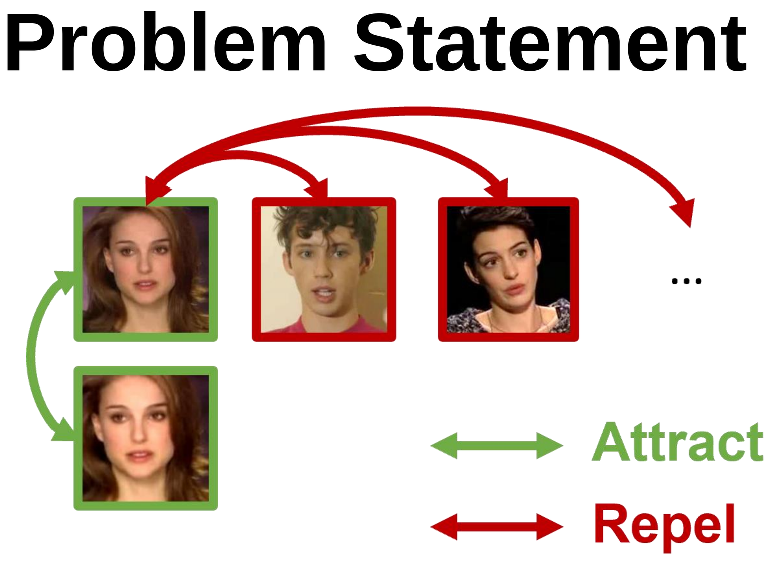
Revisiting Self-Supervised Contrastive Learning for Facial Expression Recognition

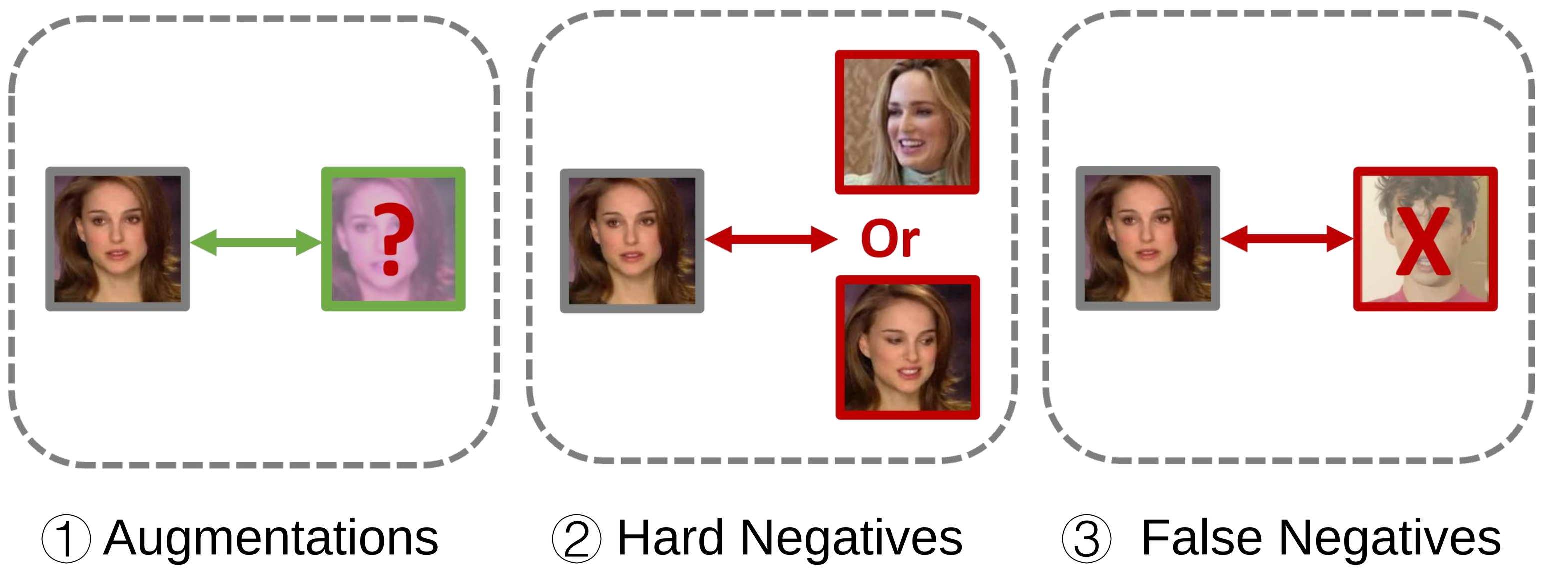
Yuxuan Shu[1], Xiao Gu[1] , Guang-Zhong Yang[2] & Benny Lo[1]
1 The Hamlyn Centre, Imperial College London, London, United Kingdom
2 The Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China

