Convolutional Network with Densely Backward Attention for Facial Expression Recognition

Cam-Hao Hua*, Thien Huynh-The†, Hyunseok Seo*, and Sungyoung Lee*
*Department of Computer Science and Engineering
Kyung Hee University, Gyeonggi-do, 17104, Republic of Korea
Email: {hao.hua,shs,sylee}@oslab.khu.ac.kr
†ICT Convergence Research Center
Kumoh National Institute of Technology, Gumi, Republic of Korea
Email: thienht@kumoh.ac.kr

***Abstract*—The emergence of convolutional neural network (CNN) has enabled facial expression recognition to accomplish significant outcomes nowadays. However, while existing multi-stream networks are subject to costly computation, the attention-embedded approaches do not involve multiple levels of semantic context in the predefined CNN. Based on the observation that emotions via a person's face are fusion of various muscular modalities, relying upon the outputs and corresponding attentional features of the deepest layer in the CNN is insufficient due to loss of informative details through multiple sub-sampling stages. Therefore, this paper introduces a CNN with densely backward attention to leverage the aggregation of channel-wise attention at multi-level features in a backbone network for reaching high recognition performance with cost-effective resource consumption. Particularly, cross-channel semantic information in high-level features are exploited densely to recalibrate fine-grained details in low-level versions. Then, a step of multi-level aggregation is further executed for thorougly involving spatial representations of important facial modalities. As a consequence, the proposed approach gains highest mean class accuracy of 79.37% on RAF-DB, which is competitive with the state-of-the-arts.**

***Index Terms*—facial expression recognition, convolutional neural network, densely backward attention**

I. Introduction

Recently, extraordinary advancement of computing resources and visual data regarding both quantity and quality has facilitated deep learning technique to be widely applied into numerous areas, especially computer vision. To this end, Convolutional Neural Network (CNN) [1]–[4], a well-known deep learning architecture, has attracted a great number of researchers thanks to its impressive performance enhancement in different recognition-based issues, such as human activity recognition [5], semantic scene understanding [6], [7], disease progression identification [8], and especially facial expression recognition (FER) [9].

In fact, FER has long been an active research field due to its diverse applications related to human-computer interactions [10]. Nowadays, with an increasing number of images collected from laboratory [11] and the wild [12], the power of CNN is exhaustively exploited in this image classification-related domain, of which further achievements are significantly accomplished. Particularly, there are three major CNN-based approaches proposed in the literature of FER: (i) ensemble of multiple deep networks [13]–[15]; (ii) algorithms of specialized objective function or statistical modules [12], [16], [17] attached to a conventional CNN; and (iii) attention mechanism embedded to pretrained CNNs [18], [19].

Since emotions via a person's face are represented by the combination of various muscular modalities (e.g., shape of eyes, eyebrow, nose, mouth, facial wrinkle, to name a few), several researches as shown in the first group aggregated multiple deep networks to express potentially facial features as well as contextual information for high recognition performance. In concrete, the authors in [13] took into account capsule, facial-attribute, and holistic-feature networks for coordinating spatial details with deep context smoothly throughout the whole architecture. Meanwhile, MRE-CNN [14] firstly divided the original input into multiple regions of interest based on predefined facial landmarks, then fed those patches into different VGG-16 [1] models for ensemble learning. Another noticeable architecture, called ResiDen [15], is the mixture of two well-known concepts in deep learning-based computer vision, i.e., residual connections [2] and dense blocks [3] in a single network. Obviously, expensive computation is the major limitation of these approaches. Hence, instead of involving additional sub-networks, methods in the second group mainly introduce locality preserving loss [12] or designated cluster loss [16] to minimize intra-category variation while maximize inter-category discrimination. Recently, SPDNet [17] offered specialized modules of covariance matrices for spatiotemporal pooling to combat distortion of facial landmarks during the learning process. However, the utilization of these objective functions or statistical modules sometimes results in trivial performance since certain discriminative features might not

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-0-01629) supervised by the IITP (Institute for Information & communications Technology Promotion). This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00655). This research was financially supported by NRF-2019R1A2C2090504.

