Leveraging Semantic Scene Characteristics and Multi-Stream Convolutional Architectures in a Contextual Approach for Video-Based Visual Emotion Recognition in the Wild

Ioannis Pikoulis, Panagiotis P. Filntisis and Petros Maragos
School of ECE, National Technical University of Athens, Athens 15773, Greece
pikoulis.giannis@gmail.com, filby@central.ntua.gr, maragos@cs.ntua.gr

***Abstract*— In this work we tackle the task of video-based visual emotion recognition in the wild. Standard methodologies that rely solely on the extraction of bodily and facial features often fall short of accurate emotion prediction in cases where the aforementioned sources of affective information are inaccessible due to head/body orientation, low resolution and poor illumination. We aspire to alleviate this problem by leveraging visual context in the form of scene characteristics and attributes, as part of a broader emotion recognition framework. Temporal Segment Networks (TSN) constitute the backbone of our proposed model. Apart from the *RGB* input modality, we make use of dense *Optical Flow*, following an intuitive multi-stream approach for a more effective encoding of motion. Furthermore, we shift our attention towards skeleton-based learning and leverage action-centric data as means of pre-training a Spatial-Temporal Graph Convolutional Network (ST-GCN) for the task of emotion recognition. Our extensive experiments on the challenging Body Language Dataset (BoLD) verify the superiority of our methods over existing approaches, while by properly incorporating all of the aforementioned modules in a network ensemble, we manage to surpass the previous best published recognition scores, by a large margin.**

I. INTRODUCTION

The interpretation, perception and recognition of human affect has been the subject of rigorous studies and analyses across several scientific disciplines such as biology, psychology, sociology, neurology and last but not least, computer science. While the aforementioned cognitive sciences focus on the extraction of the available affective information, the fields of computer vision and machine learning aim at automating the recognition process through the development of novel techniques and algorithms which are capable of producing effective and robust encodings of such information.

The majority of past efforts in visual emotion recognition have been mostly limited to the analysis of facial expressions [37], [11], [14], [25], [44], while some studies have either incorporated information relative to body pose [12], [6], [2], [9] or have attempted to perform emotion recognition exclusively on the basis of body movements and gestures [1], [15], [28], [31], [32]. While some of these approaches perform well in certain specified settings, they fail to interpret real-world scenarios. This is because emotion recognition systems are, more often than not, expected to operate on instances of people whose facial features are fully visible and their body joints are unoccluded, something which does not generally conform to reality.

Evidence from psychology related studies suggest that visual context, in addition to facial expression and body pose, provides important information to the perception of people's emotions. Dudzik *et al.* [8] propose two sources of context as means of interpreting emotional behavior, namely *perceiver knowledge/experience* and *perceivable encoding context*. Wieser and Brosch [39] highlight situational context as the primary aspect of the latter, with features that mainly revolve around the visual scenes in which the depicted emotional behaviors are embedded. Barrett and Kensinger [3] report that the structural features of the face, when viewed in isolation, often prove to be insufficient for perceiving emotion. Furthermore, empirical findings suggest that the categorization of facial expressions is sped up at the sight of congruent scenes [29], while both positive and negative contexts result in significantly different ratings of faces compared with those presented in neutral contexts [24].

In this work, we aim at extending the concept of context-based visual emotion recognition in the dynamic setting of video sequences. Our approach to the problem rests on the late fusion of Temporal Segment Networks (TSN) [38] and a Spatial-Temporal Graph Convolutional Network (ST-GCN) [41]. We extend the original TSN framework by incorporating multiple input streams that encode bodily, contextual, facial and generic scene-related features, enhancing our model's joint understanding of emotion and the depicted surrounding environments. To the best of our knowledge, our approach is the first to explicitly infuse scene characteristics as well as make effective use of multi-stream optical flow in an emotion recognition process. Extensive ablation experiments, based on the recently assembled and challenging Body Language Dataset (BoLD), are carried out so as to study the various contributions of our methods.

The remainder of the paper is structured as follows: Firstly, we provide an overview of the most notable related work in the domain of context and skeleton-based visual emotion recognition. Subsequently, we analyze our proposed model architecture. Next, we present our experimental results on the BoLD dataset, followed by conclusive remarks.

II. RELATED WORK

Kosti *et al.* [17] made one of the first notable contributions towards context-based emotion recognition by introducing the EMOTIC dataset as well as providing a baseline model that was trained and evaluated on the latter. Their baseline model consisted of three modules: two ConvNet feature extractors (one for each of the body and context input streams) and one fusion network. A notable improvement

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

Proc. 16th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India (Virtual Event), Dec. 2021

in recognition performance over the baseline model came along the EmotiCon framework, as it was introduced by Mittal *et al.* [23]. Their main contributions are associated with the incorporation of multiple modalities in the task of context-based emotion recognition, including the face, pose, inter-agent interactions and socio-dynamic context.

Subsequently, researchers shifted their attention towards video-based emotion recognition in context. Lee *et al.* [19] introduced the Context-Aware Emotion Recognition benchmark which is comprised of 13.201 TV video clips and a total of 1.1M frames. Moreover, a baseline model was proposed, featuring a face and a context encoding stream which were merged using an adaptive-fusion network.

After the recent assemble of the Body Language Dataset (BoLD), Luo *et al.* [22] furthered their contributions by comparing various network configurations and finally providing a baseline model for the task of categorical and continuous emotion prediction. Among the examined methodologies were: motion-based descriptors, i.e. histograms of optical flow and motion boundary histograms, skeleton-based learning through a ST-GCN and Laban Movement Analysis [18] as well as pixel-level learning through two-stream convolutional and TSN architectures. Filntisis *et al.* [10] proposed the incorporation of a contextual feature encoding branch and the addition of visual-semantic embedding loss based on Global Vectors (GloVe) [27] word embeddings, achieving state-of-the-art performance on BoLD.