Imperial College London
The Hamlyn Centre
The Institute of Global Health Innovation

BMVC 2022
The 33rd British Machine Vision Conference
21st - 24th November 2022, London, UK

Problem Statement





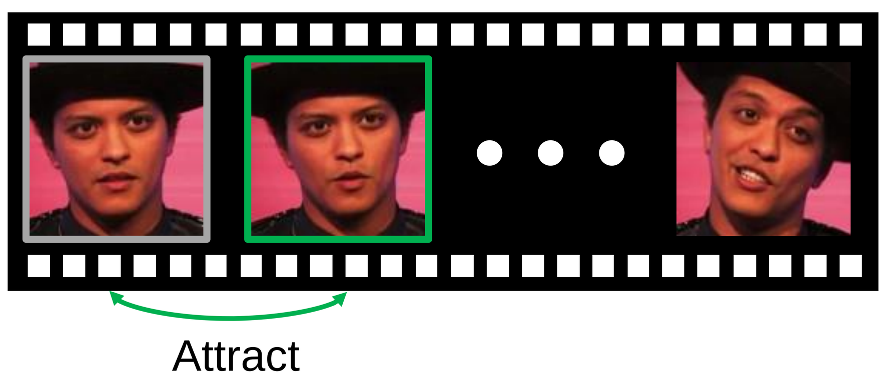
① What augmentation serves better for facial expression recognition?

② What act better as negative pairs?

③ How to reduce false negative pairs during pre-training stage?

