Self-supervised learning of a facial attribute embedding from video

BMVC'18 (Oral)

2020. 02. 18

Presented by Junsoo Lee

Motivation

- Video data contain a large collection of images of the same person from different viewpoints and with varied expressions.
- We take advantage of these data to learn an low-dimensional embedding of attributes that encodes landmarks, pose, emotion in a self-supervised manner **without any** manually-annotated labels.

What to do

- The learned embedding can then be used for another tasks (downstream tasks) using a linear layer.

  => Representation Learning

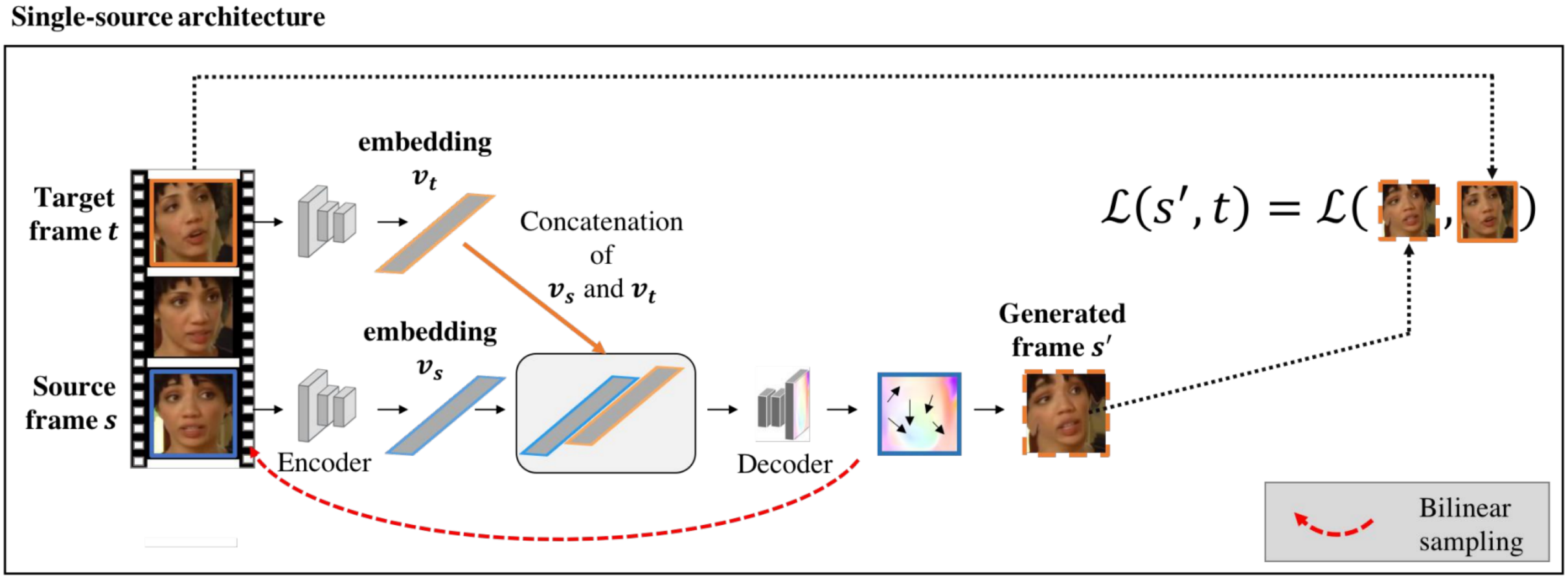
To do this

- Given only the embeddings corresponding to a source and target frame, the networks predict a flow field between the source frame and the target frame.
- This proxy task encourages the networks to distil the information, e.g., the head pose and expression, into the source and target embeddings.

Contributions

- This paper presents that the network can map an image to a meaningful representation learning by self-supervised manner without labels.
- This paper demonstrates that the network can leverage information from multiple source frames by predicting confidence/attention masks for each frame.
- Authors show that using a curriculum learning regime improves the learned embedding.

