Less is More: Adaptive Feature Selection and Fusion for Eye Contact Detection

Fuyan Ma
Chinese Academy of Military Science
Beijing, China
mafuyan@hnu.edu.cn

Yiran He
Changchun University of Science and Technology
Changchun, China
heyiran@mails.cust.edu.cn

Bin Sun*
Hunan University
Changsha, China
sunbin611@hnu.edu.cn

Shutao Li
Hunan University
Changsha, China
shutao_li@hnu.edu.cn

ABSTRACT

Detecting eye contact is essential for embodied robots to engage in natural interactions with humans, enhancing the intuitiveness and comfort of these exchanges. However, eye contact detection often presents a significant challenge due to a variety of factors, such as low contrast and various forms of occlusions. Existing methods incorporate convolutional neural networks (CNNs) or Transformers to learn discriminative representations, but usually ignore the influence of noisy or less relevant regions in facial images. To address this gap, we propose the deep feature selection and fusion network (FSFNet) for eye contact detection in multi-party conversations. Our proposed method adaptively selects fine-grained visual features and reduces the impacts of irrelevant features. Specifically, we present a local feature selection scheme that leverages the attention scores to progressively concentrate on the most informative features. By integrating the carefully selected features into the multi-head self-attention module, we can maintain the superior properties of Transformers while simultaneously reducing the overall computational demands. We evaluate the proposed method on the official eye contact detection datasets, which achieves promising results of 0.8174 and 0.79 on the validation and test sets, respectively. We have made the source code publicly accessible in https://github.com/ma-hnu/FSFNet.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; *Supervised learning*; **Neural networks**; • **Human-centered computing** → HCI theory, concepts and models.

KEYWORDS

Feature selection and fusion, Transformer, eye contact detection, multi-party conversation

*Bin Sun is the corresponding author.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3688987

ACM Reference Format:
Fuyan Ma, Yiran He, Bin Sun, and Shutao Li. 2024. Less is More: Adaptive Feature Selection and Fusion for Eye Contact Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3664647.3688987

1 INTRODUCTION

Eye contact is a fundamental aspect of non-verbal communication that helps in establishing and maintaining social bonds. It helps participants in a multi-party conversation to understand when they are being addressed or when it might be their turn to speak. Automatic eye contact detection holds significant importance for various applications and scenarios, such as human-robot interaction, emotional intelligence and healthcare. In medical settings, it can be used to monitor patients for signs of cognitive decline or other conditions that affect social interaction. For embodied robots or AI systems designed to interact with humans, the ability to detect eye contact can make the interactions feel more natural and human-like. Consequently, automatic eye contact detection has received growing attention from multidisciplinary researchers.

While there are extensive studies [6, 22–24] on human behavior analysis, limited effort has been made for eye contact detection in multi-person conversations, especially in the computer vision domain. Recent advancements in deep learning have significantly improved the accuracy of eye contact detection systems. Convolutional neural networks (CNNs) are particularly effective due to their ability to learn complex patterns from large datasets. For instance, the study by Chong et al. [3] demonstrates a deep neural network model that achieved accuracy comparable to human experts by training on millions of annotated eye contact events. However, CNNs may struggle with detecting eye contact when the subject's head pose is extreme or when there is occlusion of the eyes due to glasses, hats, or other obstructions. Ignoring these issues during the deep model design phase can invariably result in suboptimal classification outcomes.

Recent flourishing of Vision Transformers (e.g., ViT [7] and its variants [17, 18]) has considerably deepened our understanding about visual feature representation. Consequently, it is an open question how to leverage Transformers to adaptively capture the contextual information within facial features for eye contact detection. For example, Ma *et al.* [16] propose a unified network for both eye contact detect and next speaker detection. The visual

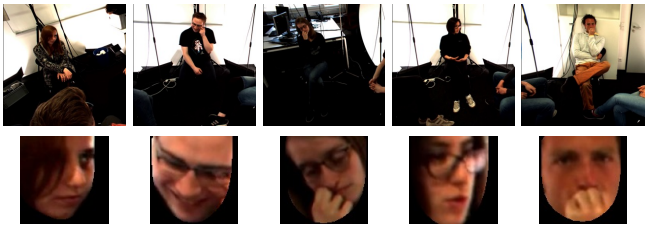


Figure 1: For the eye contact detection task, we utilize the OpenFace [1] toolkit to detect and align the facial regions of each participant. Non-frontal head poses, diverse occlusions, and different illumination conditions are commonly seen in the MPIIGroupInteraction dataset [22].

