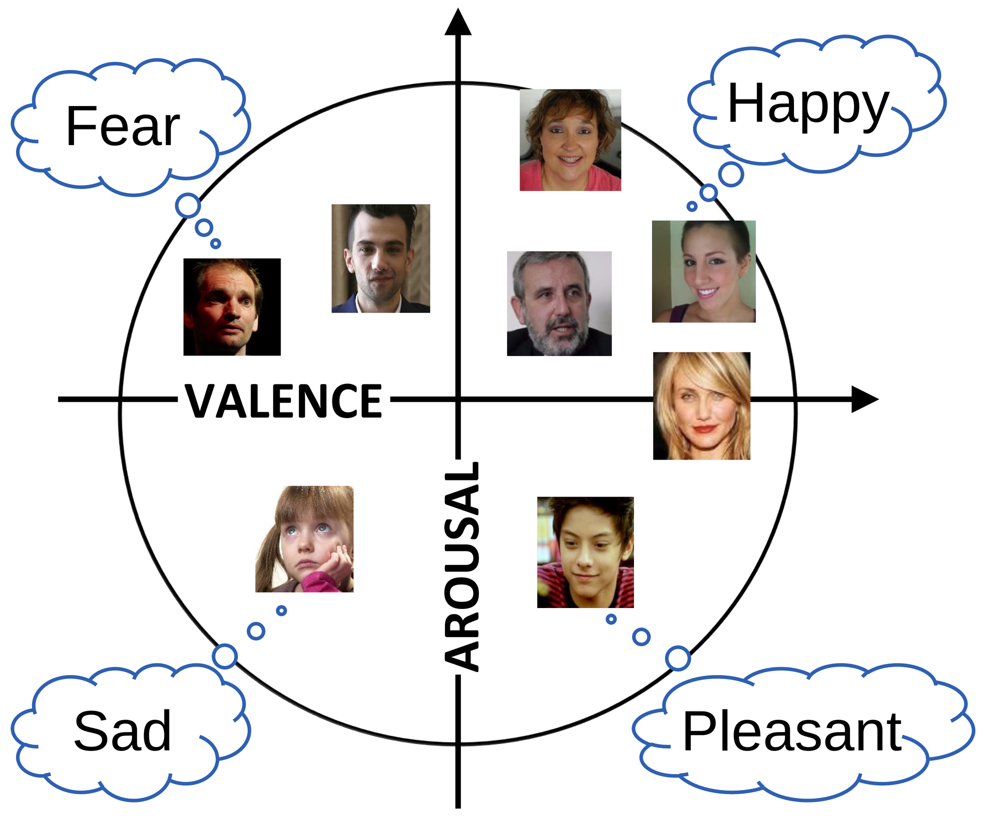
Emotion-aware Multi-view Contrastive Learning for Facial Emotion Recognition

Daeha Kim* Byung Cheol Song
Department of Electrical and Computer Engineering, Inha University, Republic of Korea

TEL AVIV 2022

Motivation

Definition Arousal Valence (AV)-based facial emotion (or expression) recognition is to perform emotional regression in a two-dimensional space with Arousal and Valence axes.

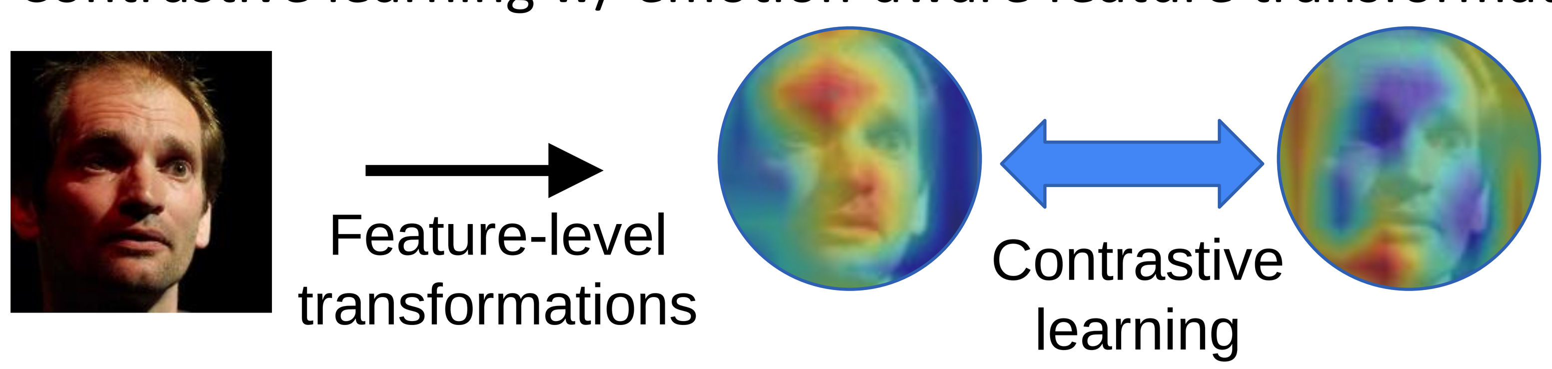


Problem formulation Previous AV FER methods have not yet technically dealt with the following concerns.

