze+HiddenDetect	0.917	0.395
Gaze+HiddenDetect+Saliency	0.920	0.408
Eunji Chong [12]	0.824	0.335
Dongze Lian [9]	0.949	0.378

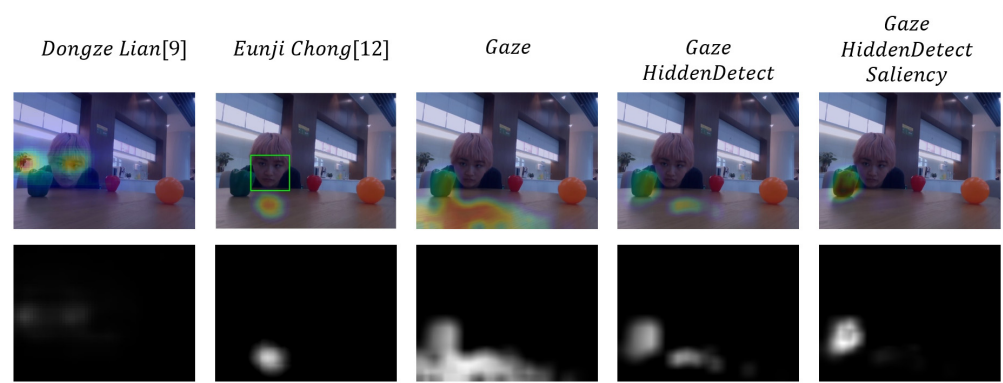


Figure 6. Application of the robot in autism rehabilitation training.

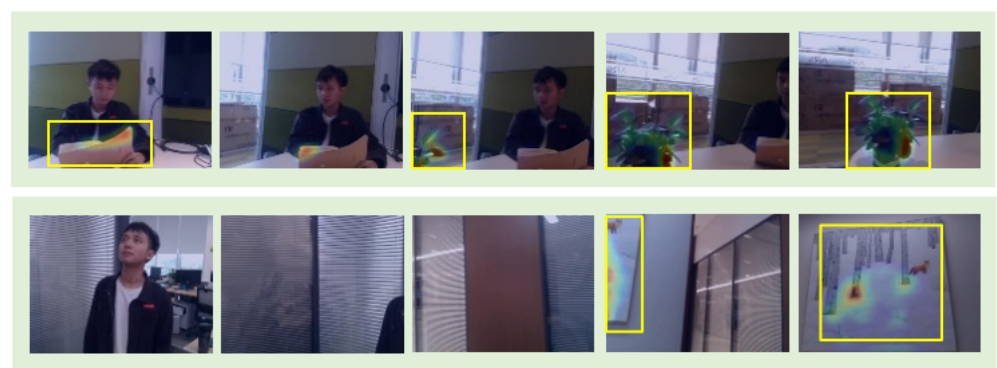


Figure 7. Illustration of cross-frame gaze estimation.

3.3. Object Description Evaluation

To assess the performance of our proposed method, we compared it with that of other state-of-the-art methods. Reference [30] proposes an adaptive attention algorithm, [21] combines object detection with image caption to achieve improved performance, and [31] combines object detection with an adaptive attention algorithm to obtain more specific descriptions. The dataset is collected from the perspective of a robot.

Figure 8 shows the comparison between the performance of our method and that of [31]. Our method generates the description of the area focused by human eyes rather than all objects in the scene. It is an advantage that our method takes the perspective of the person, and the descriptions are more specific and are currently not possible with other image caption algorithms. We apply widely used evaluation metrics, BLEU (bilingual evaluation understudy), CIDER (consensus-based image description evaluation), METEOR (metric for evaluation of translation with explicit ordering), ROUGE (recall-oriented understudy for gisting evaluation), and SPICE (semantic propositional image caption evaluation). Furthermore, due to matching problems of these matrices, we also include BERT (bidirectional encoder representations from transformers) [36] for a better assessment of sentence semantics. Meanwhile, refs. [21,30] generate object descriptions and utilize the mentioned evaluation matrices. By comparing the sentence evaluation scores of the description results in Table 2, we find that our method outperformed other algorithms, with the highest evaluation scores and the best descriptions.