As graph-based neural networks have proven to be powerful tools for determining human actions [41], [34], [20], [35], [21], there have also been attempts to adapt them for the task of emotion recognition. Bhattacharya *et al.* [4] introduced a classifier network for the task of emotion recognition through gaits, as well as a realistic gait generator, with the ST-GCN architecture being the common denominator between the two. Sheng *et al.* [33] proposed an Attention Enhanced Temporal Graph Convolutional Network (AT-GCN) as part of a multi-tasking framework which can jointly learn representations relative to both emotion and identity recognition.

III. MODEL ARCHITECTURE

A complete schematic diagram of our proposed network ensemble is shown in Fig. 1. The backbone of our network implementation resides in a combination of the two-stream convolutional [36] and TSN [38] architectures, both of which were initially proposed for video-based action recognition. During TSN training, any given input video sequence $\mathcal{V}$ is firstly divided into K segments $\{S_1, ..., S_K\}$ of equal durations. The TSN operates on a set of K snippets, with each snippet constituting an instance that has been randomly sampled from the corresponding segment. More formally, the output of a temporal segment network is modeled as follows:

$$\text{TSN}(T_1, ..., T_K) = \mathcal{H}\Big(\mathcal{G}\big(\mathcal{F}(T_1; \mathbf{W}), \ldots, \mathcal{F}(T_K; \mathbf{W})\big)\Big) \quad (1)$$

where $\{T_1, ..., T_K\}$ denote the snippets, $\mathbf{W}$ denotes the network trainable parameters, $\mathcal{F}$ denotes the snippet-level network predictions, $\mathcal{G}$ denotes a segmental consensus function and $\mathcal{H}$ denotes a prediction function.

Firstly, we will present the TSN structure regarding the *RGB* modality, along with our proposed extensions for the enhancement of emotion understanding. Next, we will do the same for the *Optical Flow* modality, and finally we will present the part of the architecture relative to skeleton-based learning. We choose to utilize 18-layer ResNets [13] as our primary feature extractors for all the subsequent convolutional branches. ResNets constitute state-of-the-art ConvNet backbones, offering a valuable trade-off between performance and computational complexity. Moreover, the ResNet-18 variant produces 512-dim deep feature vector representations for each given input image.

A. TSN-RGB

1) Body: A single RGB image usually encodes static appearance at a specific point in time, lacking the contextual information about previous and next frames. We begin by training a standard TSN using the *RGB* modality and the body crops of each frame instance. For the calculation of the necessary body bounding boxes, we make use of the coordinates of 18 body joints that have been successfully tracked along the entirety of each video sequence and are being provided by the distributors of the dataset. The body branch feature extractor is pre-trained on ImageNet [7].

2) Context: We incorporate a context stream in the form of RGB frames whose primary depicted agents have been masked out. For the acquisition of the masks we use the body bounding boxes that we have previously calculated and multiply them element-wise with a constant value of zero. Contextual feature extraction is a scene-centric task, and therefore we choose to initialize the corresponding ConvNet backbone using the Places365-Standard [45], a large-scale database of photographs, labeled with scene categories.

3) Face: We introduce an input stream which explicitly operates on extracted face crops. For the localization and extraction of faces we use the 2D coordinates that correspond to the detected eyes, ears and nose of each depicted instance. These facial landmarks are used for the calculation of a bounding box that effectively contains the head of each primary depicted agent. As the pose of an agent might result in partial or complete occlusion of their facial features, the successful extraction of the face region is not guaranteed. The ConvNet backbone of the face branch receives manual pre-training on AffectNet [25] which constitutes the largest facial expression database, containing over 1M images, annotated on both categorical and dimensional level.

4) Scenes and Attributes: The depicted scene along with its attributes hold valuable information relative to human emotion understanding, especially in cases where primary sources of affective information, such the face and body, are occluded. Therefore, we aspire to enrich our model's perception of context by directly extracting the Places365 scene-specific scores and the corresponding Scene UNderstanding (SUN) [26] attributes through an 18-layer Wide-ResNet [43] which has been jointly pre-trained on both of the aforementioned databases.

