From Pixels to Semantics: Unified Facial Action Representation Learning for Micro-Expression Analysis

Yicheng Deng, Hideaki Hayashi, and Hajime Nagahara

The University of Osaka

ICLR 2026 Poster

大阪大学
The University of Osaka

ICLR

Micro-Expression Recognition




Onset Apex Offset

Given an ME clip, we aim to recognize the emotion type

- Happiness;
- Disgust;
- Surprise;
- …
- Others. (ambiguous expressions)

Motivation

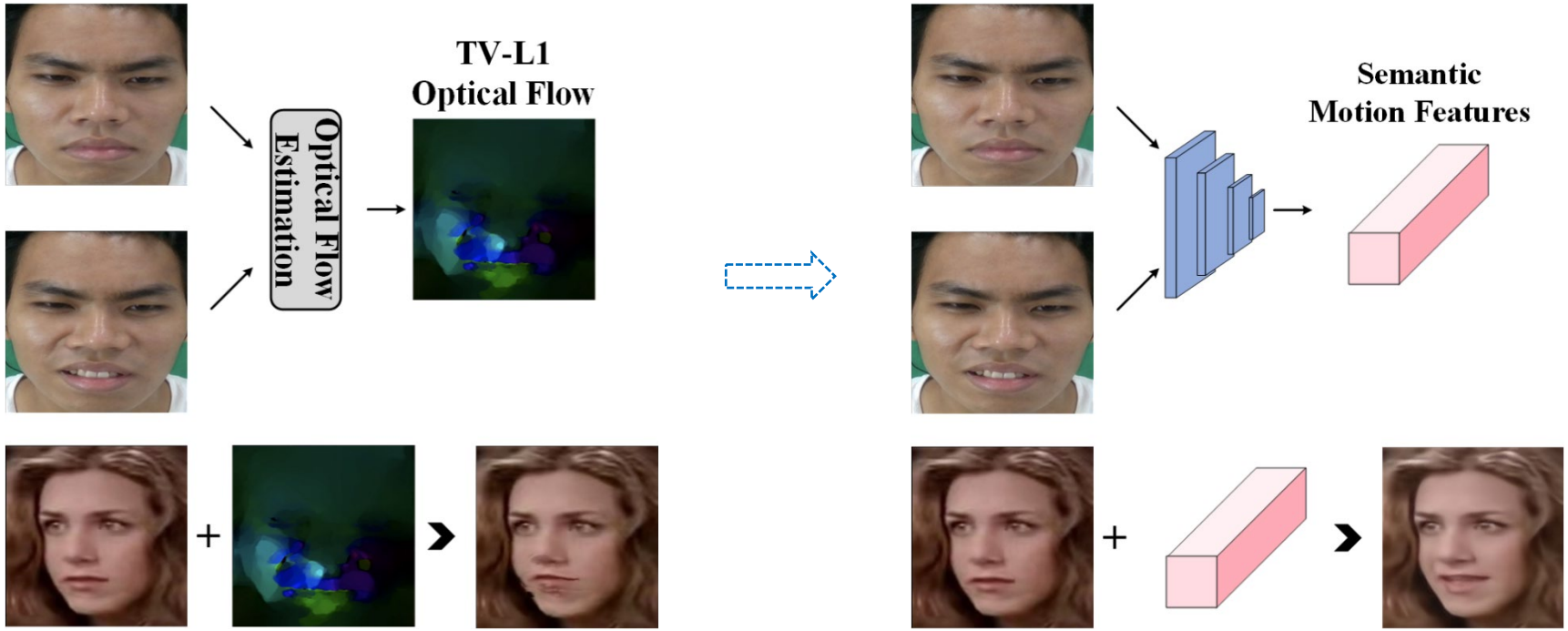
Typical existing methods rely on ***pixel-level motion descriptors***, especially ***optical flow***.

We do ***NOT*** care about ***how each pixel moves***.

