Emotion-aware Multi-view Contrastive Learning for Facial Emotion Recognition (Appendix)

Daeha Kim and Byung Cheol Song

Inha University, Incheon, Republic of Korea
kdhht5022@gmail.com,bcsong@inha.ac.kr

1 Neural Network Architecture Design

The proposed method consists of a total of five neural networks: encoder, compressor, regressor, projection head, and additional head for discriminative learning. Here, encoder and compressor are divided into two types according to the backbone. As shown in Figure 1, the compressor (AL) of AlexNet (AL) [12] is designed with average pooling with flatten and a single fully-connected (FC) layer, and requires about 294k learnable parameters. The compressor (R18) of ResNet18 (R18) [7] is designed with average pooling with flatten and two FC layers, and has about 545k learnable parameters. And, the regressor, projection heads are all composed of two FC layers with batch normalization, and only require a small amount of learnable parameters of 10k or less. Those networks differ from each other in the types of dropout and activation function. Please refer to Figure 1 for details of dimensions and training parameters.

2 Forward/backward Pass for Optimization Process

Overview. The formulas of SparseMax (Sp) and SoftMax (Sm) used in the feature transformation of the latent feature $\mathbf{z}$ are as follows.

$$\mathrm{Sp}(\mathbf{z}) := \underset{\mathbf{p}\in\Delta^{d-1}}{\arg\max}\, \langle \mathbf{z}, \mathbf{p}\rangle - \frac{1}{2}\|\mathbf{p}\|^2 = \underset{\mathbf{p}\in\Delta^{d-1}}{\arg\min} \|\mathbf{p}-\mathbf{z}\|^2 \tag{1}$$

$$\mathrm{Sm}(\mathbf{z}) := \underset{\mathbf{p}\in\Delta^{d-1}}{\arg\max}\, \langle \mathbf{z}, \mathbf{p}\rangle + \mathcal{H}(\mathbf{p}) = \frac{e^{\mathbf{z}}}{\sum_i e^{\mathbf{z}_i}} \tag{2}$$

where $\Delta^{d-1} = \left\{\mathbf{p}\in\mathbb{R}^d_+ \mid \|\mathbf{p}\|_1 = 1\right\}$, $\mathcal{H}(\mathbf{p}) = -\sum_i p_i \ln p_i$, *i.e.*, the negative Shannon entropy. The output $\mathbf{z}_a$ of Sp (Eq. (1)) generates sparse facial features associated with emotional expression. The output $\mathbf{z}_n$ of Sm (Eq. (2)) generates the average attention information of the face, which shows a relatively contrasting characteristic to $\mathbf{z}_a$.

Eqs. (1) and (2) can be generalized to a differentiable (convex) optimization problem based on the (arbitrary) objective function f and the constraint functions g and h as follows.

$$\begin{aligned} &\mathbf{p}^*(\mathbf{z}) = \underset{\mathbf{p}}{\arg\min}\, f(\mathbf{p}, \mathbf{z}) \\ &\text{subject to } g(\mathbf{p}, \mathbf{z}) \le \mathbf{0} \text{ and } h(\mathbf{p}, \mathbf{z}) = \mathbf{0} \end{aligned} \tag{3}$$

For example, in Eq. (3), the objective function f corresponds to $\|\mathbf{p}-\mathbf{z}\|^2$ in Eq. (1). The constraint functions h and g correspond to $\|\mathbf{p}\|_1 = 1$ and $\mathbf{p} \geq \mathbf{0}$, respectively. Eq. (2) can also be interpreted in the same way as Eq. (1). After all, the optimal solution $\mathbf{p}^*$ in Eq. (3) corresponds to $\mathbf{z}_a$ and $\mathbf{z}_n$ of Eqs. (1) and (2), respectively.

