Expression Recognition using a Flow-based Latent-space Representation

Saandeep Aathreya
Department of Computer Science and Engineering
University of South Florida
Tampa, Florida 33620
Email: saandeepaath@usf.edu

Shaun Canavan
Department of Computer Science and Engineering
University of South Florida
Tampa, Florida 33620
Email: scanavan@usf.edu

Abstract—**Facial expression Recognition is a growing and important field that has applications in fields such as medicine, security, education, and entertainment. While there have been encouraging approaches that have shown accurate results on a wide variety of datasets, in many cases it is still a difficult problem to explain the results. To enable deployment of expression recognition applications in-the-wild, being able to explain why an particular expression is classified is an important task. Considering this, we propose to model flow-based latent representations of facial expressions, which allows us to further analyze the features and grants us more granular control over which features are produced for recognition. Our work is focused on posed facial expressions with a tractable density of the latent space. We investigate the behaviour of these tractable latent space features in the case of subject dependent and independent expression recognition. We employ a flow-based generative approach with minimal supervision introduced during training and observe that traditional metrics give encouraging results. When subject independent expressions are evaluated, a shift towards a stochastic nature, in the probability space, is observed. We evaluate our flow-based representation on the BU-EEG dataset showing our approach provides good separation of classes, resulting in more explainable results.**

I. Introduction

Facial expression recognition (FER) has a broad range of applications in medicine [1], security [2], and education [3] to name a few. There have been encouraging results in the field through investigation of computer vision and machine learning to encode expression information from facial features [4]. Earlier methods were devised with the notion of Ekman and Friesen [5], that stated emotions are perceived in the same way regardless of the culture. Conversely, recent developments in psychology and neuroscience argue otherwise which indicates that emotions are not universal and are highly subjective per person, context, and expression [6], [7]. The attempt to generalize expression is a challenging problem [8], however, to have real-world applications in human affect analysis [9] and human-computer interactions [10], it is a necessary step. With the advent of deep learning, researchers have been able to achieve state-of-the-art results on FER problems [11], [12], [13], but the interpretation of the complex relationship between the deep features, of different classes, is still a black-box concept. Although most of the literature has shown to perform reliably even on in-the-wild data [14] [15], these methods primarily focus on the output and the validation of the their output. This paper deviates from this traditional approach towards a more explicit way of modelling facial expressions with more control over the deep features that the model is able to produce and classify.

