BIAS: A Body-based Interpretable Active Speaker Approach

Tiago Roxo, Joana C. Costa, Pedro R. M. Inácio, *Senior Member, IEEE,* Hugo Proença, *Senior Member, IEEE*
Instituto de Telecomunicações, University of Beira Interior, Portugal
{tiago.roxo, joana.cabral.costa}@ubi.pt, {inacio, hugomcp}@di.ubi.pt

***Abstract*—State-of-the-art Active Speaker Detection (ASD) approaches heavily rely on audio and facial features to perform, which is not a sustainable approach in *wild* scenarios. Although these methods achieve good results in the *standard* AVA-ActiveSpeaker set, a recent wilder ASD dataset (WASD) showed the limitations of such models and raised the need for new approaches. As such, we propose BIAS, a model that, for the first time, combines audio, face, and body information, to accurately predict active speakers in varying/challenging conditions. Additionally, we design BIAS to provide interpretability by proposing a novel use for Squeeze-and-Excitation blocks, namely in attention heatmaps creation and feature importance assessment. For a full interpretability setup, we annotate an ASD-related actions dataset (ASD-Text) to finetune a ViT-GPT2 for text scene description to complement BIAS interpretability. The results show that BIAS is state-of-the-art in challenging conditions where body-based features are of utmost importance (Columbia, open-settings, and WASD), and yields competitive results in AVA-ActiveSpeaker, where face is more influential than body for ASD. BIAS interpretability also shows the features/aspects more relevant towards ASD prediction in varying settings, making it a strong baseline for further developments in interpretable ASD models, and is available at https://github.com/Tiago-Roxo/BIAS.**

***Index Terms*—Active speaker detection, body-based analysis, interpretability, text descriptions, visual surveillance.**

I. Introduction

CURRENT Active Speaker Detection (ASD) models are known to perform reliably using only audio and face-based information. This is mainly due to the fact that state-of-the-art ASD datasets have good audio and face quality, yielding from controlled setups (movies, from AVA-ActiveSpeaker [1]) and cooperative settings (interviews, from AWS [2]). Recently, WASD [3] has been announced as a more challenging set, with degraded audio and face data quality, corresponding to less constrained data acquisition scenarios.

In this context, the existing models are not suitable for wilder settings, where audio quality might be poor and faces occluded (Figure 1). As such, we propose BIAS, an approach for ASD, that uses, for the first time, body data to complement face and audio-based features, achieving state-of-the-art results in challenging sets (WASD and open-settings of Columbia), and competitive results in more cooperative conditions (AVA-ActiveSpeaker), where the body relevance is reduced since the face is the predominant feature in this data. Furthermore, we propose a novel use of Squeeze-and-Excitation (SE) blocks [4] to provide reasoning for model decision and analyze the importance of different features. This way, we are able to obtain visual interpretability and assess feature influence in varying ASD settings, via SE vector manipulation. Finally, to improve the interpretability of ASD-related scenarios, we complement the attention of BIAS (visual) with descriptions of existing actions, via caption generation (textual). Given the absence of training data of subtle actions related to ASD tasks (*e.g.*, raise hand, cross arms, open mouth to talk), we

Fig. 1. Illustration of the **BIAS insight**: in surveillance settings, where facial and audio-based features might not be always available, body data should be crucial to accurately detect the active speakers. In such challenging conditions, providing reliable explanations for the reasoning behind the provided responses is also an important feature. This paper describes BIAS, which singularly uses facial, audio, and body-based features, also providing visual interpretability and feature importance assessment for its responses.

Manuscript received XX, 2023; revised XX, 2023.

use WASD data to annotate an ASD-related actions dataset (ASD-Text) and finetune a Generative Pretrained Transformer (GPT) model on it which, in conjunction with the visual interpretability of BIAS, can be used for a full interpretability setup. To summarize, the main contributions are:

- We propose BIAS, an ASD model based on a long temporal context approach, using audio, face, and body-based information. With respect to the state-of-the-art, BIAS is innovative by being the first method to use body-based information for ASD, yielding competitive results in AVA-ActiveSpeaker and is state-of-the-art in WASD and Columbia (open settings version);
- We propose a novel use of SE blocks for attention heatmap creation (visual interpretability) and perceive feature importance in the decisions, which can be easily included in common and customized architecture, for image and video settings;
- To improve the interpretability of ASD-related scenarios, we annotate an ASD-related actions dataset (ASD-Text) for text scene captions, providing a training setup for fine scene description in the ASD context.

II. Related Work

ASD Context. Given the task to determine the talking speaker from a set of admissible candidates, various state-of-the-art ASD datasets have been recently proposed [1], [2], [5]–[7]. Columbia [5] contains 87 minutes of a panel discussion and Talkies [6] focuses on low duration videos, totalling 4 hours, with off-screen speaking. Easycom [7] is designed for multiple augmented reality tasks, composed of various speaking sessions with background noise. AVA-ActiveSpeaker [1] contains Hollywood videos totalling almost 38 hours, with demographic diversity and Frames Per Second (FPS) variation, and has application in other topics such as audio anomaly detection [8]. ASW [2] has over 30 hours, with videos randomly selected from the VoxConverse [9], containing various sets of interviews. Recently, WASD [3] has been announced, with 30 hours of data grouped based on the audio and face quality, with balanced demographic diversity and body annotations data. For a broader overview of the ASD context, Robi *et al.* [10] review the main ASD modalities, applications and challenges.

ASD Models. Based on the available data for ASD, current state-of-the-art heavily rely on face and audio data, with audiovisual combination using 3D architectures [11], hybrid 2D-3D models [12], large-scale pretraining [13], [14], feature embedding improvement [15], and attention mechanisms [16]–[18]. Various ASD works [6], [19]–[21] based their model on a two-step process, where the first focuses on short-term analysis (audio with face combination) and the second on multi-speaker analysis. ASC [19] focuses on multi speaker analysis via temporal refinement, ASDNet [20] uses a similar approach for inter-speaker relations with improved visual backbones, and UniCon [21] relies on audio-visual relational contexts with various backbones. The improvement of ASD performance by assessing contextual information via speaker relation using Graph Convolutional Networks (GCN) [22] has also been explored [6], [23], [24]. Diverging from two-step training, end-to-end models have also emerged for ASD [23]–[26]. TalkNet [25] focused on improving long-term temporal context with audio-visual synchronization, EASEE [24] included GCN to complement spatial and temporal speaker relations, and Light-ASD [26] proposed a lightweight model by splitting 2D and 3D convolutions for audio-visual feature extraction, and applied Bidirectional Gated Recurrent Units (BGRU) for cross-modal modeling.

Body Information for Attribute Recognition. Although recent works on ASD do not use body information, this data contains information that could contribute to improve model performance, particularly in wilder conditions (*e.g.*, surveillance settings), where face is not reliably accessed. Pedestrian Attribute Recognition (PAR) datasets [27]–[29] are examples of these scenarios, containing person cropped images from surveillance settings, used to identify attributes (*e.g.*, clothing, accessory usage, gender, age) under challenging covariates such as occlusion, pose, image resolution, and luminosity. Works in this area focused on different strategies ranging from different architecture combination [30]–[32], attention-based approaches [33], [34], and attribute relation importance [35]–[37].

Model Interpretability. For visual interpretability, we can group methods into two main categories: *gradient based* [38]–[42] (gradients of each layer, computed through backpropagation) and *attribution propagation* [43]–[47] (recursive decomposition of layers contributions, all the way to model's input). Saliency based methods [48], [49], Excitation Backprop [50], and Perturbation methods [51], [52] are also visual interpretable approaches in computer vision, with Transformer-based interpretability [53] being recently explored. Although most works explore model interpretability in object classification datasets, its use in face [54], [55], body [56], and PAR [57], [58] data is not unprecedented. Contrary to current approaches, we propose a SE block-based visual interpretation, obtainable in inference time, without requiring additional computational cost for attention heatmaps creation.

III. BIAS Approach

We propose BIAS, an interpretable model for ASD, based on a long temporal context approach, using audio, face, and body data. We process data with modified encoders and customized feature vector combination, using SE blocks to provide visual interpretability and feature importance assessment, respectively. BIAS is distinctive from other state-of-the-art models by using, for the first time, body data in ASD tasks, translating into state-of-the-art performance in challenging sets and competitive results in cooperative conditions. The overall architecture is displayed in Figure 2, with details of each part in the following subsections. Note that, although body data also contain facial cues, these are not as easily perceived as body movements (given the face/body proportion), which creates a stronger focus on hands and arms movements.