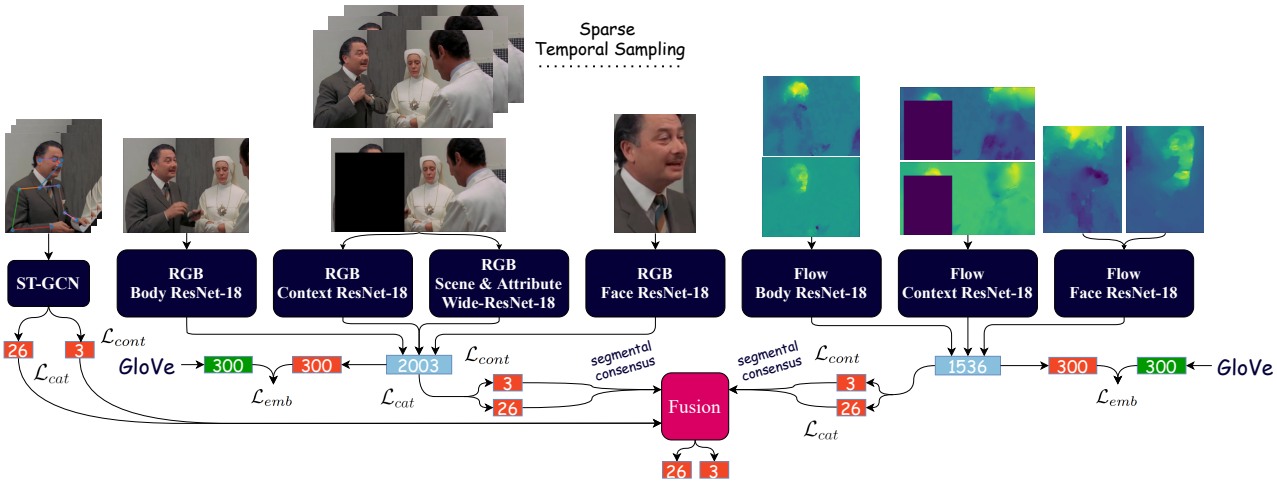

Fig. 1. Complete schematic diagram of the proposed network ensemble, featuring a ST-GCN module and three TSN input streams (*body*, *context*, *face*) for both the *RGB* and *Optical Flow* modalities, plus a *scene* & *attribute* related stream, especially for the *RGB* modality. Concatenated feature vectors are depicted in cyan, fully-connected layers are depicted in orange and GloVe word embeddings are depicted in green, along with their dimensionality or number of hidden units. The ST-GCN inherently produces video-level predictions, while in the case of the TSN-RGB and TSN-Flow, this requires the prior application of segmental consensus upon the corresponding snippet-level predictions (26 confidence scores for discrete emotions, 3 regressed values for VAD dimensions). Final predictions are obtained through late score fusion.

The Places [45] database is a quasi-exhaustive repository of 10M scene photographs, labeled with 476 scene semantic categories. We use only a subset of the latter, namely the Places365-Standard which features 1.8M images and 365 scene categories. Moreover, the SUN attribute database [26] constitutes a subset of the SUN categorical database [40], comprised of 14,000 images that are annotated using a taxonomy of 102 scene attributes. Some of the categories that are included in the Places365 dataset are: amusement park, basketball court, cemetery, jail, cell, lecture room, museum, office, sauna, soccer field, etc. Some of the scene attributes included in the SUN dataset are: competing, socializing, working, exercise, praying, open-area, enclosed-area, stressful, etc. It is quite evident that the environment and scene depicted in an image can be closely related with the emotions of the people that are present. For example, an image of a funeral that is located at a cemetery, suggests a strong correlation between the above oppressive setting and the generally negative and sad feelings shared among the depicted people. Provided that our model can leverage the hinted correlations, incorporating scene specific information can potentially boost its overall recognition performance.

Given an input image, the feature extractor produces feature maps $\mathbf{Z} \in \mathbb{R}^{512\times14\times14}$, through its last convolutional block. After the application of an average pooling layer, a deep feature vector representation is formed and fed into a FC layer with weights $\mathbf{W}_{\text{scenes}} \in \mathbb{R}^{512\times365}$, producing class confidence scores $\hat{\mathbf{y}}_{\text{scenes}} \in \mathbb{R}^{365}$. An additional set of pre-trained weights, namely $\mathbf{W}_{\text{attr}} \in \mathbb{R}^{512\times102}$ can be used for the prediction of confidence scores $\hat{\mathbf{y}}_{\text{attr}} \in \mathbb{R}^{102}$ for 102 scene attributes that are included in the SUN dataset. The corresponding scene and attribute classification probabilities, i.e. $\mathbf{p}_{\text{scenes}} \in [0,1]^{365}$ and $\mathbf{p}_{\text{attr}} \in [0,1]^{102}$, are calculated after the row-wise application of the softmax function. Subsequently, the produced scene and attribute probability scores are concatenated with the extracted deep features from all the aforementioned input streams. After the initialization of the feature extractor with the aforementioned pre-trained models, weight parameters are kept frozen during the training phase.

The inclusion of all input streams of the TSN-RGB results in a 2003-dim concatenated feature vector.

5) Loss Functions: For the training of the continuous emotion prediction branch, we use a standard *mean squared error* (MSE) loss $\mathcal{L}_{\text{cont}}$ along the three emotional dimensions of *valence*, *arousal* and *dominance*. As far as categorical emotion prediction is concerned, the ground truth targets are provided in the form of confidence scores. Firstly, we apply a sigmoid function to the barebones extracted class scores and then impose an MSE loss between the predicted and ground truth confidence scores. We denote this loss term as $\mathcal{L}_{\text{cat}_1}$. Secondly, after binarizing the ground truth confidence scores with a threshold of 0.5, we apply a binary cross-entropy loss between the produced and ground truth multi-hot target labels. We denote this term as $\mathcal{L}_{\text{cat}_2}$. We also enforce semantic congruity between the extracted visual embeddings and the categorical label word embeddings from a 300-dim GloVe [27] model, pre-trained on Wikipedia and Gigaword 5 data, in the same manner as in [10]. More specifically, given an input image $\mathbf{X}$, we transform the concatenated visual embeddings $f_v(\mathbf{X})$ into the same dimensionality as the word embeddings $f_t(y^i)$ through a linear transformation $\mathbf{W}_{\text{emb}}$, with $\mathbf{y}$ being the corresponding multi-hot target vector and i being the categorical label class index. We later apply an MSE loss between the transformed visual embeddings and the average of word embeddings that correspond only to the positive labels of the given ground truth target vector and denote this term as $\mathcal{L}_{\text{emb}}$:

$$\mathcal{L}_{\text{emb}} = \|\mathbf{W}_{\text{emb}} f_v(\mathbf{X}) - \frac{1}{|\mathcal{P}|}\sum_{y^i\in\mathcal{P}} f_t(y^i)\|_2^2 \qquad (2)$$

where $\mathcal{P}$ denotes the set of positive class labels for a given target vector $\mathbf{y}$ and $|\mathcal{P}|$ denotes the cardinality of that set. The whole network can be trained in an end-to-end manner by minimizing the combined loss function:

$$\mathcal{L} = \mathcal{L}_{\text{cat}_1} + \mathcal{L}_{\text{cat}_2} + \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{emb}} \tag{3}$$

B. TSN-Flow

Similarly to the case of temporal stream ConvNets in the original two-stream convolutional architecture [36], we experiment with training a TSN on stacked optical flow fields. Optical flow extraction is carried out via the TV-L^1 algorithm [42]. This form of dense optical flow is known to effectively encode motion between consecutive frames. We denote this model as TSN-Flow.