feature token sequence is modelled by the vanilla Transformer encoders, and the relationships among feature tokens are captured with the multi-head self-attention (MHSA) mechanism. The eye contact detection task faces the challenges of occlusions and poses variations. The deep models may mistakenly focus on some occlusion or background areas and generate wrong results. Meanwhile, the linear increase of the tokens incurs a quadratic computation cost for calculating token relationships with self-attention.

In this paper, we present our solution for the eye contact detection task in the MultiMediate challenge [19–21]. The challenge task is formulated to predict four categories (i.e., left, right, frontal and no eye contact), while the previous methods [29, 35] output whether the subject is looking at a target or not. As shown in Figure 1, we first detect and align the facial regions by OpenFace to remove redundant background information. These facial images in unconstrained environments frequently encounter unforeseen challenges, including occlusions caused by hair, eyeglasses, and hands. To improve the performance of eye contact detection, we argue that it is necessary and important for deep models to focus on foreground areas and discriminative regions, regardless of occlusions and less informative facial parts. In this work, we propose to identify the attentive feature tokens by the attention scores during each feed-forward process. Specifically, the attentiveness (importance) of the classification [CLS] token with respect to each feature token is calculated between MHSA and FFN (i.e., feed-forward network) modules. We select the top-k attentive tokens and discard the other tokens by ranking the attentiveness scores in descending order. The selected tokens are sent into the subsequent FFN and MHSA modules for aggregating global visual information and making the final classification. For the eye contact detection task, our proposed FSFNet achieves the result of 0.8174 on the validation set and the performance of 0.79 on the test set.

2 RELATED WORK

2.1 Eye contact detection and gaze estimation

Eye contact detection and gaze estimation are related but distinct concepts within the field of computer vision and human-computer interaction. Both of two tasks are integral to advancing our understanding of human interaction and have significant implications for technology that interfaces with human users. Eye contact detection typically involves binary classification (eye contact present or not), while gaze estimation is a regression problem that predicts a continuous gaze direction.

Early works in eye contact detection relied on heuristics and simple image processing techniques to identify eye contact. Ye *et al.* [34] utilize wearable eye-tracking glasses to determine the gaze of the parent and his child, and detect eye contact by the adult's point of gaze and the child's gaze direction with pre-defined rules. Meanwhile, the wearable devices are both expensive and burdensome to the subject. Previous appearance-based methods (e.g., [15, 36]) operate under the assumption that participants are facing the camera directly, which does not readily extend to scenarios involving multi-person conversations. With the advent of deep learning, there has been a shift towards using CNNs for eye contact detection, leveraging large datasets of labeled eye contact instances. Otsuka *et al.* [25] and Fu *et al.* [9] both use CNNs to extract discriminative features and improve the detection performance.

Traditional gaze estimation systems (e.g.,[27, 28]) use specialized hardware like infrared cameras and require user calibration for accurate measurements. In contrast, appearance-based gaze estimation methods (e.g.,[33, 36]) use standard RGB or RGBD cameras and rely on the visual appearance of the eyes and face to infer gaze direction. Notable works in gaze estimation include the MPIIGaze [37] and EYEDIAP [10] datasets, which have been used to train CNNs for gaze direction prediction in the wild. In summary, while eye contact detection and gaze estimation share some underlying techniques and challenges, they differ in their goals and applications. Eye contact detection focuses on identifying direct gaze at a camera or another person, whereas gaze estimation aims to determine the direction of gaze within a broader visual scene. Both fields have seen significant advancements with the incorporation of deep learning and computer vision techniques.

2.2 Transformers in computer vision

Transformers have become a pivotal architecture in the realm of computer vision, offering a versatile framework that can be adapted to a myriad of tasks. Initially introduced to the field of natural language processing (NLP), the transformative success of models like BERT [5] demonstrate the power of the self-attention mechanism. The transition of these models into computer vision has been marked by significant milestones. For instance, ViT [7] introduces a paradigm shift by applying the Transformer directly to image patches, sidestepping the traditional reliance on convolutional layers for feature extraction. This pre-trained model shows competitive performance on image classification tasks and sets a new benchmark for subsequent research.

