Comparing Facial Expressions for Face Swapping Evaluation with Supervised Contrastive Representation Learning

Felix Rosberg[1,3] and Cristofer Englund[2,3]
[1] Berge Consulting, Lindholmspiren 3A SE-417 56 Gothenburg, Sweden
[2] RISE - Research Institutes of Sweden, Lindholmspiren 3A SE-417 56 Gothenburg, Sweden
[3] Halmstad University, Center for Applied Intelligent Systems Research (CAISR), SE-301 18 Halmstad, Sweden

Abstract— **Measuring and comparing facial expression have several practical applications. One such application is to measure the facial expression embedding, and to compare distances between those expressions embeddings in order to determine the identity- and face swapping algorithms' capabilities in preserving the facial expression information. One useful aspect is to present how well the expressions are preserved while anonymizing facial data during privacy aware data collection. We show that a weighted supervised contrastive learning is a strong approach for learning facial expression representation embeddings and dealing with the class imbalance bias. By feeding a classifier-head with the learned embeddings we reach competitive state-of-the-art results. Furthermore, we demonstrate the use case of measuring the distance between the expressions of a target face, a source face and the anonymized target face in the facial anonymization context.**

I. INTRODUCTION

Privacy-aware data collection is an emerging research area within traffic safety. In this paper it concerns the task of anonymizing image data without destroying information. In the context of traffic video data collection where there are plenty of people moving around in a frame one have to, for instance, hide the identity to maintain high sequrity of the data. One approach to achieve this is to use face swapping algorithms like Face Swapping GAN (FSGAN) [15] or FaceShifter [12] to hide the identity of the people in a frame while at the same time trying to maintain facial expression and eye gaze. Keeping those attributes makes it possible to maintain a realistic behaviour of the road users, also after anonymization. To either deploy these methods, improve them or introduce new methods we need a way to evaluate the anonymizaton process. This includes for example how well the identity is obscured, how well the eye gaze is preserved and finally how the facial expression is preserved. In this work we will focus on the task of (a) representing facial expressions and (b) measuring the facial expression preservation after performing the anonymizaton process. To achieve this we want to focus on providing strong embeddings that then can be used to calculate distances between expression embeddings.

Contrastive loss has shown great promise in extracting information rich embeddings [10]. Supervised contrastive loss is utilized as it allow us to deal with class imbalance in facial expression recognition. Furthermore, it showed that it can deal with classes that have arguably sparse differences, in this case all images being facial images with small changes in the main facial area.

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

II. DATA

For all training and evaluation we used the AffectNet data set [14]. The data set is extensive and contains 287,651 facial images. The images are already cropped and aligned, along with labels for 8 different emotions in addition to valence and arousal values. Valence and arousal is a way of describing emotion in a continuous manner with two values. Valence describes the comfort of the emotion where a low value describe emotions as anger and disgust, and a high value describe emotions as happy. Arousal describes the excitation of the emotion where a low value describe emotions as sad and calm, and a high value describe emotions as anger and happy. The different discrete emotions includes neutral, happy, sad, surprise, fear, disgust, anger and contempt. The whole data set extends to 440,000 facial images. For training identity swaps, we used the FFHQ data set [9] and the method of choice was the FaceShifter.

III. RELATED WORK AND CONTRIBUTION

A. Related Work

There are four main ways of approaching the expression recognition task. There is (1) end-to-end expression classification [17], [16]. Then there is (2) representation learning [17]. Finally there is the approach of (3) regressing 3D morphable models (3DMM) [1], [2] of the face and (4) 2D landmarks [15].

Previous work [17], [20], [23], [16], [21], [18], [22], [19], [5] mostly focus on end-to-end classification and uses arguably complex neural network schemes. Many of the aforementioned approaches utilizes several datasets. Most of which achieve impressive state-of-the-art results on AffectNet [14]. We want to focus on addressing the complex training schemes and provide information rich feature vectors that can be used to make comparisons between faces that take the continuous nature of facial expressions into account, something which class labels cannot really address.

