# Saliency Prediction on Omnidirectional Images with Generative Adversarial Imitation Learning

Mai Xu

School of Electronic and Information Engineering, Beihang University
Beijing, 100191

Li Yang

School of Electronic and Information Engineering, Beihang University
Beijing, 100191

## Abstract

*When watching omnidirectional images (ODIs), subjects can access different viewports by moving their heads. Therefore, it is necessary to predict subjects' head fixations on ODIs. Inspired by generative adversarial imitation learning (GAIL), this paper proposes a novel approach to predict saliency of head fixations on ODIs, named SalGAIL. First, we establish a dataset for attention on ODIs (AOI). In contrast to traditional datasets, our AOI dataset is large-scale, which contains the head fixations of 30 subjects viewing 600 ODIs. Next, we mine our AOI dataset and determine three findings: (1) The consistency of head fixations are consistent among subjects, and it grows alongside the increased subject number; (2) The head fixations exist with a front center bias (FCB); and (3) The magnitude of head movement is similar across subjects. According to these findings, our SalGAIL approach applies deep reinforcement learning (DRL) to predict the head fixations of one subject, in which GAIL learns the reward of DRL, rather than the traditional human-designed reward. Then, multi-stream DRL is developed to yield the head fixations of different subjects, and the saliency map of an ODI is generated via convoluting predicted head fixations. Finally, experiments validate the effectiveness of our approach in predicting saliency maps of ODIs, significantly better than 10 state-of-the-art approaches.*

## 1. Introduction

In recent years, omnidirectional images (ODIs) have become increasingly popular, along with the rapid development of virtual reality (VR). Different from the traditional 2D images, ODIs provide an immersive and interactive VR viewing experience. Moreover, they enable spherical stim-uli, meaning that the range of $360° \times 180°$ can be accessible to subjects through the head-mounted display (HMD). In other words, humans can freely move their heads to change viewports for viewing attractive regions. Hence, head fixations play a vital role in modeling visual attention on ODIs. Accordingly, it is necessary to predict head fixations, which can be widely used in many applications of ODIs, e.g., compression [12], rendering [44] and visual quality assessment [13].

Most recently, there have emerged several works [3, 27, 30, 59] on predicting the saliency maps of head fixations on ODIs. For example, Lebreton *et al.* [27] developed two new models for saliency prediction of head fixations on ODIs, that are based on the traditional 2D saliency prediction models: Boolean Map based Saliency model (BMS) [55] and Graph-Based Visual Saliency (GBVS) [15]. Thus, these models are called "BMS360" and "GBVS360", respectively. Zhu *et al.* [59] proposed a multi-plane projection method to predict head fixations on omnidirectional scene, in which several blocks are generated to simulate viewports. Then, the low-level features (spatial frequency, orientation and color) and high-level semantic features (car and person) are extracted in each block, which are further fused and mapped into the overall saliency map. The above works can be seen as heuristic methods, as their features for predicting head fixations are hand-crafted. In fact, the great success of deep learning has boosted the development of saliency prediction on 2D images, which is a closely related area of head fixation prediction on ODIs. However, none of the existing saliency prediction approaches for ODIs is based on deep learning; therefore, their performance is fair.

Further complication are that the existing head fixation datasets for ODIs are all small-scale collections, which can hardly be used to train the deep learning models. Specifically, [37] is the first ODI dataset with human attention, which is composed of the head and eye fixations of 63 sub-

Table 1: Basic properties of the existing ODI/ODV datasets.

| | Dataset | Scene | Images/videos | Subjects | Resolution | Durations (s) | Ground-truth recorded | HMD/Eye-tracker |
|---|---|---|---|---|---|---|---|---|
| IMAGE | Rai *et al.* [37] | Static | 98 | 40 | ≤18,332×9,166 | 25 | Head and eye fixations | Oculus Rift DK2/SMI Eye-tracker |
| | Sitzmann *et al.* [42] | Static | 22 | 169 | ≤ 8,192×4,096 | 30 | Head and eye fixations | Oculus Rift DK2/Tobii EyeX Eye-tracker |
| | Upenik *et al.* [46] | Static | 104 | 40 | 1,334×750 | - | Head fixations | MERGE VR Goggles[1] plus iPhone 6 |
| | Hu *et al.* [17] | Static | 70 | 27 | 640×480 | 10 | Head fixations | Google Cardboard |
| | Abreu *et al.* [11] | Static | 21 | 32 | 4,096×2,048 | 10/20 | Head fixations | Oculus Rift DK2 |
| VIDEO | Yu *et al.* [54] | Dynamic | 10 | 10 | ≤ 6,144×3,072 | 10 | Head fixations | Oculus Rift DK2 |
| | Lo *et al.* [31] | Dynamic | 10 | 50 | 4,096×2,048 | 60 | Head fixations | Oculus Rift DK2 |
| | Xu *et al.* [52] | Dynamic | 208 | 31 | 4,096×2,048 | 20-60 | Eye fixations | HTC Vive/aGlass Eye-tracker |
| | Zhang *et al.* [56] | Dynamic | 104 | 27 | - | 20-60 | Head and eye fixations | HTC Vive/aGlass Eye-tracker |
| | Ozcinar *et al.* [34] | Dynamic | 6 | 17 | ≤ 8,192×4,096 | 10 | Head fixations | WebVR [49] |
| | Corbillon *et al.* [8] | Dynamic | 7 | 59 | 3,840×2,048 | 70 | Head fixations | Razer OSVR HDK2 HMD |
| | Xu *et al.* [51] | Dynamic | 76 | 58 | ≤ 8,192×4,096 | 10-80 | Head and eye fixations | HTC Vive/aGlass Eye-tracker |
| | Deep 360 Pilot [18] | Dynamic | 342 | 5 | - | - | Annotate salient object in panorama | Without using HMD |
| | David [10] | Dynamic | 19 | 57 | ≤3,840×1,920 | 20 | Head and eye fixations | Oculus Rift DK2/SMI Eye-tracker |
| | **Our dataset** | Static | 600 | 30 | ≤24,028×12,014 | 22 | Head and eye fixations | HTC Vive/aGlass Eye-tracker |

jects on 98 ODIs. In addition, both the head and eye fixations of 169 subjects on 22 ODIs are available in the dataset of [42]. In [46], the dataset has 104 ODIs viewed by 40 subjects. However, only the head fixations data are available in [46] without any eye fixations data. Moreover, the dataset of [17] includes the attention data of 27 subjects who were asked to view a total of 70 different ODIs in the VR environment.

In this paper, we establish a large-scale dataset for attention on ODIs (called the AOI dataset), which is comprised by head and eye fixations data of 30 subjects viewing 600 ODIs. Note that the 600 ODIs of our AOI dataset are diverse in both the resolution and content. By mining our dataset, we find that high consistency exists for head fixations among subjects when viewing ODIs. Besides, we have some additional findings. (1) The distribution of head fixations are variant between individual subjects; however, the consistency of head fixations tends to increase and converge when the number of subjects increases. (2) The front center bias (FCB) characteristic exists for the head fixations. (3) The magnitude of head movement (HM) is similar across all subjects over all ODIs in our AOI dataset. Based on the above findings, this paper proposes a generative adversarial imitation learning (GAIL) based approach for saliency prediction of head fixations on ODIs, which is called SalGAIL.

Specifically, our SalGAIL approach predicts the head fixations through a deep reinforcement learning (DRL) model. In the DRL model, we regard the directions of head trajectories as the *actions* of the DRL model and take the viewed omnidirectional content as the *observation* of the *environment*. As such, the DRL model can be learned to predict the head fixations of one subject on an ODI. Then, multi-stream DRL is used to generate the head fixations of different subjects, and the predicted head fixations are convoluted to generate the saliency map of the input ODI. However, different from the traditional DRL tasks, the *reward* is intractable to be obtained and quantified in our task for saliency prediction on ODIs. Instead, we propose to learn *reward* by imitating the head trajectories of subjects in the
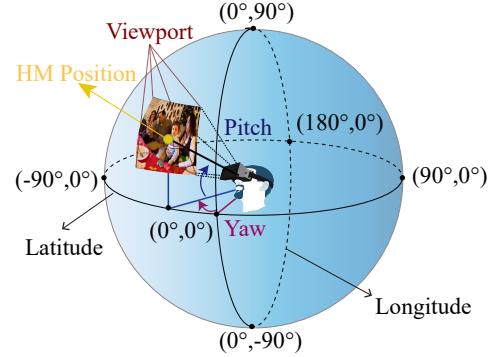


Figure 1: An example of HM in the sphere. The latitude and longitude of the HM position, i.e., the center of a viewport, are only dependent on the angles of pitch and yaw, respectively.

training stage. This strategy benefits from the most recent success of GAIL [16, 29].

In brief, the main contributions of this paper are summarized as follows.

- We establish a large-scale AOI dataset with several findings about human attention on ODIs.

- We propose a multi-stream DRL model to predict head fixations on ODIs.

- We apply GAIL to learn the *reward* of our DRL model by imitating the head trajectories of subjects.

## 2. Related Works

In this section, we review the approaches and datasets for saliency prediction on ODIs.

### 2.1. Saliency prediction approaches for ODIs

**Saliency prediction on 2D images.** The past two decades have witnessed extensive works on saliency prediction for 2D images, such as [55], [7, 14, 22, 23, 39, 53] . In

Figure 2: Some examples of ODIs in our AOI dataset.

the task of image saliency prediction, many effective spatial features have been proposed in predicting human attention with either a top-down or a bottom-up strategy. Specifically, Itti *et al.* [20] considered low-level features at multiple scales and combined them to form the saliency map of an image. Harel *et al.* [15] introduced a graph-based visual saliency (GBVS) model that defines Markov chains over various image maps, and treated the equilibrium distribution over map locations as activation and saliency values. Considering top-down image semantics, Judd *et al.* [22] proposed a saliency model based on low-, middle- and high-level image features. Moreover, Borji *et al.* [4] proposed combining low-level features of the bottom-up models with top-down cognitive visual features, and then learning a direct mapping from those features to eye fixations. Inspired by deep learning, deep neural networks (DNNs) have been successfully used to predict image saliency in an end-to-end manner, such as [9, 19, 25, 26, 35, 36, 41, 57]. Specifically, a DNN-based structure was proposed in Deepfix [26] to learn a multiscale semantic representation for image saliency. Moreover, saliency in context (SALICON) was proposed in [19], which fine-tunes the existing DNNs with an effective saliency-related loss function. In [25], a readout architecture was proposed for image saliency prediction, in which both low-level and DNN features are considered.

**Saliency prediction of eye fixations on ODIs.** Although saliency prediction has been well developed for 2D images, there are only a few approaches for predicting the saliency maps of ODIs. Different from 2D images, the saliency of ODIs refers to two forms: head fixations and eye fixations. Most of the existing saliency prediction approaches for ODIs focus on eye fixations, including [1, 3, 27, 30, 32, 43, 59]. In particular, Battisti *et al.* [3] presented a saliency model for predicting the saliency maps of eye fixations on ODIs, which is based on the combina-

tion of low-level and semantic features. Startsev *et al.* [43] proposed a new saliency prediction approach by considering projection distortions, equator bias and vertical border effects, for predicting saliency of eye fixations on ODIs. In addition, Ling *et al.* [30] took human color perception into account and proposed a model using color dictionary-based sparse representation for ODI saliency prediction. Besides, DNNs have also been successfully applied [1, 32] for saliency prediction on ODIs. "SalNet360" [32] was proposed to fine-tune traditional CNN models of 2D saliency prediction for the task of ODI saliency prediction. Additionally, "SaltiNet" [1] was developed to train a DNN model for eye fixation prediction on ODIs, which is based on a temporal-aware representation of saliency information called saliency volume.

