Context-empowered Visual Attention Prediction in Pedestrian Scenarios

Igor Vozniak, Philipp Müller, Lorena Hell, Nils Lipp, Ahmed Abouelazm, Christian Müller
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{igor.vozniak,philipp.mueller,lorena.hell,nils.lipp,ahmed.abouelazm,christian.mueller}@dfki.de

Abstract

Effective and flexible allocation of visual attention is key for pedestrians who have to navigate to a desired goal under different conditions of urgency and safety preferences. While automatic modelling of pedestrian attention holds great promise to improve simulations of pedestrian behavior, current saliency prediction approaches mostly focus on generic free-viewing scenarios and do not reflect the specific challenges present in pedestrian attention prediction. In this paper, we present Context-SalNET, a novel encoder-decoder architecture that explicitly addresses three key challenges of visual attention prediction in pedestrians: First, Context-SalNET explicitly models the context factors urgency and safety preference in the latent space of the encoder-decoder model. Second, we propose the exponentially weighted mean squared error loss (ew-MSE) that is able to better cope with the fact that only a small part of the ground truth saliency maps consist of non-zero entries. Third, we explicitly model epistemic uncertainty to account for the fact that training data for pedestrian attention prediction is limited. To evaluate Context-SalNET, we recorded the first dataset of pedestrian visual attention in VR that includes explicit variation of the context factors urgency and safety preference. Context-SalNET achieves clear improvements over state-of-the-art saliency prediction approaches as well as over ablations. Our novel dataset will be made fully available and can serve as a valuable resource for further research on pedestrian attention prediction.

1. Introduction

The visual behavior of pedestrians in a street crossing situation is influenced by the concrete layout of the street [13, 41], but also to a large extent by the existence of the time pressure [48, 2]. Due to its importance to traffic safety, pedestrian attention has been studied extensively in human science [2, 14, 20, 61]. Automatic prediction of pedestrian attention can open up the possibility to create more realistic training environments both for humans and autonomous agents. Furthermore, it will help to more accurately model and understand critical traffic scenarios [25]. Automatic prediction of human attention has received great interest in the computer vision community since more than two decades [4, 3]. Significant progress has been made especially on datasets employing a context-agnostic, free viewing paradigm with static images [28, 37, 18]. These models predict saliency maps that are averages of gaze behavior obtained from several observers for a given static image. Much fewer works proposed visual attention prediction models in an interactive environment that take into account navigation or search task characteristics [51, 6]. Until now, no approach for the prediction of pedestrian attention in an interactive environment exists that is able to account for the context factors that are specific to pedestrian behavior (i.e. urgency and safety). Likewise, to the best of our knowledge, no publicly available dataset to train such a model exists.

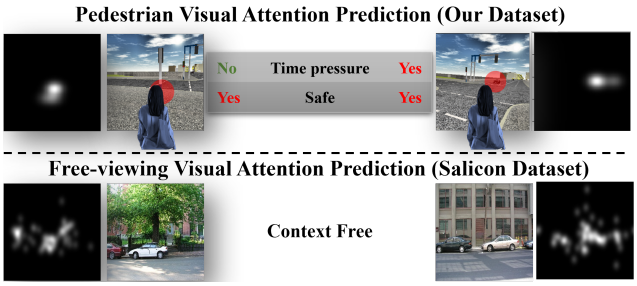

Figure 1. In contrast to classical free-viewing visual attention prediction on static images, pedestrian visual attention prediction is highly context dependent. Furthermore, saliency maps generated from pedestrian attention are more sparse compared to free-viewing saliency maps that are aggregated over several subjects viewing the same image.

We close this gap by proposing the first method and dataset for pedestrian attention prediction in street-crossing scenarios. Whereas, we do not address the task of salient object detection[1], which is a well-established area. Our approach consists of an encoder-decoder architecture and addresses three key challenges that distinguish pedestrian

[1] https://paperswithcode.com/task/salient-object-detection

attention prediction from the classical scenario of saliency prediction on static images. First, to capture the context dependence of pedestrian attention, we augment the hidden state of the encoder-decoder with information on the urgency and the safety preference of the pedestrian. Second, as opposed to the static image scenario, only a few pixels are activated in the saliency maps of visual attention in an interactive environment. To better cope with this fact, we propose the exponentially weighted Mean Squared Error (*ew-MSE*). This loss punishes the network less for wrong high-saliency predictions. Third, neural saliency models are commonly trained on multiple datasets to reduce model uncertainty and achieve the highest performance. As only our novel dataset for pedestrian attention prediction is available as of now, we explicitly model the epistemic uncertainty of the model [33].

The specific contributions of this work are threefold: **First**, we propose Context-SalNET, the first approach that addresses the task of pedestrian attention prediction. **Second**, we record the first publicly available dataset pedestrian attention prediction. The dataset consists of diverse street-crossing scenarios recorded in virtual reality and explicitly varies the context factors urgency and safety preference. The dataset consists of 528 different scenarios formed based on German In-depth Accident Study (GIDAS) report with a different street layouts and considered in this work context factors. Additionally, the complexity has been extended with layout components like safety-island and multiple lanes in moving directions [62, 63]. Thus, the total number of recorded frames is 35K, which are additionaly labelled with the context information of 11 participants in total. The full dataset will be made publicly available for future research. **Third**, we conduct comprehensive quantitative and qualitative evaluations on this novel dataset, showing the effectiveness of our context modelling approach as well as our proposed ew-MSE loss and the utility of modelling epistemic (statistical) uncertainty. In addition, Context-SalNET outperforms a current state-of-the-art saliency prediction approach [18] trained on the same dataset and improves over the current best saliency prediction approach on the MIT/Tübingen benchmark [36] which was trained on a much larger collection of datasets (no training code available for a direct comparison) [44].

2. Related Work

Our work is related to the state of the art in human attention prediction, and, more specifically to task-dependent visual attention prediction.

