"Looking at the right stuff" - Guided semantic-gaze for autonomous driving

Anwesan Pal
UC San Diego
a2pal@eng.ucsd.edu

Sayan Mondal
UC San Diego
samondal@eng.ucsd.edu

Henrik I. Christensen
UC San Diego
hichristensen@eng.ucsd.edu

Abstract

*In recent years, predicting driver's focus of attention has been a very active area of research in the autonomous driving community. Unfortunately, existing state-of-the-art techniques achieve this by relying only on human gaze information, thereby ignoring scene semantics. We propose a novel Semantics Augmented GazE (**SAGE**) detection approach that captures driving specific contextual information, in addition to the raw gaze. Such a combined attention mechanism serves as a powerful tool to focus on the relevant regions in an image frame in order to make driving both safe and efficient. Using this, we design a complete saliency prediction framework - **SAGE-Net**, which modifies the initial prediction from SAGE by taking into account vital aspects such as distance to objects (depth), ego vehicle speed, and pedestrian crossing intent. Exhaustive experiments conducted through four popular saliency algorithms show that on **49/56 (87.5%)** cases - considering both the overall dataset and crucial driving scenarios, SAGE outperforms existing techniques without any additional computational overhead during the training process. The augmented dataset along with the relevant code are available as part of the supplementary material.*[1]

1. Introduction

Cameras are one of the most powerful sensors in the world of robotics as they capture detailed information about the environment, and thus can be used for object detection [51, 33] and segmentation [48, 49] - something that is much harder to achieve with a basic range sensor. However, an image/video may contain some irrelevant information. Therefore, there is a need to filter out these unimportant regions and instead, learn to focus our "attention" on parts of the image which are necessary to solve the task at hand. This is crucial for autonomous driving scenarios, where a vehicle should pay more attention to other vehicles, pedestrians and cyclists present in its vicinity, while ignoring the inconsequential objects. Upon successfully identifying the objects of interest, the controller driving the vehicle only needs to attend to them in order to make optimal decisions.

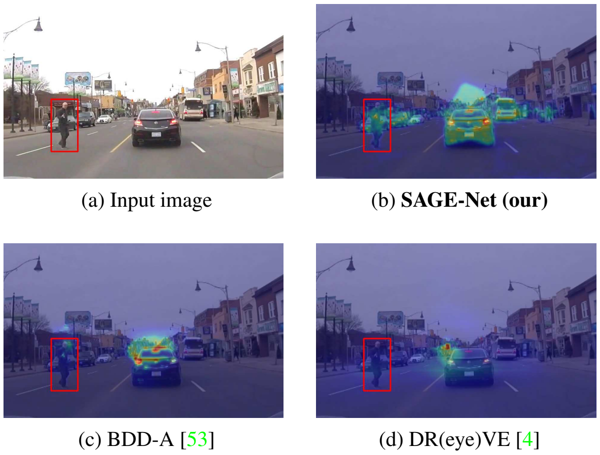

(a) Input image
(b) **SAGE-Net (our)**
(c) BDD-A [53]
(d) DR(eye)VE [4]

Figure 1: Predicted saliency map for different models (Best viewed in color). The bounding box shows a pedestrian illegally crossing the road and is prone to accident. While other models only capture the car ahead (partially), our proposed model can **completely** learn to detect both the car and the crossing pedestrian.

We propose a novel framework for predicting driver's focus of attention through a learnt saliency map by taking into consideration the semantic context in an image. Typical saliency prediction algorithms [39, 40, 53, 45] in driving scenarios rely only on human-gaze information, either through an in-car [4], or in-lab [53] setting. However, gaze by itself does not completely describe everything a driver should attend to, mainly due to the following reasons:

(i) **Peripheral vision:** Humans have a tendency to rely on peripheral vision, thus giving us the ability to fixate our eyes on one object while attending to another. This cannot be captured by an eye-tracking device. Thus, only in-car driver gaze [4] does not convey sufficient information. While the in-lab annotation does alleviate this problem to some extent [53] by aggregating the gazes of multiple in-

[1] Supplementary material including code and videos are available at https://sites.google.com/eng.ucsd.edu/sage-net.

dependent observers, it does not completely remove it since that relies on real human gaze too.

(ii) **Single focus:** When a human driver realizes that the trajectory of an incoming car or pedestrian is not likely to collide with that of the ego-vehicle, their tendency is to shift the gaze away from the oncoming traffic as it approaches. This is a major cause of accidents. To address that, we propose a method of tracking the motion of every driving-relevant object by detecting it's instances until it goes beyond the field of view of the camera. This is possible because the limitation of a human's ability of single focus does not apply to an autonomous vehicle system.

