MAAD: A Model and Dataset for "Attended Awareness" in Driving - Supplementary Material

In this supplementary material, we provide more details regarding the network modules, annotation dataset collection procedures and statistics, additional visualizations of network capabilities, and results from ablation experiments.

1. Model details

1.1. Network Architecture

1.1.1 Encoder Structure

The encoder consists of 5 stacked layers of 2 different types of spatio-temporal convolutional modules. The first two layers (denoted as a 'layer1' and 'layer2') are taken from a pretrained ResNet18 structure. The weights of these two layers are frozen during training. The remaining 3 layers of the encoder (denoted as 'S3D_1', 'S3D_2', 'S3D_3') are separable 3D convolutional (S3D) modules. Each S3D module consists of two separate 3D convolutions, one for spatial and the other for temporal processing. Detailed structure of the separable convolutional encoder modules are shown in Table 1. The output of 'S3D_3' undergoes a 3D convolution post-processing step in order to reduce the number of features from 512 to 128. The output of this post-processing step is then fed into the first decoder unit (DU5) of the decoder along with the side-channel information.

Encoder S3D ID	Structure
S3D_1	S3D(in=128, out=256)
S3D_2	S3D(in=256, out=512)
S3D_3	S3D(in=512, out=512)

Table 1. Detailed structure of the 3D convolution modules used in the Decoder Units. The spatial Conv3d in the encoder S3D modules uses kernel size of $1\times3\times3$ and a stride length of 1. Similarly, the Conv3D responsible for temporal processing relies on a kernel of $3\times1\times1$. A replication pad of size 1 is applied to the input before being processed by each of the Conv3D modules.

1.1.2 Gaze Transform Module

The gaze transform module consists of a single layer MLP whose output is encoded as a multi-channel Voronoi map which then is provided as a side channel input to the decoder units. The number of gaze points used per frame (for supervision as well as the side channel information) is fixed to be 3. The side-channel gaze input was corrupted by a spatially varying zero mean Gaussian white noise with $\sigma = 0.0347$, to account for the uncertainty due to both the foveal center location and eye tracker error; both treated as two Gaussian independent sources. Each Voronoi channel encodes a particular distance related feature, such as dx, dy, dx^2, dy^2, $dxdy$, $\sqrt{dx^2 + dy^2}$. Additionally, we also provide a bit to encode whether a particular instance of the gaze input is dropped out (as a result of the dropout applied during training) and also whether the gaze value is a valid input or not (to indicate NaNs that occur in the gaze data primarily due to eye blinks and tracker error). The total number of channels for the gaze side information is 8.

1.1.3 Optic Flow Module

The optic flow is provided as a 2-channel input, where the channels encode the flow in the horizontal and vertical direction respectively. We apply an adaptive average pool operator on the optic flow input to match the resolution of the decoder unit.

1.1.4 Decoder Unit

Each Decoder Unit (DU) can receive up to three sources of input, 1) the skip connections from the encoder, 2) the side channel information (gaze information, and optic flow) and 3) the output of the previous decoder unit, when available.

All S3D modules (for skip modules as well as side-channel+previous output modules) in each of the Decoder Unit uses a S3D unit with a kernel size of $1\times3\times3$ for spatial and $3\times1\times1$ for temporal processing. The input to each of the spatial and the temporal modules in the S3D uses a replication pad of size 1. An InstanceNorm3D and a ReLU nonlinearity is applied after the spatial and temporal processing.

The output of the skip connection module is concatenated channel-wise to the side-channel information and the output of the previous decoder unit. The concatenated input is processed by another S3D module finally undergoes a bilinear upsampling to match the resolution size of the next

Decoder Unit Id	Skip Module	Concatenated Module
DU5	NA	S3D(in=266, out=128)
DU4	NA	S3D(in=138, out=64)
DU3	NA	S3D(in=74, out=32)
DU2	S3D(in=128, out=128)	S3D(in=170, out=16)
DU1	S3D(in=64, out=64)	S3D(in=90, out=16)

Table 2. Detailed structure of the S3D convolution modules used in the Decoder Units. The spatial Conv3d in the S3D modules uses kernel size of 1×3×3 and a stride length of 1. Similarly, the Conv3D responsible for temporal processing relies on a kernel of 3×1×1. A replication pad of size 1 is applied to the skip connection input to ensure that the output can be concatenated channel-wise to the other side-channel input and the previous decoder unit output.

decoder unit. The output of last decoder unit (DU1) undergoes a final bilinear upsampling stage to match the resolution of the size of model input (240×135). The detailed structure of all the decoder units in the decoder is presented in Table 2.

The total number of channels from the side information is 10 (Voronoi gaze maps=8, and optic flow=2). In general, the following relationship holds for the feature sizes:

$$n_{in}^{concat,DU(l)} = n_{out}^{skip,DU(l)} + n_{out}^{concat,DU(l+1)} + 10$$

where n_{in} and n_{out} are the number of input and output features respectively and $l \in [1, 2]$ denotes the decoder unit id. For DU5, $n_{in} = n_{out}^{encoder,postproc} + 10$.

1.1.5 Gaze and Awareness Convolutional Modules

The output of the decoder is processed using a Conv2D with kernel operator of size 5×5 and 6 output features to generate a feature map M. The 1D gaze heatmap (p_G) is produced from M by a Conv1D operator with a kernel size of 1 followed by softmax operator to ensure that the heatmap is a valid probability distribution. Likewise, the awareness heatmap (M_A) is generated from M by another Conv1D operator with a kernel size of 1 followed by a sigmoid operator to ensure that each pixel value remains between 0 and 1. Note that, the awareness map is NOT a probability distribution.

