GazeD: Context-Aware Diffusion for Accurate 3D Gaze Estimation

Riccardo Catalini[1,*] Davide Di Nucci[1] Guido Borghi[1] Davide Davoli[2,†]
Lorenzo Garattoni[2] Giampiero Francesca[2] Yuki Kawana[3]
Roberto Vezzani[1]

[1]University of Modena and Reggio Emilia [2]Toyota Motor Europe [3]Woven by Toyota

{rcatalini, ddinucci, gborghi, rvezzani}@unimore.it,
{davide.davoli, lorenzo.garattoni, giampiero.francesca}@toyota-europe.com, yuki.kawana@woven.toyota

Abstract

We introduce GazeD, a new 3D gaze estimation method that jointly provides 3D gaze and human pose from a single RGB image. Leveraging the ability of diffusion models to deal with uncertainty, it generates multiple plausible 3D gaze and pose hypotheses based on the 2D context information extracted from the input image. Specifically, we condition the denoising process on the 2D pose, the surroundings of the subject, and the context of the scene. With GazeD we also introduce a novel way of representing the 3D gaze by positioning it as an additional body joint at a fixed distance from the eyes. The rationale is that the gaze is usually closely related to the pose, and thus it can benefit from being jointly denoised during the diffusion process. Evaluations across three benchmark datasets demonstrate that GazeD achieves state-of-the-art performance in 3D gaze estimation, even surpassing methods that rely on temporal information. Project details will be available at https://aimagelab.ing.unimore.it/go/gazed

1. Introduction

The importance of 3D gaze estimation lies in its ability to unlock deeper insights into human attention [47], and cognition [12], which are central to a wide range of applications such as human-computer interaction [51], behavioral analysis [30], and extended reality systems [54], surveillance [62], autonomous driving [43], and robotics [44].

Computer vision researchers have traditionally approached automated gaze analysis by dividing it into two main tasks [59]: *gaze estimation* and *gaze target detection*, also referred to as *gaze following*. Specifically, gaze estimation aims to predict the direction of a person's gaze, while gaze target detection aims to pinpoint the exact location a person is looking at within the scene.

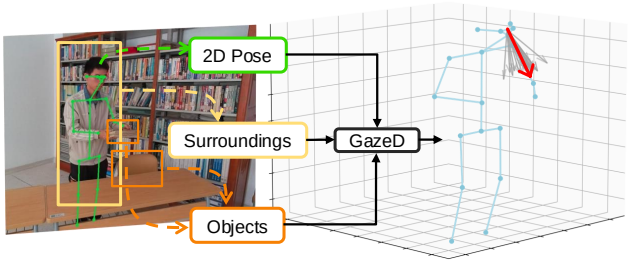

Figure 1. GazeD method jointly predicts 3D gaze and body pose analyzing the 2D pose, the surroundings of the subject and the context, in terms of objects in the scene.

Methods for 3D gaze estimation are often based on the availability or extraction of detailed information about the human face or upper body [5, 14, 19, 29], ranging from the positions of the pupils to the exact location of the eyes. Instead, only a few methods take advantage of the context, and even fewer works attempt to combine it with the human pose [58]. These elements are used more often for gaze target detection [17, 59], as they are required to relate the target of the gaze to the elements in the scene.

However, we believe that the scene context and the human pose contain knowledge useful also for 3D gaze estimation: the context influences the gaze, and the gaze itself strictly depends on the body pose. Indeed, previous studies have shown that gaze direction and body pose are closely interrelated [25]. Some works [41, 58] utilize the 2D pose or the head or body orientation to estimate the 3D gaze. However, these methods lack a mechanism to directly correlate the pose with the final gaze output.

Therefore, in this paper, we introduce GazeD, that efficiently combines different elements from the scene, *i.e.* the 2D body pose, the subject's surroundings, and the global context with objects, to output the 3D gaze direction (see

*Corresponding author: riccardo.catalini@unimore.it
†Providing contracted services for Toyota Motor Europe

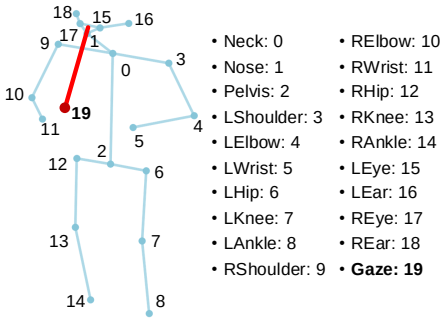

Figure 2. Human body skeleton with our additional gaze joint.

Fig. 1). A key idea of GazeD is to model the gaze as a virtual protrusion from the forehead of the person, between the eyes, where we placed **an additional joint** here referred to as **gaze joint** (see Fig. 2). The gaze joint has a variable direction (the gaze angle), while its distance from the head is fixed. Therefore, the gaze joint is not positioned in correspondence with the target object, as required to solve the gaze target detection task, but it acts as a proxy to compute the gaze direction.

Having modeled the gaze direction using an additional joint enables the resolution of the problem as an extension of 3D human pose estimation. GazeD is thus based on a regression head, which outputs both pose and gaze working on a common embedding.

Because of the intrinsic ambiguity of lifting 2D information to the 3D world, as well as the multiple possible gaze directions given the body posture and the context of the scene, GazeD regresses the 3D gaze and pose using a diffusion model. By conditioning the denoising process using 2D pose, surroundings, and context features, GazeD models the uncertainty in the data and generates multiple plausible output hypotheses. To our knowledge, we are the first to adopt a diffusion model to regress 3D gaze direction.

The embeddings used as conditioning for the diffusion model are generated by GazeD as two consecutive steps. Starting from the 2D pose estimated by an off-the-shelf method, the first step recovers information from the context close to the subject. The second step extracts additional cues from the objects in the scene, *i.e.* it captures the context of the scene far from the subject.