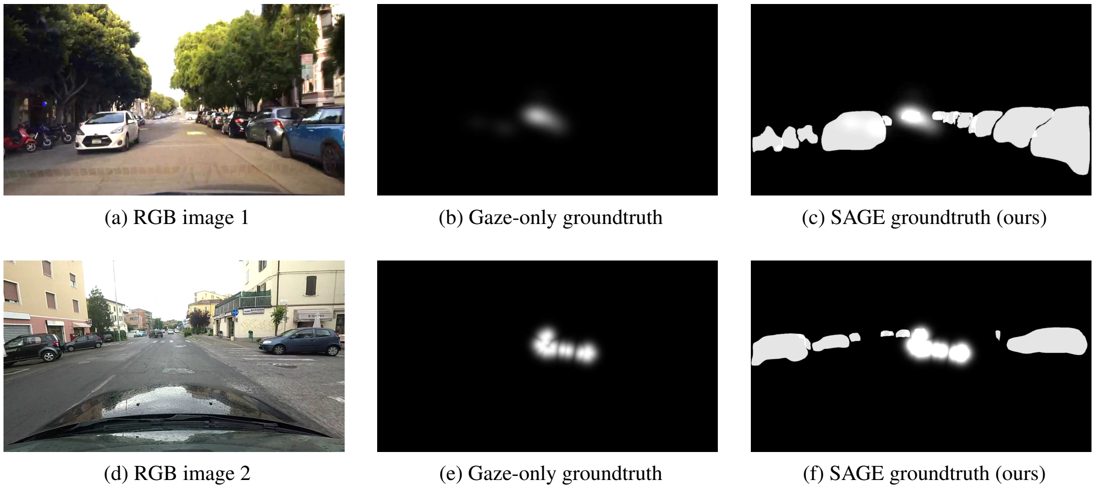
(iii) **Distracted gaze:** A human driver while driving the car might often get distracted by some road-side object - say a brightly colored building, or some attractive billboard advertisement etc. We take care of this issue by only training to detect those objects which influence the task of driving. The in-lab gaze [53] also eliminates this noise by averaging the eye movements of independent observers. However, they assume that the people annotating are positioned in the co-pilot's seat, and therefore cannot realistically emulate a driver's gaze.

(iv) **Center-bias:** For majority of a driving task, human gaze remains on the road in front of the vehicle as this is where the vehicle is headed to. When deep learning models are trained on this gaze map, they invariably recognize this pattern and learn to keep the focus there. However, this is not enough since there might be important regions away from the center of the road which demand attention - such as when cars or pedestrians approach from the sides. Thus, relying only gaze data does not help capture these important cues.

Figure 1 shows an example of an accident-prone situation, where the predicted saliency maps from an algorithm trained using different target labels are shown. Gaze-only models were able to detect the car ahead, but completely missed the pedestrian jaywalking. In contrast, our approach successfully detects both objects since it has learnt to predict semantic context in an image.

It is important to note, however, that semantics alone does not completely provide insights into the action that a driver might take at run-time. This is because a saliency map obtained only from training on semantics will give an equal-weighted attention on all the objects present. Also, when there is no object of relevance (*i.e.* an empty road near the countryside), this saliency map will not provide any attention. In reality, here the focus should be towards road boundaries, lane dividers, curbs etc. These regions can be effectively learnt through gaze information which is an indicator of a driver's intent. Thus, we design a Semantics Augmented GazE (SAGE) *ground-truth*, which successfully captures both gaze and semantic context. Figure 2 shows how our proposed ground-truth looks as compared to the existing gaze-only ground-truths.

There are three novel contributions made in this paper. Firstly, we propose SAGE - a combined attention mechanism, that can be used to train saliency models for accurately predicting an autonomous vehicle's (hereafter termed as driver) focus of attention. Secondly, we provide a thorough saliency detection framework - SAGE-Net, by including important cues in driving such as distance to objects (depth), speed of ego-vehicle and pedestrian crossing intent to further enhance the initial raw prediction obtained from SAGE. Finally, we conduct a series of experiments using



(a) RGB image 1

(b) Gaze-only groundtruth

(c) SAGE groundtruth (ours)

(d) RGB image 2

(e) Gaze-only groundtruth

(f) SAGE groundtruth (ours)

Figure 2: Comparison of SAGE with the existing gaze-only groundtruths. The top row [a-c] is for the BDD-A dataset [53] while the bottom row [d-f] is for the DR(eye)VE dataset [4]. The gaze-only maps indicate the heading of the ego-vehicle, but completely ignore the nearby and incoming cars. In contrast, SAGE captures both the driver's intent and the relevant objects.

multiple saliency algorithms on different driving datasets to evaluate the flexibility, robustness, and adaptability of SAGE - both over the entire dataset, and also specific important driving scenarios such as intersections and busy traffic regions. The rest of the paper is organized as follows. Section 2 discusses the existing state-of-the-art research in driver saliency prediction. Section 3 then provides details of the proposed framework, followed by the extensive experiments conducted in Section 4. Finally, Section 5 concludes the discussion and mentions the real-world implication of the conducted research.

2. Related Work

Advances in Salient Object Detection: Detection [51, 33] and segmentation [48, 49] of salient objects in the natural scene has been a very active area of research in the computer vision community for a long time. One of the earliest works in saliency prediction, by Itti *et al.* [22], considered general computational frameworks and psychological theories of bottom-up attention, based on center-surround mechanisms [46, 52, 24]. Subsequent behavioral [41] and computational investigations [6] used “fixations” as a means to verify the saliency hypothesis and compare models. Our approach differs from them as we incorporate both a bottom-up strategy by scanning through the entire image and detecting object features that are relevant for driving, as well as a top-down strategy by incorporating human gaze which is purely task driven. Some later studies [33, 2] defined saliency detection as a binary segmentation problem. We adopt a similar strategy, but instead of using handcrafted features that do not generalize well to real world scenes, we use deep learning techniques for robust feature extraction. Since the introduction of Convolutional Neural Networks (CNNs), a number of approaches have been developed for learning global and local features through varying receptive fields, both for 2D image datasets [51, 32, 9, 15], and video-based saliency predictions [50, 34, 14, 38]. However, these algorithms are either too heavily biased towards image datasets, or involve designs of complicated architectures which make them difficult to train. In contrast, our approach helps to improve existing architectures without any additional training parameters, thereby keeping the complexity unchanged. This is very important for an autonomous system since we want to make it as close to real-time as possible. For a detailed survey of salient object detection, we refer the reader to the work by Borji *et al.* [5].