1.2. Cost Weights and Parameters

Table 3 contains all the parameters and coefficients used for model training. These coefficients were chosen so that the relative magnitudes of the different supervisory terms were comparable. The regularization terms are roughly an order of magnitude lower than the supervisory cues. The gaze and awareness supervision costs are computed only on valid gaze points (gaze points that are not NaNs).

2. Annotation Dataset Details

Table 4 shows the breakdown of the labelled set. Table 5 contains information regarding the time of the day and the weather condition for all the 8 video sequences (from the Dr(Eye)ve dataset) we used for MAAD model training.

α_{G}	1.2
α_{ATT}	12.0
α_{AA}	1.0
$\alpha_{\text{S-A}}$	100.0
$\alpha_{\text{S-G}}$	5×10^{10}
α_{T}	600.0
α_{DEC}	1.5×10^{6}
α_{CAP}	0.01
$\alpha_{\text{CON-G}}$	1×10^{7}
$\alpha_{\text{CON-A}}$	10.0
w_{OF}	0.5
ϵ_{DEC}	0.2

Table 3. Cost term coefficients and parameters used for training.

Modifier	Num. annotations	Mean awareness
Null	16,366	0.719
Blurred	8,346	0.657
Flipped	8,311	0.674
Road-only	9,665	0.542
Reading-text	11,295	0.593

Table 4. Number of annotations and mean awareness from annotations grouped according to cognitive task modifier. Annotations reflect the variability in awareness of locations under certain cognitive modifiers, including conditions where we expect reduced awareness of annotated locations (e.g. reading text and road-only conditions).

We randomly sampled approximately 10s clips from these 8 videos from within the data we collected for third-party attended awareness annotation. The gaze data was overlaid on the video clip and in the last frame of the clip a random location was chosen and marked with a red cross. This random location was chosen equi-probably from objects, edges or anywhere in the image. After the annotators watched the video, they were asked whether they believed the subject had attended to the location marked with the red cross. More specifically, the annotators answered the fol-

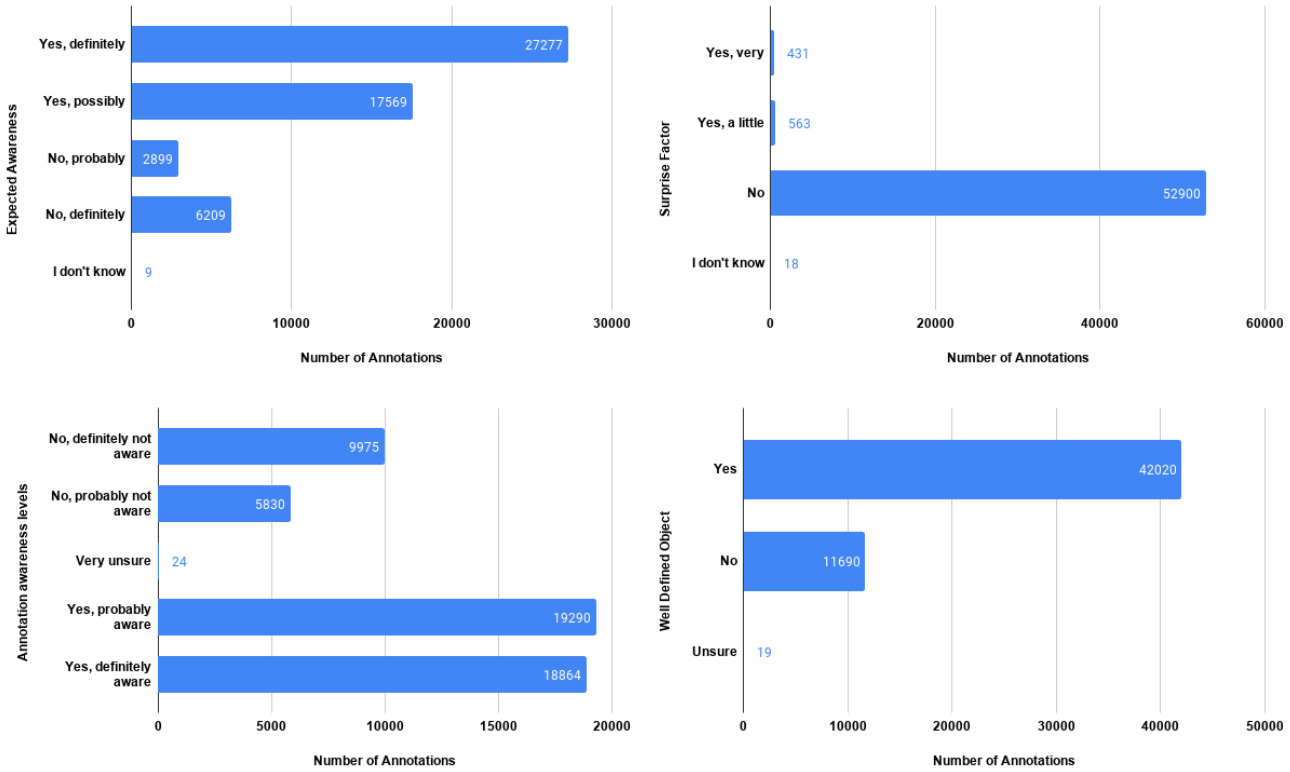

Figure 1. Annotator response distribution. Very few awareness labels were truly unsure (probably due to the explicit definition of the questions).