As an additional advantage, GazeD works on a single RGB image, avoiding the computational complexity of processing video sequences and the need for specific hardware. This streamlines its adoption in real-world applications, simplifies the training procedure, and facilitates the acquisition of new datasets. In contrast, 3D estimators based on sequences of frames [18, 27, 41] or specific data modalities, such as depth maps or point clouds [14, 24, 58], have more limited applicability in real-world scenarios.

In summary, the contributions of our paper are: i) We introduce GazeD, a method for 3D gaze estimation that combines surroundings, context with objects, and 2D human pose features to condition a diffusion model. The denoising process produces multiple plausible hypotheses of 3D gaze and human pose. ii) We propose a novel representation of the gaze as an additional joint; as a consequence, the proposed method neatly outputs both the 3D gaze and the 3D human pose. iii) Experimental evaluations on several datasets demonstrate that GazeD achieves state-of-the-art performance in 3D gaze estimation, even surpassing methods that use multiple input modalities. Additionally, our method also achieves high accuracy in predicting the 3D human pose.

2. Related Work

2.1. 3D Gaze Estimation

Recently, research in 3D gaze estimation has evolved significantly. Approaches are broadly categorized into two categories [41]: geometry-based and appearance-based ones.

Geometry-based methods. These methods [20, 33, 37, 77] rely on constructing a 3D model of the eye using optical or geometric properties. These techniques are accurate in controlled environments (*e.g.* good light conditions [40]) with consistent subject characteristics (*e.g.* head position [61]). Unfortunately, they often require specialized and expensive hardware – such as infrared cameras or eye-tracking devices – extensive calibration, limiting their applicability in real-world settings.

Appearance-based methods. These methods [5, 13, 68, 69] have gained popularity due to their reliance on standard RGB cameras to estimate the gaze from eye and face images directly. Early appearance-based approaches used hand-crafted features, such as pixel intensity or eye shape, but suffered from limited robustness in unconstrained environments. The advent of deep learning has significantly improved the performance of these methods, enabling more robust gaze estimation across varying lighting conditions, head poses, and subjects [6].

Some methods [14, 23, 24, 58] use RGB and depth data to recover scene depth, but this requires specialized hardware and is less suitable for outdoor use due to limitations like sunlight interference [49]. Alternatively, temporal information modeled with RNNs or LSTMs has improved gaze estimation by capturing movement patterns [41, 45, 75], though such approaches demand significant computational resources to handle long video sequences.

2.2. 3D Human Pose Estimation

3D Human Pose Estimation (HPE) typically involves estimating 2D poses and then lifting them to 3D [9, 11], a step that remains challenging due to the ambiguity of inferring 3D from 2D [2, 4]. To address this, some methods use

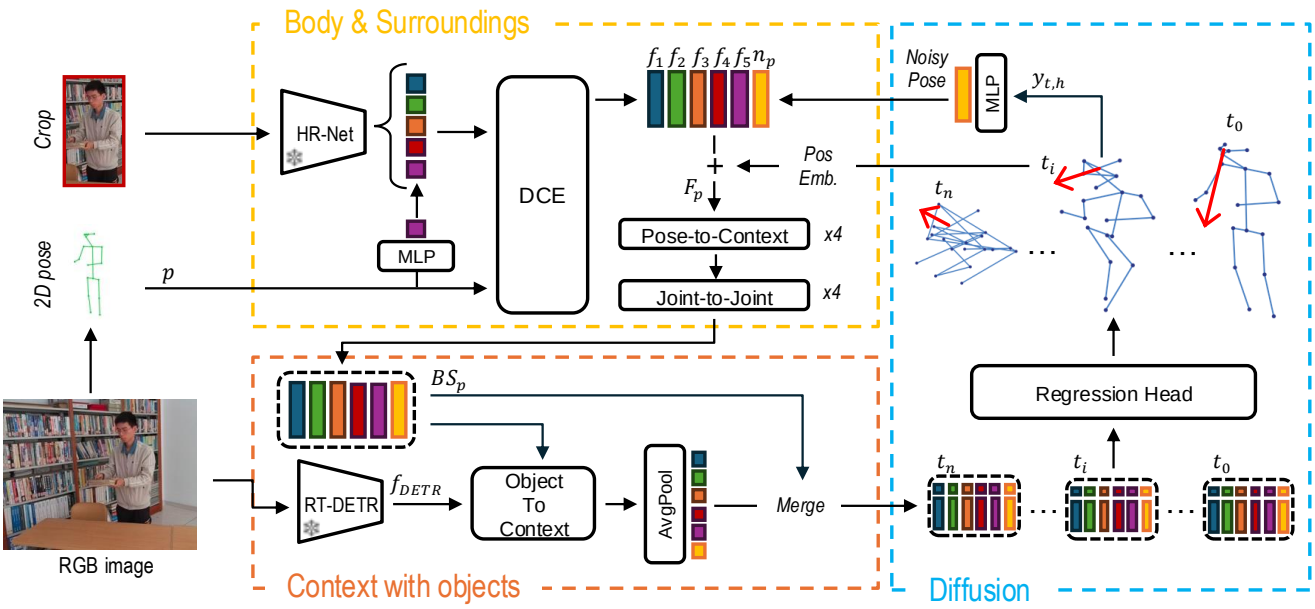

Figure 3. Overview of the proposed GazeD method that predicts the 3D gaze and human pose starting from a single input RGB image, combining information from the 2D body pose, surroundings, and context with objects.

temporal information [34, 74], although this adds latency. Given the under-constrained nature of the problem [53], multihypothesis approaches generate multiple plausible 3D poses instead of a single estimate, using techniques such as mixture density networks [42] or conditional variational autoencoders [52].