978-1-7281-5453-4/20/$31.00 ©2020 IEEE

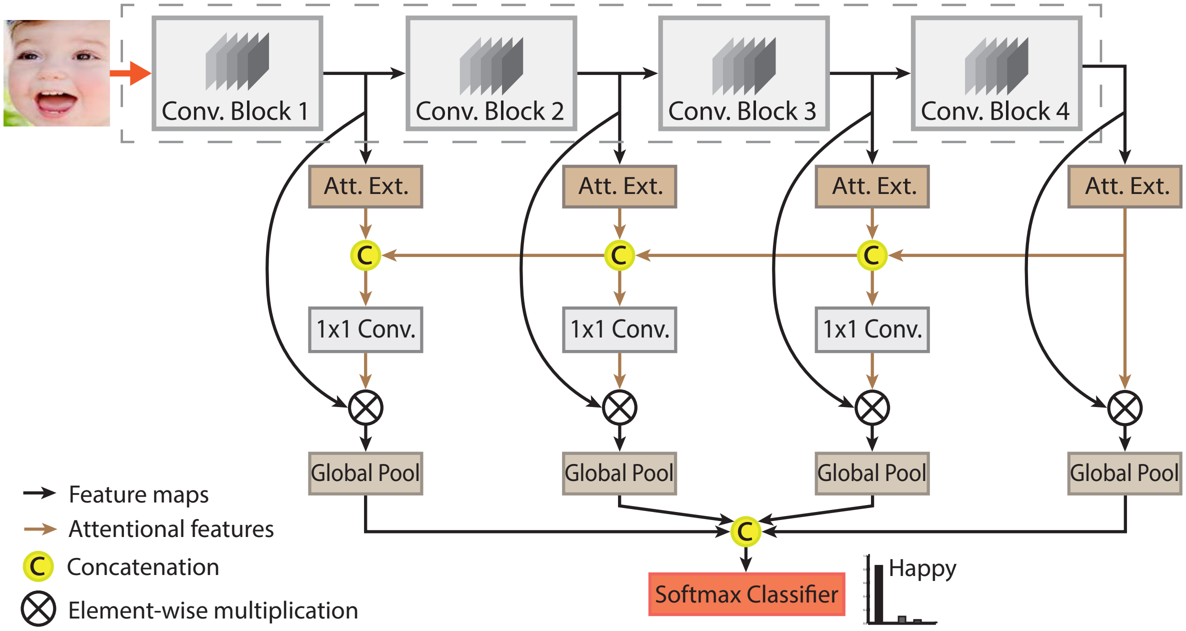


Fig. 1. Architecture of the proposed DBA-Net for facial expression recognition. Note that 'Conv. Block' and 'Att. Ext.' represent block of multiple convolutional layers and Attentional Features Extractor. Besides that, '1x1 Conv.' and 'Global Pool' stand for the convolutional layer with kernel size of 1×1 and *Average Global Pooling* layer, respectively. Color view is recommended for the best visualization.

be focused properly. Accordingly, in order to express essential features extracted by trainable layers, attention scheme is of great interest in the third group. For instance, ACNN [18] introduced patch- and global-based attention networks to re-calibrate acquired feature responses at local regions and image level, respectively. On the other hand, FERAtt [19] involved an attention module with encoder-decoder structure to effectively reconstruct facial information from the output of a CNN-based feature extractor for further classification step. It can be realized that the attention mechanism is only applied to high-level feature maps in these techniques.

In brief, existing multi-stream networks are subject to costly computation while attention-embedded models do not involve multiple levels of semantic context in a predefined CNN for FER. As aforementioned, the output emotion is represented by the fusion of different muscular modalities, which are exhaustively acquired at multiple levels by a CNN. Therefore, manifold sub-sampling stages along feedforward pass of the CNN leads to the loss of certain spatial correlations between several facial tissues, which are hardly encoded in channel dimension. Consequently, it is hypothesized that only relying upon the outputs and corresponding attentional features of the deepest layer for the classifier is insufficient.