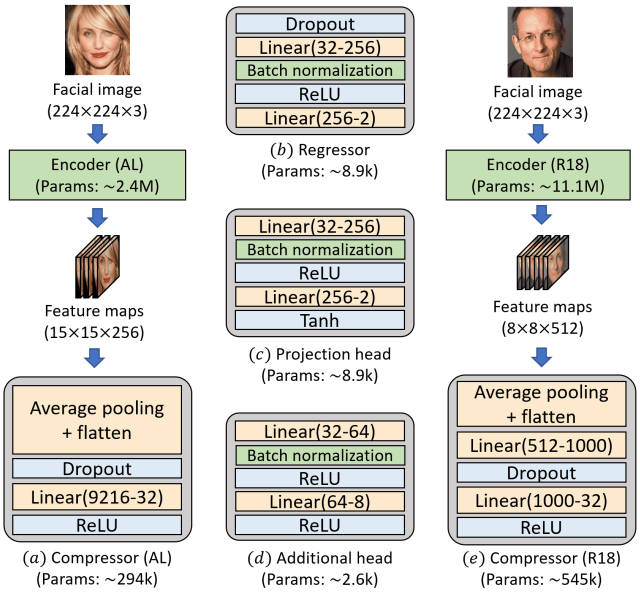

Fig. 1. Details of network architecture of the proposed method. Here, 'flatten' indicates the process of vectorizing feature maps

Forward pass. Obtaining $\mathbf{p}^*$ is to find solutions $(\mathbf{p}^*, \lambda^*, \nu^*)$ that satisfy the equality constraints of the KKT conditions of Eq. (4) below.

$$G(\mathbf{p}, \lambda, \nu) = \begin{pmatrix} \nabla_{\mathbf{p}} f(\mathbf{p}, \mathbf{z}) + \partial_{\mathbf{p}} g(\mathbf{p}, \mathbf{z})^T \lambda + \partial_{\mathbf{p}} h(\mathbf{p}, \mathbf{z})^T \nu \\ \lambda \circ g(\mathbf{z}) \\ h(\mathbf{p}) \end{pmatrix} \tag{4}$$

Here, λ and ν are dual solutions of Eq. (4). And, the optimal solution $(\mathbf{p}^*, \lambda^*, \nu^*)$ of Eq. (1) must satisfy the following conditions.

$$G(\mathbf{p}^*, \lambda^*, \nu^*) = 0, g(\mathbf{p}^*, \mathbf{z}) \leq \mathbf{0}, \lambda^* \geq \mathbf{0} \tag{5}$$

As a result, the embedded conic solver (ECOS) of the CVXPY library [3] calculates $\mathbf{p}^*$ through an iterative optimization process while satisfying the conditions of Eqs. (4) and (5).

Backward pass. In the proposed method designed in end-to-end fashion, the backward propagation of Sp and Sm is as follows. First, calculate $\partial_{(\mathbf{p},\lambda,\nu)} G$, *i.e.*,

the Jacobian matrix of $G(\mathbf{p}^*, \lambda^*, \nu^*)$. Then, register this matrix value in the backward hook of the Pytorch library. Here, the backward hook that operates during back-propagation of the end-to-end neural network allows to compute the gradients of Sp and Sm together with the gradient values of the previous layer. Finally, the computed gradients are propagated to the next layer.

In Listing 1.1, we can see the Pytorch-like pseudocode declaring Sp and Sm in less than 20 lines. For more information on forward and backward propagation, see [4].

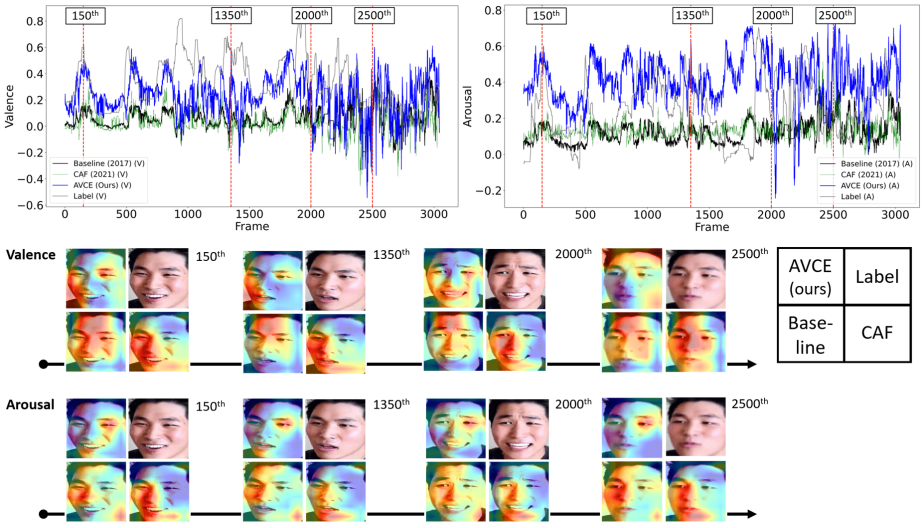

Fig. 2. Additional qualitative results of frame unit emotion fluctuations and neural activation maps on validation split of Aff-wild dataset