Saliency for driving scenario: Lately, there has been some focus on driver saliency prediction due to rise of the number of driving [25, 55, 43, 42, 37] and pedestrian tracking [11, 13, 25] datasets. Most saliency prediction models are trained using human gaze information, either through in-car eye trackers [4, 39], or through in-lab simulations [53, 45]. However, as discussed above, these methods only give an estimate of the gaze, which is often prone to center bias, or distracted focus. In contrast, our approach involves combining scene semantics along with the existing gaze data. This ensures that the predicted saliency map can effectively mimic a real driver’s intent, with the added feature of also being able to successfully detect and track important objects in the vicinity of the ego-vehicle.

3. SAGE-Net: Semantic Augmented GazE detection Network

Figure 3 provides a simplified illustration of the entire SAGE-Net framework, which comprises of three components: a SAGE detection module, a distance-based attention update module, and finally a pedestrian intent-guided saliency module. We begin by firstly describing how the SAGE maps are obtained in §3.1. Next, in §3.2, we describe how relative distances of objects from ego-vehicle should impact saliency prediction. Lastly, in §3.3, we highlight the importance of pedestrian crossing intent detection and how it influences the focus of attention.

3.1. SAGE saliency map computation

We propose a new approach to predicting driving attention maps which not only uses raw human gaze information, but also learns to detect the scene semantics directly. This is done using the Mask R-CNN (M-RCNN) [20] object detection algorithm, which returns a segmented mask around an object of interest along with it’s identity and location.

We used the Matterport implementation of M-RCNN [1] which is based on Feature Pyramid Network (FPN) [28] and uses ResNet-101 [21] as backbone. The model is trained on the MS-COCO dataset [29]. However, out of the total 80 objects in [29], we select 12 categories which are most relevant to driving scenarios - `person`, `bicycle`, `car`, `motorcycle`, `bus`, `truck`, `traffic light`, `fire hydrant`, `stop sign`, `parking meter`, `bench` and `background`. For each video frame, M-RCNN provides an instance segmentation of every detected object. However, as the relative importance of different instances of the same object is not a significant cue, we stick to a binary classification approach where we segment all objects vs the background. This object-level segmented map is then superimposed on top of the existing gaze map provided by a dataset, so as to preserve the gaze information. This gives us the final saliency map as seen in Fig 2. Upon inspection, it can be clearly seen that our ground-truth has managed to capture a lot more semantic context from the scene, which gaze-only maps have missed.

3.2. Does relative distance between objects and ego-vehicle impact focus of attention?

Depth estimation through supervised [12, 30, 36] and unsupervised [16, 47] learning methods as a measure of rela-

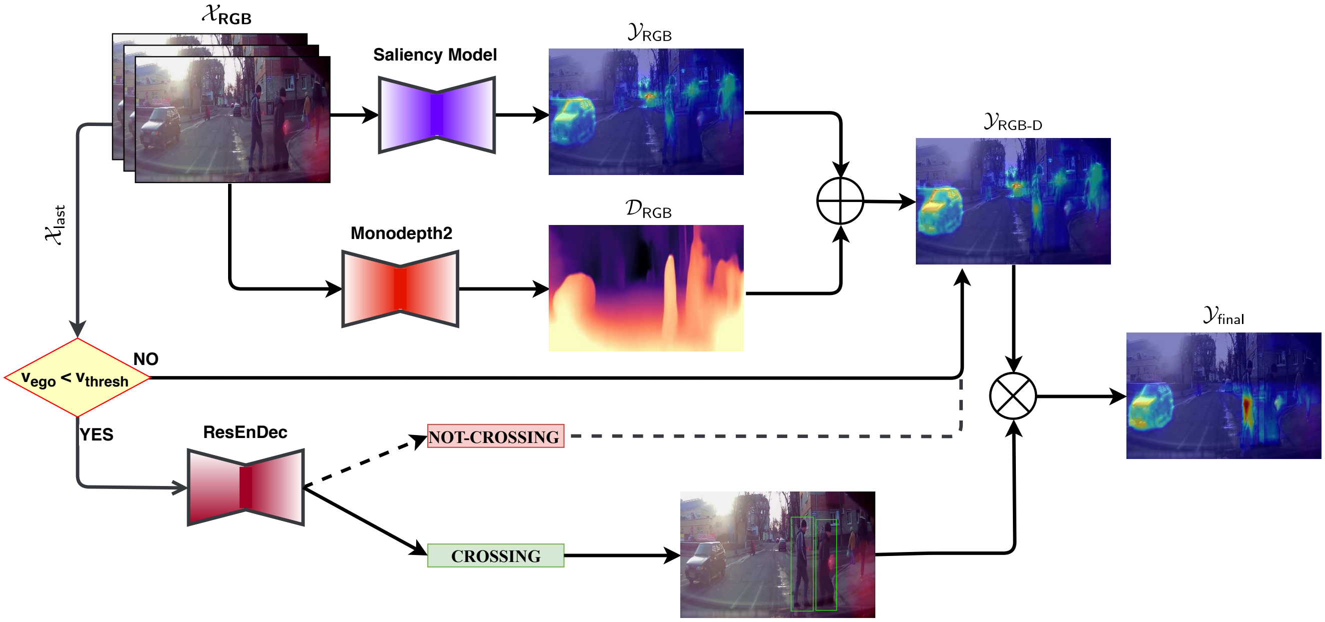

$\mathcal{X}_{\text{RGB}}$
Saliency Model
$\mathcal{Y}_{\text{RGB}}$
$\mathcal{Y}_{\text{RGB-D}}$
Monodepth2
$\mathcal{D}_{\text{RGB}}$
$\mathcal{X}_{\text{last}}$
$\mathcal{Y}_{\text{final}}$
$v_{ego} < v_{thresh}$
NO
YES
ResEnDec
NOT-CROSSING
CROSSING