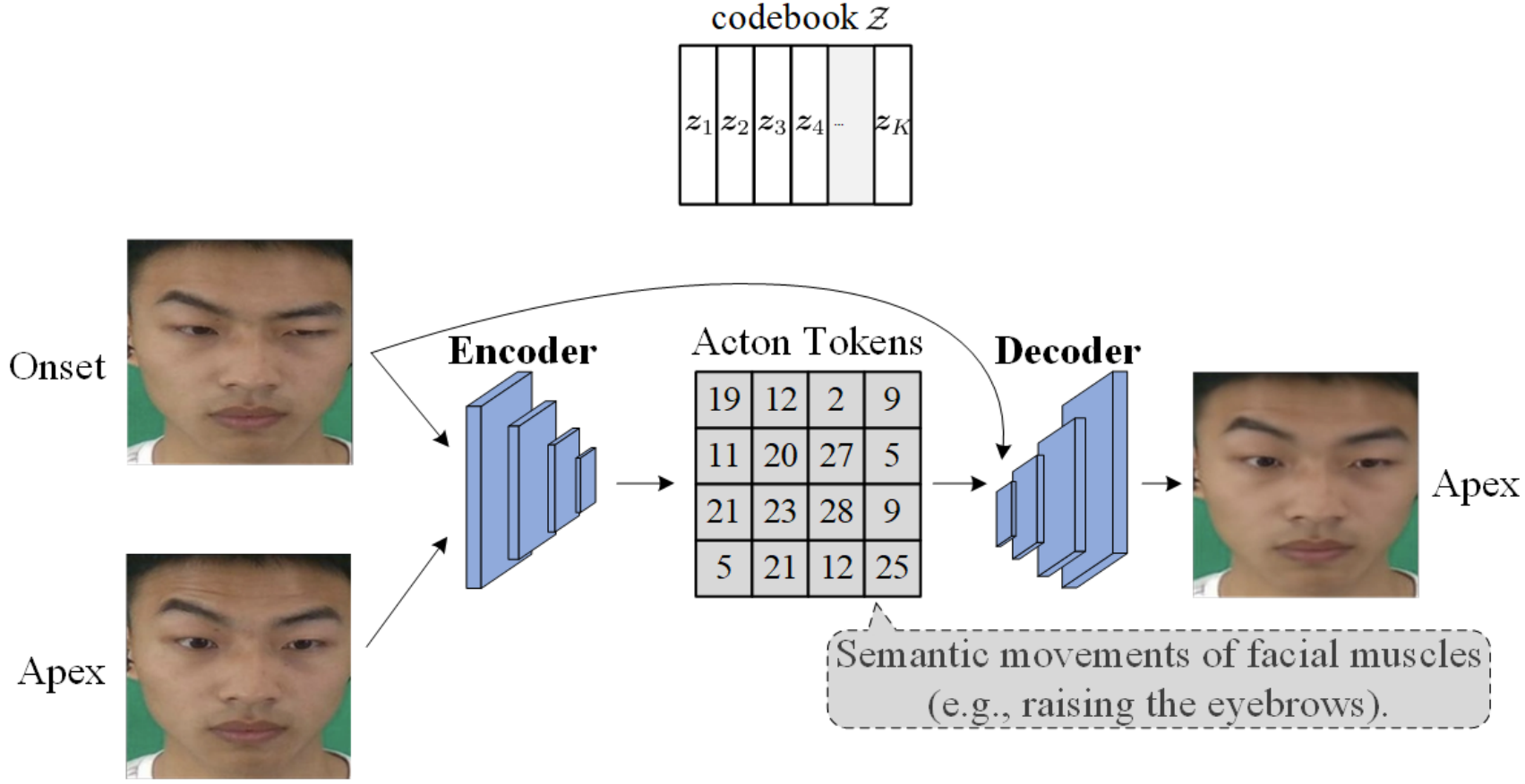
We just need to know ***what facial action occurs*** between onset and apex frames.

Such a semantic motion feature:

- is identity- and even domain-invariant;
- can be generalized to a new face to generate new MEs.



Conditional Vector-Quantized Variational Autoencoder



Limited codebook (32 in our method) is used to encode any possible facial actions.

