IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 February 2023, accepted 1 April 2023, date of publication 10 May 2023, date of current version 17 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3274744

RESEARCH ARTICLE

Transformer-Based Feature Fusion Approach for Multimodal Visual Sentiment Recognition Using Tweets in the Wild

FATIMAH ALZAMZAMI[1] AND ABDULMOTALEB EL SADDIK[1,2], (Fellow, IEEE)
[1]Multimedia Communication Research Laboratory, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[2]Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Corresponding author: Fatimah Alzamzami (falza094@uottawa.ca)

ABSTRACT We present an image-based real-time sentiment analysis system that can be used to recognize in-the-wild sentiment expressions on online social networks. The system deploys the newly proposed transformer architecture on online social networks (OSN) big data to extract emotion and sentiment features using three types of images: images containing faces, images containing text, and images containing no faces/text. We build three separate models, one for each type of image, and then fuse all the models to learn the online sentiment behavior. Our proposed methodology combines a supervised two-stage training approach and threshold-moving method, which is crucial for the data imbalance found in OSN data. The training is carried out on existing popular datasets (i.e., for the three models) and our newly proposed dataset, the Domain Free Multimedia Sentiment Dataset (DFMSD). Our results show that inducing the threshold-moving method during the training has enhanced the sentiment learning performance by 5-8% more points compared to when the training was conducted without the threshold-moving approach. Combining the two-stage strategy with the threshold-moving method during the training process, has been proven effective to further improve the learning performance (i.e. by $\approx 12\%$ more enhanced accuracy compared to the threshold-moving strategy alone). Furthermore, the proposed approach has shown a positive learning impact on the fusion of the three models in terms of the accuracy and F-score.

INDEX TERMS Transformers, ViT, sentiment, online social media, transfer learning, threshold moving, tweets, images, feature extraction, multimodality, fusion, big data, deep learning.

I. INTRODUCTION

People express their feelings, thoughts, and opinions in different visual ways, such as posting laughing faces, sunny beaches, or memes, on online social networks (OSNs), which contain diverse types of images. Images and visuals sometimes express feelings more clearly than words. On Twitter alone, tweets with images receive 89% more likes and 150% more retweets. In this work, we focus on exploiting different pieces of information within images. The aim is to analyze online social behavior (i.e., sentiment in this paper) as a support tool for textual-based analysis or to understand online social behaviors on online social networks (OSNs) when textual-based tools are limited or even unavailable. One major challenge of developing systems for OSNs is their uncontrolled content; users have the freedom to populate data under open circumstances in an unstructured manner. OSN users have the control to start a trend or share a thought, which makes social media an open platform that is not restricted to any particular domain. This has raised a need to build flexible models and systems that are able to adapt or generalize to different domains -which is the main objective of this work. This introduces a technical challenge that arises from the lack of available datasets that can be used to train models that can be adapted or generalized to different domains. Researchers tend to construct datasets according to their specific domain of interest. We recognized this issue in our earlier study [1] and attempted to resolve it by proposing Domain

The associate editor coordinating the review of this manuscript and approving it for publication was Barbara Guidi.

Free-Multimedia-Sentiment Dataset (DFMSD). Our findings in [1] revealed that specific domain sentiment models do not generalize well on different domain-specific data. Unlike domain-specific models, general sentiment models have been shown to adapt well to domain-specific sentiment datasets. Note that our previous models were designed for textual tweets only. In this paper, we investigate domain-independent sentiment extraction from visual content and conduct experimental evaluations using our image dataset (DFMSD). The images were collected under criteria-free conditions for use in models that work in the wild. The dataset contains different types of images that are categorized into three types: images containing faces, images containing text, and images containing no faces/text. Our assumption is that each image type contributes valuable information that is embedded within the images and can be used to enhance the learning of sentiment behavior on online social media. A more robust performance in image classification tasks can be achieved using visual transformers (ViTs), which have been shown to outperform CNNs [2], [3]. The attention mechanism has been shown to be the key element in achieving such high-performance robustness. ViT uses the attention mechanisms directly on image patches without depending on a CNN, where attention is either used to replace some components of the CNN or in conjunction with it (i.e., with the CNN). Furthermore, the multihead self-attention layer allows the ViT to embed information globally (i.e., attend to global features) across the overall image, which has been shown to improve the learning performance of image classification by four times compared to that of the CNN. Similar achievement is seen with textual-based transformers models (i.e., BERT-based models) [4], [5]. In this work, we utilize the transformer architecture to train three models, one for each image type, to extract high-quality features for each task (i.e., image type). Then, we use these features to model our final multimodal visual sentiment classifier. Transformers have been shown to perform best when trained on large-scale datasets [2], [5]. This means that transformers are able to generalize well on classification tasks when trained on large-scale datasets compared to when trained on small datasets. With the limitation of existing datasets, which are small in size, researchers have exploited the transfer learning approach to benefit from transformer models pretrained on large-scale datasets. Given that it is expensive to curate large-scale datasets of high quality in terms of time, cost, and human labor, transfer learning offers a reliable solution for learning classification tasks using small datasets. To overcome the differences between source and target tasks, the strategy of two-stage learning (i.e., finetuning) has been recommended in the literature [6], [7] to compensate for the limitation of small datasets.

Data imbalance is a common phenomenon in sentiment datasets collected through OSNs [1], [8], [9]. This is also observed in our DFSMD dataset. Although the literature [10] suggests that a balanced dataset would improve model learning, it is too expensive to balance the data while simultaneously preserving the natural distribution to avoid biases. To overcome the class imbalance issue in this work, we propose fusing the threshold-moving approach with the sentiment learning process. It has also been proven that the transformer architecture [4] is effective in dealing with class imbalance.