3 Details of Evaluation Metrics

Root mean-squared error (RMSE) measures the point-wise difference: $RMSE = \sqrt{\mathbb{E}(\mathbf{x} - \hat{\mathbf{x}})^2}$. Here, $\mathbf{x}$ and $\hat{\mathbf{x}}$ refer to GT and prediction, respectively. Sign agreement (SAGR) measures the overall positive/negative degree of emotion: $SAGR = \frac{1}{N}\sum_{i=1}^{N} \Gamma(\text{sign}(\mathbf{x}_i), \text{sign}(\hat{\mathbf{x}}_i))$. Here, Γ is an indicator function that outputs 1 if two values have the same sign, and 0 otherwise. In addition, in order to overcome the disadvantages of the previous metrics that cannot measure the correlation between two variables, the Pearson correlation coefficient (PCC), which can measure linearity, was used: $PCC = \frac{\mathbb{E}[(\mathbf{x}-\mu_{\mathbf{x}})(\hat{\mathbf{x}}-\mu_{\hat{\mathbf{x}}})]}{\sigma_{\mathbf{x}}\sigma_{\hat{\mathbf{x}}}}$. Here, $\mu_{\mathbf{x}}$ and $\sigma_{\mathbf{x}}$ indicate the mean and standard deviation of $\mathbf{x}$. The concordance correlation

Table 1. Results on the AffectNet. Red and blue indicate the first and second-ranked value, respectively

Methods	Backbone	Params.	RMSE (↓)		PCC (↑)		CCC (↑)	
			(V)	(A)	(V)	(A)	(V)	(A)
Baseline [15]	AlexNet	61M	0.37	0.41	0.66	0.54	0.60	0.34
Jang et al. [8]	SSD w/ VGG16	-	0.44	0.39	0.58	0.50	0.57	0.47
Kollias et al. [10]	VGG16	-	0.37	0.39	0.66	0.55	0.62	0.54
Barros et al. [1]	AlexNet	-	-	-	-	-	0.67	0.38
Kossaifi et al. [11]	ResNet18	-	0.35	0.32	0.71	0.63	0.71	0.63
Hasani et al. [6]	ResNeXt50	3.1M	0.267	0.248	0.78	0.86	0.74	0.85
CAF [9]	ResNet18 (R18)	11M	0.219	0.187	0.86	0.85	0.83	0.84
	AlexNet (tuned) (AL)	3.6M	0.222	0.192	0.81	0.86	0.80	0.85
AVCE (Ours)	ResNet18 (R18)	11M	0.191	0.174	0.903	0.848	0.865	0.840
	AlexNet (tuned) (AL)	2.7M	0.198	0.180	0.908	0.832	0.860	0.826

coefficient (CCC), which measures the agreement of two variables, is defined as follows: $CCC = \frac{2\sigma_{\mathbf{x}}\sigma_{\hat{\mathbf{x}}}PCC(\mathbf{x},\hat{\mathbf{x}})}{\sigma_{\mathbf{x}}^2+\sigma_{\hat{\mathbf{x}}}^2+(\mu_{\mathbf{x}}-\mu_{\hat{\mathbf{x}}})^2}$.

4 Full Comparison on AffectNet

In Table 1, AVCE (R18) improved RMSE (V) by 0.028 and improved PCC (V) by 4.3% in comparison with CAF (R18) [9]. Even though CCC (A) of AVCE (AL) was about 2.4% lower than that of CAF (AL), AVCE showed superior performance in the aspect of RMSE, which is very important in evaluating AffectNet consisting only of static images. In addition, it is worth noting that the number of learnable parameters of AVCE using projection heads amounted to only 75% of that of discriminator-based CAF.