From such observations, this paper introduces a CNN with Densely Backward Attention, namely DBA-Net, to leverage the aggregation of depth-wise attention at multi-level features in a pretrained CNN for attaining high recognition performance with cost-effective resource consumption. In particular, according to the fact that coarser and deeper feature maps hold much more informative context along the depth dimension than the finer and shallower ones do, we opt for conducting the impact of such channel-wise semantic information on the finer-grained features by the attention blocks (inspired from [4]) in backward manner. Moreover, for the purpose of smoothly coordinating the finely-patterned (low-level) and semantically-rich (high-level) features, a dense re-calibration procedure is taken into account. As a result, this forms into a light-weight stream of densely backward attention (DBA) for thoroughly involving spatial representations of important facial modalities, which are densely refined by semantic context of higher-level features beforehand. Apparently, such effective aggregation scheme of various semantic information from the multi-level feature maps in a CNN is the principal key for recognizing corresponding expression accurately. In order to show the effectiveness of the proposed approach, RAF-DB dataset [12] is utilized for the evaluation, of which the experimental results in terms of mean class accuracy (sum of diagonal elements in a confusion matrix) are competitive with the state-of-the-arts.

II. Methodology

This section delivers details of the proposed DBA-Net, with corresponding demonstration in Fig. 1, for FER as follows. We firstly elaborate the overall architecture wherein a mechanism of densely backward-attention attached to a pretrained CNN. Then, an in-depth description of the Attentional Features Extractor (brown module in Fig. 1) is provided. Finally, the proposed attention-based aggregation in dense manner for the improvement of prediction performance is delivered.

A. Overall Architecture

Generally, the proposed architecture consists of two parts, i.e., a backbone CNN pretrained with ImageNet [20] and the associated light-weight stream of DBA. As correspondingly illustrated in Fig. 1, convolution blocks in the dashed box represent the fundamental components of the backbone CNN while the remaining stands for the attention-embedded stream of aggregating multi-scale information for the recognition of facial emotion.

We apply different base CNNs comprising VGG [1], ResNet [2], and DenseNet [3] to show the flexibility of the proposed attention-embedded stream with respect to different capacities of feature representation. Typically, in these classification networks, layers in each convolution block learn and perform acquired features at a specific scale corresponding to a semantic level. For instance, both ResNet and DenseNet consist of four basis convolution blocks (as shown in Fig. 1), of which the final outputs have strides of 4, 8, 16, and 32 in comparison with the input's spatial size, respectively. Note that the total number of convolution and non-linear activation layers is varied in each block. Accordingly, in order to ensure the reasonable increment of computation amount, only four feature maps, which are ultimate outputs of the aforementioned learnable blocks, are taken into account for the stage of attentional features extraction. Meanwhile, since there is no explicit definition of convolution blocks in the VGG architecture, outputs of Rectified Linear Unit (ReLU) activation layers preceding the last four max-pooling layers, which also correspond to same strides as specified previously, are chosen for further processes.

It is obvious that along the feedforward flow between the convolution blocks, spatial resolution of the extracted feature maps is reduced by half while the corresponding depth size grows rapidly. Moreover, since the outcomes at later layers contain semantically-richer context in channel dimension compared to those obtained earlier, they can be utilized to re-calibrate (i.e., strengthen the informative and weaken the less-productive) feature responses extracted at shallower layers in backward fashion. By such operation, spatial details of the considered low-level feature maps are fully embedded semantic information for eliminating available ambiguities. As a consequence, it is advantageous to involve finely-patterned (high-resolution) feature maps, which possess well-organized representation of muscular modalities, in company with the semantically-rich (low-resolution) versions for the high recognition performance of facial expression. Next, the following sub-sections further characterize the algorithm to extract attention features and how the corresponding results are embedded to desired feature maps, respectively.