In this paper, we fuse the threshold-moving approach with the learning process conducted using transformer architecture as an attempt to address the class imbalance problem and to enhance the learning performance of our models. The learning of multimodal sentiment recognition requires two components: a feature extraction method and a fusion strategy. In this work, we adapt the multimodality approach in [11] and [12] and propose training three separate transformer-based deep models to extract features from three types of images (i.e., tasks): images containing faces, images containing text, images containing no faces/text. We adopt the intermediate fusion approach to fuse the extracted features and feed them to an MLP architecture to build our final multimodal sentiment model. To the best of our knowledge, we are the first to use a transformer-based fusion approach with pretrained transformer-based models to extract features for multimodal sentiment analysis on OSNs. Additionally, we believe that we are the first to fuse the threshold-moving approach with the learning process using the transformer architecture.

We summarize the contributions of this work as follows:

- A transformer-based two-stage learning strategy is utilized to alleviate the issue of our small sentiment dataset (DFMSD).
- The threshold-moving approach is fused with the learning process using the transformer architecture to solve the data imbalance of sentiment datasets.
- Three transformer-based models are developed using the proposed approach to extract deep features from DFMSD images to enhance the feature representation to be used for learning the visual sentiment model.
- A multimodality visual sentiment predictive model is designed and implemented using three types of deep features extracted based on our transformer-based pretrained models. Our DFMSD is used for modeling and evaluation in this work.

The rest of this paper is organized as follows. Section II presents details of the related work and Section III explains the datasets used for the modeling and analysis. Our proposed framework and methodology is presented in Section IV. Section V explains the experimental design and evaluation protocol along with the results and analysis. Finally, in Section VI we conclude our proposed work and findings and discuss future directions.

II. RELATED WORK

Facial expression recognition (FER) for online social networks (OSNs) in the wild is extremely challenging due to the uncontrolled condition of the images that can be shared. In addition to real faces, animated faces can be seen in image-like memes. Additionally, various head poses, occlusion, and

face deformation and blur may be observed for faces under unconstrained conditions. However, in the past few decades, great progress has been made in FER, where different learning methods that achieve good performance have been used. Some methods, such as SVM and Bayesian networks, require a preprocessing step to extract facial features before they are used for facial emotion classification, which adds substantial effort and computational overhead [13], [14], [15], [16]. In deep learning-based methods, both facial feature extraction and facial emotion classification are combined into one single stage, breaking the dependency on hand-crafted features [17], [18]. Convolutional neural networks (CNNs) have a natural inductive bias for learning feature representations from images and thus have shown promising performance in FER [19], [20], [21], [22]. However, CNN-based models can be sensitive to complex image backgrounds with occlusion or variant head poses [23]. Recent studies have shown that vision transformers (ViTs) are robust against image occlusion and disturbance [24], [25], [26], which justifies our decision to use ViTs as the backbone of our visual models to be used on OSN data.

Novel transformers [2] have become the state-of-the-art method in NLP tasks, and recently, they have been applied in computer vision tasks [27], [28]. Visual transformers have achieved remarkable performance in image classification tasks [2], [3] and outperformed CNNs in terms of computational efficiency and accuracy [2]. ViT was designed based on the attention mechanism, which has proven to be a key element for image classification to achieve high-performance robustness. ViT uses the attention mechanisms directly on a sequence of input image patches without depending on a CNN, where attention is either used in conjunction with the CNN or to replace some CNN components. When trained on a sufficient amount of data, ViT outperforms similar state-of-the-art CNNs with four times fewer computational resources and four times better efficiency and accuracy [2]. Unlike CNNs, which have small local receptive fields in each layer, the multihead self-attention layer allows the ViT to embed information globally (i.e., attend to global features) across the overall image. Moreover, the model learns to encode the relative location of image patches to reconstruct the image structure.

ViT has been shown to perform best when trained on large-scale datasets, which was manifested in its performance on ResNet against ImageNet [2]. This means that ViT is able to generalize well on image classification tasks when trained on large-scale datasets compared to when trained on small datasets. Researchers [23] can benefit from vision transformer models trained on large-scale datasets by exploiting a transfer learning approach that uses pretrained weights to finetune transformer architectures on smaller datasets. The authors in [23] adopted ViT transfer learning to learn an FER model through one-stage finetuning using pretrained weights from a transformer-based Deit-S model. Ma et al. [29] showed a positive impact of using pretrained weights (i.e., obtained from ImageNet-21K) when training their transformer-based FER model compared to when the training was conducted from scratch. While vision transformer-based FER models have been introduced, to the best of our knowledge, we could not find studies that applied vision transformers to visual sentiment classification tasks. Transfer learning has addressed the problem of small datasets (i.e., given that it is time, cost, and resource consuming to build large-scale manual annotated datasets for image classification problems, including FER and sentiment recognition). To further compensate for small datasets, existing studies suggest overcoming the difference between the source task and target task when using transfer learning through a two-stage finetuning strategy [30], [31]. In the first stage of the finetuning strategy, there is a learning shift from the source task to the target task, while in the second stage, the learning is refined in the target task [7]. The two-stage strategy has been shown to outperform one-stage finetuning on FER tasks that use small datasets [6], [7]. In this work, we adopt a two-stage strategy to build our ViT-based FER and sentiment recognition models.