The adaptability of Transformers has led to their successful deployment in various high-level vision tasks such as object detection, semantic segmentation, and video understanding. For example, the DETR [2] redefines the object detection task by formulating it as a set prediction problem, thereby removing the need for complex hand-crafted components like anchor boxes and non-maximum suppression. This end-to-end paradigm not only simplifies the pipeline but also enhances the interpretability of the model. Similarly, Swin Transformer [14] introduces a hierarchical vision Transformer using shifted windows, which significantly boost the performance of the model on various visual recognition tasks. Swin Transformer

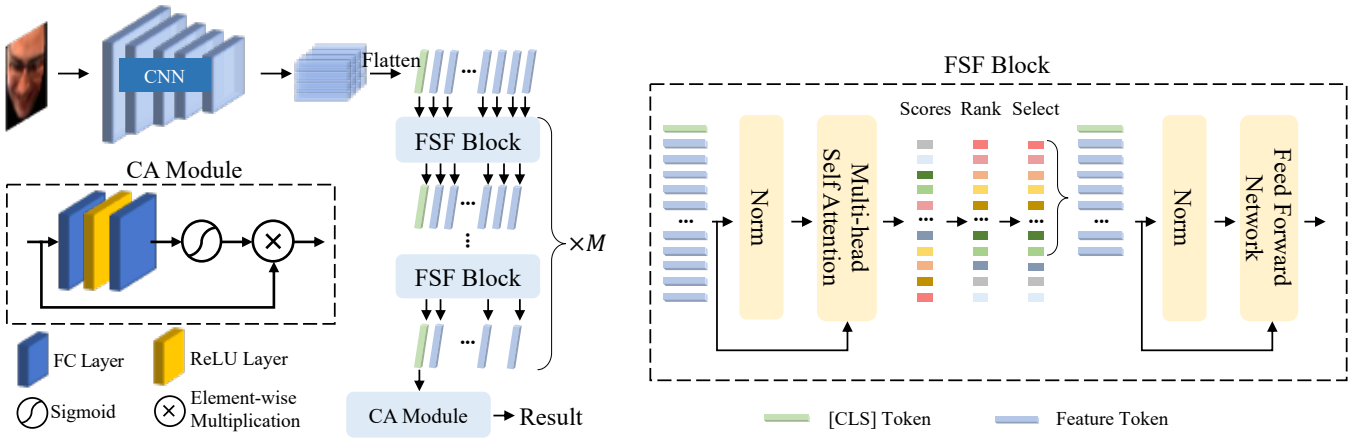


Figure 2: Overview of our method for eye contact detection. The CNN backbone first extracts feature maps for the aligned facial image. The feature maps are flattened into a sequence of feature tokens. The attentiveness scores are leveraged to adaptively select more relevant tokens and drop irrelevant tokens. The kept tokens are sent into the subsequent layers to fuse and exchange information through the [CLS] token. After the FSF blocks, the channel attention (CA) module is applied to enhance the discriminative feature learning.

demonstrates the effectiveness of incorporating Transformer architecture in handling spatial hierarchies in images.

Moreover, the field has seen the development of efficient Transformer models to address the computational intensity of these architectures. For instance, Wang *et al.* [32] propose a self-attention mechanism with linear complexity, making it more scalable for large sequences. The exploration of Transformers in computer vision is still in its ascendancy, with continuous efforts directed towards enhancing their efficiency, interpretability, and generalization capabilities. While the Transformer architecture has demonstrated remarkable success in a multitude of vision tasks, the application of Transformers to eye contact detection has not been extensively covered in the literature.

3 METHODOLOGY

3.1 Overview

As discussed above, our primary motivation is to select informative feature tokens and discard features associated with occlusions or background noise during each feed-forward process, thereby reducing computational cost without increasing the number of learnable parameters. We present the overall framework, depicted in Figure 2, based on the hybrid CNN-Transformer architecture. Given a pre-aligned facial image, we first extract features maps by a CNN backbone, and then flatten the maps into a sequence of feature tokens for the subsequent FSF blocks. Our proposed FSFNet introduces the importance scores for each token, and those tokens with higher scores are kept. The selected top-k feature tokens and the [CLS] token are jointly sent into the following FSF blocks. As the network deepens, the most informative tokens are gradually selected, while the inattentive tokens are identified and discarded. In this way, our proposed FSFNet not only gradually decreases the number of tokens, but also facilitates the self-attention module in focusing on eye contact-related tokens. After the FSF blocks, the final [CLS] token is through the squeeze-and-excitation module to make predictions (i.e., left, frontal, right and no eye contact).