B. Attentional Features Extractor

This section describes the depth-wise attentional features extractor adopted from [4] as shown in Fig. 2. Basically, the major purpose of attention mechanism is to enable the important features to be focused more intensively for a better learning process. In the original work [4], the feature weights are self-recalibrated through the capture of globally contextual information across the channels. However, in the proposed network, cross-channel attributes of feature maps chosen from the core CNN are utilized not only to enhance their own representational capability but also to embed semantically-rich details to lower-level versions.

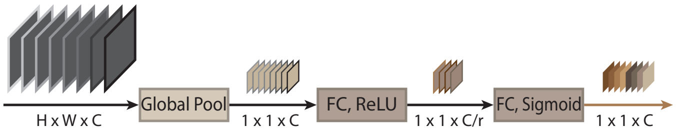


Fig. 2. Attentional Features Extractor.

As depicted in Fig. 2, at first, each channel of the concerned feature map having size of $H \times W \times C$ is spatially averaged to generate a C-length vector. Let f_i be the feature maps output from the aforementioned convolution blocks, where $i = 1, 2, 3, 4$ corresponding to the feedforward direction in the backbone network. In other words, f_1 and f_4 represent the lowest- and highest-level feature maps of interest, respectively. Subsequently, the obtained vector, denoted as $g_i \in \mathrm{R}^{\mathrm{C}}$, conducts information of cross-channel inter-dependencies as follows

$$g_i = [\mathcal{G}_1(f_i), \ldots, \mathcal{G}_c(f_i), \ldots, \mathcal{G}_C(f_i)]^T \tag{1}$$

where $\mathcal{G}_c(.)$ indicates the *Global Pool* operator that processes the c^{th} channel of a considered feature map f as following equation

$$\mathcal{G}_c(f_i) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} [f_i(h, w)]_c \tag{2}$$

where $h = 1, \ldots, H$ and $w = 1, \ldots, W$ are pixel coordinates in the feature map f_i and $[.]_c$ stands for the c^{th} channel of the concerned feature map or vector. As a result, depth-wise semantic information is encoded into the d-length vector g comprehensively. Next, to correspondingly model the underlying correlations across the channels, we feed the vector g into two *Fully Connected (FC)* layers centered by a *ReLU* activation function. It should be noted that the size of these trainable layers is equivalent to C/r and C, respectively, where r is the compression ratio fixed at 16 in this work for reducing the computational complexity. These learning procedures can be exhibited as below

$$g_{int_i} = \boldsymbol{W_{fc_2}} \left(ReLU(\boldsymbol{W_{fc_1}} g_i + \boldsymbol{b_{fc_1}}) \right) + \boldsymbol{b_{fc_2}} \tag{3}$$

where $\{\boldsymbol{W_{fc_1}} \in \mathrm{R}^{\frac{\mathrm{C}}{\mathrm{r}} \times \mathrm{C}}, \boldsymbol{b_{fc_1}} \in \mathrm{R}^{\frac{\mathrm{C}}{\mathrm{r}}}\}$ and $\{\boldsymbol{W_{fc_2}} \in \mathrm{R}^{\mathrm{C} \times \frac{\mathrm{C}}{\mathrm{r}}}, \boldsymbol{b_{fc_2}} \in \mathrm{R}^{\mathrm{C}}\}$ are trainable parameters of the first and second *FC* layers, respectively, and g_{int_i} is the intermediate depth-wise feature vector having length of C. Then, *Sigmoid* activation function is adopted to re-weight the vector g_{att_i}'s responses in the range of $(0, 1)$, which is expressed by

$$g_{att_i} = \frac{1}{1 + e^{-[g_{int_i}]_c}} \tag{4}$$

where $g_{att_i} \in \mathrm{R}^{\mathrm{C}}$ is the channel-wisely attentional feature vector, which is then used for the enrichment of informative context in the feature maps of interest learned from the backbone CNN.