A. Visual and Audio Encoders

Backbone. We obtain a sequence of frame-based embedding using a customized backbone, based on the ResNet18

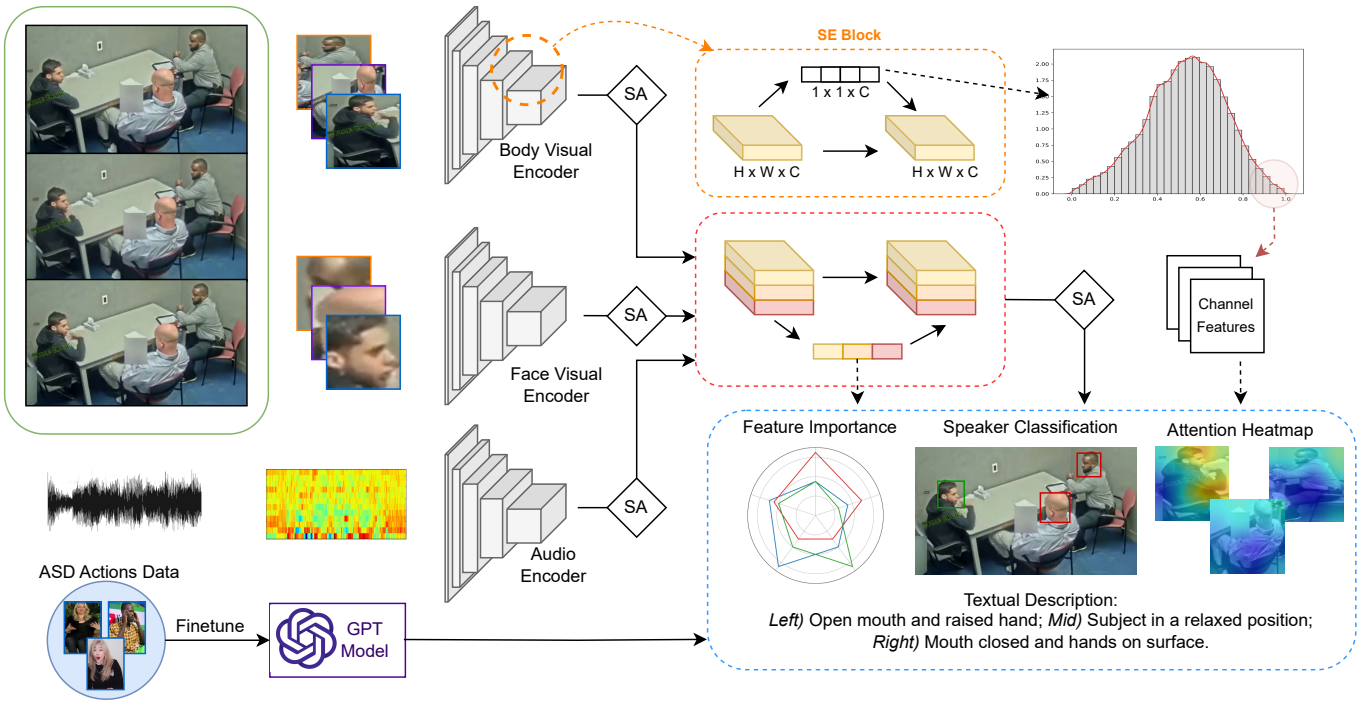


Fig. 2. Overview of the BIAS architecture and pipeline, with GPT model integration: body and face-based data is fed into the respective visual encoders, while audio is processed into MFCC before encoding. SE blocks are used in visual encoders and feature combination for attention heatmaps and feature relative importance, respectively. SA refers to self-attention blocks. Heatmaps are created by combining channel features of the respective top 10% SE vector values. BIAS prediction is based on feature combination, accompanied with visual interpretability and feature importance assessment, complemented by text descriptions from a GPT model finetuned in ASD-related actions data (ASD-Text).

architecture, composed of a 3D convolutional layer followed by ResNet18 [59] layers and a SE block, before average pooling. The inclusion of a SE block is to provide a way to retrieve informative features, via assessment of SE vector values, useful for visual interpretability, which we discuss in Section III-B.

Temporal Enconding. For long-term visual spatio-temporal representation, we use a visual temporal network, composed of 5 depth-wise separable convolutional layers (DS Conv1D [60]) with residual connections, followed by a Conv1D layer for feature dimension reduction (to 128-feature embedding).

Audio Encoder. We obtain audio temporal encoding using a ResNet34 with SE blocks in its layers (SE-ResNet34). Audio frames are represented as 13-dimensional Mel-Frequency Cepstral Coefficients (MFCCs), with the audio encoder outputting them as a 128-dimensional audio embedding for subsequent visual and audio conjunction.

B. Squeeze-and-Excitation for Interpretability

The novel use of SE block in BIAS is included in two parts of its architecture: end of visual backbones (visual interpretability) and feature combination (perceived feature importance).

Visual Interpretability. In our model, we use SE in encoders to obtain the features (channels) perceived as highly influential towards ASD (*i.e.*, high value in SE vector). However, the perception of *high* is relative to the video/categories used, so we can not use a default/hard-coded value to define what is high or low importance. As such, given that the values of the SE vector follow a Normal distribution (Figure 4), we use its formula to obtain the top 10% values of the SE vector (and, consequently, the top 10% channels) to use for ASD evaluation. The reason for considering the top 10% in our experiments is to balance model performance and visual interpretability (higher % leads to worse performance, and lower % leads to less clear visual interpretability). Finally, we conjugate the selected channels by bicubic interpolation on top of the original image to provide a heatmap of the most important regions for ASD.

Perceived Importance and Feature Influence. Another use of SE in BIAS is to combine audio, face, and body information (embeddings of 128 dimensions). For feature importance assessment, we create three different graphs that display the importance of each feature (audio, face, and body) for the categories of WASD (Table II). For each feature, we normalize the SE values for visualization purposes to better display the feature importance on WASD categories. The normalization is done based on the values obtained by the model for a video and it is not included in the BIAS training process, meaning that BIAS does not rely on additional/external information and can be used for other domains without requiring adaptations. Regarding feature influence in model performance, when the experiments do not contain a certain feature it means that we set the SE vector values to 0 for the channels of the considered feature (*e.g.*, when we consider *only face*, we set the values to 0 for audio and body channels).

TABLE I
ASD-TEXT DATASET ACTION LABELS, GROUPED BY BODY PARTS.

Body Part	#	Action Label
Hand	1	Hand raised in the air
	2	Hand touching the face
	3	Hand raised with object
	4	Hand movement (not raised)
Mouth	5	Mouth occlusion with object
	6	Mouth occlusion with hand
	7	Mouth move from speech
	8	Mouth move from expression
	9	Mouth not moving
Arms	10	Crossed arms
	11	Arms behind back
Body	12	Body in relaxed position
	13	Body facing forward
	14	Wild body movement

TABLE II
CATEGORY FEATURE MATRIX. FEATURE DESCRIPTION: FA, FACE AVAILABILITY; SO, SPEECH OVERLAP; DS, DELAYED SPEECH; FO, FACIAL OCCLUSION; HV, HUMAN VOICE AS BACKGROUND NOISE; SS, SURVEILLANCE SETTINGS. THE ABSENCE OF A CERTAIN FEATURE IS PRESENTED WITH ×, WHILE ITS PRESENCE WITH ✓. FEATURES CONTAINING ? REFER TO NON-GUARANTEE OF ITS PRESENCE OR ABSENCE. GREEN CELLS REFER TO FEATURES FAVORABLE FOR ASD, WHILE RED ONES ARE UNFAVORABLE. RETRIEVED FROM [3].

Category	FA	SO	DS	FO	HV	SS
Optimal Conditions (OC)	✓	×	×	×	×	×
Speech Impairment (SI)	✓	✓	✓	×	×	×
Face Occlusion (FO)	✓	×	×	✓	×	×
Human Voice Noise (HVN)	✓	×	×	×	✓	×
Surveillance Settings (SS)	?	?	?	?	?	✓

C. Self-attention and Loss Function

Self-attention. To improve feature importance assessment, we use an attention layer based on query, key, value strategy [16], after visual encoders and feature conjugation. Initially used in the Transformer architecture, this approach has been extensively used for attention purposes, with a role in BIAS of improving audio-visual correlation. This translates into more accurate distinction between speaking and non-speaking frames, given each speaker.