3.2 Input Embedding Generation

For the stem CNN with L blocks, the mapping function of the backbone can be represented as $\mathcal{F}_L = f_1 \circ f_2 \cdots \circ f_L$, where f denotes the mapping function of each block and $f_i \circ f_{i+1}$ is the function composition: f_{i+1} after f_i. Given a pre-aligned facial image I, the intermediate feature map of the j-th block can be obtained by

$$X_j = \mathcal{F}_j(I; f_{\theta_1}, f_{\theta_2}, \cdots, f_{\theta_j}), j \in \{1, 2, \cdots, L\}, \tag{1}$$

where f_θ denotes the learnable parameter for the mapping function f. The initial three stages of ResNet50[4], pretrained on the MS-Celeb-1M face recognition dataset [11], are utilized as the backbone to extract visual feature maps. Therefore, the feature maps of the facial image I are generally formulated as :

$$X_3 = \mathcal{F}_3(I; f_{\theta_1}, f_{\theta_2}, f_{\theta_3}). \tag{2}$$

Here we assume the size of the output feature maps X_3 is $\mathbb{R}^{H\times W\times C}$, and then X_3 is flattened into a feature sequence $X_s \in \mathbb{R}^{(H\times W)\times C}$. An additional classification token [CLS] is prepended to each feature sequence for aggregating abstract global feature representations and final classification. We train the whole FSF blocks from scratch, and the [CLS] token is expected to contain the information about the sequence at the early stage. Therefore, we initialize the [CLS] token with the average of token embeddings X_s instead of a normal distribution. Afterwards, all of these tokens are added by a learnable vector (i.e., positional encodings) and sent into the sequentially-stacked FSF blocks.

3.3 Feature Selection and Fusion

Observing the redundancies of the aligned facial images in the spatial domain, we aim to select a few informative tokens and discard the less relevant ones for further training or evaluation, which could

prevent the model from focusing on occlusion or other noisy areas. To this end, we introduce a dynamic token selection mechanism to identify different types of tokens, so that only the tokens with higher scores will be kept and forwarded into subsequent layers.

Each FSF block consists of a MHSA layer and a FFN layer. In the first forward process, the normalized feature tokens are linearly projected and packed into three matrices in MHSA, denoted as queries ($Q \in \mathbb{R}^{(n+1)\times d}$), keys ($K \in \mathbb{R}^{(n+1)\times d}$), and values ($V \in \mathbb{R}^{(n+1)\times d}$). $n = H \times W$ is the sequence length of the input tokens, and d is the embedding dimension of these tokens. Therefore, the self-attention is generally conducted as:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V. \tag{3}$$

The coefficients in $Softmax(\frac{QK^T}{\sqrt{d}})$ determine how much information of V should be fused into the self-attention output through Eq. 3.

In the FSF block, the interactions between the [CLS] token and other tokens are calculated as:

$$S_{cls} = Softmax(\frac{q_{cls}K^T}{\sqrt{d}}), \tag{4}$$

where q_{cls} denotes the query vector of the [CLS] token, and S_{cls} represents the relationships between q_{cls} and keys K. The higher the S_{cls} value is, the more relevant between the q_{cls} and the key k. In other words, the [CLS] token pays more attention(i.e., having a higher attention score) to these tokens that contribute more to the eye contact detection.

