ObjectVisA-120: Object-based Visual Attention Prediction in Interactive Street-crossing Environments

Igor Vozniak[1], Philipp Müller[1,2], Nils Lipp[1], Janis Sprenger[1], Konstantin Poddubnyy[1], Davit Hovhannisyan[1], Christian Müller[1], Andreas Bulling[3], Philipp Slusallek[1]

***Abstract*— The object-based nature of human visual attention is well-known in cognitive science, but has only played a minor role in computational visual attention models so far. This is mainly due to a lack of suitable datasets and evaluation metrics for object-based attention. To address these limitations, we present ObjectVisA-120 [4] – a novel 120-participant dataset of spatial street-crossing navigation in virtual reality specifically geared to object-based attention evaluations. The uniqueness of the presented dataset lies in the ethical and safety affiliated challenges that make collecting comparable data in real-world environments highly difficult. ObjectVisA-120 not only features accurate gaze data and a complete state-space representation of objects in the virtual environment, but it also offers variable scenario complexities and rich annotations, including panoptic segmentation, depth information, and vehicle keypoints. We further propose object-based similarity (oSIM) as a novel metric to evaluate the performance of object-based visual attention models, a previously unexplored performance characteristic. Our evaluations show that explicitly optimising for object-based attention not only improves oSIM performance but also leads to an improved model performance on common metrics. In addition, we present SUMGraph, a Mamba U-Net-based model, which explicitly encodes critical scene objects (vehicles) in a graph representation, leading to further performance improvements over several state-of-the-art visual attention prediction methods. The dataset, code and models will be publicly released.**

I. Introduction

Visual attention is a fundamental process that allows humans to focus their limited processing resources on the most relevant stimuli in the environment [1] and has been extensively studied both in cognitive science [2]–[7] and in computer vision [8]–[13]. A major insight from cognitive science is that attention is not only driven by spatial but also object-based factors [1], [3], [4]. Object-based attention is guided by the perceptual grouping of features into coherent objects [14], [15] and allows for selective enhancement of entire objects rather than just isolated spatial locations, facilitating more efficient visual search, recognition, and interaction with the environment. Consequently, cognitive models of visual attention have long incorporated explicit notions of objects [7], [16]. In contrast, research on attention prediction in computer vision largely ignores the object-based nature of human attention allocation. While recent years have seen tremendous progress in saliency prediction [9]–[13], these approaches are limited to predicting attention spatially. Apart from disregarding insights from cognitive science, this focus on spatial attention is also at odds with important requirements of application scenarios. For instance, when pedestrians cross a busy street, it is crucial to predict whether they are likely to see an important object, such as an approaching vehicle, while their spatial attention distribution is less important. A key reason for this is the lack of suitable metrics to evaluate object-based attention prediction methods. Existing metrics such as Normalised Scanpath Similarity (NSS), Kullback-Leibler Divergence (KLD), or Similarity (SIM) only spatially compare predicted and ground truth gaze data without any reference to the objects contained in the scene [19]. A major challenge for establishing a novel, object-based evaluation metric is the need for highly accurate object segmentations, which are usually not available in common human attention datasets [9], [20]–[24]. We address these limitations with three major contributions:

- **First**, we present ObjectVisA-120 (cf. Figure 1) – a novel 120-participant human attention dataset recorded using synthetic street-crossing scenarios in virtual reality (VR), therefore allowing access to highly accurate ground truth segmentations of object instances.
- **Second**, we introduce Object-based Similarity (oSIM), a metric designed to measure the accuracy of attention predictions with respect to scene objects. We show that incorporating oSIM into the training objective consistently improves model performance.
- **Third**, we introduce SUMGraph, to the best of our knowledge, the first attention prediction model for interactive environments that makes use of explicit object information. Evaluations on our novel dataset, demonstrate that SUMGraph outperforms on par with state-of-the-art methods and its ablated versions.

II. Related Works

Human visual attention prediction is studied in a wide variety of scenarios, including scanpath prediction [25], [26], gaze anticipation [27], [28], or gaze following [29]. Most studies on human visual attention prediction, however, focus on generating context-free saliency maps on still images, using ground-truth maps averaged across multiple observers [12], [30]–[32]. Salient features are often extracted by CNN backbone models [33]–[36]. The usage of recurrent models like Long-Short Term Memory (LSTM)-based architectures [32], [37], demonstrated effectiveness in processing both local and long-range visual information. Recently, Transformer-based models [10], [38]–[40] have demonstrated substantial

[1]German Research Center for Artificial Intelligence (DFKI) GmbH, Campus D32, 66123 Saarbruecken, Germany
[2]Max Planck Institute for Intelligent Systems, 70569 Stuttgart, Germany
[3]Institute for Visualization and Interactive Systems (VIS) at Stuttgart University, 70569 Stuttgart, Germany
[4]https://www.kaggle.com/datasets/igorvozniak/objectvisa-120

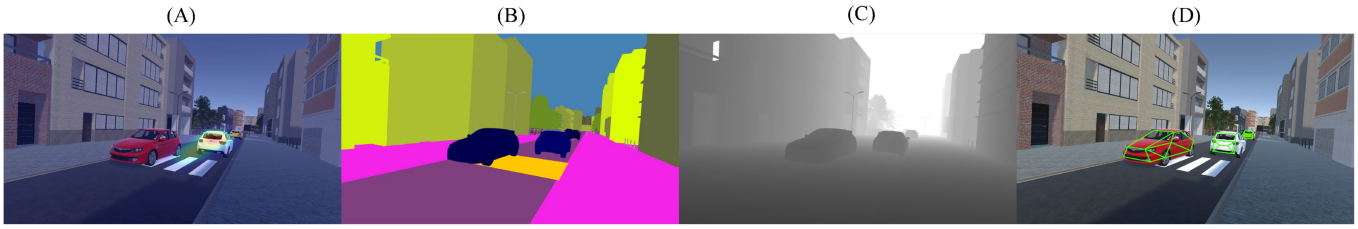

Fig. 1. ObjectVisA-120 dataset: (A) displays a FoV image with an overlaid visual attention map during street-crossing navigation task; (B) presents the affiliated panoptic segmentation, following the CityScapes [17] labeling policy; (C) shows the corresponding depth map; and (D) the keypoints and edges annotations aligned with OpenPifPaf [18] labeling policy.

