Unsupervised Self-Driving Attention Prediction via Uncertainty Mining and Knowledge Embedding

Pengfei Zhu[1], Mengshi Qi *[1,2], Xia Li[1], Weijian Li[3], and Huadong Ma[1]

[1]Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, 100876
[2]Key Laboratory of Artificial Intelligence, Ministry of Education, Shanghai, 200240
[3]Department of Computer Science, University of Rochester
{zhupengfei2000,qms,mhd}@bupt.edu.cn; lixia@bupt.cn; wli69@cs.rochester.edu

Abstract

Predicting attention regions of interest is an important yet challenging task for self-driving systems. Existing methodologies rely on large-scale labeled traffic datasets that are labor-intensive to obtain. Besides, the huge domain gap between natural scenes and traffic scenes in current datasets also limits the potential for model training. To address these challenges, we are the first to introduce an unsupervised way to predict self-driving attention by uncertainty modeling and driving knowledge integration. Our approach's Uncertainty Mining Branch (UMB) discovers commonalities and differences from multiple generated pseudo-labels achieved from models pre-trained on natural scenes by actively measuring the uncertainty. Meanwhile, our Knowledge Embedding Block (KEB) bridges the domain gap by incorporating driving knowledge to adaptively refine the generated pseudo-labels. Quantitative and qualitative results with equivalent or even more impressive performance compared to fully-supervised state-of-the-art approaches across all three public datasets demonstrate the effectiveness of the proposed method and the potential of this direction. The code is available at https://github.com/zaplm/DriverAttention.

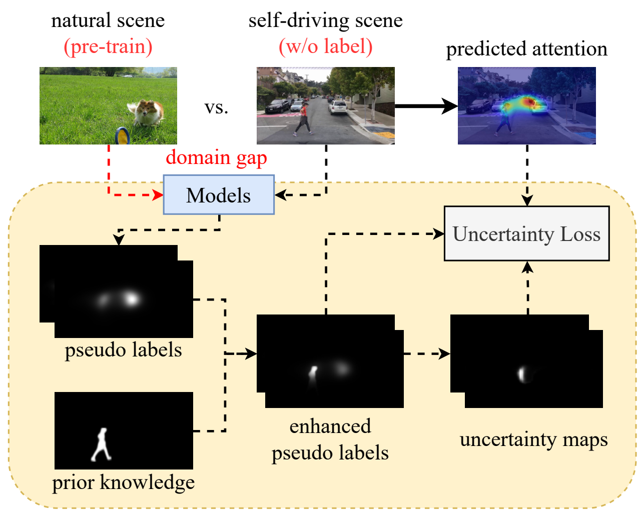

Figure 1. Illustration of the proposed unsupervised self-driving attention prediction model. Instead of relying on the ground truth labels provided by traffic datasets, our method only uses pseudo-labels generated from models pre-trained on natural scenes, and then refined the results by uncertainty mining and knowledge embedding. The red dashed line corresponds to the pre-training stage, the black dashed line refers to the training process, and the black solid line means the testing process.

1. Introduction

With the huge development of autonomous driving, predicting attention regions for self-driving systems [2; 41] has drawn rapid interest in the community. The predicted attention region provides rich contextual information to assist autonomous driving systems by locating salient areas in the traffic scene [49; 50; 60]. Most importantly, these salient areas are always the riskiest areas, where small perception errors can cause great harm to driver's safety [25]. Therefore, with a successful attention area prediction, computation resources can be reallocated to enhance the perception accuracy in these fatal areas to reduce driving risks, as well as increase the explainability and improve the reliability of autonomous driving systems [28].

Numerous datasets [1; 13; 59] and methods [2; 27; 30;

*Corresponding author.

38; 42; 59] have been proposed to address self-driving attention prediction task. Though achieving encouraging performance, these methods are trained in fully-supervised ways on large-scale labeled datasets which are hard to build and unreliable. For example, one of the widely-used datasets in self-driving named DR(eye)VE [1] was collected in two months, by recording eight drivers taking turns driving on the same route to obtain fixation data. However, simply averaging the attention of eight drivers into one driving video will lead to the wrong attention target. Another common difficulty is the huge mismatch between the collected data and real-world environments. Another self-driving dataset BDD-A [59] was constructed by asking 45 participants to watch the same recorded video and imagine themselves as the drivers. But, these simulated virtual environments inevitably brought inconsistencies to real-world conditions for human labeling. Therefore, current fully-supervised methods suffer from potential biases in public datasets and then are too hard to extend to new environments. Furthermore, large-scale pre-trained models [4] have already demonstrated strong capability in representation learning, which can be beneficial to lots of downstream tasks. But how to bridge the domain gap between the specific situation (*e.g*. self-driving scenes) and the common data pre-trained model used (*e.g*. natural scenes) is still a challenge.

To address the above-mentioned issues, we propose a novel unsupervised framework to self-driving attention prediction, which means **1)** we do not use any ground-truth labels given by self-driving datasets, **2)** we only use pseudo-labels generated from models pre-trained on natural scene datasets, **3)** we train a model on the source domain and to adapt it to the samples in the target domain (in our case is from natural to traffic scenes) following unsupervised domain adaptation [16; 22; 53]. Specifically, our proposed model is achieved with two newly-designed parts: an uncertainty mining branch is proposed to exploit pseudo-labels' uncertainties by aligning the various distributions and thus make the result reliable; another is a knowledge embedding block which is introduced to transfer the traffic knowledge into the natural domain by segmenting the focal traffic objects with Mask-RCNN [17] pre-trained on MS-COCO [31] and then enhance each pseudo-label's attention region.

In summary, our contributions can be listed as follows:

(1) We propose a novel unsupervised framework to predict self-driving attention regions, which is not relying on any labels on traffic datasets. To the best of our knowledge, this is the first work to introduce such an unsupervised method to this specific task.