**Saliency prediction of head fixations on ODIs.** In particular, there are relatively few works [11, 27, 59] to predict the saliency maps of head fixations for ODIs. Zhu *et al.* [59] employed a method of multiview projection to generate the saliency maps of head fixations on ODIs. In their work, an ODI is first projected into multiview blocks to simulate viewports. Then, both bottom-up and top-down features of all blocks are extracted and fused to generate the final saliency map of head fixations. In addition, [11] proposed adding the center bias of human attention into the saliency maps of ODIs through a postprocessing method. Lebreton *et al.* [27] developed new models for saliency prediction on ODI, called "BMS360" and "GBVS360", which are based on the traditional 2D saliency prediction models, boolean map based saliency model (BMS) [55] and GBVS [15]. More specifically, "BMS360" applied multiple fusion saliency (FMS) [11] to remove the border constraints. In "GBVS360", the input ODI in equirectangular format is projected into several rectilinear images, corresponding to different viewports, and then feature extraction is performed according to each rectilinear image. Finally, the resulting

4323

feature maps are back-projected to the equirectangular domain to yield the saliency map. In contrast to ODIs, more works [18, 51, 52, 56] have been proposed for predicting head fixations on omnidirectional videos (ODVs). For example, Xu *et al.* [51] proposed a DRL approach for saliency prediction of head fixations on ODVs. Additionally, Zhang *et al.* [56] presented a spherical CNN-based scheme for saliency prediction of ODVs. In the following, we overview the existing datasets for attention modeling on ODIs/ODVs.

## 2.2. Attention datasets for ODIs

To learn saliency models on ODIs/ODVs, datasets with head fixations and eye fixations are urgently required. Along with saliency prediction approaches, several ODIs/ODVs datasets have been recently established to collect the head fixation/eye fixation data of subjects when viewing omnidirectional scenes. Table 1 summarizes the basic properties of these datasets. To the best of our knowledge, Salient360 (Rai *et al.* [37]) and Saliency in VR (Sitzmann *et al.* [42]) have been widely used in the recent ODI saliency prediction works. These datasets are reviewed in more details as follows.

**Salient360** (Rai *et al.* [37]) is one of the earliest ODI datasets for saliency prediction. It contains 98 stimuli, which mainly include indoor, outdoor and people scenes. For each ODI, at least 40 subjects were asked to view the stimuli with free head movement in the range of $360° \times 180°$. The maximum resolutions of these stimuli are $18,332 \times 9,166$. Each ODI was presented for 25 seconds with an identical initialized viewport for all subjects. Then, the eye fixations and head fixations were recorded. Finally, the ground truth saliency maps of head fixations and eye fixations were both converted into the equirectangular format.

**Saliency in VR** (Sitzmann *et al.* [42]) is also a public dataset that records 1,980 trajectories of head fixations and eye fixations, obtained from 169 subjects viewing 22 static ODIs. In their experiment, the data of head fixations and eye fixations were captured using an HMD in both standing (called *VR standing* ) and seated (called *VR seated*) conditions. In addition, [42] also collected the data of observing the same scenes through a desktop monitor, called *desktop* condition. The dataset offered the ground truth saliency maps of head fixations and eye fixations using three different projections from sphere to plane, i.e., equirectangular, cube map and patch-based projection.

The attention dataset can benefit the saliency prediction approaches for ODIs/ODVs. In particular, the deep learning approaches require large-scale data for training the DNN models. Unfortunately, as shown in Table 1, the existing datasets lack sufficient data, especially for ODIs. Therefore, we establish a large-scale dataset for saliency prediction on ODIs, namely the AOI dataset. The details about our dataset are discussed in Section 3.

## 3. Dataset

### 3.1. Data collection

**Stimuli.** First, we collected 600 ODIs from Flickr [47], the resolution of which ranges from $4,000 \times 2,000$ to $24,028 \times 12,014$. Each ODI was downloaded in the equirectangular format and at the maximum resolution. Note that all 600 ODIs were available under the creative commons copyright. To enrich the diversity of the content in our dataset, four categories of ODIs were collected including *cityscapes, natural landscapes, indoor scenes and human scenes*. Figure 2 shows some examples for each category of ODIs in our dataset.

**Equipment.** We obtained the HM and eye movement (EM) data of the subjects through the HTC vive and aGlass. Here, the HTC vive is used as an HMD to view ODIs. The HM data can be captured by the HTC vive, while the aGlass device is able to capture the EM data within FoV. Note that the aGlass device is embedded in the HTC vive. When the subjects viewed ODIs, the "virtual desktop" was used to display all images, and meanwhile the software of [28] was applied to record both HM and EM data. Note that the whole HM data along with the time stamps form the head trajectory of viewing an ODI, from which we can extract the head fixations. Similarly, the eye fixations can also be obtained from the EM data.

**Subjects.** There were in total 30 subjects (19 females and 11 males) involved in our experiment, and their ages ranged from 18 to 30, with an average of 21 years old. Note that all subjects have normal or corrected-to-normal eyesight[2]. Before viewing ODIs, a simple training session was conducted to familiarize the subjects with the HTC vive. Furthermore, the procedure of the experiment was explained to all subjects. Finally, the subjects underwent the experiment of viewing ODIs with the following procedure.

**Procedure.** The 600 ODIs were randomly and equally divided into 2 equal groups. Two groups of ODIs were viewed by each subject on different days to avoid the fatigue. According to the Salient360 dataset [37], the duration was set to 22 seconds for viewing each ODI. After viewing each image, we inserted a gray ODI with a red dot located at longitude $= 0°$ and latitude $= 0°$, and the subjects pressed to enter the next ODI once they fixated on the red dot. Consequently, when the subjects viewed the next ODI, their HM and EM can be re-initialized to the center of the corresponding equirectangular image. Note that the subjects were allowed to have a rest when they felt fatigue. When viewing ODIs, the subjects wearing the HTC vive were asked to sit in a comfortable swivel chair, allowing them to rotate $360°$ freely. As such, all panoramic regions in the image can be easily accessed.

**Raw data.** Then, the raw HM data are recorded in the

---

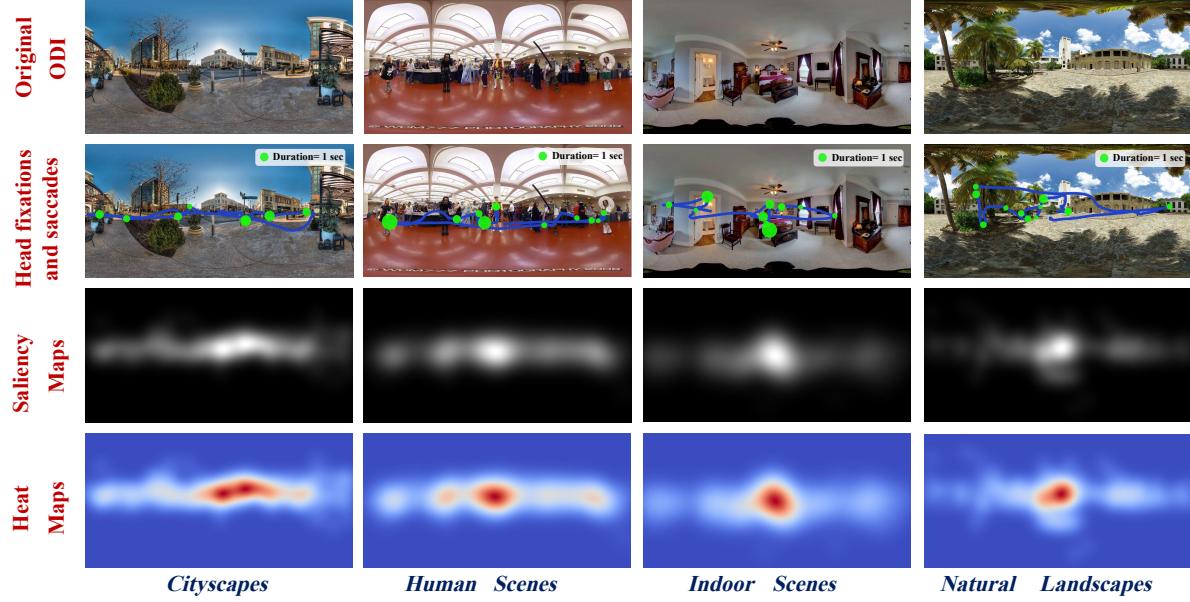[2]The device of aGlass can also be used to correct eyesight.

Figure 3: Examples for head fixations and saccades as well as saliency maps (in the form of gray maps and heat maps).
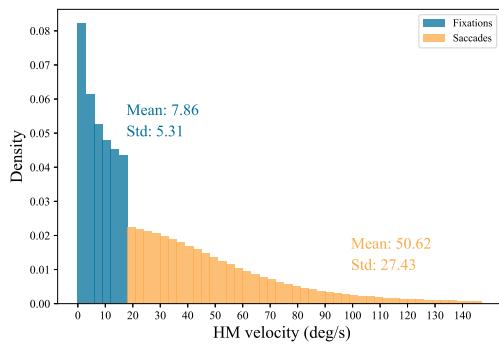


Figure 4: Histogram distribution of the velocities of fixations (blue) and saccades (orange).

following format. Note that our AOI dataset does not process the EM data, as this paper only focuses on predicting head fixations on ODIs. In fact, the HM data of a subject at one ODI can be represented by a vector: [Time stamp, HM pitch, HM yaw]. Specifically, the above vector is composed of the time stamp and HM position. (1) Time stamp: The interval time between two neighboring sample points of HM are recorded and represented in milliseconds for each ODI. (2) HM position: Two elements are related to the HM position, including 2 Euler angles: the angles of pitch and yaw. As shown in Figure 1, the location of the viewport can be represented in latitude and longitude, corresponding to the angles of pitch and yaw, respectively.

## 3.2. Data processing

Given the above raw HM data, we need to distinguish head fixations and saccades. In this paper, we mainly focus on predicting the saliency maps of head fixations. However, our dataset can also be used to predict HM saccades of HM for the future work. Our algorithm for distinguishing head fixations and saccades is presented as follows.

**Head fixations and saccades.** When viewing ODIs, one subject may move his/her head along with saccades, and then fix on the regions that are attractive to him/her, seen as head fixations. Next, we focus on extracting head fixations and saccades from the HM data. First, the velocity of HM is measured through the *orthodromic distance* [33] between two successive HM data divided by the corresponding time stamp. Mathematically, the HM velocity $v_i$ at the $i$-th sample can be denoted as

$$v_i = \frac{d_i}{\Delta T}, \tag{1}$$

where $\Delta T$ is the duration of the time stamp, and $d_i$ is the *orthodromic distance* between the $(i$-1$)$-th and $i$-th samples. Here, $d_i$ is defined as follows:

$$d_i = r \cdot \Delta\sigma_i, \tag{2}$$

where $\Delta\sigma_i$ is the *spherical distance*:

$$\Delta\sigma_i = 2\arcsin\sqrt{\sin^2(\frac{\Delta\psi_i}{2}) + \cos\psi_i \cdot \cos\psi_{i-1} \cdot \sin^2(\frac{\Delta\theta_i}{2})}. \tag{3}$$

In addition, $r$ is the radius of the omnidirectional sphere. For the $i$-th HM sample, $\Delta\psi_i$ is the difference in latitude;

$\Delta\theta_i$ is the difference in longitude; and $\psi_{i-1}$ and $\psi_i$ are latitudes of two successive samples.