Some recent work for the 3DMM approach consist of ExpNet [1] and [2]. The 3DMM approaches aim to represent the face (along with expression and emotion) with a 3D representation. Recent work to our knowledge, and on the public leader boards, do not show any results for the AffectNet data set. However for the context of face swapping and anonymization, [2] has been used in FaceShifter [12] for measuring the expression preservation and [4] was used for the MegaFS [24] face swapping model. FSGAN [15], which uses facial reenactment and blending techniques for face swapping, describes using the euclidean distance of 2D landmarks generated from dlib [11] directly between target face and the changed target face. We argue that 2D and 3D representation is not the best method for this metric as the shape of the face is expected to change when swapping identity.

Khosla et al. [10] introduced the supervised contrastive loss. They point out that modern batch-based contrastive loss functions outperforms traditional contrastive losses such as triplet loss and n-pair loss. They extend the self-supervised version to a supervised setting to exploit label information. Supervised contrastive loss showed effectiveness in pulling clusters of the same class together and pushing clusters of different classes apart.

B. Contribution

We propose a simple pre-training scheme that both learns to represent the facial expressions well and take class imbalance into account. Our work utilizes a weighted supervised contrastive representation learning [10] to train an EfficientNetB0 network [19]. Furthermore, we strengthen the performance with a multi-task approach by forcing the network to predict valence and arousal from the extracted features during pre-training. This pre-trained network can then be used to perform facial expression classification by quickly training a classification head or to be used as a facial expression embedder. We demonstrate how measuring facial expression can be used both as a metric for determining how well face swapping algorithms preserve facial expression.

IV. Proposed Method

A. Weighted Supervised Contrastive Representation Learning

We adapt the methodology from [10] with a weighting modification. We use a weighted version of the contrastive loss due to the data set being imbalanced. The non-weighted loss function $\mathcal{L}_c$ is described below in part:

$$\mathcal{L}_c = CE(softmax(y_l), softmax(z_l)), \tag{1}$$

where the cross entropy CE is calculated between the logits z_l and truth logits y_l both passed through a softmax function ($softmax$). The logits z_l is calculated with

$$z_l = \frac{(\frac{z}{\sum z^2}) \times (\frac{z}{\sum z^2}^{\top})}{\tau}, \tag{2}$$

where z is the embedding vector produced from the neural network and τ is a temperature parameter in (2). τ is set to 0.05 and is chosen with intuition regards to a lower τ punishes hard negatives further. The truth logits y_l is calculated with

$$y_l = \frac{A}{\sum A_j}, \tag{3}$$

where A in (3) is calculated with an equality tensor operation that compares two tensors element-wise and returns a tensor of zeros and ones. The resulting tensor has ones where the equality is satisfied. In our case we obtain A by calculating equality between y and transposed y ($y^{\top}$), where y is the labels.
For the weighted loss function, we calculate the weight for each class $\mathcal{W}$ before training:

$$\mathcal{W} = \frac{n_s}{n_c \cdot v_o}, \tag{4}$$

where n_s is the number of samples, n_c is the number of classes and v_o is an occurrence vector of the labels in (4). Equation (4) assigns higher weights to labels of low frequency like contempt and disgust, and small weights to common labels like neutral and happy. $\mathcal{W}$ is then used as a lookup table for the labels to assign the correct weight to each sample in the batch. The weighted contrastive loss function is then finally defined as:

$$\mathcal{L}_{wc} = CE(softmax(y_l), softmax(z_l)) \cdot \mathcal{W}(y), \tag{5}$$

where y denotes the labels, and is used to retrieve a class weight vector from the class weight look up table $\mathcal{W}$. This loss function severely punishes the neural network for classifying the uncommon labels wrong.