- *How can we extract facial emotion-aware features?*
- *What is the key for feature learning of facial emotions?*

Key Idea

✓Contrastive learning w/ emotion-aware feature transformations



Contributions

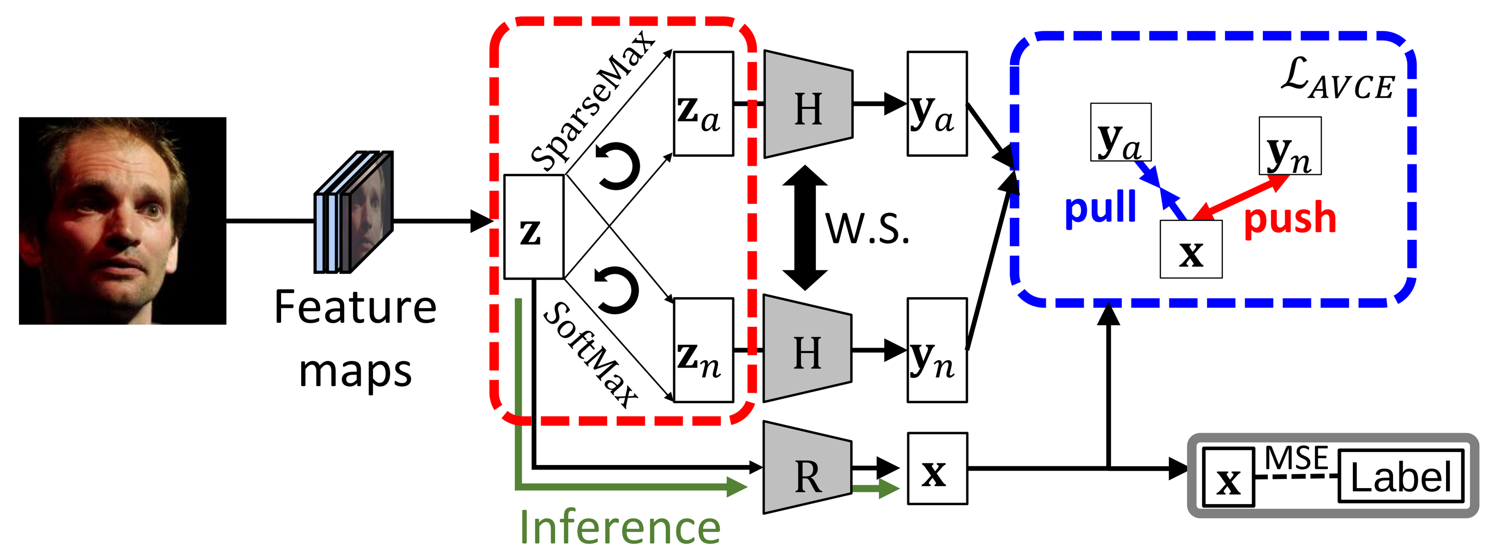
1) The proposed feature transformations enable to focus on semantic regions that are important for emotional representation.

2) We succeeded in incorporating visual perception ability into representation learning for the first time in this field.

→ <u>Given the nature of continuous AV labels, this is a challenging task.</u>

AVCE: Contrast of Emotions in AV Space

- Emotion-aware feature transformations (red box)
- Multi-view contrastive learning (blue box)
- Conventional supervised learning (gray box)



✓Multi-view contrastive learning ($\mathcal{L}_{AVCE}$)

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{(\mathbf{x},\mathbf{y}_a)\sim P_{XY}} f(\mathbf{x},\mathbf{y}) - \alpha \mathbb{E}_{(\mathbf{x},\mathbf{y}_n)\sim P_X P_Y} f(\mathbf{x},\mathbf{y}) - \frac{\beta}{2} \mathbb{E}_{(\mathbf{x},\mathbf{y}_a)\sim P_{XY}} f(\mathbf{x},\mathbf{y})$$

$$- \frac{\gamma}{2} \mathbb{E}_{(\mathbf{x},\mathbf{y}_n)\sim P_X P_Y} f(\mathbf{x},\mathbf{y}) \quad \text{s.t.} \quad f(\mathbf{x},\mathbf{y}) = \left(1 - \frac{\theta(\mathbf{x},\mathbf{y})}{\pi}\right)$$

※ See **Definition 1** and **Lemma 1** for relationship between the loss function and emotional contrast property.