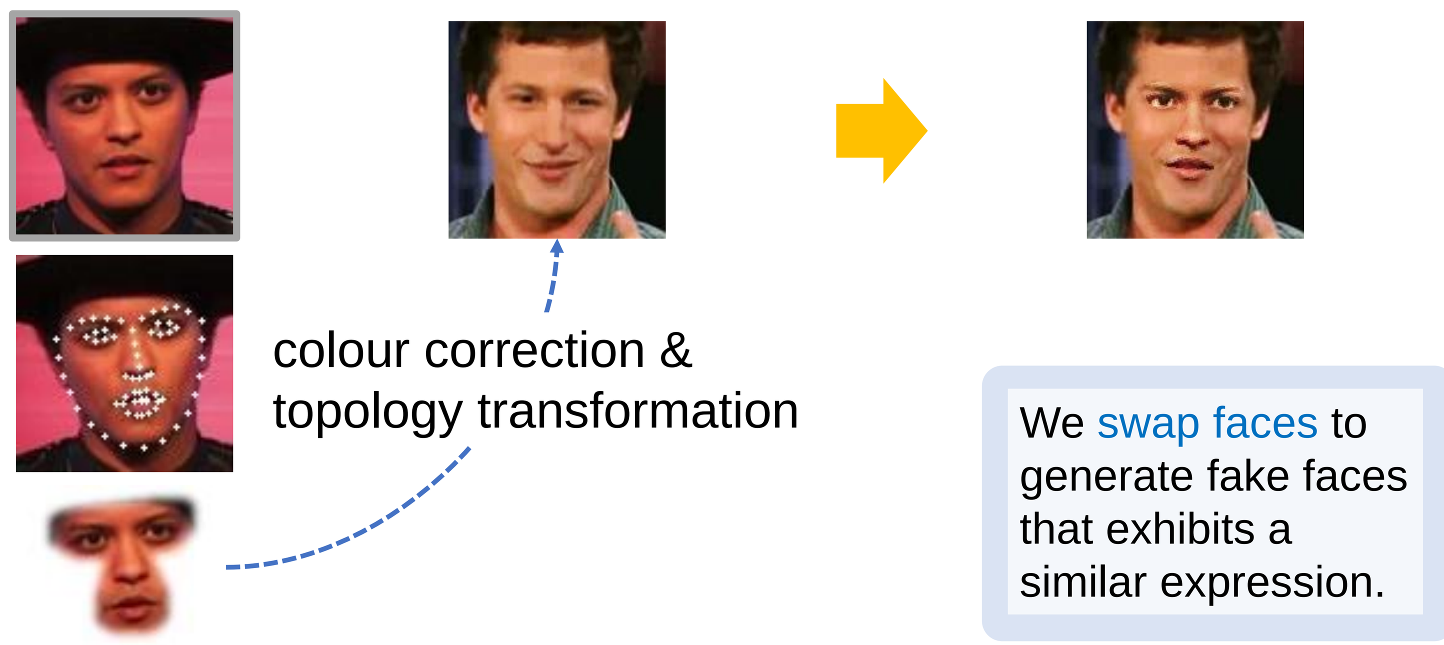
① Positives with Same Expression

Temporal Augmentation (TimeAug)



We sample images along the time domain for augmentation.

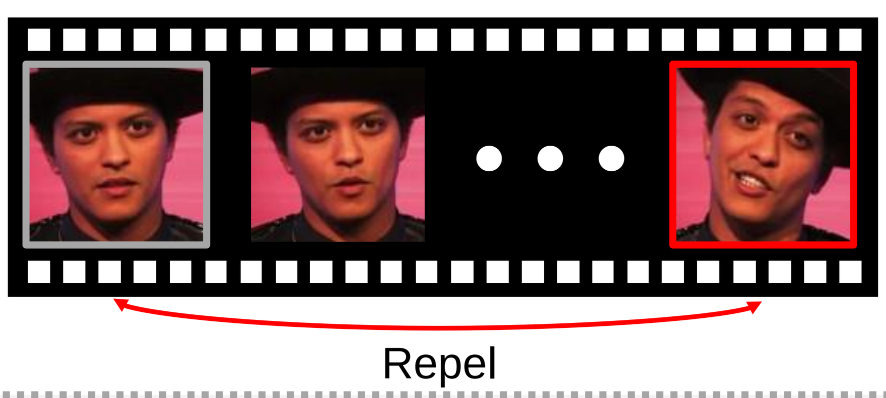
Face Swap (FaceSwap)



We swap faces to generate fake faces that exhibits a similar expression.

② Negatives with Same Identity

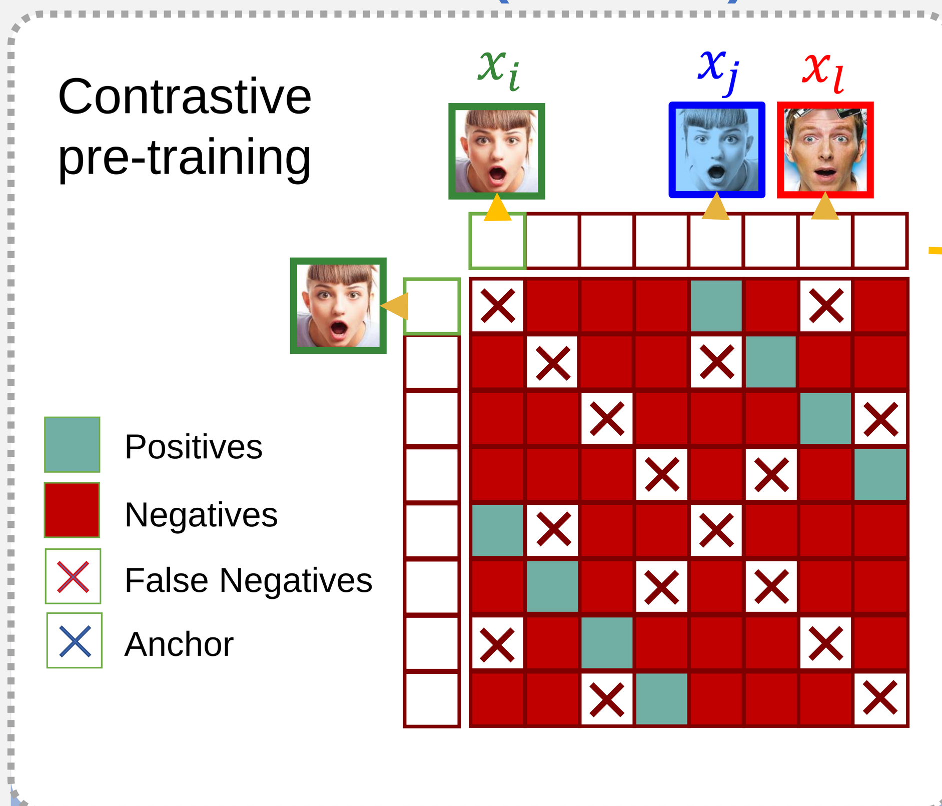
Hard Negative Sampling (HardNeg)