Then, using the *velocity-threshold identification algorithm (I-VT)* [40], we separate the head fixations and saccades based on the sample-to-sample velocities of HM. In this paper, we follow [46] to set the velocity threshold to be 18 degrees/second. In other words, if the velocity of an HM sample is below 18 degrees/second, it belongs to the head fixations; otherwise it belongs to saccades. Figure 4 shows the histograms of head fixations and saccades, calculated over all HM data of 600 ODIs in our dataset. Since this paper mainly focuses on predicting head fixations, all saccades (HM with the speed above the threshold) are discarded prior to the further analysis. After this process, we obtain the head fixations for our dataset. Figure 3-(Second row) shows a raw head trajectory of one subject when viewing an ODI, which is composed of saccades and head fixations.

**Saliency maps of head fixations.** For obtaining the 2D saliency maps of head fixations, we apply equirectangular projection to process the sphere-format data according to [58]. In equirectangular projection, the yaw and pitch of the $i$-th head fixation on the sphere coordinate (in degrees) are mapped to a 2D pixel in the equirectangular image, i.e., a head fixation denoted by $(x_i, y_i)$ in the equirectangular coordinate. Here, the origin $(0, 0)$ of the equirectangular coordinate is located at the lower left corner of the ODI. Then, $(x_i, y_i)$ can be obtained by

$$x_i = \left(\frac{\theta_i}{360} + \frac{1}{2}\right) \times W,$$
$$y_i = \left(\frac{\psi_i}{180} + \frac{1}{2}\right) \times H. \tag{4}$$

For the $i$-th head fixation, $\theta_i$ and $\psi_i$ are its yaw and pitch, respectively; $W$ and $H$ denote the width and height of the equirectangular image, in the form of pixel numbers.

Then, the head fixations of all subject are convolved with a Gaussian kernel to generate the saliency map for each ODI. According to [10, 21], a 2D Gaussian kernel with 1.5° visual angle centered at the head fixation is used in this paper. The kernel that locates at head fixation $(x_i, y_i)$ can be represented as:

$$G(x_i, y_i) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{x_i^2 + y_i^2}{2\sigma^2}}, \tag{5}$$

where the standard deviation is $\sigma = \frac{W}{2\sqrt{2\ln 2}}$ with $W$ a constant value of 90 pixels [38]. Figure 3 shows some examples of the original equirectandular images and saliency maps.

## 4. Dataset analysis

In this section, we mine our dataset to investigate human behavior in viewing ODIs. Specifically, we have the following findings.
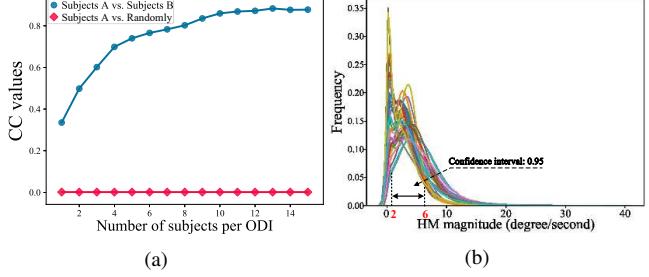


(a)                    (b)

Figure 5: (a): Average CC values alongside increased numbers of subjects per ODI over our AOI dataset; The two curves denote CC values between A and B, randomly sampled maps, respectively. (b): Probability distribution for the HM magnitude between adjacent HM across 30 subjects, over 600 ODIs of our AOI dataset. Each curve stands for the distribution of one subject.

*Finding 1*: *The distribution of head fixations are variant between individual subjects; however, the consistency of head fixations among subjects increases and converges when the number of subjects increases.*

*Analysis*: We analyze the consistency of the head fixation distribution across subjects when the number of subjects increases. To this end, we randomly divide the subjects into two equal groups, denoted as $A$ and $B$, and the number of subjects in these two groups progressively increases from 1 to 15. For each ODI in our dataset, the saliency maps of $A$ and $B$ are generated by convolving with the 2D Gaussian kernel (see (5)) over the corresponding head fixations, which are denoted as $\mathbf{S}_A$ and $\mathbf{S}_B$, respectively. Then, the consistency of head fixations between two groups is measured by calculating the linear correlation coefficient (CC) of saliency maps between $\mathbf{S}_A$ and $\mathbf{S}_B$. Figure 5-(a) shows the average CC values between $\mathbf{S}_A$ and $\mathbf{S}_B$ along with the increased number of subjects. We also plot the CC values between saliency map $\mathbf{S}_A$ and the saliency map of randomly generated head fixations (with the same number as group $A$). Note that in Figure 5-(a), the CC values are calculated and averaged over all ODIs, after randomly dividing $A$ and $B$ 20 times. We can see from this figure that the CC value is 0.335 when the subject number is 1 in each group. This result indicates that the head fixations are variant between two subjects, despite a certain consistency that exists (larger than that between the subject and the randomly generated head fixations). In addition, the CC value increases and converges along with the increased subject number. Therefore, this completes the analysis of *Finding 1*.

*Finding 2*: *There exists FCB for the head fixations on ODIs.*

*Analysis*: Given all collected fixations in our dataset, we calculate the distribution of their locations along with the longitude and latitude. The results are shown in Figure 6-(a) and (b), respectively. We can see from this figure that the head fixations tend to be attracted by the regions near
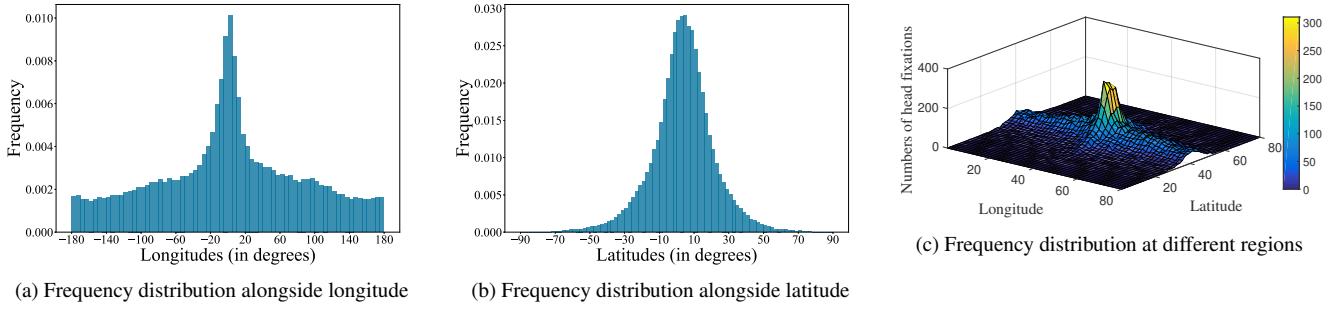
(a) Frequency distribution alongside longitude

(b) Frequency distribution alongside latitude

(c) Frequency distribution at different regions

Figure 6: (a): Numbers of head fixations alongside longitude over (from -180° to 180°), (b): Numbers of head fixations alongside latitude (from -90° to 90°), (c): Numbers of head fixations at different omnidirectional regions, calculated over all 30 subjects and all ODIs in our AOI dataset.

$0°$ longitude (i.e., the front region) and $0°$ latitude (i.e., the equator). Hence, high probability exists that the head fixations fall into the front center region. In other words, the FCB holds for the head fixations on ODIs. In addition, Figure 6-(c) counts the numbers of head fixations in different omnidirectional regions over our dataset. In this figure, the full equirectangular region of $360° \times 180°$ panorama is equally segmented to $4.5° \times 2.25°$ grids. Then, the numbers of head fixations of all 30 subjects are counted in each grid. As observed in Figure 6-(c), head fixations are more likely to be attracted by the equator (i.e., the latitude is close to $0°$), especially the center of the equator (i.e., the longitude is also close to $0°$). Again, this observation verifies that the FCB exists for the head fixations in our dataset. Therefore, the analysis of *Finding 2* is substantiated.

*Finding 3*: The HM magnitude is similar across all subjects over all ODIs in our AOI dataset.

*Analysis*: For each subject, we calculate the magnitude between two HM positions of two adjacent samples through the *spherical distance*. Figure 5-(b) shows the distributions of HM magnitudes for all subjects in our AOI dataset, and in this figure each curve stands for the distribution of one subject. As can be seen in this figure, the distributions of the HM magnitudes are similar among subjects. In particular, most of the HM magnitude values locate at the range of $2° \sim 6°$ for almost all subjects (confidence interval: 95%). Consequently, there exists similarity for the HM magnitude across all subjects when viewing ODIs. This completes the validation of *Finding 3*.

## 5. SalGAIL approach

### 5.1. Framework

In this section, we present our SalGAIL approach that aims to predict head fixations on ODIs in the form of saliency maps. Figure 7 shows the overall framework of the SalGAIL approach. As seen in this figure, our SalGAIL approach is composed of two stages: training and test. In

the training stage, we propose a GAIL method for learning the *reward* of imitating the head fixations of each subject. Then, the learned *reward* of each subject is used in the corresponding DRL stream to predict a head trajectory. In the test stage, an ODI is input to a multi-stream DRL model, and given the learned *reward*; then, each DRL stream predicts one head trajectory. Consequently, the predicted head fixations can be obtained from the head trajectories of all DRL streams. Finally, the predicted head fixations are convoluted to generate the saliency map of the input ODI.

### 5.2. Test: multi-stream DRL for saliency prediction

**Problem formulation**. First, we formulate the problem of saliency prediction on the input ODI (denoted by $\mathbf{I}$) as follows. Assume that there are in total $N$ DRL streams, each of which corresponds to one subject. Since there are 30 subjects in our AOI dataset, $N$ is chosen to be 30 in this paper. Note that the head fixations of 30 subjects have converged to consistency, according to *Finding 1*. Then, we establish an $N$-stream DRL model to predict the head trajectories of $N$ subjects, in which the $n$-th stream aims to generate the head trajectory of subject $n$, denoted as $\boldsymbol{\tau}^n = \{(\hat{x}_t^n, \hat{y}_t^n)\}_{t=1}^{T}$. Here, $\hat{x}_t^n$ and $\hat{y}_t^n$ are the 2D coordinates of the HM position obtained from the $n$-th DRL stream at time step $t$, and $T$ means the total duration of viewing each ODI. Next, we need to extract the head fixations from all predicted head trajectories $\{\boldsymbol{\tau}^n\}_{n=1}^{N} = \{(\hat{x}_t^n, \hat{y}_t^n)\}_{t=1,n=1}^{T,N}$. Let $\{\mathbf{p}_k^n\}_{k=1,n=1}^{K_n,N}$ denote all extracted head fixations, where $\{\mathbf{p}_k^n\}_{k=1}^{K_n}$ is the set of all head fixations from the $n$-th DRL stream, and $K_n$ is the total number of fixations output by this DRL stream. Then, saliency map $\tilde{\mathbf{S}}$ can be generated for the input ODI, via convoluting all fixations $\{\mathbf{p}_k^n\}_{k=1,n=1}^{K_n,N}$ with a 2D Gaussian kernel (see (5)). Since *Finding 2* reveals that the FCB exists for head fixations, FCB map $\mathbf{C}$ is added into the generated saliency map $\tilde{\mathbf{S}}$ to output the final saliency map $\mathbf{S}$:

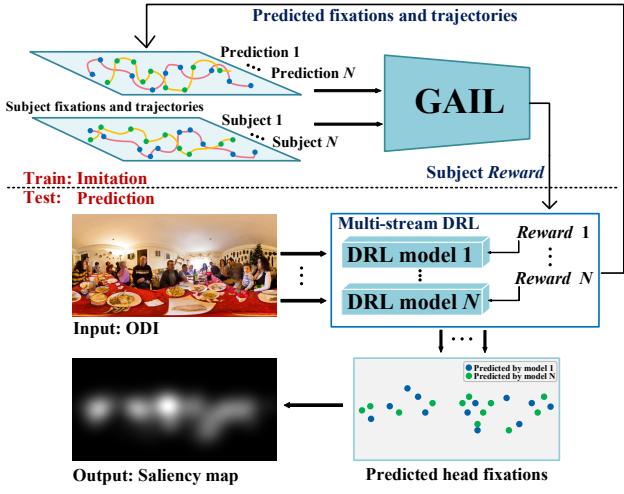$$\mathbf{S} = \text{Norm}(\mathbf{C} + \tilde{\mathbf{S}}), \tag{6}$$

Figure 7: Framework of the SalGAIL approach.

where Norm($\cdot$) is the normalization operation that ensures all saliency values range from 0 to 1.

