In: *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, pp.584-591, Virtual Event / Montreal, Canada, 19-27 August 2021.

GASP: Gated Attention for Saliency Prediction

Fares Abawi[*] , **Tom Weber** and **Stefan Wermter**
University of Hamburg
{abawi, tomweber, wermter}@informatik.uni-hamburg.de

Abstract

Saliency prediction refers to the computational task of modeling overt attention. Social cues greatly influence our attention, consequently altering our eye movements and behavior. To emphasize the efficacy of such features, we present a neural model for integrating social cues and weighting their influences. Our model consists of two stages. During the first stage, we detect two social cues by following gaze, estimating gaze direction, and recognizing affect. These features are then transformed into spatiotemporal maps through image processing operations. The transformed representations are propagated to the second stage (*GASP*), where we explore various techniques of late fusion for integrating social cues and introduce two sub-networks for directing attention to relevant stimuli. Our experiments indicate that fusion approaches achieve better results for *static* integration methods, whereas non-fusion approaches for which the influence of each modality is unknown result in better outcomes when coupled with recurrent models for *dynamic* saliency prediction. We show that gaze direction and affective representations contribute a prediction to ground-truth correspondence improvement of at least 5% compared to dynamic saliency models without social cues. Furthermore, affective representations improve GASP, supporting the necessity of considering *affect-biased attention* in predicting saliency.[1]

1 Introduction

Attending to regions or objects in our perceptual field implies an interest in acting towards them. Humans convey their attention by fixating their eyes upon those regions. By modeling fixation, we gain an understanding of the events that attract attention. These attractors are represented in the form of a *Fixation Density Map* (FDM), displaying blurred peaks on a two-dimensional map, centered on the eye *Fixation Point* (FP) of each individual viewing a frame. The FDM is a visual representation of saliency, a useful indicator of what attracts human attention.

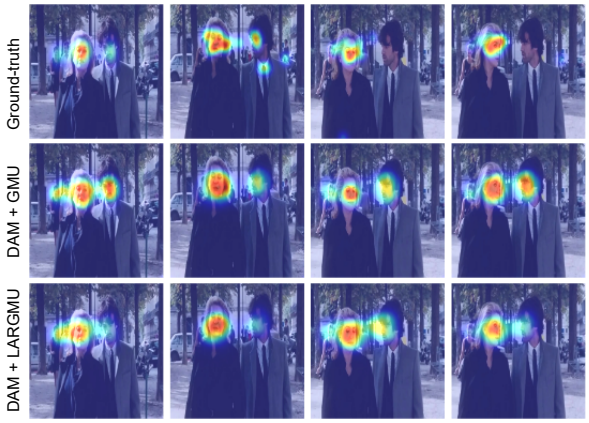

Figure 1: Frame predictions on the Coutrot Database 1 [Coutrot and Guyader, 2014]. DAM + GMU: Directed Attention Module followed by the Gated Multimodal Unit for static integration (middle); DAM + LARGMU: Directed Attention Module followed by the Late Attentive Recurrent GMU for sequential integration (bottom).

Early computational research focused on bottom-up saliency, by which the conspicuity of regions in the visual field was purely dependent on the stimuli [Itti and Koch, 2001; Bruce and Tsotsos, 2009]. On the other hand, task-driven approaches are top-down models utilizing supervised learning for performing tasks and allocating attention to regions or objects of interest. Combining face detections with low-level features has been shown to outperform bottom-up saliency models agnostic to social entities in a scene. Birmingham *et al.* [2009] corroborate the advantage of facial features in modeling saliency. They establish that when social stimuli are present, humans tend to fixate on facial features, a phenomenon weakly portrayed by bottom-up saliency detectors. Moreover, studies on human eye movements indicate that bottom-up guidance is not strongly correlated with fixation, which is rather influenced by the task [Foulsham *et al.*, 2011]. The existence of social stimuli in a scene alters fixation patterns, supporting the notion that even with the lack of an explicit task, we form intrinsic goals for guiding our gaze.

Although facial features attract attention, studies show that humans tend to follow the gaze of observed individu-

[*]Contact Author
[1]Code: http://software.knowledge-technology.info#gasp

als [Bylinskii *et al.*, 2016]. Additionally, psychological studies [Pritsch *et al.*, 2017] indicate a preference in attending towards emotionally salient stimuli over neutral expressions, a phenomenon described as *affect-biased attention*. By augmenting saliency maps with emotion intensities, affect-biased saliency models show significant improvement over affect-agnostic models [Fan *et al.*, 2018; Cordel *et al.*, 2019]. These approaches, although exclusive to static saliency models, are not limited to facial expressions, allowing for a greater domain coverage irrespective of the presence of social entities in a scene.

In light of the social stimuli relevance to modeling attention, we design a model to predict the FDM of multiple human observers watching social videos as shown in Figure 1. Such models employ top-down and bottom-up strategies operating on a sequence of images, a task referred to as *dynamic saliency prediction* [Bak *et al.*, 2017; Borji, 2019]. Our model utilizes multiple social cue detectors, namely gaze following and direction estimation, as well as facial expression recognition. We integrate the eye gaze and affective social cues, each with its spatiotemporal representation, as input to our saliency prediction model. We describe the resulting output from each social cue detector as a *feature map* (FM). We also introduce a novel FM weighting module, assigning different intensities to each FM in a competitive manner representing its priority. Each representation is best described as a *priority map* (PM), combining top-down and bottom-up features to prioritize regions that are most likely to be attended. We refer to the final model output as the Predicted FDM (PFDM).