Method: single-source architecture



- $s'(x, y) = s(x + \delta x, y + \delta y)$

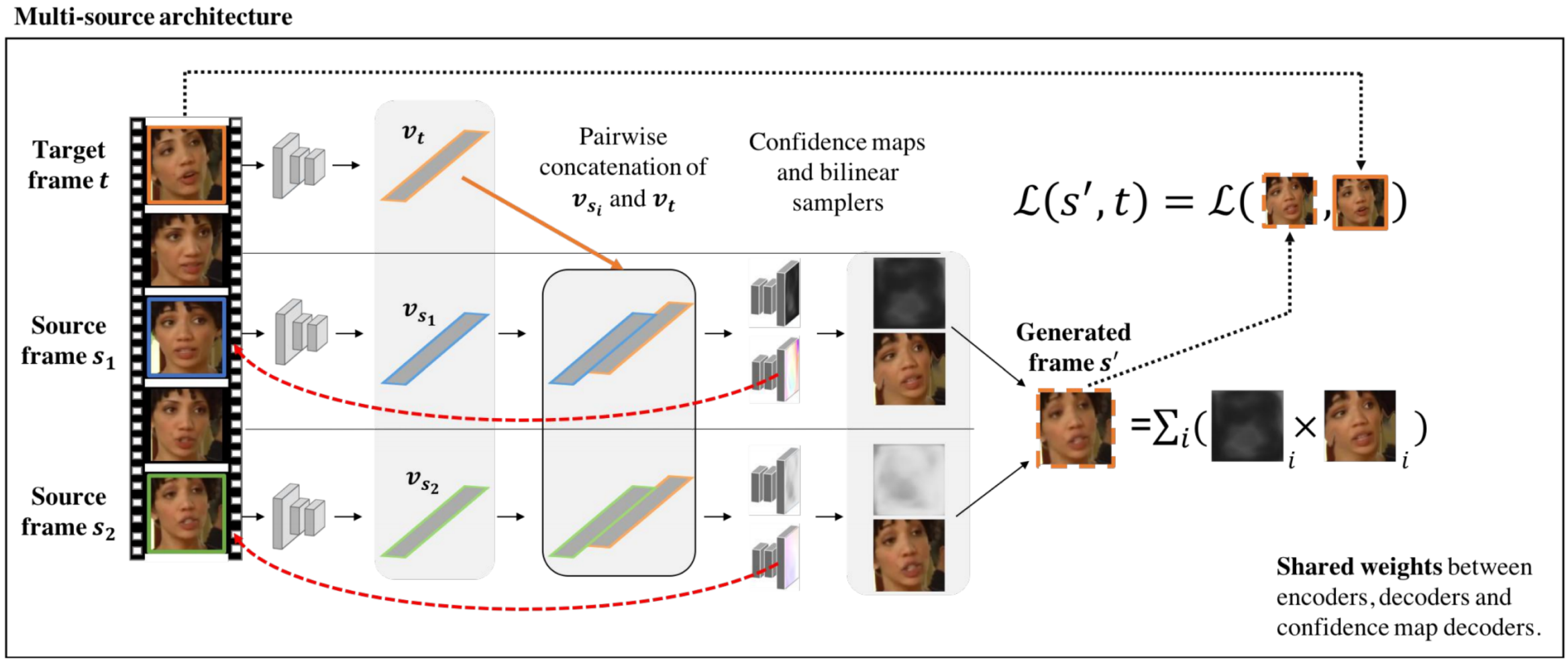
Method: single-source architecture

- The decoder learns a mapping g from the concatenated embeddings to a bilinear grid sampler, which samples from the source frame to create a new, generated frame $s' = g(v_t, v_s)(s)$.
- Precisely, g predicts offsets $(\delta x, \delta y)$ for each pixel location.