advancements in saliency prediction, achieving state-of-the-art performance by effectively capturing spatial long-range dependencies. SUM [11] integrates the efficient long-range dependency modeling of Mamba with a U-Net-like architecture, and proposes a Conditional Visual State Space (C-VSS) block which enables token-driven adaptation across various domains and data types (natural scenes, webpages, commercial imagery). SUM achieved state-of-the-art performance on several saliency prediction datasets, including SALICON [9], CAT2000 [21] and MIT1003 [20]. ContextSalNet [8] presents an approach to predict attention distributions in an interactive traffic environment conditioned on task- and personal context information. While deep neural networks can learn object-sensitive features implicitly [41], to the best of our knowledge, no previous works on visual attention prediction in interactive environments combined an explicit object representation with deep image features. Furthermore, due to the lack of an adequate metric and loss function, no previous works were directly optimised for object-based attention prediction.

Metrics for evaluating attention prediction. Depending on the precise task formulation, a wide variety of metrics to evaluate predictions of human visual attention are used. Methods predicting scanpaths (ordered fixation sequences), are often evaluated with sequence comparison metrics such as Dynamic Time Warping or Levenshtein similarity [42]. More relevant to our work are metrics from the field of saliency prediction, as they compare a predicted attention distribution with ground truth gaze [19], [43]. These metrics can be broadly separated into two categories. Distribution-based metrics such as Similarity (SIM), Pearson's Correlation Coefficient (CC), and Kullback-Leibler Divergence (KLD) compare predicted with ground truth attention distributions. Location-based metrics such as Area Under the Curve (AUC) and Normalized Scanpath Saliency (NSS) compare predicted attention distributions to ground truth fixation locations. A common feature of all these metrics is that they operate on an abstract image plane without any notion of the attended objects or their semantics.

Datasets. VR datasets of human gaze behaviour have the advantage of allowing for a wider, more natural field of view compared to what is achievable with screen-based recordings. In addition, they offer the advantage that gaze behaviour is integrated with natural body movements. In case the presented scenes are synthetic, it is possible to access highly accurate panoptic segmentation information, which is crucial for evaluating the quality of object-based attention predictions. In Table I we provide an overview of the most popular VR-based datasets focused on saliency prediction in real and synthetic environments. The free-viewing [27], [44]–[49] and search [27], [44] tasks are well-supported by existing datasets. The recently introduced MoGaze datasets explored an interaction task [50]. HOT3D [51] and AEA [52] are datasets focusing on routine eye-gaze attention with hand movements and navigation/interaction between multiple participants, respectively. ContextSalNet [8] explored pedestrian saliency prediction in street-crossing navigation tasks. However, with only 11 participants, this dataset is significantly smaller compared to ObjectVisA-120 with 120 participants. Moreover, it suffers from limited visual realism and insufficient data annotation support. While offering highly accurate panoptic segmentation annotations would be feasible in the case of synthetic datasets, none of the previous datasets provides this information, making it impossible to effectively study object-based attention on these datasets. Importantly, the street-crossing navigation task involves significant safety and ethical challenges, making it impractical to conduct such studies in real-world environments due to the setup's complexity and potential pedestrian bias.

III. ObjectVisA-120 Dataset

Study. This work builds upon an immersive street-crossing study with bidirectional traffic flow within the Unity 3D engine [56]. It includes scenes with different levels of traffic density and dynamics, presence and absence of crosswalks, as well as different numbers of non-playable characters with different behaviours (risky vs. cautious) added to influence the participants. The diversity of the setup and resulting dataset is ensured by including 120 participants, emphasizing variation across individual characteristics. HTC Vive Pro Eye head-mounted VR goggles has been used, equipped with a wireless adapter to enable unrestricted movement within the effective tracking area (9×8 meters footprint). This makes it the first VR dataset of its kind, pushing the boundary between virtual and real environments to ensure both realism and uniqueness. In total, eye-gaze fixations were recorded from 120 participants, representing a diverse

TABLE I

THE COMPARISON OF SALIENCY AND GAZE PREDICTION DATASETS (VR). IN THE COLUMN TASK, WE USE A SEPARATE NOTATION TO REPRESENT THE TASK, WHERE F STANDS FOR THE FREE VIEWING TASK, S CORRESPONDS TO THE SEARCH TASK, N STANDS FOR NAVIGATION TASK, SC DENOTES STREET-CROSSING CONSTRAINTS, AND I STANDS FOR THE INTERACTION TASK. THE (*) SYMBOL INDICATES PARTIAL OR LIMITED ANNOTATION.