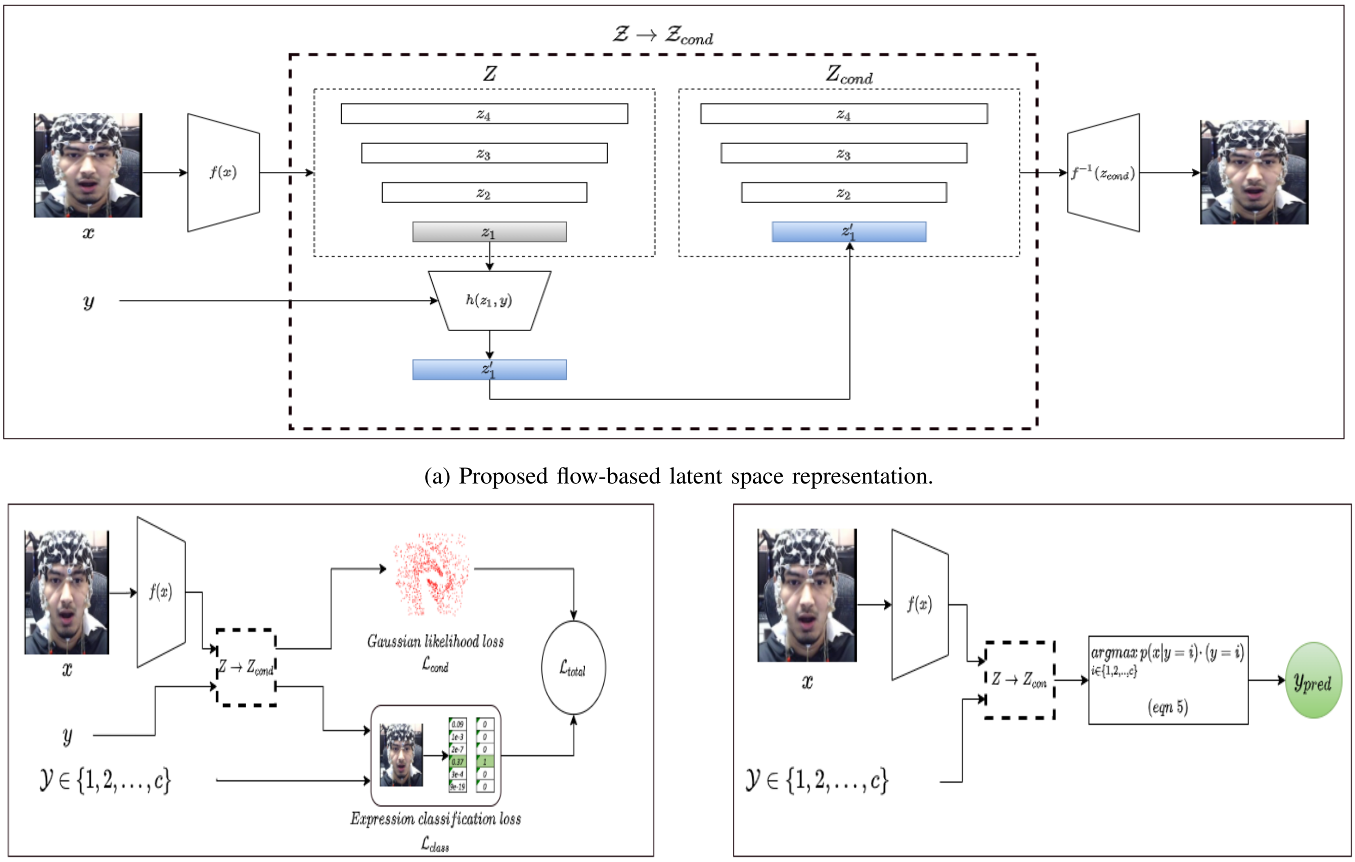
Traditional classification methods [17] [18] make the underlying assumption that the given dataset is a normal distribution. This is mostly false for high-dimensional natural signals such as images [19], which can result in an overall decrease in the accuracy of recognition models [20]. Recent advances in generative modelling has taken the field one step closer to a more tractable density estimation of high-dimensional images. For this purpose, we employ state-of-the-art generative models because of their ability to find meaningful distributions in the latent space [21], [22]. Specifically, flow-based generative models [23] have gained traction in recent years due to their ability to find explicit densities of the given dataset. Normalizing flows [24] are a simple, yet powerful technique which are capable of transforming densities of complex data into simpler forms using bijective and differentiable series of functions. Once the multiplex data has been transformed to a simple distribution, techniques such as Gaussian mixture modelling, and maximizing the log likelihood can be applied for classficiation problems (see Section III-A). Considering this, the contributions of this work are 3-fold:

1) A flow-based latent representation of facial expression is proposed. We model these representations for the task of expression recognition, in both a subject dependent and independent manner.
2) We visualize the high-dimensional features in the latent space and explore the subjective nature of expressions and observe notable differences between expressions of different subjects.
3) Using a state-of-the-art generative model, GLOW [16], we demonstrate the ability to interpolate between different expressions validating that the flow-based latent vectors form meaningful representations of expression.

II. Related Works

A. Generative Models

A range of generative models for classification purposes have been explored. GANs [25] work by finding the implicit



(a) Proposed flow-based latent space representation.

(b) Training for facial expression recognition.

(c) Testing for facial expression recognition.