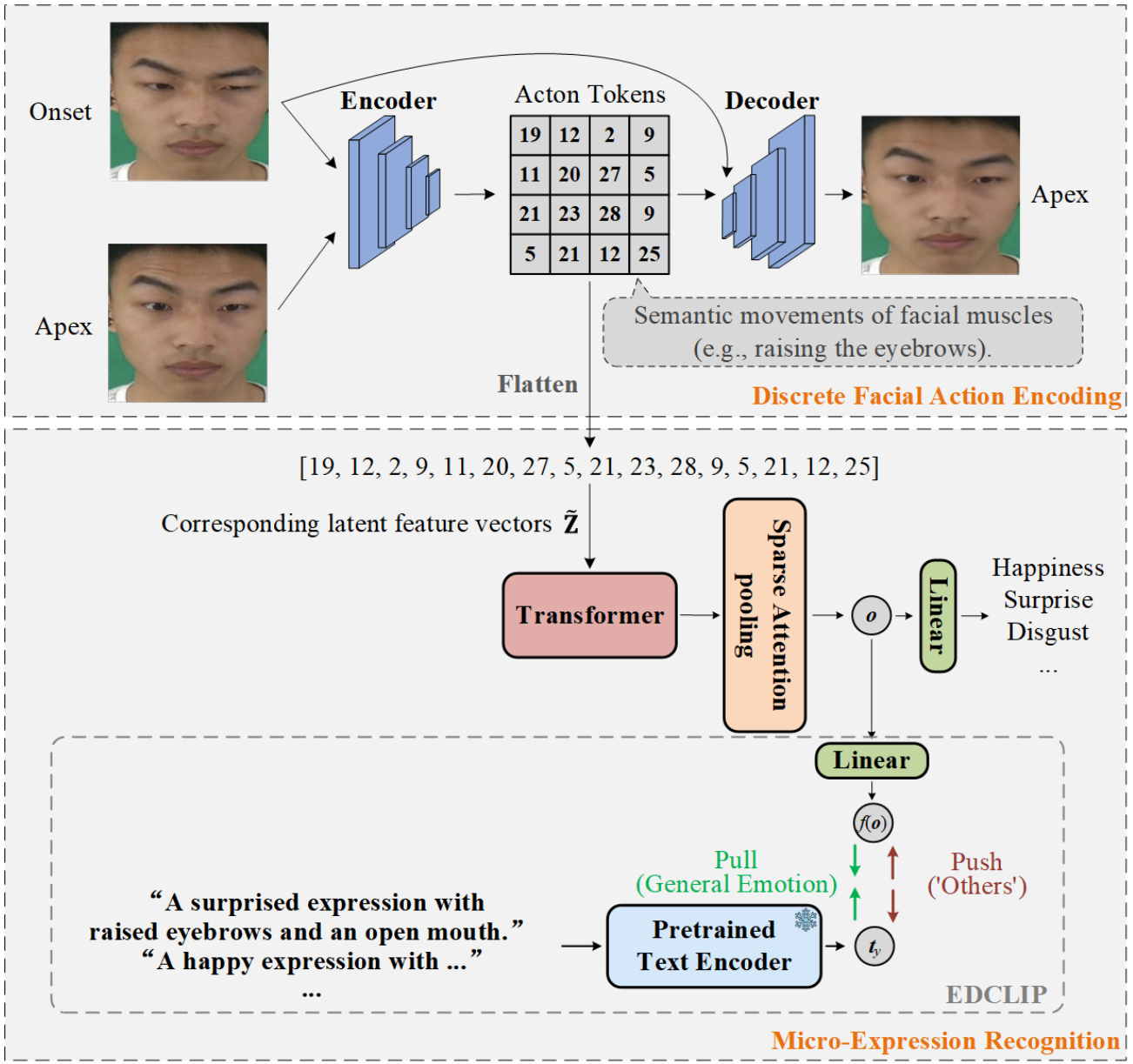
With the pretraining from large-scale facial action data (VOX-Celeb),

feature vectors in the codebook provide **identity- and domain-invariant** facial action representation.

The Exploration of Action Tokens

- One token in **different position** activates **different facial actions;**
- Tokens in **every region** could control **both upper and lower** muscle movements;
  - 2D spatial structure is not so important, but 1D order is crucial.