(2) We introduce an uncertainty mining branch to produce highly plausible attention maps by estimating the commonality and distinction between multiple easily obtained pseudo-labels from models pre-trained on natural scenes.

(3) We design a knowledge embedding block by incorporating rich driving knowledge to refine the produced pseudo-labels, which bridges the domain gap between autonomous driving and common domains (*e.g*. natural scene, daily life, and sports scene).

(4) Extensive experiments on three public benchmarks with comparable or even better results compared with fully-supervised state-of-the-art approaches demonstrate the effectiveness and superiority of the proposed method.

2. Related Work

Self-Driving Attention Prediction. With the rise of deep learning, several attempts [13; 38; 42; 59] have been made to introduce various deep learning methods into the field of self-driving attention prediction. Palazzi *et al*. [42] employed a multi-branch video understanding method to predict the driver's attention in a hierarchical manner from coarse to fine. Xia *et al*. [59] addressed the center bias problem in attention prediction by assigning varying weights to each training sample based on the KL divergence between the attention map and the average attention map. Meanwhile, Baee *et al*. [2] leveraged an inverse reinforcement learning (IRL) approach to improve the accuracy of attention prediction by incorporating task-specific information. All previous studies relied on large-scale in-lab or in-car annotated datasets [1; 13; 59]. DR(eye)VE [1] presented an in-car dataset that includes dozens of segments, which record driver's attention changes during prolonged driving in the car. BDD-A [59] and DADA-2000 [13] are presented as in-lab datasets that synthesize attention changes of several volunteers, providing more than 1000 clips, containing both normal and multiple emergent driving situations. To overcome the unreliable dependency of self-driving datasets, our model is the first to address self-driving attention prediction in an unsupervised manner by leveraging pseudo-labels generated by models pre-trained on natural scenes.

Saliency Detection. Predicted saliency regions in images or videos [12; 20; 40] can approximate human's visual attention. It has been used to evaluate the explainability of deep models [28; 60] and to assist other tasks, i.e., photo cropping [55], scene understanding [45; 46; 47; 48], vehicle re-identification [33; 34; 35] and object segmentation [60]. However, most existing datasets [5; 21; 52; 56] and methods [9; 10; 12; 20; 40; 54; 54; 61] are mainly focusing on natural scenes or common objects, not specially tailored into self-driving scenarios. In this work, we propose an uncertainty mining branch and a knowledge embedding strategy to bridge the domain gap between natural scenes and self-driving situations.

Uncertainty Estimation. Early uncertainty estimation works in deep learning mainly focus on model uncertainty, which is crucial for evaluating the accuracy and robustness of the model. A pioneer work is that Gal and Ghahramani [15; 24] use dropout to represent model uncertainty.

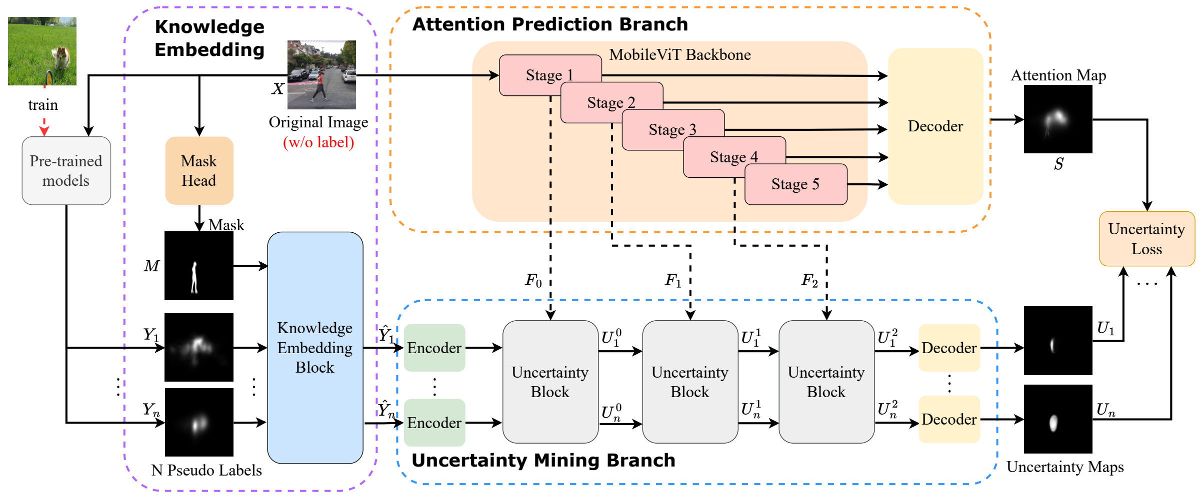

Figure 2. An overview of our proposed unsupervised self-driving attention prediction model. Our approach leverages pseudo-labels generated from models pre-trained on natural scene datasets for unsupervised training. To introduce additional semantic information for the self-driving scenario, we propose a Knowledge Embedding Block (KEB). Meanwhile, the Attention Prediction Block (APB) which uses Mobile-ViT [39] backbone comprises five stages for image feature extraction, with each stage producing features subsequently fed to the decoder. Note that features extracted in stages 1, 2, and 4 are sent to three Uncertainty blocks for multi-scale feature fusion. Our Uncertainty Mining Block (UMB) employs multiple pseudo-labels with multi-scale features for fusion and mining to generate an uncertainty map for each pseudo-label. Finally, we optimize the network structure using uncertainty loss.

Lately, Kendall *et al.* [25] constructs a new loss that combines data uncertainty and model uncertainty for multi-task learning [26]. Nowadays, uncertainty methods have been widely used in various autonomous driving tasks such as target detection [8; 36], motion prediction [11; 14], semantic segmentation [3; 58], and etc. In the field of self-driving attention prediction, there has been no prior work that incorporates uncertainty estimation. We are the first to introduce an uncertainty mining branch to estimate the commonality and distinction between multiple pseudo-labels, and then produce plausible attention maps.

