RESEARCH REPORT: DISENTANGLED VAE

Presentation by Lucas Massa

PPGI | 2024

Institute of Computing | UFAL

INTRODUCTION

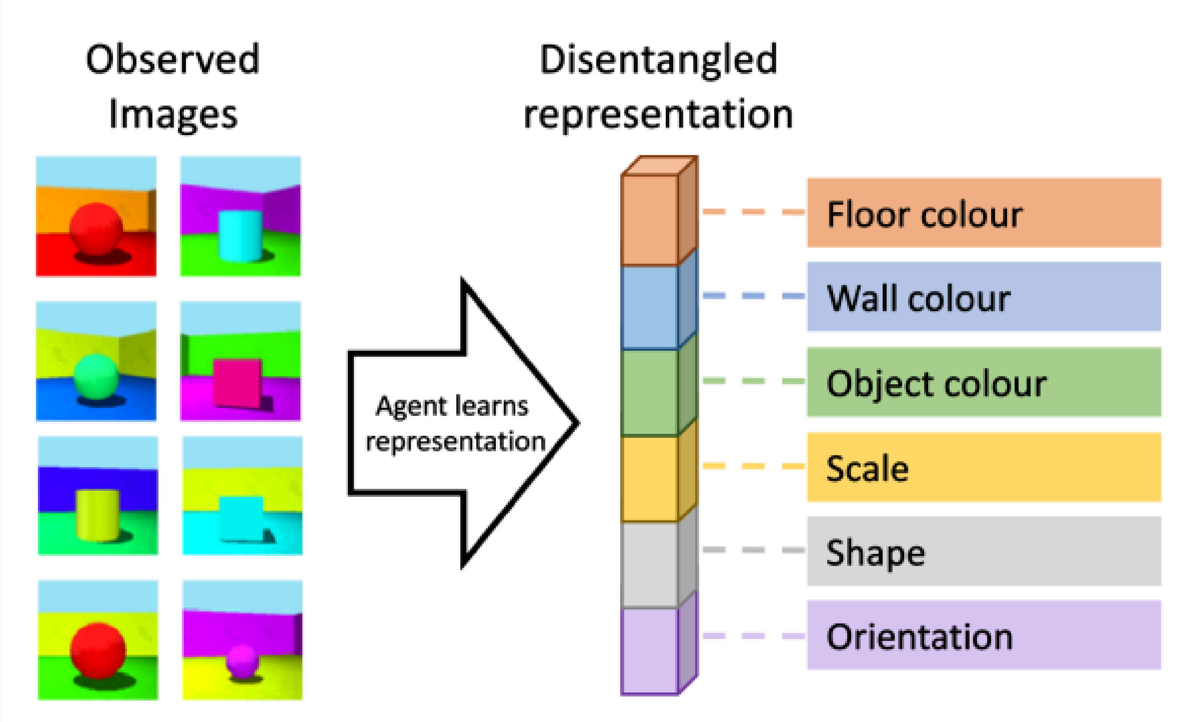
In the process of human-computer interaction, the ability to capture human emotional changes is particularly important (Wang *et. al*):

- Expression recognition and editing gained importance;
- Major challenge: high-quality expression feature extraction;
- High variations in skin color, gender, age and appearance;
- Components entangled with expression features nonlinearly.

INTRODUCTION

An expression-identity disentanglement method is of vital importance:

- Separate identity features from expression features in latent space;
- Generative adversarial methods were already applied: difficult to converge;
- Necessity of simple and effective method for facial expression tasks.

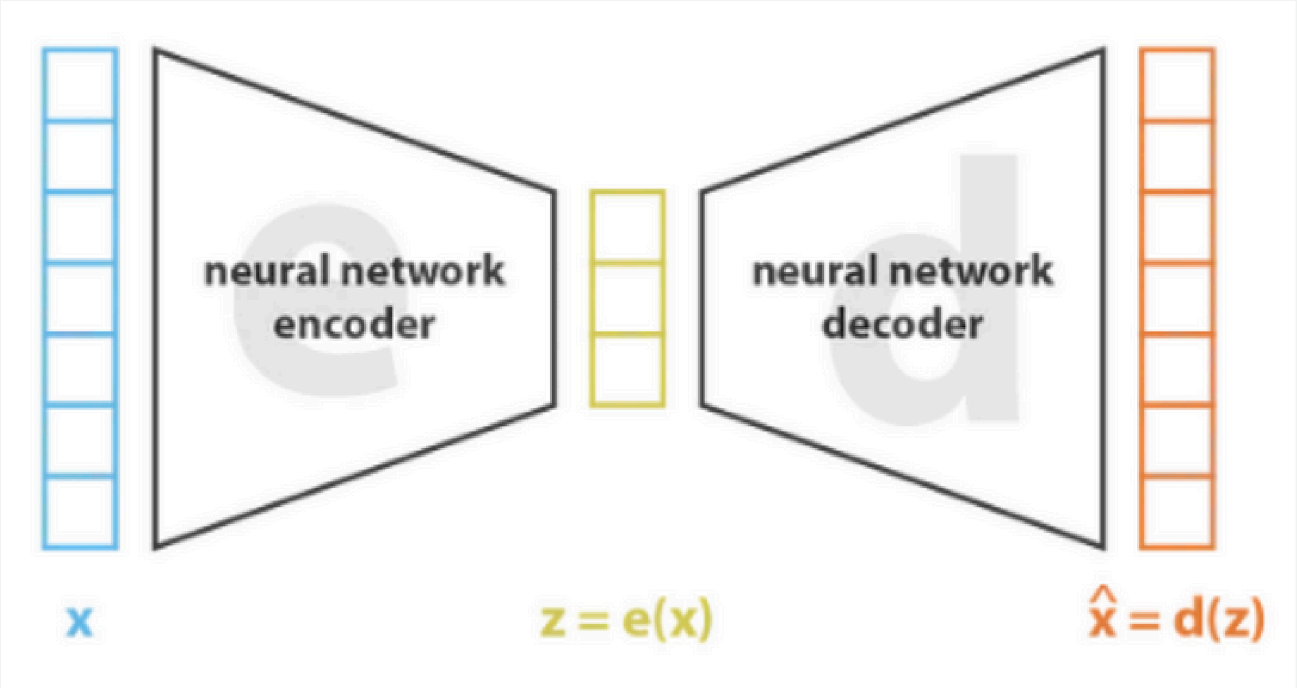


CONTRIBUTIONS

Wang *et. al* propose the Disentangled Variational Autoencoder (DisVAE) to separate expression and identity attributes:

- The proposed DisVAE can achieve explicit feature disentanglement;
- Disentangled expression features can greatly improve the performance of facial expression recognition;
- Facial expression editing can be performed by fusing identity and expression features

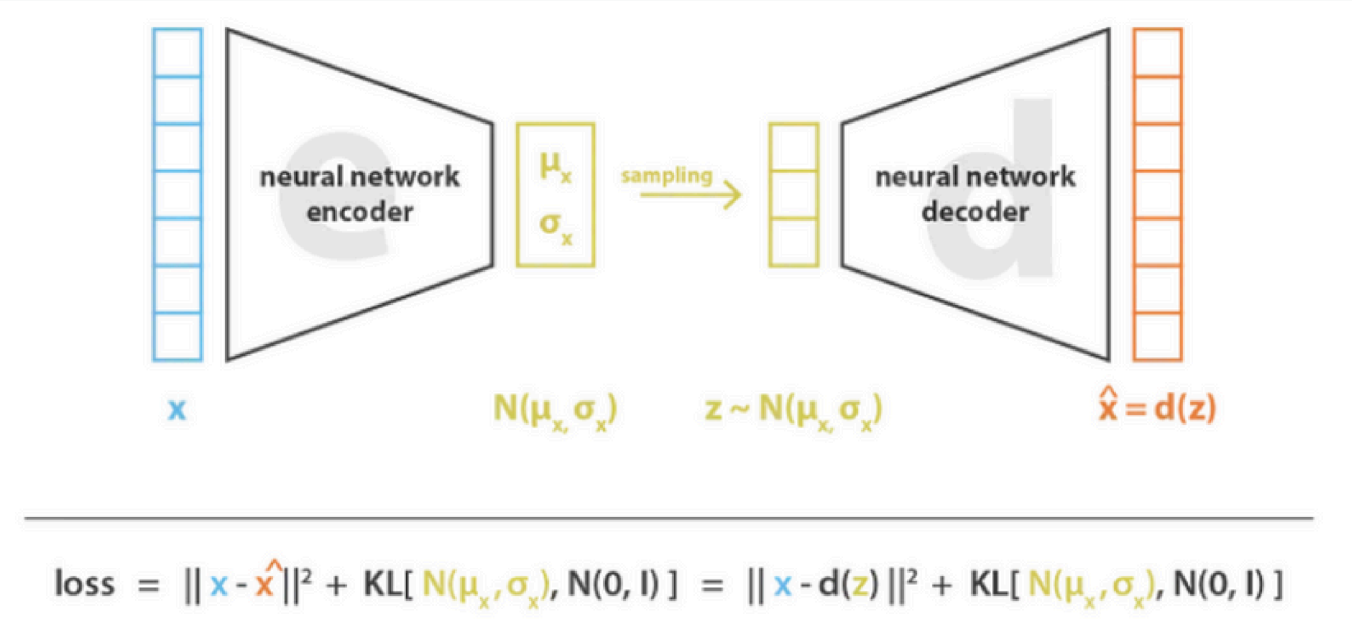
BACKGROUND



Autoencoders enable Representation Learning with neural networks:

- Various combinations of layers;
- Learn more complex patterns;
- Latent Space: where intermediary representations are projected;
- Reconstruction Loss Function.