Therefore, we can take the attentiveness scores S_{cls} of the [CLS] token as the criterion to identify the most important tokens. Specifically, the attention value S^i_{cls} indicates the importance of the i-th token. There are multiple heads performing the self attention parallelly. We average the scores $S^{(h)}_{cls}, h \in \{1, 2, \ldots, N_H\}$ of N_H heads by $\sum_{h=1}^{N_H}(S^{(h)}_{cls})/N_H$. As shown in Figure 2, we rank the scores in descending order and keep the tokens corresponding to the top-k scores, which we call the attentive tokens. As there are M stacked FSF blocks in the encoder, we utilize a token keep rate r for all M blocks to adjust the token number:

$$N_i = r * N_{i-1}, i \in \{1, 2, \ldots, M\}, \tag{5}$$

where $N_0 = H \times W + 1$ denotes the sequence length of the input tokens. The token number gradually decreases as the depth increases. It is noted that the [CLS] token is always reserved, and only the feature tokens are selected or discarded. After removing the less relative tokens, the rest tokens are sent to the subsequent layers (i.e., layer normalization and a feed-forward network). And in the next forward operation, the [CLS] token aggregates and fuses the selected tokens via MHSA.

3.4 Prediction and Optimization

After the last FSF block, the [CLS] token is used to generate the final eye contact detection result through the channel attention module. We build the CA module on top of the FSF blocks, where two fully connected layers and corresponding activation functions are used on the [CLS] token, as shown in Figure 2. This process can be formulated as:

$$x_{output} = Sigmoid(FC_2(ReLU(FC_1(x_{cls})))) * x_{cls}. \tag{6}$$

The CA module is to optimize the learning of whole network by capturing more global relationships among these local feature tokens. Different from the self-attention module in the FSF blocks, the CA module is applied to recalibrate the feature responses explicitly among feature channels. The network is optimized by the standard cross entropy loss after projecting the [CLS] token from feature space to label space. During the inference phase, the token keep rate r remains consistent with that used during the training process.

4 EXPERIMENTS

4.1 Dataset and Evaluation Metrics

We evaluate our proposed method on the official MPIIGroupInteraction dataset [22], which consists of 22 group interactions and 78 German-speaking participants. Specifically, the dataset contains 4504 training samples, 1672 validation samples and 1848 testing samples for eye contact detection. As the eye contact detection task can be formulated as single-label classification, the overall accuracy is used to evaluate the eye contact detection performance.

4.2 Implementation Details

Data Preprocessing: We use the OpenFace toolkit [1] to detect and align facial images. Due to the limitations of the toolkit, certain facial images fail to be detected. As a result, there are 4486 samples for training, 1665 samples for validation and 1835 samples for testing. We resize all facial images to 112×112 pixels and apply the data argumentation techniques (e.g., random erasing) to prevent overfitting.

Experimental Settings: We use the PyTorch framework [26] to implement our proposed method[1]. The Adam optimizer [12] and the Sharpness-Aware Minimization algorithm [8] are used to optimize the network. We train the model for 300 epochs on one NVIDIA GTX 4090 GPU card with an initial learning rate of 4e-6. It is noted that the parameters in FSF blocks are trained from scratch. In addition, the learning rate is decayed by the gamma of 0.98 every epoch, and the batch size is set to 128. The backbone outputs the features maps with the size of $14 \times 14 \times 512$, namely $H = W = 14$ and $n = 196$. The number of heads N_H in MHSA is empirically assigned to 8, and the embedding dimension d is 512. We further investigate the number of FSF blocks M in the ablation study.

4.3 Comparison with State-of-the-Art Methods

We compare our proposed FSFNet with several state-of-the art methods on the test set in Tab. 1. The organizers provide a strong baseline [21] by training RBF-SVMs on head pose and eye gaze direction feature vectors. By re-implementing the official baseline method [20] and tuning hyperparameters, the organizers improve the performance from 0.52 to 0.576 on the test set. Fu *et al.* [9] use motion histories to represent the participants' behavioral features and train a CNN to extract appearance features, which achieves the result of 0.56. Different from previous methods, TA-CNN [16] is

[1] https://github.com/ma-hnu/FSFNet

Table 1: Performance comparison with SOTA methods on the test set of MPIIGroupInteraction.