3. Method

3.1. Overview

Figure 2 shows an overview of our proposed unsupervised driving attention prediction network. Our network consists of an Attention Prediction Branch (APB), a Knowledge Embedding Block (KEB) as well as an Uncertainty Mining Branch (UMB).

Our method learns to predict self-driving attention in an unsupervised way. To achieve unsupervised learning, a naive way is to train the model with the generated pseudo-labels from a single source model pre-trained on natural scenes. However, the large domain gap between natural environments and self-driving scenes brings strong uncertainty. Meanwhile, each single source label from a specific domain shall correspond to a different distribution, in which some particular areas may lead to strong uncertainty. Encouraged by the recent development of uncertainty estimation [25; 26], we propose to improve the accuracy and robustness of our prediction by modeling uncertainty from multi-source pseudo-labels. Through the evaluation of uncertainties across various distributions, we can effectively alleviate potential discrepancies and inconsistencies. Moreover, since the generated pseudo-labels we used are directly transferred from the natural domain, they lack relevant knowledge of autonomous driving scenarios. Thus, we perform a knowledge enhancement pre-processing operation in KEB on each input pseudo-label to improve prediction results. Meanwhile, we designed a novel uncertainty mining branch (UMB) to densely acquire soft uncertainty from multi-source pseudo-labels. The UMB consists of multiple Uncertainty Blocks (UB) and recursively analyzes the commonalities and differences among multiple noisy labels to infer the pixel-level uncertainty map for each label.

Problem Formulation. Given an RGB input frame $X \in \mathbb{R}^{H\times W\times 3}$, following PSPNet [62] and DeepLabv3+ [7], APB extracts pyramid features in five levels and passes the features F from the 1st, 2nd, and 4th stages as $\{F^0, F^1, F^2\}$ to explore pseudo-labels' uncertainty in UMB. APB follows the structure of U-Net [51], feeds the extracted features from the last layer into the decoder and concatenates them with the features at corresponding granularity, and outputs the final attention prediction result as $S \in \mathbb{R}^{H\times W\times 1}$ through a Readout module. In addition, before feeding pseudo-labels into UMB, we perform a knowledge enhancement process to get pseudo-labels adapted to autonomous driving scenarios with an off-the-shelf Mask Head. Then, UMB takes N knowledge-embedded pseudo-

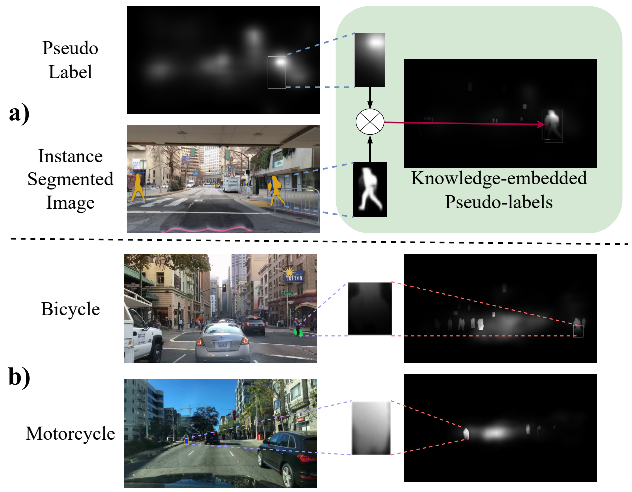

Figure 3. Illustration of the knowledge embedding strategy: a) the process of knowledge embedding for a single pseudo-label, where the salient region can be enhanced by adding the self-driving-related instance (*e.g.* pedestrian) where the operator $\otimes$ means the operation in Eq. 1; b) two other examples of knowledge embedding for bicycles and motorcycles.

labels $\hat{Y} = \left\{\hat{Y}_1, \cdots, \hat{Y}_n\right\}$ as input and estimates the uncertainty maps correspondingly, which have the same size with the final output attention map S. These pseudo-labels are fused with three different levels of features from APB to output the uncertainty maps $U = \{U_1, \cdots, U_N\}$. Finally, the model is trained by optimizing the uncertainty loss between the attention map and the uncertainty map.

3.2. Knowledge Embedding Block (KEB)

With prior knowledge, humans are able to disambiguate and discover relevant objects centered at the visual complex scenes [23]. Inspired by these findings, we design KEB to enhance prior driving knowledge and bridge the domain gap between natural scenes and self-driving environments. To be specific, we use the off-the-shelf Mask R-CNN pre-trained on the MS-COCO dataset [31] to segment the most representative traffic objects as prior knowledge, *i.e.,* pedestrians, signals, bicycles, motorcycles, and traffic signs (*e.g.*, stop signs, road signs, etc.). During the knowledge embedding, we freeze the parameters of Mask R-CNN with the open-source checkpoints to make the knowledge embedding process practically unsupervised. Through the segmenting of the input frame with Mask-RCNN, we merge the obtained masks of different categories into a single binary mask map. Note that we explore two strategies to embed prior knowledge into different pseudo-labels: 1) concatenating them at the channel dimension and 2) fusing them to a one-channel segmentation map. For the first strategy, each pseudo-label is concatenated with the binary mask and then fed into UMB, allowing the model to learn the relationship adaptively. For the second strategy, we compose each pseudo-label with the binary mask using the following formulation:

$$\hat{Y}_n = Y_n \cdot (M + \alpha)\,, \tag{1}$$

where α is a hyper-parameter that is empirically set to 0.3, Y_n denotes the n-th pseudo-label, and M denotes the segmented map of the corresponding image. We adopt the second strategy in our approach for better performance (for more experimental results please refer to Sec 4.4). However, after knowledge embedding, those pseudo-labels can be enhanced at the pixel level to have a more robust ability to identify significant traffic objects.