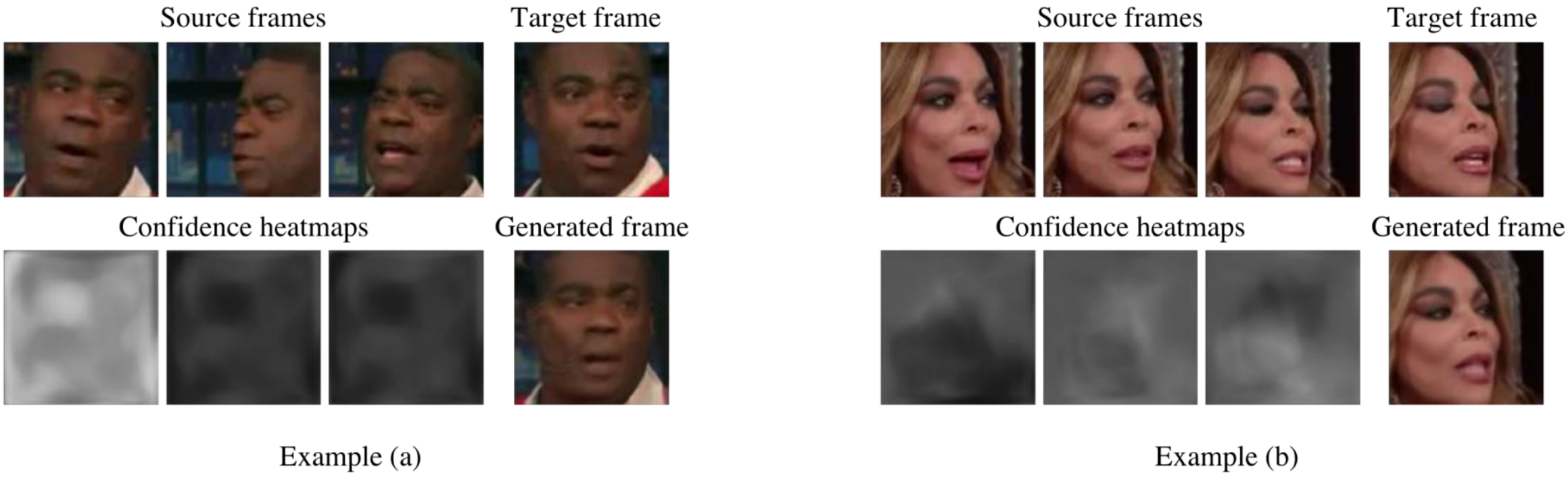
Method: single-source architecture

- The target attribute embedding must encode information about expression and pose in order for the decoder to know where to sample from in the source frame and where to place this information in the generated frame.

Method: multi-source architecture



Method: multi-source architecture



ıre 4: Confidence heatmaps learned by FAb-Net. Higher intensity corresponds to higl

$$\mathcal{L} = ||t - \frac{\sum_{i=1}^{n} e^{C_i} * (g(v_t, v_{s_i})(s_i))}{\sum_{i=1}^{n} e^{C_i}}||_1.$$

Curriculum strategy (1)

- Curriculum strategy divides the training process into several stages so that knowledge can be built up over time as the given examples become progressively more difficult.
- To do this, authors use the loss computed by a forward pass to rank samples retrospective to the difficulty.

Curriculum strategy (2)

- Given randomly chosen N samples in a batch, the loss of the each sample is computed during a forward pass.
- The samples are sorted according to this loss.
- Initially, the loss is back-propagated only on the samples in the batch which are in the 50^{th} percentile. (These are assumed to be easier samples)
- The subset is shifted by 10. (repeated 4 times)

Quantitative results: facial landmark, pose detection

Method	300-W	MAFL
Self-supervised		
Trained on VoxCeleb+		
FAb-Net	6.31	3.78
FAb-Net w/ curric.	5.73	3.49
FAb-Net w/ curric., 3 source frames	**5.71**	3.44
Trained on CelebA		
Jakab *et al.* [21] (2018)	–	**3.08**
Zhang *et al.* [64] (2018)	–	3.15
Thewlis *et al.* [51] (2017)	9.30	6.67
Thewlis *et al.* [52] (2017)	7.97	5.83
Supervised		
Trained on CelebA		
MTCNN [65] (2014)	–	**5.39**
LBF [45] (2014)	6.32	–
CFSS [69] (2015)	5.76	–
cGPRT [29] (2015)	5.71	–
DDN [60] (2016)	5.65	–
TCDCN [66] (2016)	5.54	–
RAR [57] (2016)	**4.94**	–
VGG-Face descriptor [41]	11.16	5.92

Table 1: Landmark prediction error on 300-W and MAFL datasets. Lower is better.

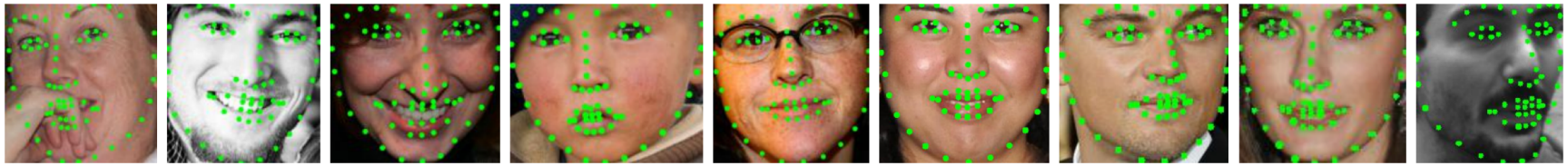
Method	Roll	Pitch	Yaw	MAE
Self-supervised				
FAb-Net	5.54°	7.84°	12.93°	8.77°
FAb-Net w/ curric.	5.33°	7.21°	11.34°	7.96°
FAb-Net w/ curric., 3 source frames	**5.14°**	**7.13°**	**10.70°**	**7.65°**
Supervised				
VGG-Face descriptor [41]	**8.24°**	8.36°	18.35°	11.65°
KEPLER [26] (2017)	8.75°	**5.85°**	**6.45°**	**7.02°**

Table 2: Pose prediction error on the AFLW test set from [26]. Lower is better.