Fig. 1: Overview of proposed architecture for flow-based latent space representation of facial expressions. (a) $f(x)$ is the original Glow model [16] which outputs the latent vector $Z = \{z_4, z_3, z_2, z_1\}$. $h(z_1, y)$ is a conditional flow model which takes as input the latent sub-vector z_1 and the one-hot encoded class label y and outputs z'_1, which is then used as a sub-vector in the new latent space Z_{cond}. (b) x is the batch of input images and y is the set of known expression classes (e.g. happy) which are fed into the proposed model. Each iteration outputs two loss values, $\mathcal{L}_{cond}$ is the conditional loss (Equation 3) and $\mathcal{L}_{class}$ is the class loss (Equation 5). $\mathcal{L}_{total}$ is the combination of L_{cond} and $\mathcal{L}_{class}$ (Equation 6). (c) y_{pred} is computed by running the test input x_{test} through $f(x)$ once and $h(z_1, y_c)$ c times where $c \in \{1, 2, .., c\}$ and taking the argmax of the probabilities of the models output (Equation 4). We refer the reader to Section III for more details.

density of the dataset through their adverserial structure and classification is done via the discriminator. Another type of generative models are VAEs [26], which inexplicitly optimize the log-likelihood by maximizing the ELBO. These models are not suitable for classification as they suffer from posterior collapse [27] wherein the density of the model closely matches the uninformative prior of the subset of latent data.

Yang et al. [28] used conditional generative adversarial networks to generate six prototypical expressions, which are then used to fine-tune convolutional neural networks. They look at the minimum distance between an input image and the generated images for classification. They report state-of-the-art results on multiple publicly available datasets. Xie et al. [29] proposed a 2-branch generative adverserial network that disentangles identity and expression information. They showed that this approach learns a discriminative representation of expression that is well suited for classfication.

B. Flow-based Modeling

Semi-conditional Normalizing flows [30] employed a combination of unconditional (f_w) and conditional flows (h_θ) wherein they concatenated the hidden features with the one-hot encoded vector of labels. This new vector was then passed on to the conditional flow (h_θ) which was used for classification purposes. This is similar to semi-supervised conditional GANs [31]. Experiments were performed on toy datasets and the MNIST classification problem. Inspired by the findings of this work which utilizes only last k dimensions of hidden data for conditional flow transformation, we employ a similar approach to avoid overfitting the model and maintain the balance between maximizing likelihood and minimizing the classification loss. Given k classes, this approach requires only 1 forward pass to classify a new test data.

FlowGMM [32] utilized the RealNVP [33] model to train each class to be associated with a different mean (μ_k) and standard deviation (σ_k). Post training, Bayes' decision rule was applied on any new test point to gather the class with max probability. This involved random generation of means and standard deviation to be assigned to different classes. Training was semi-supervised with only 10% of the data being labelled at each epoch. They were also able to retain the quality of the generated images and applied the method on several image classification problems (e.g. MNIST [34]). They also show that classification can be extended beyond images by performing text classification on different dataset such as UCI, AG-News, Yahoo answers and found that the methods outperform other traditional classification methods.

C. *Expression and Explainability*

Although there are less works that focus on explainability and expression, there are some interesting works that do focus on this area. For example, Kandeel et al. [35] used explainability to determine the best convolutional neural network architecture for recognizing the expression of drivers. They investigated the saliency maps of the output from the networks to determine the best architecture to use. They found that using this approach resulting in improved architecture selection, which ultimately lead to improved driver expression recognition accuracy. Weitz et al. [36] investigated using Layer-wise Relevant Propagation and Local Interpretable Model-agnostic Explanations to help explain how neural networks distinguished between expressions of pain and other expressions such as happy. They were able to distinguish key areas of the face that separated painful expressions from the other expression classes. Escalante et al. [37] designed a challenge around explaining video interviews. Their challenge proposes that an explainable system must be understandable by people in affective computing, signal processing social sciences, and psychology. Considering this, the challenge evaluation criteria included clarity, explainability, and soundness of the result.

We are motivated by these works, as being able to explain how different facial expressions are recognized can help the system better communicate with it's users [38], which can lead to to more public trust in real-world affective systems [39]. Considering this we extend the state of the art by incorporating generative models, along with flow-based models to give more robust visualizations and explanations for facial expression recognition. We show that the proposed approach allows for clear visualization of clusters of subject's expressions.