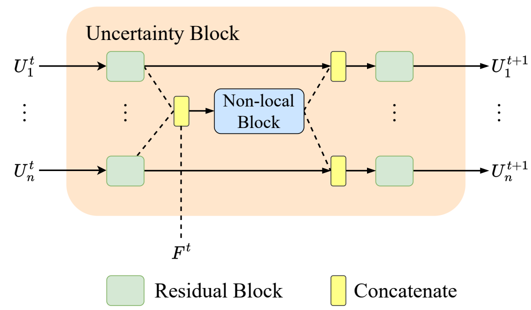

Figure 4. Illustration of the proposed Uncertainty Block. In each stage, the input uncertainty maps Us from the previous stage pass through a residual block, then are concatenated with another uncertainty map and are fed into the Non-local Block. The results are concatenated with the original uncertainty map and are passed through a residual block as the input of the next stage.

3.3. Uncertainty Mining Branch (UMB)

In our work, UMB is introduced to mine the uncertainty from multi-source pseudo-labels that are generated from multiple pre-trained models. Notice that these models are pre-trained on natural scenes, not self-driving, i.e. ML-Net [9], SAM [10], and UNISAL [12] are pre-trained on SALICON [21], while TASED-Net [40] is pre-trained on DHF-1K [56]. As is shown in Figure 4, the Uncertainty Block is proposed to exchange information between pseudo-labels and multi-scale features extracted by APB, which consists of the non-local self-attention operations and merge/split mechanism [57; 58]. In our UMB, we adopt three such blocks to gather information from both pseudo-labels and multi-scale image features and enable long-range interactions among pixels.

Specifically, in the uncertainty block, for the n-th knowledge-embedded pseudo-label $\hat{Y}_n \in \mathbb{R}^{H\times W\times 1}$, we first pass it through a convolutional layer and a downsampling layer, resulting in $\frac{1}{4}$ of the original size. Then we feed it into a residual block [18] to exchange information with pseudo-labels and features maps from other sources at the same stage. The obtained results are concatenated with the input multi-source pseudo-labels and then are passed through the non-local self-attention to obtain a coarse uncertainty map U_n corresponding to the n-th pseudo label, formulated as:

$$U_n^0 = f_{attn}^0 \left(\text{Concat} \left(\hat{Y}_1, \cdots, \hat{Y}_n, F^0 \right) \right) + \hat{Y}_n, \quad (2)$$

where the superscripts denote the stage index, and $f_{attn}^t(\cdot)$ refers to non-local self-attention. Then we gradually refine U_n^0 to U_n^{t+1} as follows:

$$U_n^{t+1} = f_{attn}^t \left(\text{Concat} \left(U_1^t, \cdots, U_N^t, F^t \right) \right) + U_n^t. \quad (3)$$

Finally, through three uncertainty blocks, the fine-grained uncertainty map $U_n^2 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 1}$ can be obtained and then be upsampled to $U_n \in \mathbb{R}^{H \times W \times 1}$ in the decoder as the same size as the original input.

3.4. Loss Function

We treat the predicted attention map S as a distribution over the spatial dimension and we need to normalize the generated pseudo-labels accordingly. To satisfy this requirement, we apply a spatial softmax layer after APB. Inspired by the uncertainty loss in [25], we assume a Boltzmann distribution under the Bayesian theory for each pseudo-label map $\hat{Y}_n \in \mathbb{R}^{H \times W \times 1}$. Therefore, the probability of the final prediction S with respect to the label $\hat{Y}_n$ can be calculated as follows:

$$p(\hat{Y}_n | S, u_n) = \prod_i \text{Softmax}(\frac{S_i}{u_n^2}), \quad (4)$$

where $u_n = 1/(H \times W) \sum_i^{H \times W} U_n^i$ is the final uncertainty estimation for the n-th pseudo-label, i denotes the pixel index of S. Also, u_n can be regarded as the temperature parameter whose magnitude determines how 'uniform' (flat) the distribution is. The negative log-likelihood of the whole pseudo-label map is calculated as:

$$\begin{aligned} & - \log p(\hat{Y}_n | S, u_n) \\ = & - \sum_i \frac{S_i}{u_n^2} + \log \sum_i \exp(\frac{S_i}{u_n^2}) \\ \approx & \frac{L_{\text{CE}}(S, \hat{Y}_n)}{u_n^2} + \log(u_n), \end{aligned} \quad (5)$$

where $L_{\text{CE}}(S, \hat{Y}_n)$ denotes the spatial cross entropy loss. In practice, we can instead predict the log variance $e_n = \log(u_n)^2$ to increase the numerical stability [26] during the training process. Now, the loss can be re-formulated as follows:

$$L(S, u_n, \hat{Y}_n) = L_{\text{CE}}(S, \hat{Y}_n) \cdot \exp(-e_n) + \frac{1}{2} e_n. \quad (6)$$

Besides, we can reformulate the cross-entropy loss $L_{\text{CE}}(S, Y_n)$ as follows:

$$\begin{aligned} L_{\text{CE}}(S, \hat{Y}_n) &= - \sum_i \hat{Y}_{n,i} \log(S_i) \\ &= - \sum_i \hat{Y}_{n,i} \log(S_i) + H(\hat{Y}_n) - H(\hat{Y}_n) \\ &= \sum_i \hat{Y}_{n,i} (\log(\hat{Y}_{n,i}) - \log(S_i)) - H(\hat{Y}_n) \\ &= L_{\text{KLD}}(\hat{Y}_n, S) - H(\hat{Y}_n), \end{aligned} \quad (7)$$

where $L_{\text{KLD}}(\hat{Y}_n, S) = \sum_i \hat{Y}_{n,i} (\log(\hat{Y}_{n,i}) - \log(S_i))$ is the KL-divergence between the pseudo-label distribution and the predicted attention map distribution. $H(\hat{Y}_n)$ is the information entropy of the distribution $\hat{Y}_n$, which is non-related to the optimization and thus can be regarded as a constant. Therefore, according to Eq. 6 and Eq. 7, and extending the calculation to all N-source pseudo-labels, we obtain the final loss as:

$$L = \sum_{n=1}^N \{ L_{\text{KLD}}(\hat{Y}_n, S) \cdot \exp(-e_n) + \frac{1}{2} e_n \}. \quad (8)$$

Notice that our KLD uncertainty loss differs from formulas of [29] that we assume a spatial distribution instead of a single per-channel counterpart. This assumption is crucial for derivation of Eq. 7.