5 Additional Qualitative Results

This section analyzes the qualitative results of AVCE through neural activation maps [21] for arousal and valence, respectively. First, let's take a look at the frame unit emotion fluctuations in the Aff-wild dataset (see Fig. 2). AVCE captured the label change at the emotional peak point (*e.g.*, the 150th) better than CAF [9] and baseline [15]. This visualization demonstrates the ability of AVCE to track the fine-grained fluctuation of labels on the arousal and valence axes.

Then, the attention regions of AVCE and other techniques are compared through neural activation maps obtained from high-dimensional images of AffectNet. As shown in the rightmost column of Fig. 3(a), AVCE captured facial regions precisely where emotions were activated even in images with high brightness saturation. Also, similar result is observed in the rightmost column in Fig. 4(d). Overall, AVCE showed activation results near facial landmarks such as eyes, nose, and mouth. On the other hand, CAF and baselines often focused on backgrounds or areas away from facial landmarks.

Table 2. Ablation study according to the negative mining methods on the Aff-wild

CRL formula	Negative mining		CCC (↑)	
	Debiased	Hard	(V)	(A)
InfoNCE [16]			0.637	0.546
	✓		0.631	0.552
		✓	0.646	0.577
Barlow-Twins [19]			0.653	0.566
	✓		0.626	0.564
		✓	0.660	0.606
AVCE (AL) (Ours)			0.682	0.594
	✓		0.686	0.617
		✓	0.691	0.592

6 Synergy Effect with Negative Mining

This section analyzes the synergy effect of AVCE using the latest negative mining methods [2,17] for contrastive representation learning (CRL). A way of applying negative mining is as follows: From the negative pairs $(\mathbf{x}, \mathbf{y}_n) \sim P_X P_Y$ sampled as many as the number of mini-batches (N), $\bar{N}(< N)$ pairs are selected by [2,17]'s mining criterion. Specifically, Debiased [2] reduces the false negatives, *i.e.*, $\mathbf{y}_n$s similar to $\mathbf{x}$ expressing global facial characteristics, and then gives a debiasing effect to CRL. Hard [17] selects hard negatives $\mathbf{y}_n$s with adjustable hardness through importance sampling strategy.

Table 2 shows the performance of AVCE and other CRL methods to which each mining method is applied. All CRL methods (InfoNCE [16], Barlow-Twins [19], and AVCE) achieved significant synergy with the Hard method. For example, Barlow-Twins with Hard showed 0.04 improved CCC (A), and AVCE with Hard showed 0.009 improved CCC (V). However, Debiased could not confirm synergy with both InfoNCE and Barlow-Twins. On the other hand, AVCE with Debiased showed 0.023 improved CCC (A), achieving state-of-the-art performance on the arousal axis.

7 Derivation of AVCE

First, InfoNCE in main body is rearranged as follows:

$$\sup_{f \in \mathcal{F}} \mathbb{E}_{P_{XY}} f(\mathbf{x}, \mathbf{y}) - \log \left(\mathbb{E}_{P_X} \mathbb{E}_{P_Y} e^{f(\mathbf{x}, \mathbf{y})} \right). \tag{6}$$

Exponential or logarithmic operations of Eq. (6) can cause learning instability. So, we remove log and exp from Eq. (6), and add a regularization term to prevent excessive increase of f as follows:

$$\sup_{f \in \mathcal{F}} \mathbb{E}_{P_{XY}} f(\mathbf{x}, \mathbf{y}) - \mathbb{E}_{P_X P_Y} f(\mathbf{x}, \mathbf{y}) + (reg.). \tag{7}$$

When the regularization term of Eq. (7) is defined from the second moment of f, the same formula as AVCE is derived. The difference in the lower bound between Eq. (6) and Eq. (7) is negligible, and the contrastive properties can be preserved.