Several efforts have been made to analyze sentiment and emotion using textual and visual modalities [32]. In the context of textual sentiment analysis, Alzamzami and El Saddik [4] attempted domain-independent sentiment analysis using a DL transformer network and showed that their model can be used to adapt to various domains, including sports and movie reviews. Image sentiment recognition is also an area of research. Sun et al. [33] designed an algorithm that discovers affective regions and supplements local features in images to improve the performance of visual sentiment analysis. Multimodal emotion and sentiment recognition has been an equally active research area in the last few years. Fortin et al. [12] proposed a multimodal architecture for emotion recognition systems, where predictions in the absence of one or two modalities are achieved by using a classifier for each combination of text, image, and tags. In another work by Xu et al. [34], an interplay of visual and textual content for sentiment recognition was modeled based on a comemory network.

The learning of multimodal sentiment recognition requires a feature extraction method and fusion strategy [35]. Most of the previous work on multimodal emotion and sentiment recognition uses low-level features (e.g., SIFT for visual modalities and GloVe for textual modalities) or deep features [35]. Features that are extracted from pretrained deep learning models are called deep features. They are extracted after a DL model is trained using a labeled dataset. In existing studies on facial recognition, deep features were extracted from pretrained facial recognition networks, and similarly pretrained text deep models have been used to extract text features for emotion and sentiment analysis [35]. Such studies highlight that deep features yield better performance than low-level features. CNN pretrained models are widely used for deep feature extraction due to their natural inductive bias. The authors in [12] used DenseNet-121 pretrained on

ImageNet to extract deep features for image, text, and tag models before they concatenated these features and fed them to two fully connected layers to obtain final predictions. Similarly, the authors in [11] used CNN-based pretrained models for deep feature extraction, VGG16 for image features and BalanceNet for textual features. A hybrid of intermediate and late fusion approaches was implemented based on CNNs to concatenate the features for the final sentiment prediction. In this work, we follow the multimodal approach in [11] and [12] and propose the use of three transformer-based pretrained deep models to extract features from images, faces and texts within the images. We adopt the intermediate fusion approach to fuse image, facial emotion, and textual features and feed them to an MLP architecture to build our multimodal sentiment classifier. To the best of our knowledge, we are the first to use a transformer-based fusion approach with three pretrained transformer-based models to extract features for multimodal sentiment analysis on OSNs.

III. DATASETS

This section presents popular datasets used for textual and visual sentiment analysis and facial emotion recognition on big data, in addition to our new sentiment multimedia dataset (DFSMD) [9].

- **Twitter for Sentiment Analysis (T4SA)** [36]: T4SA is a multimedia (i.e., texts and images) sentiment dataset that contains 1 million tweets with 1.5 million images. Texts were noisily annotated with three sentiment classes: positive, negative, and neutral. The images were annotated based on the sentiment associated with the texts. Due to the quality of the neutral class annotation observed during initial experiments, there was a confusion between the neutral class and both positive and negative classes. As a result, we removed the neutral class from T4SA dataset for training purposes. The T4SA dataset is used for the first stage finetuning of the basic visual sentiment modeling in this work.
- **FER-2013** [37]: FER-2013 is a well-known facial emotion recognition dataset that has been used extensively in FER models and applications. It consists of 35K images that are grayscale with 48*48 pixels. The facial images were manually annotated with seven emotions: anger, disgust, fear, happy, sad, surprise, and neutral. The FER-2013 dataset was used to build our FER model.
- **AffectNet** [38]: AffectNet is currently the largest manually annotated facial emotion recognition in-the-wild dataset. It consists of 1 M facial images in greyscale with 48*48 pixels. 44 K of which were manually annotated with eight facial emotions: neutral, happy, sad, surprise, fear, disgust, anger, and contempt. Since we are studying sentiment in-the-wild on social media, animated and cartoon images exist, especially in memes. Therefore, we converted AffectNet images into animated versions and combined them with the original images. The modified version of the AffectNet dataset was used for the first stage finetuning of our facial emotion recognition modeling in this work. Note that we converted the images of AffectNet to grayscale since we used FER-2013 data to build our final FER model.
- **DFMSD** [9]: The domain-free multimedia sentiment dataset (DFMSD) is our newly introduced dataset for visual and textual sentiment analysis. It was designed and constructed to work in the wild and to be able to deal with uncontrolled conditions in online social media. Data in DFSMD was collected using Twitter Stream API. The protocol followed to collect and annotate DFSMD distinguishes it from other datasets, as the data collection process was not restricted to any keywords, domains, locations, or predefined retrieval criteria. The annotation questions and dataset annotators were selected carefully to minimize any possible biases during the annotation. Moreover, the annotators were selected on the basis of sentiment agreement using three expert psychologists. The DFSMD consists of 14,488 tweets that contain 10,244 images; 46% (i.e., 6683 tweets) of the tweets are positive, 33% (i.e., 4822) are negative, and 21% (i.e., 2983) are neutral. The image distribution is as follows: 47% belonging to the positive class, 10% belonging to the negative class, and 43% belonging to the neutral class. Note that texts and images were annotated separately in a way that does not affect the annotation of the images. We decided to extend our sentiment image dataset by following the same collection and annotation approach used earlier as an attempt to improve the deep learning performance and to minimize the problem of a severe class imbalance. The first version was published in an earlier study [9].

IV. METHOD

This section presents in detail the method we follow to model our multi-model visual social behavior analyzer; it includes data preprocessing, datasets, and all the approaches we adopted for building our final model.