Method	Year	Accuracy
Baseline [21]	2021	0.52
Motion [9]	2021	0.56
Baseline [20]	2022	0.576
TA-CNN [16]	2022	0.7261
MSCFN [30]	2023	0.65
DA [13]	2023	0.777
FSFNet	2024	0.79

Table 2: Evaluation of different token keep numbers (r) on the validation set, and corresponding GFLOPs compared with the baseline.

r	Accuracy	GFLOPs
1.0	0.8135	100%
0.9	0.8162	66%
0.8	0.8148	45%
0.7	0.8156	33%
0.6	0.8174	24%
0.5	0.8150	19%
0.4	0.8132	15%
0.3	0.8103	13%

designed to detect eye contact in an end-to-end manner. TA-CNN utilizes the detected facial images to directly extract discriminative features by a hybrid CNN-Transformer architecture. As a result, TA-CNN significantly improves the eye contact detection performance to 0.7261. Li *et al.* [13] further investigate the effect of different data argumentation techniques for eye contact detection. They achieve the performance of 0.777 by fine-tuning a pretrained Swin Transformer with the cropped facial images. MSCFN [30] utilizes the conformer blocks to generate multi-scale features and obtains the result of 0.65. Although these previous methods achieve promising results, they commonly fail to overcome the difficulties under unconstrained conditions (e.g., occlusions and complex backgrounds). With the help of our proposed FSF block to progressively discard irrelevant features, our method outperforms these state-of-the-art methods with an accuracy of 0.79 on the test set.

4.4 Ablation Study

Impact of the Token Keep number r in FSF Blocks: The token keep number r is a hyperparameter to control how many feature tokens should be kept. To explore the impact of the parameter r, we fix the block number $M = 8$ and conduct experiments with different values r from 1.0 to 0.3. We also calculate the computational costs (i.e., GFLOPs) of the FSF blocks by the fvcore toolkit [2] and report them in Table 2. The CNN backbone extracts feature maps in a 5.493 GFLOPs cost, and the baseline Transformer encoder has 0.99 GFLOPs. When $r = 1.0$, it represents the baseline model without dropping any redundant tokens. The detailed results are shown in Table 2, and our proposed FSFNet framework achieves good accuracy/GFLOPs trade-offs. As shown in Table 2, when $r = 0.6$, it boosts the performance from 0.8135 to 0.8174, but with only 24% GFLOPs in FSF blocks compared with the baseline model. The smaller r is, the fewer feature tokens are kept for eye contact detection, which forces the model to focus on more informative features in the spatial domain. However, we find that too less information may make it harder to generate correct predictions, so the results decrease when r is too small (e.g., 0.3, 0.4).

[2]https://github.com/facebookresearch/fvcore

Table 3: Evaluation of different block numbers (M) on the validation set.

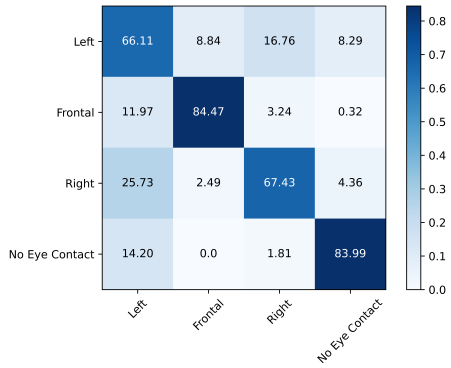
Scale	r	Accuracy
small	0.6	0.8078
medium	0.6	0.8168
large	0.6	0.8174
small	0.7	0.8072
medium	0.7	0.8140
large	0.7	0.8156

Table 4: Evaluation of different modules on the validation set and test set. The best result on the test set is obtained when the token keep ratio $r = 0.6$.

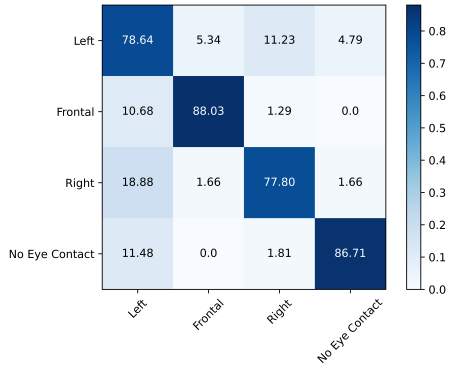
Method	Accuracy (val)	Accuracy (test)
ResNet50[1]	0.7345	-
ResNet50[2]	0.7785	-
FSFNet w/o CA	0.7994	0.76
FSFNet ($r = 0.6$)	0.8174	0.79
FSFNet EMA	0.8203	-

Impact of the Block Number M in FSF Blocks: We construct large/medium/small-scale networks (M=8, 6, 4) to capture the relationships among feature tokens. The comparison results are shown in Table 3. For each scale and r value, the accuracy our the network is presented. When $r = 0.6$, the accuracy for small-scale networks is 0.8078, for medium-scale networks is 0.8168, and for large-scale networks is 0.8174. We find that with the increase of M, the feature representation ability greatly enhances, and the performance significantly improves. Similar experimental results can be seen from these large/medium/small-scale networks when $r = 0.7$.