In all our subsequent implementations, we stack bidirectional optical flow fields from $L = 5$ consecutive frames for each snippet. After decomposing each displacement vector into its horizontal and vertical components, we end up with a 10-channel input volume per segment, per input stream. To begin with, we train a standard TSN using the *Optical Flow* modality and the body crops of each frame instance. The usage of body joint coordinates for the localization and extraction of the necessary bounding boxes remains the same as in the case of the *RGB* modality. Body-oriented dense optical flow encodes the movement of the primary depicted agent in each instance. Additionally, we incorporate a context stream, in the form of stacked optical flow fields whose primary depicted agents have been masked out. The context input stream effectively encodes the motion of any occasional secondary agent or object. Lastly, we introduce an input stream that focuses solely on the head and face movements of the primary agent. This is achieved by training an additional temporal ConvNet on small fragments of dense flow that correspond to the head region of each agent.

The features extracted using optical flow streams have distributions that greatly differ from their RGB counterparts. As optical flow values are discretized in the interval $[0, 255]$, therefore sharing the same range with RGB images, we use RGB models to initialize the parameters of the temporal ConvNets. The feature extractors for all TSN-Flow streams were pre-trained on ImageNet [7]. Consequently, the weights of the first convolutional layer are modified so as to handle the input of optical flow fields. More specifically, the weights are averaged across the RGB channels and replicated by the number of channels of the temporal stream inputs.

The inclusion of all the aforementioned input streams results in a 1536-dim concatenated feature vector per input volume. During training we employ the same combined loss function as the one used for its RGB counterpart.

C. Skeleton-Based Learning

As for the final source of affective information, we shift our attention towards the *Human Skeleton* and attempt to incorporate a Spatial-Temporal Graph Convolutional Network (ST-GCN) [41], as it was originally proposed for skeleton-based action recognition. We choose to deploy the vanilla ST-GCN consisting of 9 layers of spatial-temporal graph convolution operators (ST-GCN units). The features extracted from the last ST-GCN unit undergo average pooling and with the use of 1×1 convolutions, the final predictions are produced.

1) Joint Labeling Strategies: The main variable setting of the ST-GCN configurations is the joint labeling strategy that is being used during the construction of the graph adjacency matrix, namely *uniform*, *distance* or *spatial*. With *uniform* being the simplest labeling strategy, all joints that are connected through a limb belong in the same subset, resulting in $K_v = 1$ total subsets. The *distance* labeling strategy extends the concept of neighboring joints, as pairs of joints that are connected through a sequence of limbs are also taken into consideration, leading to a total of $K_v = D + 1$ subsets, where D is the maximum allowed distance between two neighboring joints (we choose $D = 1$ for simplicity). Lastly, according to the *spatial* labeling strategy, neighboring joints are distinguished based on their individual distances from a fixed root (the neck), resulting in $K_v = 3$ subsets.

2) Forward Propagation: In the spatial-temporal case, the input feature map $\mathbf{H}_{\text{in}}$ of a ST-GCN unit is represented as a tensor of shape $(C_{\text{in}}, T_{\text{in}}, V)$, where C_{in} denotes the number of input channels, T_{in} denotes the number of frames in the skeleton sequence and V denotes the number of nodes. Firstly, the input tensor undergoes a $(K_v \cdot C_{\text{out}}) \times 1 \times 1$ spatial graph convolution operation, with C_{out} being the desired number of output channels and K_v being the number of joint subsets that are formed based on the chosen labeling strategy. The resulting tensor is reshaped into $(K_v, C_{\text{out}}, T_{\text{in}}, V)$ and multiplied with the normalized adjacency matrix $\mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}$, where $\hat{\mathbf{A}} = \mathbb{I} + \mathbf{A}$ ($\mathbb{I}$ denotes the identity matrix) and $\mathbf{D}$ is a diagonal matrix with elements $D^{ii} = \sum_j \hat{A}^{ij}$. In case of the distance and spatial partitioning strategies ($K_v > 1$), the adjacency matrix $\mathbf{A}$ is formed by stacking K_v matrices $\mathbf{A}_k$, with each one corresponding to one of the K_v joint subsets. If we ignore interlayer nonlinearities, then the aforementioned spatial convolution operation is equivalent to the original GCN [16] formula:

$$\mathbf{H}_{\text{out}} = \sum_k \mathbf{W}_k \mathbf{H}_{\text{in}} \mathbf{D}_k^{-\frac{1}{2}} \hat{\mathbf{A}}_k \mathbf{D}_k^{-\frac{1}{2}} \tag{4}$$

where $\mathbf{W}_k$ are $C_{\text{out}} \times C_{\text{in}} \times 1 \times 1$ weight matrices (the multiplication is replicated T_{in} times in the temporal dimension and V times in the spatial dimension). $D_k^{ii} = \sum_j \hat{A}_k^{ij} + \alpha$ is the normalized diagonal matrix and α is set to 0.001 to avoid empty rows. Additionally, learnable edge importance weighting can be implemented simply by multiplying element-wise the adjacency matrices $\hat{\mathbf{A}}_k$ of Eq. 4 with a weight mask $\mathbf{M}$, namely $\hat{\mathbf{A}}_k \odot \mathbf{M}$. The output feature map resulting from the spatial graph convolution undergoes a $C_{\text{out}} \times \Gamma \times 1$ temporal convolution, with Γ denoting the temporal kernel size, completing the processing pipeline of a ST-GCN unit.

Training in the case of the ST-GCN is driven by a combined loss function similar to the one described in Eq. 3, with the sole difference being the exclusion of the categorical label embedding loss $\mathcal{L}_{\text{emb}}$.