Diffusion Models. More recently, Denoising Diffusion Probabilistic Models [21] have been applied to 3D HPE. These models treat 3D pose estimation as a reverse diffusion process, where a highly uncertain 3D pose distribution is progressively refined toward a more accurate pose. Methods like DiffPose [15] leverage spatial-temporal context from 2D pose sequences to guide this diffusion process. A key advantage of diffusion models in this context is their ability to generate multiple hypotheses, providing a probabilistic framework naturally, and this allows for improved performance by aggregating multiple outputs, effectively reducing the impact of outliers. Furthermore, graph convolutional neural networks have been integrated with diffusion models [7] to explicitly capture the correlations between joints, enhancing pose estimation accuracy.

3. Method

Given a single RGB image $I \in \mathbb{R}^{H \times W \times 3}$ as input, our goal is to predict the 3D gaze direction together with the 3D pose of the person in the scene. To this end, we define an additional gaze joint and we concatenate it to the list of body joints to provide the output $\mathbf{y} \in \mathbb{R}^{J \times 3}$, where J is the number of skeleton's joints, including the gaze joint (see Fig. 2). The gaze unit vector v is defined as the direction from the midpoint between the eyes and the gaze joint:

$$v = \left\| \mathbf{y}^{Gaze} - \left(\frac{\mathbf{y}^{LEye} + \mathbf{y}^{REye}}{2} \right) \right\| \quad (1)$$

As shown in Figure 3, GazeD is composed of three modules: **Body & Surroundings**, responsible for the feature extraction from the human pose and surroundings, **Context with objects** to integrate the general context, including the location of objects, and a **Diffusion** module that contains a regression head for the 3D gaze and pose prediction as well as the diffusion scheduler.

3.1. Body & Surroundings

This module takes as input a cropped image of the person, along with its 2D pose $\mathbf{p} \in \mathbb{R}^{J \times 2}$. The pose $\mathbf{p}$ is obtained by concatenating the output of a 2D pose estimator with an additional joint representing the 2D gaze. However, as the gaze joint is virtual and lacks a correspondence in the image, we set the 2D gaze point as the midpoint between the eyes. Although it is not the 2D projection of the corresponding 3D gaze joint, such a point has proven to be a good and consistent initialization. For this joint, it will not only be necessary to perform a third-dimensional lifting, but all three coordinates must be correctly estimated by the method.

We use a HR-Net[56] backbone to extract intermediate hierarchical features $\mathcal{H} = \{\mathbf{H}_l \in \mathbb{R}^{H_l \times W_l \times C_l}\}_{l=1}^{L}$, where L is the number of feature maps ($L = 4$ in our experiments).

Deformable Surroundings Extraction. As shown in [38, 71], it is possible to encode fine-grained visual cues – *i.e.*, the joint locations – and extract high-level semantics, *i.e.*, the spatial configuration of the joints, via the high- and low-level features of a stacked network based on down-sampling operations [56, 66]. Therefore, following [71], we leverage a Deformable Context Extraction (DCE) module, based on the deformable attention mechanism [76]. DCE extracts spatial contextual cues from the intermediate feature maps using the initial 2D pose joints as reference points. Linear projections of the 2D input poses are concatenated to the hierarchical features $\mathcal{H}$ as an additional channel.

The output of the DCE module $F'_p \in \mathbb{R}^{(L+1)\times J\times d}$ is an embedding containing near context (*i.e.*, surroundings) and body pose features. We fixed $d = 128$ in our experiments. A linear projection of the noisy 3D poses at the current timestep coming from the diffusion scheduler (see Sect. 3.3) is then concatenated to F'_p. Moreover, a positional encoding of the diffusion timestep is added in order to generate $F_p \in \mathbb{R}^{(L+2)\times J\times d}$ and to make the model aware of the current diffusion step.

Pose-to-Context and Joint-to-Joint Modules. Drawing inspiration from multi-modality models [1, 31] that employ a transformer encoder, we use a similar architecture to learn a joint representation. F_p is a multichannel descriptor, which contains two channels for the pose and L channels for the context. The Pose-to-Context Attention Module performs a self-attention among the $L + 2$ descriptors (tokens) of size d for each joint. The Joint-to-Joint Attention Module considers J tokens of size $d' = d \cdot (L + 2)$ and computes self-attention among them. The Pose-to-Context module enriches the embeddings of each joint with contextual information, while the Joint-to-Joint module enables data sharing between the different joints. $BS_p \in \mathbb{R}^{d'\times J}$ is the final output of the Body & Surroundings module, with a descriptor of size d' for each joint.

3.2. Context with Objects

The goal of this module is to extract information from the whole image, particularly focusing on elements that can affect or guide the person's gaze – *i.e.*, the objects. To this aim, we use a DETR-like object detector, which is able to provide a descriptor of the objects in the image, with knowledge of both their location and their class. Let $F_{DETR} \in \mathbb{R}^{Q\times d'}$ be the last hidden states (removing the localization and classification heads and projecting to the common size d') of the detector obtained with Q input queries. The Object-To-Context block performs a cross-attention between F_{DETR} and BS_p. An average pooling operation is applied along the query dimension to merge all the important information related to the scene objects. The obtained embedding $CO \in \mathbb{R}^{d'}$ is merged with BS_p^{Gaze} to generate the final Pose&Gaze embedding $PG \in \mathbb{R}^{d'\times J}$.

3.3. Diffusion-based Multi-hypothesis Generation

Estimating the 3D gaze and pose of people from RGB is inherently challenging. Major issues are the partial or complete occlusion of the eyes and the lack of depth information. Therefore, we propose the use of diffusion models to estimate the gaze direction, as their ability to generate multiple plausible hypotheses based on the person's pose and contextual information becomes highly valuable. By modeling various potential gaze directions, the diffusion process accommodates the inherent uncertainties and ambiguities.