Figure 3: The complete **SAGE-Net** framework (Best viewed in color), comprising of a saliency model trained on **SAGE** groundtruth, and added parallel modules for depth estimation and pedestrian intent prediction based on ego-vehicle speed (v_{ego}).

tive distance between objects and ego-vehicle has been a long studied problem in the autonomous driving community [35, 17, 18]. Human beings inherently react and give more attention to vehicles and pedestrians which are "closer" to them as opposed to those at a distance, since chances of collision are much higher for the former case. Unfortunately, this crucial information is yet to be exploited for predicting driving saliency maps to the best of our knowledge. In this paper, we consider this through the recently proposed self-supervised monocular depth estimation approach - Monodepth2 [18]. However, SAGE-Net is not restricted to just this algorithm, but can effectively inherit stereo or LiDAR-based depth estimators into its framework as well.

We considered two methods of incorporating depth maps into our framework. The first involves taking a parallel depth channel which does not undergo any training, but is simply used to amplify nearby regions of the predicted saliency map. The second method is to use it as a separate trainable input to the saliency prediction model along with the raw image, in a manner similar to how optical flow and semantic segmentation maps are trained in [39]. We decided to go with the first strategy because in addition to being much simpler and faster to implement, it also removes the issue of training a network only on depth map which has a lot less variance in data, thus leading to overfitting towards the vanishing point in the image.

Given an input clip of 16 image frames, $\mathcal{X}_{\mathbf{RGB}} \in \mathbb{R}^{16\times3\times h\times w}$, we obtain the raw prediction $\mathcal{Y}_{\text{RGB}} \in \mathbb{R}^{h\times w}$. In addition, for each frame, we also compute the depth map $\mathcal{D}_{\text{RGB}} \in \mathbb{R}^{h\times w}$. Finally, we combine the raw prediction with the depth map to obtain $\mathcal{Y}_{\text{RGB-D}}$ using the $\oplus$ operator, which is defined as

$$\mathcal{Y}_{\text{RGB}} \oplus \mathcal{D}_{\text{RGB}} = \mathcal{Y}_{\text{RGB}} * \mathcal{D}_{\text{RGB}} + \mathcal{Y}_{\text{RGB}} \tag{1}$$

3.3. Should we pay extra attention to pedestrians crossing at intersection scenarios?

Accurate pedestrian detection in crosswalks is a vital task for an autonomous vehicle. Thus, we include an additional module which focuses solely on the crossing intent of pedestrians at intersections, and correspondingly updates the saliency prediction. It should be noted that even though SAGE does capture pedestrians in its raw prediction in general driving scenarios, it does not distinguish between them and other objects in crowded traffic conditions such as intersections. This is critical since the chances of colliding with a pedestrian are much higher around intersection regions than at other roads. However, this is a slow process since it involves detecting pedestrians and predicting their pose at run-time. Fortunately, this situation only occurs when the speed of the ego-vehicle itself is less. Thus, we only include this in our framework when the speed of the ego-vehicle ($\mathtt{v_{ego}}$) is below a certain threshold velocity $\mathtt{v_{thresh}}$. It is not very difficult to obtain $\mathtt{v_{ego}}$ since most driving datasets provide this annotation [4, 53]. Also, for an autonomous vehicle, the odometry reading contains this. $\mathtt{v_{thresh}}$ is a tunable hyper-parameter which can vary as per the road and weather conditions. When $\mathtt{v_{ego}} < \mathtt{v_{thresh}}$, we look to see if there are pedestrians crossing the road. This is done using the recently proposed algorithm ResEnDec [19] which predicts the intent $\mathcal{I}$ of pedestrians as "`crossing`" or "`not crossing`" through an encoder-decoder framework using a spatio-temporal neural network and ConvLSTM. We trained this algorithm on the JAAD [25] dataset, considering 16 consecutive frames to be the temporal strip while making a prediction on the last frame $\mathcal{X}_{last}$. Our framework is designed such that if the prediction is "`crossing`", we use an object detector $\mathcal{O}$ such as YOLOv3 [44] to get the bounding box of the pedestrians from that last frame. Consequently, we amplify the predicted attention for pixels inside the bounding boxes, while leaving the rest of the image intact. This is given by the $\otimes$ operator, defined as follows

$$\mathcal{Y}_{\text{RGB-D}} \otimes \mathbf{bbox} = \begin{cases} \mathcal{Y}_{\text{RGB-D}}[x,y] * k \; \forall \; (x,y) \in \mathbf{bbox} \\ \mathcal{Y}_{\text{RGB-D}}[x,y] * 1/k, \text{ else} \end{cases} \tag{2}$$

where k is an amplification factor (> 1) by which the predicted map is strengthened. If the predicted intent is "`not crossing`", we simply stick with the original prediction $\mathcal{Y}_{\text{RGB-D}}$. The summary of the entire SAGE-Net algorithm is depicted in 1.