**Multi-stream DRL model**. Now, we focus on the multi-stream DRL model for predicting all head trajectories $\{(\hat{x}_t^n, \hat{y}_t^n)\}_{t=1,n=1}^{T,N}$. In our approach, each multi-stream DRL shares the same framework, but with different *rewards*. See Section 5.3 for more details about the *reward* modeling. Here, we take the $n$-th DRL stream as an example. Specifically, we define the terms of the DRL stream as follows.

- *Observation* $\hat{\mathbf{O}}_t^n$ is the viewport at time step $t$ for the $n$-th DRL stream.

- *Action* $\hat{a}_t^n$ at time step $t$ is formulated by the HM from $t-1$ to $t$. Similar to [51], the *action* space includes 8 discrete directions $\{0°, 45°, 90°, \cdots, 315°\}$. Different from [51], one additional *action* (denoted as *stay*) is added to indicate that the HM is fixed without any change in the viewport. Thus, the *action* space is $\{0 : stay, 1 : 0°, 2 : 45°, 3 : 90°, \cdots, 8 : 315°\}$.

- *Policy* $\pi_{\boldsymbol{\omega}}^n$ is modeled as the predicted probability distribution over the *actions* of HM across time steps with $\boldsymbol{\omega}$ as its parameters.

- *Reward* $r_t^n$ denotes the *reward* of the *action* made at time step $t$ in the $n$-th DRL stream. In our approach, the *reward* function is learned to imitate the human *actions* of head trajectories, to be described in Section 5.3.

- *Environment* $\mathbf{E}$ is composed of the *reward* estimator and viewport extractor, such that the *reward* and *observation* can be obtained for the *agent* in action-making.

Given the above terms, the procedure of our multi-stream DRL is summarized in Figure 8. Specifically, for the $n$-th

DRL stream, *observation* $\hat{\mathbf{O}}_t^n$ at time step $t$ is obtained from the viewport. That is, the viewport is extracted to make its center locate at the HM position $(\hat{x}_t^n, \hat{y}_t^n)$. It is worth mentioning that the size of the viewport is determined by the HMD. Then, the viewport is projected onto the 2D plane with the size of $84 \times 84$, as *observation* $\hat{\mathbf{O}}_t^n$. Subsequently, *observation* $\hat{\mathbf{O}}_t^n$ is input into a CNN (see Figure 8-(b) for the structure of CNN) to produce a *policy* $\pi_n$, which maximizes *reward* $r_t^n$. Given the *policy*, the *agent* follows $\epsilon$-*greedy* [2] to randomly sample an action $\hat{a}_t^n$ from *stay* or 8 discrete directions. Based on action $\hat{a}_t^n$, *environment* $\mathbf{E}$ updates the current HM position with a fixed HM magnitude, from $(\hat{x}_t^n, \hat{y}_t^n)$ to $(\hat{x}_{t+1}^n, \hat{y}_{t+1}^n)$ for the next time step. Here, we set the fixed HM magnitude to be averaged magnitude across all subjects on the ODIs of the training set according to *Finding 3*. Then, the new *observation* $\hat{\mathbf{O}}_{t+1}^n$ can be obtained upon $(\hat{x}_{t+1}^n, \hat{y}_{t+1}^n)$ for making the *action* at time step $t+1$. The transition of the *observation* is then defined as $T$: $\hat{\mathbf{O}}_{t+1}^n \sim T(\hat{\mathbf{O}}_t^n \mid \hat{a}_t^n)$.

### 5.3. Training: GAIL for reward modeling

In this section, we focus on modeling the *reward* for the DRL model of our SalGAIL approach, which is based on GAIL. Specifically, GAIL is applied to make the predicted head trajectories of our DRL model imitate ground truth head trajectories of subjects. The framework of the training stage can be seen in Figure 9-(a). For GAIL, our multi-stream DRL model acts as the generator, outputting the head trajectories. Then, the discriminator distinguishes whether a head trajectory is a predicted one or the ground truth. Consequently, the probability that an input head fixation is the ground truth can be obtained from the discriminator, viewed as the *reward* for the DRL model. Furthermore, we propose a *policy* selector to match the *policy* of one DRL model to the corresponding subject, when maximizing the *reward* in the discriminator. In the following, more details about the generator, discriminator and selector are presented.

- Generator. The generator of our SalGAIL approach is the multi-stream DRL model described in Section 5.2 , which aims at learning *policies* to imitate the head trajectories of subjects. The input to the generator is the training ODIs. Then, the *policy* of one DRL stream can be updated at each episode for optimizing the corresponding *reward*. Note that each DRL stream learns one *policy*, corresponding to the head trajectories of a subject. Consequently, the generator outputs the predicted head trajectories, as the input to the discriminator. Here, the predicted head trajectories are $\{\boldsymbol{\tau}^n\}_{n=1}^N$ correspond to the ground truth head trajectories of subjects $\{\boldsymbol{\tau}_S^n\}_{n=1}^N$, which are obtained under the *policies* of these subjects: $\boldsymbol{\pi}_S = \{\pi_S^n\}_{n=1}^N$.

---

**Algorithm 1:** Training stage based on GAIL.

---
**Input:** The training ODIs; the head trajectories of $N$ subjects $\{\boldsymbol{\tau}_S^n\}_{n=1}^N$ over the training ODIs.
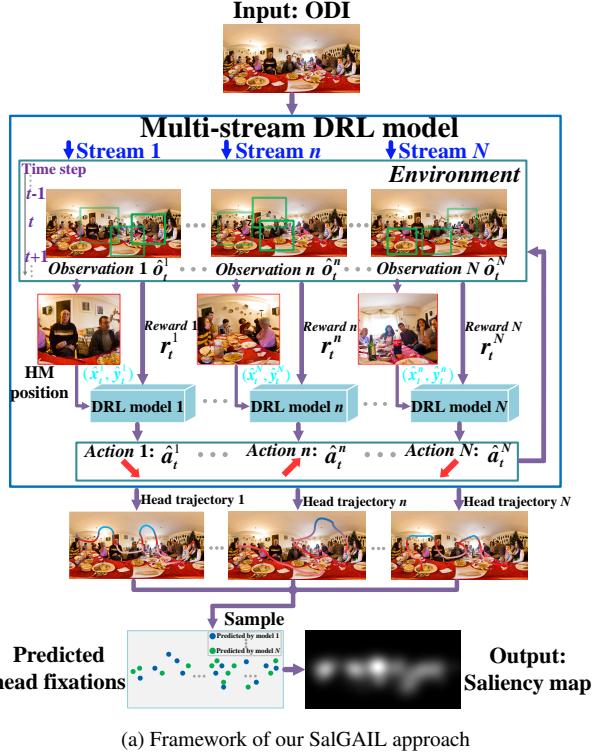**Output:** The learned *policies* $\{\pi_{\boldsymbol{\omega}}^n\}_{n=1}^N$ of the multi-stream DRL model.

1 **Initialize:** Maximum number of training cycles $H$; the number of episodes $I$; the step size of one episode $B$; random initial parameters
   $\boldsymbol{\omega}_0 = \{\boldsymbol{\omega}_\pi^0, \boldsymbol{\omega}_V^0\}$, $\boldsymbol{\phi}_0$ and $\boldsymbol{\eta}_0$ for the generator, discriminator and *policy* selector.
2 Run the generator, discriminator and *policy* selector with the initial parameters $\boldsymbol{\omega}_0$, $\boldsymbol{\phi}_0$ and $\boldsymbol{\eta}_0$, respectively.
3 **while** $h < H$ **do**
4      $h \leftarrow h+1$.
5      Randomly sample $N$ ODIs into the multi-stream DRL model.
6      Initialize HM positions $\{(\hat{x}_t^n, \hat{y}_t^n)\}_{t=1,n=1}^N \leftarrow (0,0)$, as *observations* $\{\hat{\mathbf{O}}_t^n\}_{t=1,n=1}^N$.
7      **for** $i = 1 \rightarrow I$ **do**
8          Initialize empty sets $\boldsymbol{\chi}$ and $\boldsymbol{\chi}_S$ for collecting the *observation-action* pairs of prediction and subjects, respectively.
9          **for** $n = 1, 2, \ldots, N$ **do**
10              Set latent vector $\mathbf{c}_n$ in an one-hot form, corresponding to $n$-dimension.
11              Obtain *observation-action* pairs $\boldsymbol{\chi}_S^n$ (size: $B$) of the $n$-th subject from $\boldsymbol{\tau}_S^n$: $\boldsymbol{\chi}_S^n \subset \boldsymbol{\tau}_S^n$.
12              Sample predicted *observation-action* pairs $\boldsymbol{\chi}^n$ (size: $B$) under the $n$-th DRL *policy*: $\hat{a}_t^n \sim \pi_{\boldsymbol{\omega}}^n(\hat{a}_t^n \mid \hat{\mathbf{O}}_t^n, \mathbf{c}_n)$, $\hat{\mathbf{O}}_{t+1}^n \sim T(\hat{\mathbf{O}}_t^n \mid \hat{a}_t^n)$.
13              Calculate the *rewards* $\{r_t^n\}_{t=(i-1)\cdot B+1}^{i\cdot B}$ according to (7) and deliver the *rewards* into *rollout* for optimization.
14              **if** $\{r_t^n\}_{t=(i-1)\cdot B+1}^{i\cdot B}$ *tend to converg* **then**
15                  Break from the while loop.
16              **end**
17              Append $\boldsymbol{\chi}_S^n$ and $\boldsymbol{\chi}^n$ into $\boldsymbol{\chi}_S$ and $\boldsymbol{\chi}$, respectively.
18          **end**
19          Update $\boldsymbol{\phi}_i$ to $\boldsymbol{\phi}_{i+1}$ by ascending with gradients $\Delta_{\boldsymbol{\phi}_i}$ in (18).
20          Update $\boldsymbol{\eta}_i$ to $\boldsymbol{\eta}_{i+1}$ by decreasing with gradients $\Delta_{\boldsymbol{\eta}_i}$ in (19).
21          **for** $n = 0, 1, 2, \ldots, N$ **do**
22              Compute $R_t^n$ using *rewards* $\{r_t^n\}_{t=(i-1)\cdot B+1}^{i\cdot B}$ from *rollout* according to (10).
23              Update $\boldsymbol{\omega}_V^i$ to $\boldsymbol{\omega}_V^{i+1}$ by decreasing with gradients $\Delta_{\boldsymbol{\omega}_V^i}$ in (16).
24              Update $\boldsymbol{\omega}_\pi^i$ to $\boldsymbol{\omega}_\pi^{i+1}$ by increasing with gradients $\Delta_{\boldsymbol{\omega}_\pi^i}$ in (17).
25          **end**
26          Assign parameters $\boldsymbol{\omega}_V^{i+1}$ and $\boldsymbol{\omega}_\pi^{i+1}$ to *policies* $\{\pi_{\boldsymbol{\omega}}^n\}_{n=1}^N$; assign parameters $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ to $D_{\boldsymbol{\phi}}$ and $S_{\boldsymbol{\eta}}$, respectively.
27          **return** $\{\pi_{\boldsymbol{\omega}}^n\}_{n=1}^N$.
28      **end**
29 **end**