Loss Function. The main source for ASD is the conjunction of 3 data sources: audio, face, and body. To improve independent feature relevance for ASD, and to not motivate ASD prediction solely based on their conjunction, we include weighted cross entropy losses using individual data on ASD training prediction. For inference, only the conjunction of 3 data sources is considered.

D. ASD-Text Dataset

For text scene description, we finetune a ViT-GPT2 [61], [62] model on WASD data, with ViT trained in ImageNet [63] and GPT2 in WebText [62]. Before annotations, we define a set of 14 admissible ASD-related actions, in Table I. Then, for each action we create 3 admissible captions for each gender to ease data annotation, and construct a Graphical User Interface (GUI) to annotate. For each subject image, the annotators select the gender and all existing actions on it, using the GUI to create annotations with predefined admissible captions from the selected actions and gender. The total annotated data is composed of 47 246 captions, and 11 733 images from WASD subvideos (up to 30s), randomly selected from all videos of the WASD train set. Similar to the COCO captions dataset [64], we divide our annotations into 90/10 train/test [65]. The caption prediction of ViT-GPT2 model in Section V-C is done on images from the test set of WASD.

E. Implementation Details

BIAS is trained with an Adam optimizer, with a initial learning rate of 10^{-4}, decreasing 5% for each epoch. All visual data is reshaped into 112 x 112, audio data is represented by 13-dimensional MFCC, and both visual and audio features have an encoding dimension of 128. Self-attention uses a transformer layer with 8 attention heads. For visual augmentation, we perform random flip, rotate and crop, while for audio augmentation, we use negative audio sampling [25]. In sum, given a video data during training, a audio track of a new one is randomly selected from the same batch as noise, maintaining the same speaking label of the original soundtrack. Since AVA-ActiveSpeaker does not have body data annotations, we obtain body bounding box annotations from AVA Actions Dataset [66] and complement them with speaking labels of AVA-ActiveSpeaker. ViT-GP2 model is finetuned for 3 epochs, using AdamW optimizer, with a learning rate of 5×10^{-5}, without weight decay or warmup steps.

IV. EXPERIMENTS

A. Datasets, Models, and Evaluation Metrics

Datasets. The AVA-ActiveSpeaker dataset [1] is an audio-visual active speaker dataset from Hollywood movies. With 262 15 minute videos, typically only train and validation sets are used for experiments: 120 for training, and 33 for validation, corresponding to 29,723 and 8,015 video utterances, respectively, ranging from 1 to 10 seconds. The main challenges of this dataset are related to language diversity, FPS variation, the existence of faces with low pixel numbers, blurry images, noisy audio, and dubbed dialogues. Similar to other works, we report the obtained results on the AVA-ActiveSpeaker validation subset.

The WASD dataset [3] compiles a set of videos from real interactions with varying accessibility of the two components for ASD: *audio* and *face*. With 30 hours of labelled data, WASD is divided into 5 categories with varying degrees of audio and face quality, grouped into categories: Optimal Conditions (OC), Speech Impairment (SI), Face Occlusion (FO), Human Voice Noise (HVN), and Surveillance Settings (SS). Table II presents the main characteristics of WASD categories. WASD contains 164 videos, with varying FPS, averaging 28 second duration, with balanced demographics, and similar train/test division as AVA-ActiveSpeaker (80/20). We report the results on each category and Easy-Hard grouping, following WASD experiments (Easy: OC and SI, Hard: FO, HVN, and SS).

TABLE III
BIAS AND STATE-OF-THE-ART MODELS PERFORMANCE ON THE DIFFERENT CATEGORIES OF WASD, GROUPED BY EASY AND HARD, USING THE MEAN AVERAGE PRECISION (MAP) METRIC.

Model	Easy		Hard		
	OC	SI	FO	HVN	SS
ASC [19]	91.2	92.3	87.1	66.8	72.2
MAAS [6]	90.7	92.6	87.0	67.0	76.5
ASDNet [20]	96.5	97.4	92.1	77.4	77.8
TalkNet [25]	95.8	97.5	93.1	81.4	77.5
Light-ASD [26]	97.8	98.3	95.4	84.7	77.9
BIAS	**97.8**	**98.4**	**95.9**	**85.6**	**82.5**

TABLE IV
COMPARISON OF BIAS AND STATE-OF-THE-ART MODELS ON THE AVA-ACTIVESPEAKER, GROUPED BY THE VISUAL ENCODER USED. MODELS WITH ∗ CUSTOMIZED THE REPORTED BACKBONES.

Model	Visual Encoder	Par(M)	Body Data	mAP
ASC [19]	RN18 2D	23.3	×	87.1
MAAS [6]	RN18 2D	21.7	×	88.8
TalkNet [25]	RN18* 2D-3D	15.0	×	92.3
BIAS	RN18* 2D-3D	31.6	✓	**92.4**
ASDNet [20]	RNx101 3D	49.7	×	93.5
EASEE-50 [67]	RN50 3D	74.7	×	94.1
Light-ASD [26]	Conv 2D-1D	1.0	×	94.1

We also consider Columbia [5] following the methodology of Light-ASD [26] where models are trained in AVA-ActiveSpeaker, without any additional fine-tuning. Columbia consists of an 87-minute panel discussion video, with five speakers (Bell, Boll, Lieb, Long, and Sick) taking turns speaking, with 2-3 speakers visible at any given time.

Models. The considered models are the ones with state-of-the-art results and publicly available implementations: ASC [19], MAAS [6], TalkNet [25], ASDNet [20], and Light-ASD [26]. ASC, MAAS, and ASDNet are trained in a two-step process, while TalkNet and Light-ASD are trained end-to-end. MAAS did not provide its Multi-modal Graph Network setup so we present the results from the available implementation.

Evaluation Metrics. For AVA-ActiveSpeaker and WASD, we use the official ActivityNet evaluation tool [1] that computes mean Average Precision (mAP), while for Columbia we use F1 score. Following the Microsoft COCO Image Captioning Challenge approach, caption generation is evaluated by ROUGE-L [68], METEOR [69], and BLEU-1 to 4 [70].

B. BIAS Performance in WASD

To assess the importance of body information for ASD we compare BIAS with the reported results of state-of-the-art models in WASD [3], divided by categories, in Table III.

Similar Performance in Easy. The inclusion of body information in BIAS culminates in state-of-the-art results across all categories, obtaining slightly better results in Easy setups. In these scenarios, the reliability of face access and sound quality, with minor degradation (OC and SI), is enough to warrant a good performance from state-of-the-art models. As such, the complement of body information from BIAS does not translate into a significant improvement over state-of-the-art models, given the cooperative settings of the Easy group for ASD.

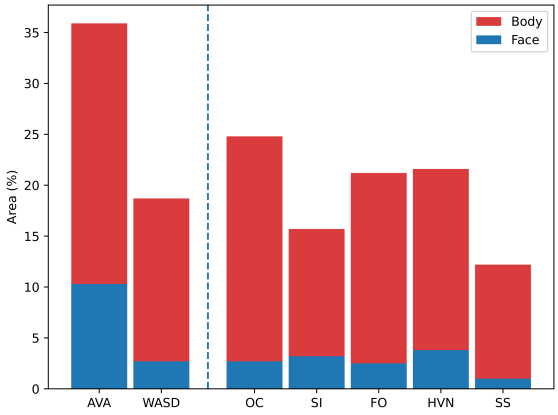


Fig. 3. Comparison of face and body area, relative to image dimension, in percentage. AVA-ActiveSpeaker contains data with subjects closer to the camera, expressed by higher face and body percentage, relative to WASD and any of its categories. Surveillance Settings (SS) is the category with further distance of subjects from camera.

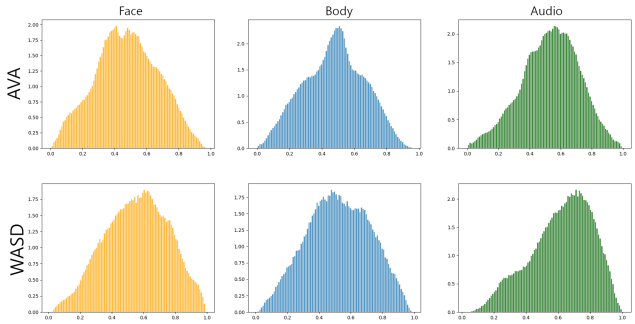


Fig. 4. SE vector values from feature (audio, body, and face) combination, for BIAS trained in AVA-ActiveSpeaker and WASD. Both datasets follow a normal distribution. AVA refers to AVA-ActiveSpeaker.