Dataset	Real/Synth.	Task	Depth	Segment	Instance	Panoptic	Skeleton	Sound	Participants	Size
SalientVR [49]	Mixed	F							45	103 videos
SaliencyInVR [53]	Real	F					✓*		169	22 frames
VR-EyeTracking [48]	Real	F							45	208 videos
Salient360! [47]	Real	F							48	60 frames
DeepIntoVS [54]	Synthetic	F							50	6,978 frames
Panonut360 [46]	Mixed	F							50	15 videos
D-SAV360 [45]	Real	F	✓					✓	87	85 videos
DGaze [55]	Synthetic	F							43	5 scenes
FixationNet [27]	Synthetic	S							27	162 trials
MoGaze [50]	Real	I+N	✓*	✓*			✓		7	3 hours
EHTask [44]	Real	F+S							30	15 videos
HOT3D [51]	AR/VR	I					✓*		19	3.7M frames
AEA [52]	Real	I+N						✓	4	∼1Mil. frames
ContextSalNet [8]	Synthetic/VR	N+SC	✓	✓					11	35K frames
ObjectVisA-120 (Ours)	Synthetic	N+SC	✓	✓	✓	✓	✓*	✓	120	7,200 videos, 6.14M frames

range of personal attributes: **nationality** *(even split between German and Japanese)*, **age** *(Range: 20–50 years, μ: 30.56, σ: 9.02)*, **gender** *(60 male, 60 female)*, **height** *(Range: 151.50 - 190.50 cm, μ: 172.46 cm, σ: 9.47 cm)*, **weight** *(Range: 41.60 - 119.60 kg, μ: 66.18 kg, σ: 14.57 kg)*, **driving experience in years** *(Range: 0 - 32, μ: 9.42, σ: 8.85)*, and **VR familiarity** *(81 yes, 39 no)*. Each participant performed 60 street-crossing trials [56]. Prior to the study and after eye-tracking sensor calibration, each participant had a warm-up phase to get used to the environment, physical and virtual boundaries, and adjust to the correspondence between their physical movements and virtual navigation. A detailed description of the study design is available in the corresponding paper [56]. Importantly, none of the contributions presented in this work overlap with those in [56], which describes the general setup and study design, but neither presents a publicly available dataset for attention prediction, nor proposes a framework for object-based attention prediction.

ObjectVisA-120 dataset. While data was recorded at 90Hz, for the purpose of this work, we chose to consider every 3rd frame, resulting in 6.142.167 frames (sampling rate of 30Hz). In addition to the raw field-of-view RGB images (cf. Figure 1, A), the dataset comprises corresponding eye-gaze fixation information (2D/3D) and derived saliency maps (cf. Figure 1, A overlaid) computed individually, panoptic segmentation (covering vehicles, pedestrians, crossings, road and pavement surfaces, buildings, vegetation, signage, and other street furniture as unary objects), depth maps (cf. Figure 1, B-C), and skeleton labels (cf. Figure 1, D) for vehicles. Notable, since the entire state space of the study has been saved, it is possible to generate additional modalities like 2D/3D bounding boxes or add another layer of segmentation. To maintain consistency with previous work, the skeleton labelling corresponds to the sparse variation of the Apollo-Car3D dataset as used in OpenPifPaf [18], which contains 24 keypoints. Incorporating pedestrian skeleton representations is beneficial due to their strong relevance for behavioural and intent inference, whereas extending graph representations to static scene elements offers limited added value and incurs unnecessary computational overhead. The panoptic segmentation on ObjectVisA-120 follows the CityScapes [17] labeling policy. The extensive labeling of ObjectVisA-120 enables multiple computer vision tasks to be performed, including human-like trajectory generation empowered by visual attention [57], providing a significant value to the community. Note that the annotation policy for occluded objects closely reflects real-world perception, where the occluded parts of objects are excluded from all labels (cf. Figure 1). Where necessary, state-of-the-art methods may be used to generate on-demand segmentation labels for real-world datasets and applications. Subsequent fine-tuning on the ApolloCar3D dataset can be applied to derive the required skeleton representations of vehicles, people and animals supporting deployment in real-world settings. To generate ground truth attention maps, we follow previous work [8], where 2D sparse ground truth attention maps were created using the last three fixation points. According to [58], the eccentric angle of the foveal region is approximately 2°, additionally, the eye-tracking accuracy of VR headset[5] ranges from 0.5°-1.1°. Thus, the ground-truth attention maps were generated using Gaussians

[5]HTC Vive Pro Eye: https://developer.vive.com/resources/hardware-guides/vive-pro-eye-specs-user-guide/

with a standard deviation of $3 \times dva$ (degree of visual angle).

IV. Object-based Similarity (oSIM) Metric

We introduce Object-based Similarity (oSIM), the first metric to specifically evaluate object-based attention prediction, relevant for safety-critical settings like pedestrian street-crossing and its interaction with the approaching vehicle (cf. Figure 2). Object-based Similarity is based on the classical Similarity (SIM) metric used to evaluate saliency predictions given image space. Importantly, instead of operating on the raw pixel values without any notion of objects and its semantics, oSIM operates on the level of objects. In essence, the panoptic-based histograms are being compared based on the aggregated sum of pixel intensities across different classes:

$$oSIM = \sum_{mask \in Image} \min\left(\sum_{i \in mask} S_i, \sum_{i \in mask} S_i^{gt}\right) \quad (1)$$

where $mask \in Image$ is a mask corresponding to a single object in the image. $\sum_{i \in mask} S_i$ corresponds to the predicted saliency sum over pixels i in the $mask$, where $\sum_{i \in mask} S_i^{gt}$ stands for the ground truth saliency sum over pixels i in the $mask$, respectively. Intuitively, instead of measuring similarity between predicted and ground truth attention on the image pixels, oSIM measures their similarity on the objects basis. As such, each object has the same influence on the overall oSIM metric, irrespective of its size. In Figure 2, we illustrate the differences between the classic image Similarity (SIM) and the object-based Similarity (oSIM) scores. As shown in columns (B-C), when the predicted saliency remains within the same object as the ground truth, oSIM yields a significantly higher value. In this way, oSIM accounts for the relevance of objects for the human perceptual system (approaching vehicle while crossing in Figure 2), whereas SIM is solely based on spatial image alignment. The proposed metric is applicable beyond the present context and can be extended to other research domains (e.g., driver attention, human–robot collaboration). Beyond panoptic-level segmentation, oSIM supports hierarchical, part-level segmentation, enabling objects to be decomposed into semantically meaningful internal components (e.g., body parts of pedestrians).