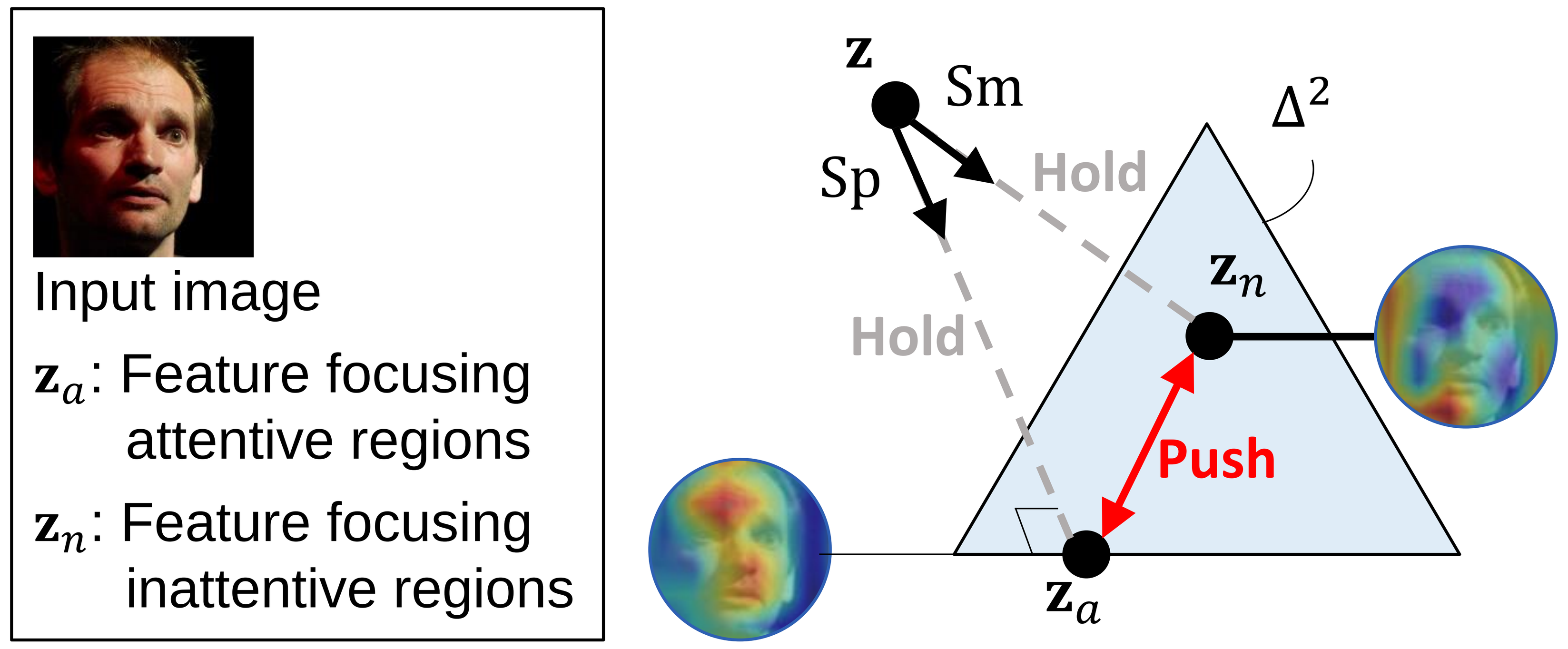
✓Feature transformations (SparseMax and SoftMax)

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

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

$\Delta^{d-1} = \{\mathbf{p} \in \mathbb{R}_+^d \mid \|\mathbf{p}\|_1 = \mathbf{1}\}$ and $\mathcal{H}(\mathbf{p}) = -\sum_i p_i \ln p_i$ (Shannon entropy)

※ **Implementation.** We utilized CVXPY library. ECOS (embedded conic solver) takes about 1.3 seconds per mini-batch on Xeon® E5-1650 CPU to generate $\mathbf{z}_a$ and $\mathbf{z}_n$.