Improvement in Hard. The major difference of BIAS relative to other models is in Hard categories, with degraded audio and face image quality. In particular, audio degraded categories (HVN and SS) are the ones where BIAS obtains the biggest improvement over state-of-the-art, which is linked to body information access from BIAS, diminishing the dependence of audio cues for ASD. Regarding scenarios without reliable access to mouth movement (FO), conjunction of body cues with audio information is a more reliable approach for ASD, translated by the increased performance.

C. BIAS Performance in Other Datasets

Although we show that inclusion of body information is important for ASD in WASD, in particular for categories with more degraded audio and face data, we also assess its importance in other setups using AVA-ActiveSpeaker and Columbia in Tables IV and V, respectively.

BIAS in AVA-ActiveSpeaker. We compare BIAS with the other state-of-the-art models, by grouping them into sets, based on the backbone used. The results show that body information inclusion contributes to state-of-the-art performance

TABLE V
COMPARISON OF F1-SCORE (%) ON THE COLUMBIA DATASET.

Model	Speaker					
	Bell	Boll	Lieb	Long	Sick	Avg
TalkNet [25]	43.6	66.6	68.7	43.8	58.1	56.2
LoCoNet [71]	54.0	49.1	80.2	80.4	76.8	68.1
Light-ASD [26]	82.7	**75.7**	87.0	74.5	85.4	81.1
BIAS	**89.3**	75.4	**92.1**	**88.8**	**88.6**	**86.8**

TABLE VI
ABLATION STUDIES ON THE EFFECT OF SE FOR FEATURE COMBINATION AND THE CONTRIBUTION OF DIFFERENT FEATURES TOWARDS BIAS PERFORMANCE (MAP) IN AVA-ACTIVESPEAKER (AVA) AND WASD.

Settings				AVA	WASD
SE	Audio	Face	Body		
✓	×	×	✓	62.7	84.6
✓	×	✓	×	79.9	90.2
✓	✓	×	✓	75.3	86.2
✓	×	✓	✓	81.2	92.1
✓	✓	✓	×	91.9	92.2
×	✓	✓	✓	91.3	91.8
✓	✓	✓	✓	92.4	94.1

in ResNet18 grouping, but it is slightly worse than superior or custom backbones, which suggests that body data can be used but it is not as relevant in scenarios where face access and audio quality are more reliable [3]. Figure 3 shows the face/body area proportion in AVA-ActiveSpeaker and WASD, grouped by categories, corroborating the reduced importance of body in AVA-ActiveSpeaker data. In this context, the proximity to camera and reliable face access makes body information less contributing to improve ASD, since most of the information usable already exists in the face data. As such, BIAS approach to include body does not translate to a state-of-the-art performance in such settings.

Robustness of BIAS in Columbia. We also assess the performance of BIAS in Columbia, following the methodology of Light-ASD [26] where models are trained in AVA-ActiveSpeaker, without any additional fine-tuning, and compare with the results reported on Light-ASD, in Table V. In this more challenging setting, BIAS approach to combine body with face and audio information leads to a state-of-the-art performance and highlights its robustness to perform ASD in cross-domain settings. Given the results, BIAS is a resilient state-of-the-art ASD model, applicable in scenarios with varying data quality.

D. Ablation Studies

SE Feature Combination. To complement ASD feature importance for BIAS, we also assess the effect of SE for feature combination and using different features towards BIAS performance in AVA-ActiveSpeaker and WASD, in Table VI. The results show that: 1) When using only one visual feature, face is more relevant than body for both datasets; 2) For two feature combination, audio with face is the approach with better results, particularly for AVA; and 3) The aggregated feature combination is better for both datasets, with increased performance in WASD due to the importance of body information in its most challenging categories; and 4) Regarding the effect of SE, its inclusion translates into improved results for both datasets, with a bigger improvement on WASD, which is linked to its more challenging data where adequate feature selection is of utmost importance (*i.e.*, careful selection of face *vs.* body features is more relevant when face may not be reliably accessed).

TABLE VII
BIAS PERFORMANCE USING DIFFERENT BACKBONES IN AVA-ACTIVESPEAKER (AVA), WHILE MAINTAINING THE REMAINING OF THE ARCHITECTURE.

Visual Backbone	Par(M)	AVA
ResNet18	31.6	92.4
ResNet50	55.1	92.2
ResNet101	91.3	92.4

TABLE VIII
BIAS PERFORMANCE (MAP) IN AVA-ACTIVESPEAKER (AVA) AND WASD, WITH VARYING BODY INPUTS AND AS THE ONLY MODEL INPUT FOR ASD.

Model Input	AVA	WASD
Body w/ Face Region	62.7	84.6
Body w/o Face Region	44.8	76.0

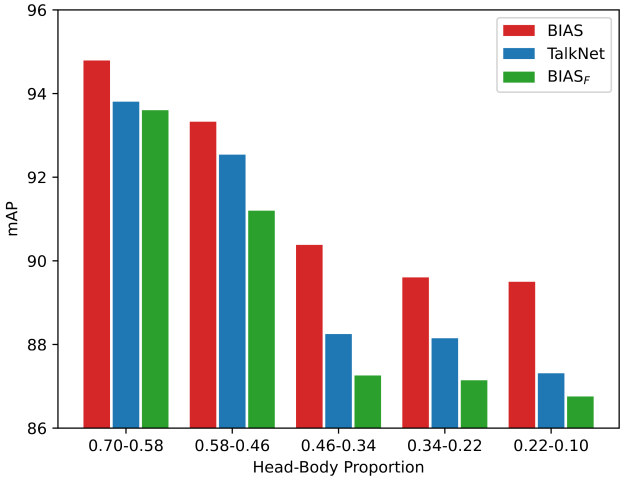


Fig. 5. Performance of BIAS, TalkNet, and BIAS$_F$ relative to Head-Body Proportion (HBP) in WASD, across 5 equidistant intervals based on minimum (0.1) and maximum (0.7) HBP. BIAS$_F$ refers to BIAS with only face as visual input.

Backbone Variance. We explore the influence of having bigger backbones for visual feature extraction in Table VII, which shows varying backbones do not influence BIAS performance. With ASDNet and EASEE, we see that their higher extraction power is also accompanied with additional computationally heavy components to assess actors' relation (*e.g.*, GCN), and customized lightweight models [26] also achieve state-of-the-art performance which confirms that simply having higher extraction power does not necessarily translate into better performance. This also occurs in BIAS since the combination of face and body features is not directly influenced by varying the backbone given that the combination