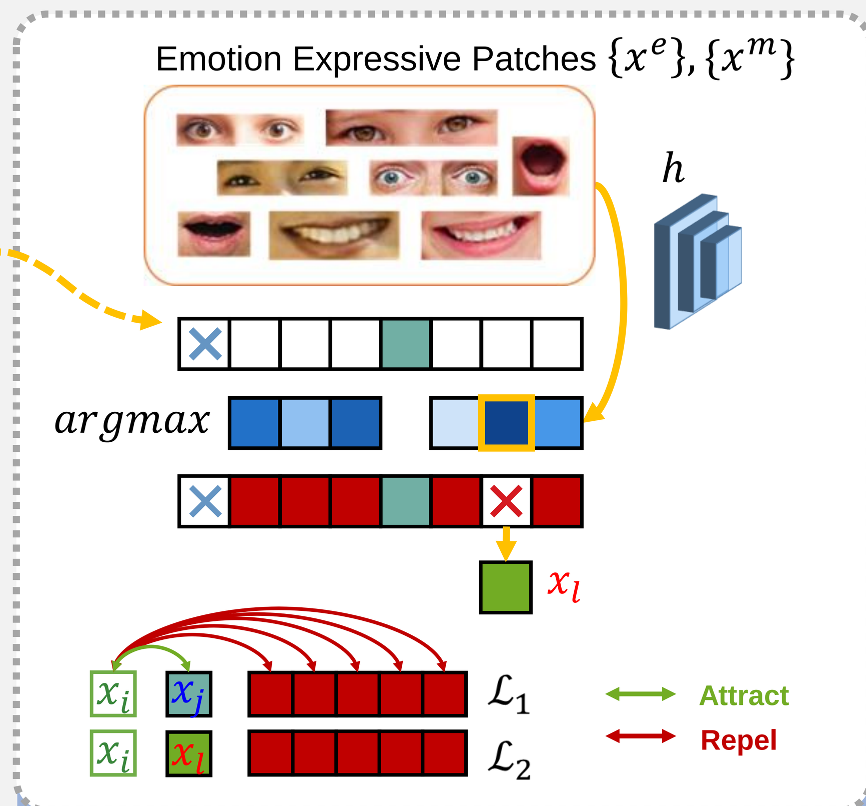
We sample images with large time interval as "hard negative".