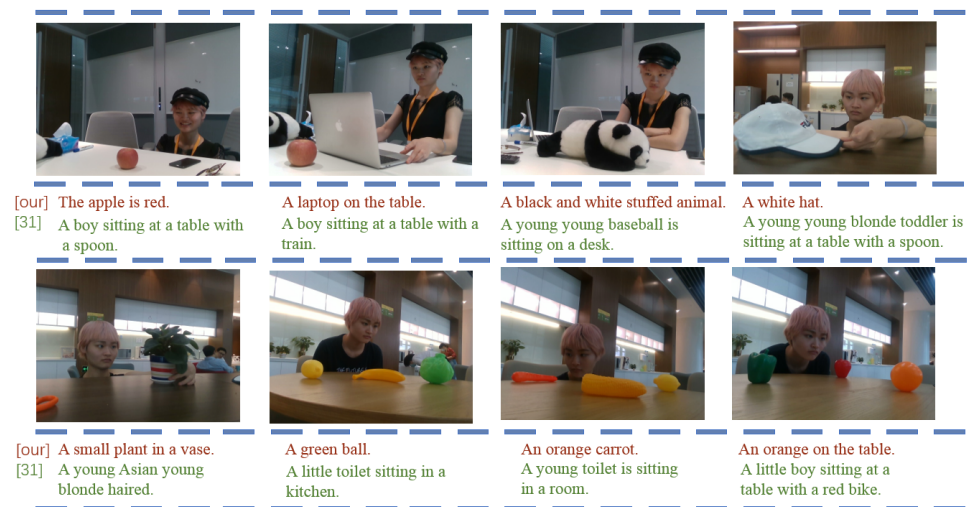


Figure 8. Comparison between our method and [31].

Table 2. Evaluation scores between our method and [21,30].

Method	Ours	DenseCap [21]	NIC [30]
BLEU1	0.251	0.176	0.154
BLEU2	0.133	0.07	0.075
BLEU3	0.08	0.035	0.043
BLEU4	0.053	0.019	0.027
CIDER	0.558	0.257	0.373
METEOR	0.132	0.088	0.043
ROUGE	0.239	0.155	0.163
SPICE	0.182	0.096	0.068
BERT	0.885	0.863	0.803

4. Conclusions and Future Work

This paper creatively integrates third-person visual attention into object description and proposes a human–robot interaction system applied to the rehabilitation of children with autism. This system is composed of two main parts, attention prediction and object description. The first part utilizes the spatial relationship between the sightline and the map to predict the interested area. The second part utilizes the attention probability map to describe objects. The experiments demonstrate that our method predicts the objects interested by children with high efficacy. However, there are some limitations in our paper. The existing image description accuracy is insufficient. Moreover, the efficiency of gaze tracking and SLAM algorithm needs further optimization. Apart from this, we have several desirable extensions. For example, robots use real-time tracking of areas of interest to the human eye to conduct conversations with people, helping empty-nest elderly, white-collar workers, and other people to relieve their worries, prevent depression, and so on. In terms of robot applications, our system provides new ideas for human–robot interaction. In the future, we will focus on the work of improving recognition accuracy and enriching the forms of sentences.