VIDEO ID	Time	Weather
VID06	Morning	Sunny
VID07	Evening	Rainy
VID10	Evening	Rainy
VID11	Evening	Cloudy
VID26	Morning	Rainy
VID35	Morning	Cloudy
VID53	Evening	Cloudy
VID60	Morning	Cloudy

Table 5. Time of the day and the weather condition for each of the videos (from the Dr(Eye)ve dataset) used for MAAD training. All driving sequences occur in an urban (downtown) setting.

lowing questions:

- Do you think the driver is aware of the object/area? (red cross; must be near the green circle at some point in the video, not being near at the end of the video is fine, if it is close and moving along with the object, we want a human judgment of someone who has the extra knowledge and is focusing on this) a) Yes, definitely aware b) Yes, probably aware c) Very unsure d) No, probably not aware e) No, definitely not aware
- Is the red cursor on a well-defined object such as a car or person? (not well defined: exit, piece of road, something you cannot put a boundary around. If it is part of an object, then it is still well defined. For example, building is not well defined because its a large area and cannot be separated from the ground) a) Yes b) No c) Unsure.
- If you are driving and are concentrated on driving, would you expect to be aware of this object? (red cross. Based on everything you see in the video) a) Yes, definitely b) Yes, possibly c) No, probably d) No, definitely e) I don't know.
- Were you surprised by the behavior or appearance of the highlighted object/region in the video? (red cross; jumped suddenly, didnt expect to see it, didnt see it coming, near accidents) a) Yes, very b) Yes, a little c) No d) I don't know.

Figures 1 and 2 the responses from the annotators and the distribution of annotation video snippets respectively.

3. Examples of Calibration Optimization

Figure 4 shows more examples of how the network successfully corrects a miscalibrated side-channel gaze input.

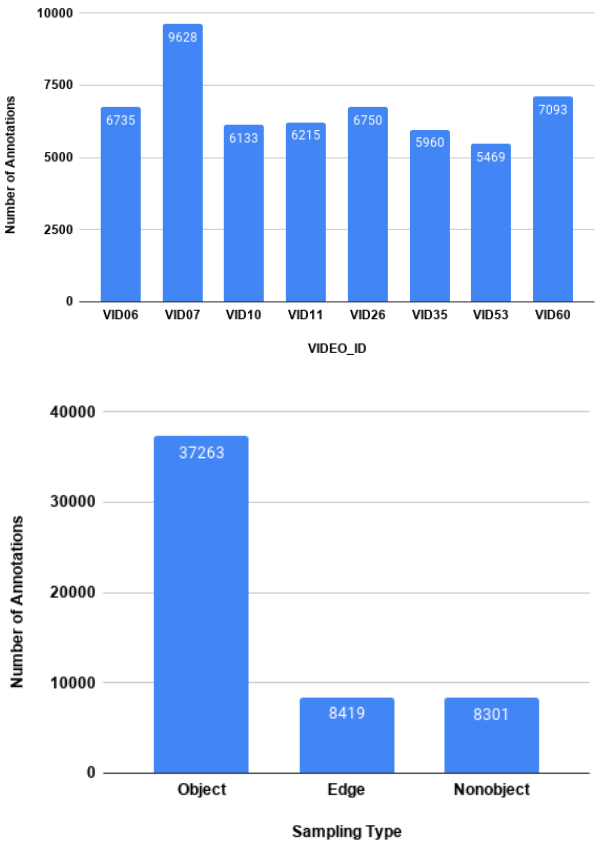

Figure 2. Top: Distribution of annotation video snippets. All video sequences were well represented, aside from some variance due to the criteria used to select desired and undesired sampling points. The video ids on the horizontal axis refer to the original video sequence ids in the Dr(Eye)ve dataset. Bottom: Distribution of final location for annotation according to sampling type.

In each of the examples in the figure, before correction the miscalibrated gaze distorts the heatmap and pulls it away from the ground truth gaze. As the networks learns the correction transform (for this experiment, the correction transform was learned by training the network on the test split that was used during the original model training phase), it corrects for the miscalibration and the heatmap begins to align closely with the ground truth gaze. Note that, during the optimization procedure for learning the correction the weights of the entire network except that of $T_{correct}$ are kept frozen.

4. Visualization of Denoising Mean Shift Traces

The meanshift algorithm is a procedure for locating the local maxima—the modes—of a density function. For the gaze denoising experiment, we perform the exact same meanshift procedure on three different density maps a) the gaze-conditioned saliency map, b) pure saliency map and c) the mask image that encodes the objects in the scene. Figure 5 shows different examples of the traces of the mean shift procedure on the mask image, gaze map without and with side channel gaze. In general, we see that when mean shift is performed on the gaze-conditioned saliency maps the resulting mode is closer to the ground truth (right-most column in Figure 5).

Ablation	Awareness Estimate
$\mathcal{L}_{ACAP}$	0.167
$\mathcal{L}_{DEC}$	0.073*
$\mathcal{L}_{S-A}$	0.157
$\mathcal{L}_{S-G}$	0.143
$\mathcal{L}_{ACAP}, \mathcal{L}_{AA}$	0.138*
$\mathcal{L}_{ACAP}, \mathcal{L}_{S-G}$	0.270
$\mathcal{L}_{ACAP}, \mathcal{L}_{ATT}$	0.444
$\mathcal{L}_{ACAP}, \mathcal{L}_{T}$	0.165
$\mathcal{L}_{ACAP}, \mathcal{L}_{S-G}, \mathcal{L}_{S-A}$	0.134*
$\mathcal{L}_{T}, \mathcal{L}_{S-G}$	0.264
$\mathcal{L}_{T}, \mathcal{L}_{S-G}, \mathcal{L}_{S-A}$	0.146
Full model	**0.138**

Table 6. Attended awareness estimation (mean squared error) on the test set using different ablations of MAAD. The testing noise level was set to be $\sigma_n = 0.1$. The result highlighted in red indicates the anomalous case in which the awareness heatmap is no longer spatially localized and hence results in gross overestimation of attended awareness. The results highlighted in blue indicate ablations for which the results were comparable to the full model but resulted in training instability. For more discussion on results with asterisk please refer to the text in Section 5.