---

- Discriminator. The discriminator is a binary classifier determining whether an input head trajectory is the ground truth, which is also used to yield the *reward* of the DRL model in predicting head trajectories. The parameters of the discriminator are updated by distinguishing samples of the head trajectories red in each episode between subjects and the generator. Specifically, the CNN structure of the discriminator can be seen in Figure 9-(b). The input to the discriminator is the *observations* sampled from the head trajectories concatenated with the corresponding *actions* that are encoded by one-hot vector (9 dimensions). Here, the samples from the head trajectories are the *observation-action* pairs in each episode extracted from $\{\boldsymbol{\tau}_S^n\}_{n=1}^N$ and $\{\boldsymbol{\tau}^n\}_{n=1}^N$, defined as $\boldsymbol{\chi}_S$ and $\boldsymbol{\chi}$, respectively. Hence, the probability of $\boldsymbol{\chi}$ being $\boldsymbol{\chi}_S$ seen as the output of the discriminator is viewed as the *reward* for our DRL model. Unfortunately, the original GAIL approach can only learn the *reward* by imitating a single subject [16]. It is difficult for each DRL stream to make correct *actions* to imitate the head tra-

jectories of multiple subjects. Therefore, we propose a *policy* selector that can be used to solve this problem.

- *Policy* selector. Given a subject, the *policy* selector is used to find a suitable *policy* from one among multiple DRL models, via maximizing the *reward* of imitating the subject's head trajectories. This is achieved by adding a latent vector $\mathbf{c}$ into our *policy* function $\pi(\hat{a} \mid \hat{\mathbf{O}}, \mathbf{c})$ as shown in Figure 8-(b). Here, $\pi(\hat{a} \mid \hat{\mathbf{O}}, \mathbf{c})$ is based on InfoGAN [6] and differs from the traditional *policy* function $\pi(\hat{a} \mid \hat{\mathbf{O}})$, which only relies on *observation* $\hat{\mathbf{O}} = \{\hat{\mathbf{O}}_t^n\}_{t=1,n=1}^{T,N}$ and *action* $\hat{\mathbf{a}} = \{\hat{a}_t^n\}_{t=1,n=1}^{T,N}$. In our approach, the latent vector $\mathbf{c}$ is represented by a one-hot vector, of which the $n$-th dimension is represented by 1 or 0, corresponding to the $n$-th stream DRL model. Guided by $\mathbf{c}$, the *policy* selector can select a specific *policy* from the mixture of *policies* of multiple DRL models. Specifically, the *policy* selector acts as a posterior estimation, encouraging the maximum of the mutual information between $\mathbf{c}$ and $\boldsymbol{\chi}$. In our approach, the posterior estimation can be modeled by the CNN of Figure 9-(b) as the *policy* se-

**Input: ODI**

**Multi-stream DRL model**

↓**Stream 1**   ↓**Stream n**   ↓**Stream N**

Time step

*t-1*
*t*
*t+1*

*Environment*

**Observation 1** $\hat{o}_t^1$  · · ·  **Observation n** $\hat{o}_t^n$  · · ·  **Observation N** $\hat{o}_t^N$

**Reward 1** $r_t^1$   **Reward n** $r_t^n$   **Reward N** $r_t^N$

**HM position**  $(\hat{x}_t^1, \hat{y}_t^1)$   $(\hat{x}_t^n, \hat{y}_t^n)$   $(\hat{x}_t^N, \hat{y}_t^N)$

**DRL model 1** · · · **DRL model n** · · · **DRL model N**

**Action 1:** $\hat{a}_t^1$ · · · **Action n:** $\hat{a}_t^n$ · · · **Action N:** $\hat{a}_t^N$

Head trajectory 1   Head trajectory n   Head trajectory N

**Sample**

Predicted by model 1
Predicted by model N

**Predicted head fixations**

**Output: Saliency map**

(a) Framework of our SalGAIL approach

**DRL model n**

Convolution  Linear  Leaky Relu  Softmax

*observation* $\hat{o}_t^n$

Latent variables

State value

Policy $\pi_n$   **action** $\hat{a}_t^n$
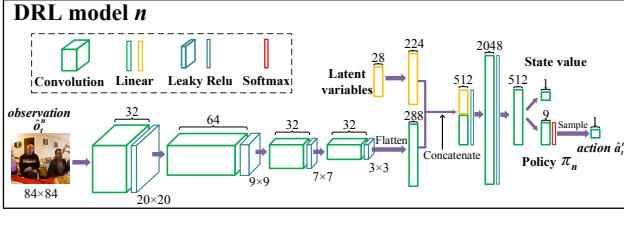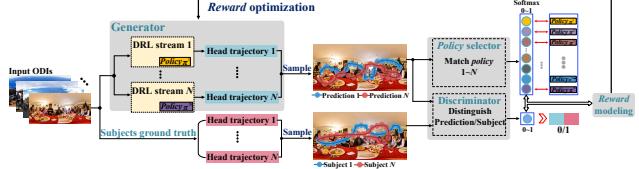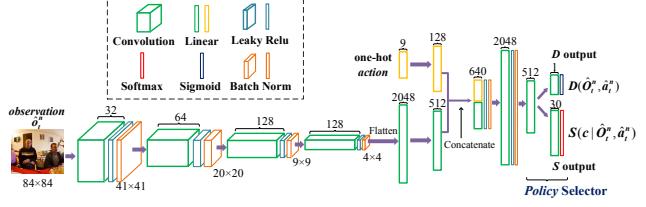
(b) Structure of CNN in DRL model $n$

Figure 8: (a): Framework of our SalGAIL approach for predicting the saliency maps of head fixations on ODIs. (b): Taking the $n$-th DRL stream for an example, we show the internal structure of the DRL network.

lector. Note that our discriminator and *policy* selector share the same parameters in first few layers, and the last layer outputs different values for the discriminator and *policy* selector. In addition, the mutual information is viewed as one part of the *reward* for the DRL model to be maximized. In the following, the details about *reward* modeling are presented.

**Reward modeling.** We take into account two components for modeling the *reward*. The first component is the probability of the predicted head trajectories being those of the subjects, as presented for the discriminator. The second component is the mutual information between the predicted head trajectories and the latent vector of the *policy* selector. Specifically, for the $n$-th DRL model of the generator, the



(a) The training stage of our SalGAIL approach



(b) Structure of CNN in the discriminator and the *policy* selector

Figure 9: (a): The training stage of SalGAIL includes *reward* modeling and optimization. Note that the *reward* I is the output of the discriminator and the *reward* II is the output of the selector. (b): The CNN structure of the discriminator network $D$ and posterior approximation network $S$.

*reward* $r_t^n$ at time step $t$ can be defined as follows:

$$r_t^n = r(\hat{\mathbf{O}}_t^n, \hat{a}_t^n; \boldsymbol{\phi}) + \lambda_1 r(\mathbf{c}; \boldsymbol{\eta}). \quad (7)$$

where $\lambda_1$ ($> 0$) is the hyperparameter that controls the trade-off between two components of the *reward*. In the above equation, $r(\hat{\mathbf{O}}_t^n, \hat{a}_t^n; \boldsymbol{\phi})$ is the first component of the *reward*, which is obtained from the output of the discriminator. Mathematically, it is formulated by

$$r(\hat{\mathbf{O}}_t^n, \hat{a}_t^n; \boldsymbol{\phi}) = -\log(1 - D_{\boldsymbol{\phi}}(\hat{\mathbf{O}}_t^n, \hat{a}_t^n)), \quad (8)$$

where $D_{\boldsymbol{\phi}}$ denotes the CNN of the discriminator with $\boldsymbol{\phi}$ as its parameters. In addition, $r(\mathbf{c}; \boldsymbol{\eta})$ is the second component of the *reward* obtained from the *policy* selector, and it is a mutual information as follows,

$$r(\mathbf{c}; \boldsymbol{\eta}) = \sum_{k=1}^{K} p(c_k \mid \mathbf{O}_t^n, a_t^n) \log S_{\boldsymbol{\eta}}(c_k \mid \hat{\mathbf{O}}_t^n, \hat{a}_t^n), \quad (9)$$

where $S_{\boldsymbol{\eta}}$ is the CNN of the *policy* selector with $\boldsymbol{\eta}$ as its parameters. In addition, $c_k$ is the $k$-th element in the $\mathbf{c}$, and $p(c_k \mid \mathbf{O}_t^n, a_t^n)$ is the probability distribution of the latent vector matching the DRL model to a given subject. Recall that $K$ is the number of streams in our DRL model.

Next, the accumulated discount *reward* of time step $t$ at each episode can be calculated as follow:

$$R_t^n = \sum_{b=t}^{B} \gamma^{b-t} r_b^n, \quad (10)$$

where $\gamma$ is the discount factor of *Q-learning* [48] and $B$ is the step size of one episode. Finally, $r_t^n$ and $R_t^n$ are delivered into the *rollout* and used for updating the parameters of

the generator through *reward* optimization. The optimization procedure is discussed in the following.

**Optimization.** Given (10), we optimize the *rewards* of all DRL streams in our SalGAIL approach by maximizing the expectation of the accumulated discount *rewards* at each episode,

$$\max_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi}}[R_t^n] = \max_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi}}[\sum_{b=t}^{B} \gamma^{b-t} r_b^n], \qquad (11)$$

such that *policies* $\boldsymbol{\pi}$ can be learned in the generator. Here, $\boldsymbol{\pi} = \{\pi_{\boldsymbol{\omega}}^n\}_{n=1}^N$ denote the *policies* of $N$ DRL streams, which can be learned by updating the CNN parameters of DRL. Note that a causal entropy regularization term [5] is added in (11) to ensure the exploration in the decision making of DRL. Mathematically, it is written as

$$H(\boldsymbol{\pi}) \triangleq \mathbb{E}_{\boldsymbol{\pi}}[-\log \boldsymbol{\pi}(\hat{\mathbf{a}} \mid \hat{\mathbf{O}}, \mathbf{c})]. \qquad (12)$$

Recall that $\hat{\mathbf{O}}$ and $\hat{\mathbf{a}}$ are the sets of *observations* $\hat{\mathbf{O}}=\{\hat{\mathbf{O}}_t^n\}_{t=1,n=1}^{T,N}$ and *actions* $\hat{\mathbf{a}}=\{\hat{a}_t^n\}_{t=1,n=1}^{T,N}$, respectively. Thus, the optimization formulation of (11) is rewritten as

$$\max_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi}}[R_t^n] = \max_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi}} \left[ \sum_{b=t}^{B} \gamma^{b-t} \left( -\log(1 - D_{\boldsymbol{\phi}}(\hat{\mathbf{O}}_b^n, \hat{a}_b^n)) \right. \right.$$
$$\left. \left. + \lambda_1 \sum_{k=1}^{K} p(c_k \mid \mathbf{O}_b^n, a_b^n) \log S_{\boldsymbol{\eta}}\left(c_k \mid \hat{\mathbf{O}}_b^n, \hat{a}_b^n\right) \right) \right] - \lambda_2 H(\boldsymbol{\pi}),$$
$$(13)$$

based on (8) and (9). Here, $\lambda_2$ ($> 0$) is the hyperparameter for balancing the trade-off of the first two terms and the regularization term in (13). Consequently, the generator is capable of imitating the head trajectories of subjects through *reward* optimization (13), once $D_{\boldsymbol{\phi}}(\hat{\mathbf{O}}_b^n, \hat{a}_b^n)$ and $S_{\boldsymbol{\eta}}(c_k \mid \hat{\mathbf{O}}_b^n, \hat{a}_b^n)$ have been obtained. In the following, we present the details about calculation on $D_{\boldsymbol{\phi}}(\hat{\mathbf{O}}_b^n, \hat{a}_b^n)$ and $S_{\boldsymbol{\eta}}(c_k \mid \hat{\mathbf{O}}_b^n, \hat{a}_b^n)$, which can be achieved through the optimization on the discriminator and the *policy* selector, respectively.