BACKGROUND



VAE are a probabilistic version of conventional Autoencoder:

- Latent space learns a probability distribution;
- More organized latent space;
- Loss function comprised of Reconstruction and KL Divergence.

LEARNING APPROACHES

UNSUPERVISED

Does not receive any supervision. Only information is input data and output target.

SEMI-SUPERVISED

Receives weak supervision with class related labels. Uses this information to learn similar intra-class features.

LEARNING APPROACHES

UNSUPERVISED

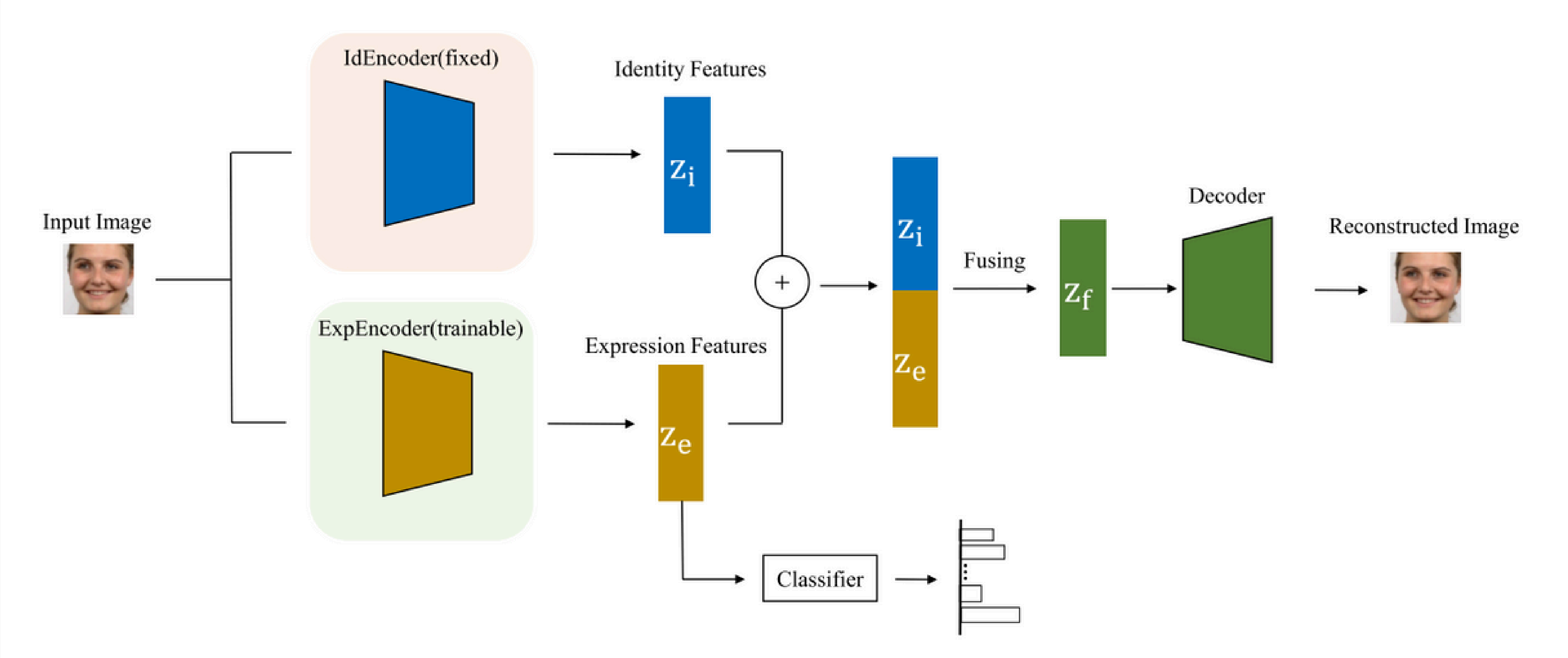
Does not receive any supervision. Only information is input data and output target.

SEMI-SUPERVISED

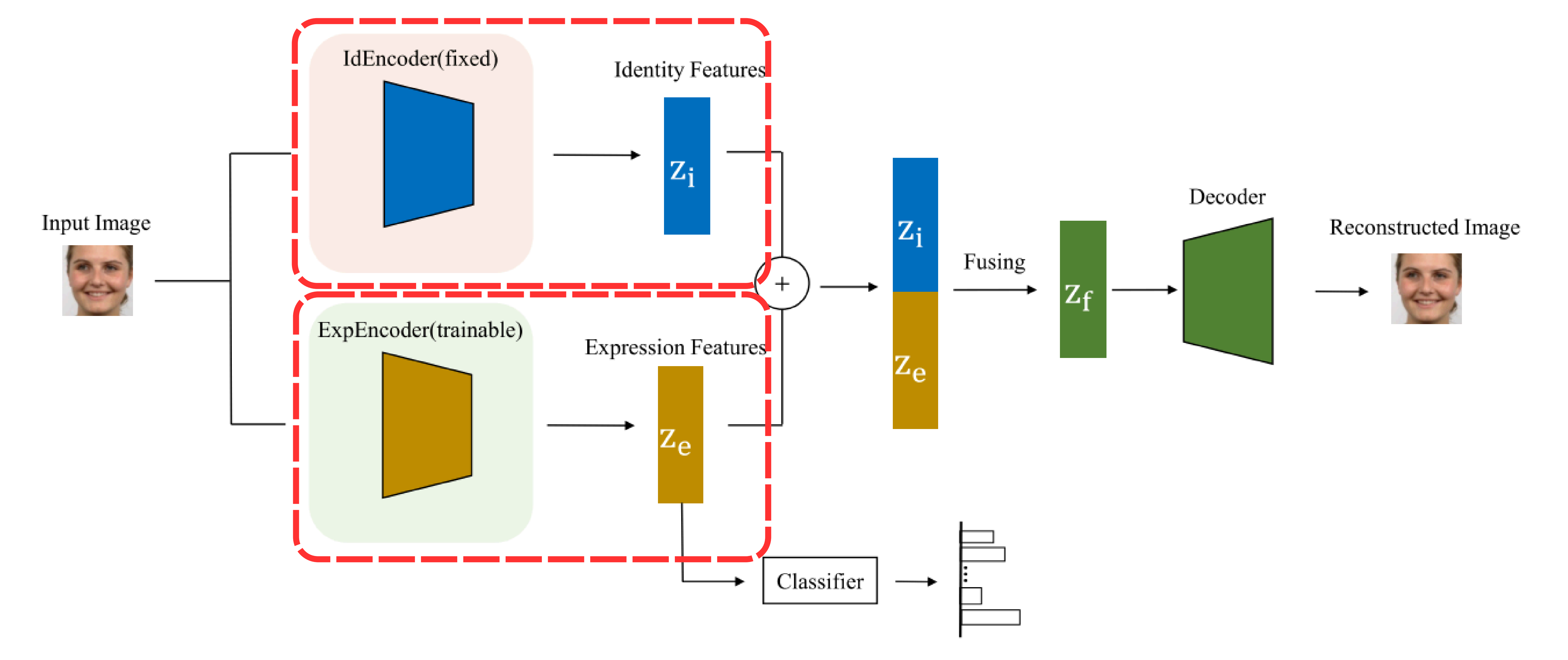
Receives weak supervision with class related labels. Uses this information to learn similar intra-class features.

METHODOLOGY

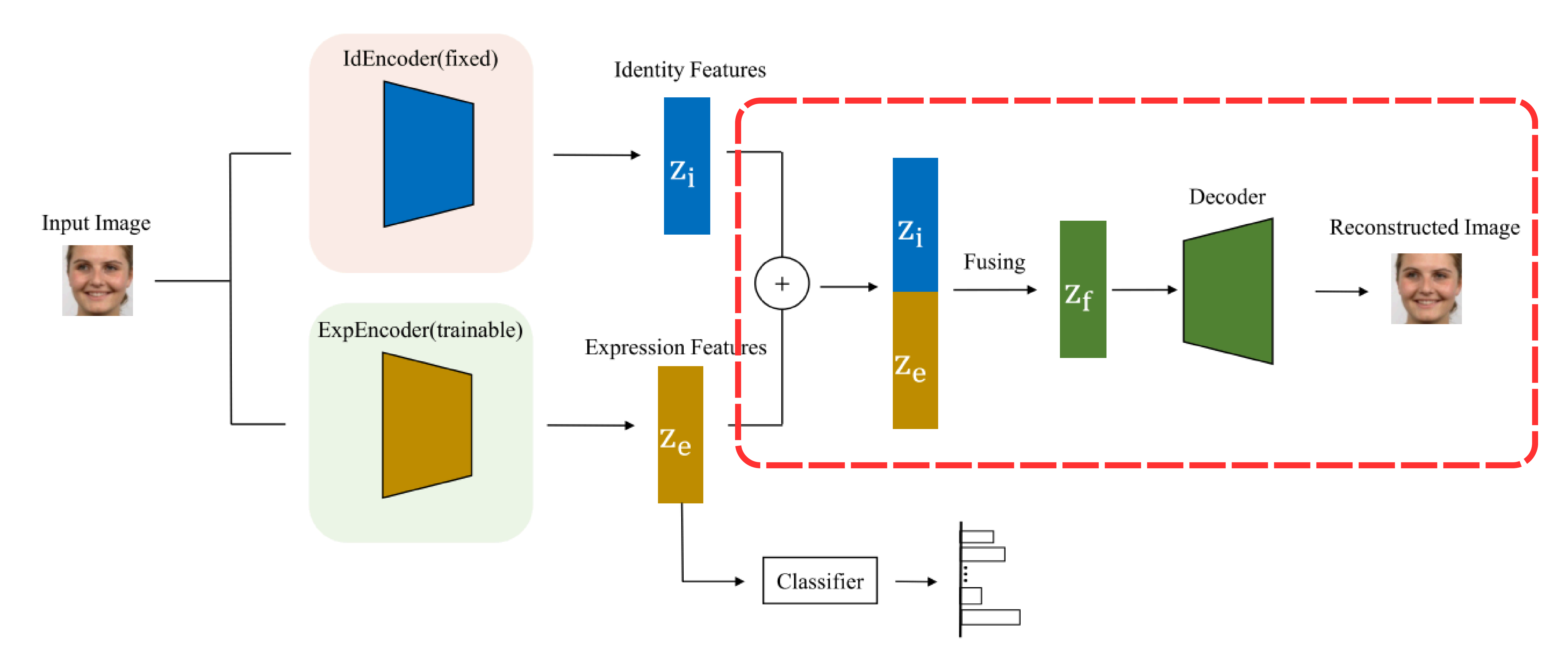
DisVAE is composed of two encoders and one decoder:



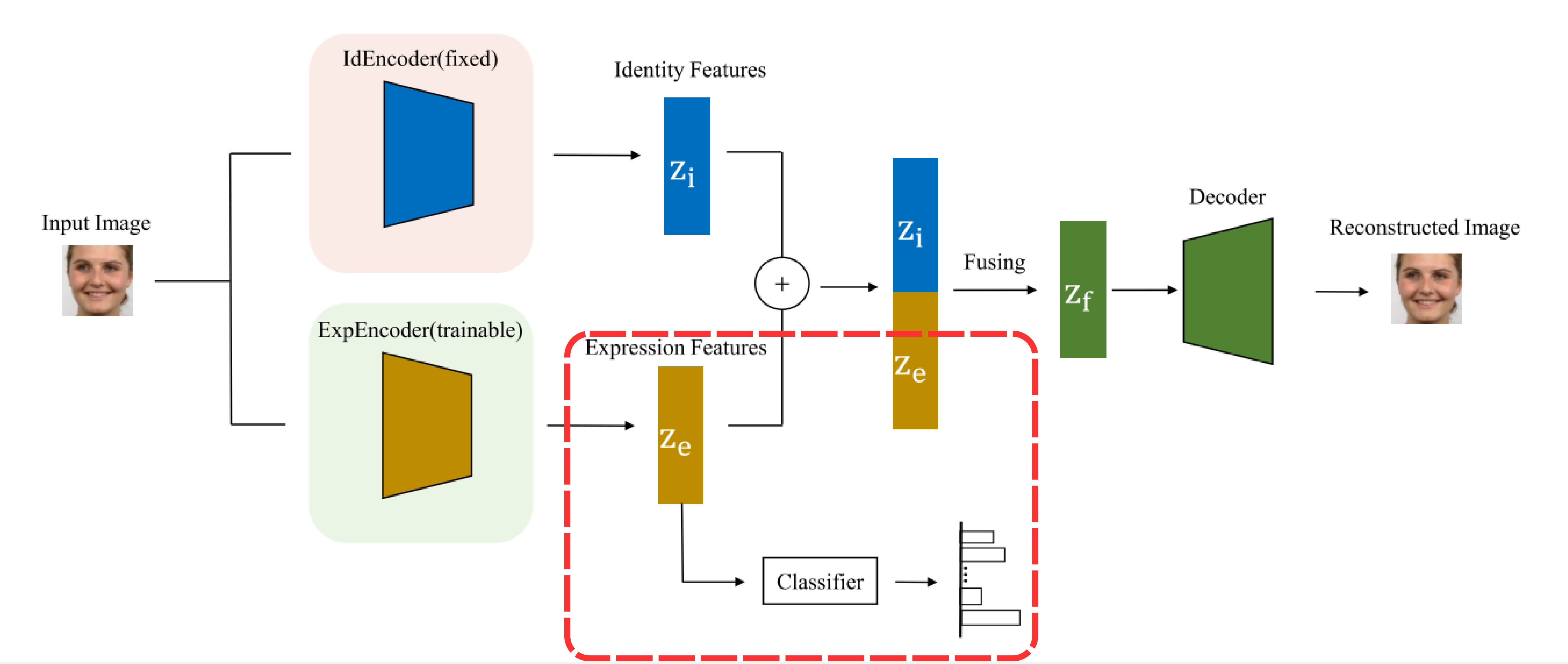
METHODOLOGY



METHODOLOGY



METHODOLOGY



METHODOLOGY

PRE-TRAIN

Identity Disentanglement Stage:

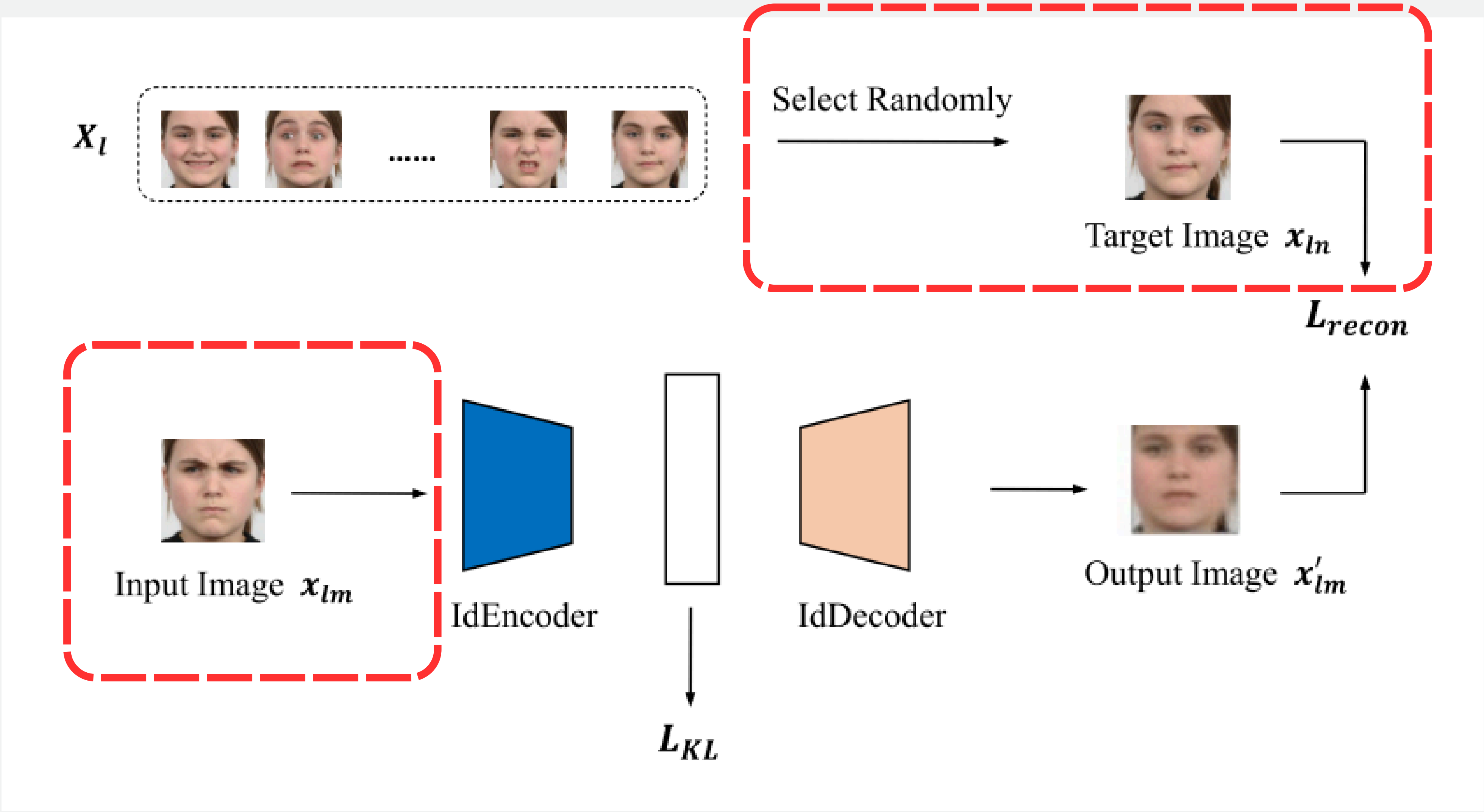
- IdEncoder is pre-trained to extract identity features.