V. Methodology

Overview. The overall architecture of SUMGraph is shown in Figure 3 (left). Our approach builds upon the recent Mamba-U-Net-based model of [11] and introduces novel Graph VSS and Graph C-VSS blocks, which extend the Visual State Space (VSS) and Conditional Visual State Space (C-VSS) blocks with a context graph that captures object-based scene information (cf. Figure 3, right). In the SUM architecture, the C-VSS block provides a unified scene representation by encoding semantic and spatial structure, while an attention mechanism selectively focuses on the most relevant objects and regions to support interaction-aware downstream reasoning, those characteristics were transferred to SUMGraph. The encoder produces four hierarchical output representations,

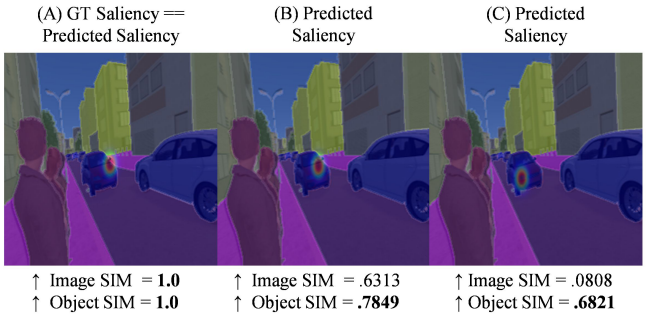

Fig. 2. Comparison of oSIM with the SIM metric. Column (A) shows perfect alignment with the ground truth (SIM = oSIM = 1.0). In (B–C), predictions diverge, and both metrics decline. Unlike SIM, which drops quickly due to spatial misalignment, oSIM also accounts for object-level semantics (e.g., safety-critical approaching vehicle).

where each block is followed by a downsampling layer that halves the spatial dimensions and doubles the number of channels. The decoder contains four Graph C-VSS blocks with two sub-blocks each, whereas the last block contains a single Graph C-VSS block. The patch-expanding layers perform upsampling to match the corresponding initial resolution, and the linear saliency head layer generates the output.

Graph C-VSS block. The novel graph-empowered Graph C-VSS block is shown in Figure 3 (right). $G = (V, E, A)$ denotes the graph structure for each object in the scene, where V represents nodes (e.g. vertices of vehicles), E represents edges (connections between vertices), A corresponds to the global object-based attributes, such as the speed, the distance to the user, and object's direction (towards the user or away). $X_{\text{graph}} \in \mathbb{R}^{M \times f}$ represents node features of the graph, where M is the number of vertices in each object's skeleton, and f is the feature dimension of each node. $A_{\text{object}} \in \mathbb{R}^p$ is the global attribute vector for each object, where p stands for the number of attributes. E_{graph} denotes the edge index representing the graph structure for each object, where $C \in \mathbb{R}^c$ is the residual condition vector for conditional modulation [11], with c dimensions. In the first step (cf. Figure 3, Graph embedding), we process object skeletons and local attributes (e.g. front_wheel_left, as in [18]) using two consecutive graph convolutions. We apply the initial graph convolution to transform node features, followed by the second graph convolution to refine node embeddings:

$$\begin{aligned} X_{\text{graph}}^{(1)} &= \sigma(\text{GCNConv}(X_{\text{graph}}, E_{\text{graph}})) \\ X_{\text{graph}}^{(2)} &= \sigma(\text{GCNConv}(X_{\text{graph}}^{(1)}, E_{\text{graph}})) \end{aligned} \quad (2)$$

where $X_{\text{graph}}^{(1)}, X_{\text{graph}}^{(2)} \in \mathbb{R}^{M \times h}$ represent the hidden states with h as the number of hidden dimension, and σ stands for ReLU activation function. Later, we project the object's global attribute vector A_{object} to the same hidden dimension h: $A'_{\text{object}} = \sigma(W_{\text{attr}} A_{\text{object}})$, where $W_{\text{attr}} \in \mathbb{R}^{h \times p}$ is a learnable weight matrix. Upon completion, we aggregate node features