3) Data Augmentation: Joint coordinates are normalized using the largest joint bounding box within each sequence and subsequently centralized in the range $[-0.5, 0.5]$. We then partly follow the proposed methodologies of [41]. After finding the maximum sequence length T, we proceed to pad all joint sequences with zeros until they reach that specified length, instead of repeating them from the beginning. During training, paddings are applied randomly within the sequence, while during inference, paddings are placed always at the end for consistency. During training, we also perform random affine transformations on the skeleton sequences of all frames, with the aim of simulating camera movement.

After augmentation, input data is represented by tensors of size (C, T, V). For each video frame, BoLD provides 18 tuples that contain 2D joint coordinates, plus a detection confidence score associated with each joint, therefore in our case $C = 3$ and $V = 18$. In order to further reduce the effect of over-fitting, we also pre-train the ST-GCN on the Kinetics dataset [5] which has been extensively used for skeleton-based action recognition.

IV. EXPERIMENTAL EVALUATIONS

A. Dataset

All of our upcoming experimental results are based on the standard train, validation and test splits of the Body Language Dataset (BoLD)[1] which was assembled by Luo *et al.* [22] and constitutes a database that focuses on bodily expressions of emotion. BoLD is comprised of 9,876 movie video clips of body movements, depicting a total of 13,239 human characters. The annotation of the dataset was performed using a crowdsourcing pipeline based on the Amazon Mechanical Turk. Instances are annotated in both categorical and dimensional level, utilizing the 26 emotional categories of the EMOTIC [17] dataset and the VAD [30] dimensional model, respectively.

As far as evaluation metrics are concerned, for categorical emotion prediction, *average precision* (AP), i.e. the area under the precision-recall curve as well as the *area under the receiver operating characteristic* (ROC-AUC) are used. For continuous emotion regression along the VAD dimensions, the *coefficient of determination* (R^2) is used. Performance comparison among different models is carried out on the basis of an aggregatory *emotion recognition score* (ERS) which is calculated as follows:

$$\text{ERS} = \frac{1}{2}\big(\text{m}R^2 + \frac{1}{2}(\text{mAP} + \text{mRA})\big) \tag{5}$$

where mR^2 denotes the mean R^2 score among the VAD dimensions while mAP and mRA denote the mean AP and mean ROC-AUC scores over the 26 emotion categories, respectively.

B. Configuration Details

For all experiments regarding TSN configurations, we use $K_{\text{train}} = 3$ segments during training and $K_{\text{val}} = 25$ segments during validation while the segmental consensus function has been chosen to be average pooling. Both the TSNs and ST-GCN are trained for 30 epochs with a batch size of 16, using the SGD optimizer, momentum equal to 0.9 and weight decay equal to 10^{-5}. The initial learning rate is set to 10^{-3} in the case of TSNs and $5 \cdot 10^{-3}$ in the case of the ST-GCN. The learning rates are reduced by a factor of 0.1 whenever the monitored loss on the validation set plateaus. No data augmentation technique was applied for either the TSN-RGB or TSN-Flow, as the built-in variations of the BoLD dataset proved sufficient for avoiding over-fitting.

Apart from the previously described methodologies, we experiment with the partial batch-normalization scheme (*Partial BN*), as proposed in [38]. After the initialization with pre-trained models, in every ConvNet feature extractor, we freeze the mean and variance of parameters of all batch normalization layers, except for the first one. This method is expected to work especially well in the case of temporal ConvNets and reduce the effect of over-fitting. More specifically, as the distribution of optical flow is different from the RGB images, the distribution of activation values in the first convolutional layer will also differ from the ones inherited through their initialization with RGB pre-trained models.

All results were generated on a single NVIDIA GeForce RTX 2080 Ti. Our PyTorch code and pre-trained models are publicly available[2].

C. Ablation Studies

Tables I and II present performance comparisons among all TSN-RGB and TSN-Flow configurations, respectively, which we have considered. The second columns describe the various input streams that are being included, with "B" denoting the body, "C" denoting the context, "F" denoting the face, "S" denoting the Places365 scene categories and "A" denoting the corresponding SUN attributes.

TABLE I
PERFORMANCE COMPARISON OF VARIOUS TSN-RGB MODEL CONFIGURATIONS ON THE BOLD VALIDATION SET.

Method	Features	$\mathcal{L}_{\text{emb}}$	*Partial BN*	Regression	Classification		ERS
				mR^2↑	mAP↑	mRA↑	
TSN-RGB	B	No	No	0.0300	0.1419	0.5910	0.1983
	BC			0.0362	0.1468	0.6021	0.2053
	BCF			0.0647	0.1756	0.6414	0.2366
	BCFS			0.0679	0.1746	0.6414	0.2379
	BCFA			0.0685	0.1763	0.6417	0.2388
	BCFSA	No	No	0.0710	0.1762	0.6435	0.2404
		Yes	No	0.0713	0.1779	0.6457	0.2416
		Yes	Yes	**0.0969**	**0.1839**	**0.6537**	**0.2579**
	CFSA	Yes	Yes	0.0804	0.1821	0.6485	0.2479

TABLE II
PERFORMANCE COMPARISON OF VARIOUS TSN-FLOW MODEL CONFIGURATIONS ON THE BOLD VALIDATION SET.