TRAIN

Expression Disentanglement Stage:

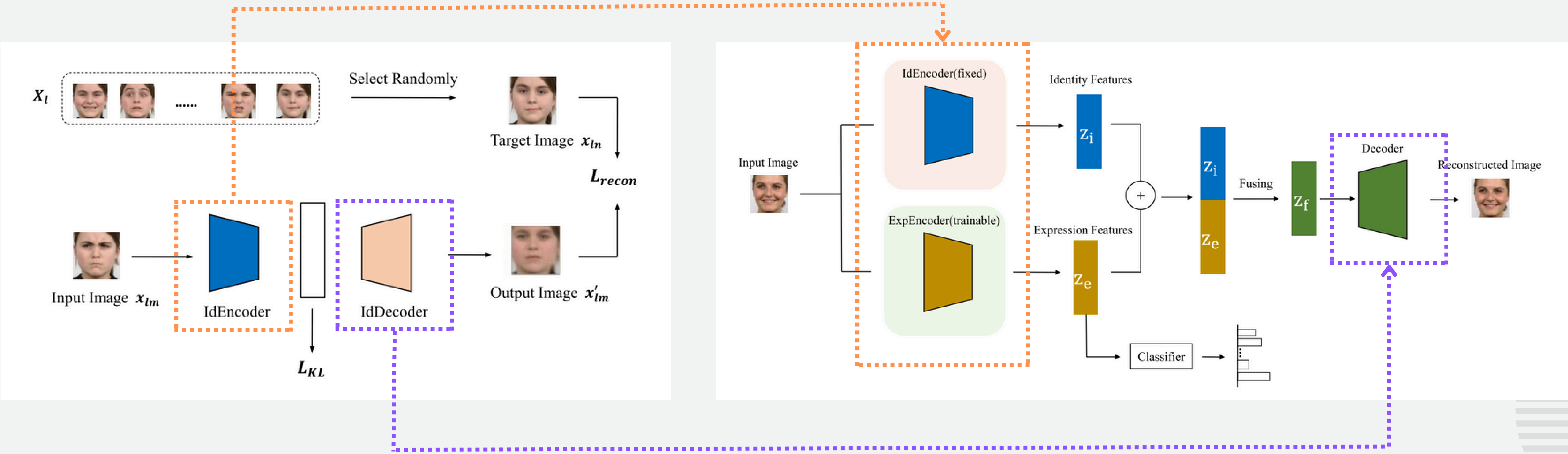
- DisVAE is trained to extract identity-unrelated expression features.

PRE-TRAIN


X_l
......
Select Randomly
Target Image x_{ln}
L_{recon}
Input Image x_{lm}
IdEncoder
IdDecoder
Output Image x'_{lm}
L_{KL}

TRAIN

Weight initialization:



TRAIN

DisVAE is trained in a multi-task learning fashion to extract identity-unrelated expression features:

- Pre-trained IdVAE is used to initialize weights;
- IdEncoder is fixed;
- A expression classification task is used to enforce expression feature learning;
- Identity and expression features are recoupled in order to reconstruct the input image.

DATASETS

The experiments make use of three "open" face expression datasets:

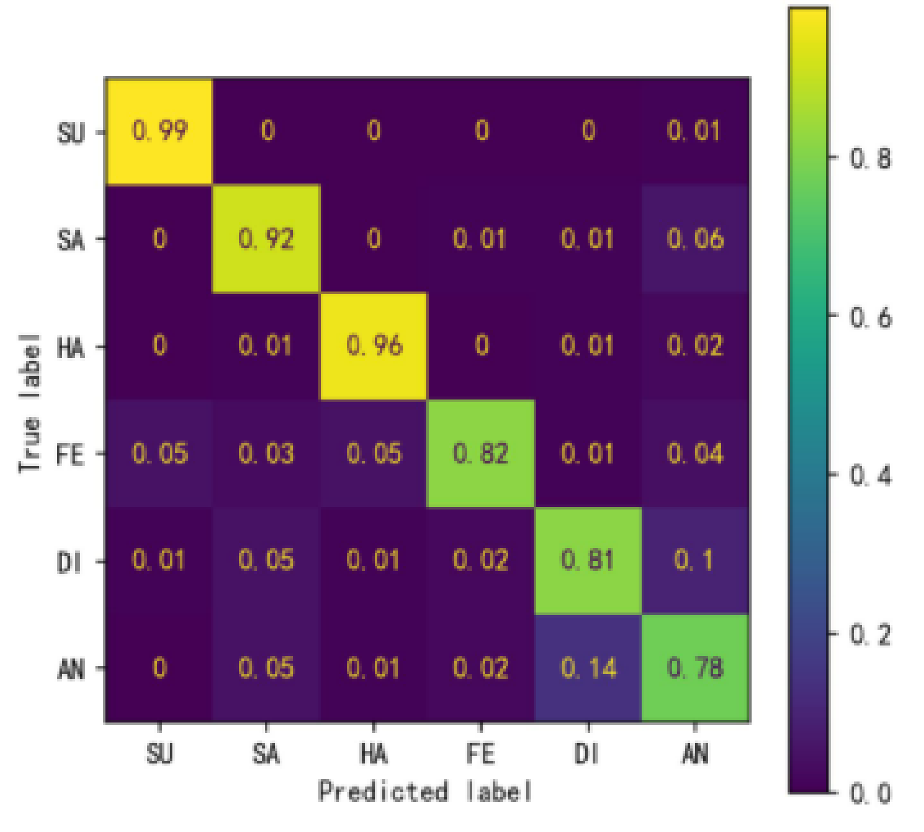
- CK+;
- Oulu-Casia;
- RaFD.

RESULTS

Expression recognition:

ACCURACY ON CK+ AND OULU-CASIA

Models	Input	CK+	Oulu-CASIA
LBP-TOP [27]	Sequence	88.99	68.13
STM-Explet [14]	Sequence	94.19	74.59
DTAGN [8]	Sequence	97.25	81.46
LOMo [19]	Sequence	95.10	82.10
FN2EN [4]	Static	96.80	87.71
PPDN [29]	Static	97.30	84.59
DeRL [23]	Static	97.30	**88.00**
ADFL [2]	Static	98.17	87.90
CNN baseline	Static	84.38	77.78
Our DisVAE	Static	**98.37**	87.90

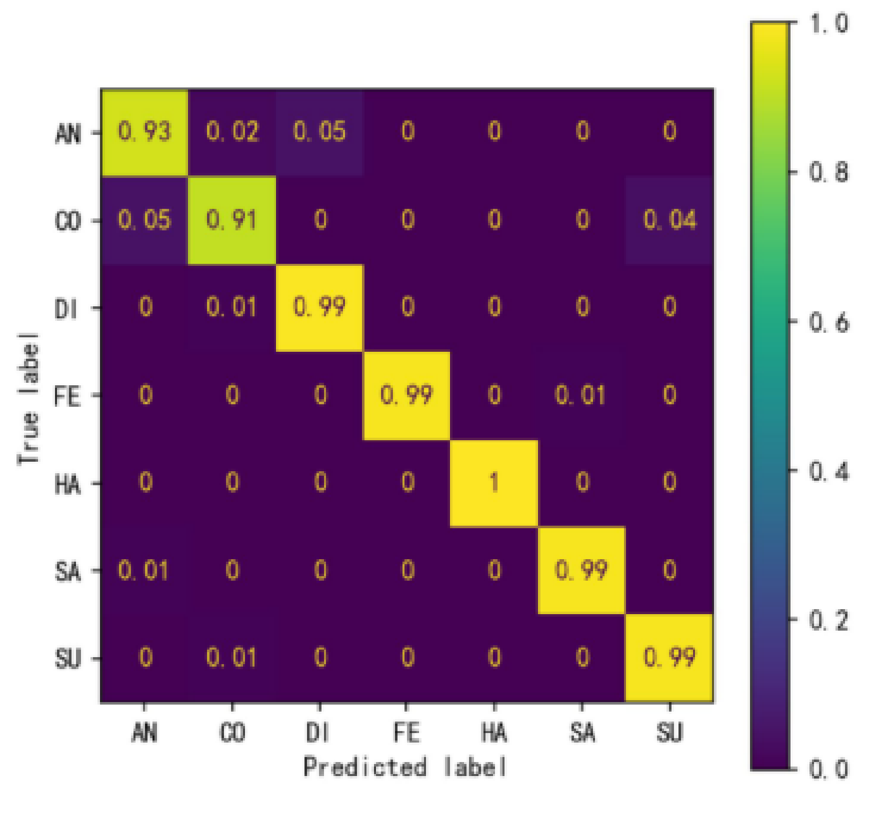


RESULTS

Expression recognition:

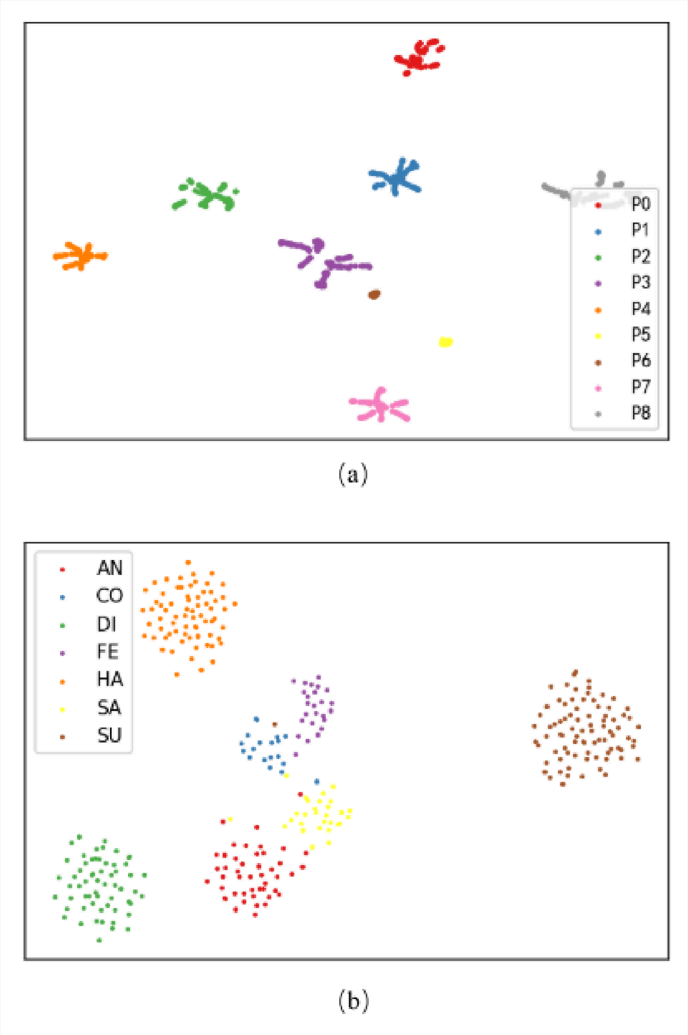
ACCURACY ON RAFD

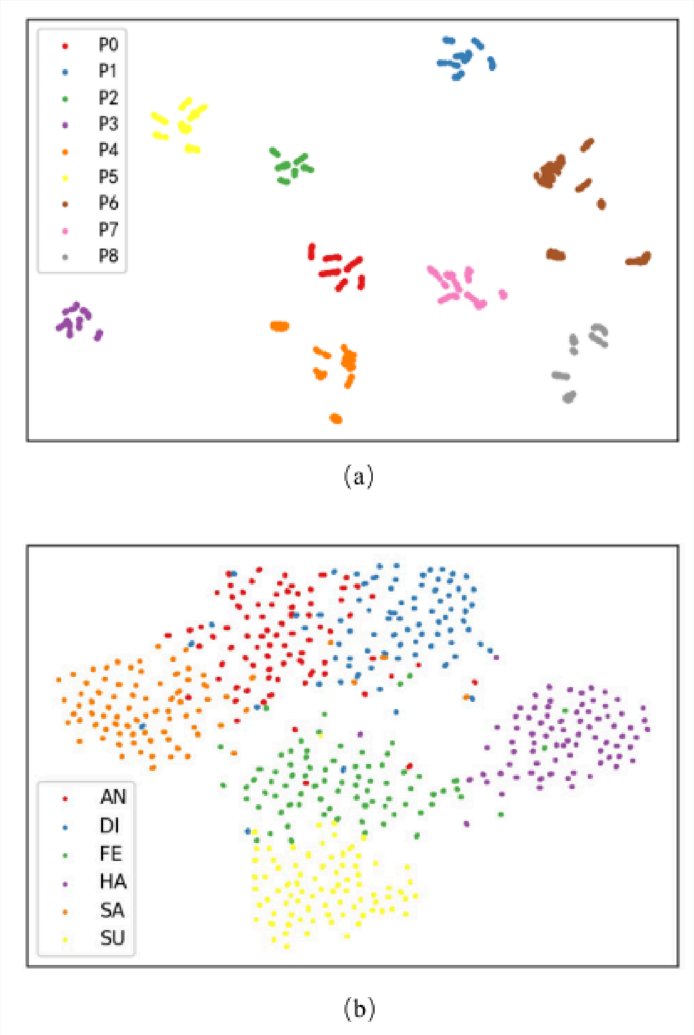
Models	Input	Accuracy
SURF [18]	Static	90.64
VisAtt [16]	Static	93.10
SVM [12]	Static	94.51
ANN-Gabor [6]	Static	99.15
TDGAN [22]	Static	99.32
CNN baseline	Static	94.16
Our DisVAE	Static	**99.78**



RESULTS

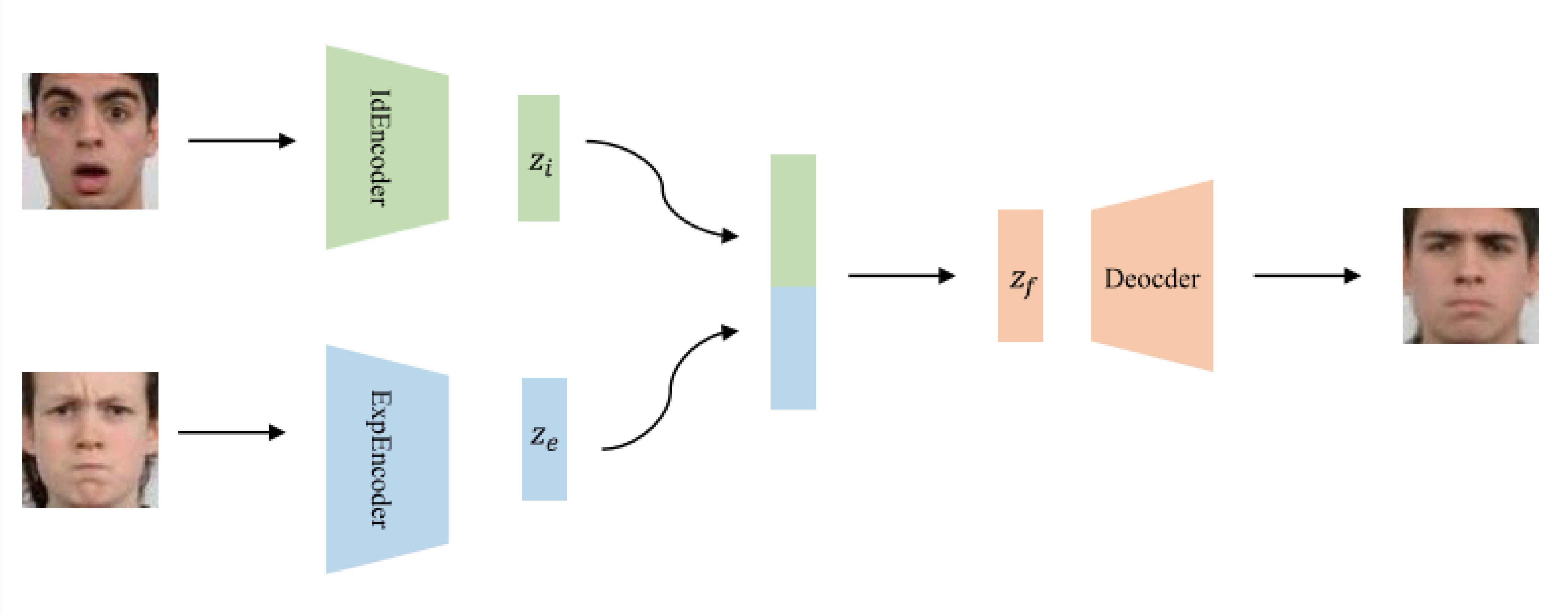
Learned features:



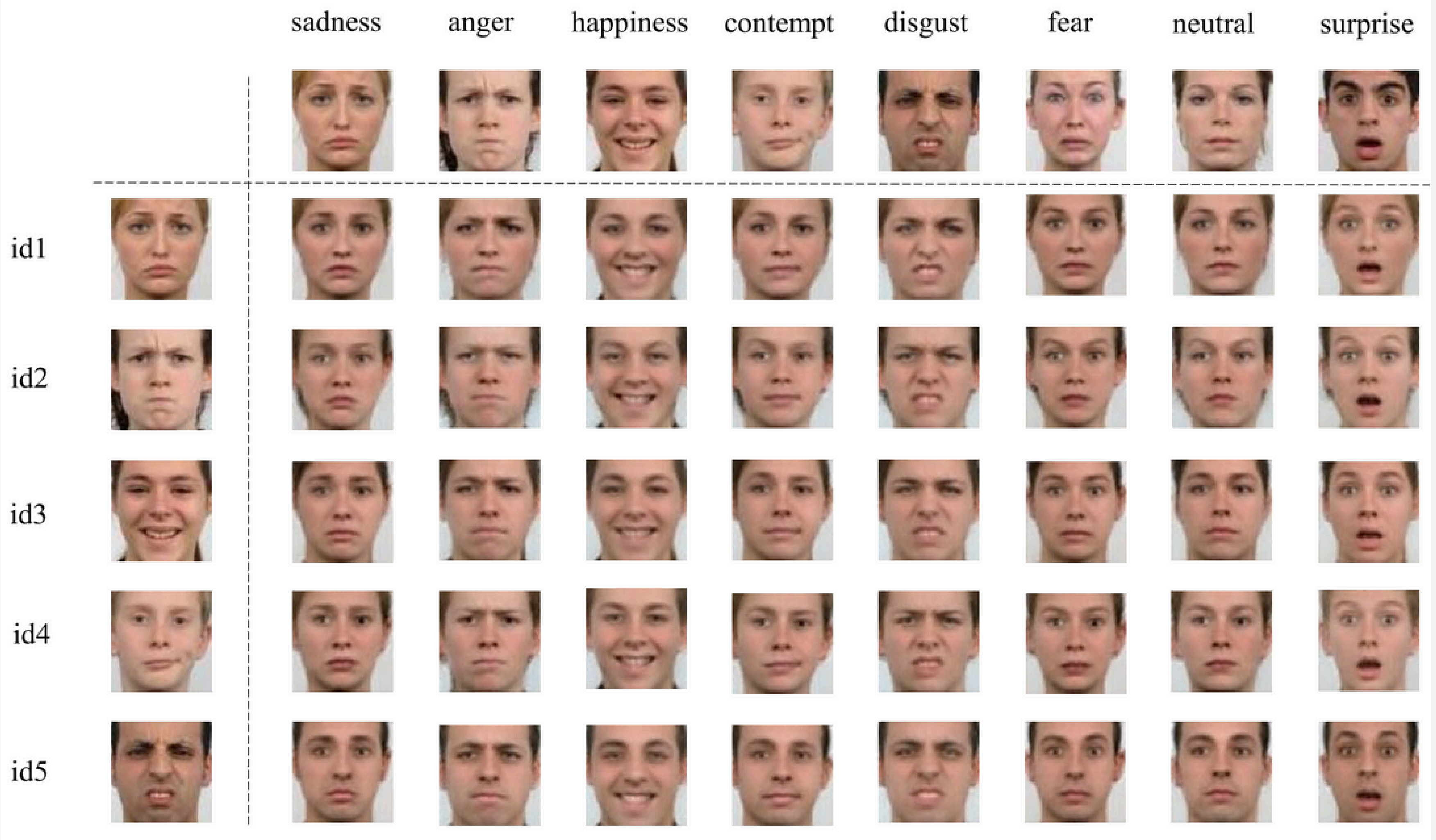


RESULTS

Facial expression editing:



RESULTS



sadness
anger
happiness
contempt
disgust
fear
neutral
surprise
id1
id2
id3
id4
id5

PROBLEMS

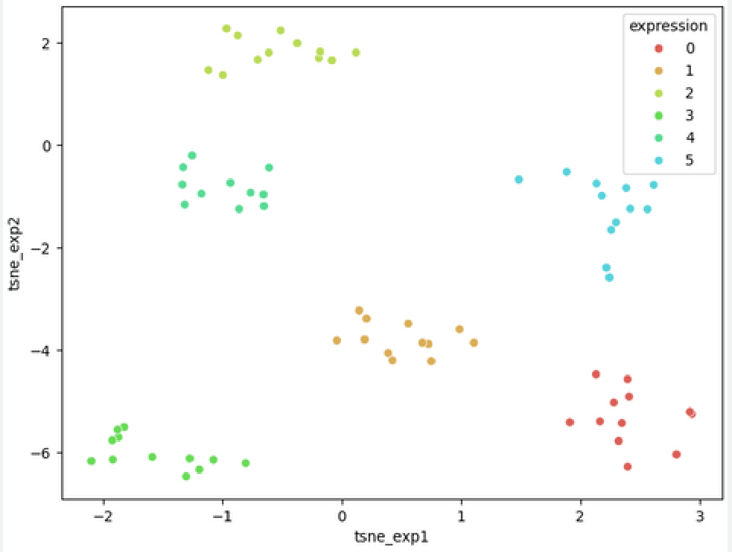
DETAILS MISSING	There is no GitHub link for code inspection. Some architecture details and hyperparameter values are missing.
DATASET ACCESS	Despite being listed as public, all the datasets need to be requested to the owners.

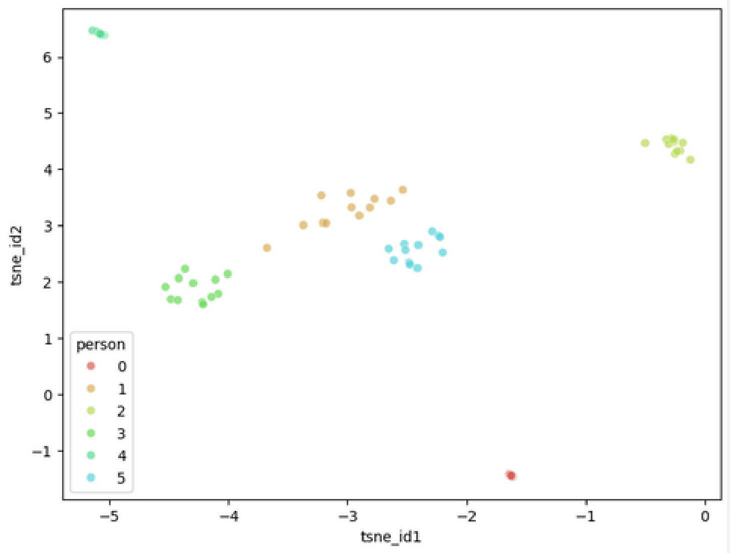
IMPLEMENTATION

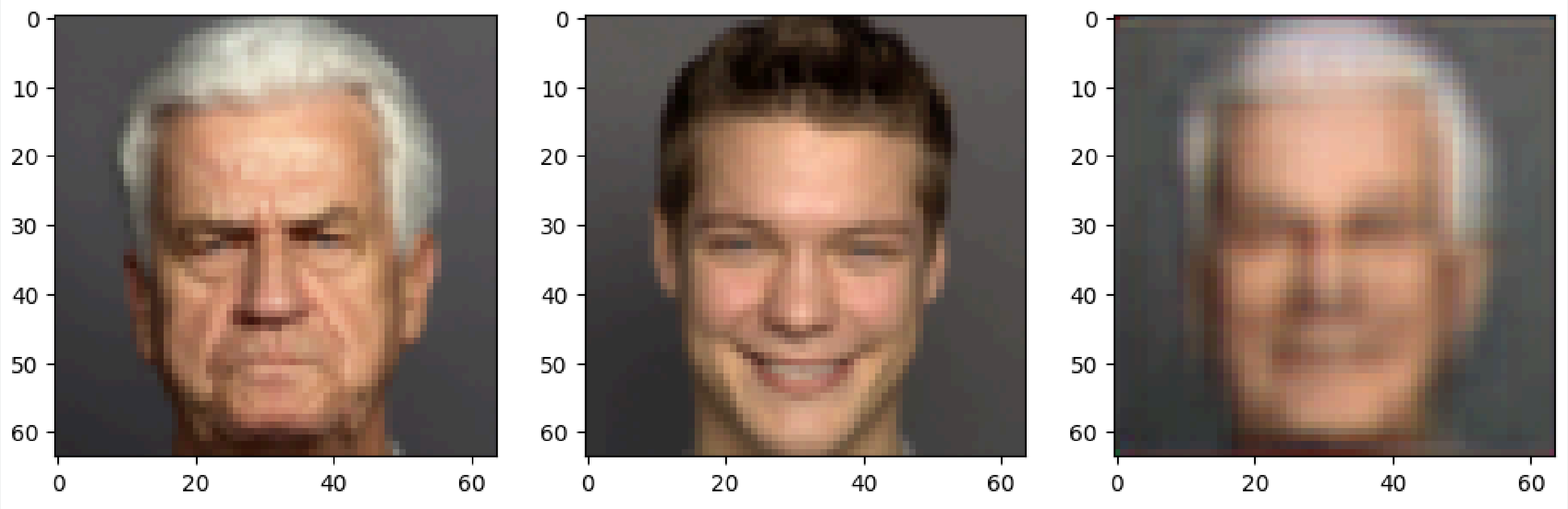
The implementation used for reproduction considered the following aspects:

- Model layer configurations listed in referenced paper;
- Latent feature fusion done by a Fully Connected Layer followed by ReLU activation;
- Optimizer hyperparameters given by Wang *et. al* with a lower learning rate;
- Public datasets: FACES and CK+48.