C. Densely Backward Attention scheme

To this end, four attentional feature vector g_{att1}, g_{att2}, g_{att3}, and g_{att4} (demonstrated by outbound brown arrows of Att. Ext. modules in Fig. 1) corresponding to the four chosen feature maps f_1, f_2, f_3, and f_4, respectively, are inferred by the attentional features extractor presented previously. With respect to various backbone models, the vectors g_{att_i} have different lengths of C as reported in Table I. Clearly, later g_{att} with larger dimension comprises more informative specificity, which can utilized for re-calibrating finer-resolution features learned from prior convolution blocks. Therefore, for the purpose of extensively involving finely-patterned feature without contextual ambiguities to the final emotion prediction, it is necessary to embed the channel-wisely semantic details from higher-level features in a backbone CNN in backward style. Furthermore, since each attentional feature vector g_{att_i} contains a certain level of descriptive statistics, we opt for a dense concatenation manner to exhaustively integrate those multi-level semantic representational features, by which spatial details of lower-level features are refined more flexibly.

As manifested in Fig. 1, following operations are executed after the extraction of attentional features

$$\begin{aligned} f_{att4} &= f_4 \otimes g_{att4} \\ f_{att3} &= f_3 \otimes \boldsymbol{W_{113}} \left(\mathcal{C}[g_{att3}, g_{att4}]\right) \\ f_{att2} &= f_2 \otimes \boldsymbol{W_{112}} \left(\mathcal{C}[g_{att2}, g_{att3}, g_{att4}]\right) \\ f_{att1} &= f_1 \otimes \boldsymbol{W_{111}} \left(\mathcal{C}[g_{att1}, g_{att2}, g_{att3}, g_{att4}]\right) \end{aligned} \tag{5}$$

where $\otimes$ symbolizes the element-wise multiplication operator, $\mathcal{C}[.]$ denotes the vector concatenation procedure, and $\{\boldsymbol{W_{113}}, \boldsymbol{W_{112}}, \boldsymbol{W_{111}} \in \mathrm{R}^{\mathrm{C}\times1\times1\times\mathrm{D}}\}$ are the trainable parameters of C convolution filters with size of $1 \times 1 \times D$. Obviously, C is the channel size of the considered feature map f_i while the value of D depends on dimension of the output concatenated by involved vectors $g_{att_i}, g_{att_{i+1}}, \ldots, g_{att_4}$. For example, if using VGG-16 as the core model, values D of $\boldsymbol{W_{113}}$, $\boldsymbol{W_{112}}$, and $\boldsymbol{W_{111}}$ are 1024, 1280, and 1408, respectively (based on values C in Table I). In short, the function of such learnable 1×1 convolution layers is to effectively reduce the dimension of concatenated feature vector for appropriate recalibration of considered feature maps. Moreover, by such densely backward scheme, features captured at shallower layers can still obtain much higher-level representations for a more efficient refinement.

Then, the re-weighted feature maps f_{att_i} continuously pass through the *Global Pool* modules followed by a concatenation operator for the collection and comprehensive aggregation of globally essential context at multiple levels, respectively, as below

$$f_{fer} = \mathcal{C}\left[G(f_{att1}), G(f_{att2}), G(f_{att3}), G(f_{att4})\right] \tag{6}$$

where f_{fer} refer to as the ultimate features of our DBA-Net and G represents the *Global Pool* operator defined in (1) and (2). Finally, we employ a Softmax classifier to recognize the corresponding facial expression given a predefined number of supervised emotion labels.

TABLE I
LENGTHS C OF EXTRACTED ATTENTIONAL FEATURE VECTORS g_{att_i} ($i = 1, 2, 3, 4$) WITH RESPECT TO DIFFERENT BACKBONE CNNS.

Backbone CNN	g_{att1}	g_{att2}	g_{att3}	g_{att4}
VGG-16 [1]	128	256	512	512
ResNet-101 [2]	256	512	1024	2048
DenseNet-161 [3]	384	768	2112	2208

Note: These values of C are also identical to the depth size of corresponding feature maps f_1, f_2, f_3, f_4 extracted from the backbone networks.