A. PREPROCESSING

The preprocessing step is very important for the learning process; we cleaned, denoised, and prepared the data before we fed them to the VIT for training. Image preprocessing consists of the following steps:

- Face detection: We used the face detection algorithm proposed in [39] that uses facial keypoints to detect faces. The face detector finds four coordinates for the region of interest (ROI) of faces. Then, the detected faces are cropped, and all irrelevant background is discarded. Additionally, faces that are far away or not clear, with respect to the ratio of the face and the image size, are discarded. If the ratio is less than a predefined threshold with respect to the image, and if the value is less than a predefined threshold, we discarded the face. This step was applied to the facial emotion recognition modeling part.

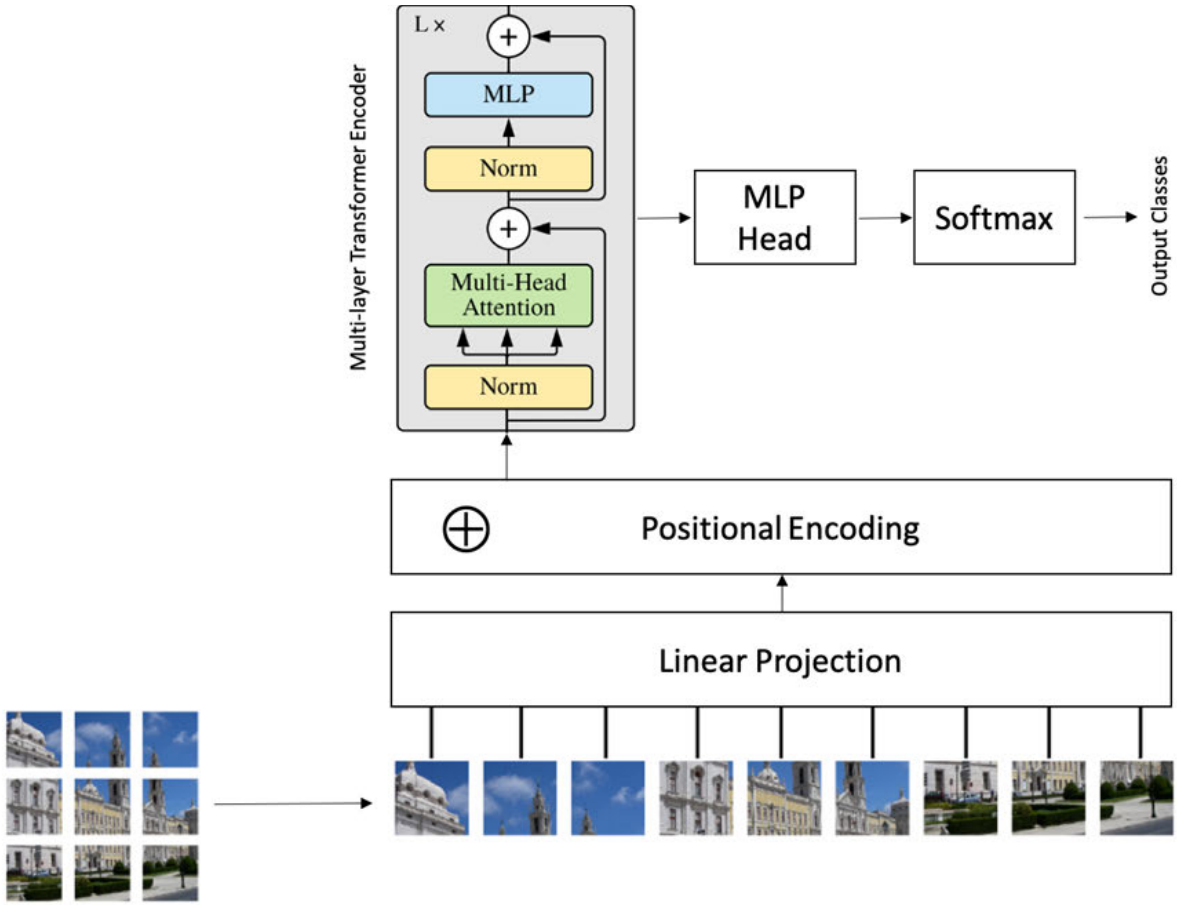

FIGURE 1. The architecture of visual Transformers [2] used in training our models.

- Color to grayscale conversion. This step was applied in the facial emotion recognition modeling part.
- Data augmentation: The size of the dataset is increased, and the imbalance between classes is addressed since deep learning works better with more data. We used 2 types of augmentation, one for training and another for testing:
  - Training time: Images were randomly resized in the range of (224,350) and then center-cropped by a cropping size of (224, 224). Then, the images were randomly flipped.
  - Testing time: Images were resized to (256,256) and then cropped using the 10-crop technique. Using ten-crop technology, an image is resized to (256,256), and 5 crops (upper-left, upper-right, lower-left, lower-right, center) with a cropping size of (224,224) are made. Then, L-R flipping is applied, resulting in 10 cropped-flipped images. Finally, we used the average prediction of these 10 images [40]. This step was applied in the basic visual sentiment modeling and facial emotion recognition modeling parts.
- Text detection and extraction: We used optical character recognition (OCR) to detect and recognize texts from images. Extracted words are not in sorted order after OCR extraction; hence, we sorted the extracted words in order of their occurrence using contour detection to separate the different lines. Then, we simply processed the contours left to right to sort the words within lines. Finally, we applied quantization on heights with a predefined threshold to group words in the same line together. This step was applied in the textual-images part.