2.1. State of the Art in Visual Attention Prediction

Most work on human attention prediction has focused on the task of predicting context-free saliency maps on images [28, 38, 12, 18, 60]. The ground truth for this task is a gaze density map averaged over many observers for a given image. The current state-of-the-art approaches on the influential MIT saliency benchmark [36] are DeepGaze IIE [44] (1st), UniSal [18] (2nd) and SalFBNet [16] (3rd). DeepGaze IIE improves over its previous version DeepGazeII [38] by fusing different backbone networks, thus, the exact training setup is essential to avoid performance bias. At the time of submission, no open-source implementation of DeepGaze IIE was available that would allow us to train the network on our dataset. [16] proposed SalFBNet which learns a saliency distribution using pseudo-ground-truth, and subsequent fine-tuned on existing datasets. No implementation was publicly available at the time of submission. UniSal [18] on the other hand utilizes domain adaptation to train a single model for both image- and video based saliency generation. We choose UniSal as a context-free baseline method, as the authors provide an open-source implementation, allowing for training on our dataset. The majority of saliency generation models [44, 18, 15] are following similar architectural designs with encoder and decoder. UniSal [18], for instance, consists of a MobileNet V2 [56] encoder, followed by concatenation with learned priors, Bypass-RNN, and a decoder with skip connections, fusion and smoothing layers. The usage of domain-adaptive modules allows for domain-shift between the image and video saliency datasets. Note that a large body of work exists on video saliency prediction [46, 68, 75, 30, 47, 45, 39], as well as on egocentric saliency prediction [64, 26, 70]. Recent works in this field commonly extract temporal features like optical flow, recurrences, or 3D convolutions [47, 45, 39, 64]. While these techniques are applicable to our scenario, our focus in this work is to investigate pedestrian attention prediction informed by context attributes, as well as our proposed ew-MSE loss that addresses the challenge of sparse ground truth saliency. To isolate these aspects and to increase the comparability to the current state of the art in saliency prediction, we choose to leave the integration of temporal features to future work.

2.2. Task-dependent Visual Attention Prediction

A large number of works show the importance of the task context in human visual attention allocation [73, 5, 40, 22, 21]. For example, [21] studied the effects of free-viewing, as well as search- and navigation tasks on visual attention in a virtual environment.They found that navigation, in contrast to free-viewing and search tasks, produces fixations which are more center located. Moreover, in [40], authors studied the relation between eye movements and day-to-day activities like food preparation tasks, indicating nearly all eye movements are made to task-relevant objects. It confirms the high effect of the ”top-down” component, whereas the bottom-up attributes like color, shape, and size

contribute very little to "intrinsic saliency". Interestingly, authors in [7] classified the type of driving (manual vs autonomous) given gaze patterns recorded in a virtual study. All these works scientifically confirm the importance and influence of contextual factors on visual attention.

Despite the importance of context in human attention allocation, only few attention prediction methods explicitly model context. An early computational model predicting task-dependent visual attention prediction was introduced in [51]. The authors incorporated task-dependent top-down modulation with bottom-up saliency extraction to model participants' attention when playing video games. Later, [6] instructed subjects to navigate in simulated environments (2D and 3D). Task attributes were modelled by a gist descriptor [65, 55], as well as by the subjects' current motor actions. More recently [74] proposed a task-dependent saliency prediction model for web pages. Their CNN models task-specific and a task-free aspects of attention in separate branches of the network. In contrast to previous work which explicitly modelled different kinds of tasks (e.g. navigation versus free viewing [6]), we for the first time explicitly model the qualitative aspects urgency and safety preference within the framework of pedestrian navigation tasks.

3. Method

The overall architecture of Context-SalNET (Figure 2) consists of an encoder-decoder neural network, which is conditioned on the context attribute information (Input 2, Figure2). To cope with the fact of sparse saliency maps in the interactive pedestrian scenario, we introduce the exponentially weighted MSE (*ew-MSE*) loss. Furthermore, we model epistemic uncertainty in accordance to [33] to account for the fact that the available data for pedestrian attention prediction is limited.

3.1. Context-SalNET Architecture

Our encoder-decoder architecture is inspired by [50], but introduces a novel concatenation layer between encoder and decoder that introduces context information (see Figure 2). The encoder consists of blocks of CNN layers. Each block is followed by a max-pooling layer. The **concatenation** bottleneck layer is composed of an embedding layer to encode context information, followed by a fully connected layer with dropout and batch normalization in order to improve the optimization landscape [57] and to solve for the internal covariate shift. The **decoder** mirrors the encoder except for the addition of upsampling layers to achieve the corresponding resolution. In order to maintain fine-grained spatial resolution, we add skip connections as described in [23] between blocks 5 and 6 of the encoder and blocks 1 and 2 of the decoder, respectively. In preliminary experiments, these skip connections proved to have a large impact on performance. Except for the Sigmoid output layer, we use ReLU activation functions [49]. The output of the Context-SalNET is a saliency map indicating the attention focus of the pedestrian. To avoid overfitting and enable probabilistic inference, dropout is applied to blocks 4 and 6 of the encoder and 1-3 blocks of the decoder.

3.2. Exponentially Weighted MSE Loss

Compared to classical saliency prediction, where ground truth saliency maps are aggregated over several observers of a static image, ground truth saliency maps in pedestrian attention prediction are much more sparse, only containing few non-zero entries. To account for this, we modify the mean squared error (MSE) loss that is commonly used in saliency prediction by exponentially weighting it with the magnitude of the prediction. The resulting exponentially weighted MSE (*ew-MSE*) loss penalises high predictions less, combating the tendency of vanilla MSE to resort to predicting zeros as a result of the sparse ground truth. Formally,

$$\text{ew-MSE} = \frac{1}{N}\sum_{i=1}^{N} \exp(-\hat{y}_i)(y_i - \hat{y}_i)^2 \qquad (1)$$

where $\hat{y}$ denotes the model output, y the ground-truth, and N corresponds to the number of output pixels of $\hat{y}$.