III. EXPERIMENTS

In this section, we firstly provide an overview of the RAF-DB dataset [12], which is used for evaluating the proposed methodology. Then, corresponding implementation details are given. Afterwards, ablation study with corresponding discussion and comparison with state-of-the-art models are performed to show the remarkable capability of our DBA-Net in terms of FER.

A. Benchmark Dataset

RAF-DB [12], standing for Real-world Affective Faces Database, is a large-scale dataset of in-the-wild facial expression. This database is challenging in the literature since its 30,000 images carries out a tremendous diversity of ages, genders and ethnicity, head poses, lighting conditions, occlusions, specialized manipulations, and so on. In this paper, we only experiment with the single-label set, i.e., each image exclusively indicates one of the seven basic classes of emotion (angry, disgust, fear, happy, neutral, sad, and surprise). Accordingly, 12,271 training and 3,068 testing images, which are prior cropped into the resolution of 100×100 around the regions of face, are involved for the designated experiments. Moreover, it should be noted that mean class accuracy (i.e., sum of diagonal elements in the resulting confusion matrix) is the golden metric to benchmark the classification performance due to the between-class imbalance issue stated in [12].

B. Implementation Details

The proposed model and corresponding evaluations are implemented using Pytorch [21] and Scikit-learn [22] frameworks, respectively. Same as existing work, we also apply following augmentation such as random change of hue and saturation, horizontal flipping, and rotation in range of (-20° , 20°) to the training images with mini-batch size of 64. In addition, weight decay of 0.0005 is employed generalize the proposed model more robustly.

About the training stage, we initialize the learning rate at 0.005 and use Softmax loss to assess the quality of DBA-Net's parameters given ground-truth labels. Then, in order to accordingly minimize the calculated loss with respect to those trainable parameters, we follow the optimization procedure in [23] wherein stochastic gradient descent with momentum of 0.9 is utilized along with the 'poly'-style schedule of learning rate decay. Notably, the training process takes place in 50 epochs on one NVIDIA 1080TI GPU.

TABLE II
MEAN CLASS ACCURACY ON RAF-DB [12] TEST SET WITH VARIOUS SETTINGS OF BACKBONE CNN AND ATTENTION STRATEGY.

Backbone CNN	Strategy			Mean Class Accuracy (%)	Number of Parameters
	Baseline	Aw/oDB	DBA-Net		
VGG-16 [1]	✓			74.96	134.30M
		✓		77.35	14.81M
			✓	78.81	15.94M
ResNet-101 [2]	✓			77.10	42.51M
		✓		77.48	43.23M
			✓	79.33	49.19M
DenseNet-161 [3]	✓			77.21	26.49M
		✓		77.75	27.78M
			✓	79.37	42.91M

TABLE III
COMPARISON OF MEAN CLASS ACCURACY ON RAF-DB [12] TEST SET.

Approach	Mean Class Accuracy (%)
DLP-CNN [12]	74.20
3DMFA [16]	75.73
ResiDen [15]	76.54
MRE-CNN [14]	76.73
Capsule-based Net [13]	77.48
Double C*d*-LBP [24]	78.60
SPDNet [17]	79.43
DBA-Net (VGG-16)	**78.81**
DBA-Net (ResNet-101)	**79.33**
DBA-Net (DenseNet-161)	**79.37**

C. Ablation Study

For the purpose of showing robustness of the proposed architecture regarding facial expression prediction, we experiment three strategies, i.e., Baseline, Aw/oDB, and DBA for each backbone network. Note that the Baseline refers to as finetuning the pretrained model end-to-end. Meanwhile, the Aw/oDB corresponds to the involvement of attentional features extractor at the end of the basis convolution blocks but without densely backward concatenation scheme. Accordingly, quantitative performance of these strategies with different backbone CNNs on the testing images is reported in Table II.