Representing pose and gaze direction using a single skeleton with an additional joint brings two advantages. First, it enables the formalization of the global inference process as a denoising task, starting from a completely random pose sampled from a unique Gaussian distribution. Second, it simplifies the optimization function: we adopted a single MSE loss between the predicted and the ground-truth joint coordinates, implicitly incorporating and standardizing the contributions of the pose and the gaze.

The iterative denoising procedure can be applied in parallel to H initial hypotheses $\hat{y}_{N,h} \sim \mathcal{N}(0; 1)$ in order to generate H final predictions $\hat{y}_{0,h} \in \mathbb{R}^{J\times 3}$ after N denoising iterations. For efficient inference, we employ the optimized DDIM [55] denoising scheduler. A regression head is included in the diffusion module, and it is trained to perform the denoising task.

Gaze and Pose aggregation. GazeD generates H hypotheses of the gaze and the pose, each one representing a plausible 3D solution. The distribution itself contains additional information about the real gaze (and posture). Therefore, a correct aggregation of the generated hypotheses allows for reducing the prediction error of a single hypothesis.

As the aggregation function A, we adopt the average operation (**AVG**) at joint level. Despite its simplicity, this aggregation has proven to be effective and accurate, as reported in experiments. For the sake of completeness, we also compute the "Supervision from an Oracle" [52] aggregation ($\textbf{ORC}_\text{G}$) that selects the closest hypothesis with respect to the ground truth annotation. This aggregation is useful to highlight the upper-bound performance of the proposed method, but it is limited in its applicability when ground truth annotations are not available. Additional aggregation functions based on oracle are investigated in Section 4.6. We avoided using additional aggregation techniques that required ground-truth information or camera calibration parameters [50], which are not always available or predictable in single-frame methods.

4. Experimental Evaluation

4.1. Datasets

As GazeD is based on the rich information extracted from the surroundings and global context, we focus on datasets

Method	Office	Living Room	Kitchen	Library	Courtyard	All
Fixed bias	88.0/76.0	85.5/76.7	86.0/82.4	89.0/85.1	89.7/88.7	88.1/79.7
Frontal gaze	22.6/21.9	36.6/35.4	**17.9**/19.6	27.1/25.8	30.5/33.8	28.8/28.8
Dias *et al.* [10]	−/27.2	−/25.2	−/19.8	−/24.9	−/36.1	−/27.1
XGaze [69]	24.2/23.0	42.0/40.9	23.3/22.9	24.6/22.3	30.2/31.9	29.2/28.4
Nonaka *et al.* [41]	20.0/18.1	25.6/25.5	21.5/18.6	21.9/20.1	28.4/30.5	24.1/23.3
Gaze360 [29]†	24.0/19.2	41.1/31.3	32.4/21.2	27.5/20.7	28.2/28.3	30.4/24.5
Nonaka *et al.* [41]†	**14.4**/**14.3**	25.1/22.6	20.4/19.6	19.8/18.4	25.4/**26.9**	21.7/20.9
Ours$_{(H=20,\ A=AVG)}$	15.8/16.3	**19.3**/**20.6**	18.2/**19.5**	**17.6**/**16.9**	**25.3**/29.1	**19.5**/**20.5**
Ours$_{(H=20,\ A=ORC_G)}$	11.6/11.6	13.2/13.8	14.6/13.7	14.2/12.9	23.9/27.9	15.9/16.3

Table 1. Experimental results on GAFA dataset expressed as MAE_{3D}/MAE_{2D}. † indicates methods leveraging temporal information.

containing 3D gaze annotations and full images, thus excluding those only containing crops around the faces, eyes or body of the subject [13, 29, 69]. Additional details about datasets are reported in the Supplementary.

GAFA [41] (Gaze from Afar) dataset is designed for 3D gaze estimation in surveillance scenarios, capturing freely moving people in natural settings. It includes more than 850k video frames from 5 different daily environments. It features a wide range of head poses, including back views and high-pitch angles, reflecting realistic conditions. GAFA is annotated with 3D gaze directions and body orientations, using wearable cameras and AR marker-based positioning systems for ground truth.

GFIE [24] is a dataset introduced for 2D and 3D gaze-following tasks, created using a system that combines a laser rangefinder and an RGB-D camera to record and annotate gaze behaviors in natural indoor environments. The system guides the subject's gaze target using a laser spot, which is then detected in the RGB images to generate precise annotations, and then removed using image inpainting [60]. The 3D gaze target is reconstructed using the distance measured by the laser rangefinder and the camera's intrinsic parameters. The dataset includes about 71k frames of 61 subjects, performing a wide range of activities.

Ego-Gaze. We create this dataset starting from the multimodal Ego-Exo4D dataset [16]. Specifically, we select frames from the Ego-Pose subset in which the 3D annotation of the human pose is available. Then, we compute 3D gaze annotations from the data acquired with the Aria glasses. The dataset includes a wide range of skilled activities— such as sports, music, dance—performed in natural settings. Because the Ego-Pose dataset is still used for competitions, the official test set has not been made available. Therefore, we use the official validation split as test set and we sample validation instances from the training set. We will release splits, enabling future comparisons.

Method	RGB	Crops	Depth	MAE_{3D}
Random				84.4
Center				87.2
GazeFollow [48]	✓	✓		41.5
Lian *et al.* [35]	✓	✓		26.7
Rt-Gene [13]	✓	✓		21.0
Hu *et al.* [24]	✓	✓	✓	17.7
Toaiari *et al.* [58]			✓	15.9
Chong *et al.* [8]†	✓	✓		20.8
Gaze360 [29]†	✓			19.8
Ours$_{(H=20,\ A=AVG)}$	✓			**13.6**
Ours$_{(H=20,\ A=ORC_g)}$	✓			9.9

Table 2. Quantitative results on GFIE dataset. For each method, the input data is reported: **RGB** for color images, **Crops** for head, face, or eye crops, and **Depth** for depth maps. † indicates methods leveraging temporal information.