Algorithm 1 SAGE-Net($\mathcal{X}_{\mathbf{RGB}}, v_{thresh}$)

```
$\mathcal{Y}_{\text{RGB}} \leftarrow \text{Saliency model}(\mathcal{X}_{\mathbf{RGB}})$
$\mathcal{X}_{\text{last}} \leftarrow \mathcal{X}_{\mathbf{RGB}}[-1]$
$\mathcal{D}_{\text{RGB}} \leftarrow \text{Monodepth2}(\mathcal{X}_{\text{last}})$
$\mathcal{Y}_{\text{RGB-D}} \leftarrow \mathcal{Y}_{\text{RGB}} \oplus \mathcal{D}_{\text{RGB}}$
if $v_{ego}(\mathcal{X}_{\text{last}}) > v_{thresh}$ then return $\mathcal{Y}_{\text{RGB-D}}$
else
    $\mathcal{I}_{\mathcal{X}_{\text{last}}} \leftarrow \text{ResEnDec}(\mathcal{X}_{\mathbf{RGB}})$
    if $\mathcal{I}_{\mathcal{X}_{\text{last}}} = \texttt{not crossing}$ then return $\mathcal{Y}_{\text{RGB-D}}$
    else
        $\mathbf{bbox} \leftarrow \mathcal{O}(\mathcal{X}_{\text{last}})$
        $\mathcal{Y}_{\text{final}} \leftarrow \mathcal{Y}_{\text{RGB-D}} \otimes \mathbf{bbox}$
        return $\mathcal{Y}_{\text{final}}$
```

4. Experiments and Results

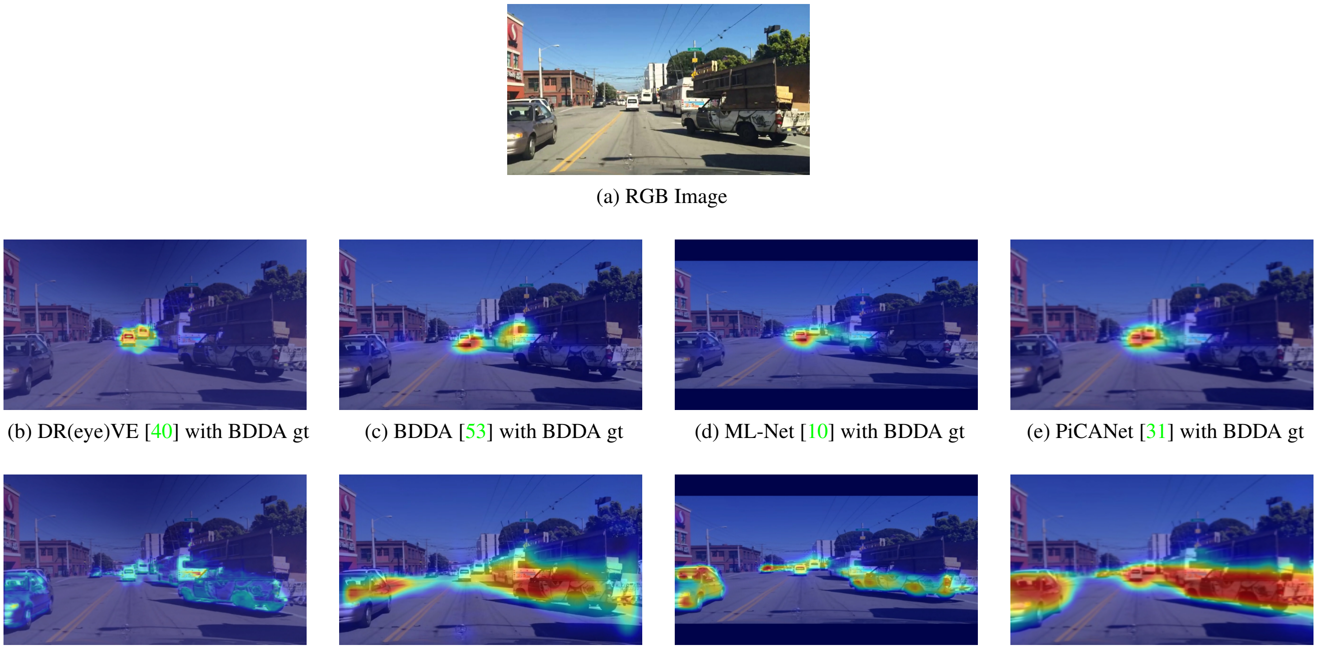
Due to the simplicity of computation of our proposed ground-truth, several experiments can be run using it. These experiments can be split into a two-stage hierarchy - (i) conducted over the entire dataset comprising of multiple combinations in driving scenarios - day vs night, city vs countryside, intersection vs highway etc. and (ii) those over specific important driving conditions such as intersection regions and crowded streets. The reason for the latter set of experiments is that averaging out the predicted results over all scenarios is not always reflective of the most important situations requiring maximum human attention [53]. For all the experiments, we describe the evaluation metrics used for comparison, and using those, compare the results of the gaze-only groundtruth and our proposed SAGE groundtruth for the different algorithms and datasets.

4.1. Some popular saliency prediction algorithms

We selected four popular saliency prediction algorithms from an exhaustive list for training with SAGE groundtruth and compared their performance against those trained with gaze-only maps. The first set of algorithms, DR(eye)VE [40] and BDD-A [53], were created exclusively for saliency prediction in the driving context. For DR(eye)VE, we only consider the image-branch for our analysis instead of the multi-branch network [39] due to two main reasons which make real-time operation possible. Firstly, it has a fraction of the number of trainable parameters and hence is faster to train and evaluate. Secondly, the latter assumes that the optical flow and semantic segmented maps are pre-computed even at test time, which is difficult to achieve online. The BDD-A algorithm is more compact and it consists of a visual feature extraction module [26], followed by a feature and temporal processing unit in the form of 2D convolutions and Convolutional LSTM (Conv2D-LSTM) [54] network respectively. However, both these algorithms combine the features extracted from the final convolution layers to make the saliency maps. This mechanism ignores low-level intermediate representations such as edges and object boundaries, which are important detections for driving scenario. Thus, we also consider ML-Net [10], which achieved best results on the largest publicly available image saliency dataset SALICON [23]. It extracts low, medium, and high-level image features and generates a fine-grained saliency map from them. Finally, PiCANet [31] extends this notion further by generating an attention map at each pixel over a context region and constructing an attended contextual feature to further enhance the feature representability of ConvNets. Figure 4 shows a comparison of the predicted saliency maps trained on gaze-only ground-truth, and those obtained from SAGE. For nearly every gaze-only model, the focus of attention is entirely towards the center of the image, thereby ignoring other cars. In contrast, SAGE-trained models have managed to successfully capture this vital information. We refer the reader to Appendix B of the supplementary material for implementation details of these four algorithms.