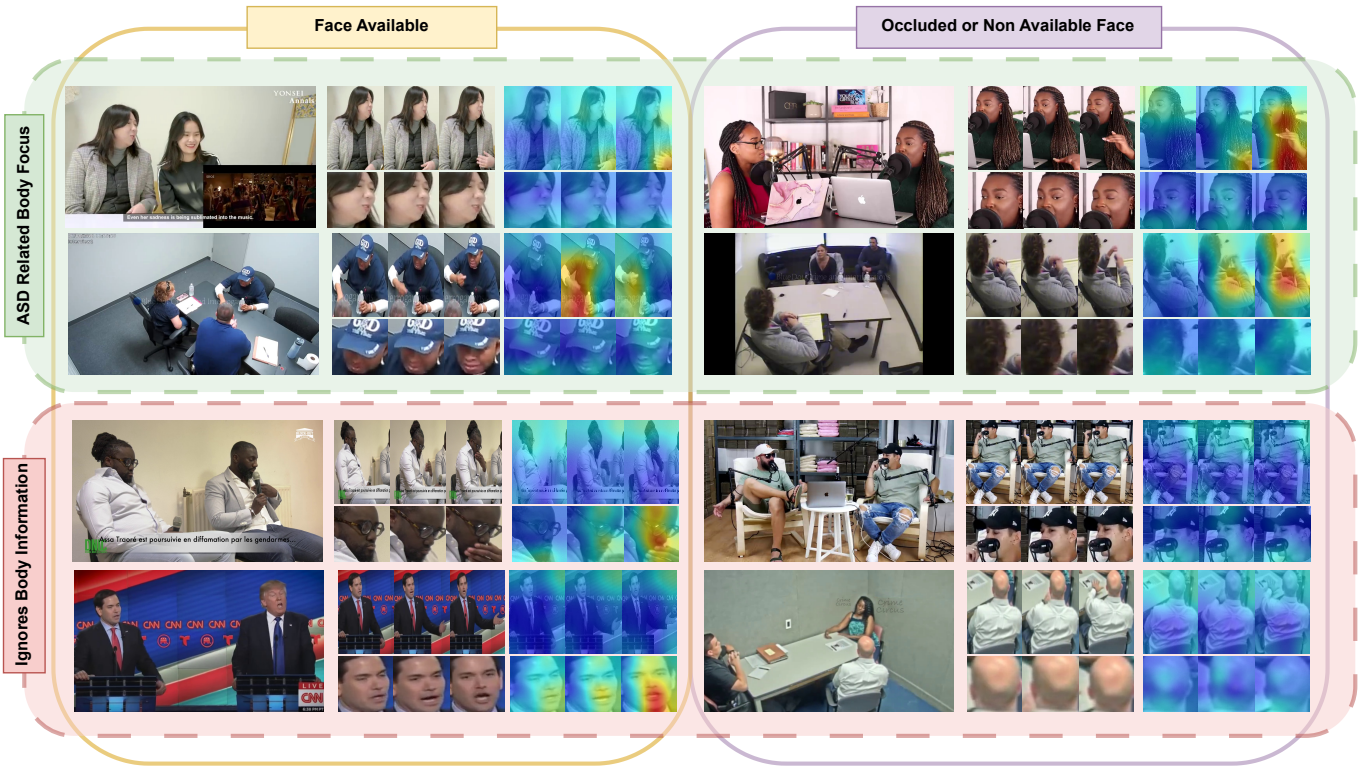


Fig. 6. Context, body and face crops, and respective BIAS attention heatmaps of various scenarios. We group the examples into 4 overlapping sets: 1) face available; 2) occluded or non available face; 3) focus on ASD-body movements, and 4) ignoring body information.

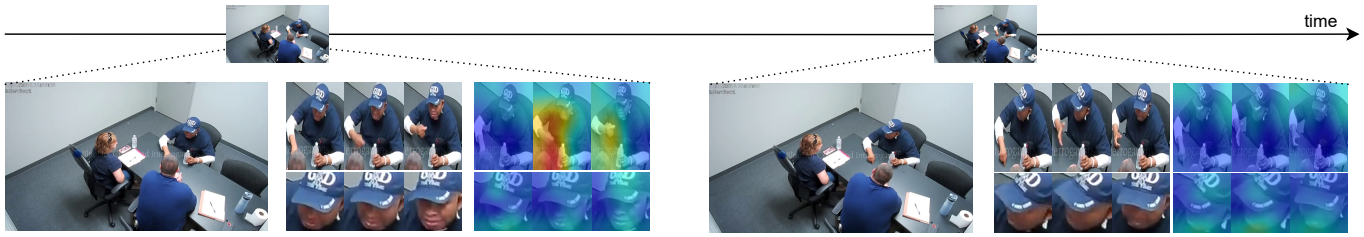


Fig. 7. Variation of body movement attention in different time lapses of the same context. Although sudden hand movements related to speech are considered by BIAS, subtle ones are not perceived as important, within the same challenging set.

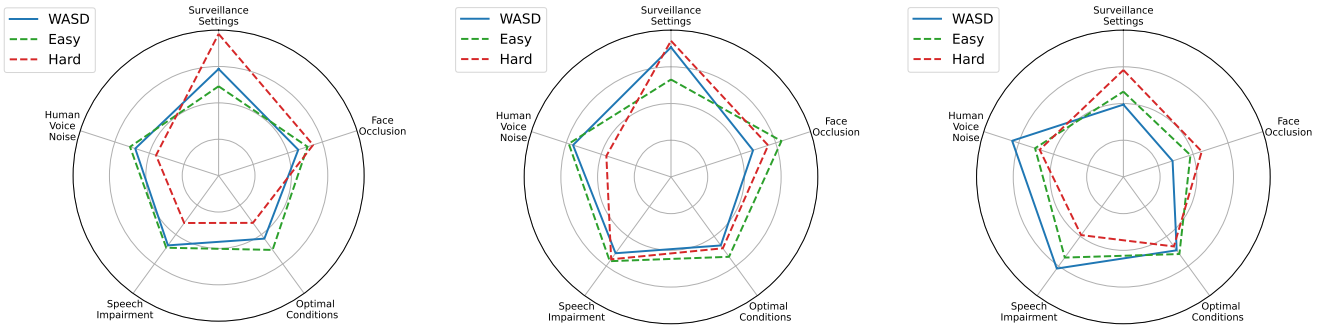
is done after feature extraction.

Body *vs.* Face Influence. Given the novelty of BIAS to include body for ASD, we explore the importance of body *vs.* face influence, in Table VIII, where we ignore facial cues of body data by including black boxes in the face region. The results show that, although facial cues are important for ASD, body-related actions also contribute towards BIAS performance. However, for AVA-ActiveSpeaker a significant portion of body information comes from the face region, mainly due to the proximity of subjects to the camera in its data (Figure 3), which further supports the inability of BIAS to achieve a state-of-the-art performance in such settings. Our approach in this experiment was to deliberately omit information, which can be seen as an adversarial attack, thus decreasing the model performance [72], [73]. As such, we also assess the importance of face *vs.* body with varying Head-Body Proportion (HBP), in Figure 5, for WASD. The results show that when body is significantly predominant relative to face (*i.e.*, low HBP, meaning that face is small), the model that uses body information (BIAS) is superior to others that only use face (BIAS_F and TalkNet), and this discrepancy is bigger the lower the HBP is. This shows that the performance discrepancy is independent of the model and is mainly due to the difference of data inputs (*i.e.*, it is not facial cues that justify the differences but rather the information present in body data). As such, even if body information also contains facial cues, body is the most relevant feature when face is small and/or not easily accessed.

V. BIAS Interpretability

A. Squeeze-and-Excitation Visual Interpretability

To assess the importance of body for ASD, we create attention heatmaps using SE channels, following the methodology described in Section III-B. These heatmaps highlight the region where BIAS is focusing on when predicting the active speaker of a given scene. We consider different scenarios to



(a) Audio (b) Body (c) Face

Fig. 8. Relevance of audio, body, and face-based features for ASD on the five different categories of WASD. Hard-BIAS gives higher audio importance in audio impaired categories, perceives body importance similar to WASD-BIAS for the most body-reliant scenario (Surveillance settings) and gives, overall, less face importance to face-dependent contexts, relative to WASD and Easy-BIAS.

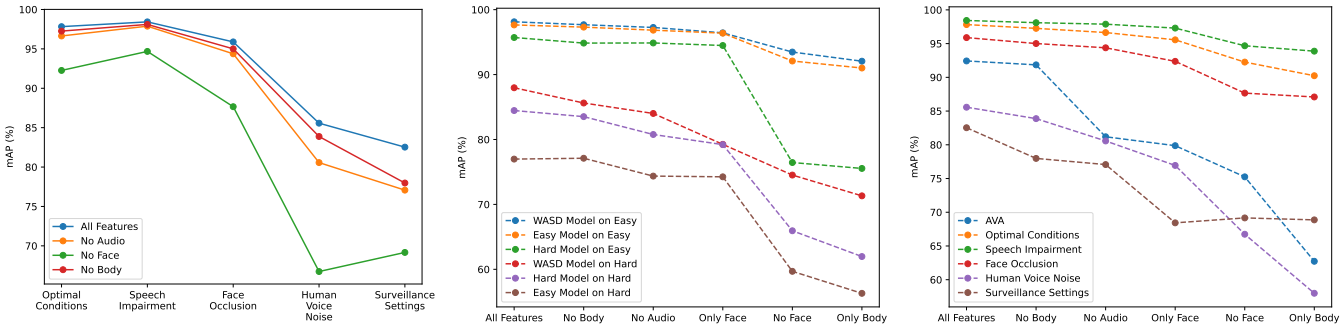


Fig. 9. Influence of the different features on BIAS performance (mAP): *left*) Effect of audio, face and body absence in WASD categories; *middle*) Feature variation influence relative to different training and testing sets (Easy, Hard, and WASD); and *right*) Feature effect on AVA-ActiveSpeaker (AVA) and WASD categories.

assess the relative importance of face and body via BIAS attention heatmaps, in Figure 6, where each scenario contains the initial setting (bigger image) and the next three frames for body and face movements (and their respective heatmaps).