Our work is motivated by the following findings: **1)** Task-driven strategies are pertinent to predicting saliency [Foulsham *et al.*, 2011]; **2)** Changes in motion contribute to the relevance of an object, underlining the importance of spatiotemporal features for predicting saliency [Min *et al.*, 2020]; **3)** Psychological studies indicate that attention is driven by social stimuli [Salley and Colombo, 2016]. To address the first finding, we state that our approach is task-driven by virtue of supervision since the objective is predicated on modeling multiple observer fixations. Although the datasets employed in this study were collected under a free-viewing condition, the top-down property is arguably maintained due to the intrinsic goals of the observer. These goals are driven by socially relevant stimuli addressed in our model through its reliance on multiple social cue modalities and facial information. We detect the social and facial features in a separate stage, hereafter described as the *Social Cue Detection* (SCD) stage.

To address the second finding, our model learns temporal features in two stages. Sequential learning in SCD is not a necessity but a result of the models employed for social cue detection, e.g., recurrent models or models pre-trained on optical flow tasks. In the second stage (GASP), we integrate social cues as illustrated in Figure 2. GASP also employs sequential learning, not only registering environmental changes such as color and intensity but also features pertaining to motion.

Finally, we consider social attention by employing an audiovisual saliency prediction modality, as well as social cue detectors (also described as modalities) that specialize in performing distinct tasks. Each of these tasks is highly relevant to visual attention, both from a behavioral and a computational perspective. We aim to explore feature integration approaches for combining social cues. We present gated attention variants and introduce a novel approach for directing attention to all modalities. To the best of our knowledge, our model is the first to consider affect-biased attention by using facial expression representations for dynamic saliency prediction based on deep neural maps.

Algorithm 1 SCD sampling and generation

Input:
Video and audio frames sampled from $\mathtt{ds}$ = AVE dataset
Parameters:
Window sizes $\mathtt{W}_{\mathtt{SP}} = 15, \mathtt{W}_{\mathtt{GE}} = 7, \mathtt{W}_{\mathtt{GF}} = 5, \mathtt{W}_{\mathtt{FER}} = 0$
O/P steps $\mathtt{T'}_{\mathtt{SP}} = 15, \mathtt{T'}_{\mathtt{GE}} = 4, \mathtt{T'}_{\mathtt{GF}} = 0, \mathtt{T'}_{\mathtt{FER}} = 0$
Output:
Modality windows $\mathtt{mdl}_{\mathtt{win}}$
O/P buffers $\mathtt{buf}_{\mathtt{mdl}}$

for $\mathtt{vid}$ in $\mathtt{ds}$ **do**
  Let $\mathtt{t} = 0$
  for $\mathtt{frm}$ in $\mathtt{vid}$ **do**
    Let $\mathtt{fcs}$ = face crops & bounding boxes in $\mathtt{frm}$
    for $\mathtt{mdl}$ in $\in \{$SP, GE, GF, FER$\}$ **do**
      if $\mathtt{W}_{\mathtt{mdl}} > \mathtt{t}$ **then**
        Let $\Delta = \mathtt{W}_{\mathtt{mdl}} - \mathtt{t}$
        for δ in $\in \{\Delta, ..., \mathtt{W}_{\mathtt{mdl}}\}$ **do**
          Let $\mathtt{mdl}_{\mathtt{win}}[\delta] = <\mathtt{frm}, \mathtt{fcs}>$
        end for
      else
        Shift $\mathtt{mdl}_{\mathtt{win}}$ by 1 to the left
        Let $\mathtt{mdl}_{\mathtt{win}}[\mathtt{W}_{\mathtt{mdl}}] = <\mathtt{frm}, \mathtt{fcs}>$
      end if
      Execute $\mathtt{buf}_{\mathtt{mdl}}[\mathtt{t}] = \mathtt{mdl}(\mathtt{mdl}_{\mathtt{win}})[\mathtt{T'}_{\mathtt{mdl}}]$
    end for
    Propagate $\mathtt{buf}[\mathtt{t}]$ to GASP
    Let $\mathtt{t} = \mathtt{t} + 1$
  end for
end for

2 Social Cue Detection

In the first stage (SCD), we extract high-level features from three social cue detectors and an audiovisual saliency predictor. We utilize the S^3FD face detector [Zhang *et al.*, 2017] for acquiring the face locations of actors in an image. The cropped face images are passed to the social cue detectors as input. The window size W, i.e., the number of frames fed simultaneously as input to each model, varies according to the requirements of each model. We generate modality representations at output timestep T' for each social video in AVE [Tavakoli *et al.*, 2020] as shown in Algorithm 1.

Our motivation behind selecting social cue detectors lies in their ability to generalize to various "in-the-wild" settings, regardless of the surrounding environment or lighting conditions. All chosen models were trained on datasets containing social entities captured from different angles and distances.

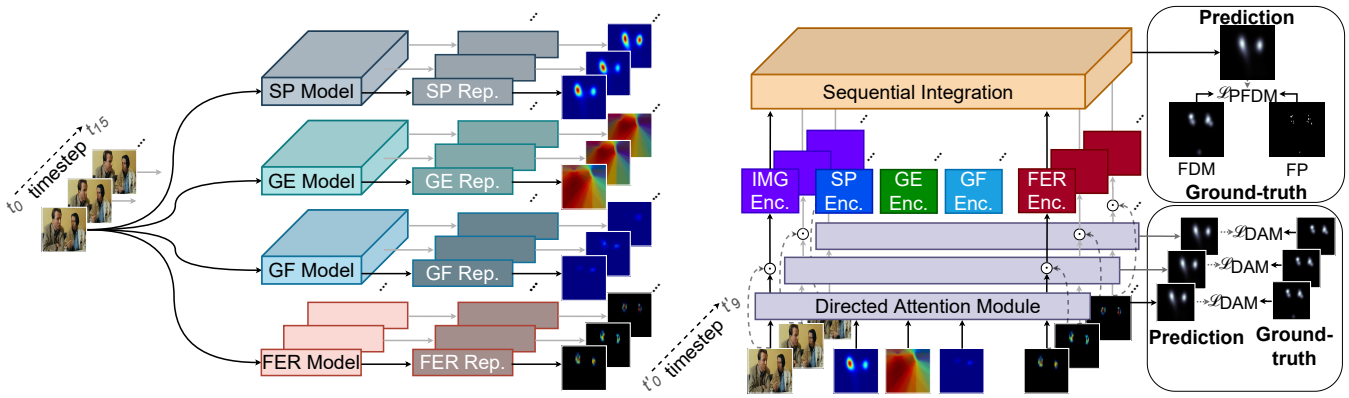

Figure 2: Overview of our sequential two-stage model. SCD (left) extracts and transforms social cue features to spatiotemporal representations. GASP (right) acquires the representations and integrates features from the different modalities.