4.2. Evaluation metrics

We consider a set of metrics which are suitable for evaluating saliency prediction in the driving context, as opposed to general saliency prediction. More specifically, for driving purpose, we want to be more careful about identifying "False Negatives (FN)" than "False Positives (FP)", since the former error holds a much higher cost. As illustrated in Section 3, our proposed ground-truth has both a gaze component and a semantic component. Thus, we classify the set



(a) RGB Image

(b) DR(eye)VE [40] with BDDA gt

(c) BDDA [53] with BDDA gt

(d) ML-Net [10] with BDDA gt

(e) PiCANet [31] with BDDA gt

(f) DR(eye)VE [40] with SAGE gt

(g) BDDA [53] with SAGE gt

(h) ML-Net [10] with SAGE gt

(i) PiCANet [31] with SAGE gt

Figure 4: Comparison of the prediction of four popular saliency models trained on the BDD-A ground-truth (middle row) and our SAGE groundtruth (bottom row). It can be seen that for each model, SAGE trained results can capture more detailed semantic context (Best viewed in color).

of metrics broadly into two categories - (i) fixation-centric and (ii) semantic-centric.

For the first category, we choose two distribution-based metrics - Kullback-Leibler Divergence ($\mathrm{D_{KL}}$), and Pearson's Cross Correlation (CC). $\mathrm{D_{KL}}$ is an asymmetric dissimilarity metric, that penalizes FN more than FP. CC, on the other hand is a symmetric similarity metric which equally affects both FN and FP, thus giving an overall information regarding the misclassifications that occurred. Another variant of fixation metrics are the location-based metrics, such as Area Under ROC Curve (AUC), Normalized Scanpath Saliency (NSS) and Information Gain (IG), which operate on the ground-truth being represented as discrete fixation locations [7]. But for the driving task, it is crucial to identify every point on a relevant object, especially their boundaries, in order to mitigate risks. Thus, continuous distribution metrics are more appropriate here as they can better capture object boundaries.

In the second category, we again consider two metrics - namely $\mathrm{F\text{-}score}$, which measures region similarity of detection, and Mean Absolute Error (MAE), which gives pixel-wise accuracy. $\mathrm{F\text{-}score}$ is given by the formulae,

$$F_\beta = \frac{(1+\beta^2) * precision * recall}{\beta^2 * precision + recall} \tag{3}$$

where β^2 is a parameter that weighs the relative importance of precision and recall. In most literatures [50, 3, 27], β^2 is taken to be 0.3, thus giving a higher weightage to precision. However, following the earlier discussion regarding varying costs associated with FN and FP for the driving purpose, we consider β^2 to be 1, thereby assigning equal weightage to each. For a formal proof of this, we refer the reader to Appendix A of the supplementary material.

4.3. Results and Discussion

In this section, we discuss the experiments and results of algorithms trained on our proposed SAGE groundtruth, along with how they compare to the performance of the same algorithms, when trained on existing gaze-only ground-truths [4, 53]. We compare our results with that of BDD-A gaze in most of the experiments, since it is more reflective of scene semantics than the DR(eye)VE gaze. For fair comparison, we adopt different strategies for evaluating the fixation centric and semantic centric metrics. Since both the traditional gaze-only approach and SAGE contain gaze information, we use the respective ground-truths to evaluate the fixation metrics (*i.e.* gaze for the gaze-only trained model, and SAGE for our trained model). However, for the semantic metrics, we use the segmented maps generated by Mask RCNN as ground-truth to evaluate how well each of the methods can capture semantic context. The first set of comparisons, given by Table 1 and Figure 5, are calculated by taking the average over the entire test set, while the re-

	Fixation-centric metrics				Semantic-centric metrics			
	D_{KL}		CC		F_1 score		MAE	
Model	Gaze gt	**SAGE gt**	Gaze gt	**SAGE gt**	Gaze gt	**SAGE gt**	Gaze gt	**SAGE gt**
DREYEVE [39]	1.28±0.43	**0.73±0.38**	0.58±0.13	**0.75±0.13**	0.1±0.06	**0.37±0.14**	0.11±0.06	**0.08±0.05**
BDDA [53]	1.34±0.67	**1.02±0.49**	0.54±0.23	**0.6±0.18**	0.12±0.11	**0.46±0.19**	**0.12±0.09**	0.13±0.07
ML-Net [10]	**1.1±0.32**	1.35±0.51	**0.64±0.13**	0.6±0.14	0.12±0.07	**0.43±0.14**	0.12±0.06	**0.1±0.06**
PiCANet [31]	1.11±0.28	**0.83±0.31**	0.64±0.11	**0.73±0.11**	0.15±0.08	**0.64±0.15**	0.11±0.06	**0.11±0.05**

Table 1: Comparison of different saliency algorithms trained on BDD-A gaze gt and SAGE gt. All experiments are conducted on the BDD-A dataset.