3.3. Model Uncertainty

While human attention is influenced by image evidence as well as context factors, the non-deterministic simulation state space including dynamic vehicles, pedestrian, dynamic traffic lights, and obstacles introduces substantial stochastic components. Due to the interactive nature of our environment, resulting in different head angles, height, and body orientations, every FoV image and corresponding eye-gaze fixation is unique. In contrast to classical saliency prediction [28], this stochasticity can not be averaged out and the resulting sparseness of data leads to a large model (i.e. epistemic) uncertainty [27]. To address this challenge, we for the first time propose to model uncertainty in a human attention prediction model. We report Epistemic uncertainty inline with [33], where the dropout variational inference is adopted during the inference phase to approximate the distributions over the network parameters. Hence, both the training and inference phases are conducted with activated dropout in order to sample from the stochastic posterior, thus, to derive mean and variance over each predicted pixel. In preliminary experiments, we also evaluated the effects of modelling aleatoric uncertainty, but we observed no performance improvements.

3.4. Training Details

We train Context-SalNET according to Equation 1, where AUC metrics is utilized as an early stop criteria. Given a total of $\sim 35K$ pedestrian visual attention images, we set the ratio of 80% to 20% for the training-

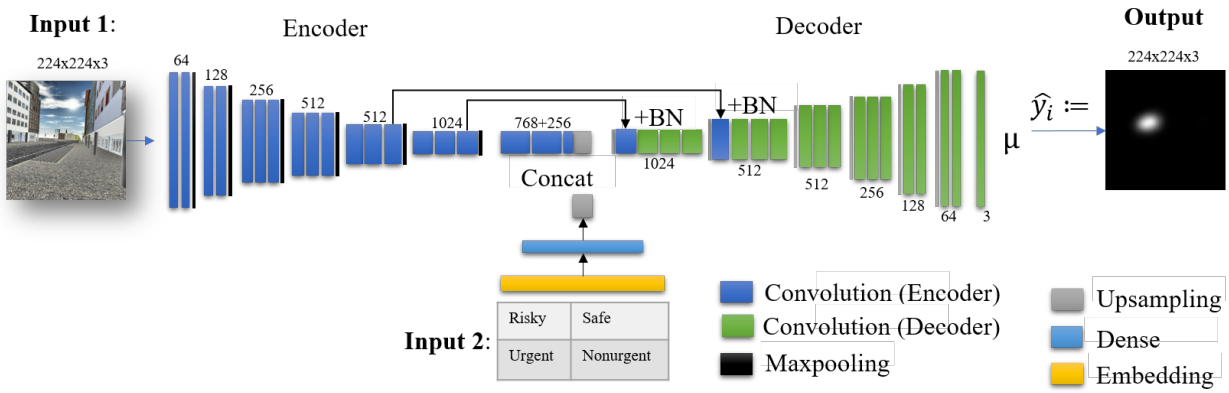

Figure 2. CNN encoder-decoder architecture of the generator network. The input consists of: 1) field-of-view image and; 2) sample of corresponding frame specific context attributes. The objective is to output corresponding attention map. Skip connections are indicated with arrows between encoder and decoder, where batch normalization (BN) is applied to account for different data distributions.

validation data splits, where the testing is performed on the unseen and subject-specific dataset. We utilized Adam [34] optimizer with the loss rate of 10^{-5} and the batch size of 96 images across the entire workflow of this work. The input image resolution is set to $224x224x3$, as in line with VGG16 architecture. During leave-one-subject-out cross-validation training, we used clusters with Tesla A100 (40vGB) and Quadro RTAX6000 (48vGb), 2-Core CPU, and 128GB RAM, each. The weights of Encoder (batches of convolution layers 1-5) are initialized from VGG16 [59] for faster convergence and to overcome insufficient gradients. UniSal [69] is trained on our dataset in line with the initial training pipeline, allowing for a fair comparison to Context-SalNET. Both UniSal and introduced Context-SalNET rely on backbone networks, ModelNet V2 [56] and VGG16 [59] respectively, where both pre-trained on the ImageNet dataset[2].

4. Dataset

The focus of this research is goal-directed pedestrian behavior in traffic scenarios and the influence of context attributes, i.e. high-level aspects. Thus, the targets of the research are achievable by utilizing synthetic environments even if a lack in photorealism introduces a domain gap to real images. However, on a more general note, a domain gap exists in any combination of training vs testing settings [71] and is beyond the scope of this work.

4.1. Context Attributes and Scenarios

We manipulated two context factors in the navigation task that are highly relevant to pedestrian scenarios. First, we vary the *time pressure* to which participants are exposed. Second, we instruct participants to perform their task in either a *risky or a safe* way. To avoid ambiguity we would like to stress that in our work, we use the notion of *context attributes* (time pressure, riskiness), which differs from the notion of *task* (e.g. free-viewing vs. search vs. navigation) used in some previous works [21].

To record a realistic dataset with a high relevance to challenging real-life traffic situations, we base our scenarios on the German in-Depth Accident Study[3] (GIDAS) which identified nine classes of critical street-crossing scenarios (see Figure 4) that specify street layouts and traffic participants (pedestrian, vehicle, and potential obstacles). We added three more additional scenarios to cover additional urban scene complexities like safety island between two opposite direction lanes, multiple lanes in each driving direction, and crossings involving consecutive traffic lights. This helps to additionally increase the variation in participants' visual behavior and the number of opportunities to realise safe or unsafe street crossing behavior. To further increase the realism, we embedded these scenarios into a virtual reconstruction (digital-twin) of a real city with accurate street layouts including traffic lights, pedestrian street crossings, bicycle lanes, parking spots, as well as reconstructions of the actual buildings.