Overall Framework



C-VQ-VAE for facial action representation extraction

1D Transformer with sparse attention pooling for classification

EDCLIP for bridging action tokens with human-understandable emotions

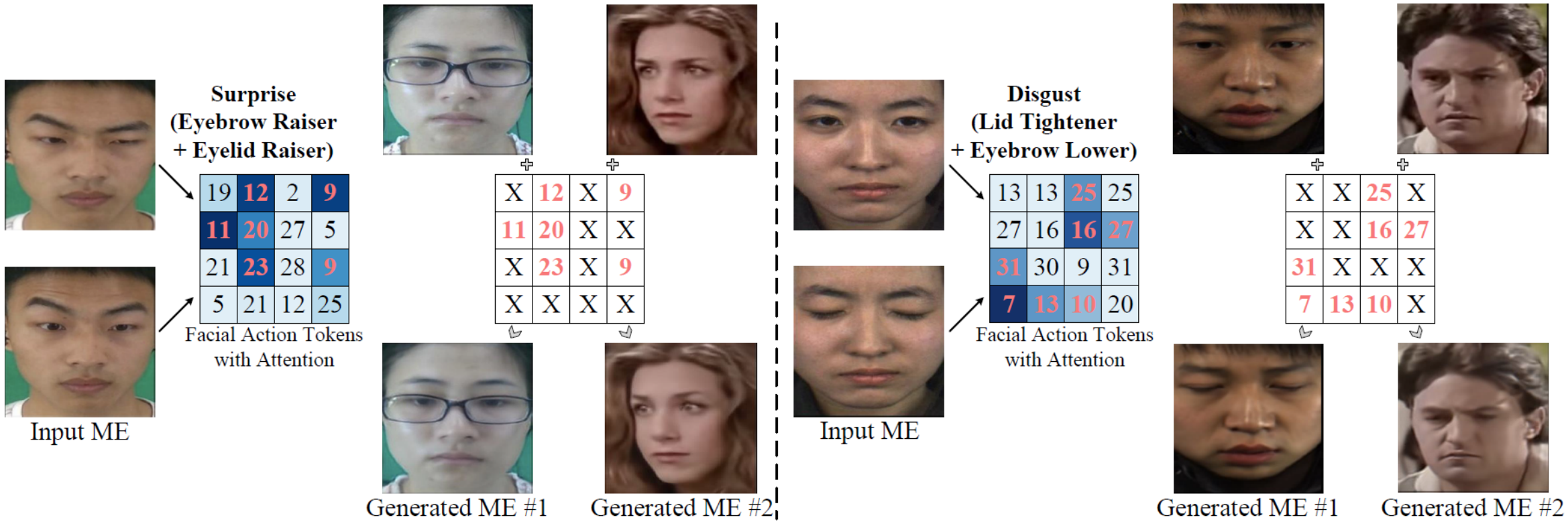
Experimental Results

Methods	Full		CASME-II		SMIC-HS		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
OFF-ApexNet (Gan et al., 2019)	0.7196	0.7096	0.8764	0.8681	0.6817	0.6695	0.5409	0.5392
STSTNet (Liong et al., 2019)	0.7353	0.7605	0.8382	0.8686	0.6801	0.7013	0.6588	0.6810
Graph-AU (Lei et al., 2021)	0.7914	0.7933	0.8798	0.8710	0.7192	0.7215	0.7751	0.7890
SLSTT (Zhang et al., 2022)	0.8160	0.7900	0.9010	0.8850	0.7400	0.7200	0.7150	0.6430
FRL-DGT (Zhai et al., 2023)	0.8120	0.8110	0.9190	0.9030	0.7430	0.7490	0.7720	0.7580
SRMCL (Bao et al., 2024)	0.8630	0.8830	0.9635	0.9649	0.7946	0.8053	0.8470	**0.8866**
MFDAN (Cai et al., 2024)	0.8453	0.8688	0.9134	0.9326	0.6815	0.7043	0.7871	0.8196
HTNet (Wang et al., 2024)	0.8603	0.8475	0.9532	0.9516	0.8049	0.7905	0.8131	0.8124
LTR3O (Zhu et al., 2025)	0.8931	0.8819	0.9578	0.9487	0.8336	0.8298	**0.8912**	0.8526
Ours	**0.8943**	**0.8967**	**0.9738**	**0.9754**	**0.8422**	**0.8476**	0.8716	0.8513

Results on composite dataset for 3-class recognition.
positive, negative, surprise

Gray-scale images.

Experimental Results



Tokens highlighted by the attention mechanism can be applied to new facial images.

From Pixels to Semantics: Unified Facial Action Representation Learning for Micro-Expression Analysis

Yicheng Deng, Hideaki Hayashi, and Hajime Nagahara

The University of Osaka

Thank you!

大阪大学
The University of Osaka

ICLR