B. VISUAL TRANSFORMERS MODEL

Visual transformer (ViT) [2], which was introduced in 2020, is used as the deep learning architecture in our work. ViT has been a competitive alternative to CNNs for image recognition tasks. It outperforms the current state-of-the-art CNNs by four times in terms of computational efficiency and accuracy [2], especially on big data regimes. In big data regimes, the inductive biases of CNNs are not needed; instead, ViT can learn those biases by itself. Shallower layers of the ViT are able to localize attention (i.e., attend to local pixels) and globalize attention (i.e., attend to global pixels)

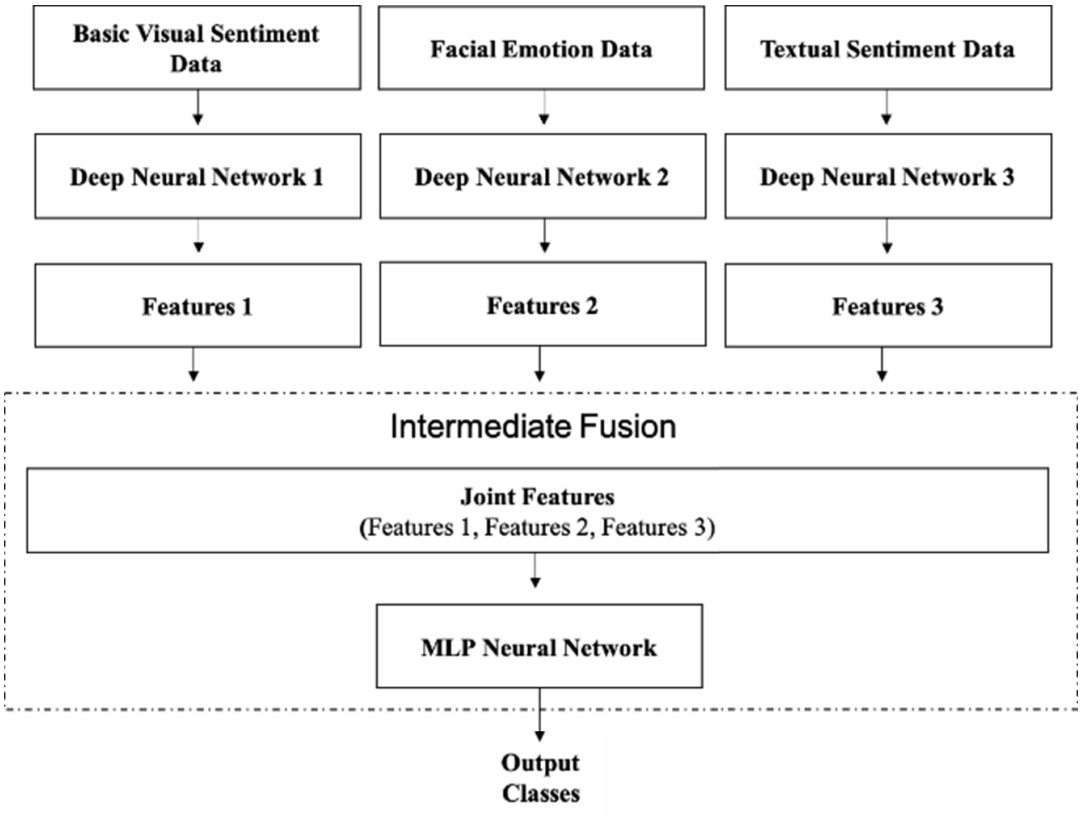

FIGURE 2. The proposed architecture for ViT-based multi-modality fusion for visual online social behavior analysis.

compared to CNNs, which only have local small receptive fields in shallower layers. The advantage of shallower layers being able to attend to local and global pixels in images is that it allows the ViT model to learn how and when to attend, and the bias is no longer needed. Because of its efficiency in handling big data regimes [2], it can be widely used in systems and applications related to online social networks—where data sizes are very large—with remarkable performance.

A high-level overview of the ViT architecture is given in Figure. 1. An input image is split into patches and then flipped in order and flattened out. A linear projection is applied to the flattened patches, and then positional encoding is added before the data are fed to the multilayer transformer encoder. The structure of a single encoder consists of a multi-head attention module and multilayer perceptron module. The output of the encoder is then fed to a multilayer perceptron head, which is used as a classification module that yields class predictions.

C. THRESHOLD-MOVING

The default decision threshold (i.e., 0.5) for classification problems with class imbalance might negatively impact the learning performance and hence yield poor results. The default decision threshold might not represent the optimal interpretation of a model's predicted probabilities. As such, a simple approach to improve the classification performance on imbalanced data is to tune the hyperparameter (i.e., threshold) that is used to map the predicted probabilities to class labels. The process of tuning this hyperparameter is called threshold moving. In this work, we calculated the optimal threshold using a grid search approach. We searched threshold values for a model and considered the best value, i.e., the value that yielded the best performance in terms of our evaluation metric. We applied threshold moving during the training process in such a way that in each validation iteration (i.e., epoch), we examined a range of threshold values on the predicted class probabilities to find the best threshold. The threshold that achieved the best performance (i.e., in terms of the evaluation metric) was then adopted for the model (i.e., at the current iteration or epoch). Upon the completion of the training, the model with the best performance was chosen to make predictions on new data.