③ False Negatives Cancellation

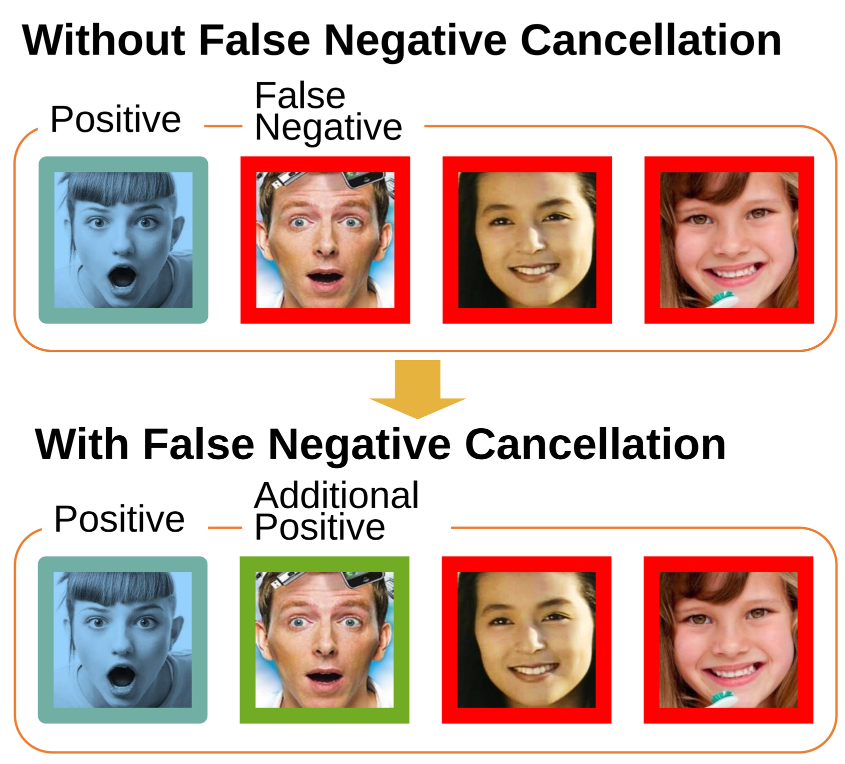
Model structure (MaskFN)



In **contrastive-learning pre-training** stage, false negative pairs are difficult to avoid without the actual labels.



Emotion expressive patches – Eyes and mouths are cropped as mouth-eye descriptors.