Method	Basket	Dance	Various	All
XGaze	21.6/20.6	23.8/27.1	21.6/20.5	23.1/24.9
Gaze360	21.8/17.0	21.0/21.8	22.1/17.6	21.3/20.4
Ours	**15.4**/**14.7**	**18.6**/**18.9**	**15.3**/**12.7**	**17.5**/**17.4**

Table 3. Quantitative results on the Ego-Gaze dataset. GazeD is tested with H=20, A=AVG.

4.2. Implementation Details and Training

As backbones, we use different pre-trained models. For 2D human pose estimation, we use HRNet [56], capable of maintaining high-resolution representations throughout the whole architecture and achieving great accuracy. As an object detector, we use RT-DETR [73], a recent end-to-end architecture with good accuracy and real-time performance for the object detection task from single RGB images. Both models are frozen during the training phase and are used

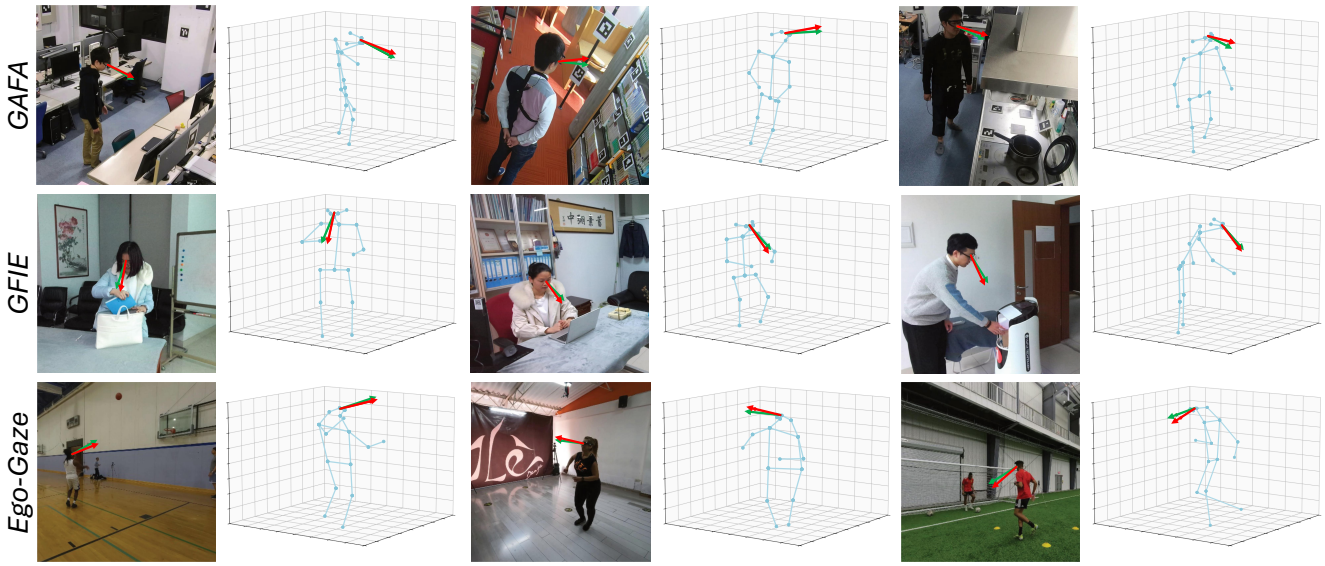

Figure 4. Qualitative Results on the three datasets. Green arrow represents ground truth gaze direction, red arrow is the prediction

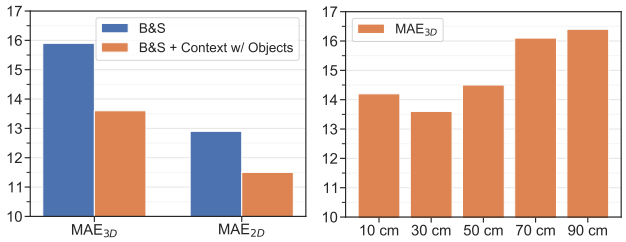

Figure 5. Ablation study on GFIE dataset: (a) the contribution of the "Context with objects" module in addition to the "Body&Surroundings". (b) Performance of GazeD in terms of MAE_{3D} by varying the distance of the gaze joint from the head.

with their original weights and parameters.

We train GazeD with a batch size of 64 for 100 epochs on all datasets. We use Adam optimizer [32] with a starting learning rate of $6e^{-4}$ using a linear decay with factor 0.993. No data augmentation is applied to input images.

4.3. Baselines and Competitors

We compare GazeD with several 3D gaze estimation baselines and competitors.

For the GAFA dataset [41], we compute two baselines. The first is *fixed bias*, *i.e.* the mean gaze direction is obtained from the training set, and the error metric is computed using this mean value over the test set [41]. This baseline is intended to show the lower bound accuracy on this dataset. The second baseline is *frontal gaze*, where we compute the angular error assuming that the predicted gaze direction is always orthogonal to the line between the two eyes. This is a useful reference for understanding the precision that a method based solely on the pose of the head would achieve. As competitors, we use a variety of state-of-the-art methods from the literature. The approach proposed by Dias *et al.* [10] estimates 2D gaze on the image plane using facial keypoints detected by OpenPose [3]. Gaze360 [29] takes a sequence of full-head images as input and provides the 3D gaze direction. XGaze [69] uses facial images as input and assumes high-resolution facial images.