Table 3. Recognition performance on RAF-DB test set

Methods	Accuracy (%)
FSN [20]	81.10
DLP-CNN [13]	84.13
RAN [18]	86.90
DACL [5]	87.78
AVCE-discrete	84.70

Table 4. Comparison of different backbones on Aff-wild

Backbones	PCC		CCC	
	(V)	(A)	(V)	(A)
ResNet18	0.600	0.621	0.552	0.583
ResNet50	0.621	0.609	0.562	0.553
ResNet101	0.610	0.632	0.581	0.580

8 Other Experimental Results

Results on RAF-DB. In order to validate whether AVCE works in the discrete FER task, we additionally utilized RAF-DB. RAF-DB is annotated with six discrete basic expressions (i.e., happy, surprise, angry, fear, disgust, and sad) and neutral expressions. And, it consists of 12,271 training images and 3,068 testing images. Technically, after modifying the output dimension of the regressor and projection head from 2 to 7, the last activation function was removed, and categorical crossentropy was used for learning instead of MSE. All other settings were the same. The experimental results are shown in the Table 3. AVCE showed about 0.6% improvement in accuracy than DLP-CNN [13]. Although the performance was about 3% lower than that of the SOTA methods, if appropriate AVCE modification according to the understanding of discrete emotion category is involved, it will be possible to achieve SOTA performance even in RAF-DB.

Results of different backbones. Table 4 shows the performance tendencies according to the increase in the number of ResNet layers. The number of layers and the value of the Valence axis tended to be proportional to each other, and the highest PCC (A) performance was observed in ResNet101.

Effect of AVCE. We plotted some samples of AFEW-VA evaluation split in a two-dimensional space using t-SNE visualization tool [14] to analyze the implicit effect of feature learning by AVCE. The result is shown in Fig. 5. In general, it was observed that samples with small expression changes (e.g., neutral) were concentrated in the middle. On the other hand, samples corresponding to angry emotion were mainly located near the bottom of the projection space.

In addition, an experiment was performed to analyze the explicit effect of feature learning by AVCE. The result is shown in Table 5. Thanks to the triplet-based regularization term and the well-optimized training settings, RMSE values

of 0.2 or less were observed even in the absence of $\mathcal{L}_{AVCE}$. However, the PCC/CCC values on Valence axis decreased significantly by about 0.2. Values on Arousal axis also showed similar trends.

Inference speed. During the training phase, it takes about 213.7, 529.1, and 451.6 (ms) to update the parameters of Baseline, CAF, and AVCE, respectively (per 1 iteration). During the testing phase, the operating times of Baseline, CAF, and AVCE are 1.46, 1.10, and 1.07 (ms), respectively. The speed of each technique based on the image size of 224×224 and AlexNet was measured by averaging 100 times, and all configurations not mentioned were set the same.

Table 5. Ablation study according to the $\mathcal{L}_{AVCE}$ on Aff-wild

Methods	RMSE		PCC		CCC	
	(V)	(A)	(V)	(A)	(V)	(A)
AVCE (AL) w/ $\mathcal{L}_{AVCE}$	0.154	0.154	0.713	0.632	0.682	0.594
AVCE (AL) w/o $\mathcal{L}_{AVCE}$	0.206	0.181	0.520	0.574	0.476	0.523

```
import cvxpy as cp
from cvx_utils import OptLayer

# SparseMax
z = cp.Variable(32)
x = cp.Parameter(32)

f = lambda z,x:cp.sum_squared(z - x)
g = lambda z,x:-z
h = lambda z,x:cp.sum(z) - 1
SP = OptLayer([z],[x],f,[g],[h])

# SoftMax
w = cp.Variable(32)
y = cp.Parameter(32)

fs = lambda w,y:-w@y-cp.sum(cp.entr(w))
hs = lambda w,y:cp.sum(w) - 1
SM = OptLayer([w],[y],fs,[],[hs])

# Forward pass
feat = compressor(encoder(img))
sp_feat = SP(feat)  # Aug. 1 (SparseMax)
sm_feat = SM(feat)  # Aug. 2 (SoftMax)
```