Using the above weighted loss function, we train the EfficientNetB0 network from scratch, adding a projection head of dimension size 128 and two linear layers for continuously predicting the arousal and valence values. Arousal and valence loss is calculated with mean squared error. In Fig. 1 the entire training scheme is illustrated for the weighted contrastive representation learning process. The total loss used is:

$$\mathcal{L}_{wtot} = \mathcal{L}_{wc} + MSE_{aro} + MSE_{val}, \tag{6}$$

where MSE_{aro} denotes the mean squared error for the arousal value and MSE_{val} denotes the mean squared error for the valence value. We compare the weighted loss function with the non-weighted version:

$$\mathcal{L}_{tot} = \mathcal{L}_c + MSE_{aro} + MSE_{val}, \tag{7}$$

The entire pretraining process for the weighted contrastive representation learning is shown in Fig. 1. For data augmentation horizontal flip and random crop was used. The random crop augmentation crops a 200x200 image from the 224x224 input data and then resizes it back to 224x224. For training parameters a learning rate of 0.001 was used for the representation learning. Exponential decay of the learning rate was used, decaying the learning rate with a factor of 0.9 every 15000 steps. Batch size of 128. Global average pooling is used in the end of the EfficientNetB0 encoder, before the projection layer. Finally we used an early stopping callback

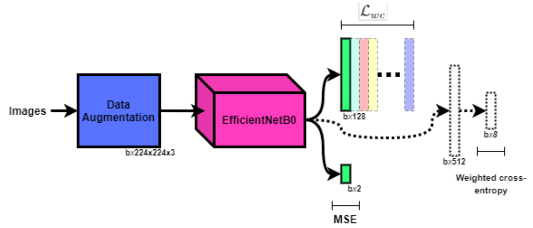

Fig. 1. The training scheme for the pretraining representation learning step. b represents the batch size. Multi-colored boxes illustrates embeddings of different classes in a batch. Dotted branch illustrates the classification training step.

TABLE I

ACCURACY ON AFFECTNET VALIDATION DATA SET FOR 8 EMOTIONS.

Methods	Accuracy	Extra data
Schoneveld et. al (Multimodal) [17]	**61.60%**	*yes*
Savchenko et. al (Multi-task) [16]	61.32%	*yes*
Vo et. al (PSR) [21]	60.68%	*yes*
Shi et. al (ARM) [18]	59.75%	*yes*
Wang et. al (RAN) [22]	59.50%	*yes*
Ours (EfficientNetB0) + $\mathcal{L}_{wtot}$	**59.58%**	*no*
Ours (ResNet50V2) + $\mathcal{L}_{wtot}$	58.51%	*no*
Ours (ResNet50V2) + $\mathcal{L}_{tot}$	57.76%	*no*
Ours (ResNet50V2) + $\mathcal{L}_{wc}$	48.96%	*no*
Siqueira et. al [19]	59.30%	*no*
Mollahosseini et. al [14]	58.00%	*no*
End-to-end classification (EfficientNetB0)*	15.15%	*no*
SimSiam (EfficientNetB0)*	12.50%	*no*
Autoencoder (EfficientNetB0)*	12.50%	*no*

* Trained with the same configuration and hyper-parameters as our best method.

to retrace the weights with lowest loss. The representation network trained for 50 epochs and the classification for 25 epochs.

B. Expression Classification

When the network is pretrained with the total loss function in (6) we use it for measuring the accuracy and representation capabilities. For this task we remove the projection head and add one layer with 512 units along with a linear classification layer. Dropout rate of 0.5 was used before and after the hidden layer in aforementioned classification head. While keeping the entire representation network frozen, we train the network to classify expressions. We deploy the cross entropy loss together with the same weight scaling as in (5) for this task. The performance is compared to other methods along with a ResNet50V2 baseline and a EfficientNetB0 baseline. The same augmentation process as the pretraining step is used. The accuracy of the classification also represents the performance of the representation learning performed in the pretraining.

C. Measuring Expression Preservation

For measuring the distance in representation space we look into the euclidean distance as this was used for the 2D landmark approach of measuring expression preservation [15]. The idea is to compare the distances between a target face, a source face and the changed target face for demonstrating the use as a metric when face swapping. By target face we mean the face of which we want to change identity, by source face we mean the face whose identity we want to use to impose on the target face and by changed target face we mean the manipulated face of the target face. The identity swapping used for demonstration is a FaceShifter implemented to generate 128x128 face swaps [12]. To assure fair comparisons we qualitative judge the identity swaps whether FaceShifter is able to maintain the facial expression so one can compare expression distance between target face, source face and the changed target face.