Body Importance for ASD. The scenarios in the intersection of *ASD Related Body Focus* with *Face Available* show that BIAS gives higher attention to body when the scenario contains challenging features, even if face is available or partially accessible. In particular, with audio quality subpar (video playing) or limited face availability (hat), body movement importance increases for ASD. When considering the intersection of *ASD Related Body Focus* with *Occluded or Non Available Face*, we are limiting the assess to facial cues which further increases the relative importance of body, translated by BIAS attention to hand movements (in these examples). This shows that body information complement is mainly required in more challenging scenarios, where audio and face are not easily retrieved.

Face over Body. The intersection of *Ignores Body Information* with *Face Available*, shows examples where BIAS opts to give less relevance to body information (heatmaps of body with little/no coloring). In these cooperative conditions, face is the most relevant/reliable data, leading BIAS to rely exclusively on it and ignore body cues (hand movement in both cases).

No Body Focus. The intersection of *Ignores Body Information* with *Occluded or Non Available Face* shows situations where BIAS can also ignore body information in occluded or non-visible faces. In most cases, BIAS attention to body is mainly linked to pronounced (hand) movement. However, in the scenarios of the intersection, subjects only perform subtle movements when speaking and BIAS response to it is (wrongly) ignoring body information. This shows that BIAS body attention is not entirely linked to speaking activities but more on pronounced/abrupt movements, as exemplified by Figure 7. With different time frames of the same challenging set, body focus varies in similar scenes (hand movement while talking), with the difference being on the second timeframe having more subtle movement than the first.

B. Feature Importance Assessment

Although visual explanation provides some reasoning behind BIAS prediction, we also assess the relative importance of features (audio, face, and body data). We compare the effect of training sets on the perceived feature importance, using WASD and Easy/Hard group training, in Figure 8: WASD-BIAS, Hard-BIAS, and Easy-BIAS refer to BIAS training in WASD, Hard, and Easy data, respectively.

Relative Feature Importance. The main conclusions are: 1) For WASD-BIAS and Easy-BIAS, all the categories have

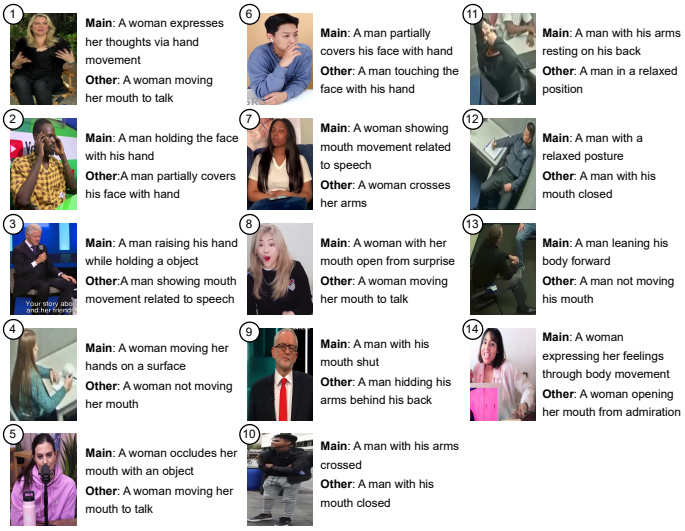


Fig. 10. ASD-Text dataset finetuned ViT-GPT2 model predictions in representative images of the 14 considered ASD-related actions of Table I. Each figure number relates to the respective ASD-action number. ViT-GPT2 predictions contain the main caption and a related one (other).

similar audio importance (Figure 8a). The most divergent variation is Hard-BIAS, which gives less importance to audio in categories with faces available and higher importance in scenarios with occluded faces; 2) Both WASD-BIAS and Hard-BIAS give higher importance to body in surveillance scenarios relative to other categories (Figure 8b), given the various challenges in this setting. Furthermore, with Hard-BIAS, body is less important for the category with more close-up faces and less visible body. This trend is not seen in the other models given their different training conditions (*i.e.*, they have access to other categories with close-up faces and less visible body), leading to a well rounded body importance; and 3) Regarding face importance (Figure 8c), Hard-BIAS is the more evenly distributed model, meaning that face is overall important for ASD regardless of the category. Regarding WASD-BIAS (and Easy-BIAS to some extent), face is more relevant for scenarios with close-up faces and impaired audio, leading to increased face importance relative to the other categories.

Performance Influence. We explore the effect of group training and feature influence in model performance, in Table IX and Figure 9, respectively. Overall, face is the most influential feature for BIAS in WASD, and the removal of body information has higher impact in surveillance settings, which highlights the body importance in these conditions (Figure 9a). When considering the impact of different features (audio, face, and body) in different training sets (Figure 9b), we observe that subsequent removal of available features translates into decreased performance, with training with all data (WASD) translating in more resilient results than training only on Easy or Hard data alone. This and the results in Table IX highlight the importance of robust training data for improved resilience, particularly to avoid situations where training may lead to a false sense of face reliability (drop of performance from Hard-BIAS on Easy data when not using

TABLE IX
EFFECT OF GROUP TRAINING, IN MAP, ON THE CATEGORIES OF WASD. EASY GROUP TRAINING CONTAINS DATA OF OPTIMAL CONDITIONS AND SPEECH IMPAIRMENT, WHILE HARD GROUP CONTAINS DATA OF THE REMAINING WASD CATEGORIES.

Category	Train Set		
	Easy	Hard	WASD
Optimal Conditions	97.1	94.7	97.8
Speech Impairment	98.2	96.7	98.4
Face Occlusion	90.4	93.5	95.9
Human Voice Noise	77.3	83.4	85.6
Surveillance Settings	63.2	79.4	82.5

TABLE X
VIT-GPT2 PERFORMANCE INCREASE WITH FINETUNE ON ASD-TEXT DATASET. METRIC DESCRIPTION: RL, ROUGE-L; M, METEOR; B1-4, BLEU1-4.

Model	RL	M	B1	B2	B3	B4
ViT-GPT2$_{Base}$	0.28	0.18	0.14	0.03	0.01	0
ViT-GPT2$_{ASD}$	0.61	0.58	0.61	0.50	0.37	0.31

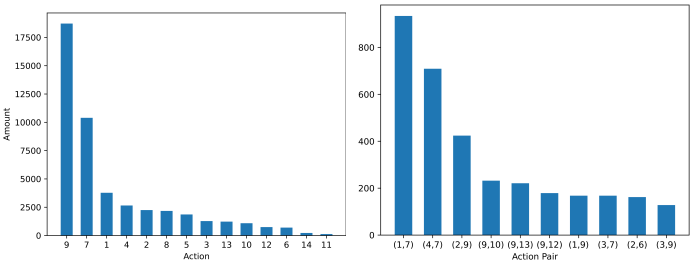


Fig. 11. Distribution of ASD-Text dataset actions (left) and pair of actions more commonly associated (right). Action labels refer to their numbers in Table I.

face). We also extend the analysis of feature influence in AVA and WASD training (Figure 9c). With subsequent removal of features, BIAS underperfoms the most on categories with audio impairment (HVN) and unreliable face access and audio quality (SS). Furthermore, surveillance settings is where body information is of utmost importance, based on the performance stability with using *Only Face*, *No Face* and *Only Body* (*i.e.* good face/body importance balance). Finally, BIAS training in AVA is not reliant on body, and has greater focus on audio and face features, heavily degrading in performance without these features. This shows that AVA data does not promote adequate feature combination, translating in models less resilient to varying data quality (*i.e.*, when audio and/or face quality are affected).

C. ASD-Text Dataset

To improve the interpretability of ASD-related scenarios, we complement the visual interpretability of attention heatmaps with text scene description, via caption generation of a ViT-GPT2 model. Given the absence of relevant data to train models for captions on ASD-related actions, we start by showing the relevance of the proposed ASD-Text dataset (Section III-D), annotation distribution, and its importance towards a full interpretability setup.