We introduce adversarial training to learn the CNN parameters of the discriminator, such that $D_{\boldsymbol{\phi}}(\hat{\mathbf{O}}_b^n, \hat{a}_b^n)$ can be obtained for (13). In adversarial training, the discriminator tries to make the predicted head trajectories distinguishable from the corresponding ground truth trajectories. Mathematically, $D_{\boldsymbol{\phi}}(\hat{\mathbf{O}}_b^n, \hat{a}_b^n)$ can be obtained by solving the following optimization formulation:

$$\max_{\boldsymbol{D}, \hat{\boldsymbol{D}}} \mathbb{E}_{\boldsymbol{\pi}_S}[\log \boldsymbol{D}] + \mathbb{E}_{\boldsymbol{\pi}}[\log(1 - \hat{\boldsymbol{D}})]. \qquad (14)$$

Recall that $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_S$ are the *policies* of the generator and subjects, respectively. Here, $\boldsymbol{D} = \{D_{\boldsymbol{\phi}}(\mathbf{O}_t^n, a_t^n)\}_{t=1,n=1}^{B,N}$ and $\hat{\boldsymbol{D}} = \{D_{\boldsymbol{\phi}}(\hat{\mathbf{O}}_t^n, \hat{a}_t^n)\}_{t=1,n=1}^{B,N}$ are the sets of outputs from the discriminator in one episode, in which the inputs

Table 2: Settings of hyperparameters in our SalGAIL approach.

| | | |
|---|---|---|
| Generator | Maximum number of training cycles $H$ | $5 \times 10^4$ |
| | The number of episodes $I$ | 42 |
| | The step size of one episode $B$ | 5 |
| | Mini-batch size | 6 |
| | Discount factor $\gamma$ in (10) | 0.99 |
| | Initial learning rate | $7 \times 10^{-4}$ |
| | The angle of the negative slope in LeakyReLU | 0.01 |
| Discriminator & *Policy* selector | Initial learning rate | $2 \times 10^{-4}$ |
| | Batch size | 150 |
| | The angle of the negative slope in LeakyReLU | 0.2 |
| | Numerical stability value in BatchNorm | $1 \times 10^{-5}$ |
| | Momentum in BatchNorm | 0.1 |
| | Weight decay | $2 \times 10^{-3}$ |
| Others | Trade-off hyperparameter for *reward* $\lambda_1$ in (13) | 0.7 |
| | Causal entropy coefficient $\lambda_2$ in (13) | 0.01 |

are the *observation-action* pairs for the ground truth and the prediction.

Next, we learn the CNN parameters of the *policy* selector by optimizing the mutual information between the predicted head trajectories and latent vector to maximum. Then, we can obtain $S_{\boldsymbol{\eta}}(c_k \mid \hat{\mathbf{O}}_t^n, \hat{a}_t^n)$ by solving the following optimization formulation:

$$\max_{\mathbf{S}} \mathbb{E}_{\boldsymbol{\pi}}[\log \mathbf{S}], \qquad (15)$$

where $\mathbf{S}$ denotes the set: $\{S_{\boldsymbol{\eta}}(c_n \mid \hat{\mathbf{O}}_t^n, \hat{a}_t^n)\}_{t=1,n=1}^{B,N}$.

After obtaining $D_{\boldsymbol{\phi}}(\hat{\mathbf{O}}_b^n, \hat{a}_b^n)$ and $S_{\boldsymbol{\eta}}(c_k \mid \hat{\mathbf{O}}_b^n, \hat{a}_b^n)$, we can solve the optimization problem of (13). This is achieved by updating the parameters of $\boldsymbol{\pi}$. As for $\boldsymbol{\pi}$, its parameters $\boldsymbol{\omega}=\{\boldsymbol{\omega}_V^i, \boldsymbol{\omega}_{\pi}^i\}$ are composed of two parts, where the first part is used to update the state value in the DRL model, and the second part is used to update the *policy* in the DRL model. Therefore, we can obtain the gradient $\boldsymbol{\omega}_V^i$:

$$\Delta_{\boldsymbol{\omega}_V^i} = \mathbb{E}_{\boldsymbol{\chi}}[\nabla_{\boldsymbol{\omega}_V^i}(R_t^n - V_{\boldsymbol{\omega}_V^i}(\hat{\mathbf{O}}^n, \hat{a}^n))^2]. \qquad (16)$$

Here, $V_{\boldsymbol{\omega}_V^i}(\hat{\mathbf{O}}^n, \hat{a}^n)$ is one part of output of the DRL model. Based on Lemma 1, the gradient $\boldsymbol{\omega}_{\pi}^i$ can be obtained to optimize (13), which is as follows:

$$\Delta_{\boldsymbol{\omega}_{\pi}^i} = \mathbb{E}_{\boldsymbol{\chi}}[\nabla_{\boldsymbol{\omega}_{\pi}^i} \log \pi_{\boldsymbol{\omega}_{\pi}^i}^n(\hat{a}_t^n \mid \hat{\mathbf{O}}_t^n, \mathbf{c}_n) \cdot (\lambda_2 + R_t^n)]. \qquad (17)$$

**Lemma 1** *Consider that $R_t^n$ is the accumulated discount reward. The gradient $\boldsymbol{\omega}_{\pi}^i$ for optimizing* (13) *at episode $i$ can be calculated by* (17).

*Proof: See Appendix A.*

To solve the optimization problem of (14), the parameters of $D_{\boldsymbol{\phi}}$ can be updated through the following gradient:

$$\Delta_{\boldsymbol{\phi}_i} = \mathbb{E}_{\boldsymbol{\chi}_S}[\nabla_{\boldsymbol{\phi}_i} \log D_{\boldsymbol{\phi}_i}(\mathbf{O}, a)] + \mathbb{E}_{\boldsymbol{\chi}}[\nabla_{\boldsymbol{\phi}_i} \log(1 - D_{\boldsymbol{\phi}_i}(\hat{\mathbf{O}}, \hat{a}))]. \qquad (18)$$

Table 3: CC, KL divergence, NSS and AUC results of saliency prediction by our and other approaches over the test set of our AOI dataset.

| Categories | Approaches | 2D images saliency models | | | | ODIs saliency models | | | | | | Our model |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DVA | BMS | SALICON | MLNet | BMS360 | GBVS360 | Startsev | Zhu | DHP | Battisti | SalGAIL |
| *Cityscapes* | CC | 0.667 | 0.643 | 0.547 | 0.566 | 0.721 | 0.642 | 0.691 | 0.747 | 0.653 | 0.596 | **0.766** |
| | KL divergence | 0.578 | 0.549 | 0.626 | 0.994 | 0.755 | 0.626 | 0.496 | 0.432 | 0.569 | 0.939 | **0.345** |
| | NSS | 1.063 | 0.975 | 0.620 | 0.771 | 1.268 | 1.051 | 0.992 | 1.176 | 0.975 | 0.926 | **1.447** |
| | AUC | 0.768 | 0.762 | 0.731 | 0.754 | 0.825 | 0.784 | 0.790 | 0.806 | 0.788 | 0.744 | **0.841** |
| *Indoor Scenes* | CC | 0.594 | 0.536 | 0.512 | 0.662 | 0.655 | 0.551 | 0.484 | 0.636 | 0.579 | 0.535 | **0.686** |
| | KL divergence | 0.662 | 0.642 | 0.683 | 0.689 | 0.546 | 0.627 | 0.665 | 0.552 | 0.649 | 0.835 | **0.458** |
| | NSS | 1.115 | 0.789 | 1.109 | 1.258 | 1.209 | 0.917 | 0.768 | 1.230 | 0.954 | 0.846 | **1.449** |
| | AUC | 0.782 | 0.747 | 0.803 | 0.831 | 0.829 | 0.762 | 0.758 | 0.805 | 0.774 | 0.759 | **0.848** |
| *Human Scenes* | CC | 0.607 | 0.548 | 0.512 | 0.611 | 0.712 | 0.567 | 0.600 | 0.735 | 0.621 | 0.598 | **0.757** |
| | KL divergence | 0.545 | 0.530 | 0.586 | 0.727 | 0.639 | 0.513 | 0.472 | **0.354** | 0.545 | 0.722 | 0.352 |
| | NSS | 1.209 | 1.072 | 0.858 | 1.249 | 1.430 | 0.965 | 1.259 | 1.417 | 1.318 | 1.203 | **1.603** |
| | AUC | 0.794 | 0.770 | 0.767 | 0.831 | 0.846 | 0.762 | 0.817 | 0.843 | 0.802 | 0.799 | **0.859** |
| *Natural Landscapes* | CC | 0.492 | 0.443 | 0.401 | 0.366 | 0.734 | 0.532 | 0.613 | 0.725 | 0.512 | 0.606 | **0.756** |
| | KL divergence | 0.625 | 0.779 | 0.827 | 1.381 | 0.401 | 0.664 | 0.572 | 0.478 | 0.688 | 0.929 | **0.228** |
| | NSS | 0.875 | 0.809 | 0.438 | 0.489 | 1.502 | 0.894 | 0.931 | 1.205 | 0.881 | 0.847 | **1.725** |
| | AUC | 0.713 | 0.704 | 0.700 | 0.630 | 0.842 | 0.722 | 0.776 | 0.796 | 0.718 | 0.757 | **0.863** |
| **Overall** | CC | 0.590 | 0.557 | 0.511 | 0.589 | 0.714 | 0.590 | 0.595 | 0.727 | 0.591 | 0.589 | **0.742** |
| | KL divergence | 0.603 | 0.584 | 0.637 | 0.844 | 0.584 | 0.566 | 0.532 | 0.420 | 0.613 | 0.786 | **0.345** |
| | NSS | 1.066 | 0.975 | 0.856 | 1.064 | 1.378 | 0.995 | 1.052 | 1.295 | 1.032 | 1.014 | **1.556** |
| | AUC | 0.764 | 0.758 | 0.757 | 0.784 | 0.841 | 0.766 | 0.793 | 0.821 | 0.771 | 0.775 | **0.853** |

Similarly, the parameters of $S_{\boldsymbol{\eta}}$ can be updated through the gradient of objective (15):

$$\Delta_{\boldsymbol{\eta}_i} = -\lambda_1 \mathbb{E}_{\boldsymbol{\chi}}[\nabla_{\boldsymbol{\eta}_i} \log S_{\boldsymbol{\eta}_i}(\mathbf{c} \mid \hat{\mathbf{O}}, \hat{a})]. \qquad (19)$$

Finally, we can train the CNN parameters of all networks in our SalGAIL approach by solving the optimization problem of (13)-(15). In summary, the training stage of our SalGAIL approach can be seen in Algorithm 1.