2.1 Gaze Following Model (GF)

VideoGaze [Recasens *et al.*, 2017] receives the source image frame containing the gazer, the target frame into which the gazer looks, and a face crop of the gazer in the source frame along with the head and eye positions as input to its pathways. All frames are resized to 227 $\times$ 227 pixels. The model acquires five consecutive frames (W_{GF}) at timestep T'_{GF} and returns a fixation heatmap of the most probable target frame for every detected face in a source frame.

Representation. The mean fixation heatmaps resulting from each face in the source frame are overlaid on a single feature map in the corresponding target frame timestep. We transform the fixation heatmaps using a jet colormap.

2.2 Gaze Estimation Model (GE)

Gaze360 [Kellnhofer *et al.*, 2019] estimates the 3D gaze direction of an actor. The model receives face crops of the same actor over a predefined period, covering seven frames (W_{GE}) centered around timestep T'_{GE}. Each crop is resized to 224 $\times$ 224 pixels. The model predicts the azimuth and pitch of the eyes and head along with a confidence score.

Representation. We generate cones and position their tips on detected face centroids. The cones are placed upon a zero-valued map with identical dimensions to the input image. The cone base is rotated towards the direction of gaze. The apex angle of the cone is set to 60°. The face furthest from the lens is projected first with an opacity of 0.5, followed by the remaining faces ordered by their distances to the lens. A jet colormap is then applied to the cone map.

2.3 Facial Expression Recognition Model (FER)

We employ the facial expression recognition model developed by Siqueira *et al.* [2020]. The model is composed of convolutional layers shared across 9 ensembles. The model receives all face crops in a frame as input, each resized to 96 $\times$ 96 pixels, and recognizes facial expressions from 8 categories. Since the model operates on static images, we set the window size W_{FER} and output timestep T'_{FER} to 0.

Representation. Grad-CAM [Selvaraju *et al.*, 2017] features are extracted from all 9 ensembles. We take the mean of the features for all faces in the image and apply a jet colormap transformation on them. A 2D Hanning filter is applied to the features to mitigate artifacts resulting from the edges of the cropped Grad-CAM representations. We center the filtered representations on the face positions upon a zero-valued map with dimensions identical to the input image.

2.4 Audiovisual Saliency Prediction Model (SP)

In the SCD stage, we utilize DAVE [Tavakoli *et al.*, 2020] for predicting saliency based on visual and auditory stimuli. Separate streams for encoding the two modalities are built using a 3D-ResNet with 18 layers. The visual stream acquires 16 images (W_{SP}), each resized to 256 $\times$ 320 pixels. The auditory stream acquires log Mel-spectrograms of the video-corresponding audio frames, re-sampled to 16kHz. The model produces an FDM at the final output timestep T'_{SP} considering all preceding frames within the window W_{SP}.

Representation. We transform the resulting fixation density map from DAVE using a jet colormap.

3 Gated Attention for Saliency Prediction

We standardize all SCD features to a mean of 0 and a standard deviation of 1. The input image (IMG) and FMs are resized to 120 $\times$ 120 pixels before propagation to GASP.

3.1 Directed Attention Module (DAM)

The Squeeze-and-Excitation (SE) [Hu *et al.*, 2018] layer extracts channel-wise interactions, applying a gating mechanism for weighting convolutional channels according to their informative features. The SE model, however, emphasizes modality representations having the most significant gain, mitigating channels with lower information content. For our purpose, it is reasonable to postulate that the most influential FM channels are those belonging to the SP since it would result in the least erroneous representation in comparison to the ground-truth FDM. However, this causes the social cue modalities to have a minimal effect, mainly due to their low correlation with the FDM as opposed to the SP.

To counter bias towards the SP, we intensify non-salient regions such that the model learns to assign greater weights to modalities contributing least to the prediction. Alpay *et al.* [2019] propose a language model for skipping and preserving activations according to how surprising a word is, given its context in a sequence. In this work, we assimilate surprising words to channel regions with an unexpected contribution to the saliency model.

We construct a model for emphasizing unexpected features using two streams as shown in Figure 3: **1)** The inverted stream with output heads; **2)** The direct stream attached to the modality encoders of our GASP model. The inverted stream is composed of an SE layer followed by a 2D convolutional layer with a kernel size of 3×3, a padding of 1, and 32 channels. A *max pooling* layer with a window size of 2×2 reduces the feature map dimensions by half. Finally, a 1×1 convolution is applied to the pooled features, reducing the feature maps to a single channel. To emphasize weak features, we invert the input channels:

$$\mathbf{u}_{c'}^{-1} = \log\left(\frac{1}{Softmax(\mathbf{u}_{c'})}\right) = -\log[Softmax(\mathbf{u}_{c'})] \tag{1}$$

where $\mathbf{u}_{c'}$ represents the individual channels of all modalities. The spatially inverted channels $\mathbf{u}_{c'}^{-1}$ are standardized and propagated as input features to the inverted stream. The direct stream is an SE layer with its parameters tied to the inverted stream and receives the standardized FM channels $\mathbf{u}_{c'}$ as input. Finally, the direct stream propagates the channel parameters multiplied with each FM to the modality encoders of GASP. The resulting weighted map is the priority map (PM).