Qualitative results: facial landmark

Figure 2: Landmark prediction visualisation for FAb-Net on the MAFL dataset. A dot denotes ground truth and the cross FAb-Net's prediction. A failure case is shown to the right.

(a) Ground truth landmarks



(b) FAb-Net's predicted landmarks

Figure 3: Landmark prediction visualisation for FAb-Net on the 300-W dataset.

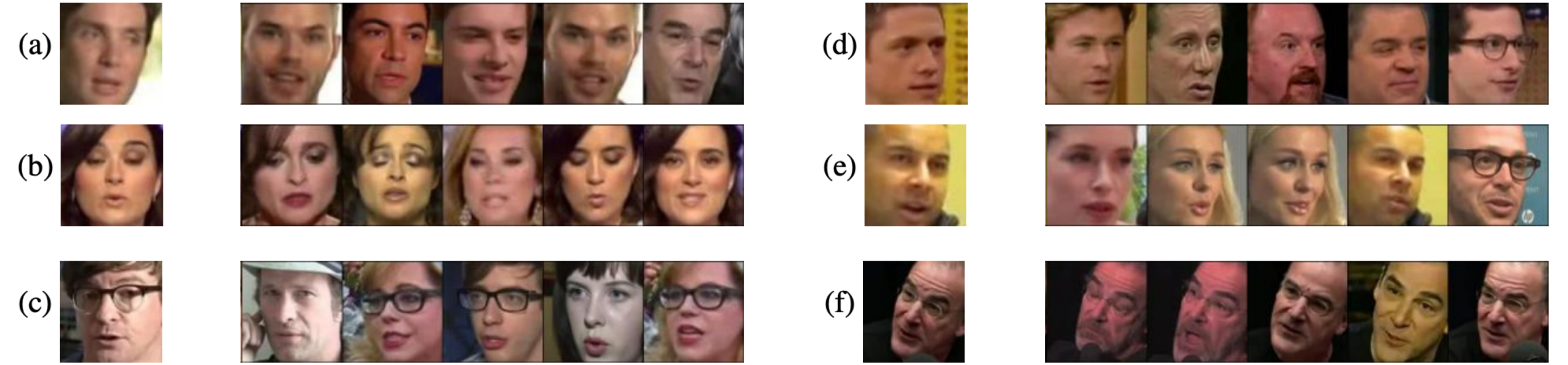
Quantitative results: expression classification

	AUC for different AUs											
	1	2	4	5	6	9	12	17	20	25	26	avg.
Self-supervised												
FAb-Net	72.0	68.9	73.2	69.4	88.2	78.6	89.5	71.0	75.9	81.4	72.0	76.4
FAb-Net w/ curriculum	73.4	71.8	75.3	67.8	90.4	78.8	91.9	72.4	74.5	**83.7**	73.3	77.6
FAb-Net w/ curriculum, 3 source frames	**74.1**	**72.3**	**75.8**	**68.8**	**90.7**	**81.8**	**92.5**	**73.7**	**77.2**	83.6	**73.6**	**78.6**
Gidaris *et al.* [19]	68.6	64.0	72.8	70.0	83.9	78.1	83.8	68.4	72.6	73.1	67.2	72.9
SplitBrain [63]	65.5	59.8	66.7	60.8	71.8	65.8	73.3	64.5	57.4	68.1	61.1	65.0
Autoencoder	67.2	60.5	70.1	65.1	79.6	70.4	80.1	68.3	66.5	70.5	64.1	69.3
Supervised												
VGG-Face descriptor [41]	**81.8**	**83.0**	83.5	81.8	92.0	**90.9**	95.7	**80.6**	85.2	86.5	73.0	84.9
VGG-11 (from scratch)	74.7	77.2	**85.8**	**83.7**	**93.8**	89.7	**97.5**	78.3	**86.9**	**96.4**	**81.5**	**86.0**

Table 3: Expression classification results for state-of-the-art self-supervised and supervised methods on EmotioNet [6] for multiple facial action units (AUs). Higher is better for AUC.

Application: image retrieval

- Applications of the learned embedding: retrieving images with similar pose but across different identities.



Thank you!