D. TWO-STAGE STRATEGY

Transfer learning offers a rich set of benefits, including improving the efficiency of model training and saving time and resources, since building a high-performance model from scratch requires a large amount of data, time, resources, and effort. Therefore, we used the two-stage learning approach [6] to train ViT models as an attempt to solve the limitation of small labeled datasets. We implemented the first-stage

finetuning step on very large datasets to maximize the benefits of transfer learning. We used a ViT architecture with pretrained weights from the ImageNet-21K dataset [2]. The whole ViT architecture was retrained for the first stage of finetuning. For the basic sentiment model, we used the TS4A dataset to finetune the ViT pretrained model. The model learns two sentiment classes: positive and negative. To overcome the class imbalance between the two classes, we fused the threshold-moving approach with the sentiment learning process in this part. For the facial emotion expression model, we finetuned the pretrained ViT model using the modified version of the AffectNet dataset. The modified version of the AffectNet dataset includes its original images in addition to the same images converted into animated versions. The model learns eight classes: neutral, happy, sad, surprise, fear, disgust, anger, and contempt.

For the second-stage finetuning step, we initialized the ViT architecture with the weights obtained from our first-stage finetuning step. The last fully connected layer was replaced by a new MLP head for both models: basic sentiment and facial emotion models.The basic sentiment classifier outputs two classes: positive and negative. Our DFSM dataset was used for the second-stage fine-tuning step by adopting the threshold-moving approach during the process of sentiment learning on the positive and negative classes. The finetuning step in the second stage was performed by training the whole ViT architecture. We took the output of the last layer (i.e., the high-level representations of the mode) and fed it as features to our visual multimodal sentiment classifier. The facial emotion classifier outputs seven classes: anger, disgust, fear, happy, sad, surprise, and neutral. The FER-2013 dataset was used to finetune the AffectNet pretrained model after the first-stage finetuning step. The finetuning step in the second stage was performed by training the whole ViT architecture. We took the output of the last layer (i.e., the high-level representations of the mode) and fed it as features to our visual multimodal sentiment classifier.

E. VISUAL DEEP MULTIMODAL FUSION

We adopted the architecture of intermediate fusion to fuse our deep learning-based models with the goal of building a multimodality online social behavior model. Figure. 2 shows an illustration of our proposed architecture for developing a multimodal online social behavior classifier (i.e., a sentiment case study in this work). Intermediate fusion allows for data fusion at different stages of model learning as it offers flexibility to fuse features at different depths. Deep learning-based multimodal data fusion has shown great improvement in learning performance [12], [35]. The input for the intermediate fusion is the higher-level representations (i.e., features) obtained through multiple layers of deep learning. Hence, the intermediate fusion in the context of multimodal deep learning is the simultaneous fusion of different model representations into a hidden layer so that the model learns a representation from each of the individual models. The layer where fusion is performed is called the fusion layer. In this work a ViT-based fusion is proposed for a multimodality visual online social behavior analysis. Three models, namely: single-modality sentiment, facial emotion, and textual sentiment models are trained using Transformer backbones (i.e., ViT for visual content and BERT for textual content). Then these models are used to extract deep features before all are fused to form one joint feature that will be fed into an MLP classification head (Figure. 2).

V. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental results and analysis for the visual models. The first-stage fine-tuning step of the ViT architecture was implemented with pretrained weights obtained from the ImageNet-21K dataset for both single-modality sentiment and FER models. Note that single modality model refers to the images as they are without extracting facial nor textual features.

A. PERFORMANCE OF THE BASIC VISUAL SENTIMENT MODEL

1) FIRST STAGE FINETUNING

We implemented the first-stage fine-tuning step of ViT architecture using T4SA dataset. Due to the poor quality annotation of the neutral class (i.e., after preliminary experimentation), we decided to train the model on the positive and negative classes. Table. 1 shows the performance of our ViT-based single-modality sentiment model in the first-stage fine-tuning step. Table. 1 shows the results of the performance when fusing threshold-moving with the training. Threshold-moving has been shown to absolutely enhance the learning performance by six points in terms of accuracy, eight points in terms of the positive F-score, and five points in terms of the negative class. This model was used for image generic deep feature extraction to learn our multimodality visual sentiment classifier.

Note that we present the learning performance (i.e., without threshold moving applied) for three classes. During the training, it was observed that the model confused the neutral with both positive and negative classes.

2) SECOND STAGE FINETUNING

In Table. 2, we demonstrate the performance of our second-stage ViT model using the images from our DFMS dataset. Based on the first-stage pretrained model, The performance accuracy of learning two classes (i.e., positive and negative classes) i.e., with the threshold-moving technique fused during the training is 81% with F-scores of 0.86, 0.7 for the positive and negative classes, respectively. Further, the results of the second-stage fine-tuning step have shown the effectiveness of the two-strategy fine-tuning approach on learning three classes (positive, negative, and neutral); the learning performance greatly improved by 12 points in terms of accuracy, 29 points in terms of positive F-score, and 8 for

TABLE 1. The performance of first-stage single-modality ViT sentiment model on T4SA dataset.

	Precision			Recall			F-score			Accuracy
	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	
Binary classes	0.64	0.61	-	0.58	0.67	-	0.61	0.64	-	63
Binary classes with threshold moving	0.69	0.69	-	0.69	0.69	-	0.69	0.69	-	69
3 classes	0.51	0.48	0.47	0.37	0.54	0.55	0.43	0.51	0.51	49

TABLE 2. The performance of second-stage single-modality ViT sentiment model on images from our DFMSD dataset.