3.2 Modality Encoder (Enc.)

The modality encoder is a convolutional model used for extracting visual features from the priority maps. The first two layers of the encoder have 32 and 64 channels respectively. A maximum pooling layer reduces the input feature map to half its size. The pooled layer is followed by two layers with 128 channels each. Finally, the representations are decoded by applying transposed convolutions with 128, 64, and 32 channels. The last layer has a number of channels equivalent to the input channels. All convolutional kernels have a size of 3×3, with a padding of 1. For GASP model variants operating on single frames (static integration variants), all modalities share the same encoder. For sequential integration variants, each modality has a separate encoder shared across timesteps.

3.3 Recurrent Gated Multimodal Unit (RGMU)

Concatenating the modality representations could lead to successful integration. Such a form of integration is commonly used in multimodal neural models, including audiovisual saliency predictors. We describe such approaches as non-fusion models, whereby the contribution of each modality is unknown. To account for all modalities, we employ the Gated Multimodal Unit (GMU) [Arevalo *et al.*, 2020]. The GMU learns to weigh the input features based on a gating mechanism. To preserve the spatial features of the input, the authors introduce a convolutional variant of the GMU. This model, however, disregards the previous context since it does not integrate features sequentially. Therefore, we extend the convolutional GMU with recurrent units and express it as follows:

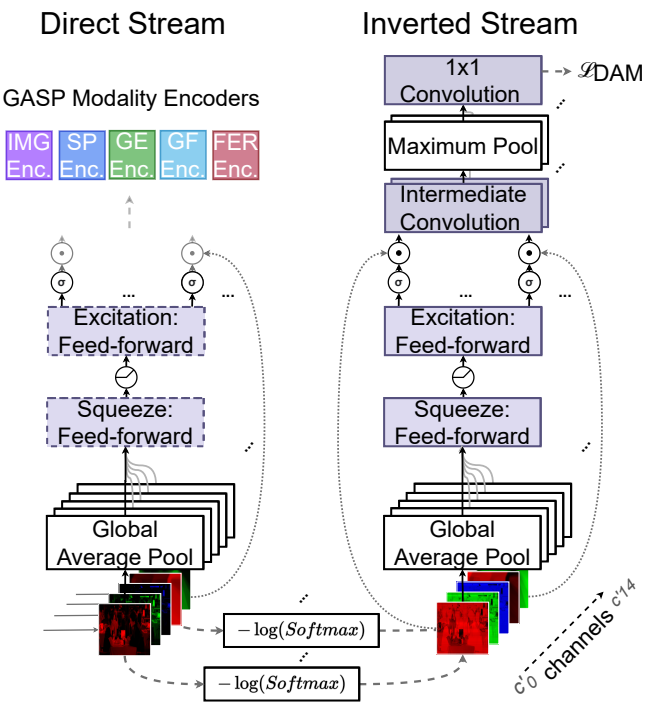

Figure 3: The direct (left) and inverted (right) streams of our Directed Attention Module (DAM). The parameters of the direct stream are frozen and tied to the inverted stream as indicated by the dashed borders.

$$\begin{aligned}
\mathbf{h}_t^{(m)} &= \tanh\left(\mathbf{W}_x^{(m)} * \mathbf{x}_t^{(m)} + \mathbf{U}_h^{(m)} * \mathbf{h}_{t-1}^{(m)} + b_h^{(m)}\right) \\
\mathbf{z}_t^{(m)} &= \sigma(\mathbf{W}_z^{(m)} * [\mathbf{x}_t^{(1)}, .., \mathbf{x}_t^{(M)}] + \mathbf{U}_z^{(m)} * \mathbf{z}_{t-1}^{(m)} + b_z^{(m)}) \\
\mathbf{h}_t &= \textstyle\sum_{m=1}^{M} \mathbf{z}_t^{(m)} \odot \mathbf{h}_t^{(m)}
\end{aligned} \tag{2}$$

where $\mathbf{h}_t^{(m)}$ is the hidden representation of modality m at timestep t. Similarly, $\mathbf{z}_t^{(m)}$ indicates the gated representation. The total number of modalities is represented by M. The parameters of the Recurrent Gated Multimodal Unit (RGMU) are denoted by $\mathbf{W}_x^{(m)}$, $\mathbf{W}_z^{(m)}$, $\mathbf{U}_h^{(m)}$, and $\mathbf{U}_z^{(m)}$. The modality inputs $\mathbf{x}^{(m)}$ at timestep t are concatenated channel-wise as indicated by the $[\cdot,\cdot]$ operator and convolved with $\mathbf{W}_z^{(m)}$. The $\mathbf{z}_t^{(m)}$ representation is acquired by summing the current and previous timestep representations, along with the bias term $b_z^{(m)}$. A sigmoid activation function denoted by σ is applied to the recurrent representations $\mathbf{z}_t$. The final feature map $\mathbf{h}_t$ is the Hadamard-product between $\mathbf{z}_t^{(m)}$ and $\mathbf{h}_t^{(m)}$ summed over all modalities.