III. Flow-based Latent Representation of Facial Expressions

A. *Normalizing flows*

Normalizing flows [24] are powerful distribution approximators which are comprised of a chain of transformations that transform a complex distribution into a simple one. Mathematically, it is defined as a bijective mapping $f : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{X}$ defines the data space and $\mathcal{Z}$ defines the density of the latent space, which is typically chosen to be Gaussian. To infer the unknown probability density $\mathcal{X}$, we apply the inverse of the transformations $f^{-1} : \mathcal{Z} \to \mathcal{X}$ to generate new data from the data space $\mathcal{X}$ using change of variables theorem. Fig **??** provides a brief description of the functioning of normalizing flows, as applied to facial expressions.

Given a multivariate random variable z with a probability density $z \sim \pi(z)$, which is a function of unknown variable x given by $z = f(x)$ (and so, $x = f^{-1}(z)$), we need to infer the probability density of $x \sim p(x)$. *Change of Variable* theorem states that during any transformation, the total probability mass must be preserved, therefore the density of both z and x must always sum up to 1.

$$\int p(x)dx = \int \pi(z)dz = 1$$

The new density $p(x)$ is then the product of original density $\pi(z)$ and ratio of the volumes, which is typically given by calculating the determinant of Jacobian dz/dx.

$$p(x) = \pi(z) \cdot \left| det \frac{dz}{dx} \right|$$

Substituting $z = f(x)$, we get

$$p(x) = \pi(z) \cdot \left| det \frac{df(x)}{dx} \right|$$

Applying log on both sides, we get

$$\log(p(x)) = \log(\pi(z)) + \log \left| det \frac{df(x)}{dx} \right| \tag{1}$$

The calculation of log determinant at each step is expensive and therefore research involves finding an efficient way to avoid the direct calculation of log determinants. The functions $f(x)$ and $f^{-1}(z)$ are parametrized by deep neural networks whose core components are called *affine coupling layers*. These layers are defined by affine transformations of the input x as $y = s \odot x + t$, where s and t are neural networks. For more details on normalizing flows, we refer the reader to works from Dinh et al. [33], Kobyzev et al. [23], and Kingma et al. [16].

B. *Supervised Learning*

We examine the conditions where we are trying to solve a supervised classification task and learn a generative model simultaneously. For example, we may want to be able to generate new face images with an arbitrary expression and be able to classify the kind of expression that was generated. Subsequently, we use the feature space generated by Glow [16] to take full advantage of the labelled data available to us. Current works include the multi-scale architecture of Glow where the latent space Z is comprised of multiple sub-vectors $Z = \{z_4, z_3, z_2, z_1\}$. This kind of architecture enables fine-grained intermediate features which adds value to the intermediary representations [33]. We use this to add conditionality and supervision to the model. The overall architecture is shown in Fig. 1. It's a three-fold approach consisting of modifying the current architecture to include *conditionality and supervision*, *training* the modified architecture and *inferring* class labels from the final model.

Fig. 1a shows the overview of the proposed architecture. Similar to the work from Atanov et al. [30], the architecture consists of two parts, the original flow model $f(x)$ which maps $x \rightarrow Z$ and the smaller, conditional flow model $h(z_1, y)$ which maps $z_1 \rightarrow z_1'$. Here, h is a subset of f with much fewer layers and blocks (Section IV-B). We use a one-hot encoded vector of labels $\mathbf{y}$ and concatenate it with the latent sub-vector as $z_1 = concat(z_1, \mathbf{y})$ before passing it through h. Once we have the new latent sub-vector z_1', we concatenate it back with the original Z vector which now becomes $Z = \{z4, z3, z2, z_1'\}$. The proposed approach extends the work of Atanov et al. [30] in terms of optimizations. They compute the marginal likelihood $p(x)$ by optimizing the joint density $E_y p(x, y)$. We split this approach by first optimizing the conditional likelihood $p(x|y)$ through the loss function $\mathcal{L}_{cond}$ and implicitly optimizing $p(y)$ by minimizing the classification loss $\mathcal{L}_{class}$. This kind of decoupled approach allows us to independently monitor the key objectives involved in classification tasks, which in this case are the two losses $\mathcal{L}_{cond}$ and $\mathcal{L}_{class}$.