# 6. Experimental results

## 6.1. Settings

In this section, we validate the effectiveness of the proposed SalGAIL approach. To this end, each category of ODIs in our AOI dataset is randomly divided into training and test sets in a ratio of $5:1$. As a result, there are 500 training ODIs and 100 test ODIs. Then, we compare the performance of our SaGAIL approach with other state-of-the-art approaches, including DVA [50], BMS [55], SALICON [19], MLNet [9], BMS360 [27], GBVS360 [27], Startsev *et al.* [43], Zhu *et al.* [59], DHP [51] and Battisti *et al.* [3]. Among these approaches, DVA, BMS, SALICON, and MLNet are the latest saliency prediction approaches for 2D images. The remaining approaches are the recent approaches for predicting saliency maps of head fixations on ODIs. Since only the training models of SALICON, ML-Net and DHP are available online, they are retrained over our training set for a fair comparison.

In our SalGAIL approach, the input to the generator at each time step is the predicted viewport at the last time step, which has been projected onto a 2D plane and downsampled to $84 \times 84$. The number of streams $N$ was set to 30 in our SalGAIL approach, the same as the subject number. When training the generator of our SalGAIL approach,

the hyperparameters were tuned to optimize the accumulated discount *rewards* over the training set. The values of these hyperparameters can be found in Table 2. Table 2 further tabulates the key hyperparameter settings of the discriminator and the *policy* selector, also tuned over the training ODIs. Then, the RMSprop optimizer [45] and the Adam optimizer [24] were used to update the parameters of the generator and the discriminator/*policy* selector, respectively. All experiments were conducted on a computer with an Intel(R) Core(TM) i7-6700K CPU@4.0 GHz, 32 GB of RAM and a single Nvidia GeForce GTX 1080Ti GPU.

## 6.2. Performance evaluation on SalGAIL

**Objective evaluation**. In our experiments, we objectively evaluate the accuracy of saliency prediction of head fixations in terms of four metrics: CC, KL divergence, normalized scanpath saliency (NSS) and the area under the receiver operating characteristic curve (AUC). The larger values of CC, NSS or AUC indicate higher accuracy in saliency prediction, while a smaller KL divergence means better performance of saliency prediction. Table 3 tabulates the results of CC, KL divergence, NSS and AUC for our own and 10 other approaches over each category of ODIs and all ODIs. We can see from this table that our SalGAIL approach performs much better than other approaches in terms of different metrics. In particular, our SalGAIL approach has a 0.015 increase in CC, 0.075 reduction in KL divergence, 0.178 increase in NSS and 0.012 increase in AUC, over Zhu *et al.* and BMS360 which perform the best among all compared approaches. In addition, our approach is also superior to other approaches in four metrics for categories of *cityscapes, indoor scenes, human scenes* and *textnatural landscapes*. For *human scenes*, our approach achieves the best performance in terms of CC, NSS and AUC, while its KL divergence result is slightly worse than those of Zhu

Figure 10: Examples of saliency maps of head fixations, predicted by our SalGAIL and other approaches.

*et al.*. Generally, our SalGAIL approach is effective in saliency prediction of head fixations on ODIs and is superior to other state-of-the-art approaches.

**Subjective evaluation**. Next, we compare the subjective results of saliency prediction on ODIs. For each category of ODIs, 2 test ODIs are randomly selected from our AOI dataset. Figure 10 visualizes the saliency maps of the selected ODIs generated by our own and 10 other approaches. As seen in this figure, the saliency maps of our SalGAIL approach are much closer to those of the ground truth, when compared with other approaches. This result implies that our SalGAIL approach performs well in the subjective results of saliency prediction. In summary, both objective and subjective results show that our SalGAIL approach outperforms other state-of-the-art approaches for predicting

saliency of head fixations on ODIs.

### 6.3. Generalization ability test

**Evaluation over the Salient360 dataset**. Now, we assess the generalization ability of our approach by testing over the Salient360 dataset [37]. Here, the SalGAIL model, learned from the training set of our AOI dataset, is directly used to predict the saliency maps of head fixations on all ODIs of the Salient360 dataset. We also compare the performance of our SalGAIL approach with 10 other approaches in terms of CC, KL divergence, NSS and AUC. The average results are reported in Table 4. As shown in this table, our SalGAIL approach again outperforms all other approaches. Specifically, there is at least a 0.021 increase in CC, 0.095 reduction in KL divergence, 0.085 increase in NSS, and

Table 4: Mean (standard deviation) values for saliency prediction accuracy of our and other approaches over Salient360 and Saliency in VR datasets.

| | Salient360 (*Seated* condition) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ours | DVA | BMS | SALICON | MLNet | BMS360 | GBVS360 | Startsev | Zhu | DHP | Battisti |
| CC | **0.757** | 0.534 | 0.502 | 0.467 | 0.429 | 0.736 | 0.502 | 0.612 | 0.678 | 0.565 | 0.563 |
| KL divergence | **0.366** | 0.727 | 0.597 | 0.764 | 1.367 | 0.647 | 0.642 | 0.555 | 0.461 | 0.738 | 0.742 |
| NSS | **0.893** | 0.643 | 0.665 | 0.387 | 0.462 | 0.808 | 0.607 | 0.479 | 0.720 | 0.658 | 0.652 |
| AUC | **0.708** | 0.661 | 0.655 | 0.633 | 0.638 | 0.695 | 0.642 | 0.637 | 0.691 | 0.664 | 0.665 |
| | Saliency in VR (*Standing* condition) | | | | | | | | | | |
| | Ours | DVA | BMS | SALICON | MLNet | BMS360 | GBVS360 | Startsev | Zhu | DHP | Battisti |
| CC | **0.641** | 0.502 | 0.475 | 0.444 | 0.458 | 0.627 | 0.286 | 0.441 | 0.544 | 0.475 | 0.523 |
| KL divergence | **0.425** | 0.618 | 0.644 | 0.686 | 1.169 | 0.495 | 0.989 | 0.815 | 0.586 | 0.764 | 0.626 |
| NSS | **1.467** | 1.086 | 1.072 | 1.061 | 1.052 | 1.413 | 0.624 | 1.059 | 1.198 | 1.124 | 1.150 |
| AUC | **0.747** | 0.705 | 0.694 | 0.678 | 0.665 | 0.738 | 0.608 | 0.665 | 0.711 | 0.689 | 0.702 |
| | Saliency in VR (*Seated* condition) | | | | | | | | | | |
| | Ours | DVA | BMS | SALICON | MLNet | BMS360 | GBVS360 | Startsev | Zhu | DHP | Battisti |
| CC | **0.603** | 0.376 | 0.356 | 0.295 | 0.344 | 0.518 | 0.370 | 0.261 | 0.586 | 0.421 | 0.503 |
| KL divergence | **0.454** | 1.157 | 1.102 | 0.946 | 1.489 | 1.861 | 0.905 | 1.286 | 0.492 | 0.923 | 1.110 |
| NSS | **1.042** | 0.635 | 0.576 | 0.510 | 0.567 | 0.970 | 0.624 | 0.571 | 0.977 | 0.745 | 0.815 |
| AUC | **0.772** | 0.671 | 0.665 | 0.669 | 0.665 | 0.760 | 0.669 | 0.641 | 0.742 | 0.695 | 0.718 |

0.013 increase in AUC. This result demonstrates the high generalization ability of our SalGAIL approach.

**Evaluation over the VR dataset**. We further test the performance of our SalGAIL approach over the VR dataset [42]. In the VR dataset, two groups of head fixations are collected, for viewing the same ODIs in the *standing* and *seated* conditions, respectively. Note that the head fixations are obtained in the *seated* condition for both our AOI dataset and the Salient360 dataset. Consequently, the generalization ability of our SalGAIL approach can be evaluated for different viewing conditions. The average results of the CC, KL divergence, NSS and AUC are also reported in Table 4. We can see from this table that our SalGAIL approach also performs better than all other approaches, for both *standing* and *seated* conditions. In summary, our SalGAIL approach has higher generalization ability on different datasets and in different viewing conditions.

## 6.4. Ablation analysis

**Ablation on DRL.** We evaluate the effectiveness of DRL applied in our SalGAIL approach, via replacing it by the supervised learning baseline. Specifically, the supervised baseline acts as a classifier, which is modeled by the CNN, to predict the 8 discrete HM directions or *stay* in the *action* space. For a fair comparison, the CNN structure of the supervised baseline is the same as that of the DRL model in our SalGAIL approach. Meanwhile, the input to the supervised baseline is the viewport extracted from the ODI, which is also the same as the input to SalGAIL. Additionally, the output of the supervised baseline is the probabil-

Table 5: Comparison of saliency prediction results among our SalGAIL approach, the supervised baseline and the random baseline.

| | SalGAIL | Supervised | Random |
|---|---|---|---|
| CC | **0.742** | 0.524 | 0.235 |
| KL divergence | **0.345** | 0.754 | 1.867 |
| NSS | **1.556** | 0.612 | 0.325 |
| AUC | **0.853** | 0.593 | 0.387 |

ity distribution over 8 discrete HM directions and *stay* in the *action* space. Similar to our SalGAIL approach, the supervised baseline also runs 30 streams of classifiers to predict the head fixations of 30 subjects. In the training stage, for one stream, the supervised baseline learns one classifier model from the ground truth data of the corresponding subject. In the test stage, the baseline selects an *action* for one stream at each time step, based on the corresponding trained classifier. In addition to the supervised baseline, we also compare SalGAIL with the random baseline, in which the action of HM directions and *stay* are randomly generated. Figure 11 shows the accuracy of predicting HM directions along with time steps, for our SalGAIL approach, the supervised baseline and the random baseline.

We can see that the prediction accuracy dramatically decreases when replacing DRL by the supervised baseline and random baseline. Moreover, the predicted head trajectories can be obtained upon the predicted *actions* from all streams, and then the saliency map is generated by convoluting all head fixations which are sampled from the predicted head
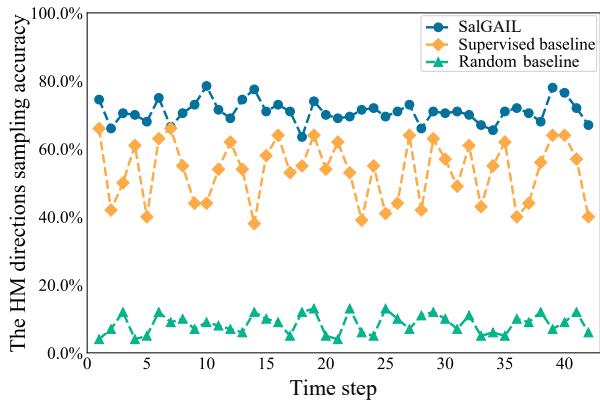
Figure 11: Accuracy of predicting HM directions alongside time steps.



Figure 12: Performance of SalGAIL at different numbers of DRL streams.

trajectories. Table 5 tabulates the CC, KL divergence, NSS and AUC values of SalGAIL, the supervised baseline and the random baseline. As shown in Table 5, the proposed SalGAIL approach performs much better than both the supervised and random baselines. This result validates the effectiveness of DRL applied in our SalGAIL.