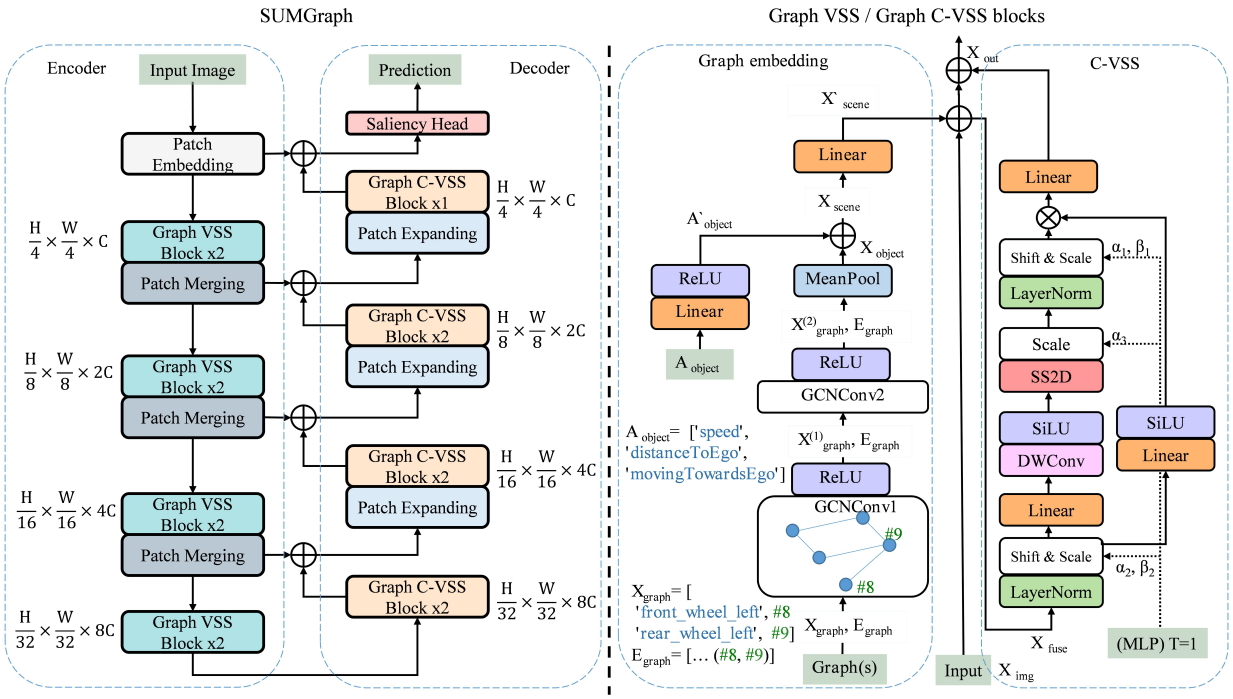

Fig. 3. Left: Overall architecture of the proposed SUMGraph model for visual attention prediction. Right: novel Graph VSS (fusion of graph and image features, in encoder) and Graph C-VSS (fusion of graph and conditional image features, in decoder) blocks for integration of additional contextual information. $\otimes$ stands for the element-wise produce operation, $\oplus$ is the element-wise addition.

within each object's skeleton using global mean pooling:

$$X_{\text{object}} = \text{MeanPool}(X^{(2)}_{\text{graph}}) + A'_{\text{object}} \tag{3}$$

where $X_{\text{object}} \in \mathbb{R}^h$ is the combined feature vector for each object, integrating both node and global attribute information. To obtain a scene-level context vector X'_{scene}, we aggregate and project across all object embeddings:

$$X_{\text{scene}} = \frac{1}{K}\sum_{k=1}^{K} X^{(k)}_{\text{object}} \qquad X'_{\text{scene}} = W_{\text{scene}} X_{\text{scene}} \tag{4}$$

where K stands for the number of objects. The scene-level context vector X'_{scene} is fused with the visual features X_{img} obtained from the C-VSS block as described in [11].

Conditional VSS. Lastly (cf. Figure 3, C-VSS), the conditional modulation step remains unchanged [11], allowing for adaptations to various domains. The C-VSS block is in line with [11], which allows for further extensions since the modulation of the graph-empowered feature maps through dynamic scaling and shifting operations are adjusting feature activations. α_i is a scaling factor, β_i is a shifting factor, where the modulated feature map is equal to $\alpha_i \odot$ (original feature map) $+ \beta_i$. In this work, we used a single set token $T = 1$. The token is fed into the Multi-Layer Perceptron (MLP) model, allowing it to be conditioned on the characteristics of natural scene-based eye tracking data type. The main objective of MLP is to regress α_i and β_i parameters based on the data types [11]. The regressed parameters dynamically adjust the model's behaviour as in [11], enhancing its performance and generalization.

Loss function. Following previous work [11], our model employs a loss function that integrates several components [10]–[12], [59]. We extend the loss with the $oSIM$ term to optimize for object-based attention explicitly:

$$\begin{aligned} Loss = \lambda_1 \cdot KLD(S^{gt}, S) + \lambda_2 \cdot CC(S^{gt}, S) + \\ \lambda_3 \cdot SIM(S^{gt}, S) + \lambda_4 \cdot NSS(F^{gt}, S) + \\ \lambda_5 \cdot MSE(S^{gt}, S) + \boldsymbol{\lambda_6 \cdot oSIM(S^{gt}, S, P)} \end{aligned} \tag{5}$$

where S^{gt} stands for the ground truth attention map, S corresponds to the predicted attention map, F denotes the binary fixation map, and P is the panoptic segmentation. The λ_i are the weighting coefficients. Each term targets a specific evaluation criterion commonly used to compare predictions of human attention maps to ground truth [10]–[12], [59]. KLD stands for Kullback-Leibler Divergence, CC for the correlation coefficient, SIM for the Similarity metric, NSS for the Normalized Scanpath Similarity, and MSE denotes Mean Squared Error.

VI. Experiments

Implementation details. To model the relevant scene aspects in street-crossing scenarios, we encode all vehicles that are in the participant's field of view into an object graph. In particular, depending on their visibility to the user, we encode up to 24 2D keypoints as nodes in a graph. Each node represents a local attribute, such as *front_wheel_left*, following the labelling scheme presented in [18]. Furthermore, each

TABLE II

THE UPPER SECTION PRESENTS RESULTS FOR MODELS WITHOUT FINE-TUNING ON OBJECTVISA-120, WHILE THE MIDDLE SECTION INCLUDES FINE-TUNED MODELS, INCLUDING OURS. SUMGRAPH (WITH SCALE) REFERS TO THE EXPLICIT UP/DOWNSAMPLING OF GRAPH INFORMATION (SEE FIGURE 3, LEFT). THE THIRD EVALUATION HIGHLIGHTS PERFORMANCE GAINS ON OBJECTVISA-120 OBTAINED BY INTEGRATING THE OSIM LOSS INTO THE OVERALL TRAINING OBJECTIVES.