Training the model consists of maximizing the log likelihood in Equation 1 and also minimizing the classification loss at each epoch. We now focus on formulating the two losses $\mathcal{L}_{cond}$ and $\mathcal{L}_{class}$ under the new supervised conditions. Fig. 1b shows the overview of model training with 2 different losses while adopting the $\mathcal{Z} \rightarrow \mathcal{Z}_{cond}$ module from Fig. 1a. The new model h has dependence on the label y, so instead of maximizing the likelihood $\log p(x)$ from Equation 1, we now maximize the conditional $\log p(x|y)$ as

$$\log(p(x|y)) = \log(\pi(z)) + \log|det \frac{\partial f(x)}{\partial x}| + \log|det \frac{\partial h(z,y)}{\partial z}|. \tag{2}$$

This likelihood is then maximized by minimizing the conditional loss which is given by

$$\mathcal{L}_{cond} = -\log(p(x|y)). \tag{3}$$

Next, the classification loss, during training, is obtained by evaluating $p(y|x)$, on each of the c classes as

$$y_{pred} = \underset{i \in \{1,2,...c\}}{argmax}\ p(x|y=i)p(y=i), \tag{4}$$

where c is the total number of classes. It's important to note that the first two terms of $p(x|y = i)$,in Equation 2, undergo only one forward pass as it's independent of y. This approach accounts for relatively inexpensive calculations of log determinants multiple times [30].

The class loss is then calculated using the cross entropy loss on the prediction and labels

$$\mathcal{L}_{class} = CrossEntropy(y_{pred}, y). \tag{5}$$

The overall loss is given by the equation

$$\mathcal{L}_{total} = \mathcal{L}_{cond} + \lambda \mathcal{L}_{class}. \tag{6}$$

For our experiments, we have empirically found a λ of 0.3 to optimize both losses equally since the conditional loss $\mathcal{L}_{cond}$ and the classification loss $\mathcal{L}_{class}$ must converge synchronously. Fast convergence of $\mathcal{L}_{cond}$ might lead to model underfitting on classification and faster convergence of $\mathcal{L}_{class}$ might lead to overfitting the classification with poorly preserved probability density. See Fig. 1b for an overview of training. During *testing*, a new face image x_{test} is fed into f once and h for a total of c times, where we obtain the prediction by using Equation 4. See Fig. 1c for an overview of testing.

IV. Experimental Design and Results

A. Dataset

To validate the proposed flow-based latent representation of facial expressions, we evaluate the BU EEG [40] dataset. It is a multimodal emotion dataset which comprises of posed and authentic facial expressions, facial action units (FACS) [41] and EEG signals. The dataset contains data collected from 29 subjects of various ethnicity and backgrounds with 22 Asian, 2 White, 4 Mid-eastern and 1 from other ethnicity. For the facial features, there are 29 videos, for each subject, which is ~25 minutes in length at 24 fps with size $250 \times 350 \times 3$. Facial expression segments have been extracted from the videos as part of data preprocessing using the metadata files for 6 prototypical expressions - Anger, Disgust, Fear, Happiness, Sadness and Surprise, with a total of 54511 frames. These expressions are posed under the lab environment. The sequence of frames have been run through DeepFaceLab [42] face detector and cropped to a size of 256×256. The final size of the images have been kept at $64 \times 64 \times 3$. It is important to note that there is some imbalance, in terms of total number of frames for each expression, with $\sim 9\%$ difference in the minimum and maximum number of frames (surprise and disgust, respectively).

B. Implementation Details

The code is implemented in the PyTorch framework [43]. The model has been trained on 8 NVIDIA GPUs for a total of 1200 epochs with a learning rate of 1e-4 for both f and h. The batch size was kept at 32 with image size of $64 \times 64 \times 3$. The Glow model f consisted of 4 blocks of 32 stacked Conv-Relu-Conv layer (called flows). Output of each block corresponds to the latent sub-vector of $Z = \{z_4, z_3, z_2, z_1\}$. Similarly, the unconditional model h is the subset of f which consisted of 4 flows and 1 block. The model architecture has been adapted from Rosanality's[1] implementation of Glow.