The aforementioned recurrent approach suffers from vanishing gradients as the context becomes longer. To remedy this effect, we propose the integration of GMU with the convolutional Attentive Long Short-Term Memory (ALSTM) [Cornia *et al.*, 2018]. ALSTM applies soft-attention to single timestep input features over multiple iterations. We

utilize ALSTM for our static GASP integration variants. For sequential variants, we modify ALSTM to acquire frames at all timesteps instead of attending to a single frame multiple times:

$$\mathbf{x}_t = Softmax(\mathbf{z}_{t-1}) \odot \mathbf{x}_t \tag{3}$$

where $\mathbf{z}_{t-1}$ represents the pre-attentive output of the previous timestep. We adapt the sequential ALSTM to operate in conjunction with the GMU by performing the gated fusion per timestep. We refer to this model as the Attentive Recurrent Gated Multimodal Unit (ARGMU). Alternatively, we perform the gated integration after concatenating the input channels and propagating them to the sequential ALSTM. Since the modality representations are no longer separable, we describe this variant as the Late ARGMU (LARGMU). We refer to the total number of timesteps as the context size. Analogous to the sequential variants, we create similar gating mechanisms for static integration approaches. Replacing the sequential ALSTM with the ALSTM by Cornia *et al.* [2018], we present the non-sequential Attentive Gated Multimodal Unit (AGMU), as well as the Late AGMU (LAGMU).

4 Experimental Setup

We train our GASP model on the social event subset of AVE [Tavakoli *et al.*, 2020]. AVE is a composition of three datasets: DIEM [Mital *et al.*, 2011], Coutrot Databases 1 [Coutrot and Guyader, 2014] and 2 [Coutrot and Guyader, 2015]. To train the model, we employ the loss functions introduced by Tsiami *et al.* [2020], assigning the loss weights $\lambda_1 = 0.1$, $\lambda_2 = 2$, and $\lambda_3 = 1$ to *cross-entropy*, *CC*, and *NSS* losses respectively. The loss functions $\mathcal{L}_{PFDM}$ are weighted, summed, and applied to the final layer for optimizing the modality encoder and integration model parameters. The model is trained using the Adam optimizer, having a learning rate of 0.001, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All models are trained for $\sim$10k iterations with a batch size of 4. We conduct five trials acquiring the mean results for evaluation.

The models are evaluated on the test subset of social event videos in AVE. We employ five commonly used metrics in dynamic saliency prediction [Tavakoli *et al.*, 2020; Tsiami *et al.*, 2020]: *Normalized scanpath saliency* (NSS); *Linear correlation coefficient* (CC); *Similarity metric* (SIM); *Area under the ROC curve* (AUC-J); *Shuffled AUC* (sAUC). The negative fixations for the sAUC metric are sampled from all the mean eye positions in the social event subset of AVE.

The inverted stream of our DAM layer has a separate output head for each timestep. We compute the cross-entropy between the DAM prediction and the FDM. For sequential integration models, the loss is summed over all timesteps. The loss $\mathcal{L}_{DAM}$ with a weight $\lambda_{DAM} = 0.5$ is computed for optimizing the inverted stream parameters. The parameters are transferred to the direct stream with frozen parameters.

An NVIDIA RTX 2080 Ti GPU with 11 GB VRAM and 128 GB RAM is used for training all static and sequential models. To extract spatiotemporal maps in the first stage (SCD), we employ an NVIDIA TITAN RTX GPU with 24 GB VRAM and 64 GB RAM to accommodate all social cue detectors simultaneously. We perform the SCD feature extraction in a preprocessing step for all AVE dataset videos.

Model	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑
Additive	0.5842	0.5912	0.0882	1.19	0.1878
Concatenative	0.8782	0.6303	0.6614	2.71	0.4743
ALSTM	0.6881	0.5727	0.4503	2.05	0.3316
SE	0.5367	0.5597	0.0359	1.03	0.0972
LAGMU *(Ours)*	0.8347	0.6376	0.5576	2.48	0.4361
DAM + LAGMU *(Ours)*	0.8791	0.6379	0.6606	2.76	**0.5278**
GMU	0.8792	0.6374	0.6545	2.75	0.5172
AGMU *(Ours)*	0.6829	0.6359	0.2046	1.47	0.2212
DAM + GMU *(Ours)*	**0.8845**	**0.6397**	**0.6620**	**2.77**	0.5233
DAM + AGMU *(Ours)*	0.8587	0.6372	0.6372	2.71	0.5066

Table 1: Static integration results. Top rows represent non-fusion methods and bottom rows are fusion-based integration approaches.

5 Results

5.1 Static Integration

We examine integration approaches operating on a single frame in GASP. The *Additive* model refers to the integration variant in which the feature maps of all encoders are summed, followed by a 3×3 convolution with 32 channels and a padding of 1. The *Concatenative* variant applies a channel-wise concatenation to the feature maps, followed by the aforementioned convolutional layer. ALSTM, LAGMU, and AGMU employ the non-sequential ALSTM variant by Cornia *et al.* [2018]. The Squeeze-and-Excitation [Hu *et al.*, 2018] (SE) model precedes the modality encoder. We note that all models excluding SE and DAM replace the integration model. Finally, all model variants are followed by a 1×1 convolution resulting in the final output feature map.

In Table 1, we show that the best NSS scores are achieved by our DAM + GMU variant. This indicates fewer false positives predicted by fusion in comparison to other non-fusion methods. We also observe that the DAM enhances fusion models but appears to have minimal effect on the GMU variant (the difference across all metrics is insignificant).

5.2 Sequential Integration

We modify our GASP integration model to have a context greater than one. All models employ batch normalization applied to the temporal axis. The integration models are followed by a 1×1 convolution resulting in the final output feature map. In Table 2, we experiment with context sizes $\in \{2, 4, 6, 8, 10, 12\}$ and observe an overall improvement in performance with a context size of 4. The directed attention variant with late non-fusion gating (DAM + LARGMU) achieves the best scores on all metrics. This implies that gated integration is beneficial, even though the representations preceding the GMU are not separable.