4.2. Recording Setup

To simulate the traffic scenarios we chose the open-source simulation software OpenDS[4], whereas we considered other simulators like Carla[5], LGSVL[6], and GTA5[7], however by the time of conducting the recording session it was missing some important features, e.g., support of pedestrian-centric VR goggles with eye-gaze record-

[2]https://www.image-net.org/

[3]GIDAS - https://www.gidas.org/start.html
[4]OpenDS - https://opends.dfki.de
[5]Carla - https://carla.org/
[6]LGSVL - https://www.svlsimulator.com/
[7]GTA5-https://www.gta5-mods.com/scripts/driving-mode-selection

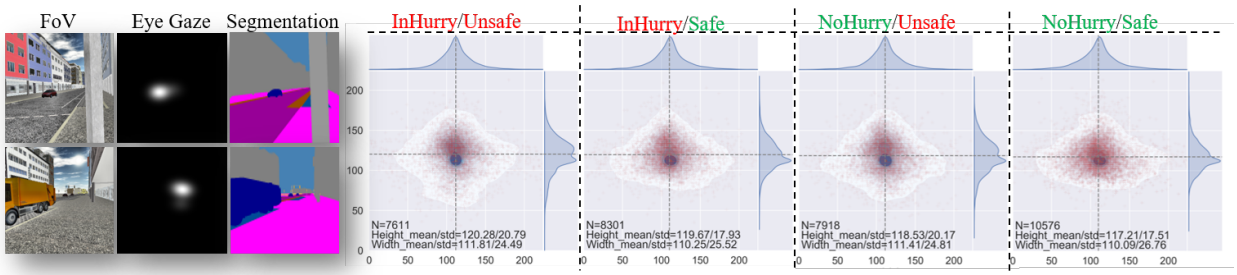

Figure 3. Samples of the recorded dataset and additionally extracted information. Left: samples of recorded field-of-view (FoV) images, corresponding eye gaze information, and segmentation maps (inline with CityScapes color schema with extensions caused by fine-grained scene-related details). Right: accumulative distribution of fixation points across all subjects with a context-based split. The N indicates the number of samples of a specific context type, where mean and std values are self-explanatory.

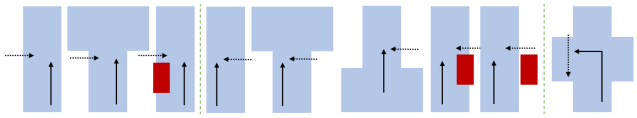

Figure 4. Traffic scenario layouts based on German In-Depth Accident Study (GIDAS). The solid vector stands for the approximated moving direction of the vehicle(s), where the dashed vector indicates an approaching direction of the subject. The red rectangle stands for an obstacle on the way.

ing, digital-twin setup and workflow control. Moreover, OpenDS has the key advantage that it will allow other researchers to replay the recorded pedestrian trajectories that we plan to publish since the raw as well as post-processed dataset will be released. This will increase both the reproducibility of our research and the value of the dataset to investigate novel research questions. Figure 3 shows samples of recorded images (top row), namely RGB frames with corresponding post-processed saliency and segmentation maps respectively. Besides, the corresponding depth maps are also recorded and to be released. Thus, might assist in future empirical studies. The bottom row in Figure 3 illustrates distributions of aggregated across all subjects fixation points based on context factors. Moreover, provided visualizations of context based eye-gaze distributions is inline with the empirical studies, where subjects tend to look further away with vertical $mean = 120, 28$ and $td = 20, 79$ in case of "InHurry/Unsafe" setup (Figure 3, bottom left) to look for more potential hazards like approaching vehicles. Thus, the perception of higher risk leads to more cautions behaviour and more detailed assessment of the traffic before crossing the street. Such factors are i.e., the absence of traffic signals and zebra crosswalks, lower time-to-collision, faster cars, wider streets with several lanes [54]. Whereas, in the case of "NoHurry/Safe" setup (Figure 3 , bottom right), we observe smallest vertical $std = 17, 51$ with $mean = 117, 21$ and highest horizontal $std = 26, 76$. The aim of our study is to model the impact of context attributes on human visual attention as opposed to low-level modelling of fine-grained image features, hence, the rendering capabilities of the gaming engine is not central for our research.

To achieve a maximum degree of realism and immersion in the simulation, we made use of virtual reality goggles. We employed the HTC Vive Eye[8] featuring an integrated eye tracker. Furthermore, we used two Base Stations 2.0 and collected user input via a Xbox One controller. The camera rotation and translation coordinates are taken directly from VR goggles. Thus, pitch, jaw, and roll angles as well as actions like jumping or squatting are supported in our setup, which makes it a well-suited benchmark test due to the underlined complexity. To balance resolution with simulation performance, we choose a sampling rate of 3 frames-per-second to record the subject's current field of view, her current attention, as well as the corresponding semantic segmentation map.

4.3. Procedure

We recruited 15 participants out of which four withdrew due to feelings of motion sickness. Prior to the study, all participants gave informed consent for participation and for inclusion of their pseudonymized data in the dataset. For each participant, the eye tracker was calibrated at the start of the recording session. Subsequently, participants spent 5 minutes in the simulation to familiarize themselves with the controls. Participants were presented with four blocks of all 12 traffic scenarios each. Each block realized one combination of time pressure (yes/no) and riskiness (high/low). Each of the 12 trials in a block started by visually indicating the target location for 5 seconds. Data recording started after these 5 seconds have passed. In each trial, participants are able to move *forward, backward, left and right* and head movements were mapped to camera movements along *pitch, yaw, and roll* angles. Hence, we collected a unique dataset with 528 scenarios, resulting in $\sim 35K$ of unique FoV images and corresponding segmentation, saliency, depth maps as well as xml files to store simulated