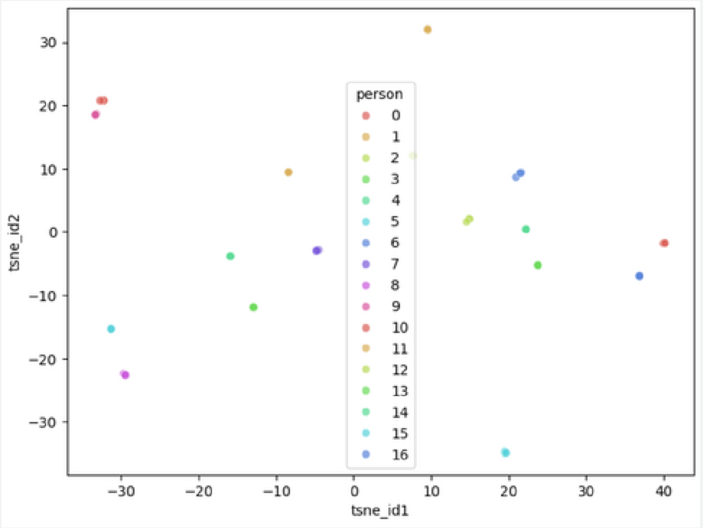
REPRODUCTION: FACES

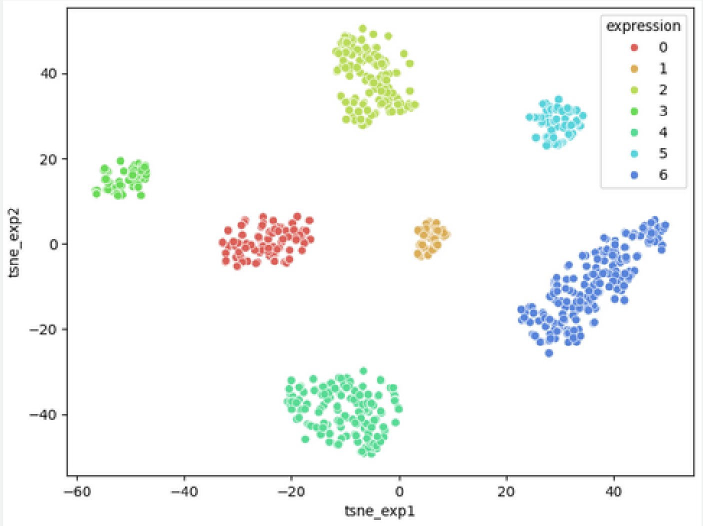


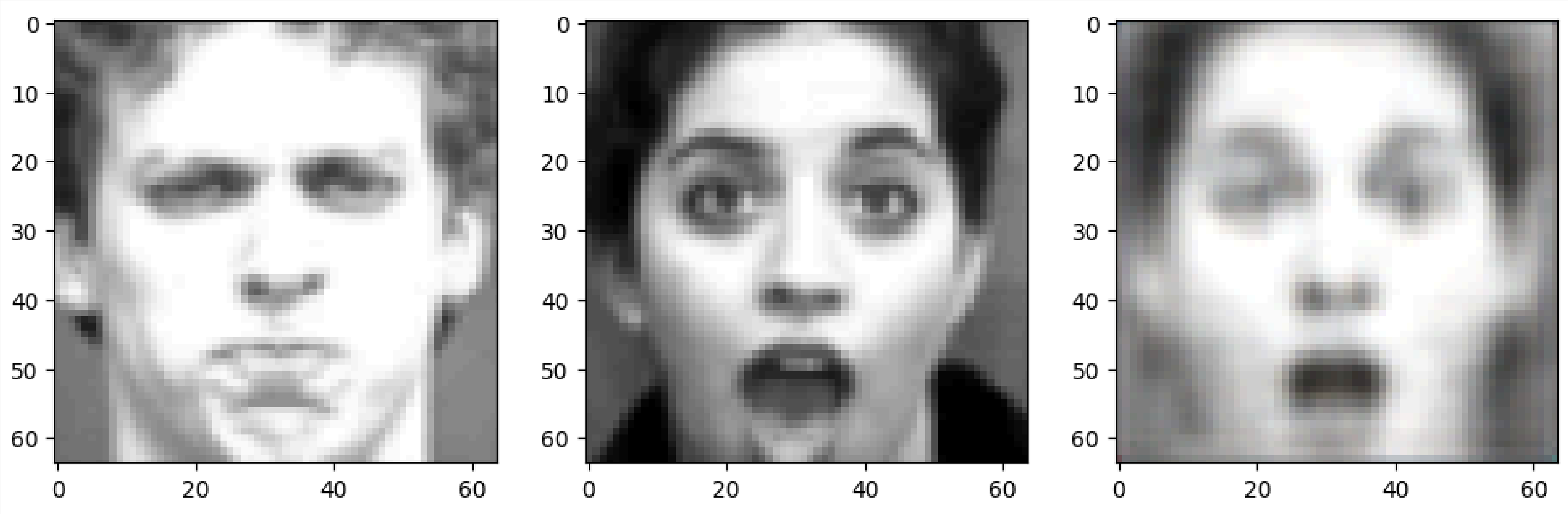




REPRODUCTION: CK+48







Presentation by Lucas Massa | PPGI | 2024 | UFAL

IMPROVEMENTS

END-TO-END

Reproduce results without pre-train.

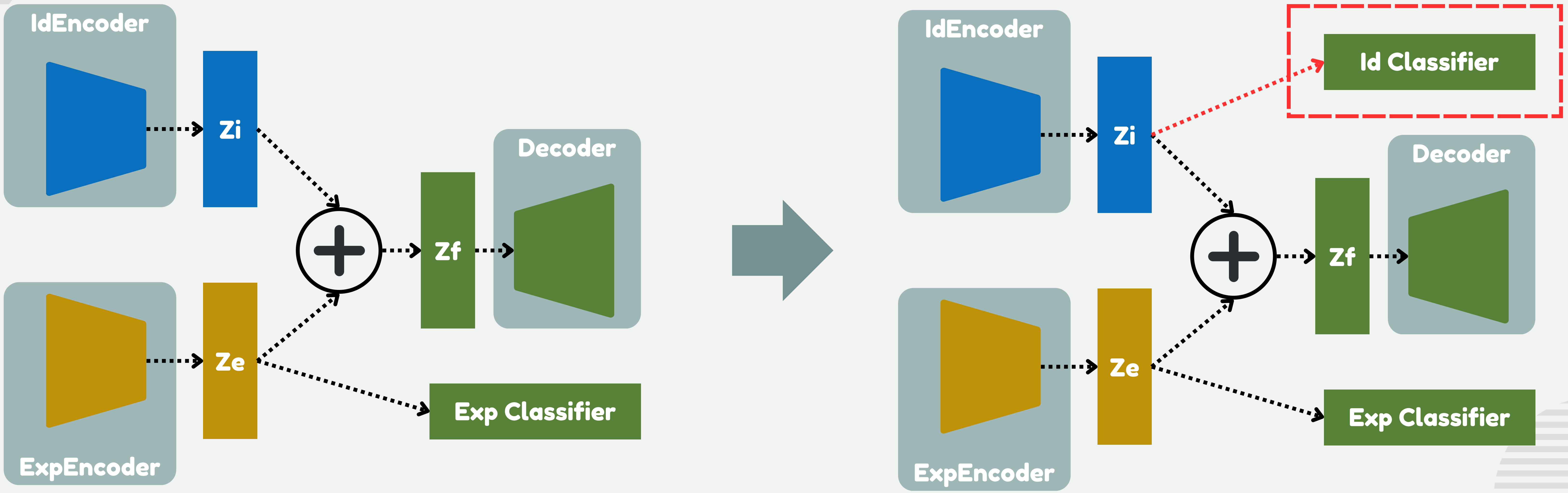
LOSS

Analyze the impact of current and new loss function terms.

TOPOLOGY

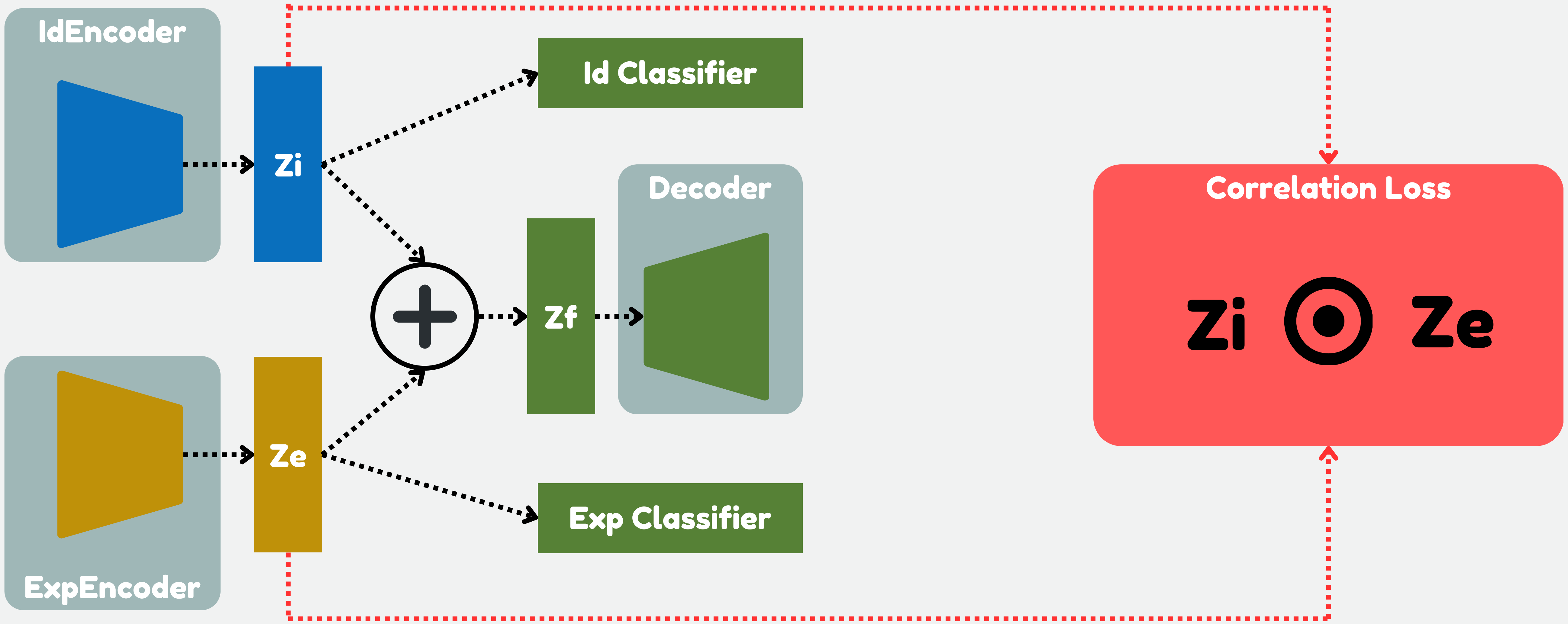
Analyze the topology learned by the latent space.