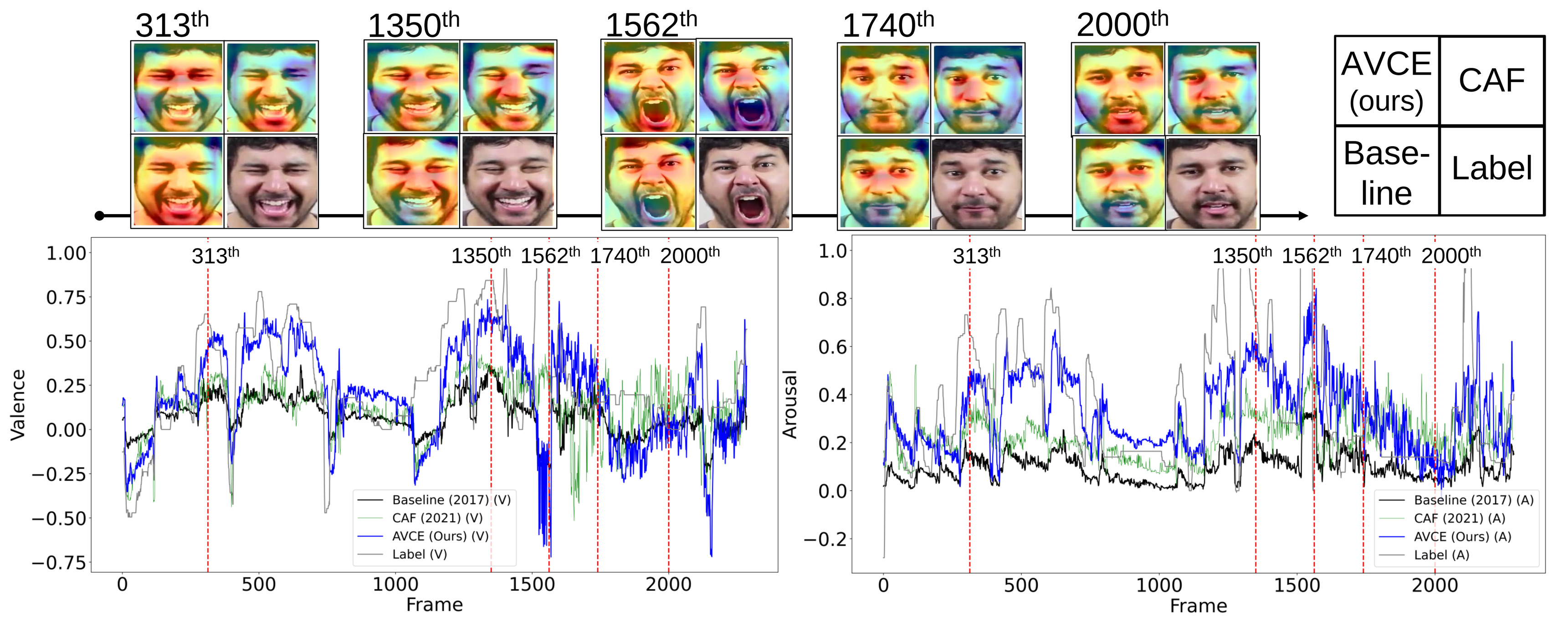
Experiments

✓Quantitative results on Aff-wild dataset

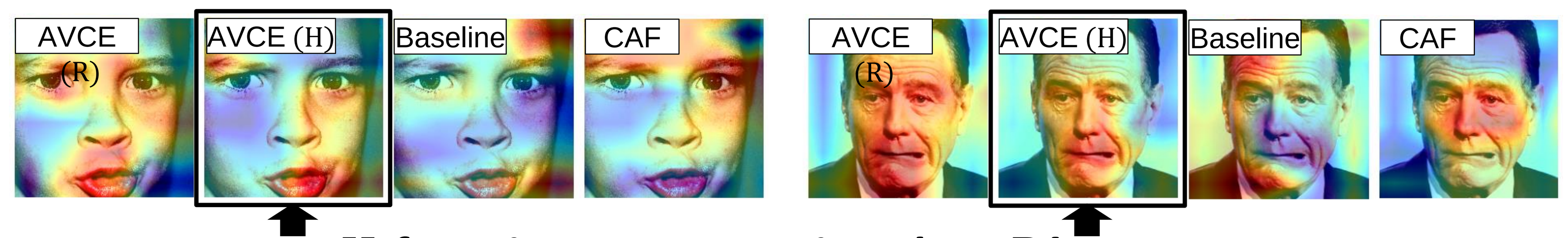
Methods	RMSE-V	RMSE-A	SAGR-V	SAGR-A	CCC-V	CCC-A
Hasani *et al.* [1]	0.27	0.36	0.57	0.74	0.36	0.19
CAF (AL) [2]	0.24	0.21	0.68	0.78	0.54	0.56
AVCE (AL)	0.154	0.154	0.849	0.795	0.682	0.594

[1] B. Hasani et al., Facial affect estimation in the wild using deep residual and convolutional networks, In CVPRW, 2017.
[2] D. Kim and BC Song, Contrastive adversarial learning for person independent facial emotion recognition, In AAAI, 2021.

✓Frame unit emotional fluctuations w/ mean neural act. maps



✓Influence analysis of self-supervision



H focusing sparse region than R!

✓Additional neural activation maps for each axis