V. RESULTS

A. Representation Learning

1) Quantitative Results: To investigate the representational power of using supervised contrastive representation learning as pretraining, we report top achieved accuracy when training with weighted cross entropy, with a classification head instead of a projection head. As seen in Table I we achieve a top 59.58% accuracy with a completely frozen EfficientNetB0 encoder pretrained with the weighted contrastive loss. EfficientNetB0 is shown to be a lot more effective for this task compared to a ResNet50V2 baseline, boosting performance by 1.07% percent units compared to the baseline. As shown in Table I, when comparing to recent work that do not uses extra training data beyond AffectNet, our approach boost the performance slightly compared to the previous best method. We also tried training the classification network end-to-end with the same configuration, only adjusting the learning rate, and was not able to learn anything i.e. it only managed to reach 15.15% accuracy. Furthermore, in the *same* setting and configuration as the EfficientNetB0 + $\mathcal{L}_{wtot}$ (Table I) tried semi-supervised contrastive learning using siamese representation learning (SimSiam) [3] and a simple autoencoder [8] built on top of the original encoder. Both these methods yielded poor performance, not capable of learning anything. We suspect that the siamese representation learning fails due to that it relies on heavy augmentation threshold for augmentations like random crop, zoom and translation. It is often preferred for both facial recognition and facial expression recognition to utilize aligned images. The heavy augmentation would break the alignment preprocessing completely.

2) Ablation Study: Ablation is done in a few step. First we *'downgrade'* the backbone from EfficientNetB0 to a ResNetV2. Then removal of the weighting of the contrastive loss and finally removal of the multi-task prediction head for arousal and valence. The accuracy is displayed in Table I. The added components in total boost performance from 48.96% to 59.58%, with the multi-task component yielding the biggest boost. In Fig. 2 we display a confusion matrix between classes. Shi et. al reports a confusion matrix as well [18]. Even if they used extra training data from the RAF-DB dataset [13] and obtain an overall better accuracy, the per class accuracy is more uneven. For instance ours maintain a 57.11% accuracy for the contempt expression while Shi et.

TABLE II

Comparison between our approach and 2D landmark approach for comparing expression for face swaps. Left to right column: Method, euclidean distance error between target face to source face and target face to changed face, mean euclidean distance for target face to changed face and ratio of $t2c < s2c$.

Method	$L2$ error ↓	Mean $t2c$ $L2$ ↓	Ratio ↑
Ours*	**0.07**	**0.17**	**0.73**
68 2D landmarks*	0.29	0.32	0.39
51 2D landmarks*	0.28	0.31	0.49
Ours$^+$	**0.07**	**0.17**	**0.70**
68 2D landmarks$^+$	0.31	0.38	0.39
51 2D landmarks$^+$	0.35	0.35	0.47

* Different class comparison. $^+$ Same class comparison.

al achieved a 39.00% accuracy for contempt. Contempt is usually omitted and deemed a hard expression to examine.

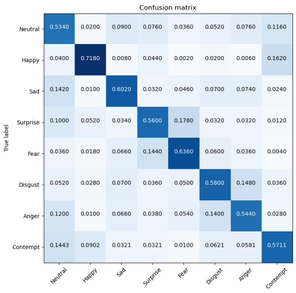

Fig. 2. Normalized confusion matrix.

3) Qualitative Results: We investigate the representation learnt by plotting the AffectNet validation set using t-SNE [6]. In Fig. 3 one can see clear and even clusters for each class. Arguably the fact that neutral facial expression lies in between all other facial expression is a sign of good representation.

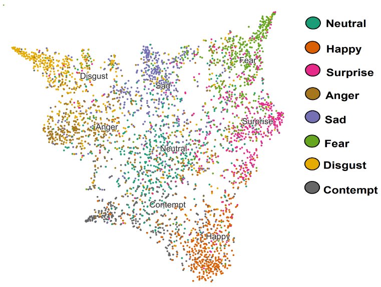

Fig. 3. T-SNE plot of embeddings (EfficientNetB0) of the AffectNet validation data set.