Comparing the results of static integration in Table 1 to dynamic integration approaches in Table 2, we observe that several static approaches perform on a par with recurrent models. Nonetheless, the sequential DAM + LARGMU with context sizes of 8 and 10 outperform all integration methods. In Table 3, we observe an insignificant difference in metric scores among the best sequential models for all context sizes. Compared to the best static model, the variances of sequential

Model	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑
Sequential ALSTM	**0.8849**	0.6428	0.6590	2.79	0.4522	0.8794	0.6439	0.6621	2.78	0.5244	0.8789	0.6425	**0.6612**	**2.76**	0.5291
LARGMU *(Ours)*	0.8791	**0.6444**	0.6572	2.76	0.5251	**0.8860**	**0.6460**	**0.6698**	**2.80**	0.5191	**0.8818**	**0.6433**	0.6556	**2.76**	0.5205
DAM + LARGMU *(Ours)*	0.8789	0.6437	**0.6703**	**2.78**	0.5354	0.8799	0.6443	0.6568	2.76	**0.5287**	0.8801	0.6423	0.6589	**2.76**	**0.5293**
RGMU *(Ours)*	0.8343	0.6239	0.6195	2.62	0.4717	0.8797	0.6350	0.6184	2.69	0.4614	0.7331	0.5851	0.5117	2.33	0.3823
ARGMU *(Ours)*	0.8793	0.6410	0.6607	2.74	**0.5359**	0.8819	0.6456	0.6556	2.75	0.5279	0.8656	0.6388	0.6534	2.72	0.5017
DAM + RGMU *(Ours)*	0.8726	0.6329	0.6547	2.73	0.5135	0.8714	0.6345	0.6539	2.73	0.5198	0.8718	0.6346	0.6510	2.73	0.5172
DAM + ARGMU *(Ours)*	0.8747	0.6421	0.6536	**2.78**	0.5305	0.8790	0.6363	0.6391	2.72	0.4934	0.8560	0.6271	0.6418	2.70	0.5133
	Context Size = 2					Context Size = 4					Context Size = 6				
Model	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑
Sequential ALSTM	0.8766	0.6389	0.6628	2.75	0.5307	0.8773	0.6412	0.6704	2.76	0.5306	0.8759	0.6352	**0.6665**	2.74	0.5275
LARGMU *(Ours)*	0.8702	0.6362	0.6529	2.72	0.5307	0.8791	0.6356	0.6511	2.72	0.5168	**0.8788**	0.6416	0.6624	**2.75**	0.5152
DAM + LARGMU *(Ours)*	**0.8872**	**0.6529**	**0.6903**	**2.84**	**0.5520**	**0.8830**	**0.6527**	**0.6980**	**2.87**	**0.5566**	0.8775	**0.6418**	0.6612	2.74	**0.5328**
RGMU *(Ours)*	0.7982	0.6031	0.5763	2.48	0.4368	0.8229	0.5981	0.5197	2.53	0.4502	0.8147	0.5717	0.4119	2.30	0.3276
ARGMU *(Ours)*	0.6892	0.5680	0.3253	1.84	0.2549	0.8467	0.6179	0.6116	2.62	0.4707	0.8457	0.6130	0.6007	2.56	0.4593
DAM + RGMU *(Ours)*	0.8678	0.6344	0.6622	2.75	0.5212	0.8579	0.6284	0.6540	2.72	0.5207	0.8759	0.6327	0.6549	2.74	0.5123
DAM + ARGMU *(Ours)*	0.8580	0.6259	0.6274	2.66	0.4874	0.8663	0.6303	0.6446	2.69	0.5161	0.8561	0.6249	0.6441	2.70	0.5117
	Context Size = 8					**Context Size = 10**					Context Size = 12				

Table 2: Sequential integration results. Top rows represent non-fusion methods and bottom rows are fusion-based integration approaches.

Model	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑	No. Parameters	Training Time	Context Size
DAM + GMU	0.8845 ± 0.1389	0.6397 ± 0.0831	0.6620 ± 0.2324	2.77 ± 0.53	0.5233 ± 0.1628	0.77M	25 mins	1
DAM + LARGMU	0.8789 ± 0.0420	0.6437 ± 0.0742	0.6703 ± 0.1068	2.78 ± 0.32	0.5354 ± 0.0675	4.28M	185 mins	2
LARGMU	0.8860 ± 0.0396	0.6460 ± 0.0778	0.6698 ± 0.1093	2.80 ± 0.32	0.5191 ± 0.0634	4.28M	130 mins	4
LARGMU	0.8818 ± 0.0400	0.6433 ± 0.0732	0.6556 ± 0.1027	2.76 ± 0.30	0.5205 ± 0.0607	4.28M	140 mins	6
DAM + LARGMU	0.8872 ± 0.0374	0.6529 ± 0.0703	0.6903 ± 0.1167	2.84 ± 0.31	0.5520 ± 0.0729	4.28M	280 mins	8
DAM + LARGMU	0.8830 ± 0.0451	0.6527 ± 0.0785	0.6980 ± 0.1164	2.87 ± 0.34	0.5566 ± 0.0753	4.28M	315 mins	10
DAM + LARGMU	0.8775 ± 0.0430	0.6418 ± 0.0747	0.6612 ± 0.1115	2.74 ± 0.32	0.5328 ± 0.0716	4.28M	330 mins	12

Table 3: Best model results for all context sizes with the standard deviation over five trials provided for each metric, along with the training duration and number of parameters. All models employ audiovisual DAVE (Context Size = 16) as the SCD stage saliency predictor.