4. Experimental Results

In the experiments, we first compare our proposed unsupervised method with other full-supervised networks on several widely-adopted datasets, *i.e.,* BDD-A, DR(eye)VE, DADA-2000. Subsequently, extensive ablation studies are conducted to verify the effectiveness of each proposed component in our proposed network.

4.1. Experimental Settings

Datasets. We evaluate the performance of our proposed model on three self-driving benchmarks: BDD-A, DR(eye)VE, and DADA-2000. **BDD-A** [59] is an in-lab driving attention dataset consisting of 1,232 short time slices (each within 10 seconds). It contains a large amount of data from driving on various urban and rural roads. We follow its split and obtain 28k frames for training, 6k frames for validating, and 9k frames for testing. **DR(eye)VE** [1] is an in-car dataset that tries to maintain consistent driving conditions under controlling variables, and it contains 74 long videos in total (each is up to 5 minutes long). We follow [1] and choose the last 37 videos as the test set. **DADA-2000** [13] is another in-lab dataset and the only one including vehicle crash cases, which offers us the possibility to predict driving attention under extreme critical scenarios.

Methods	BDD-A [59]		DR(eye)VE [1]		DADA-2000 [13]	
	KLD↓	CC↑	KLD↓	CC↑	KLD↓	CC↑
Multi-Branch [42]	1.28	0.58	**1.40**	**0.56**	2.27	0.45
HWS [59]	1.34	0.54	2.12	0.51	2.50	0.40
SAM [10]	2.46	0.25	2.56	0.38	2.85	0.27
Tased-Net [40]	1.79	0.52	<u>1.88</u>	0.47	<u>1.88</u>	<u>0.53</u>
MEDIRL [2]	2.51	**0.74**	-	-	2.93	**0.63**
ML-Net [9]	1.20	0.64	2.00	0.44	-	-
UNISAL [12]	1.49	0.58	-	-	-	-
PiCANet [32]	<u>1.11</u>	0.64	-	-	-	-
DADA [13]	-	-	-	-	2.19	0.50
Ours (unsupervised)	**1.099±0.016**	<u>0.640±0.007</u>	1.901±0.004	<u>0.510±0.005</u>	**1.677±0.007**	0.488±0.002

Table 1. Performance comparison between our proposed unsupervised method and state-of-the-art fully-supervised methods. It is worth noting that our unsupervised method achieves comparable or even better performance compared with the fully-supervised methods. The numbers in bold denote the best results, and those marked with underlines denote the second best.

Test Dataset / pseudo-labels	BDD-A [59]		DR(eye)VE [1]		DADA-2000 [13]	
	KLD↓	CC↑	KLD↓	CC↑	KLD↓	CC↑
BDD-A	**1.099±0.016**	**0.635±0.007**	1.924±0.004	0.508±0.003	**1.677±0.007**	**0.488±0.002**
DR(eye)VE	1.188±0.011	0.608±0.002	1.908±0.008	**0.517±0.005**	1.801±0.017	0.458±0.004
DADA-2000	1.242±0.021	0.578±0.009	**1.889±0.012**	0.513±0.010	1.711±0.015	0.483±0.007

Table 2. Performance comparison of our proposed unsupervised network trained with pseudo-labels generated from various self-driving datasets (BDD-A, DR(eye)VE, DADA-2000) and then test on each benchmark. The best result is highlighted in bold.

This dataset contains 2000 video clips and has over $658,746$ frames. We follow [13] to split all videos at the ratio of 3:1:1 for training, validating, and testing.

Metrics. To comprehensively evaluate our model, we utilize two common metrics, *i.e.,* Kullback-Leibler divergence (KLD) [29] as well as Pearson Correlation Coefficient (CC) [44]. KLD evaluates the similarity between the predicted driving attention map and the real distribution, and it is an asymmetric dissimilarity measure that penalizes false negative (FN) values more than false positive (FP) values. While CC evaluates how much the predicted driving attention map is linearly correlated with the real distribution, it is a symmetric similarity measure that penalizes equally for both FN and FP. Notice that we do not adopt the discrete metrics, such as Area Under ROC Curve (*AUC*) and its variants (*AUC-J*, *AUC-S*), Normalized Scanpath Saliency (*NSS*), and Information Gain (*IG*) [6], because the continuous distribution metrics are observed to be more appropriate to predict risky pixels and areas in driving scenarios [41].

Compared Methods. We compare our proposed unsupervised approach with recent fully-supervised state-of-the-art methods, including Multi-Branch [42], HWS [59], SAM [10], TASED-Net [40], MEDIRL [2], ML-Net [9], UNISAL [12], PiCANet [32] and DADA [13].

4.2. Implement Details

Our proposed network is implemented using PyTorch [43]. For each dataset, we first sample both the original video frame and the gaze annotated maps to $3Hz$, making them aligned with each other. During training, the generated pseudo-labels and the original images are resized to 224×224, and the values are normalized in the spatial dimension. Regarding the knowledge embedding strategy, we use Mask R-CNN pre-trained on the MS-COCO [31] to segment important instances and fuse them with pseudo-labels. Furthermore, we set the initial learning rate of our proposed network to 0.001, using a learning scheduler that first warm-up and then descends in a cosine fashion. Additionally, we use the Adam optimizer [37] ($\beta_1 = 0.9, \beta_2 = 0.999$) with the weight decay 0.0001. Overall, we run 10 epochs with a batch size of 32 for training, and the training time of our proposed network is approximately 50 minutes on a single RTX 3090 GPU. While it takes about 12 ms to infer attention regions per frame.