In general, Aw/oDB and DBA-Net outperform the baseline one 0.38-2.39% and 2.16-3.85%, respectively, for all backbone networks. This implies that the engagement of attention scheme at multi-scale features and subsequent depth-wise aggregation of corresponding outcomes are plausible in the scenario of facial expression identification. The major reason is arguably originated from the fact that the attention strategy at specific scales followed by a concatenation module can intensively exploit features of different muscular modalities captured at multiple levels in a CNN.

Moreover, regarding the effectiveness of the DBA compared to Aw/oDB, the mean class accuracy is further improved 1.46% (in the case of using VGG-16 as backbone network), 1.85% (ResNet-101), and 1.62% (DenseNet-161). Such superior performance points out the importance of additionally integrating higher-level attentional feature vectors for recalibrating lower-level feature maps. As discussed in Section II-C, the dense combination in backward manner helps the network flexibly express informative spatial features subject to multi-level semantic details along depth dimension.

It is also obvious that the greater capacity the core CNN has, the better performance is attained (but not significantly). Concretely, using ResNet-101 and DenseNet-161 as backbones yields the similar mean class accuracy of 79.33% and 79.37%, respectively, which are around 0.5% higher than that of employing VGG-16.

As for details of class-wise performance, we further present confusion matrices of the proposed DBA-Net corresponding to different backbone networks in Fig. 3. All of these confusion matrices deliver common outcomes as follows. The prediction of happy feeling yields highest accuracy and that of neutral, sad, and surprise also gives remarkable performance. On the other hand, the expressions of disgust and fear are misclassified with neutral/sad and sad/surprise by an average rate of about 10%, respectively. We argue that the compression layers following the dense combination of attentional features (i.e., $\boldsymbol{W}_{\mathbf{113}}$, $\boldsymbol{W}_{\mathbf{112}}$, and $\boldsymbol{W}_{\mathbf{111}}$ in (5)) have to trade-off unavoidable loss of concatenated semantic details, which leads to indistinguishable representations of facial modalities between the above-mentioned emotions.

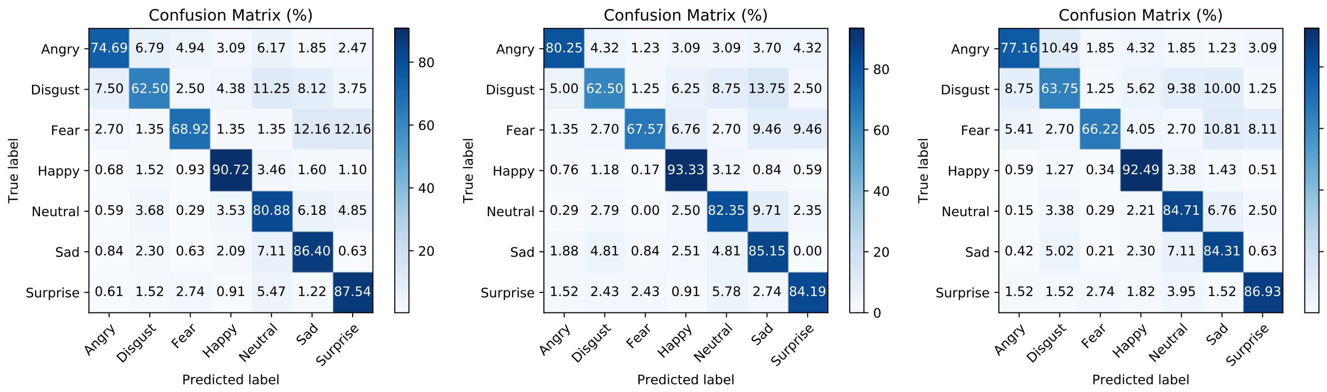
Regarding the computational costs enumerated in Table II, the DBA-Net with VGG as backbone can reduce the number of parameters by approximately 88% because of not involving expensive *FC* layers at backend of the baseline. Meanwhile, compared to original architectures of the ResNet and DenseNet, using the proposed scheme only increases the complexity by around 16%. Clearly, although the increment of parameters' amount is mainly caused by the stage of densely backward concatenation, it is worth gaining an improvement of 1.46-1.85% for mean class accuracy as aforementioned.