GE	GF	FER	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑
-	-	-	0.8767	0.6338	0.6542	2.72	0.5228
-	-	✓	0.7535	0.5951	0.4466	2.17	0.3578
-	✓	-	0.6893	0.5679	0.3222	1.84	0.2539
-	✓	✓	0.8778	0.6442	0.6652	2.76	0.5350
✓	-	-	0.8769	0.6272	0.6493	2.70	0.4798
✓	-	✓	**0.8859**	0.6505	0.6840	2.86	0.5381
✓	✓	-	0.8776	0.6367	0.6543	2.74	0.5216
✓	✓	✓	0.8830	**0.6527**	**0.6980**	**2.87**	**0.5566**

Table 4: Social cue modality ablation applied to our best GASP model (DAM + LARGMU; Context Size = 10).

Model	Test	AUC-J ↑	sAUC ↑	CC ↑	NSS ↑	SIM ↑
UNISAL (Visual only)	AVE	0.8640	0.6545	0.4243	2.04	0.3818
TASED (Visual only)	AVE	0.8601	0.6515	0.4631	2.19	0.4084
DAVE (Visual only)	AVE	0.8824	0.6138	0.5136	2.45	0.4080
DAVE (Audiovisual baseline)	AVE	0.8853	0.6121	0.5453	2.65	0.4420
STAViS (Visual only)	STA	0.8577	0.6517	0.4690	2.08	0.4004
STAViS (Audiovisual)	STA	0.8752	0.6154	0.4912	2.79	0.4774
UNISAL + GASP *(Ours)*	AVE	0.8771	0.6334	0.6494	2.70	0.5244
TASED + GASP *(Ours)*	AVE	0.8602	0.6195	0.5736	2.50	0.4725
DAVE + GASP *(Ours)*	AVE	0.8830	0.6527	**0.6980**	2.87	**0.5566**
STAViS + GASP *(Ours)*	STA	**0.8910**	**0.6825**	0.6052	**3.08**	0.4324

Table 5: Comparison with state-of-the-art by varying the SCD SP of our best GASP model (DAM + LARGMU; Context Size = 10).

model scores are lower, indicating the stabilizing influence of attentive LSTMs with the addition of context.

5.3 Modality Contribution

We evaluate the contribution of each modality to the final prediction by computing the mean activation of the gates across channels and timesteps. This evaluation is only applicable to fusion methods in either static or sequential forms of integration. We observe that the DAM does not alter the modality contribution of the static GMU. For sequential variants, introducing the DAM allows modalities to have a uniform contribution to the final output.

We examine the modalities contributing an improvement to the best non-fusion sequential model. As shown in Table 4, FER in combination with GE achieves results on par with the best model. The exclusion of GF has a minimal effect on the model due to the sparsity of its representation. Significant degradation in the model variant with social modalities ablated implies the necessity of social cues in concert.

5.4 Comparison with State-of-the-art

We compare the performance of our model with four dynamic saliency predictors. We replace DAVE [Tavakoli *et al.*, 2020] with STAViS [Tsiami *et al.*, 2020], TASED [Min and Corso, 2019] and UNISAL [Droste *et al.*, 2020] in the SCD stage during the evaluation phase. Due to the overlap in datasets between DAVE and STAViS, we retrain our GASP model with STAViS as the SCD audiovisual saliency predictor. We evaluate and train our STAViS-based model on social event videos according to the data splits concocted by Tsiami *et al.* [2020] to avoid data leakage. The average scores are computed for all split test sets over five trials.

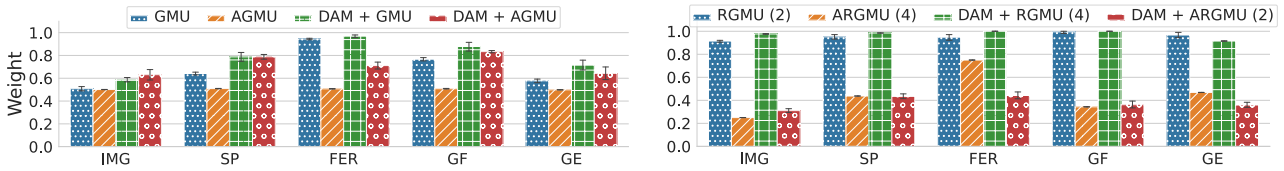

Figure 4: Aggregated modality weights of static (left) and sequential (right) fusion methods. Context sizes are shown within parentheses.

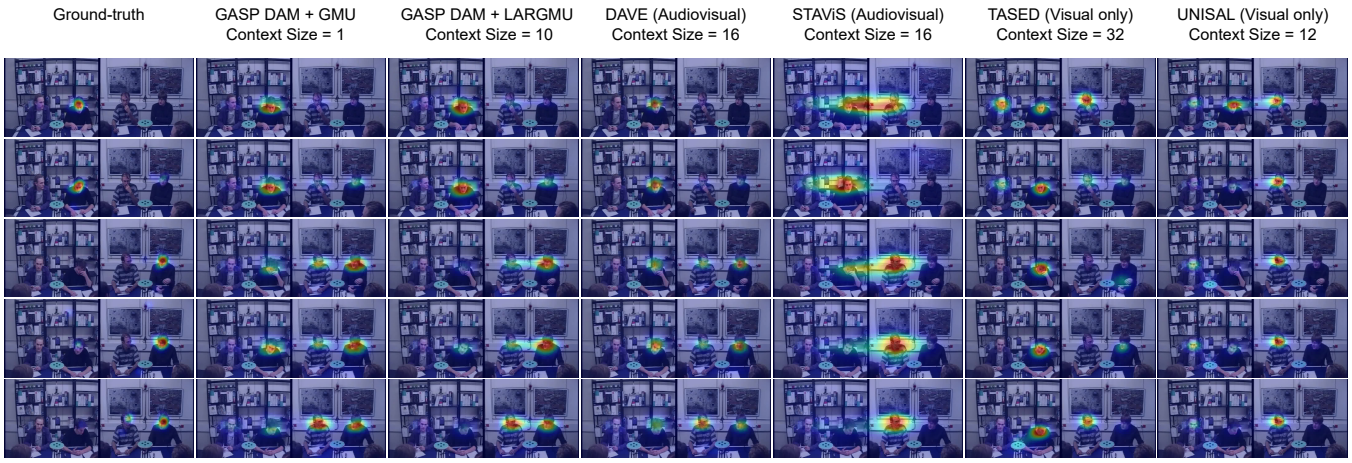

Figure 5: State-of-the-art model comparison on a Coutrot Database 2 [Coutrot and Guyader, 2015] sample. Our sequential (Context Size = 10) and static (Context Size = 1) GASP models employ the audiovisual DAVE (Context Size = 16) as the SCD stage saliency predictor.