[8]HTC Vive Eye - https://www.vive.com/

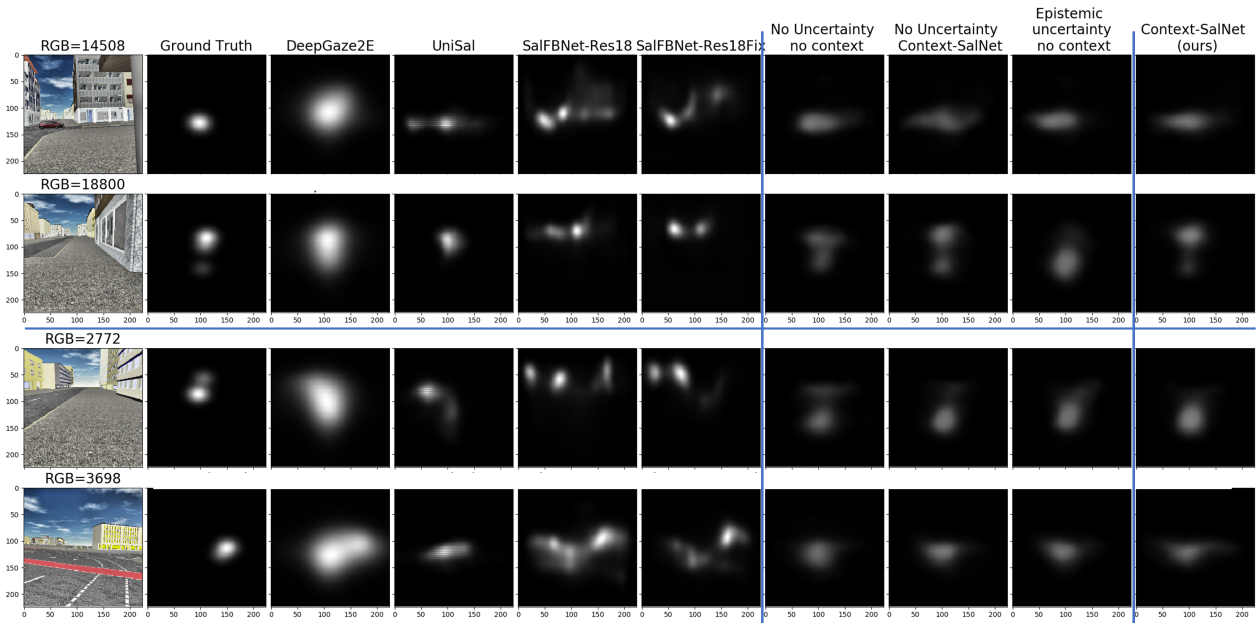

Figure 5. Qualitative analysis of randomly selected best with ($AUC > 0.99\%$) and worse ($AUC < 0.70\%$) samples. The rows 1-2 stand for the best samples, while rows 3-4 correspond to the worse visual predictions. Provided renderings serve two purposes: 1) qualitative baseline evaluations with columns 2-5, and 9; 2) qualitative ablation evaluations with columns 6-9. Column 1 stands for the input FoV image with a unique RGB sequence ID.

related information e.g., position, speed, orientation of the body, and head.

5. Experiments

5.1. Pre-Processing

Saliency ground truth information consists of fixation sequences, namely recorded X and Y coordinates projected to the image plane. Following previous work that made use of fixation maps in pedestrian navigation scenarios [66], we aggregate the gaze locations obtained from the last three frames to create a representation of participants' current focus of attention. To arrive at continuous ground truth attention maps, we follow the saliency map computation in [67] with *degree of visual angle* set to $dva = 9.3$. We discount the previous attention points in intensity to allow the neural network to account for previous information, but to also overfitting to the additional auxiliary information. On the images recorded from the simulator, we applied Contrast Limited Adaptive Histogram Equalization (AE) [76] to obtain even color distributions across images, which improves invariance to unique attributes of the scene, e.g. uniquely colored buildings. For optimal alignment with community standards, the labelling scheme of our segmentation maps matches the CityScapes [11] labelling convention except where we had to introduce new classes that are missing in CityScapes (e.g. bicycle lanes, parking slots).

5.2. Quantitative Evaluation

Using our novel dataset, we evaluate Context-SalNET on the task of pedestrian attention prediction both against state-of-the-art saliency prediction approaches as well as against ablations. We also evaluate a context-free version of Context-SalNET against state-of-the-art approaches on SALICON [32] in order to estimate its performance on an established saliency benchmark dataset.

Metrics. In line with prior works [4, 67, 8], we adopt the following evaluation metrics: AUC-Judd (AUC-J), AUC-Borji (AUC-B), shuffled AUC (s-AUC), Similarity Metric (SIM), Linear Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS) and Kullback-Leibler Divergence (KLDiv) [9].

Comparison to SOTA saliency models. Table 1 shows the evaluation results of Context-SalNET against the latest publicly available state-of-the-art approaches on the MIT benchmark[36], namely UniSal [18]. We include the evaluation results of the pre-trained model provided by DeepGaze IIE [44] (ranked 1st[9]) and SalFBNet [16] as training code is not publicly available. Note however, that these results are not comparable to the other methods as DeepGaze IIE, for instance, uses several backbone networks and is trained on different datasets as well as it utilises center bias information computed on the target dataset.

Context-SalNET clearly outperformed both the center bias baseline as well as UniSal [18] (ranked 2nd on MIT

Method	AUC-J ↑	s-AUC ↑	AUC-B ↑	NSS ↑	SIM ↑	CC ↑	KLDiv ↓
DeepGaze2E [44]	0.9526	0.6313	**0.7842**	2.7158	0.3726	0.5146	0.1326
SalFBNet-R18 [16]	0.9050	0.5418	0.5818	1.7376	0.2761	0.3225	0.2393
SalFBNet-R18Fix [16]	0.9014	0.5340	0.5591	1.6121	0.2605	0.2902	0.2646
Center Bias	0.8360	0.5101	0.5381	1.0940	0.2231	0.2130	**0.1322**
UniSal [18]	0.9388	0.5631	0.5961	2.7097	0.3978	0.4537	0.3755
Context-SalNET (ours)	**0.9605**	**0.6654**	0.7723	**3.3048**	**0.4646**	**0.5843**	0.1690

Table 1. Leave-one-subject-out baseline evaluation results using different evaluation metrics. Arrows indicate whether higher (↑) or lower (↓) is better. DeepGaze2E/SalFBNet is shown separately as it is composed of several backbone networks and trained on different training data than the other approaches. Bold numbers indicate the best results.