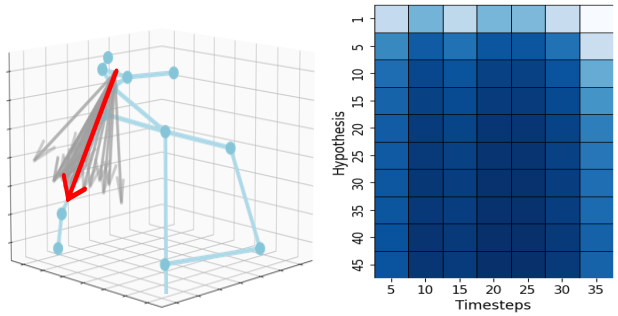

Figure 6. (a) GazeD, predicts multiple hypotheses on which aggregation functions are applied (see Sect. 3.3) (b) Investigation on the number of different hypotheses vs number of timesteps. Darker color represents a lower MAE_{3D} value.

The GFIE dataset was originally proposed for gaze target detection, and thus provides additional information on the scene – *i.e.* depth maps. For this reason, we evaluate on this dataset with other baselines and competitors. The baseline methods include the *random* approach, *i.e.* the 2D and 3D gaze directions are randomly selected within the

#	Diffusion	Objects	Near Context	MAE_{3D}	MAE_{2D}
1	✗	✓	✓	16.1	14.2
2	✓	✗	✓	15.9	12.9
3	✓	✓	✗	15.1	12.0
4	✓	✓	✓	13.6	11.5

Table 4. Ablation analysis of each component.

Aggregation method	GAFA [41]		GFIE [24]	
	MAE_{3D}	MAE_{2D}	MAE_{3D}	MAE_{2D}
AVG	19.5	20.5	13.6	11.5
ORC_P	19.7	20.4	13.4	10.8
ORC_G	15.9	16.3	9.9	7.9
ORC_J	12.6	13.3	8.7	7.5

Table 5. Impact of different aggregation methods on gaze.

image and point cloud, respectively. In addition, the *center* baseline localizes the gaze always at the center of the point cloud of the 3D space. As competitors, we also use existing 2D gaze-following methods, *i.e*. GazeFollow [48], Lian [35], and Chong [8]. To retrieve their 3D gaze angle, we first back-projected the 2D gaze target into the 3D space using the available registered depth maps. The results of Gaze360 [29] and Rt-Gene [13] are collected from [23]. Finally, we report the results of the recent work by Toaiari *et al*. [58], which utilizes upper-body skeleton data and the depth map of the scene to predict the 3D gaze.

4.4. Comparison with state-of-the-art

We report the result using the Mean Angular Error (MAE), the standard metric for the evaluation of gaze estimation methods. MAE is expressed in degrees, and it is calculated as the average of the angular difference between the predicted and ground-truth gaze directions over all the testing samples. In addition to the 3D errors (MAE_{3D}), we report the metric using the directions on the image plane (MAE_{2D}).

Table 1 reports the performance of our GazeD and other approaches. On the GAFA dataset, as shown, GazeD achieves the best performance on both metrics, even outperforming methods that leverage additional temporal information, *i.e*. [29, 41]. Both MAE_{3D} and MAE_{2D} achieved by our method are well below the *frontal gaze* baseline, indicating that the idea to model the gaze as additional joint is effective in estimating the 3D gaze direction and that the output of our method is not the mere head pose.

Table 2 reports similar results on the GFIE dataset. Also in this case, our method largely outperforms the competitors. In particular, GazeD outperforms even methods that are based on fine details, such as the face or body crops, or additional input data as depth maps, whose contribution is significant in 3D estimation tasks.

In both datasets, we also report the results obtained using the ORC_G aggregation function. As expected, these are the best results. However, the ground truth is not normally available in the inference phase, and it is not completely correct to use it to select the best hypothesis. These results show that the diffusion process can generate hypotheses close to the ground truth, suggesting future work on more sophisticated aggregation strategies.

The newly introduced Ego-Gaze dataset imposes retraining the Gaze360 and XGaze methods. Unfortunately, it was not possible to implement more recent techniques, such as the ones developed in [24, 41], due to a lack of depth maps and body or head orientation, respectively. Results are reported in Table 3, organized in three main scenes, *i.e*., basket, dance and various. The latter includes the less represented classes, such as cooking, soccer and bike repair. As shown, GazeD achieves the best results in all the scenes. These results demonstrate the robustness of the proposed approach on a challenging dataset with complex scenes.

4.5. Qualitative results

Some qualitative results are reported in Figure 4, where the input image and the predicted 3D gaze and pose are shown. The ground truth gaze direction vector is drawn as a green arrow, while the predicted vector is drawn as a red arrow. These results confirm the ability of GazeD to predict gaze direction in wide-angle ranges, also when the face is not visible or partially occluded. Additional qualitative results are reported in the Supplementary material.

4.6. Ablation Studies

Ablation studies are mainly computed on the GFIE dataset, using GazeD in the configuration described in Section 4.2.

Module contributions We investigate the contribution of each module (see Table 4). In experiment #1, we use the transformer-based model without the diffusion process, training the network to predict directly pose and gaze. In #2, we remove the module "Context with Objects" for context analysis. In #3, we remove the part of the method responsible for extracting and processing the surroundings of the person. Each module is a key part of the method.

Context with Objects module To highlight the performance improvement provided by the proposed "context with objects" module, we tested the results of GazeD directly using BS_p in input to the diffusion step (see Fig. 3). The results are reported in Figure 5a. Adding the object embeddings clearly improves the model's ability to solve the 3D gaze estimation task.

Distances of the Gaze Keypoint The gaze joint is an auxiliary point used to solve the task, but it is not physically present. Its distance from the eyes was chosen to be close enough to the body to be modeled as a joint and, at the same

Methods	Mart. [39]	Zhao [70]	Sun [57]	Yang [65]	Hoss. [22]	Liu [36]	Xu [64]	Zhao [72]	Zhao [71]	Diffu. [7]	Diffp. [15]	**Ours** (H=20, A=AVG)	**Ours** (H=20, A=ORC)
MPJPE ↓	62.9	60.8	59.1	58.6	58.3	52.4	51.9	51.8	**43.4**	<u>49.4</u>	49.7	49.7	41.1

Table 6. Results on Human3.6M dataset for the 3D Human Pose Estimation task. The best result is in **bold**, the second one is <u>underlined</u>.