Author Contributions: Conceptualization, J.Z.; methodology, Y.Z. and F.D.; validation, Y.Z. and F.D.; formal analysis, F.D. and J.S.; resources, L.C.; data curation, J.Z.; writing—original draft preparation, Y.Z. and F.D.; writing—review and editing, S.S. and Z.J.; supervision, Z.S. and J.S.; project administration, J.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shenzhen Science and Technology Innovation Commission, grant numbers JCYJ20170410172100520 and RCYX20200714114736115. This research was also funded by the Open Program of Neusoft Corporation, item number SKLSAOP1702, as well as the Shenzhen Institute of Artificial Intelligence and Robotics for Society, grant number 2019-INT020 and AC01202101014. The APC was funded by the Shenzhen Science and Technology Innovation Commission.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the reason that all the participants in this study were the authors (Y.Z. and F.R.D.) but not patients.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank Lifu Chen and his colleagues in DoGoodly International Education Center and Smart Children Education Center for providing facilitation and assistance to our facilitators in collecting experimental data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Anamaria, P.C.; Ramona, S.; Sebastian, P.; Pinteau, S.; Saldien, J.; Rusu, A.; David, D.; Vanderfaillie, J.; Lefeber, D.; Vanderborght, D. Can the social robot probo help children with autism to identify situation-based emotions? A series of single case experiments. *Int. J. Hum. Robot.* **2013**, *10*, 1350025. [[CrossRef](#)]
2. Peca, A.; Tapus, A.; Aly, A.; Pop, C.; Jisa, L.; Pinteau, S.; Rusu, A.; David, D. Exploratory Study: Children's with Autism Awareness of Being Imitated by Nao Robot. *arXiv* **2012**, arXiv:2003.03528.
3. Santner, K.; Fritz, G.; Paletta, L.; Mayer, H. Visual recovery of saliency maps from human attention in 3D environments. In Proceedings of the IEEE International Conference on Robotics & Automation, Karlsruhe, Germany, 6–10 May 2013. [[CrossRef](#)]
4. Wang, H.; Pi, J.; Qin, T.; Shen, S.; Shi, B.E. SLAM-based localization of 3D gaze using a mobile eye tracker. In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland, 14–17 June 2018; pp. 1–5. [[CrossRef](#)]
5. Fathi, A.; Hodgins, J.K.; Rehg, J.M. Social interactions: A first-person perspective. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1226–1233. [[CrossRef](#)]
6. Marin-Jimenez, M.J.; Zisserman, A.; Eichner, M.; Ferrari, V. Detecting People Looking at Each Other in Videos. *Int. J. Comput. Vis.* **2014**, *106*, 282–296. [[CrossRef](#)]
7. Park, H.S.; Jain, E.; Sheikh, Y. Predicting Primary Gaze Behavior Using Social Saliency Fields. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013. [[CrossRef](#)]
8. Recasens, Adrià Recasens Contiente. Where are they looking? Diss. Massachusetts Institute of Technology. 2016. Available online: <http://gaze-follow.csail.mit.edu/> (accessed on 2 June 2021).
9. Lian, D.; Yu, Z.; Gao, S. Believe It or Not, We Know What You Are Looking at! 2018. Available online: <https://github.com/svip-lab/GazeFollowing/> (accessed on 2 June 2021).
10. Parks, D.; Borji, A.; Itti, L. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vis. Res.* **2015**, *116*, 113–126. [[CrossRef](#)] [[PubMed](#)]
11. Chong, E.; Ruiz, N.; Wang, Y.; Zhang, Y.; Rozga, A.; Rehg, J.M. 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), Munich, Germany, 8–14 September 2018. [[CrossRef](#)]
12. Chong, E.; Wang, Y.; Ruiz, N.; Rehg, J.M. Detecting Attended Visual Targets in Video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online Conference, 14–19 June 2020. [[CrossRef](#)]
13. Recasens, A.; Vondrick, C.; Khosla, A.; Torralba, A. Following Gaze Across Views. *arXiv* **2016**, arXiv:1612.03094.
14. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.S.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
15. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [[CrossRef](#)]
16. Medsker, L.R.; Jain, L.C. Recurrent neural networks. *Des. Appl.* **2001**, *5*, 64–67.
17. Cornia, M.; Baraldi, L.; Serra, G.; Cucchiara, R. Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention. *Acm Trans. Multimed. Comput. Commun. Appl.* **2018**, *14*, 1–21. [[CrossRef](#)]
18. Cornia, M.; Baraldi, L.; Serra, G.; Cucchiara, R. Visual saliency for image captioning in new multimedia services. In Proceedings of the IEEE International Conference on Multimedia & Expo Workshops, Hong Kong, China, 10–14 July 2017. [[CrossRef](#)]
19. Liu, Y.; Wu, Q.; Tang, L.; Shi, H. Gaze-assisted Multi-stream Deep Neural Network for Action Recognition. *IEEE Access* **2017**, *5*, 19432–19441. [[CrossRef](#)]

20. Sugano, Y.; Bulling, A. Seeing with Humans: Gaze-Assisted Neural Image Captioning. *arXiv* **2016**, arXiv:1608.05203.
21. Johnson, J.; Andrej, K.; Li, F.-F. Denscap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [[CrossRef](#)]
22. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [[CrossRef](#)] [[PubMed](#)]
23. Yang, L.; Tang, K.; Yang, J.; Li, L.-J. Dense Captioning with Joint Inference and Visual Context. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 17–22 June 2017. [[CrossRef](#)]
24. Wang, E.K.; Zhang, X.; Wang, F.; Wu, T.Y.; Chen, C.M. Multilayer Dense Attention Model for Image Caption. *IEEE Access.* **2019**, *7*, 66358–66368. [[CrossRef](#)]
25. Zhang, B.; Zhou, L.; Song, S.; Chen, L.; Jiang, Z.; Zhang, J. Image Captioning in Chinese and Its Application for Children with Autism Spectrum Disorder. In Proceedings of the ICMLC 12th International Conference on Machine Learning and Computing, Shenzhen, China, 15–17 February 2020. [[CrossRef](#)]
26. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D.A.; et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [[CrossRef](#)]
27. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
28. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
29. Jia, D.; Wei, D.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Computer Vision & Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [[CrossRef](#)]
30. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [[CrossRef](#)]
31. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Neural Baby Talk. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. [[CrossRef](#)]
32. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
33. Baltrusaitis, T.; Zadeh, A.; Lim, Y.C.; Morency, L.-P. OpenFace 2.0: Facial Behavior Analysis Toolkit. In Proceedings of the IEEE 13th International Conference on Automatic Face & Gesture Recognition, IEEE Computer Society, Xi'an, China, 15–19 May 2018; pp. 59–66. [[CrossRef](#)]
34. Mur-Artal, R.; Tardós, J.D. ORB-SLAM2: An Open-source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [[CrossRef](#)]
35. Montabone, S.; Soto, A. Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image Vis. Comput.* **2010**, *28*, 391–402. [[CrossRef](#)]
36. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. 2019. Available online: https://github.com/Tiiiger/bert_score/ (accessed on 2 June 2021).