Listing 1.1. Pytorch-style Pseudocode for Feature Transformations

References

1. Barros, P., Parisi, G., Wermter, S.: A personalized affective memory model for improving emotion recognition. In: International Conference on Machine Learning. pp. 485–494 (2019)
2. Chuang, C.Y., Robinson, J., Lin, Y.C., Torralba, A., Jegelka, S.: Debiased contrastive learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 8765–8775. Curran Associates, Inc. (2020)
3. Diamond, S., Boyd, S.: Cvxpy: A python-embedded modeling language for convex optimization. The Journal of Machine Learning Research **17**(1), 2909–2913 (2016)
4. Duvenaud, D.: Deep Implicit Layer tutorial. http://implicit-layers-tutorial.org/ (2020), [Online; accessed 19-July-2008]
5. Farzaneh, A.H., Qi, X.: Facial expression recognition in the wild via deep attentive center loss. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2402–2411 (2021)
6. Hasani, B., Negi, P.S., Mahoor, M.: Breg-next: Facial affect computing using adaptive residual networks with bounded gradient. IEEE Transactions on Affective Computing (2020)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Jang, Y., Gunes, H., Patras, I.: Registration-free face-ssd: Single shot analysis of smiles, facial attributes, and affect in the wild. Computer Vision and Image Understanding **182**, 17–29 (2019)
9. Kim, D.H., Song, B.C.: Contrastive adversarial learning for person independent facial emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 5948–5956 (2021)
10. Kollias, D., Cheng, S., Ververas, E., Kotsia, I., Zafeiriou, S.: Generating faces for affect analysis. arXiv preprint arXiv:1811.05027 **16** (2018)
11. Kossaifi, J., Toisoul, A., Bulat, A., Panagakis, Y., Hospedales, T.M., Pantic, M.: Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6060–6069 (2020)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
13. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2852–2861 (2017)
14. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
15. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
16. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
17. Robinson, J.D., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. In: International Conference on Learning Representations (2020)
18. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing **29**, 4057–4069 (2020)

19. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: Proceedings of the 38th International Conference on Machine Learning, Virtual Event. vol. 139, pp. 12310–12320. PMLR (2021)
20. Zhao, S., Cai, H., Liu, H., Zhang, J., Chen, S.: Feature selection mechanism in cnns for facial expression recognition. In: BMVC. p. 317 (2018)
21. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)