Methods	Pavvlo [46]	Zheng [74]	Wang [63]	Li [34]	Zheng [74]	Zhang [67]	Zhao [71]	**Ours** (H=20, A=AVG)	**Ours** (H=20, A=ORC)
MPJPE ↓	84.0	77.1	68.1	58.0	57.7	54.9	**44.7**	<u>46.6</u>	33.8

Table 7. Results on MPI-INF-3DHP dataset for the 3D Human Pose Estimation task. The best result is in **bold**, the second one is <u>underlined</u>.

time, far enough to reduce the dependence on noise in the final conversion into angles. The measurement used, equal to 30 centimeters, is supported by an experimental analysis. MAE_{3D} errors vs distances are plotted in Figure 5b.

Number of hypotheses and timesteps A key advantage of diffusion models is their ability to generate multiple hypotheses. In Figure 6a, a real multiple-hypothesis prediction is depicted. Then, we analyze how the number of hypotheses H and the number of denoising iterations N affect the final MAE_{3D}. Figure 6b shows the matrix from which we selected the final values of H and N, where darker colors denote lower errors. Based on this analysis, we selected $H = 20$ and $N = 20$ for our evaluation, as a favorable trade-off between accuracy and computational load.

Multiple Hypothesis Aggregation Strategies Having multiple generated hypotheses allows us to explore various aggregation strategies. In Table 5, we compare the different aggregations described in Section 3.3. We also investigate different oracle selections [52] obtained in three different ways. ORC_G chooses the hypothesis with the lowest error, specifically at the gaze joint. ORC_P selects the hypothesis with the lowest Mean Per-Joint Position Error (MPJPE) relative to the ground truth; however, our results indicate that minimizing MPJPE at the pose level does not necessarily produce the most accurate gaze estimation. Finally, ORC_J employs a per-joint selection strategy in which, for each joint, the coordinates with the lowest error are independently selected, resulting in a more accurate estimation of gaze direction. Since ORC_p, ORC_g, and ORC_j rely on ground truth data, they are not applicable in real-world scenarios. Therefore, we consider **AVG** as the most appropriate baseline for fair comparison.

4.7. Additional Evaluation

As previously mentioned, GazeD predicts not only the 3D gaze, but also the 3D body pose: therefore, we also analyze performances on this task pose. Therefore, in this section, we analyze the performance of this task.

Dataset The Human3.6M dataset [26] is a well-known dataset of 3.6 million images with 3D human pose annotations. It contains 17-joint skeleton annotations for 11 subjects performing 15 activities, captured by 4 cameras in an indoor environment. For evaluation, we follow the standard protocol of training on subjects S1, S5, S6, S7, and S8, and testing on subjects S9 and S11. The MPI-INF-3DHP dataset consists of over 1.3 million frames captured from 14 cameras and it is widely used for training and evaluating 3D human pose estimation models. It contains 8 actors performing activities such as walking, sitting, and sports. The frames are annotated using a skeleton model with 17 joints.

3D Pose Evaluation For the training, we use a batch of 128 for 50 epochs. Other training settings are the same used for the gaze evaluation. Performance is evaluated using the Mean Per Joint Position Error (MPJPE) [28], which calculates the average Euclidean distance (in millimeters) between predicted and ground truth 3D joint coordinates. In Tables 6 and 7, we report the comparison for the 3D pose estimation task between our model and literature competitors, on Human3.6M and MPI-INF datasets, respectively. Among the others, Diffupose [7], Diffpose [15] are the most similar methods since they are based on a diffusion architecture. As shown, the results obtained are better than a large portion of the literature, and comparable with the most recent one. These experimental results suggest that, although our method was not specifically developed for the HPE task, it still achieves competitive results with a good level of accuracy.

5. Conclusion

We introduced GazeD, a method for 3D gaze and pose estimation from single RGB images. By modeling 3D gaze through a diffusion process, GazeD integrates 2D pose, surrounding context, and global scene cues. The use of a diffusion model addresses the inherent ambiguity of 3D gaze estimation, generating multiple plausible hypotheses. Results demonstrate the efficacy of GazeD, highlighting its potential for accurate 3D gaze and pose estimation.

References

[1] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35, 2022.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1), 2019.

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018.

[5] Yihua Cheng, Feng Lu, and Xucong Zhang. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *ECCV*, 2018.

[6] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[7] Jeongjun Choi, Dongseok Shim, and H. Jin Kim. Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. pages 3773–3780, 2023.

[8] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *CVPR*, 2020.

[9] Andrea D'Eusanio, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. Refinet: 3d human pose refinement with depth maps. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2320–2327. IEEE, 2021.

[10] Philipe Ambrozio Dias, Damiano Malafronte, Henry Medeiros, and Francesca Odone. Gaze estimation for assisted living environments. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020.

[11] Andrea D'Eusanio, Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara. Depth-based 3d human pose refinement: Evaluating the refinet framework. *Pattern Recognition Letters*, 171:185–191, 2023.

[12] Maria K Eckstein, Belén Guerra-Carrillo, Alison T Miller Singley, and Silvia A Bunge. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental cognitive neuroscience*, 25, 2017.

[13] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *ECCV*, 2018.

[14] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, 2014.

[15] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *CVPR*, 2023.

[16] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024.

[17] Jian Guan, Liming Yin, Jianguo Sun, Shuhan Qi, Xuan Wang, and Qing Liao. Enhanced gaze following via object detection and human pose estimation. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*. Springer, 2020.

[18] Yiran Guan, Zhuoguang Chen, Wenzheng Zeng, Zhiguo Cao, and Yang Xiao. End-to-end video gaze estimation via capturing head-face-eye spatial-temporal interaction context. *IEEE Signal Processing Letters*, 30, 2023.