5. Ablation Experiments

We performed a set of leave-N-out ablations to investigate the impact of different regularization terms on the network's ability to estimate attended awareness. Table 6 shows the mean squared error in the awareness estimation for different ablations that we tested.

Regularization for stability: One of the key functions of the regularization terms is to provide stability during training. In our ablation experiments we found that ablating the attention capacity regularization term ($\mathcal{L}_{ACAP}$) in general, resulted in training instability and in truncated training runs despite seemingly comparable (and at times better) awareness estimation scores to the full model.

We also experimented with a different network architecture in which the S3D modules in the decoder units were replaced with standard Conv3D modules. Due to the larger number of parameters for Conv3D modules, the number of layers in the encoder and decoder were reduced to 4. For these architectures, we found that including the spatial regularization for the gaze map ($\mathcal{L}_{S-G}$) was critical for stability

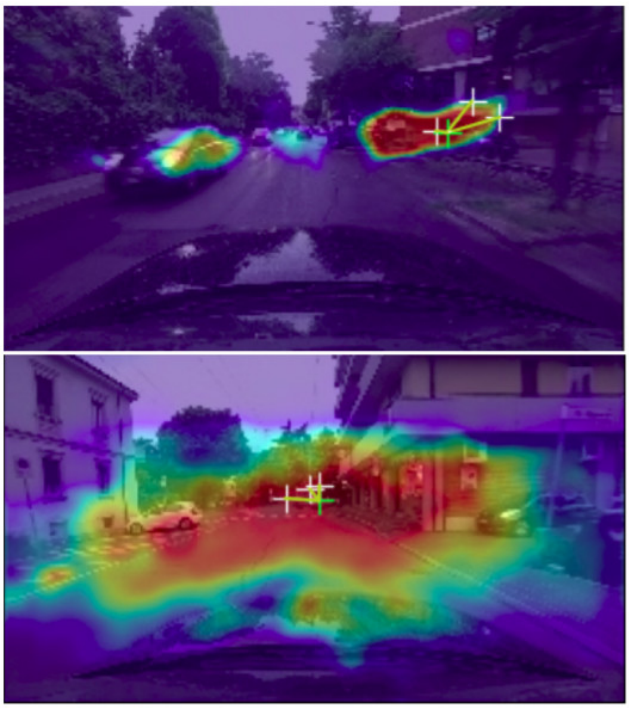

Figure 3. Comparison of awareness heatmap with (top) and without $\mathcal{L}_{DEC}$ (bottom). The heatmap is much more localized when decay term is present in the cost function.

during training.

In general, from our ablation experiments we recommend that for both the S3D and non-S3D versions of the model, the spatial regularization ($\mathcal{L}_{S-G}$) and the attention capacity ($\mathcal{L}_{ACAP}$) cost terms should be added to improve training stability.

$\mathcal{L}_{DEC}$ **ablation**: Although removing the decay term, ($\mathcal{L}_{DEC}$), resulted in better awareness estimation scores (row 2, Table 6, this was due to the fact that without $\mathcal{L}_{DEC}$ the awareness heatmap was no longer spatially localized as shown in Figure 3 essentially resulting in over-estimation of attended awareness. Over-estimation of driver awareness (model falsely predicting that the driver is aware of something when they are not) can lead to undesirable consequences when used in safety warning systems in autonomous vehicles. Additionally, utilizing $\mathcal{L}_{DEC}$ also accelerated the convergence of the model during training.

6. Influence of Cognitive Task Modifiers

During the dataset collection procedure we opted for a high-accuracy gaze tracker. However, this raises a question about the effect of the cognitive task modifier in a passive observation experiment. In order to investigate the impact of cognitive task modifiers as a latent factor that could influence awareness estimation accuracy, we trained MAAD exclusively on training data collected under the 'null condition'. This model was then evaluated on the data collected under the remaining cognitive task modifier conditions.

Task Modifier	Null Condition Model	Full Model
Null Condition	**0.110**	0.139
Reading-Text	0.211	**0.132**
Blurred	0.220	**0.140**
Flipped	0.231	**0.166**
Roadonly	0.171	**0.114**

Table 7. Awareness estimation mean squared error: Breakdown of the results according to cognitive task modifier type for full model and the model training on only null condition data.

Noise level	MSE, FG	MSE, MAAD
Null Condition	**0.110**	0.492
Reading-Text	0.211	**0.393**
Blurred	0.220	**0.425**
Flipped	0.231	**0.454**
Roadonly	0.171	**0.330**

Table 8. Mean squared error awareness estimates with spatio-temporal Gaussian with optic flow (FG) and our proposed approach (MAAD) according to cognitive task modifier type for the model trained only on null condition data.

From Table 7 we can see that a model that was trained exclusively on null condition data performed worse on the other task modifiers compared to the full model. However, as shown in Table 8, the null condition model still did considerably better than the spatio-temporal Gaussian baseline (FG) with optic flow. These results indicate that the model is sensitive to the cognitive task that the subject is executing. Future work will explore ways to disentangle this latent factor within the network capabilities.