END-TO-END

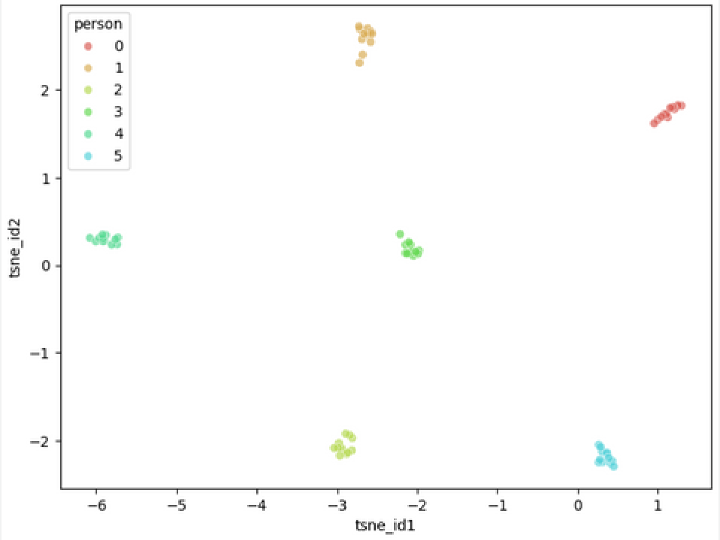


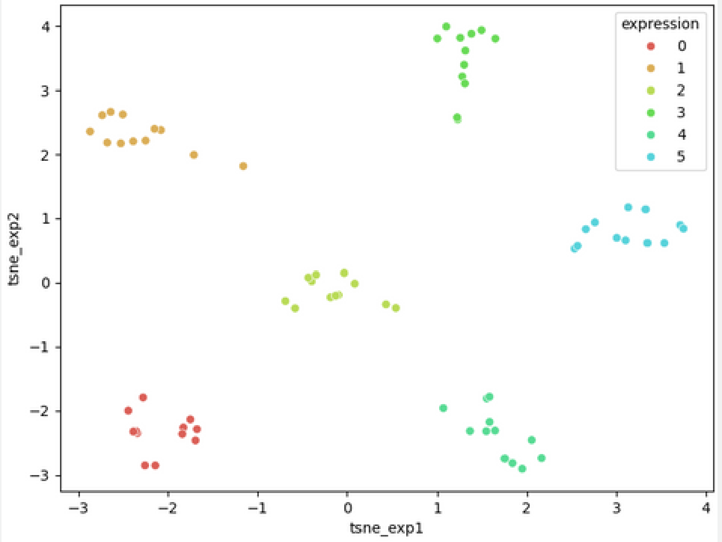
IdEncoder
Zi
ExpEncoder
Ze
Zf
Decoder
Exp Classifier
IdEncoder
Zi
Id Classifier
Decoder
Zf
ExpEncoder
Ze
Exp Classifier

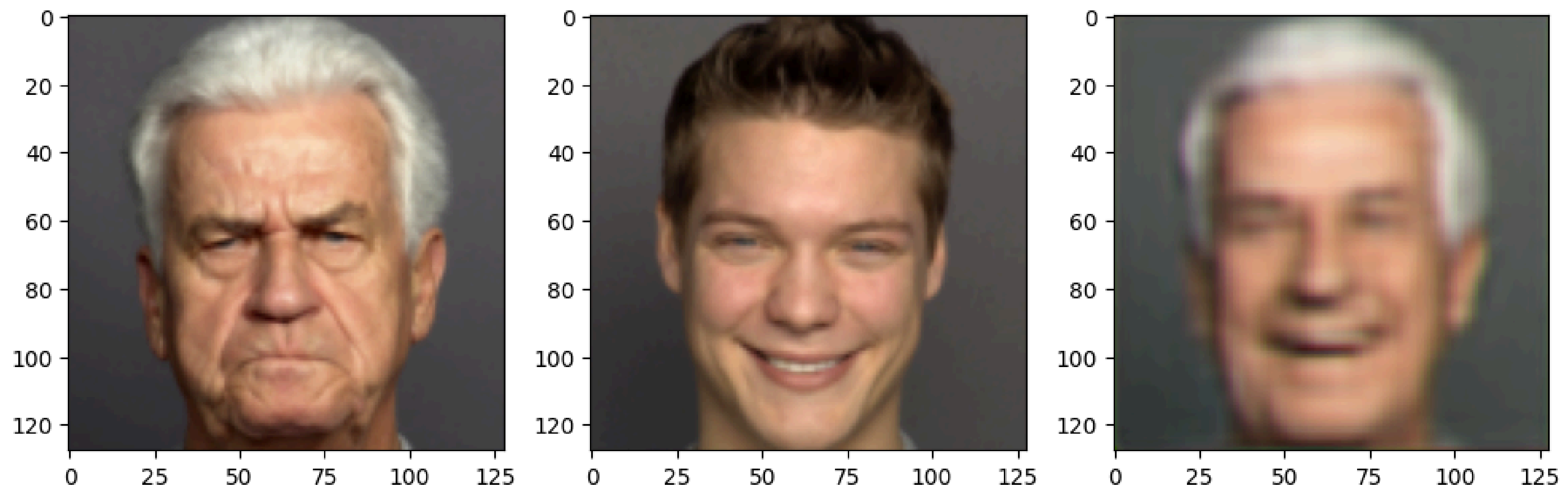
END-TO-END


IdEncoder
Zi
Id Classifier
Decoder
Zf
Correlation Loss
Zi ⊙ Ze
Ze
Exp Classifier
ExpEncoder

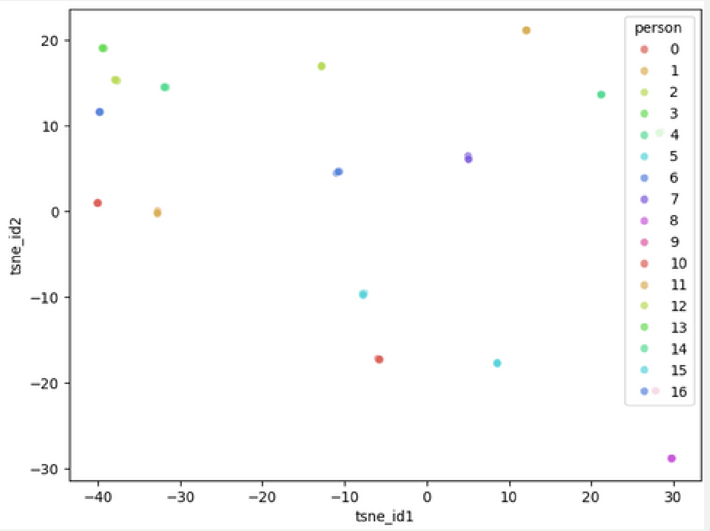
END-TO-END: FACES

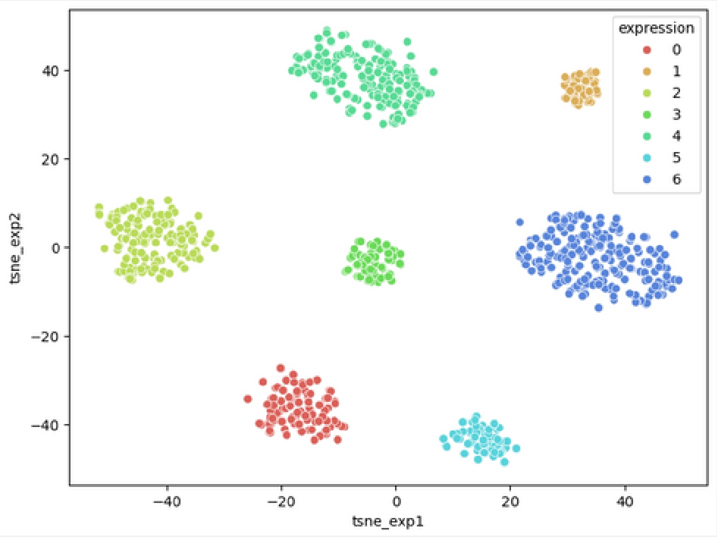


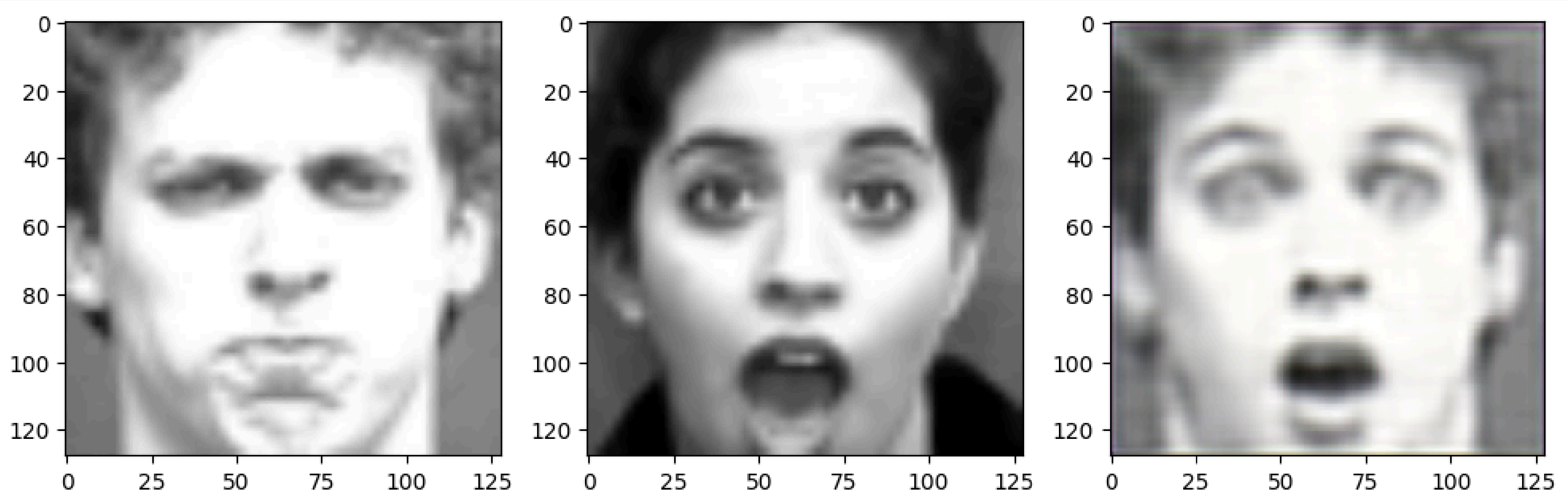




END-TO-END: FACES







REFERENCES

WANG, Tianhao; ZHANG, Mingyue; SHANG, Lin. DisVAE: Disentangled Variational Autoencoder for High-Quality Facial Expression Features. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, 2023. p. 1-8.

THANK YOU

Presentation by Lucas Massa