	Precision			Recall			F-score			Accuracy
	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	
Binary classes with threshold moving	0.88	0.67	-	0.84	0.73	-	0.86	0.7	-	81
3 classes	0.77	0.66	0.36	0.67	0.54	0.55	0.72	0.59	0.44	61

the negative F-score. We observe that the performance of the neutral class decreased in terms of precision which explains the model confusion with the other classes. Overall, the model performs well in distinguishing between positive and negative classes since it has high positive F-score scores > 0.70 and negative F-score scores ≈ 0.60 for both classes and that in the presence of the neutral class.

B. PERFORMANCE OF THE FACIAL EMOTION MODEL

To train all FER models, all faces have be to be detected and cropped as explained in the preprocessing section.

1) FIRST STAGE FINETUNING

We implemented the first-stage finetuning step of the ViT architecture using the AffectNet dataset. An accuracy of 59% and an F-score of 0.59 were obtained for 8 classes using this model.

2) SECOND STAGE FINETUNING

Based on the weights obtained from the first-stage ViT fine-tuning step using AffectNet, we implemented the second-stage fine-tuning step with the pretrained weights obtained from the first stage using the FER-2013 dataset. Table. 3 illustrates the effectiveness of the two-stage strategy in enhancing the learning performance—in terms of precision, recall, and accuracy—between multiple classes. This can be obviously observed in the recall of the model, which greatly improved by 7 points when applying the two-stage strategy compared to when only using one-stage strategy for finetuning. This model was used for facial emotion deep feature extraction to learn our multimodal visual sentiment classifier.

TABLE 3. The performance of the second-stage ViT FER model on the FER-2013 dataset.

	Precision	Recall	F-score	Accuracy
FER2013- one-stage	0.68	0.62	0.64	69
FER2013- two-stage	0.71	0.7	0.7	70

C. PERFORMANCE OF VISUAL MULTI-MODAL FUSION

Table. 4 shows the effect of using the extra information of facial emotion and texts residing in images, in addition to the information from the images themselves. Fusing ViT features from our FER and single-modality models resulted in a slight improvement in the performance compared to the performance of the single-modality ViT sentiment model, in terms of accuracy and F-score for both positive and negative classes. While fusing single-modality sentiment and FER features noticeably improved the negative precision (i.e., by 6 points), the recall of the positive class decreased (i.e., by 4 points). However, fusing textual and facial emotion features along with the single-modality sentiment stabilizes the learning and further improves the overall performance for all the classes in terms of F-score. In more detail, negative precision improved by 4 points without affecting the positive recall and similarly positive recall improves by 3 points without affecting the negative recall.

We further examined the effect of fusing facial emotion and text features with the single-modality sentiment on three classes: positive, negative, and neutral classes. It can be seen from Table. 4 that the overall F-score of the model improved especially for the negative and neutral classes when fusing

TABLE 4. Performance of fusing three types of ViT-based deep features extracted from three pretrained models: single-modality sentiment, facial emotion, and textual sentiment. The performance is evaluated in terms of accuracy, precision, recall, and F-score.

	Precision			Recall			F-score			Accuracy
	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	
Single-modality Sentiment model (ViT)	0.88	0.67	-	0.84	0.73	-	0.86	0.7	-	81
MLP (Sentiment + FER ViT deep features) - 2 classes	0.86	0.73	-	0.88	0.69	-	0.87	0.71	-	82
MLP (Sentiment + FER + text ViT+BER Tdeep features) - 2 classes	0.88	0.71	-	0.87	73	-	0.87	0.72	-	82
MLP (Sentiment + FER ViT deep features) - 3 classes	0.73	0.66	0.43	0.77	0.54	0.44	0.75	0.59	0.44	64
MLP (Sentiment + FER + text ViT+BERT deep features) - 3 classes	0.75	0.64	0.42	0.73	0.57	0.5	0.74	0.6	0.46	64

the three types of features compared fusing to only two type of features.

VI. CONCLUSION AND FUTURE WORK

Inspired by emojis, which speak a universal language through iconic expressions, we adopted a supplementary language-independent approach, where a universal language of visual images and emotions is used to model online social behavior and sentiment, in this work. Accordingly, we proposed a multimodality classifier that leverages three types of images in parallel: images with text, images with faces, and images without faces or text. The corresponding experiments show that exploiting facial emotion and textual information extracted from images contributes to enhancing the learning of visual online social sentiment. Note that sentiment behavior has been considered as a case study to examine our proposed approach.

It has been found that fusing the threshold-moving approach with the learning process using the transformer architecture has shown a great improvement in the learning performance in terms of the F-score and accuracy. It also shows robustness in handling the class imbalance problem found in OSN data. In addition, the two-stage finetuning strategy is shown to work with the transformer architecture, confirming its robustness in solving the problem of small and insufficient datasets for both binary and multiclass classification tasks. This finding is consistent with the two-stage learning strategy used with the CNN architecture [6], [7].

Our proposed transformer-based feature extraction approach has been shown to be effective in the learning of multimodal visual sentiment using social media data. The high F-score and accuracy for both binary and multiclass sentiment predictions are evidence that strong feature representations were obtained using our proposed approach. Additionally, it is evident that the consideration of image types (i.e., to extract sentiment information) is successful because it actually contributes to the learning of sentiment from visual images. Given the proposed approach, the data used for training, and results, our proposed model could eventually be used for handling the images in the wild of social media independently of particular domains they belong to.

In terms of future directions, we are interested in exploring extra types of features that can contribute to improving the learning of online sentiment behavior. We are also interested in incorporating more types of online social behaviors, such as hate speech. Furthermore, we plan to develop a lighter version of the proposed model and deploy it in mobile smart devices. The current version of our model is designed with high resource requirements for computations and memory, which makes it unsuitable for deployment on portable smart devices such as smartphones.