Method	CC ↑	KLD ↓	AUC ↑	SIM ↑	NSS ↑	oSIM ↑
1. w/o finetuning						
TempSAL [59]	0.0234	3.9136	0.6497	0.0412	0.1823	0.3457
TranSalNet [10]	0.2340	2.9500	0.8812	0.1231	2.1626	0.3766
ContextSalNet [8]	0.0093	3.9165	0.6177	0.0406	0.0760	0.3484
SUM [11]	0.2961	2.6820	0.9160	0.2003	3.0513	0.4304
2. w/ finetuning						
TempSAL [59]	0.4342	1.8510	0.9606	0.3192	5.9976	0.5767
TranSalNet [10]	0.4348[3rd]	1.7059	0.9672	0.3414	6.4809	0.5961
ContextSalNet [8] w/ [11] loss	0.4335	1.6941	0.9680	0.3462	**6.6447**	0.6042[2nd]
SUM [11]	**0.4722**	1.7062	0.9662	0.3470	6.2607	0.5909
SUMGraph (Ours)	0.4564	1.6747[2nd]	**0.9683**	**0.3568**	6.4357	**0.6086**
SUMGraph (Ours) (no global attr.)	0.4508	1.6729	0.9681	0.3482	6.3898	0.6019
SUMGraph (Ours) & Scale	0.4643[2nd]	**1.6581**	**0.9683**	0.3481[2nd]	6.5241[2nd]	0.6025
3. w/ finetuning & oSIM loss						
TempSAL [59]	+0.0044	+0.0006	+0.0101	-0.0029	+0.0943	+0.0022
TranSalNet [10]	+0.0135	-0.0234	+0.0002	+0.0085	+0.1648	+0.0072
ContextSalNet [8] w/ [11] loss	+0.0044	-0.0098	+0.0005	+0.0011	-0.0652	+0.0012
SUM [11]	-0.0034	-0.0424	+0.0019	+0.0001	+0.1112	+0.0097
SUMGraph (Ours)	-0.0023	-0.0130	+0.0001	+0.0062	+0.0148	+0.0045

graph in the scene contains global attributes: speed, distance, and direction. It is important to note that graph processing is performed only when there are relevant objects within the pedestrian's FoV. If no objects are present, the system defaults to the baseline model [11].

Train/validation/test splits. We divided the data into training, validation, and test sets using a 70%, 10%, and 20% allocation strategy, which corresponds to 84, 12, and 24 participants, respectively. To avoid introducing biases in the training, we ensure an equal number of male/female and Japanese/German participants within the splits. The splits were verified for demographic balance using a Kolmogorov–Smirnov test on key variables (age and height).

Training details. We implemented our approach in the PyTorch framework and trained it on a cluster with 10 × A100 (80vGB) for 15 epochs with early stopping after 4 epochs. We used Adam optimisation, with the initial learning rate set to 1×10^{-4}, with a learning rate scheduler that decreased learning rate by a factor of 10 after four epochs. We employed Distributed Data-Parallelization (DDP), where the overall batch size is set to 750 samples. In line with [11], we scale the resolution to 256×256. The optimal values for the weighting coefficients λ_i in the loss function (cf. Equation 5) are set to: $\lambda_1 = 10$, $\lambda_2 = -2$, $\lambda_3 = -1$, $\lambda_4 = -1$, $\lambda_5 = 1$, $\lambda_6 = -1$ and correspond to [11], whereas the λ_6 is determined through multiple test trials. The MLP architecture's layers match the baseline [11]. For hidden layers within the graph embedding network, we apply Kaiming uniform initialization. This method is specific for layers utilizing ReLU activations, maintaining the variance of input signals and preventing issues like vanishing or exploding gradients [60]. Our SUMGraph architecture uses SUM [11] weights pre-trained on six different datasets [9], [20]–[24].

Metrics. In line with previous work on visual saliency- and human attention prediction [8]–[11], [13], [31], we measure the quality of predicted attention with both location- and distribution-based metrics [19]. Location-based metrics, such as NSS and AUC (Area Under the ROC Curve), represent ground truth with a binary fixation map. Distribution-based metrics, like CC, SIM, and KLD, evaluate the similarity between predicted and ground truth attention distributions. In addition, we evaluate predictions with novel $oSIM$ metric.

A. Quantitative results

We present comparisons against state-of-the-art saliency prediction approaches in Table II. We make use of the weights published by the original authors for initialization, except in the case of ContextSalNet [8], where these weights were not available and we needed to fall back to VGG [34] weights. The low performance achieved by models without fine-tuning highlight the difference between classical saliency estimation and prediction of sparse attention maps on ObjectVisA-120. The top section of Table II presents the default performance of