Method	AUC-J ↑	s-AUC ↑	AUC-B ↑	NSS ↑	SIM ↑	CC ↑	KLDiv ↓
No uncertainty							
vanilla MSE, no context	0.9587	0.6589	0.7642	3.2163	0.4508	0.5710	0.1782
vanilla MSE	0.9580	0.6567	0.7537	3.2441	0,4544	0.5722	0.1795
no context	0.9581	0.6542	0.7595	3.1670	0.4496	0.5645	0.1764
Context-SalNET	0.9584	0.6570	0.7600	3.1879	0.4472	0.5652	0.1770
Epistemic uncertainty							
vanilla MSE, no context	0.9575	0.6552	0.7503	3.2682	0,4620	0.5711	0.1959
vanilla MSE, random context	0.9524	0.6459	0.7427	3.1028	0.4422	0.5498	0.2068
vanilla MSE	0.9588	0.6581	0.7577	3.2888	**0.4661**	0.5799	0.1933
no context	0.9592	0.6630	**0.7744**	3.2458	0.4548	0.5770	0.1679
random context	0.9599	0.6577	0.7588	3.2479	0.4566	0.5739	**0.1642**
Context-SalNET (ours)	**0.9605**	**0.6654**	0.7723^{2nd}	**3.3048**	0.4646^{2nd}	**0.5843**	0.1690^{3rd}

Table 2. Leave-one-subject-out ablation evaluation results using different evaluation metrics. We present combinations of three ablation dimensions: uncertainty modelling, context modelling (either removing the context concatenation layer or by providing random context information), and vanilla mean squared error (MSE) instead of our proposed exponentially weighted MSE.

Benchmark[9]) across 6 out of 7 metrics. Context-SalNET clearly improves over DeepGaze2E in 5 out of 7 metrics, while it is close in AUC-B.

Ablation Study. The results of our ablation study are summarized in Table 2. To quantify the effect of context modelling we created two different ablated versions: *random context* consists of the exact same architecture as Context-SalNET, but receives random context information as input. For the *no context* condition on the other hand we removed the context network and the context concatenation layer, resulting in fewer network parameters. Crucially, Context-SalNET clearly improves over both ablation conditions. It improves in 6 out of 7 metrics over the random context condition and in 5 out of 7 metrics over the no context condition. We also evaluated the impact of our novel ew-MSE loss by comparing to vanilla MSE. Here, Context-SalNET improved in 6 out of 7 metrics over the variant with vanilla MSE. Finally, we observe clear improvement for epistemic uncertainty modelling. The ablation of Context-SalNET without uncertainty modelling (i.e. no dropout at test time) is inferior in all 7 metrics.

Performance on SALICON. While general saliency prediction is not the focus of this paper, we evaluate a context-free version of Context-SalNET on SALICON[10] to obtain an estimate on how our architecture performs on this task in relation to SOTA approaches (see Table 3). More precisely, *Context-free-SalNET* consists of our encoder-decoder architecture including ew-MSE loss and epistemic uncertainty modelling, but without the context network and context concatenation layer. Context-free-SalNet shows results that are close to the state of the art, and even outperfroms the other methods by a significant margin in the CC metric.

5.3. Qualitative Evaluation

In Figure 5, we randomly selected success (top two rows) and failure cases (bottom two rows) of Context-SalNET. DeepGaze2E strongly relies on center bias priors, leading to an over-estimation of the extent of the attention focus. UniSal shows more accurate predictions in comparison to the ground-truth. SalFBNet models show comparable to

[9]https://saliency.tuebingen.ai/results.html

[10]http://salicon.net/challenge-2017/

Method	AUC-J ↑	s-AUC ↑	IG ↑	NSS ↑	SIM ↑	CC ↑	KLDiv ↓
MD-SEM [19]	0.864	**0.746**	0.660	**2.058**	0.868	0.774	0.568
EMLNet [29]	0.866	0.746	0.736	2.050	0.886	0.780	0.520
SAM-Res [12]	0.865	0.741	0.538	1.990	0.899	0.793	0.610
ACNet-V17 [42]	0.866	0.739	**0.854**	1.948	0.896	0.786	**0.228**
DI-Net [72]	0.862	0.739	0.195	1.959	**0.902**	0.795	0.864
MSI-Net [35]	0.865	0.736	0.793	1.931	0.889	0.784	0.307
GazeGAN [10]	0.864	0.736	0.720	1.899	0.879	0.773	0.376
FBNet [15]	0.843	0.706	0.343	1.687	0.785	0.694	0.708
SalFBNet-Res18 [15]	0.867	0.733	0.805	1.950	0.888	0.773	0.303
SalFBNet-Res18Fixed [15]	**0.868**	0.740	0.839	1.952	0.892	0.772	0.236
Context-free-SalNet (ours)	0.862	0.730	0.750	1.833	0.763	**0.870**	0.308

Table 3. Evaluation results comparing Context-free-SalNET (since no context factors are presented for SALICON benchmark, we removed it from the architecture) to current top methods on SALICON free-viewing saliency benchmark [31]. The lower values for KLDiv indicate better performance. For CC metric the values should approach either 1 (positive correlation) or -1 (negative correlation), where 0 means no correlation. The higher values, the better the performance rule is applied to the remaining metrics.

UniSal results, but includes more false positive predictions. Context-SalNET is able to produce predictions close to the ground truth without relying heavily on center bias or producing false positive predictions. Columns 7-9 show qualitative results for the ablation conditions, supporting the utility of each method contribution. Rows 3-4 show samples of low-performing attention predictions, which holds across baseline and ablation evaluations for all models. Confirming the statement, that visual attention prediction for pedestrians in street-crossing scenarios is indeed a challenge due to the immense state-space of this navigation task.