REFERENCES

[1] F. Alzamzami, M. Hoda, and A. El Saddik, "Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation," *IEEE Access*, vol. 8, pp. 101840–101858, 2020.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[3] L. Meng, H. Li, B.-C. Chen, S. Lan, Z. Wu, Y.-G. Jiang, and S.-N. Lim, "AdaViT: Adaptive vision transformers for efficient image recognition," 2021, *arXiv:2111.15668*.

[4] F. Alzamzami and A. El Saddik, "Monitoring cyber SentiHate social behavior during COVID-19 pandemic in North America," *IEEE Access*, vol. 9, pp. 91184–91208, 2021.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[6] H. Wang, D. P. Tobon V., M. S. Hossain, and A. El Saddik, "Deep learning (DL)-enabled system for emotional big data," *IEEE Access*, vol. 9, pp. 116073–116082, 2021.

[7] Y. Miao, H. Dong, J. M. A. Jaam, and A. E. Saddik, "A deep learning system for recognizing facial expression in real-time," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 2, pp. 1–20, May 2019.

[8] V. Athanasiou and M. Maragoudakis, "A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: A case study for modern Greek," *Algorithms*, vol. 10, no. 1, p. 34, 2017.

[9] R. Abaalkhail, F. Alzamzami, S. Aloufi, R. Alharthi, and A. El Saddik, "Affectional ontology and multimedia dataset for sentiment analysis," in *Proc. Int. Conf. Smart Multimedia*. Cham, Switzerland: Springer, 2018, pp. 15–28.

[10] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *J. Artif. Intell. Res.*, vol. 19, no. 1, pp. 315–354, Jul. 2003.
[11] P. Kumar, V. Khokher, Y. Gupta, and B. Raman, "Hybrid fusion based approach for multimodal emotion recognition with insufficient labeled data," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 314–318.
[12] M. Pagé Fortin and B. Chaib-draa, "Multimodal multitask emotion recognition using images, texts and tags," in *Proc. ACM Workshop Crossmodal Learn. Appl.*, Jun. 2019, pp. 3–10.
[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.
[14] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *Proc. IEEE Int. Conf. Image Process.*, 2005, p. 370.
[15] X. Feng, M. Pietikainen, and A. Hadid, "Facial expression recognition with local binary patterns and linear programming," *Pattern Recognit. Image Anal. C/C Raspoznavaniye Obrazov I Analiz Izobrazhenii*, vol. 15, no. 2, p. 546, 2005.
[16] I. Buciu and I. Pitas, "Application of non-negative and local non negative matrix factorization to facial expression recognition," in *Proc. 17th Int. Conf. Pattern Recognit.*, 2004, pp. 288–291.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[19] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2852–2861.
[20] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.
[21] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.
[22] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3510–3519.
[23] H. Li, M. Sui, F. Zhao, Z. Zha, and F. Wu, "MVT: Mask vision transformer for facial expression recognition in the wild," 2021, *arXiv:2106.04520*.
[24] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," 2021, *arXiv:2105.10497*.
[25] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," 2021, *arXiv:2103.16854*.
[26] Q. Huang, C. Huang, X. Wang, and F. Jiang, "Facial expression recognition with grid-wise attention and visual transformer," *Inf. Sci.*, vol. 580, pp. 35–54, Nov. 2021.
[27] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," 2021, *arXiv:2102.04378*.
[28] K. Mahmood, R. Mahmood, and M. van Dijk, "On the robustness of vision transformers to adversarial examples," 2021, *arXiv:2104.02610*.
[29] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Trans. Affect. Comput.*, early access, Oct. 26, 2021, doi: 10.1109/TAFFC.2021.3122146.
[30] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 443–449.
[31] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 173–189, Jun. 2016.
[32] A. Ortis, G. M. Farinella, and S. Battiato, "Survey on visual sentiment analysis," *IET Image Process.*, vol. 14, no. 8, pp. 1440–1456, Jun. 2020.
[33] M. Sun, J. Yang, K. Wang, and H. Shen, "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.
[34] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020.
[35] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.
[36] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 308–317.
[37] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, and D.-H. Lee, "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2013, pp. 117–124.
[38] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," 2017, *arXiv:1708.03985*.
[39] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial Landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.

FATIMAH ALZAMZAMI received the M.Sc. degree in computer science from the Faculty of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada, where she is currently pursuing the Ph.D. degree in computer science. Her research interests include machine learning, deep learning, big data, social multimedia analysis, and mining. She is supervised by Prof. Abdulmotaled El Saddik.

ABDULMOTALEB EL SADDIK (Fellow, IEEE) is currently a Distinguished Professor with the School of Electrical Engineering and Computer Science, University of Ottawa, and a Professor with the Mohamed bin Zayed University of Artificial Intelligence. He has supervised more than 150 researchers. He has coauthored ten books and more than 550 publications and chaired more than 50 conferences and workshops. He received research grants and contracts totaling more than $22 million. His research interests include the establishment of digital twins to facilitate the wellbeing of citizens using AI, the IoT, AR/VR, and 5G to allow people to interact in real time with one another and with their smart digital representations in the metaverse. He is a fellow of the Royal Society of Canada, Engineering Institute of Canada, and Canadian Academy of Engineers. He is an ACM Distinguished Scientist. He received several international awards, such as the IEEE I&M Technical Achievement Award, the IEEE Canada C. C. Gotlieb (Computer) Medal, and the A. G. L. McNaughton Gold Medal for important contributions to the field of computer engineering and science.