Method	Features	$\mathcal{L}_{\text{emb}}$	*Partial BN*	Regression	Classification		ERS
				mR^2↑	mAP↑	mRA↑	
TSN-Flow	B	No	No	0.0560	0.1431	0.5778	0.2082
	BC			0.0661	0.1415	0.5882	0.2155
	BF			0.0649	0.1497	0.5971	0.2192
	BCF	No	No	0.0795	0.1524	0.6054	0.2292
		Yes	No	0.0888	0.1563	0.6135	0.2369
		Yes	Yes	**0.0948**	0.1566	0.6172	**0.2409**
	CF	Yes	Yes	0.0799	**0.1574**	**0.6219**	0.2348

[1] https://cydar.ist.psu.edu/emotionchallenge

[2] https://github.com/GiannisPikoulis/FG2021-BoLD

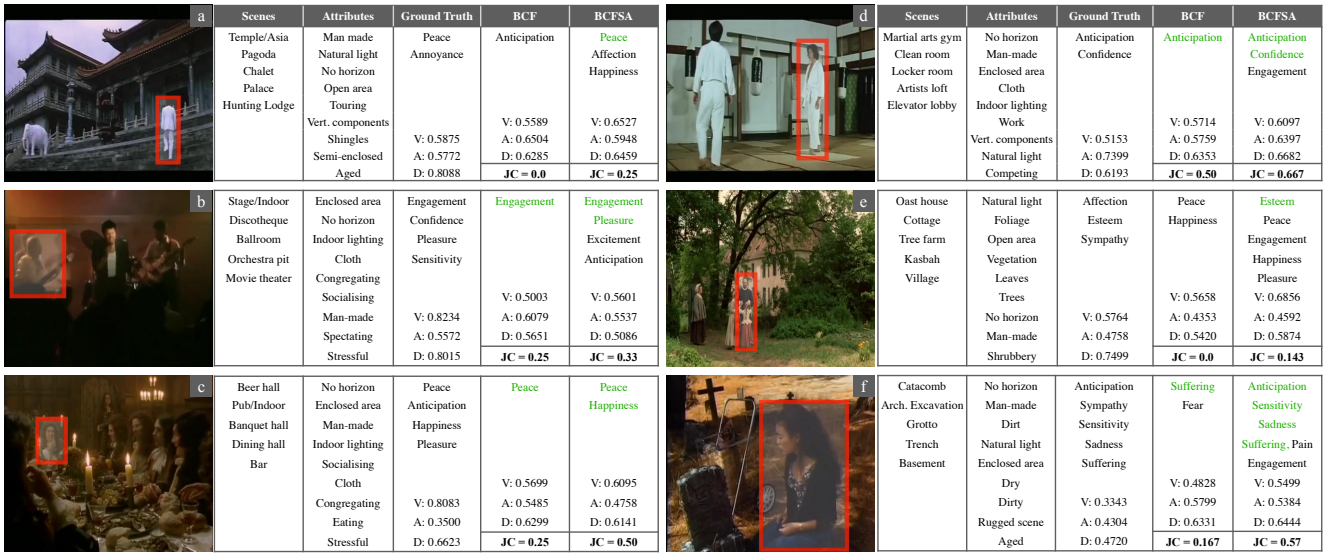

Fig. 2. Top-5 predicted scene categories, top-9 predicted attibutes, ground truth and predicted (regressed) emotion categories (VAD values) as well as *Jaccard similarity coefficient* (JC) on samples that have been randomly selected from the BoLD validation set. All predictions are made at video level.

1) TSN-RGB: Both the inclusion of the context and face streams are conducive to an increase in ERS score, with the latter showcasing a more considerable boost in performance over the bare-bones body stream. The sole inclusion of either the Places365 scene-specific or the SUN attribute scores is conducive to higher recognition scores, with the latter seemingly being more beneficial. However, it is the combined usage of the two that results in the biggest improvement in performance, in both the categorical and continuous tasks, as expected. Moreover, while the addition of $\mathcal{L}_{\text{emb}}$ leads to a trivial performance boost, the application of the *Partial BN* regularization scheme tops off our previously best performing network, reaching a maximum of 0.2579 ERS on the BoLD validation set. It seems as if the continuous re-estimation of mean and variance parameters of batch-normalization layers that are located deeper within the network, becomes obsolete and may in fact have a negative impact on generalization performance, provided that the model's parameters have been previously initialized through a proper pre-training procedure.

The beneficial influence of scene and attribute related features in human emotion understanding becomes more evident in cases where the facial characteristics and poses of the depicted agents are occluded. This is further highlighted in Fig. 2 which includes instances that were randomly selected from the validation set. Each instance is accompanied by its top-5 predicted scene categories, top-9 predicted attributes, ground truth and predicted (regressed) emotion categories (VAD values) as well as the corresponding *Jaccard similarity coefficient* (JC), for each model configuration. Correct category recognition is indicated in green. In all cases, the incorporation of scene and attribute characteristics, on top of the existing bodily, contextual and facial features, results in more emotions being correctly recognized. In addition, emotions that are semantically related, i.e. peace-happiness-pleasure (e) and sadness-suffering-pain (f), are jointly predicted, even though some of them have not been included by the annotators.

TABLE III

PERFORMANCE COMPARISON OF VARIOUS ST-GCN MODEL CONFIGURATIONS ON THE BOLD VALIDATION AND TEST SETS.

Set	Method	Pre-training	Labeling Strategy	Regression	Classification		ERS
				mR²↑	mAP↑	mRA↑	
Valid.	ST-GCN (ours)	None	*uniform*	0.0322	0.1274	0.5674	0.1898
			distance	0.0380	0.1352	0.5772	0.1971
			spatial	0.0385	0.1392	0.5871	0.2008
		Kinetics [5]	*spatial*	**0.0652**	**0.1542**	**0.6103**	**0.2237**
Test	Luo *et al.* [22]	N/A	N/A	0.0440	0.1263	0.5596	0.1940
	ST-GCN (ours)	Kinetics [5]	*spatial*	**0.0908**	**0.1694**	**0.6268**	**0.2445**

2) TSN-Flow: The introduction of either the context or face stream leads to a marginal improvement over the bare-bones temporal body stream. A more considerable boost in performance is achieved through the inclusion of all three input streams. These findings validate our intuitive decision to follow a multi-stream approach for encoding motion through optical flow, analogously to the case of the *RGB* modality. As mentioned in the case of TSN-RGB, the addition of the categorical label embedding loss $\mathcal{L}_{\text{emb}}$ improves the network's performance in both categorical and continuous tasks, while with the application of the *Partial BN* scheme, the resulting model tops off all previous configurations reaching a maximum of 0.2409 ERS.