7. Examples of Gaze and Awareness Maps

Figures 6,7 and 8 are more examples of gaze and awareness maps for different interesting scenarios that arise during driving. Figures 6 and 7 are examples of gaze conditioned saliency and highlights the spatio-temporal persistence of awareness in situations where the subject's gaze shifts between multiple driving-relevant entities (such as pedestrians, traffic lights and other cars) in the scene. Figure 8 is a unique example which illustrates how the network can handle occlusions.

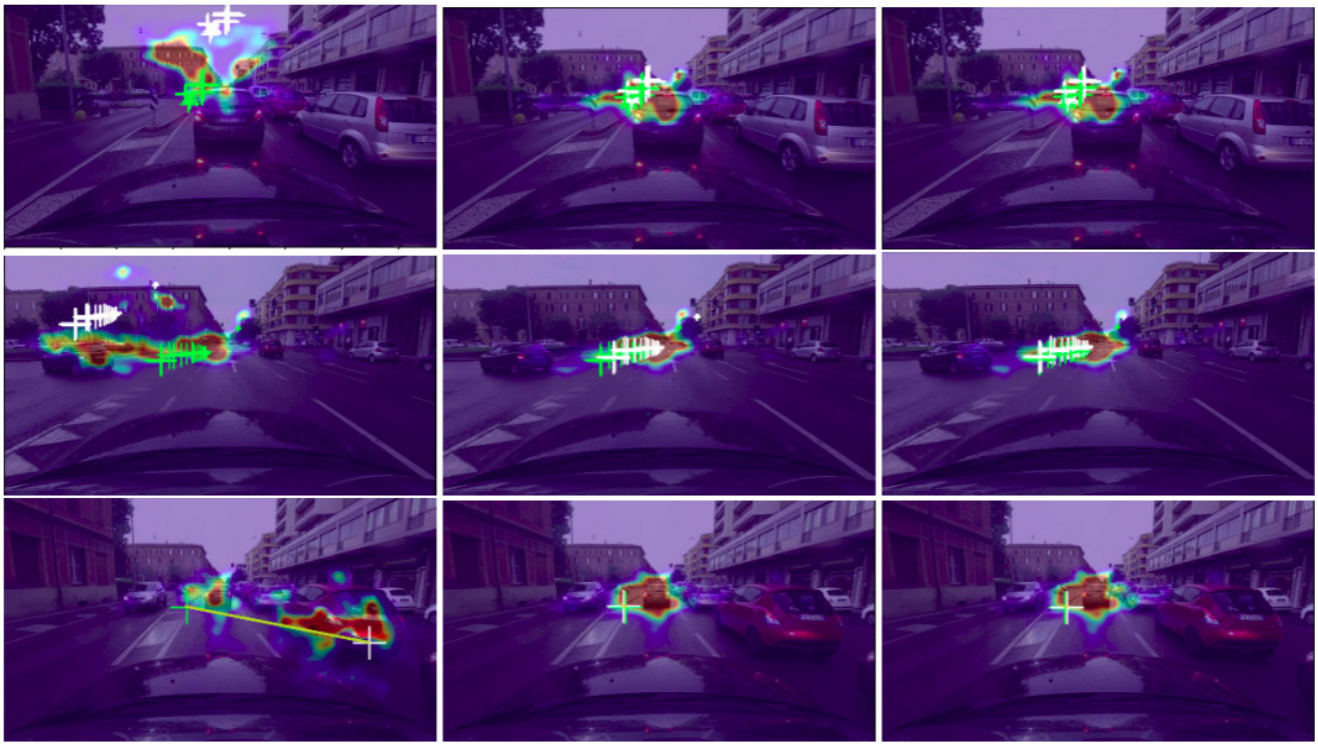

Figure 4. Examples of how the network learns a correction transform to correct for a miscalibrated gaze input. In each of the examples, the corrupted gaze input is marked as white cross hairs and the ground truth gaze is marked as lime-green cross hairs. Left column: Gaze maps before calibration. Due to the corruption applied, we can see that the side channel gaze is far away from the ground truth gaze. Middle and Right Column: As the optimization progresses, the networks learns to correct for the miscalibration and brings the side channel input close to the ground truth. In all these examples, the noise level was set to be 0.3.