Combining our best GASP model with different saliency predictors improves their performances, as shown in Table 5. Although GASP is not retrained, it extracts information from the untuned saliency predictors pertinent to the prediction. The sequential GASP also exhibits greater resistance to central bias as shown in Figure 5 (middle row) compared to other models, where the actor closest to the center is incorrectly predicted as a fixation target. The social cue integration and sequential learning in GASP contribute to such resistance.

6 Conclusion

We introduce GASP, a gated attention model for predicting saliency by integrating knowledge from two social cues using three separate detectors. We represent each social cue as a spatiotemporal map, accounting for the gaze direction, gaze target, and facial expressions of an observed individual. We examine gated and recurrent approaches for integrating the social cues. Our gated integration variants achieve better results than non-gated approaches. We also present a module for emphasizing weak features, which is effective for static and sequential integration methods. Furthermore, we show that gaze direction and facial expression representations have a positive effect when integrated with saliency models. The latter supports the importance of considering affect-biased attention. GASP improves state-of-the-art saliency model prediction performances on multiple metrics, indicating the efficacy of social cue integration in our architecture.

Acknowledgements

The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169).

References

[Alpay *et al.*, 2019] Tayfun Alpay, Fares Abawi, and Stefan Wermter. Preserving activations in recurrent neural networks based on surprisal. *Neurocomputing*, 342:75–82, 2019.

[Arevalo *et al.*, 2020] John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A González. Gated multimodal networks. *Neural Computing and Applications*, pages 10209–10228, 2020.

[Bak *et al.*, 2017] Cagdas Bak, Aysun Kocak, Erkut Erdem, and Aykut Erdem. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia*, 20(7):1688–1698, 2017.

[Birmingham *et al.*, 2009] Elina Birmingham, Walter F Bischof, and Alan Kingstone. Saliency does not account for fixations to eyes within social scenes. *Vision Research*, 49(24):2992–3000, 2009.

[Borji, 2019] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):679–700, 2019.

[Bruce and Tsotsos, 2009] Neil DB Bruce and John K Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):5, 2009.

[Bylinskii *et al.*, 2016] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where should saliency models look next? In *European Conference on Computer Vision (ECCV)*, pages 809–824. Springer, 2016.

[Cordel *et al.*, 2019] Macario O Cordel, Shaojing Fan, Zhiqi Shen, and Mohan S Kankanhalli. Emotion-aware human attention prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4026–4035. IEEE, 2019.

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

[Coutrot and Guyader, 2014] Antoine Coutrot and Nathalie Guyader. An audiovisual attention model for natural conversation scenes. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 1100–1104. IEEE, 2014.

[Coutrot and Guyader, 2015] Antoine Coutrot and Nathalie Guyader. An efficient audiovisual saliency model to predict eye positions when looking at conversations. In *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, pages 1531–1535. IEEE, 2015.

[Droste *et al.*, 2020] Richard Droste, Jianbo Jiao, and Alison J Noble. Unified image and video saliency modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[Fan *et al.*, 2018] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Juan Xu, Mohan S Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7521–7531. IEEE, 2018.

[Foulsham *et al.*, 2011] Tom Foulsham, Jason JS Barton, Alan Kingstone, Richard Dewhurst, and Geoffrey Underwood. Modeling eye movements in visual agnosia with a saliency map approach: Bottom–up guidance or top–down strategy? *Neural Networks*, 24(6):665–677, 2011.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141. IEEE, 2018.

[Itti and Koch, 2001] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[Kellnhofer *et al.*, 2019] Petr Kellnhofer, Adrià Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6912–6921. IEEE, 2019.

[Min and Corso, 2019] Kyle Min and Jason J Corso. TASED-Net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2394–2403. IEEE, 2019.

[Min *et al.*, 2020] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29:3805–3819, 2020.

[Mital *et al.*, 2011] Parag K Mital, Tim J Smith, Robin L Hill, and John M Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011.

[Pritsch *et al.*, 2017] Carla Pritsch, Silke Telkemeyer, Cordelia Mühlenbeck, and Katja Liebal. Perception of facial expressions reveals selective affect-biased attention in humans and orangutans. *Scientific Reports*, 7(1):1–12, 2017.

[Recasens *et al.*, 2017] Adrià Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1435–1443. IEEE, 2017.

[Salley and Colombo, 2016] Brenda Salley and John Colombo. Conceptualizing social attention in developmental research. *Social Development*, 25(4):687–703, 2016.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. GRAD-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2017.

[Siqueira *et al.*, 2020] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 5800–5809. AAAI, 2020.

[Tavakoli *et al.*, 2020] Hamed R Tavakoli, Ali Borji, Juho Kannala, and Esa Rahtu. Deep audio-visual saliency: Baseline model and data. In *Symposium on Eye Tracking Research and Applications*, pages 1–5. ACM, 2020.

[Tsiami *et al.*, 2020] Antigoni Tsiami, Petros Koutras, and Petros Maragos. STAViS: Spatio-temporal audiovisual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4766–4776. IEEE, 2020.

[Zhang *et al.*, 2017] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S^3FD: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 192–201. IEEE, 2017.