Discriminative and Robust Representation Learning for Facial Expression Recognition

Ming Li
ming.li@u.nus.edu
National University of Singapore

I. Introduction

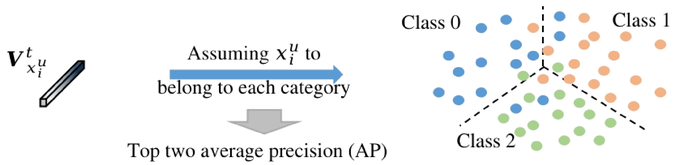
- Facial Expression Recognition (FER) targets to classify a facial image according to its emotional status.
- Local discriminative patterns are pretty important, e.g., the corner of the eyes, mouth, and nose.
- Exsiting works usually apply uncertainty modelling to tackle inconsistency of labels caused by subjective annotations.

II. Motivation

- We seek self-supervised learning (SelfSL) to assist the network to focus on informative parts.
- We employ self-paced learning (SPL) and semi-supervised learning (SemiSL) simultaneously to learn robust representation.

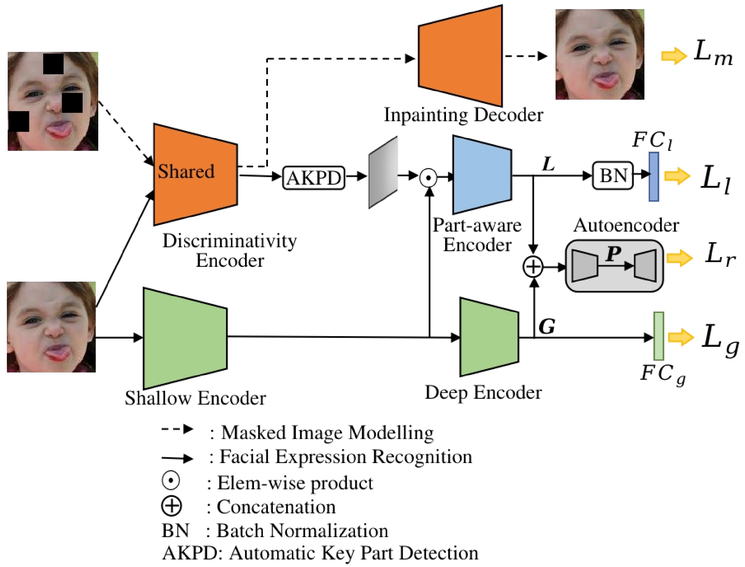
III. Methodology

- Optimizing a network to predict randomly masked image patches, i.e., Masked Image Modelling (MIM), encourages it to focus on key facial units.

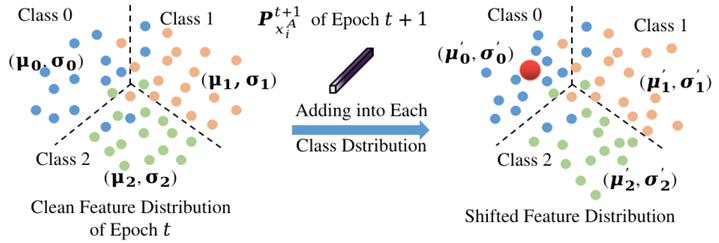


Retrieval Relabelling. Considering the class-imbalanced training data, we present a novel re-labelling method, i.e., RR, for semi-supervised learning. After the training epoch t, an embedding vector $\mathbf{V}$ is extracted for each image through l_2 normalizing the concatenation of local vector $\mathbf{L}$ and global vector $\mathbf{G}$. It retrieves a noisy sample among the clean set and adopts the average precision (AP) to measure its similarity to each class. Here we take C=3 for example.

- For self-paced learning, considering the different learning paces of different classes, we apply Class-dependent Gaussian Mixture Model to model losses of each class and separate the clean data from noisy samples.
- To learn robust representation from clean data, we propose a Distribution Contrastive Learning as the complement of cross-entropy loss.
- For semi-supervised learning, we present Retrieval Relabelling, i.e., retrieving each noisy sample among the clean data gallery and obtaining the Average Precisions, to get the soft label for each noisy image.



Overview of our framework, which consists of Local Branch and Global Branch. Our approach learns discriminative and robust FER representations simultaneously. SelfSL is applied to facilitate informative facial units detection with no more supervision. SPL and SemiSL are also performed to encourage robust representation extraction. Our network architecture incorporates MIM into discriminative FER feature learning.



Distribution Contrastive Learning. The proposed DCL regularizes representation extraction from separated clean data in a loose manner: when adding $\mathbf{P}$ into the embedding distribution of its true class, the shifts of $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ should be minimal.

III. Experiments

- We conduct experiments on four popular FER benchmarks.
- We can achieve SOTA on three testing scenarios.

Method	Venue	ExtraInfo	Backbone	RAF-DB	AffectNet7	AffectNet8	FERPlus	SFEW
gACNN (48)	TIP18	✓	VGG16	85.07	58.78	-	-	-
SCN (68)	CVPR20	✗	ResNet18	87.03	-	60.23	88.01	-
LDL-ALSG (5)	CVPR20	✓	ResNet50	85.33	58.29	-	-	55.87
RAN (69)	TIP20	✓	ResNet18	86.90	-	59.5	88.55	54.19
KTN‡ (38)	TIP21	✗	ResNet50	88.07	63.97	-	90.49	-
DACL (15)	WACV21	✗	ResNet18	87.78	65.20	-	-	-
DMUE (63)	CVPR21	✗	ResNet18	88.76	62.84	-	88.64	57.12
FDRL (61)	CVPR21	✗	ResNet18	89.47	-	-	-	**62.16**
SEIIL (47)	TCSVT21	✗	ResNet50	88.23	-	-	-	-
MANet (85)	TIP21	✗	ResNet18	88.40	64.53	60.29	-	59.40
RUL (83)	NeurIPS21	✗	ResNet18	88.98	-	-	-	-
LDL (86)	AAAI21	✗	ResNet50	-	63.70	59.89	-	-
TransFER (76)	ICCV21	✗	IR50+ViT	**90.91**	66.23	-	90.83	-
Ours	-	✗	ResNet50	90.24	**66.71**	**61.2**	**91.03**	62.06

V. Conclusion

- We are the first to successfully apply self-supervised MIM to enable discriminative FER representations learning.
- Comprehensive experiments demonstrate the effectiveness and generalization ability of our approach qualitatively and quantitatively.