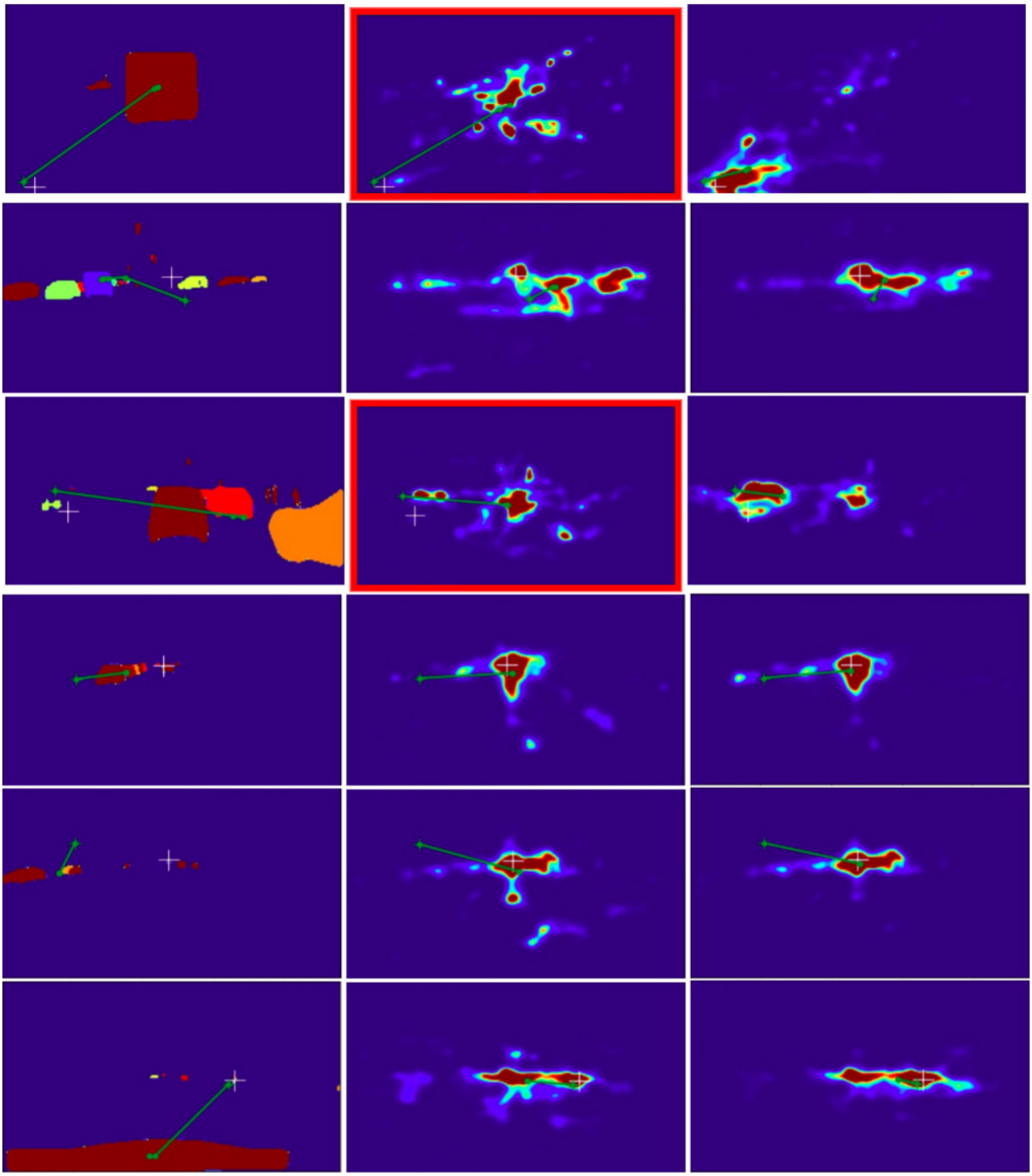

Figure 5. Examples of meanshift operation performed on Left: Object Masks, Middle: Gaze map without side-channel (pure saliency) information. The examples highlighted using the red rectangle indicate extreme failure cases. Right: Gaze map with side channel noisy gaze (our approach). The meanshift sequences is shown as green polylines. The starting point (the noisy gaze) of the sequence is indicated using a green crosshair. The ground truth gaze is denoted as white cross on the images. Our approach with noisy side channel gaze outperforms the object-based and pure saliency-based approaches.