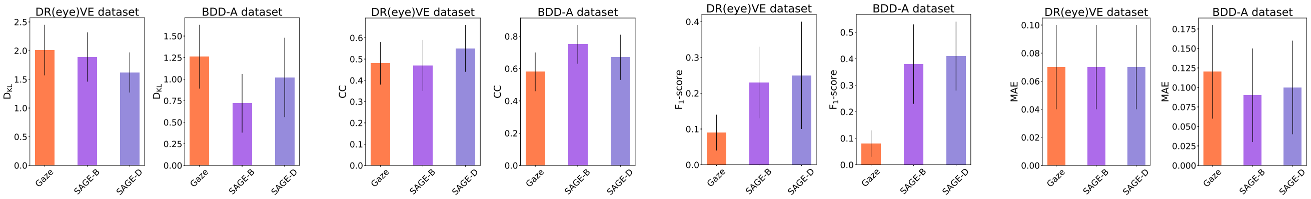
maining comparisons are for a subset of the test set involving two important driving scenarios, namely - pedestrians crossing at an intersection in Table 2, and cars approaching towards the ego-vehicle in Table 3.

Overall comparison - In Table 1, we train the four algorithms described in §4.1 on the BDD-A dataset [53]. We show the results obtained when evaluating the algorithms trained on the gaze-only data, and then on SAGE data generated by combining semantics with the gaze of [53]. As observed from the table, the D_{KL} and F_1 values obtained on SAGE are optimal for almost all the algorithms, while for CC and MAE, it either performs better or is marginally poorer in performance. Overall, this analysis shows that our proposed SAGE ground-truth performs good on a diverse set of algorithms, thus proving its flexibility and robustness.

We next consider Figure 5, where a cross-evaluation of our method with respect to different driving datasets is performed. For this set of experiments, we fix one algorithm, namely DR(eye)VE [40], while we vary the data. We evaluate two variants of SAGE - first, by combining scene semantics with the gaze of [4], and second, with the gaze of [53]. For each of these, we compare with the respective gaze-only ground-truth of the respective datasets. Like before we evaluate the performance of predicted saliency maps using the same fixation-centric and semantic-centric metrics. The results show that the proposed SAGE models are not strongly tied to a dataset and can adapt to different driving conditions. It is important to note that even though the cross evaluation (SAGE-D tested on [53], and SAGE-B tested on [4]) is slightly unfair, the results for SAGE still significantly outperforms those of the respective gaze-only models.

Comparison at important driving scenarios - In Table 2, we consider the scenarios of pedestrians crossing at intersections. For this purpose, we used a subset of the JAAD dataset [25] containing more than five pedestrians (not necessarily as a group) crossing the road. The same four algorithms described in §4.1 have been reconsidered for this case. Using M-RCNN, the segmented masks of all the crossing pedestrians were computed and the predicted saliency maps from the models were evaluated against this baseline. Upon comparison, it can be seen that models trained on SAGE surpass those trained on the gaze-only ground-truth. It is to be noted that even though none of the models were trained on the JAAD dataset [25], the results are still pretty consistent across all the algorithms. This shows that learning from SAGE indeed yields a better saliency prediction model which can detect pedestrians crossing at an intersection more reliably.

Finally, in Table 3, we took into account another important driving scenario where we consider the detection of number of cars approaching the ego-vehicle as a metric. The evaluation set was constructed by us from different snippets of the DR(eye)VE [4] and the BDD-A [53] datasets, where a single or a group of cars is/are approach-



(a) K-L Divergence (D_{KL})

(b) Cross Correlation (CC)

(c) F_1 score

(d) Mean Absolute Error (MAE)

Figure 5: Cross-evaluation of SAGE-gt by considering the gaze of two different datasets. [4] and BDD-A [53] have been used for comparison. SAGE-B/D refers to the combination of semantics with the gaze of BDD-A/DR(eye)VE dataset.

	Fixation-centric metrics				Semantic-centric metrics			
	D_{KL}		CC		F_1 score		MAE	
Model	Gaze gt	**SAGE gt**	Gaze gt	**SAGE gt**	Gaze gt	**SAGE gt**	Gaze gt	**SAGE gt**
DREYEVE [39]	3.36±0.76	**1.56±0.62**	0.19±0.09	**0.55±0.15**	0.07±0.06	**0.21±0.09**	0.08±0.04	**0.07±0.04**
BDDA [53]	2.37±0.78	**1.87±0.81**	0.28±0.16	**0.43±0.16**	0.2±0.13	**0.37±0.17**	**0.09±0.05**	0.12±0.04
ML-Net [10]	2.44±0.58	**2.27±0.67**	0.29±0.11	**0.41±0.15**	0.15±0.07	**0.31±0.13**	0.09±0.04	**0.08±0.04**
PiCANet [31]	2.97±0.68	**1.81±0.72**	0.20±0.11	**0.50±0.14**	0.13±0.07	**0.44±0.16**	**0.07±0.04**	0.11±0.03

Table 2: Comparison of SAGE with the gaze models for pedestrian crossing at intersection scenario. The clips are taken from the JAAD [25] dataset.

ing the ego-vehicle from the opposite direction in an adjacent lane. Once again, we evaluated the four algorithms on this evaluation set. Like in Table 2, here too, we analyze the detections with respect to those made by M-RCNN. The results from Table 3 show that for almost each experiment the performance of algorithms trained on SAGE is consistent in detecting the vehicles more accurately compared to the models trained by gaze-only ground-truth.

To summarize, the experiments clearly show that the proposed SAGE ground-truth can be easily trained using different saliency algorithms and the obtained results can also operate well across a wide range of driving conditions. This makes it more reliable for the driving task as compared to existing approaches which only rely on raw human gaze. Overall, the performance of our method is better than gaze-only groundtruth on **49/56 (87.5%)** cases, not only when averaged over the entire dataset, but more importantly, in specific driving situations demanding higher focus of attention.