6. Discussion

6.1. Applications

Our method can be applied in all areas where precise modelling of pedestrian behavior is desired. This includes driving simulators that can be used to train humans, but also the generation of training data for autonomous driving, where modelling and predicting pedestrian behavior is a key challenge [52]. Furthermore, it can be used for critical scenario generation as an extension to [66] in order to better understand and make predictions about dangerous traffic situations. Accurately modelling the attention of pedestrians in such scenarios can help to improve the generation of plausible walking trajectories [1]. Finally, by introducing certain extensions, it can be even applied to solve real-world problems as in [43].

6.2. Limitations and Future Work

While we showed clear improvements of our methods over previous approaches, a number of aspects need to be addressed in future work. While we evaluated the impact of context factors on pedestrian attention, in the future our approach should also be extended to include additional person-specific factors that are relevant in traffic scenarios [58, 21, 17, 5]. Joint attention between the driver of a car and the pedestrian is crucial in traffic situations [53], hence, an explicit representation could empower attention prediction models. Furthermore, it will be important to include different roles of traffic participants (e.g. driver, bicyclist) in our model. Additional challenges arise from the geographical location, since traffic scenarios can significantly differ throughout the world. Authors in [24] summarized significant cultural behaviour differences in street-crossing tasks between German and Japanese people. Moreover, while virtual reality is an effective research tool to collect close-to-natural data, future work also needs to find ways to validate results obtained in VR in the real world. The impact of different pre-trained weights, ModelNet for instance, on final performance is an interesting research question.

7. Conclusion

We introduced Context-SalNET, a novel context driven visual attention generation approach for street-crossing pedestrian scenarios. In evaluations of a newly recorded VR dataset of street crossing tasks including several task context factors, Context-SalNET outperformed a state-of-the-art saliency prediction model and ablation experiments demonstrated our methods' ability to effectively exploit task context factors. The dataset, including driving simulation setups and recorded gaze behavior will be made publicly available. Together with our novel method, this dataset will be an important building block for future research on pedestrian attention prediction.

8. Acknowledgements

This work has been funded by the German Ministry for Research and Education (BMBF) in the project REACT (grant no. 01IW17003). P. Müller was funded by BMBF (grant no. 01IS20075).

References

[1] André Antakli, Igor Vozniak, Nils Lipp, Matthias Klusch, and Christian Müller. Hail: Modular agent-based pedestrian imitation learning. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 27–39. Springer, 2021.

[2] Pavlo Bazilinskyy, Dimitra Dodou, and Joost CF De Winter. Visual attention of pedestrians in traffic scenes: A crowdsourcing experiment. In *International Conference on Applied Human Factors and Ergonomics*, pages 147–154. Springer, 2021.

[3] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[4] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012.

[5] Ali Borji and Laurent Itti. Defending yarbus: Eye movements reveal observers' task. *Journal of vision*, 14(3):29–29, 2014.

[6] Ali Borji, Dicky N Sihite, and Laurent Itti. Probabilistic learning of task-specific visual attention. In *2012 IEEE Conference on computer vision and pattern recognition*, pages 470–477. IEEE, 2012.

[7] Iuliia Brishtel, Stephan Krauß, Thomas Schmidt, Jason Raphael Rambach, Igor Vozniak, and Didier Stricker. Classification of manual versus autonomous driving based on machine learning of eye movement patterns. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 692–697, 2022.

[8] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. 2015. *URL: http://saliency. mit. edu/results_mit300. html*, 12:13, 2014.

[9] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018.

[10] Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo, and Patrick Le Callet. How is gaze influenced by image transformations? dataset and model. *IEEE Transactions on Image Processing*, 29:2287–2300, 2019.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[12] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[13] Brigitte Cambon de Lavalette, Charles Tijus, Sébastien Poitrenaud, Christine Leproux, Jacques Bergeron, and Jean-Paul Thouez. Pedestrian crossing decision-making: A situational and behavioral approach. *Safety science*, 47(9):1248–1253, 2009.

[14] Joost de Winter, Pavlo Bazilinskyy, Dale Wesdorp, Valerie de Vlam, Belle Hopmans, Just Visscher, and Dimitra Dodou. How do pedestrians distribute their visual attention when walking through a parking garage? an eye-tracking study. *Ergonomics*, 64(6):793–805, 2021.

[15] Guanqun Ding, Nevrez İmamoğlu, Ali Caglayan, Masahiro Murakawa, and Ryosuke Nakamura. Fbnet: Feedback-recursive cnn for saliency detection. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pages 1–5. IEEE, 2021.

[16] Guanqun Ding, Nevrez Imamouglu, Ali Caglayan, Masahiro Murakawa, and Ryosuke Nakamura. Salfbnet: Learning pseudo-saliency distribution via feedback convolutional networks. *arXiv preprint arXiv:2112.03731*, 2021.

[17] Aurélie Dommes, M-A Granié, M-S Cloutier, Cécile Coquelet, and Florence Huguenin-Richard. Red light violations by adult pedestrians and other safety-related behaviors at signalized crosswalks. *Accident Analysis & Prevention*, 80:67–75, 2015.

[18] Richard Droste, Jianbo Jiao, and J Alison Noble. Unified image and video saliency modeling. In *European Conference on Computer Vision*, pages 419–435. Springer, 2020.

[19] Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. How much time do you have? modeling multi-duration saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4473–4482, 2020.

[20] Duane R Geruschat, Shirin E Hassan, and Kathleen A Turano. Gaze behavior while crossing complex intersections. *Optometry and vision science*, 80(7):515–528, 2003.

[21] Jacob Hadnett-Hunter, George Nicolaou, Eamonn O'Neill, and Michael Proulx. The effect of task on visual attention in interactive virtual environments. *ACM Transactions on Applied Perception (TAP)*, 16(3):1–17, 2019.

[22] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005.