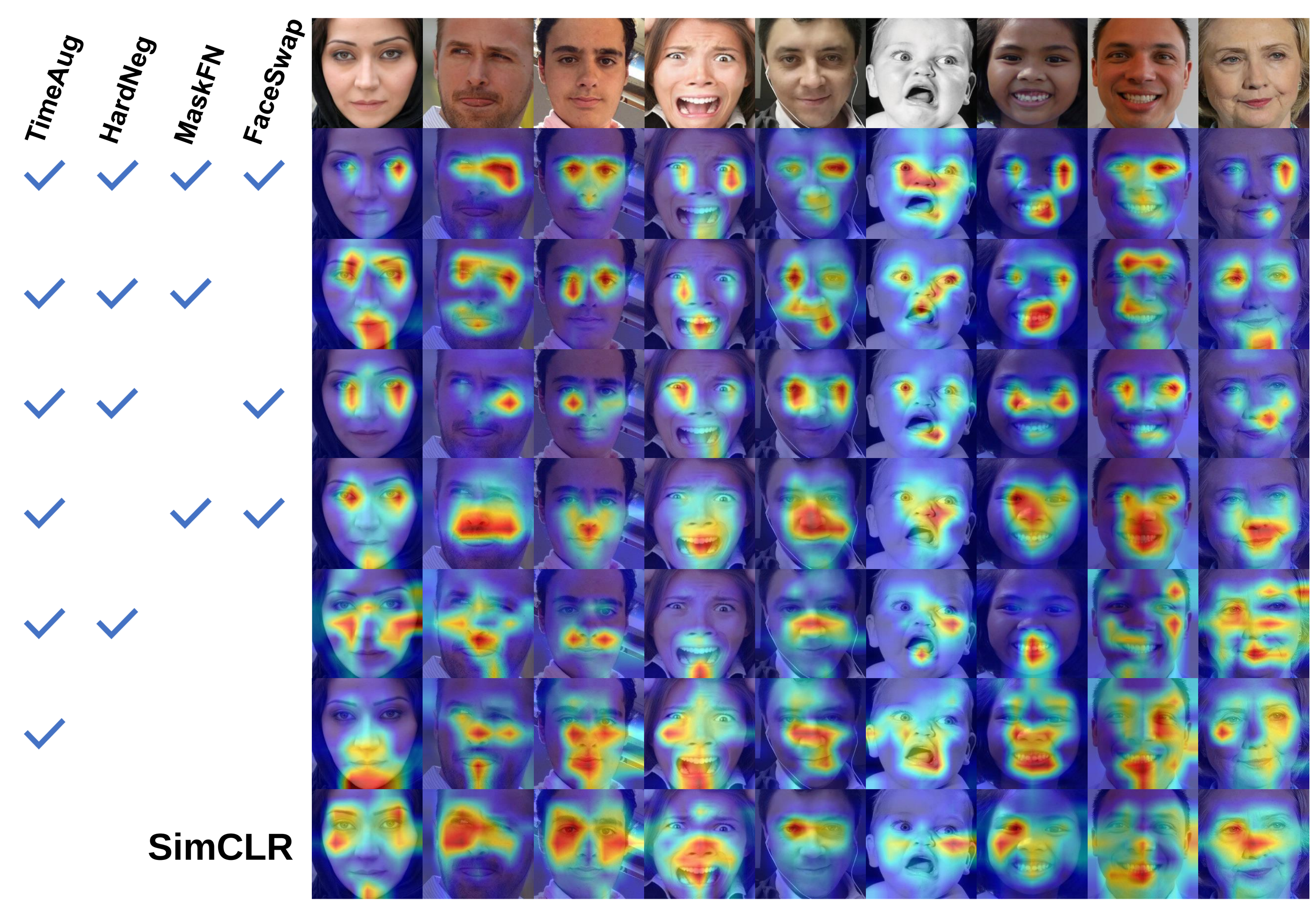


- Use the feature of mouth-eye descriptors (extracted from fixed ResNet18) as similarity indicator.
- Pick the sample with the highest similarity to the anchor and consider it as a false negative.
- With this false negative cancellation strategy, we are able to select and mask the instances that are more likely to be false negatives.

Experiment Result

Pretraining Methods					Dataset	EXPR		Valence		Arousal	
						F1↑	Acc↑	CCC↑	RMSE↓	CCC↑	RMSE↓
Supervised					ImageNet	56.7%	56.6%	0.563	0.462	0.480	0.376
BYOL					VoxCeleb1	56.3%	56.4%	0.560	0.460	0.462	0.386
MoCo-v2					VoxCeleb1	56.8%	56.8%	0.570	0.454	0.486	0.378
SimCLR					VoxCeleb1	57.5%	57.7%	0.594	0.431	0.451	0.387
CycleFace					VoxCeleb1,2	48.8%	49.7%	0.534	0.492	0.436	0.383
	TimeAug	HardNeg	FaceSwap	MaskFN							
Ours a	✓				VoxCeleb1	57.8%	57.9%	0.583	0.448	0.500	0.374
Ours b	✓	✓			VoxCeleb1	58.1%	58.3%	0.594	0.437	0.500	0.373
Ours c	✓	✓	CutMix		VoxCeleb1	58.3%	58.4%	0.542	0.463	0.508	0.368
Ours d	✓		✓	✓	VoxCeleb1	58.6%	58.7%	0.568	0.444	0.502	0.369
Ours e	✓	✓	✓		VoxCeleb1	58.8%	58.9%	**0.601**	**0.429**	**0.514**	**0.367**
Ours f	✓	✓		✓	VoxCeleb1	58.9%	58.9%	0.578	0.448	0.493	0.370
Ours g	✓	✓	✓	✓	VoxCeleb1	**59.3%**	**59.3%**	0.595	0.435	0.502	0.372

Results on Expression Classification and Valence & Arousal recognition. Our proposed strategies outperform both the ImageNet-pretrained model as well as other self-supervised methods, on all facial expression tasks of AffectNet.



Visualisation of the saliency map with different strategies. Our proposed strategies are able to regulate the network to focus more on the regions that are more expressive to emotions. Pictures are selected from AffectNet.

Conclusion

- We revisited the use of self-supervised contrastive learning, and proposed three complementary novel strategies to regulate the network to lean towards emotion related information.
- The experimental results have shown that our self-supervised training strategies outperform the state-of-the-art methods on downstream FER tasks, including both categorical expression classification and dimensional Valence & Arousal regression.

Contact Us

https://arxiv.org/abs/2210.03853
y.shu21@imperial.ac.uk
xiao.gu17@imperial.ac.uk

Project

Github