B. Identity Swap Metric

As suggested, the representation network can be used to determine how well face/identity swapping methods can maintain facial expressions by operating on the embedding vectors. We compare our approach with a 2D landmark baseline approach used to measuring expression preservation for FSGAN [15]. Comparison is done with all 68 landmarks from dlib [11] and 51 landmarks when omitting the points around the face. The comparison is made by looking at the normalized euclidean distance ($L2$) error between source face to target face ($s2t$) and source face to the changed face ($s2c$). We also report the mean distance from target face to the changed face ($t2c$) and the ratio of $t2c$ being less than the distance between the source face to the changed face ($s2c$). The values were generated by using a pretrained FaceShifter [12] to swap faces between different expression classes in the AffectNet validation set. 3000 samples were generated, out of which half has randomly different expression labels and half have the same. Results can be seen in Table II for the same expression class and different expression class comparison.

VI. Conclusions and Future Works

Using weighted supervised contrastive representation learning for pretraining and training embedding networks is a promising approach for facial expression recognition and possibly other similar tasks. Our method beats previous methods that do not use extra training data (*see* III-A) and reach competitive accuracy to state-of-the-art on AffectNet. The method is straight forward and easy to implement. Our approach also maintains an even class accuracy instead of completely favoring the common classes. Predicting expressions is arguably not a discrete classification task, as facial expressions are continuous in between each other and some overlap. For instance a surprised facial expression can be of happiness or fear. The t-SNE plot seems to suggest that our approach learns this overlap to an extent as surprise lies between happy and fear, and neutral being positioned in the middle (Fig. 3). We showed that for measuring facial expression preservation after face swaps, our approach is superior to a 2D landmark approach and is valid to have an idea of how well the expression is preserved. This is under assumption that the face swapping method (FaceShifter) preserves facial attributes well to make the comparison valid, which we qualitative deemed it was. Our conclusion is that our facial expression embeddings contains rich information as one is able to reach competitive state-of-the-art accuracy with simply the encoder feature maps, a linear layer and a classification layer.

For future work it would be interesting to utilize knowledge distillation by including more data points and distill the representation learning network [7], [23]. During experimentation we only had access to 287,651 of the 440,000 human labeled images in AffectNet. There is a great potential in our approach to reach even better accuracy if more data was accessed. For our proposed distance metric for face swapping algorithms using the embeddings, we are considering making more extensive experiments and comparisons to accurately determine what kind of failure was produced by the face swapping algorithm. One aspect that should be addressed is the need to compare our representation vectors to 3DMM expression coefficients.

References

[1] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. Expnet: Landmark-free, deep, 3d facial expressions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 122–129. IEEE, 2018.

[2] B. Chaudhuri, N. Vesdapunt, and B. Wang. Joint face detection and facial motion retargeting for multiple faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2019.

[3] X. Chen and K. He. Exploring simple siamese representation learning. In *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[5] A. H. Farzaneh and X. Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2402–2411, January 2021.

[6] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, page 857–864, Cambridge, MA, USA, 2002. MIT Press.

[7] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[8] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[9] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[11] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(60):1755–1758, 2009.

[12] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

[13] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017.

[14] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, Jan 2019.

[15] Y. Nirkin, Y. Keller, and T. Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019.

[16] A. V. Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks, 2021.

[17] L. Schoneveld, A. Othmani, and H. Abdelkawy. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 146:1–7, Jun 2021.

[18] J. Shi and S. Zhu. Learning to amend facial expression representation via de-albino and affinity. *arXiv preprint arXiv:2103.10189*, 2021.

[19] M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[20] R. Vemulapalli and A. Agarwala. A compact embedding for facial expression similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[21] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8:131988–132001, 2020.

[22] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[23] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020.

[24] Y. Zhu, Q. Li, J. Wang, C. Xu, and Z. Sun. One shot face swapping on megapixels. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4834–4844, June 2021.