3) ST-GCN: As far as skeleton-based learning is concerned, Table III provides a performance comparison among all ST-GCN configurations which we have considered. We notice that the *spatial* labeling strategy leads to better results compared to the others, confirming the findings of [41], in spite of having a relatively minor impact on the overall emotion recognition performance. Pre-training the ST-GCN on Kinetics provides a significant performance boost in both categorical and continuous tasks over all of its counterparts that have been trained on BoLD from scratch, reaching a

TABLE IV

PERFORMANCE COMPARISON OF VARIOUS NETWORK ENSEMBLES ON THE BOLD VALIDATION SET, UTILIZING LATE FUSION SCHEMES.

Method	Score Fusion	Regression	Classification		ERS
		mR^2↑	mAP↑	mRA↑	
TSN-RGB+TSN-Flow (ours)	Maximum	0.1000	0.1840	0.6549	0.2597
	Average	0.1458	0.1884	0.6671	0.2867
TSN-RGB+TSN-Flow +ST-GCN (ours)	Maximum	0.0735	0.1806	0.6478	0.2439
	Average	0.1440	**0.1933**	0.6667	0.2870
	Weighted Average	**0.1498**	0.1930	**0.6694**	**0.2905**
Filntisis *et al.* [10]	Average	0.0917	0.1656	0.6266	0.2439

TABLE V

QUANTITATIVE RESULTS ON THE BOLD TEST SET REGARDING THE PERFORMANCE AND COMPLEXITY OF OUR PROPOSED MODEL VERSUS OTHER PUBLISHED WORKS.

Method	# Parameters ($\times 10^6$)	Regression	Classification		ERS
		mR^2↑	mAP↑	mRA↑	
Luo *et al.* [22]	N/A	0.1030	0.1714	0.6352	0.2530
Filntisis *et al.* [10]	111.5	0.1141	0.1796	0.6416	0.2624
Ours	71.4	**0.1609**	**0.2187**	**0.6829**	**0.3059**

maximum ERS of 0.2237 on the validation set and 0.2445 on the test set. Therefore, pre-training plays a crucial role in the overall performance of the network and presumably constitutes the main differentiating factor between the reported results of [22], and ours.

4) Proposed Method: The proposed methodology constitutes a late fusion scheme among the best performing models from all modalities, namely *RGB*, *Optical Flow* and *Human Skeleton*, effectively forming a network ensemble. The score fusion methods which will be considered include: maximum, simple average and weighted average. Table IV summarizes the results. A weighted average of the TSN-RGB, TSN-Flow and ST-GCN, with a weight ratio of 2:2:1 respectively, leads to the best result of 0.2905 ERS on the validation set. More importantly, our implementation surpasses the current state-of-the-art of 0.2439 ERS on the BoLD validation set, as it was recently achieved in [10]. Figs. 3 and 4 summarize the results for the 26 discrete emotion categories and the continuous VAD dimensions relative to the AP and R^2 performance metrics, respectively.

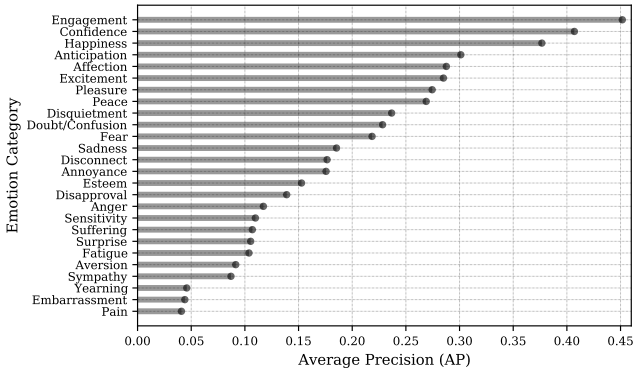

Fig. 3. *Average precision* (AP) scores per emotion category, as obtained on the BoLD validation set, using our proposed network ensemble.

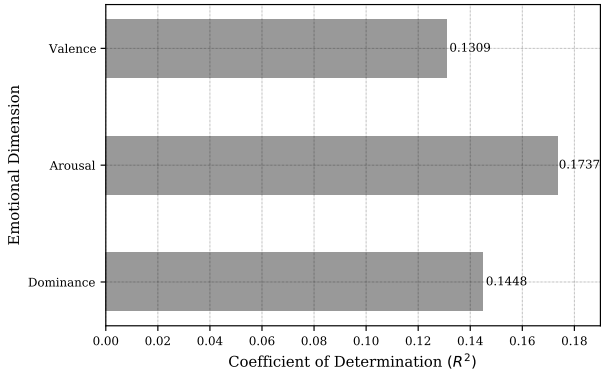

Fig. 4. *Coefficient of determination* (R^2) per emotional dimension, as obtained on the BoLD validation set, using our proposed network ensemble.

Subsequently, we evaluate our best performing network ensemble on the official BoLD test set. A comparative study regarding the performance and complexity of our proposed model and earlier published works is presented in Table V. The proposed network ensemble manages to surpass the current state-of-the-art of 0.2624 ERS, as achieved in [10], by a considerable margin on all metrics, thus verifying the superiority of our technique. As far as complexity is concerned, our model does a good job at maintaining a lower number of trainable parameters through the efficient utilization of multiple input streams and shallow feature extractors (ResNet-18, Wide-ResNet-18), in comparison with previous implementations that made use of deeper ConvNet backbones [10] (ResNet-50 & 101) and disentangled the various input streams all together [22].