4.3. Quantitative Comparisons

The quantitative performance of our proposed unsupervised network compared with other fully-supervised state-of-the-art models can be found in Table 1. Note that in our experiments, our unsupervised model does not utilize any ground-truth labels from self-driving datasets, but is only trained with the generated pseudo-labels with the input BDD-A training set, and then tested on each benchmark's test set. From Table 1, we can clearly observe that the proposed uncertainty network achieves competitive results compared to all fully-supervised methods and even outper-

Ablated Variants	BDD-A [59]		DR(eye)VE [1]		DADA-2000 [13]	
	KLD↓	CC↑	KLD↓	CC↑	KLD↓	CC↑
APB (unsupervised)	1.233	0.608	2.013	0.501	1.805	0.460
APB+UMB	1.141	0.622	1.941	0.510	1.702	0.480
APB+UMB+non-local block	1.134	0.626	1.917	0.514	1.695	0.485
Ours: APB+UMB+non-local block+KEB	**1.099**	**0.635**	**1.901**	**0.518**	**1.677**	**0.488**

Table 3. Comparison between our proposed unsupervised model and its ablated variants. All models are trained with pseudo-labels generated from BDD-A and tested on other self-driving attention datasets (BDD-A, DR(eye)VE, DADA-2000). We ablate parts of the proposed model in each iteration until the basic APB is left alone. The basic APB is trained with unsupervised learning using pseudo-labels generated from the BDD-A training set by ML-Net. The best result is highlighted in bold.

Pseudo-labels	KLD↓	CC↑
M	1.233	0.608
U	1.246	0.597
M+U	**1.099**	**0.635**
M+U+T	1.189	0.619
M+U+S	1.162	0.621
M+U+T+S	1.167	0.620

Table 4. Comparison of different sources of pseudo-labels in the UMB on the model performance. In this table, we use the following abbreviations: M for ML-Net [9], U for UNISAL [12], T for TASED-Net [40], and S for SAM [10].

Input	KLD↓	CC↑
concat (obj. & text)	1.126	0.626
concat (obj.)	1.123	0.628
single (obj. & text, $\alpha = 0.3$)	1.123	0.631
single (obj., $\alpha = 0.3$)	**1.099**	**0.635**

Table 5. Comparison of different strategies and types of knowledge embedding, where "obj." refers to the masks of segmented critical objects with Mask-RCNN, "text" refers to the masks of detected text (*e.g*. road signs, stop signs, etc.) with EAST in the traffic scene, and α means the hyper-parameter in Eq. 1.

forms previous fully-supervised methods in terms of the KLD metric on BDD-A and DADA-2000, and achieves the second-best w.r.t CC on BDD-A and DR(eye)VE, demonstrating the effectiveness and potential of our proposed unsupervised method.

In order to examine the transferability of these three self-driving benchmarks (*i.e.,* BDD-A, DR(eye)VE, DADA-2000), we report the results of our method trained with pseudo-labels generated in each dataset and tested on another dataset in Table 2. We can find that the model trained with pseudo-labels generated from BDD-A's raw images performs the best on the test sets of two other datasets (BDD-A, DADA-2000). On the test set of the DR(eye)VE dataset, the network trained with pseudo-labels generated from DR(eye)VE's raw images performs the best on the CC metric, while the network trained with pseudo-labels generated from DADA-2000's raw images performs the best on the KLD metric indicating a superior transferability of our method. Furthermore, we discover that the images from BDD-A capture more diverse and generalized self-driving scenes, resulting in more useful and reliable pseudo-labels for our unsupervised method. Hence, our final model in this work uses the pseudo-labels generated from BDD-A.

4.4. Ablation Studies

Impact of different modules. In Table 3, we examine each module of our proposed unsupervised model to verify its effectiveness. It can be seen that unsupervised training of APB with the pseudo-label generated from BDD-A achieves the worst performance. When we include UMB with multiple branches, the performance of the model improves significantly, far exceeding APB. Further, by adding the non-local block, we can also observe an obvious improvement. Finally, KEB brings a solid improvement to the model, making the results of our full model compatible with the state-of-the-art fully supervised models. In a word, each module in the study contributes to the final performance, while the proposed modules in this paper (UMB and KEB) contribute the most.

Different source of pseudo-labels. To examine the effect of different sources of pseudo-labels on the final results, we compare the performance of different pseudo-labels as is shown in Table 4. The first two rows indicate the results of training with a single source pseudo-label (*e.g.* ML-Net or UNISAL), while the third row indicates the best results of training with two sources pseudo-labels together (*i.e.,* ML-Net+UNISAL) to explore uncertainty, demonstrating our UMB is able to enhance the final performance through the interaction between multiple sources of pseudo-labels. However, more than two sources of pseudo-labels result in a performance drop, as illustrated in the subsequent few lines. Therefore we choose two source pseudo-labels (ML-Net and UNISAL) in all our experiments.