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

[24] Lorena Hell, Janis Sprenger, Matthias Klusch, Yoshiyuki Kobayashi, and Christian Müller. Pedestrian behavior in japan and germany: A review. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1529–1536. IEEE, 2021.

[25] Andreas Homann, Torsten Bertram, Markus Buß, Martin Keller, and Karl-Heinz Glander. Definition of critical traffic scenarios to evaluate trigger criteria for collision avoidance. In *18. Internationales Stuttgarter Symposium*, pages 671–681. Springer, 2018.

[26] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.

[27] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

[28] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[29] Sen Jia and Neil DB Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 95:103887, 2020.

[30] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. Deepvs: A deep learning based video saliency prediction approach. In *Proceedings of the european conference on computer vision (eccv)*, pages 602–617, 2018.

[31] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[32] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

[33] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

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

[35] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020.

[36] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/.

[37] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

[38] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016.

[39] Qiuxia Lai, Wenguan Wang, Hanqiu Sun, and Jianbing Shen. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE Transactions on Image Processing*, 29:1113–1126, 2019.

[40] Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25-26):3559–3565, 2001.

[41] Lucie Lévêque, Maud Ranchet, Jonathan Deniel, Jean-Charles Bornard, and Thierry Bellet. Where do pedestrians look when crossing? a state of the art of the eye-tracking studies. *IEEE Access*, 8:164833–164843, 2020.

[42] Pengqian Li, Xiaofen Xing, Xiangmin Xu, Bolun Cai, and Jun Cheng. Attention-aware concentrated network for saliency prediction. *Neurocomputing*, 429:199–214, 2021.

[43] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from simulation for trajectory prediction. In *European Conference on Computer Vision*, pages 275–292. Springer, 2020.

[44] Akis Linardos, Matthias Kummerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, 2021.

[45] Panagiotis Linardos, Eva Mohedano, Juan Jose Nieto, Noel E O'Connor, Xavier Giro-i Nieto, and Kevin McGuinness. Simple vs complex temporal recurrences for video saliency prediction. *arXiv preprint arXiv:1907.01869*, 2019.

[46] Sophie Marat, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *International journal of computer vision*, 82(3):231, 2009.

[47] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2394–2403, 2019.

[48] Barbara A Morrongiello, Michael Corbett, Jessica Switzer, and Tom Hall. Using a virtual environment to study pedestrian behaviors: How does time pressure affect children's and adults' street crossing behaviors? *Journal of pediatric psychology*, 40(7):697–703, 2015.

[49] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[50] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Cristian Canton Ferrer, Jordi Torres, Kevin McGuinness, and Noel E OConnor. Salgan: Visual saliency prediction with adversarial networks. In *CVPR Scene Understanding Workshop (SUNw)*, 2017.

[51] Robert J Peters and Laurent Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[52] Atanas Poibrenski, Matthias Klusch, Igor Vozniak, and Christian Müller. M2p3: multimodal multi-pedestrian path prediction by self-driving cars with egocentric vision. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 190–197, 2020.

[53] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 264–269. IEEE, 2017.

[54] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Understanding pedestrian behavior in complex traffic scenes. *IEEE Transactions on Intelligent Vehicles*, 3(1):61–70, 2017.

[55] Laura Walker Renninger and Jitendra Malik. When is scene identification just texture recognition? *Vision research*, 44(19):2301–2311, 2004.

[56] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[57] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help op-

timization? In *Proceedings of the 32nd international conference on neural information processing systems*, pages 2488–2498, 2018.

[58] John Shen and Laurent Itti. Top-down influences on visual attention during listening are modulated by observer sex. *Vision research*, 65:62–76, 2012.

[59] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[61] Hagai Tapiro, Anat Meir, Yisrael Parmet, and Tal Oron-Gilad. Visual search strategies of child-pedestrians in road crossing tasks. *Proceedings of the Human Factors and Ergonomics Society Europe*, 2014.

[62] Hagai Tapiro, Tal Oron-Gilad, and Yisrael Parmet. The effect of environmental distractions on child pedestrian's crossing behavior. *Safety science*, 106:219–229, 2018.

[63] Hagai Tapiro, Tal Oron-Gilad, and Yisrael Parmet. Pedestrian distraction: The effects of road environment complexity and age on pedestrian's visual attention and crossing behavior. *Journal of safety research*, 72:101–109, 2020.

[64] Hamed Rezazadegan Tavakoli, Esa Rahtu, Juho Kannala, and Ali Borji. Digging deeper into egocentric gaze prediction. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 273–282. IEEE, 2019.

[65] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[66] Igor Vozniak, Matthias Klusch, André Antakli, and Christian Müller. Infosalgail: Visual attention-empowered imitation learning of pedestrians in critical traffic scenarios. In *Proceedings of 12th IEEE International Conference on Neural Computation Theory and Application*. IEEE, 2020.

[67] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4903, 2018.

[68] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):220–237, 2019.

[69] Yanyu Xu, Shenghua Gao, Junru Wu, Nianyi Li, and Jingyi Yu. Personalized saliency and its prediction. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2975–2989, 2018.

[70] Kentaro Yamada, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki. Attention prediction in egocentric video using motion and visual saliency. In *Pacific-Rim Symposium on Image and Video Technology*, pages 277–288. Springer, 2011.

[71] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *CVPR*, 2021.

[72] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. A dilated inception network for visual saliency prediction. *IEEE Transactions on Multimedia*, 22(8):2163–2176, 2019.

[73] Alfred L Yarbus. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer, 1967.

[74] Quanlong Zheng, Jianbo Jiao, Ying Cao, and Rynson WH Lau. Task-driven webpage saliency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 287–302, 2018.

[75] Sheng-hua Zhong, Yan Liu, Feifei Ren, Jinghuan Zhang, and Tongwei Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *Twenty-seventh AAAI Conference on Artificial Intelligence*, 2013.

[76] Karel Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics gems*, pages 474–485, 1994.