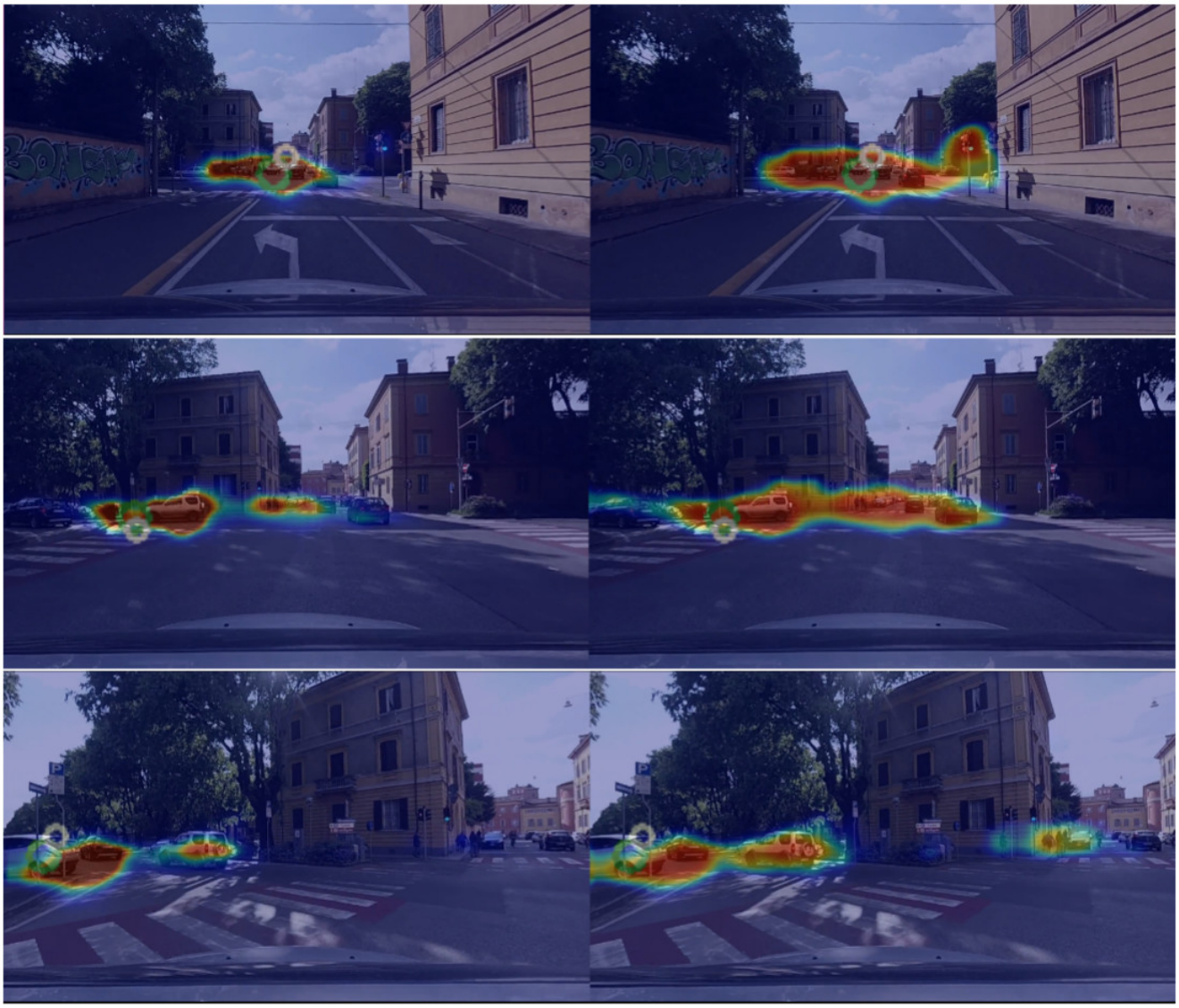

Figure 6. Gaze and awareness map as the ego car performs a left turn. Top Left: The car approaches an intersection and is about to make a left turn. The gaze map and the awareness map are primarily concentrated in the center of image. Middle: The car has started the left turn maneuver. The gaze map has started to shift leftward. However, the awareness map is much more smooth and indicates awareness of the cars ahead of the ego car in the previous frame. Bottom Middle: The left turn maneuver is almost complete and the gaze map is almost completely shifted to the left hand side. The awareness map still exhibits temporal persistence of objects that were attended to a few seconds before. In this figure, the gaze and awareness maps are in the left and right column respectively. The green circle indicate the ground truth gaze and the white circles indicate the noisy side channel information fed into the network during inference.