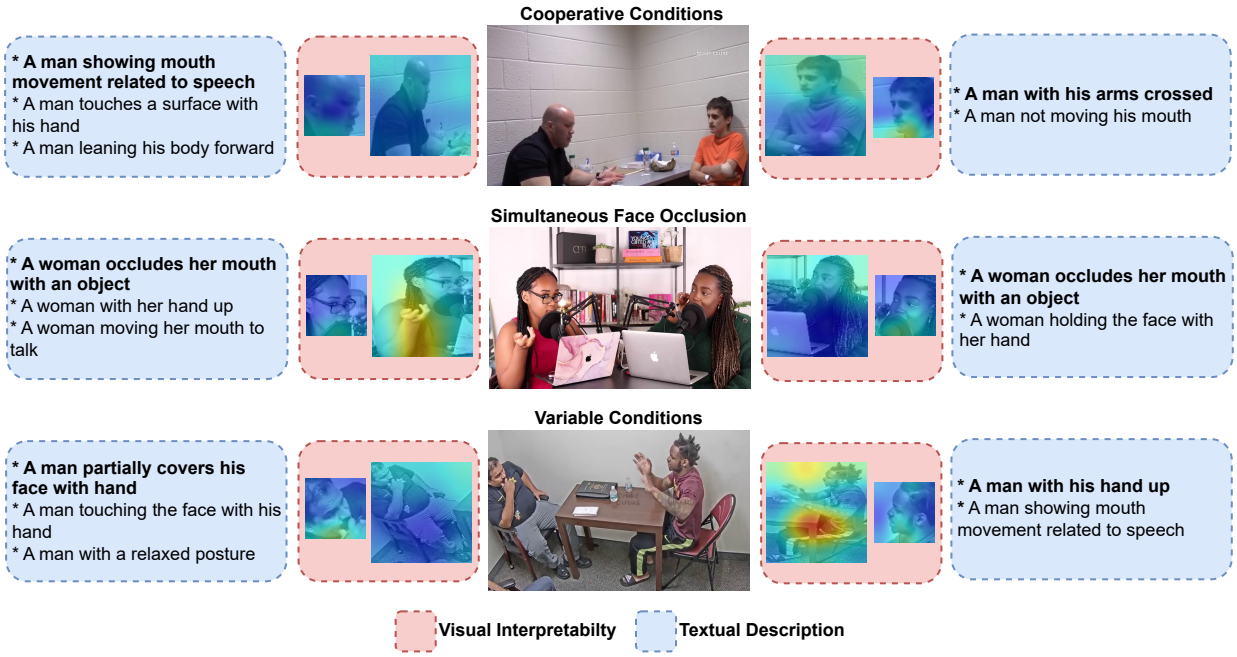


Fig. 12. Combination of SE visual interpretability with ViT-GPT2 model subject description in various scenarios with varying ASD challenges. Bold descriptions are the main prediction of ViT-GPT2.

ASD-Text Dataset Importance. We assess the performance of a pretrained (Base) and finetuned (ASD) ViT-GPT2 on ASD-Text dataset, on Table X, where the results support the need for ASD-Text dataset given the improvements of a finetuned ViT-GPT2 in predicting ASD-related captions. Based on the increase of all caption related metrics to reference levels [74], finetuned ViT-GPT2 can reliably describe ASD scenes in various challenging scenarios (WASD data). Furthermore, we also display ViT-GPT2 predictions on representative examples of the considered training actions (Table I), in Figure 10. We present two captions per image based on the top 5 predictions of ViT-GPT2, with *main* being its first predicted caption and *other* being the first caption not related to the action of the main.

Annotation Distribution. Figure 11 shows the overall distribution of ASD-Text dataset actions. The majority of subjects have mouths closed or actively moving, relating to the speaking labels considered (talking *vs.* not talking). Regarding the actions more commonly paired, we have subjects talking with hand movement, either raised or on a surface, illustrating body movement associated with talking. More linked to non-talking subjects we have hands touching faces, without occluding mouth, as the most common pair association.

Full Interpretability Setup. We can also use ASD-Text dataset to create a complete interpretability setup that combines the visual information of BIAS (attention heatmap) with the text description of ViT-GPT2, as shown in Figure 12. In the three different scenarios, with varying challenges, visual information and text scene description represent the key characteristics for the decision behind accurate ASD.

VI. Conclusion

In this paper we propose BIAS, a multi-modal approach for Active Speaker Detection (ASD) that singularly considers audio, face, and body-based information, which is state-of-the-art in challenging settings and has competitive results in more cooperative conditions. Furthermore, we propose a novel application of Squeeze-and-Excitation blocks to assess ASD feature importance in different settings and provide visual interpretability, complementing them with text descriptions from a ViT-GPT2 model (finetuned in ASD-Text dataset) for a full interpretability setup. Our work highlights the importance of body inclusion for ASD in unconstrained/challenging conditions and serves as baseline for models to perform in wilder scenarios such as surveillance settings.

Acknowledgments

This work was supported in part by the Portuguese FCT/MCTES through National Funds and co-funded by EU funds under Project UIDB/50008/2020; in part by the FCT Doctoral Grant 2020.09847.BD and Grant 2021.04905.BD.

References

[1] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi *et al.*, "Ava active speaker: An audio-visual dataset for active speaker detection," in *2020 IEEE ICASSP*. IEEE, 2020, pp. 4492–4496. 1, 2, 4, 5

[2] Y. J. Kim, H.-S. Heo, S. Choe, S.-W. Chung, Y. Kwon, B.-J. Lee, Y. Kwon, and J. S. Chung, "Look who's talking: Active speaker detection in the wild," *arXiv preprint arXiv:2108.07640*, 2021. 1, 2

[3] T. Roxo, J. C. Costa, P. R. M. Inácio, and H. Proença, "Wasd: A wilder active speaker detection dataset," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pp. 1–1, 2024. 1, 2, 4, 5, 6

[4] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. 1

[5] P. Chakravarty and T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in *Proceedings of the ECCV*. Springer, 2016, pp. 285–301. 2, 5

[6] J. L. Alcázar, F. Caba, A. K. Thabet, and B. Ghanem, "Maas: Multi-modal assignation for active speaker detection," in *Proceedings of the IEEE/CVF ICCV*, 2021, pp. 265–274. 2, 5

[7] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments," *arXiv preprint arXiv:2107.04174*, 2021. 2

[8] T. Roxo, J. C. Costa, P. R. Inácio, and H. Proença, "On exploring audio anomaly in speech," in *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2023, pp. 1–6. 2

[9] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the Conversation: Speaker Diarisation in the Wild," in *Proceedings of the Interspeech 2020*, 2020, pp. 299–303. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2337 2

[10] S. Nur Aisyah Mohd Robi, M. Atiff Zakwan Mohd Ariffin, M. A. Mohd Izhar, N. Ahmad, and H. Mad Kaidi, "Active speaker detection using audio, visual, and depth modalities: A survey," *IEEE Access*, vol. 12, pp. 96 617–96 634, 2024. 2

[11] J. S. Chung, "Naver at activitynet challenge 2019–task b active speaker detection (ava)," *arXiv preprint arXiv:1906.10555*, 2019. 2

[12] Y.-H. Zhang, J. Xiao, S. Yang, and S. Shan, "Multi-task learning for audio-visual active speaker detection," *The ActivityNet Large-Scale Activity Recognition Challenge*, pp. 1–4, 2019. 2

[13] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263. 2

[14] S.-W. Chung, J. S. Chung, and H.-G. Kang, "Perfect match: Improved cross-modal embeddings for audio-visual synchronisation," in *2019 IEEE ICASSP*. IEEE, 2019, pp. 3965–3969. 2

[15] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the 2006 IEEE Computer Society Conference on CVPR*, vol. 2. IEEE, 2006, pp. 1735–1742. 2

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 2, 4

[17] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," in *Proceedings of the ECCV*. Springer, 2020, pp. 208–224. 2

[18] Y. Cheng, R. Wang, Z. Pan, R. Feng, and Y. Zhang, "Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3884–3892. 2

[19] J. L. Alcázar, F. Caba, L. Mai, F. Perazzi, J.-Y. Lee, P. Arbeláez, and B. Ghanem, "Active speakers in context," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2020, pp. 12 465–12 474. 2, 5

[20] O. Köpüklü, M. Taseska, and G. Rigoll, "How to design a three-stage architecture for audio-visual active speaker detection in the wild," in *Proceedings of the IEEE/CVF ICCV*, 2021, pp. 1193–1203. 2, 5

[21] Y. Zhang, S. Liang, S. Yang, X. Liu, Z. Wu, S. Shan, and X. Chen, "Unicon: Unified context network for robust active speaker detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3964–3972. 2