5. Conclusion and Future Work

In this paper we introduced SAGE-Net, a novel deep learning framework for successfully predicting “where the autonomous vehicle should look” while driving, through predicted saliency maps that learn to capture semantic context in the environment, while retaining the raw gaze information. With the proposed SAGE-groundtruth, saliency models have been shown to have attention on the important driving-relevant objects while discarding irrelevant or less important cues, without having any additional computational overhead to the training process. Extensive set of experiments demonstrate that our proposed method improves the performance of existing saliency algorithms across multiple datasets and various important driving scenarios, thus establishing the flexibility, robustness and adaptability of SAGE-Net. We hope that the research conducted in this paper will motivate the autonomous driving community into looking at strategies, that are simple but effective, for enhancing the performance of currently existing algorithms.

Our future work will involve incorporating depth in the SAGE-groundtruth and then having the entire framework to be trained end-to-end. Currently this could not be achieved due to low variance in the depth data, leading to overfitting. Another possible direction that is being considered is to explicitly add motion dynamics of segmented semantic objects in the surroundings in the form of SegFlow [8]. Work in this area is under progress as we are building a campus-wide dataset with these kind of annotations through visual sensors and camera-LiDAR fusion techniques.

Acknowledgements

The authors would like to thank Army Research Laboratory (ARL) W911NF-10-2-0016 Distributed and Collaborative Intelligent Systems and Technology (DCIST) Collaborative Technology Alliance for supporting this research.

	Fixation-centric metrics				Semantic-centric metrics			
	D_{KL}		CC		F_1 score		MAE	
Model	Gaze gt	**SAGE gt**	Gaze gt	**SAGE gt**	Gaze gt	**SAGE gt**	Gaze gt	**SAGE gt**
DREYEVE [39]	3.87±0.79	**1.28±0.71**	0.18±0.11	**0.62±0.19**	0.08±0.08	**0.33±0.16**	0.08±0.05	**0.07±0.05**
BDDA [53]	2.95±0.96	**1.92±1.01**	0.19±0.16	**0.42±0.18**	0.14±0.13	**0.34±0.19**	**0.09±0.09**	0.12±0.07
ML-Net [10]	2.72±0.6	**1.94±0.9**	0.21±0.1	**0.5±0.18**	0.12±0.07	**0.37±0.14**	0.09±0.05	**0.08±0.05**
PiCANet [31]	3.17±0.6	**1.69±0.88**	0.18±0.1	**0.55±0.17**	0.12±0.07	**0.49±0.2**	**0.08±0.05**	0.1±0.04

Table 3: Comparison of SAGE with the gaze models for detecting multiple cars approaching the ego-vehicle from the opposite direction. The clips are taken from the DR(eye)VE [4] and BDD-A [53] datasets.

References

[1] Waleed Abdulla. Mask R-CNN for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

[2] Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Süsstrunk. Salient region detection and segmentation. In *International conference on computer vision systems*, pages 66–75. Springer, 2008.

[3] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1597–1604, 2009.

[4] Stefano Alletto, Andrea Palazzi, Francesco Solera, Simone Calderara, and Rita Cucchiara. DR(eye)VE: A dataset for attention-based tasks with applications to autonomous and assisted driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.

[5] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media*, pages 1–34, 2014.

[6] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2006.

[7] 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.

[8] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, pages 686–695, 2017.

[9] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019.

[10] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A Deep Multi-Level Network for Saliency Prediction. In *International Conference on Pattern Recognition (ICPR)*, 2016.

[11] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34, 2012.

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[13] Andreas Ess, Bastian Leibe, and Luc Van Gool. Depth and appearance for mobile scene analysis. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[14] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8554–8564, 2019.

[15] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[16] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.

[17] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[18] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *arXiv preprint arXiv:1806.01260*, 2018.

[19] Pratik Gujjar and Richard Vaughan. Classifying pedestrian actions in advance using predicted video of urban driving scenes. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'19)*, Montreal, Canada, May 2019.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[21] 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.

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

[23] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[24] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[25] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Joint attention in autonomous driving (JAAD). *arXiv preprint arXiv:1609.04741*, 2016.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[27] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014.

[28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[30] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2015.

[31] Nian Liu, Junwei Han, and Ming-Hsuan Yang. PiCANet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018.

[32] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019.

[33] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2010.

[34] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3623–3632, 2019.

[35] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, 2018.

[36] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

[37] Athma Narayanan, Isht Dwivedi, and Behzad Dariush. Dynamic traffic scene classification with space-time coherence. *arXiv preprint arXiv:1905.12708*, 2019.

[38] A. Pal, C. Nieto-Granda, and H. I. Christensen. DEDUCE: Diverse scene detection methods in unseen challenging environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4198–4204, Nov 2019.

[39] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver's focus of attention: the DR(eye)VE project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018.

[40] Andrea Palazzi, Francesco Solera, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Learning where to attend like a human driver. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 920–925. IEEE, 2017.

[41] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123, 2002.

[42] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The H3D dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *International Conference on Robotics and Automation*, 2019.

[43] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[44] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[45] Ashish Tawari, Praneeta Mallela, and Sujitha Martin. Learning to attend to salient targets in driving videos using fully convolutional RNN. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3225–3232. IEEE, 2018.

[46] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[47] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017.

[48] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3395–3402, 2015.

[49] Wenguan Wang, Jianbing Shen, Hanqiu Sun, and Ling Shao. Video co-saliency guided co-segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(8):1727–1736, 2017.

[50] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3064–3074, 2019.

[51] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1448–1457, 2019.

[52] Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, 15(3):419, 1989.

[53] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *Asian Conference on Computer Vision*, pages 658–674. Springer, 2018.

[54] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[55] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018.