[19] Zidong Guo, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. Domain adaptation gaze estimation by embedding with prediction consistency. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[20] Craig Hennessey, Borna Noureddin, and Peter Lawrence. A single camera eye-gaze tracking system with free head motion. In *Proceedings of the 2006 symposium on Eye tracking research & applications*, 2006.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 2020.

[22] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *ECCV*, 2018.

[23] 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*, 71, 2022.

[24] Zhengxi Hu, Yuxue Yang, Xiaolin Zhai, Dingye Yang, Bohan Zhou, and Jingtai Liu. Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In *CVPR*, 2023.

[25] Zhiming Hu, Jiahui Xu, Syn Schmitt, and Andreas Bulling. Pose2gaze: Eye-body coordination during daily activities for gaze prediction from full-body poses. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7), 2013.

[27] Swati Jindal, Mohit Yadav, and Roberto Manduchi. Spatio-temporal attention and gaussian processes for personalized video gaze estimation. In *CVPR*, 2024.

[28] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *CVPR*, 2015.

[29] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, 2019.

[30] Jess Kerr-Gaffney, Amy Harrison, and Kate Tchanturia. Eye-tracking research in eating disorders: A systematic review. *International Journal of Eating Disorders*, 52(1), 2019.

[31] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*. PMLR, 2021.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Ji Woo Lee, Chul Woo Cho, Kwang Yong Shin, Eui Chul Lee, and Kang Ryoung Park. 3d gaze tracking method using purkinje images on eye optical model and pupil. *Optics and Lasers in Engineering*, 50(5), 2012.

[34] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, pages 13147–13156, 2022.

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

[36] Kenkun Liu, Rongqi Ding, Zhiming Zou, Le Wang, and Wei Tang. A comprehensive study of weight sharing in graph networks for 3d human pose estimation. In *ECCV*, 2020.

[37] Feng Lu, Yue Gao, and Xiaowu Chen. Estimating 3d gaze directions using unlabeled eye images via synthetic iris appearance fitting. *IEEE Transactions on Multimedia*, 18(9), 2016.

[38] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016.

[39] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[40] Atsushi Nakazawa and Christian Nitschke. Point of gaze estimation through corneal surface reflection in an active illumination environment. In *ECCV*. Springer, 2012.

[41] Soma Nonaka, Shohei Nobuhara, and Ko Nishino. Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *CVPR*, 2022.

[42] Tuomas Oikarinen, Daniel Hannah, and Sohrob Kazerounian. Graphmdn: Leveraging graph structure and deep learning to solve inverse problems. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.

[43] Anwesan Pal, Sayan Mondal, and Henrik I Christensen. "looking at the right stuff"-guided semantic-gaze for autonomous driving. In *CVPR*, 2020.

[44] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. Eye gaze tracking for a humanoid robot. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015.

[45] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. *arXiv preprint arXiv:1805.03064*, 2018.

[46] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019.

[47] George E Raptis, Christina Katsini, Marios Belk, Christos Fidas, George Samaras, and Nikolaos Avouris. Using eye gaze data and visual activities to infer human cognitive styles: method and feasibility studies. In *proceedings of the 25th conference on user modeling, Adaptation and Personalization*, 2017.

[48] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Advances in neural information processing systems*, 28, 2015.

[49] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. Kinect range sensing: Structured-light versus time-of-flight kinect. *Computer vision and image understanding*, 139, 2015.

[50] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *ICCV*, 2023.

[51] Anjana Sharma and Pawanesh Abrol. Eye gaze techniques for human computer interaction: A research survey. *International Journal of Computer Applications*, 71(9), 2013.

[52] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *ICCV*, 2019.

[53] Edgar Simo-Serra, Arnau Ramisa, Guillem Alenya, Carme Torras, and Francesc Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *CVPR*. IEEE, 2012.

[54] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4), 2018.

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

[56] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[57] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, 2017.

[58] Andrea Toaiari, Vittorio Murino, Marco Cristani, and Cigdem Beyan. Upper-body pose-based gaze estimation for privacy-preserving 3d gaze target detection. *arXiv preprint arXiv:2409.17886*, 2024.

[59] Francesco Tonini, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Object-aware gaze target detection. In *ICCV*, 2023.

[60] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, 2018.

[61] Roberto Valenti, Nicu Sebe, and Theo Gevers. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2), 2011.

[62] Ulas Vural and Yusuf Sinan Akgul. Eye-gaze based real-time surveillance video synopsis. *Pattern Recognition Letters*, 30 (12), 2009.

[63] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *ECCV*, 2020.

[64] Tianhan Xu and Wataru Takano. Graph stacked hourglass networks for 3d human pose estimation. In *CVPR*, 2021.

[65] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018.

[66] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in neural information processing systems*, 34, 2021.

[67] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *CVPR*, 2022.

[68] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1), 2017.

[69] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *ECCV*. Springer, 2020.

[70] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, 2019.

[71] Qitao Zhao, Ce Zheng, Mengyuan Liu, and Chen Chen. A single 2d pose with context is worth hundreds for 3d human pose estimation. In *NeurIPS*, 2023.

[72] Weixi Zhao, Weiqiang Wang, and Yunjie Tian. Graformer: Graph-oriented transformer for 3d pose estimation. In *CVPR*, 2022.

[73] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *CVPR*, 2024.

[74] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, 2021.

[75] Xiaolong Zhou, Jianing Lin, Jiaqi Jiang, and Shengyong Chen. Learning a 3d gaze estimator with improved itracker combined with bidirectional lstm. In *ICME*. IEEE, 2019.

[76] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2020.

[77] Zhiwei Zhu and Qiang Ji. Eye gaze tracking under natural head movements. In *CVPR*. IEEE, 2005.