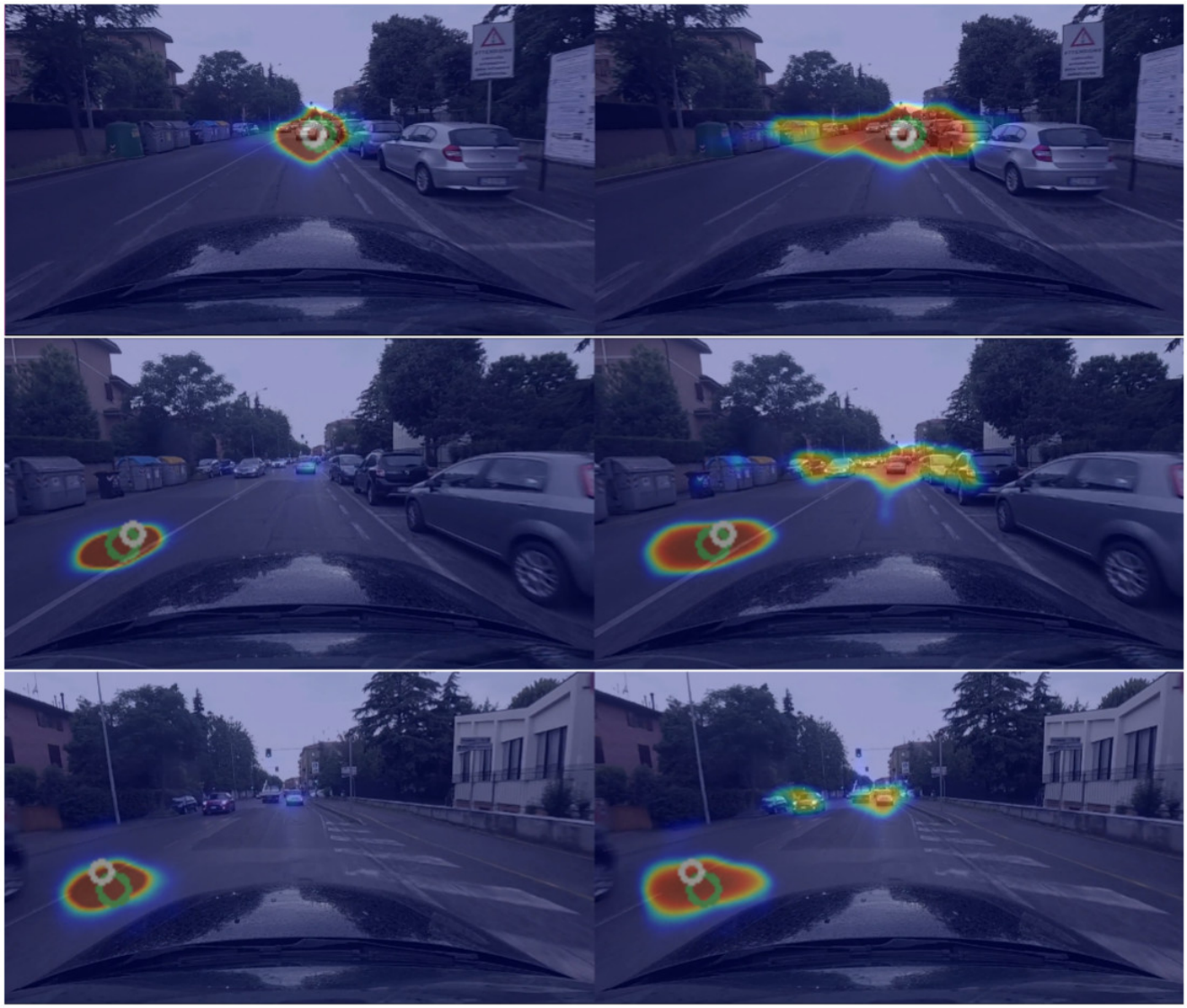

Figure 7. Gaze conditioned saliency and awareness maps with shifting gaze and incoming traffic: Top: The gaze is fixated straight ahead and on the incoming traffic. Middle: The gaze map shifts to the bottom left (reading text). The awareness map exhibits multiple regions of activation: a) for the newly attended region, b) the car straight ahead and c) the incoming traffic. Bottom: The gaze maps remains almost the same as the subject continues to gaze in the bottom left. As the incoming traffic approaches closer to the ego car, the activation levels of the awareness map have weakened and furthermore, the activation regions have separated indicating spatial and temporal persistence attached to objects.

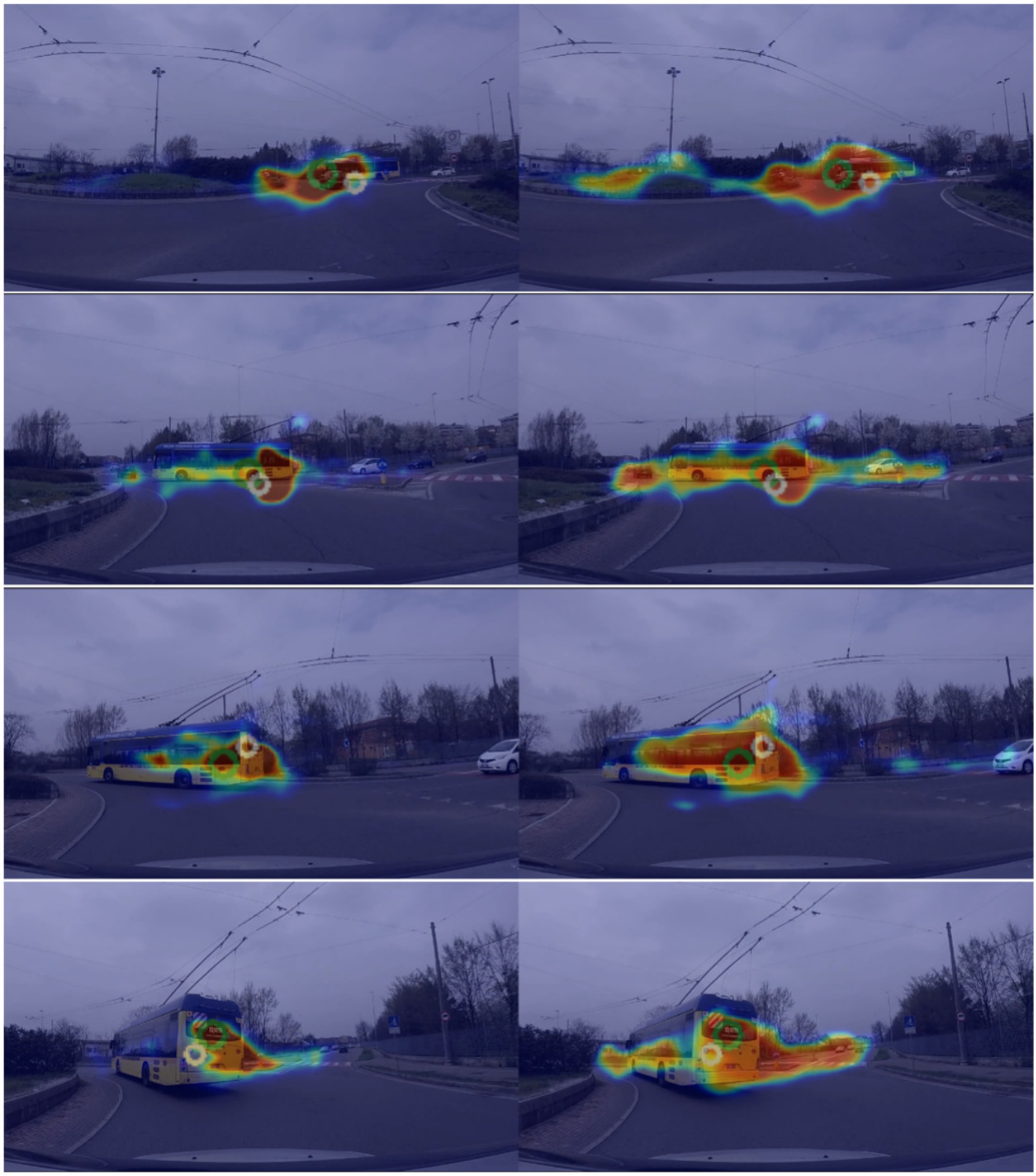

Figure 8. Gaze conditioned saliency and awareness maps with occlusion: Top: As the ego car approaches the traffic roundabout, the gaze is impinged on the car ahead. The awareness map reflects the fact that the subject is aware of the car as well. Top Middle: The gaze has shifted toward the bus on the right and the car ahead is about to be occluded by the bus. The awareness is split between the car and the vehicles on the right. Bottom Middle: The car is no longer visible due to occlusion. The gaze and the awareness activation is solely on the bus. Bottom: The car has reappeared in the visual field after occlusion. The gaze activation continues to be on the bus. The awareness map reassigns positive awareness on the previously attended car demonstrating how the model can 'remember' past attended objects despite occlusions.