[22] M. Welling and T. N. Kipf, "Semi-supervised classification with graph convolutional networks," in *Proceedings of the ICLR*, 2016. 2

[23] K. Min, S. Roy, S. Tripathi, T. Guha, and S. Majumdar, "Learning long-term spatial-temporal graphs for active speaker detection," *arXiv preprint arXiv:2207.07783*, 2022. 2

[24] J. L. Alcazar, M. Cordes, C. Zhao, and B. Ghanem, "End-to-end active speaker detection," *arXiv preprint arXiv:2203.14250*, 2022. 2

[25] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3927–3935. 2, 4, 5, 6

[26] J. Liao, H. Duan, K. Feng, W. Zhao, Y. Yang, and L. Chen, "A light weight model for active speaker detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 932–22 941. 2, 5, 6

[27] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 789–792. 2

[28] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proceedings of the IEEE ICCV*, 2017, pp. 350–359. 2

[29] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *arXiv preprint arXiv:1603.07054*, 2016. 2

[30] T. Roxo and H. Proença, "Yinyang-net: Complementing face and body information for wild gender recognition," *IEEE Access*, vol. 10, pp. 28 122–28 132, 2022. 2

[31] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Jin, "Grouping attribute recognition for pedestrian with joint recurrent learning." in *Proceedings of the IJCAI*, 2018, pp. 3177–3183. 2

[32] C. Tang, L. Sheng, Z. Zhang, and X. Hu, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization," in *Proceedings of the IEEE ICCV*, 2019, pp. 4997–5006. 2

[33] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proceedings of the ECCV*, 2018, pp. 680–697. 2

[34] H. Guo, K. Zheng, X. Fan, H. Yu, and S. Wang, "Visual attention consistency under image transforms for multi-label image classification," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2019, pp. 729–739. 2

[35] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2015, pp. 111–115. 2

[36] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, vol. 95, pp. 151–161, 2019. 2

[37] J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method," *arXiv preprint arXiv:2005.11909*, 2020. 2

[38] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE ICCV*, 2017, pp. 618–626. 2

[39] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016. 2

[40] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," *Advances in neural information processing systems*, vol. 32, 2019. 2

[41] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017. 2

[42] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the ICML*. PMLR, 2017, pp. 3319–3328. 2

[43] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015. 2

[44] J. Gu, Y. Yang, and V. Tresp, "Understanding individual decisions of cnns via contrastive backpropagation," in *Proceedings of the Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 2019, pp. 119–134. 2

[45] B. K. Iwana, R. Kuroki, and S. Uchida, "Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation," in *Proceedings of the 2019 IEEE/CVF ICCVW*. IEEE, 2019, pp. 4176–4185. 2

[46] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017. 2

[47] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the ICML*. PMLR, 2017, pp. 3145–3153. 2

[48] B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," *IEEE TPAMI*, vol. 41, no. 9, pp. 2131–2145, 2018. 2

[49] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," *Advances in neural information processing systems*, vol. 30, 2017. 2

[50] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018. 2

[51] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proceedings of the IEEE/CVF ICCV*, 2019, pp. 2950–2958. 2

[52] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE ICCV*, 2017, pp. 3429–3437. 2

[53] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2021, pp. 782–791. 2

[54] B. Yin, L. Tran, H. Li, X. Shen, and X. Liu, "Towards interpretable face recognition," in *Proceedings of the IEEE/CVF ICCV*, 2019, pp. 9348–9357. 2

[55] M. Winter, W. Bailer, and G. Thallinger, "Demystifying face-recognition with locally interpretable boosted features (libf)," in *2022 10th EUVIP*. IEEE, 2022, pp. 1–6. 2

[56] J. Fu, J. Gao, and C. Xu, "Learning semantic-aware spatial-temporal attention for interpretable action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5213–5224, 2021. 2

[57] K. Doshi and Y. Yilmaz, "Towards interpretable video anomaly detection," in *Proceedings of the IEEE/CVF WACV*, 2023, pp. 2655–2664. 2

[58] T. Roxo and H. Proença, "Is gender "in-the-wild" inference really a solved problem?" *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 573–582, 2021. 2

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 3

[60] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE TPAMI*, vol. 44, no. 12, pp. 8717–8727, 2018. 3

[61] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer, and P. Vajda, "Visual transformers: Token-based image representation and processing for computer vision," *arXiv preprint arXiv:2006.03677*, 2020. 4

[62] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. 4

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on CVPR*, 2009, pp. 248–255. 4

[64] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015. 4

[65] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on CVPR*, 2015, pp. 3128–3137. 4

[66] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6047–6056. 4

[67] J. L. Alcazar, M. Cordes, C. Zhao, and B. Ghanem, "End-to-end active speaker detection," 2022. [Online]. Available: https://arxiv.org/abs/2203.14250 5

[68] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the Text summarization branches out*, 2004, pp. 74–81. 5

[69] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72. 5

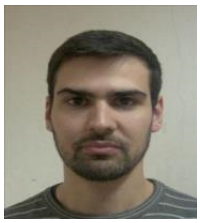
[70] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. 5

[71] X. Wang, F. Cheng, G. Bertasius, and D. Crandall, "Loconet: Long-short context network for active speaker detection," *arXiv preprint arXiv:2301.08237*, 2023. 6

[72] N. Hingun, C. Sitawarin, J. Li, and D. Wagner, "Reap: a large-scale realistic adversarial patch benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4640–4651. 7

[73] J. C. Costa, T. Roxo, H. Proença, and P. R. M. Inácio, "How deep learning sees the world: A survey on adversarial attacks & defenses," *IEEE Access*, vol. 12, pp. 61 113–61 136, 2024. 7

[74] Q. Wang, J. Wan, and A. B. Chan, "On diversity in image captioning: Metrics and methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1035–1049, 2020. 10



Tiago Roxo (Member, IEEE) obtained a bachelor's degree in Computer Science and Engineering from Universidade da Beira Interior (UBI) in 2019 and is currently pursuing a Ph.D.'s degree, with a FCT (*Fundação para a Ciência e a Tecnologia*) scholarship, in the field of Computer Vision and Artificial Intelligence.

Joana Cabral Costa obtained her bachelor's and master's degree in Computer Science and Engineering from Universidade da Beira Interior (UBI) in 2019 and 2021, respectively. She is currently pursuing a Ph.D.'s degree, with a FCT (*Fundação para a Ciência e a Tecnologia*) scholarship, in the field of Computer Vision and Adversarial Attacks.

Pedro R. M. Inácio is an associate professor of the Department of Computer Science at the University of Beira Interior (UBI), which he joined in 2010. Lectures subjects related with information assurance and (cyber)security, and computer based simulation, to graduate and undergraduate courses, namely to the B.Sc., M.Sc. and Ph.D. programmes in Computer Science and Engineering. He is currently the Pro-Rector for the Digital University and the Data Protection Officer of UBI. Holds a 5-year B.Sc. degree in Mathematics/Computer Science and a Ph.D. degree in Computer Science and Engineering, obtained from UBI, Portugal, in 2005 and 2009 respectively. The Ph.D. work was performed in the enterprise environment of Nokia Siemens Networks Portugal S.A., through a Ph.D. grant from the Portuguese Foundation for Science and Technology.

He is an IEEE senior member, an ACM professional member and a researcher of the Instituto de Telecomunicações (IT). His main research topics are information assurance and security, computer based simulation, and network traffic monitoring, analysis and classification. He has 70+ publications in the form of book chapters and papers in international peer-reviewed books, conferences and journals. He frequently reviews papers for IEEE, Springer, Wiley and Elsevier journals. He is a member of the Technical Program Committees of flagship national and international workshops and conferences, such as ACM SAC, IEEE NCA, IFIPSEC or ARES.

Hugo Proença (SM'12), B.Sc. (2001), M.Sc. (2004) and Ph.D. (2007) is an Associate Professor in the Department of Computer Science, University of Beira Interior and has been researching mainly about biometrics and visual-surveillance. He was the coordinating editor of the IEEE Biometrics Council Newsletter and the area editor (ocular biometrics) of the IEEE Biometrics Compendium Journal. He is a member of the Editorial Boards of the Image and Vision Computing, IEEE Access and International Journal of Biometrics. Also, he served as Guest Editor of special issues of the Pattern Recognition Letters, Image and Vision Computing and Signal, Image and Video Processing journals.