Prior knowledge. The proposed KEB in our model is used to migrate the self-driving or traffic knowledge to refine the generated pseudo-labels from the model pre-trained on the natural scenes. However, there remain problems, 1) what prior traffic knowledge should be added? 2) How to add such prior knowledge to the generated pseudo-

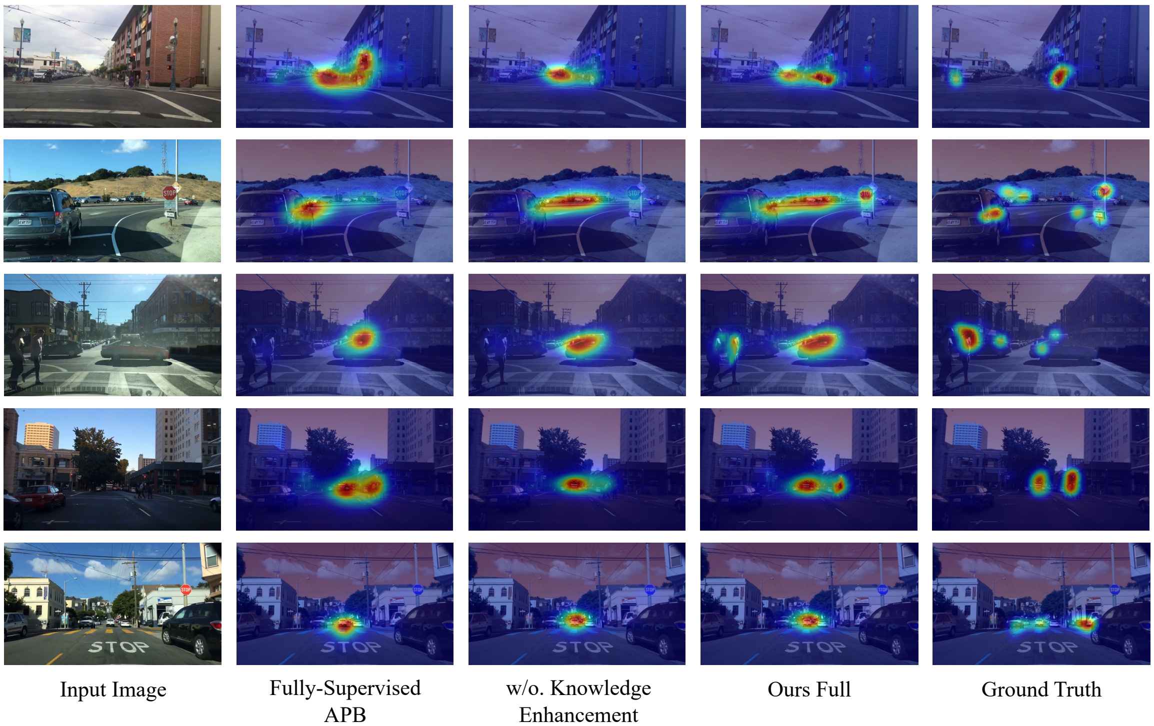

Figure 5. Visualization of the attention prediction results from different methods, *i.e.,* fully-supervised APB, our method without knowledge embedding, and our full method. The results show the effectiveness of our full model in locating critical areas in the driving scene. A failure case is shown in the last row.

label? Hence, we explore different ways of adding prior knowledge to add it more effectively. Here we use a pre-trained Mask R-CNN [17] to segment important traffic instances denoted as "objects" like pedestrians and traffic lights, and we also adopt a pre-trained OCR text detection model (EAST [63]) to segment important texts denoted as "text" like road signs and billboard. We can see in Table 5 that segmenting only important traffic instances achieve the best performance. Furthermore, we examine two different adding methods in KEB, *i.e.,* combining different categories of prior knowledge with pseudo-labels by concatenation along the channel dimension denoted as "concat", or by operation in Eq. 1 denoted as "single". As is shown in Table 5, the result indicates that using the operation in Eq. 1 works best.

Training strategy	KLD↓	CC↑
fully-supervised APB	**1.039**	**0.657**
semi-supervised v1	1.669	0.422
semi-supervised v2	1.130	0.629
unsupervised	1.099	0.635

Table 6. Comparing the different training paradigms, *i.e.,* supervised, semi-supervised and unsupervised settings.

Semi-supervised setting. In addition, we also compare the semi-supervised settings following [19] upon the same network, and the results are reported in Table 6. Specifically, we conduct two semi-supervised training schemes: 1) **Semi-supervised v1** refers to training the APB using $\frac{1}{4}$ of randomly sampled labeled data on BDD-A and then training the entire network using pseudo-labels generated from the remaining raw images; 2) **Semi-supervised v2** refers to the reversed process. However, as is shown in Table 6, we observe drastic drops in the result of the network in both Semi-supervised v1 and v2 compared with fully-supervised APB and are even inferior to our model trained in an unsupervised way. The poor performance can be explained by only using a small portion of the dataset tend to fool the model into learning a more restricted central bias, especially in self-driving. Our unsupervised method can leverage the information transferred from natural scenes by uncertainty mining, which is able to include more generalized information from non-traffic scenes to reduce bias.

4.5. Qualitative Results

Figure 5 shows visual comparisons of our model's variants on the BDD-A test set. We can observe that our full model achieves the best performance. For example, in the first row, the ground truth focuses on the pedestrians and traffic lights at the edge of the road, while the results of other methods show a strong center bias that put a lot of attention to the center of the road. Instead, our proposed model is able to reduce the central bias and assign higher attention values to the pedestrians and traffic lights in the scene which aligns with ground truth. In the second and third rows, our full model correctly focuses on the stop sign and the passing pedestrians, respectively. With an additional comparison between the third and fourth columns, we find that the proposed strategy successfully and effectively improves the final results and helps to focus on more important traffic areas of objects in the scene. To dive deep into the model's performance, a failure case is shown in the last row, where a truck tries to drive from right to left at the crossing. Our model (Ours Full) fails to focus on the truck, which is severely occluded with the nearby parked vehicles. An accurate object detection model can be further adopted to address this challenge in the future.

5. Conclusion

In this paper, we propose a novel unsupervised method for self-driving attention prediction. An uncertainty mining branch and a knowledge embedding block are introduced to generate reliable pseudo-labels and bridge the domain gap, respectively. Extensive experiments on three widely-used benchmarks demonstrate the effectiveness and superiority of our proposed method. In the future, we would incorporate the proposed method into the explainable autonomous driving control system.