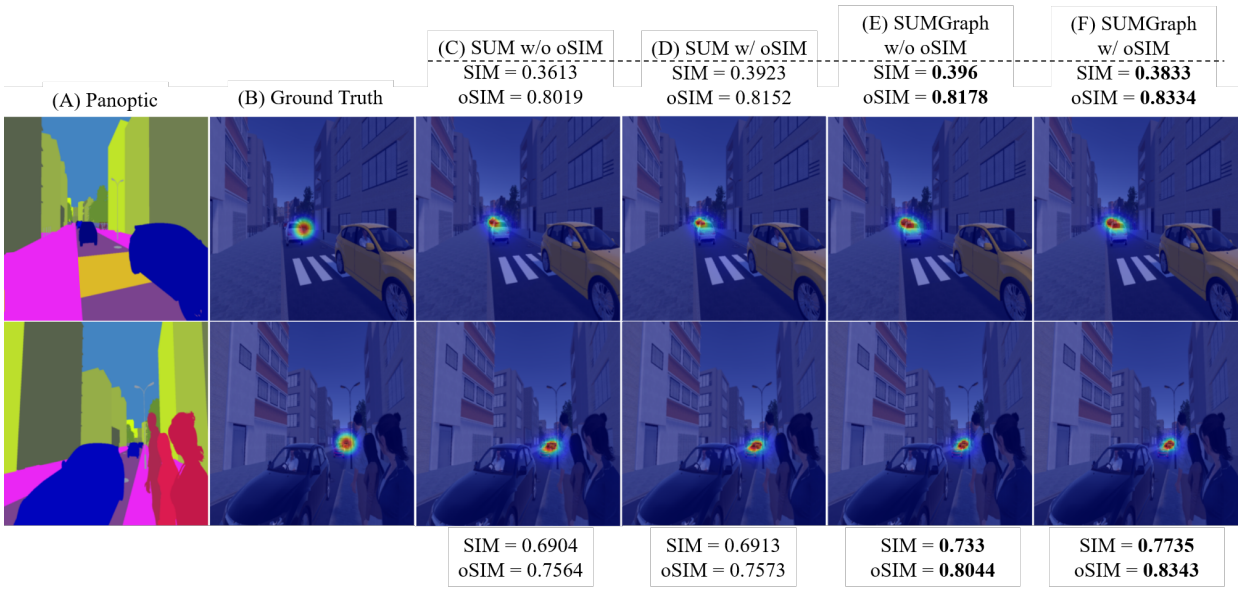

Fig. 4. Qualitative visualizations of random samples. For clarity in performance improvements, both SIM and oSIM scores are provided. Column (A) presents the panoptic labels, (B) shows the ground truth attention map. The SUMGraph model (D) demonstrates improved performance compared to the baseline, with the incorporation of oSIM (columns C vs. D, and E vs. F) leading to consistent improvements.

models without fine-tuning. In the middle section of Table II, we show the results of models after training on ObjectVisA-120 dataset. SUMGraph performs on par with state-of-the-art models and achieves the best performance in the KLD, AUC, SIM, oSIM metrics. In summary, **SUMGraph outperforms prior methods in 28 out of 30 metrics**, demonstrating its effectiveness across both pixel- and object-level benchmarks. Moreover, the inclusion of the additional oSIM loss term as part of the total loss function contributes to a **performance boost in 25 out of 30 metrics** (cf. Table II, 3rd evaluations). This ablation study highlights not only the value of SUMGraph's architectural design but also the general applicability, robustness, and consistency of our object-centric (oSIM) loss formulation. Furthermore, we investigate the impact of global attributes in our novel Graph (C-)VSS blocks. Our findings show that incorporating global attributes in the graphs improves the performance of SUMGraph. We also highlight the difference between the default and explicit graph structuring, where graphs are scaled to match block resolution.

B. Qualitative results

In Figure 4, we randomly selected samples to showcase the performance of our method versus the baseline. As observed (cf. Figure 4, columns C vs. D, E vs. E), the **utilization of oSIM as an additional training objective improves performance, while also resulted in improving SIM score.** To conclude, it guides the model's predictions toward capturing the semantic context of the object more accurately, which is especially relevant for safety-critical settings like street-crossing. **SUMGraph shows better performance in comparison to SUM** [11], reflecting the influence of the used graph structure (aligned with how pedestrians estimate approaching vehicles and its affiliated features) in safety-critical scenarios like street-crossing.

VII. Limitations and Future Work

While our novel dataset, evaluation metric, and method improve the state-of-the-art in visual attention prediction, several limitations remain. Our training data is predominantly urban-centric, which may impact generalization to rural, extreme-weather environments or daylight conditions. Future work should mitigate these shortcomings. The participants in our study were aged between 20 and 50 years, and the generalization to other age groups remains to be investigated. While our dataset features participants from two distinct cultural backgrounds (DE and JP), a wider scope of backgrounds should be investigated.

VIII. Conclusion

In this work, we introduced the ObjectVisA-120 dataset, a novel VR resource applicable to diverse tasks, with a focus on visual attention prediction in street-crossing scenarios, an area challenging due to complexity, safety and ethical considerations. We proposed an object-based Similarity (oSIM) metric to capture object-driven attention, offering a perspective more aligned with human perception. Empirical results show that integrating oSIM into the loss function improves performance and advances modeling of visual attention. We also introduced the SUMGraph model, which exploits contextual graph information to improve predictive accuracy and achieve state-of-the-art performance.

IX. Acknowledgment

This work was supported by the European Regional Development Fund (EFRE; EFRE-AuF-0000866) and in part by the

European Union's Horizon Europe research and innovation programme (No. 101076360). The initial study recording has been funded by the German Ministry for Research and Education (BMBF; 01IW17003), and partly by the New Energy and Industrial Technology Development (NEDO; JPNP18010).

REFERENCES

[1] Z. Chen, "Object-based attention: A tutorial review," *Attention, Perception, & Psychophysics*, 2012.

[2] E. Sood and at al., "Improving neural saliency prediction with a cognitive model of human visual attention," in *CogSci*, 2023.

[3] P. De Graef, "Semantic effects on object selection in real-world scene perception," 2005.

[4] N. Roth, M. Rolfs, O. Hellwich, and K. Obermayer, "Objects guide human gaze behavior in dynamic real-world scenes," *PLOS Computational Biology*, 2023.

[5] M. M. Chun and J. M. Wolfe, "Visual attention," *Blackwell handbook of sensation and perception*, 2005.

[6] P. R. Roelfsema, "Cortical algorithms for perceptual grouping," *Annu. Rev. Neurosci.*, 2006.

[7] W. X. Schneider, "Vam: A neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action," *Visual Cognition*, 1995.