D. Comparison with State-of-the-art Methods

Through the quantitative comparison shown in Table III, the proposed DBA-Net achieves mean class accuracy competitive with that of the state-of-the-arts. In details, by only applying VGG-16 as the core network in our architecture, higher rates of 0.21-4.61% than those of the compared methods (except for SPDNet [17]) are attained. Furthermore, with deeper backbone networks like ResNet-101 or DenseNet-161, performance of the proposed approach is continuously improved, and hence, comparable to that of the cutting-edge SPDNet [17] with trivially lower accuracy of 0.1-0.06%. Clearly, such impressive performance on a challenging dataset expresses the advantage of aggregating low- and high-level features by the utilization of channel-wise attention mechanism in densely backward structure.

IV. CONCLUSION

This study has introduced a cost-effective convolutional network with densely backward attention, namely DBA-Net, for FER. The proposed approach aims to aggregate low- and high-level features in an efficient way according to the hypothesis that facial emotion is represented by the fusion



(a) (b) (c)

Fig. 3. Confusion Matrices of the proposed DBA-Net on RAF-DB dataset [12] with different backbone CNN: (a) VGG-16 [1], (b) ResNet-101 [2], (c) DenseNet-161 [3].

of different muscular modalities extracted at multiple levels. For such purpose, attention mechanism is densely embedded in backward manner to a pretrained classification-based CNN for leveraging the performance of FER. The achievement of impressive experimental results enables the DBA-Net to be widely applied in practical scenarios. Lastly, the issue of several misclassification is our next focus in the future work.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[3] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 2261–2269.

[4] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] T. Huynh-The, C.-H. Hua, and D. Kim, "Encoding pose features to images with data augmentation for 3d action recognition," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2019.

[6] C.-H. Hua, T. Huynh-The, and S. Lee, "Convolutional networks with bracket-style decoder for semantic scene segmentation," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2018, pp. 2980–2985.

[7] C.-H. Hua, T. Huynh-The, and S. Lee, "Retinal vessel segmentation using round-wise features aggregation on bracket-shaped convolutional neural networks," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2019, pp. 36–39.

[8] C.-H. Hua, T. Huynh-The, K. Kim, S.-Y. Yu, T. Le-Tien, G. H. Park, J. Bang, W. A. Khan, S.-H. Bae, and S. Lee, "Bimodal learning via trilogy of skip-connection deep networks for diabetic retinopathy risk progression identification," *International Journal of Medical Informatics*, 2019.

[9] M. Wu, W. Su, L. Chen, Z. Liu, W. Cao, and K. Hirota, "Weight-adapted convolution neural network for facial expression recognition in human-robot interaction," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–12, 2019.

[10] S. Li and W. Deng, "Deep facial expression recognition: A survey," *CoRR*, vol. abs/1804.08348, 2018.

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, June 2010, pp. 94–101.

[12] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019.

[13] S. Ghosh, A. Dhall, and N. Sebe, "Automatic group affect analysis in images via visual attribute and feature networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 1967–1971.

[14] Y. Fan, J. C. Lam, and V. O. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 84–94.

[15] S. Jyoti, G. Sharma, and A. Dhall, "Expression empowered residen network for facial action unit detection," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, May 2019, pp. 1–8.

[16] F. Lin, R. Hong, W. Zhou, and H. Li, "Facial expression recognition with data augmentation and compact feature learning," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 1957–1961.

[17] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2018, pp. 480–4807.

[18] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[19] P. D. Marrero Fernandez, F. A. Guerrero Pena, T. Ing Ren, and A. Cunha, "Feratt: Facial expression recognition with attention net," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018, pp. 833–851.

[24] F. Shen, J. Liu, and P. Wu, "Double complete d-lbp with extreme learning machine auto-encoder and cascade forest for facial expression analysis," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 1947–1951.