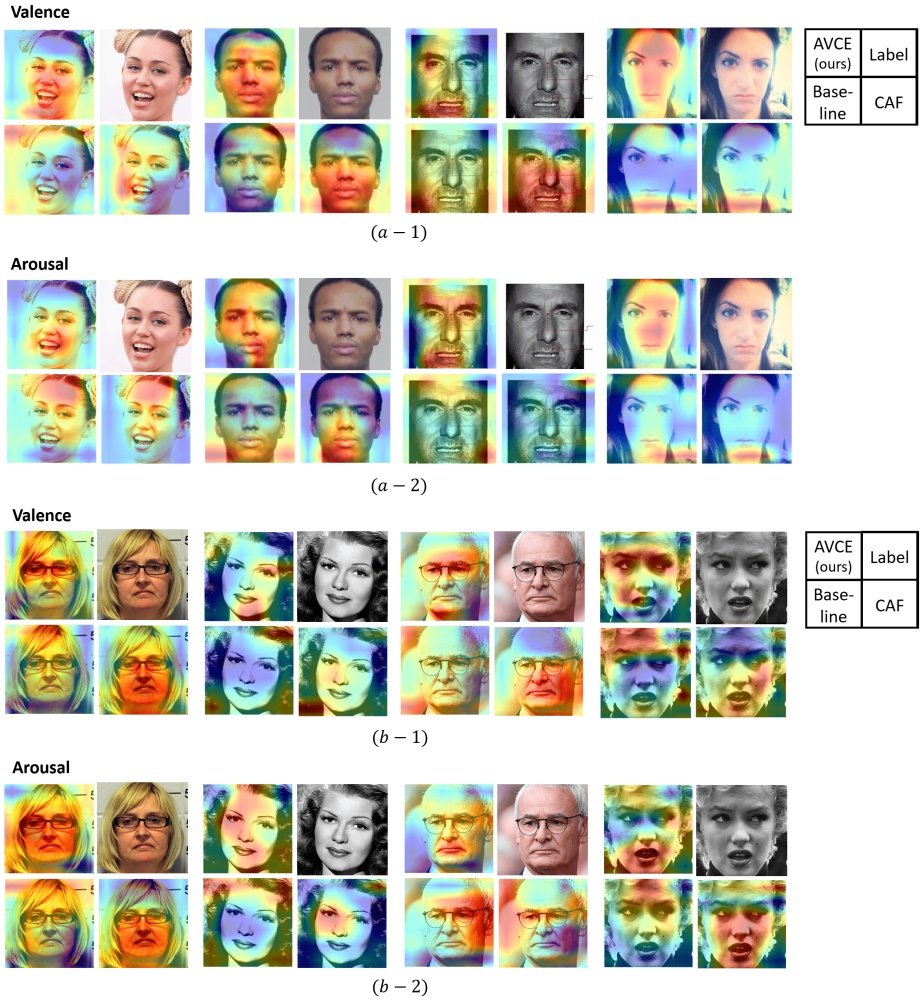

Fig. 3. Additional neural activation maps on high-dimensional AffectNet dataset.

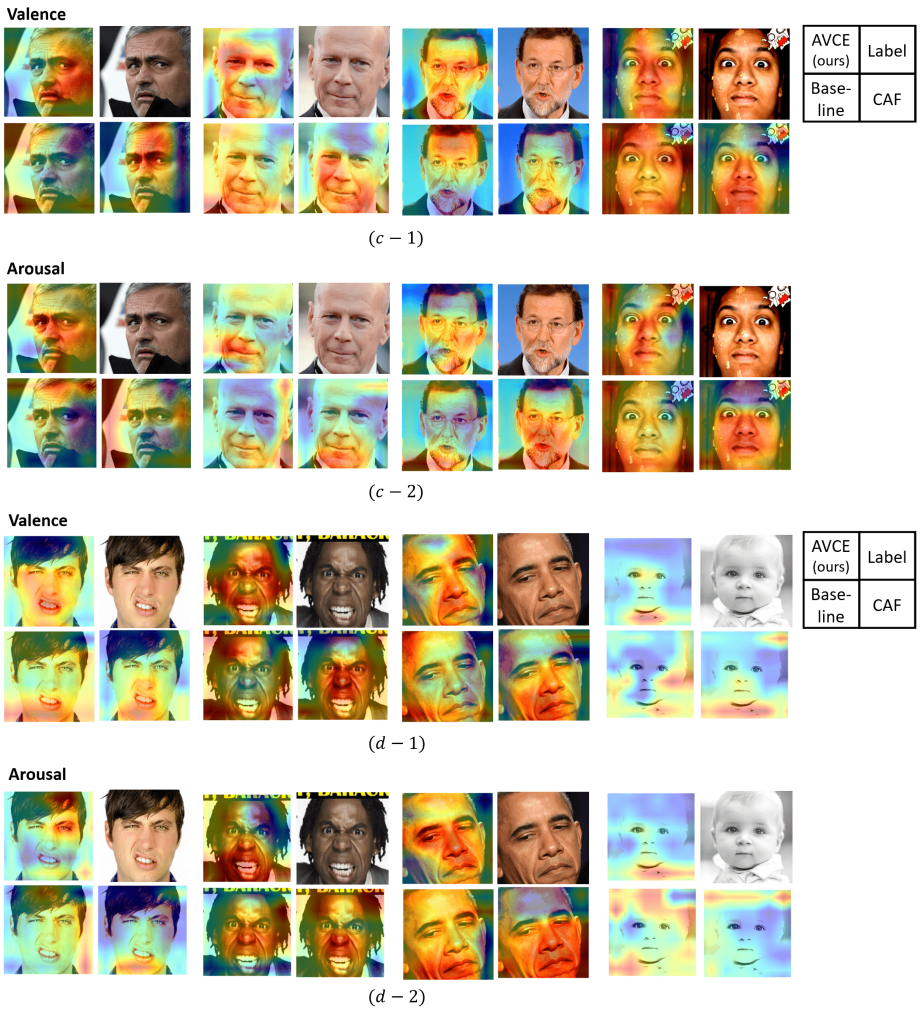

Fig. 4. Additional neural activation maps on AffectNet dataset (cont.).

Fig. 5. Two-dimensional visualization result on AFEW-VA evaluation split.