**Ablation on the stream number of DRL models.** In our SalGAIL approach, the multi-stream DRL model is used to imitate the head fixations of different subjects. Here, we conduct the ablation experiments to investigate the influence of DRL stream numbers on the performance of our SalGAIL approach. The results are plotted in Figure 12. As shown in this figure, the values of NSS, CC and AUC grow and the KL divergence decreases, along with the increased stream number. Additionally, all four metrics converge when the stream number of DRL approaches 30. This result implies the necessity of the multi-stream DRL and the reasonableness of setting the stream number to 30.

**Ablation on the learned *Reward*.** Here, we evaluate the effectiveness of the *reward* learned by the GAIL algorithm of our approach. To this end, we replace the learned *reward* by a hand-designed *reward* [51]. Then, we obtain the performance of our approach with the learned and hand-designed *rewards*. In addition, we perform a comparison with a random *reward* approach, which randomly samples *actions* at each time step during the whole training process. The comparison results are shown in Figure 13. We can see from this figure that the learned *reward* makes our approach perform significantly better than both hand-designed and random *rewards*. Therefore, the proposed GAIL algorithm of our approach is effective in modeling the *reward* of DRL for saliency prediction on ODIs.

**Ablation on the FCB.** Finally, we ablate the FCB in our SalGAIL approach to investigate its effect on saliency prediction of ODIs. Specifically, we remove the FCB map in (6), such that the results of our approach without the FCB
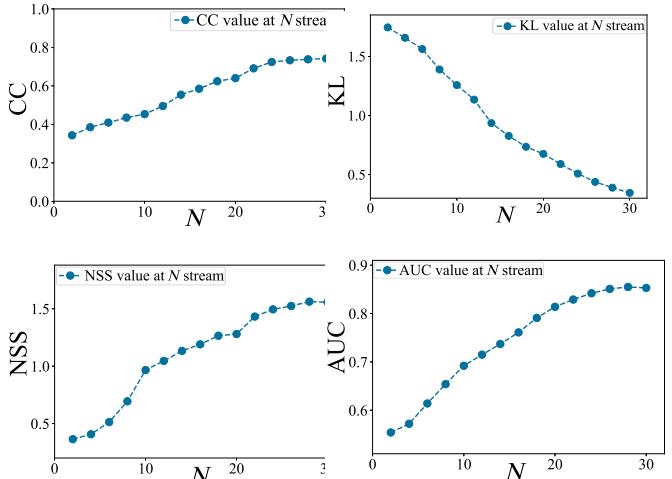
can be obtained. Consequently, after removing the FCB, there is a 0.027 reduction in CC, 0.045 increase in KL divergence, 0.085 reduction in NSS and 0.018 reduction in AUC. Thus, it is necessary to embed the FCB in our SalGAIL approach.

# 7. Conclusion

In this paper, we have proposed the SalGAIL approach for predicting the saliency maps of head fixations on ODIs. First, we established the AOI dataset, which is composed of both head fixations and eye fixations of 30 subjects on 600 ODIs. To the best of our knowledge, AOI is the largest dataset for attention modeling on ODIs. Second, we mined the AOI dataset and achieved several findings regarding the head fixations of subjects when viewing ODIs. Third, inspired by these findings, we proposed a multis-tream DRL model in our SalGAIL approach for saliency prediction on ODIs, in which the *reward* is learned by imitating head trajectories of human through the proposed GAIL algorithm. In the multi-DRL model, each DRL stream yields the HM trajectory of one subject, and then head fixations can be sampled from the yielded HM trajectories of all DRL streams. The experiment also validates the high generation ability of our SalGAIL approach across different datasets. Finally, a processing technique was presented to convolute the predicted head fixations of an input ODI with a Gaussian kernel, such that the saliency map of the ODI can be generated as the output of our SalGAIL approach. The extensive experiments showed that our SalGAIL approach is superior to 10 state-of-the-art approaches in predicting the saliency maps of head fixations on ODIs.

There are two promising research directions of future work. (1) The proposed SalGAIL in this paper mainly
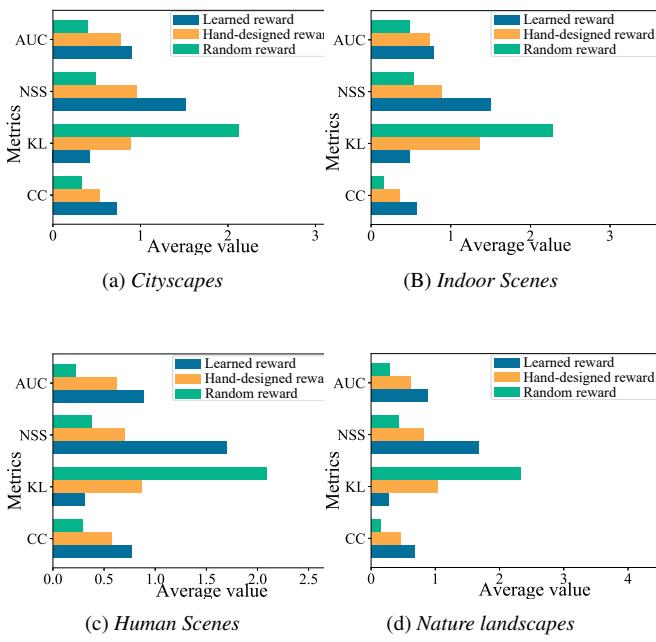
(a) *Cityscapes*　　(B) *Indoor Scenes*

(c) *Human Scenes*　　(d) *Nature landscapes*

Figure 13: Performance evaluation over the DRL model that uses learned *reward*, hand-designed *reward* and random *reward* on four categories in AOI.

focuses on saliency prediction of head fixations on ODIs. Saliency prediction of eye fixations remains to be developed for ODIs. This is an important area for future work. (2) Our SalGAIL approach may be used to remove visual redundancy for some ODI processing tasks, e.g., ODI quality assessment and compression. This is another interesting application for future work.

## References

[1] M. Assens, X. Giro-i-Nieto, K. McGuinness, and N. E. OConnor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2331–2338, Oct 2017.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, May 2002.

[3] F. Battisti, S. Baldoni, M. Brizzi, and M. Carli. A feature-based approach for saliency estimation of omni-directional images. *Signal Processing: Image Communication*, 69:53 – 59, 2018.

[4] A. Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 438–445, June 2012.

[5] A. Boularias, J. Kober, and J. Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 182–189, 2011.

[6] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, pages 2172–2180, 2016.

[7] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[8] X. Corbillon, F. De Simone, and G. Simon. 360 degree video head movement dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*, pages 199–204, 2017.

[9] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3488–3493, Dec 2016.

[10] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet. A dataset of head and eye movements for 360&deg; videos. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*, pages 432–437, 2018.

[11] A. De Abreu, C. Ozcinar, and A. Smolic. Look around you: Saliency maps for omnidirectional images in vr applications. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, May 2017.

[12] Y. S. de la Fuente, R. Skupin, and T. Schierl. Video processing for panoramic streaming using hevc and its scalable extensions. *Multimedia Tools and Applications*, 76(4):5631–5659, 2017.

[13] V. R. Gaddam, M. Riegler, R. Eg, C. Griwodz, and P. Halvorsen. Tiling in interactive panoramic video: Approaches and evaluation. *IEEE Transactions on Multimedia*, 18(9):1819–1831, Sep. 2016.

[14] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, Oct 2012.

[15] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[16] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4565–4573, 2016.

[17] B. Hu, I. Johnson-Bey, M. Sharma, and E. Niebur. Head movements during visual exploration of natural images in virtual reality. In *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, March 2017.

[18] H. Hu, Y. Lin, M. Liu, H. Cheng, Y. Chang, and M. Sun. Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1396–1405, July 2017.

[19] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 262–270, Dec 2015.

[20] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[21] M. Iwasaki and H. Inomata. Relation between superficial capillaries and foveal structures in the human retina. *Investigative ophthalmology & visual science*, 27(12):1698–1705, 1986.

[22] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pages 2106–2113, Sep. 2009.

[23] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell. Sun: Top-down saliency using natural statistics. *Visual cognition*, pages 979–1003, 2009.

[24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] M. K??mmerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge. Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4808, Oct 2017.

[26] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, Sep. 2017.

[27] P. Lebreton and A. Raake. Gbvs360, bms360, prosal: Extending existing saliency prediction models from 2d to omnidirectional images. *Signal Processing: Image Communication*, 69:69 – 78, 2018.

[28] C. Li, M. Xu, X. Du, and Z. Wang. Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*, pages 932–940, 2018.

[29] Y. Li, J. Song, and S. Ermon. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3812–3822, 2017.

[30] J. Ling, K. Zhang, Y. Zhang, D. Yang, and Z. Chen. A saliency prediction model on 360 degree images using color dictionary based sparse representation. *Signal Processing: Image Communication*, 69:60 – 68, 2018.

[31] W.-C. Lo, C.-L. Fan, J. Lee, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu. 360&deg; video viewing dataset in head-mounted virtual reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*, pages 211–216, 2017.

[32] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic. Salnet360: Saliency maps for omni-directional images with CNN. *CoRR*, 2017.

[33] ORH.

[34] C. Ozcinar and A. Smolic. Visual attention in omnidirectional video for virtual reality applications. In *2018 IEEE Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2018.

[35] J. Pan, C. Canton-Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giró i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *CoRR*, 2017.

[36] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. OConnor. Shallow and deep convolutional networks for saliency prediction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 598–606, June 2016.

[37] Y. Rai, J. Gutiérrez, and P. Le Callet. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*, pages 205–210, 2017.

[38] U. Rajashekar, I. Van Der Linde, A. C. Bovik, and L. K. Cormack. Gaffe: A gaze-attentive fixation finding engine. *IEEE transactions on image processing*, 17(4):564–573, 2008.

[39] V. Ramanishka, A. Das, J. Zhang, and K. Saenko. Top-down visual saliency guided by captions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3135–3144, July 2017.

[40] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA '00)*, pages 71–78, 2000.

[41] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[42] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, April 2018.

[43] M. Startsev and M. Dorr. 360-aware saliency estimation with conventional image saliency predictors. *Signal Processing: Image Communication*, 2018.

[44] M. Stengel and M. Magnor. Gaze-contingent computational displays: Boosting perceptual fidelity. *IEEE Signal Processing Magazine*, 33(5):139–148, Sep. 2016.

[45] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[46] E. Upenik and T. Ebrahimi. A simple method to obtain visual attention data in head mounted virtual reality. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 73–78, July 2017.

[47] F. VR.

[48] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3):279–292, May 1992.

[49] WebVR.

[50] J. S. Wenguan Wang. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 2018.

[51] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[52] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao. Gaze prediction in dynamic 360immersive videos. In *2018*

*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5333–5342, June 2018.

[53] J. Yang and M. Yang. Top-down visual saliency via joint crf and dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(3):576–588, March 2017.

[54] M. Yu, H. Lakshman, and B. Girod. A framework to evaluate omnidirectional video coding schemes. In *2015 IEEE International Symposium on Mixed and Augmented Reality*, pages 31–36, Sept 2015.

[55] J. Zhang and S. Sclaroff. Exploiting surroundedness for saliency detection: A boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):889–902, May 2016.

[56] Z. Zhang, Y. Xu, J. Yu, and S. Gao. Saliency detection in 360 degree videos. In *2018 IEEE European Conference on Computer Vision (ECCV)*, pages 504–520, 2018.

[57] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[58] C. Zhou, Z. Li, and Y. Liu. A measurement study of oculus 360 degree video streaming. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*, pages 27–37, 2017.

[59] Y. Zhu, G. Zhai, and X. Min. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication*, 69:15 – 25, 2018.