Effectiveness of the Proposed Modules: To evaluate the effect of the proposed modules, we design the ablation study on the validation set to better understand the impact of the proposed FSF block and CA module. The CNN backbone we used is ResNet50, and the backbone inherits the pretrained weights on ImageNet and MS-Celeb-1M, denoted as ResNet50[1] and ResNet50[2], respectively. As illustrated in Table 4, the pre-trained weights are critical for CNN backbone, which boosts the performance from 0.7345 to 0.7785 on the validation set. This indicates that the CNN pre-trained on facial recognition datasets can extract more useful features for eye contact detection.



(a) Confusion matrix of the baseline method on the validation set.

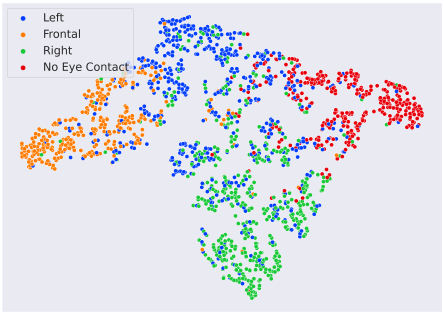


(b) Confusion matrix of our FSFNet on the validation set.

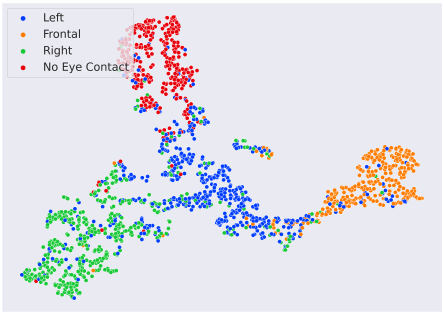
Figure 3: The confusion matrices of our method on the validation set. The diagonal values of each confusion matrix corresponds the accuracy of specific eye contact type. The darker the cell, the higher its accuracy.

Compared with the baseline method ResNet50[2], both the CA module and the FSF block increase the performance. The CA module brings a performance gain of 0.018 and 0.03 on the validation set and test set, respectively. We also introduce the exponential moving average (EMA) technique to maintain moving averages of the trained parameters. The decay rate for EMA is set to 0.999. As shown in Table 4, our FSFNet with EMA that use averaged parameters produces significantly better results than the final trained values.

Visualization Analysis: As shown in Figure 3, we illustrate the confusion matrices of the baseline ResNet50[1] and our FSFNet on the validation set. We can see from the confusion matrices that our FSFNet achieves promising results on these four different eye contact types. The visualization results in Fig. 4 also illustrates that our proposed method successfully drives the learned features to become closer within each class. Compared with the baseline method, our FSFNet achieves the clearer boundary between different classes and proves tighter intra-class feature representations.



(a) The feature distribution learned by the baseline method.



(b) The feature distribution learned by our FSFNet.

Figure 4: Visualization of the learned feature distribution by t-SNE[31] on the validation set. Note that we adopt the same color settings for all sub-figures.

5 CONCLUSION

In this paper, we present the FSFNet, a novel deep learning framework, for eye contact detection within multi-party conversations. Our approach integrates an adaptive feature selection mechanism within the Transformer architecture, ensuring focusing on relevant facial features. Specifically, we leverage the attention scores to dynamically select the most salient feature, which not only refines the model's feature representation ability but also significantly curtails computational costs. The design of our FSF blocks and the channel attention module achieves promising eye contact detection performance even under challenging real-world conditions. The experimental results on the MPIIGroupInteraction dataset demonstrate the effectiveness of our FSFNet, achieving the accuracy of 0.79 on the test set. Visualization analysis and the ablation studies provide insightful observations regarding the impact of token keep number and the block number. Our work contributes to the field by advancing eye contact detection technology, which holds profound implications for human-robot interaction and other applications.

6 ACKNOWLEDGMENTS

This work was supported by the National Natural Science Fund of China (62171183) and by Hunan Provincial Natural Science Foundation of China (2022JJ20017).