[8] I. Vozniak and et al., "Context-empowered visual attention prediction in pedestrian scenarios," in *Proceedings of the IEEE/CVF WACV*, 2023.

[9] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proceedings of the IEEE CVPR*, 2015.

[10] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, "Transalnet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, 2022.

[11] A. Hosseini, A. Kazerouni, S. Akhavan, M. Brudno, and B. Taati, "Sum: Saliency unification through mamba for visual attention modeling," in *IEEE/CVF WACV*. IEEE, 2025.

[12] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *Computer Vision–ECCV 2020*. Springer, 2020.

[13] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, 2017.

[14] S. P. Vecera, "Grouped locations and object-based attention: Comment on egly, driver, and rafal (1994)." 1994.

[15] R. Egly, J. Driver, and R. D. Rafal, "Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects." *Journal of Experimental Psychology: General*, 1994.

[16] D. D. Salvucci, "An integrated model of eye movements and visual encoding," *Cognitive Systems Research*, 2001.

[17] M. Cordts and at al., "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE CVPR*, June 2016.

[18] S. Kreiss, L. Bertoni, and A. Alahi, "Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[19] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[20] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th ICCV*. IEEE, 2009.

[21] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015.

[22] L. Jiang and et al., "Does text attract attention on e-commerce images: A novel saliency prediction dataset and method," in *CVPR*, 2022.

[23] Y. Jiang and at al., "Ueyes: Understanding visual saliency across user interface types," in *CHI*, 2023.

[24] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *Journal of Vision*, 2014.

[25] X. Chen, M. Jiang, and Q. Zhao, "Predicting human scanpaths in visual question answering," in *Proceedings of the IEEE/CVF CVPR*, 2021.

[26] Z. Yang and et al., "Predicting goal-directed human attention using inverse reinforcement learning," in *CVPR*, 2020.

[27] Z. Hu, A. Bulling, S. Li, and G. Wang, "Fixationnet: Forecasting eye fixations in task-oriented virtual environments," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[28] P. Müller, E. Sood, and A. Bulling, "Anticipating averted gaze in dyadic interactions," in *ACM Symposium on Eye Tracking Research and Applications*, 2020.

[29] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" *Advances in neural information processing systems*, 2015.

[30] 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*, 2002.

[31] M. Kümmerer, T. S. Wallis, and M. Bethge, "Deepgaze ii: Reading fixations from deep features trained on object recognition," *arXiv preprint arXiv:1610.01563*, 2016.

[32] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transactions on Image Processing*, 2018.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*.

[34] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE CVPR*, 2016.

[36] M. Kummerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low-and high-level contributions to fixation prediction," in *ICCV*, 2017.

[37] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE TIP*, 2018.

[38] A. Hosseini and at al., "Brand visibility in packaging: A deep learning approach for logo detection, saliency-map prediction, and logo placement analysis," *Discover Applied Sciences*, 2025.

[39] K. Han and et al., "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[40] Y. A. D. Djilali, K. McGuinness, and N. O'Connor, "Learning saliency from fixations," in *Proceedings of the IEEE/CVF WACV*, 2024.

[41] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," in *Proceedings of the ICLR*.

[42] R. Fahimi and N. D. Bruce, "On metrics for measuring scanpath similarity," *Behavior Research Methods*, vol. 53, pp. 609–628, 2021.

[43] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," 2018.

[44] Z. Hu, A. Bulling, S. Li, and G. Wang, "Ehtask: Recognizing user tasks from eye and head movements in immersive virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[45] E. Bernal-Berdun and et al., "D-sav360: A dataset of gaze scanpaths on 360° ambisonic videos," *IEEE TVCG*, 2023.

[46] Y. Xu and at al., "Panonut360: A head and eye tracking dataset for panoramic video," in *ACM Multimedia Systems Conference*, 2024.

[47] E. David and at al., "The salient360! toolbox: Processing, visualising and comparing gaze data in 3d," in *ETRA*, 2023.

[48] Y. Xu and et al., "Gaze prediction in dynamic 360 immersive videos," in *IEEE CVPR*, 2018.

[49] S. Wang and et al., "Salientvr: Saliency-driven mobile 360-degree video streaming with gaze information," in *MobiCom*, 2022.

[50] P. Kratzer and et al., "Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze," *IEEE RAL*, 2020.

[51] P. Banerjee and et al., "Hot3d: Hand and object tracking in 3d from egocentric multi-view videos," *CVPR*, 2025.

[52] Z. Lv and et al., "Aria everyday activities dataset," 2024.

[53] V. Sitzmann and et al., "Saliency in vr: How do people explore virtual environments?" *IEEE TVCG*, 2018.

[54] U. Celikcan, M. B. Askin, D. Albayrak, and T. K. Capin, "Deep into visual saliency for immersive vr environments rendered in real-time," *Computers & Graphics*, 2020.

[55] Z. Hu and et al., "Dgaze: Cnn-based gaze prediction in dynamic scenes," *IEEE TVCG*, 2020.

[56] J. Sprenger and et al., "Cross-cultural behavior analysis of street-crossing pedestrians in japan and germany," in *IEEE IV*, 2023.

[57] I. Vozniak, M. Klusch, A. Antakli, and C. Müller, "Infosalgail: Visual attention-empowered imitation learning of pedestrian behavior in critical traffic scenarios." in *IJCCI*, 2020.

[58] L. Wang, X. Shi, and Y. Liu, "Foveated rendering: A state-of-the-art survey," *Computational Visual Media*, 2023.

[59] B. Aydemir, L. Hoffstetter, T. Zhang, M. Salzmann, and S. Süsstrunk, "Tempsal-uncovering temporal information for deep saliency prediction," in *Proceedings of the IEEE/CVF CVPR*, 2023.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE ICCV*, 2015.