6. Acknowledgement

This work was partly supported by the Funds for Creative Research Groups of China under Grant 61921003, the National Natural Science Foundation of China under Grant 62202063, Beijing Natural Science Foundation (No.L223002), the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) under Grant 2021QNRC001, the 111 Project under Grant B18008, the Open Project Program of Key Laboratory of Advanced Information Science and Network Technology of Beijing (No.XDXX2201) and Key Laboratory of Artificial Intelligence, Ministry of Education, Shanghai (No.AI2022005).

References

[1] 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 *CVPRW*, pages 54–60, 2016.

[2] Sonia Baee, Erfan Pakdamanian, Inki Kim, Lu Feng, Vicente Ordonez, and Laura Barnes. Medirl: Predicting the visual attention of drivers via maximum entropy deep inverse reinforcement learning. In *ICCV*, pages 13178–13188, 2021.

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *ICCVW*, pages 0–0, 2019.

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[5] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPRW*, 2015.

[6] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *TPAMI*, 41(3):740–757, 2018.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018.

[8] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *ICCV*, pages 502–511, 2019.

[9] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. In *ICPR*, pages 3488–3493. IEEE, 2016.

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

[11] Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, Nitin Singh, and Jeff Schneider. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *WACV*, pages 2095–2104, 2020.

[12] Richard Droste, Jianbo Jiao, and J Alison Noble. Unified image and video saliency modeling. In *ECCV*, pages 419–435. Springer, 2020.

[13] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *TITS*, 2021.

[14] Di Feng, Lars Rosenbaum, and Klaus Dietmayer. Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In *ITSC*, pages 3266–3273. IEEE, 2018.

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059. PMLR, 2016.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[19] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NeurIPSW*, 2015.

[20] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. Deepvs: A deep learning based video saliency prediction approach. In *ECCV*, pages 602–617, 2018.

[21] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–

1080, 2015.
[22] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, pages 4893–4902, 2019.
[23] Fumi Katsuki and Christos Constantinidis. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521, 2014.
[24] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.
[25] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS*, 30, 2017.
[26] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018.
[27] Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *CVPR*, pages 9661–9670, 2020.
[28] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *ECCV*, pages 563–578, 2018.
[29] Solomon Kullback. Information theory and statistics. *New York: Dover*, 1968.
[30] Fahad Lateef, Mohamed Kas, and Yassine Ruichek. Saliency heat-map as visual attention for autonomous driving using generative adversarial network (gan). *TITS*, 2021.
[31] 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 *ECCV*, pages 740–755. Springer, 2014.
[32] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018.
[33] Wu Liu, Xinchen Liu, Huadomg Ma, and Peng Cheng. Beyond human-level license plate super-resolution with progressive vehicle search and domain priori gan. In *ACM MM*, pages 1618–1626, 2017.
[34] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle reidentification for urban surveillance. In *ECCV*, pages 869–884. Springer, 2016.
[35] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *TMM*, 20(3):645–658, 2017.
[36] Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. A general framework for uncertainty estimation in deep learning. *IEEE Robot*, 5(2):3153–3160, 2020.
[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.
[38] Kai Lv, Hao Sheng, Zhang Xiong, Wei Li, and Liang Zheng. Improving driver gaze prediction with reinforced attention. *TMM*, 23:4198–4207, 2020.
[39] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *ICLR*, 2022.
[40] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *ICCV*, pages 2394–2403, 2019.
[41] Anwesan Pal, Sayan Mondal, and Henrik I Christensen. " looking at the right stuff"-guided semantic-gaze for autonomous driving. In *CVPR*, pages 11883–11892, 2020.
[42] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver's focus of attention: the dr (eye) ve project. *TPAMI*, 41(7):1720–1733, 2018.
[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019.
[44] Karl Pearson. Notes on the history of correlation. *Biometrika*, 13(1):25–45, 1920.
[45] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *CVPR*. IEEE, 2019.
[46] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *Proc. ECCV*. springer, 2018.
[47] Mengshi Qi, Jie Qin, Yi Yang, Yunhong Wang, and Jiebo Luo. Semantics-aware spatial-temporal binaries for cross-modal video retrieval. *IEEE Transactions on Image Processing*, 30:2989–3004, 2021.
[48] Mengshi Qi, Jie Qin, Xiantong Zhen, Di Huang, Yi Yang, and Jiebo Luo. Few-shot ensemble learning for video classification with slowfast memory networks. In *MM*. ACM, 2020.
[49] Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. Stc-gan: Spatio-temporally coupled generative adversarial networks for predictive scene parsing. *IEEE Transactions on Image Processing*, 29:5420–5430, 2020.
[50] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. Ke-gan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *Proc. CVPR*. IEEE, 2019.
[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.
[52] Khurram Soomro and Amir R Zamir. Action recognition in realistic sports videos. In *CVSports*, pages 181–208. Springer, 2014.
[53] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017.
[54] Guotao Wang, Chenglizhao Chen, Deng-Ping Fan, Aimin Hao, and Hong Qin. From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach. In *CVPR*, pages 15119–15128, 2021.
[55] Wenguan Wang and Jianbing Shen. Deep cropping via attention box prediction and aesthetics assessment. In *ICCV*, pages 2186–2194, 2017.
[56] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *CVPR*, pages 4894–4903, 2018.
[57] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.
[58] Yifan Wang, Wenbo Zhang, Lijun Wang, Ting Liu, and

Huchuan Lu. Multi-source uncertainty mining for deep unsupervised saliency detection. In *CVPR*, pages 11727–11736, 2022.
[59] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. Predicting driver attention in critical situations. In *ACCV*, pages 658–674. Springer, 2018.
[60] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *CVPR*, pages 9523–9532, 2020.
[61] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency detection in 360 videos. In *ECCV*, pages 488–503, 2018.
[62] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.
[63] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *CVPR*, pages 5551–5560, 2017.