REFERENCES

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.

[3] Eunji Chong, Elysha Clark-Whitney, Audrey Southerland, Elizabeth Stubbs, Chanel Miller, Eliana L Ajodan, Melanie R Silverman, Catherine Lord, Agata Rozga, Rebecca M Jones, et al. 2020. Detection of eye contact with deep neural networks is as accurate as human experts. *Nature Communications* 11, 1 (2020), 6386.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4690–4699.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[6] Michael Dietz, Daniel Schork, and Elisabeth André. 2016. Exploring eye-tracking-based detection of visual search for elderly people. In *2016 12th International Conference on Intelligent Environments (IE)*. IEEE, 151–154.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[8] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2020. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412* (2020).

[9] Eugene Yujun Fu and Michael W Ngai. 2021. Using Motion Histories for Eye Contact Detection in Multiperson Group Conversations. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4873–4877.

[10] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 2014. EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 255–258.

[11] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*. Springer, 87–102.

[12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[13] Kun Li, Dan Guo, Guoliang Chen, Feiyang Liu, and Meng Wang. 2023. Data Augmentation for Human Behavior Analysis in Multi-Person Conversations. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9516–9520.

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[15] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. 2014. Adaptive linear regression for appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 10 (2014), 2033–2046.

[16] Fuyan Ma, Ziyu Ma, Bin Sun, and Shutao Li. 2022. TA-CNN: A Unified Network for Human Behavior Analysis in Multi-Person Conversations. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7099–7103.

[17] Fuyan Ma, Bin Sun, and Shutao Li. 2023. Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion. *IEEE Transactions on Affective Computing* 14, 2 (2023), 1236–1248. https://doi.org/10.1109/TAFFC.2021.3122146

[18] Fuyan Ma, Bin Sun, and Shutao Li. 2024. Transformer-Augmented Network With Online Label Correction for Facial Expression Recognition. *IEEE Transactions on Affective Computing* 15, 2 (2024), 593–605. https://doi.org/10.1109/TAFFC.2023.3285231

[19] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Anna Penzkofer, Dominik Schiller, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. 2024. MultiMediate'24: Multi-Domain Engagement Estimation. In *Proceedings of the 32nd ACM International Conference on Multimedia*.

[20] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2022. MultiMediate '22: Backchannel Detection and Agreement Estimation in Group Interactions. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM New York, NY, USA, 6 pages.

[21] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2021. MultiMediate: Multi-modal Group Behaviour Analysis for Artificial Mediation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4878–4882.

[22] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *23rd International Conference on Intelligent User Interfaces*. 153–164.

[23] Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. 2018. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 1–10.

[24] Philipp Müller, Ekta Sood, and Andreas Bulling. 2020. Anticipating averted gaze in dyadic interactions. In *ACM Symposium on Eye Tracking Research and Applications*. 1–10.

[25] Kazuhiro Otsuka, Keisuke Kasuga, and Martina Köhler. 2018. Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 191–199.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).

[27] Akshay Rangesh, Bowen Zhang, and Mohan M Trivedi. 2020. Driver gaze estimation in the real world: Overcoming the eyeglass challenge. In *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1054–1059.

[28] Yong-Goo Shin, Kang-A Choi, Sung-Tae Kim, and Sung-Jea Ko. 2015. A novel single IR light based gaze estimation method using virtual glints. *IEEE Transactions on Consumer Electronics* 61, 2 (2015), 254–260.

[29] Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. 2013. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 271–280.

[30] Qiya Song, Renwei Dian, Bin Sun, Jie Xie, and Shutao Li. 2023. Multi-scale Conformer Fusion Network for Multi-participant Behavior Analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9472–9476.

[31] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).

[32] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768* (2020).

[33] Xinming Wang, Jianhua Zhang, Hanlin Zhang, Shuwen Zhao, and Honghai Liu. 2021. Vision-based gaze estimation: A review. *IEEE Transactions on Cognitive and Developmental Systems* 14, 2 (2021), 316–332.

[34] Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D Abowd, and James M Rehg. 2012. Detecting eye contact using wearable eye-tracking glasses. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 699–704.

[35] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2017. Everyday eye contact detection using unsupervised gaze target discovery. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*. 193–203.

[36] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4511–4520.

[37] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 1 (2017), 162–175.