V. CONCLUSIONS AND FUTURE WORK

This study employs two major components of action recognition related literature, namely Temporal Segment Networks (TSN) and Spatial-Temporal Graph Convolutional Networks (ST-GCN) with the aim of extending the concept of context-based emotion recognition in the dynamic setting of video sequences. The most notable contribution of this paper is the extension of the original TSN architecture with the inclusion of multiple input streams that effectively encode bodily, contextual, facial and generic scene-related features, enhancing our model's perception of visual context and emotion in general. Our intuitive modifications regarding the incorporation of scene and attribute classification scores, as well as multi-stream optical flow, combined with a properly pre-trained ST-GCN, have led to significant improvements over the state-of-the-art techniques with relation to the challenging Body Language Dataset (BoLD).

Lastly, a possible future research direction might be proposals for further exploitation of the categorical label dependencies that reside within the datasets and may lead to an additional improvement in categorical emotion prediction. Also, the *Depth* modality has been left relatively unexplored on the subject of context-based emotion recognition in videos. Currently, published data with relation to BoLD are quite scarce and there is definitely a lot of room for improvement. However, our existing results undoubtedly prove that we have made significant steps in the right direction.

REFERENCES

[1] A. [Asha] Kapur, A. [Ajay] Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen. Gesture-based affective computing on motion capture data. In *Proc. Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*. Springer, 2005.

[2] T. Bänziger, M. Mortillaro, and K. R. Scherer. Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12:1161–1179, 2012.

[3] L. F. Barrett and E. A. Kensinger. Context is routinely encoded during emotion perception. *Psychological Science*, 21:595–599, 2010.

[4] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha. STEP: Spatial temporal graph convolutional networks for emotion perception from gaits. In *Proc. AAAI Conf. on Artificial Intelligence*, 2020.

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] G. Castellano, L. Kessous, and G. Caridakis. Emotion recognition through multiple modalities: Face, body gesture, speech. In *Affect and Emotion in Human-Computer Interaction*, pages 92–103. Springer, 2008.

[7] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] B. Dudzik, M. Jansen, F. Burger, F. Kaptein, J. Broekens, D. K. J. Heylen, H. Hung, M. A. Neerincx, and K. P. Truong. Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *Proc. Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019.

[9] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos. Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction. *IEEE Robotics and Automation Letters*, 4:4011–4018, 2019.

[10] P. P. Filntisis, N. Efthymiou, G. Potamianos, and P. Maragos. Emotion understanding in videos through body, context, and visual-semantic embedding loss. In *Proc. Eur. Conf. on Computer Vision Workshops (ECCVW) - Workshop on Bodily Expressed Emotion Understanding (BEEU)*. Springer, 2020.

[11] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Proc. Int. Conf. on Neural Information Processing (ICONIP)*. Springer, 2013.

[12] H. Gunes and M. Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30:1334–1345, 2007.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2015.

[15] M. Karg, K. Kühnlenz, and M. Buss. Recognition of affect based on gait patterns. *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40:1050–1061, 2010.

[16] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *Proc. Int. Conf. on Learning Representations (ICLR)*, 2017.

[17] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Context based emotion recognition using EMOTIC dataset. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 42:2755–2766, 2019.

[18] R. Laban and L. Ullmann. *The Mastery of Movement*. Boston, Plays, 1971.

[19] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn. Context-aware emotion recognition networks. In *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019.

[20] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang. ARBEE: Towards automated recognition of bodily expression of emotion in the wild. *Int. Journal of Computer Vision (IJCV)*, 128:1–25, 2019.

[23] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. EmotiCon: Context-aware multimodal emotion recognition using Frege's principle. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] D. Mobbs, N. Weiskopf, H. C. Lau, E. Featherstone, R. J. Dolan, and C. D. Frith. The Kuleshov Effect: The influence of contextual framing on emotional attributions. *Social Cognitive and Affective Neuroscience*, 1:95–106, 2006.

[25] A. Mollahosseini, B. Hasani, and M. H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. on Affective Computing*, 10:18–31, 2019.

[26] G. Patterson and J. Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[27] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[28] S. Piana, A. Staglianò, A. Camurri, and F. Odone. A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition. In *Proc. IDGEI Int. Workshop*, 2013.

[29] R. Righart and B. De Gelder. Recognition of facial expressions is influenced by emotional scene gist. *Cognitive, Affective, & Behavioral Neuroscience*, 8:264–272, 2008.

[30] J. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 1977.

[31] S. Saha, S. Datta, A. Konar, and R. Janarthanan. A study on emotion recognition from body gestures using Kinect sensor. In *Proc. Int. Conf. on Communication and Signal Processing (ICCSP)*. IEEE, 2014.

[32] T. Sapiński, D. Kamińska, A. Pelikant, and G. Anbarjafari. Emotion recognition from skeletal movements. *Entropy*, 21:646, 2019.

[33] W. Sheng and X. Li. Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network. *Pattern Recognition*, 114, 2021.

[34] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[35] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[36] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proc. Int. Conf. on Neural Information Processing Systems (NIPS)*, 2014.

[37] J. M. Susskind, A. K. Anderson, and G. E. Hinton. The Toronto face database. *Dept. of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep*, 3, 2010.

[38] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal Segment Networks: Towards good practices for deep action recognition. In *Proc. Eur. Conf. on Computer Vision (ECCV)*. Springer, 2016.

[39] M. J. Wieser and T. Brosch. Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in Psychology*, 3, 2012.

[40] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[41] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proc. AAAI Conf. on Artificial Intelligence*, 2018.

[42] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L^1 optical flow. In *Proc. DAGM Symp. on Pattern Recognition*. Springer, 2007.

[43] S. Zagoruyko and N. Komodakis. Wide residual networks. In *Proc. British Machine Vision Conf. (BMVC)*. BMVA Press, 2016.

[44] K. Zhang, Y. Huang, Y. Du, and L. Wang. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Trans. on Image Processing*, 26:4193–4203, 2017.

[45] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 40:1452–1464, 2018.