C. Expression Recognition Results

The proposed approach can be *applied* to facial expression recognition, as can be seen in Table I. It is noteworthy to point out that to the best our knowledge, this is the first work that performs facial expression recognition using normalizing flows. Since our focus is to showcase the potential *applications* of these techniques in the domain of affective computing, we present our results along with other experiments extended through our work. The first row of Table I presents average

[1] https://github.com/rosinality/glow-pytorch

Data type	Classifier	Accuracy (in %)
z-(latent)	Flow based classifier (ours)	87.5
t-SNE embeddings	Random Forest	96.5
	Extra Trees	89.23
	AdaBoost	69.12
	Decision Tree	94.12

TABLE I: Accuracy score on different classifiers using z data and t-SNE embeddings of z data

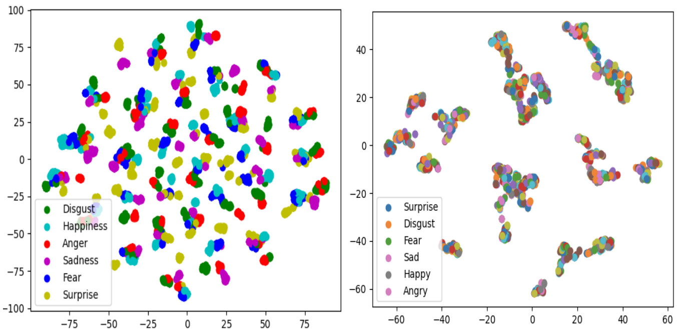
	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	820	90	12	2	7	0
Disgust	107	944	50	14	11	0
Fear	4	23	629	63	6	7
Happiness	2	10	49	926	10	3
Sadness	17	14	16	20	826	9
Surprise	6	22	67	38	27	882

TABLE II: Average confusion matrix of facial expression recognition for 10-fold cross validation.

accuracy of 10-fold cross validation on uniform folds of the entire BU EEG [40] dataset. As can be seen in Table II, the majority of expressions were recognized with relatively high accuracy. Overall, surprise had the lowest average misclassification error compared to other expressions. This can be explained, in part, by the large visual differences between surprise, and the other expressions, as can be seen in Fig. 4. These results are encouraging and further validate that meaningful information can be extracted from the low dimensional representations of the data.

D. Visualization and Explainability

We hypothesize that the proposed flow-based latent representation of facial expressions will provide accurate visualization and greater explainability. To test this hypothesis, we embed the latent representation of the z_1' latent vector into 2D space using the t-SNE method [44]. Fig. 2a shows the plot of test images belonging to 6 different expressions from 29 subjects. Each color corresponds to an expression class and it can be seen that 29 clusters are formed denoting each subject, with their expressions, per cluster. This naturally leads to the question, is this visualization any better than what we get with deep features? To answer this question, we juxtapose Fig. 2a with Fig. 2b, which is a t-SNE output of deep features. We used a convolutional neural network (CNN), on the same test data, which comprises of four stacks of residual blocks. Each block contains a pair of Conv-BatchNorm-ReLU layers followed by two dense connections. This model performed reasonably well, obtaining an accuracy of 74% on the test set, however, we can see no discernible pattern in the plot compared to the latent representation. Even though it has formed different clusters, they do not associate with either subjects or different expressions.



(a) Flow-based t-SNE.

(b) CNN t-SNE.

Fig. 2: Flow-based vs CNN comparison of t-SNE embeddings of latent data during subject dependent classification.

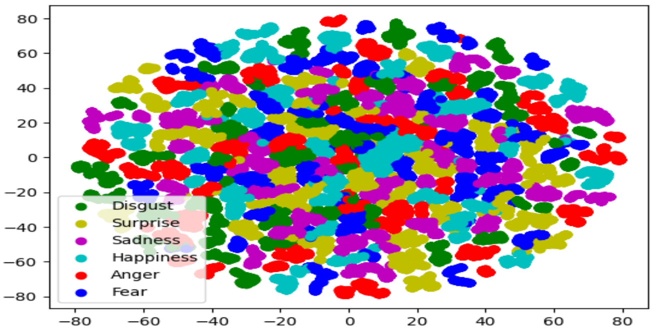


Fig. 3: Flow-based t-SNE embeddings of test data during subject independent classification.

Explainability of AI, especially in deep learning, is quite important when solving and realizing real-world problems. The proposed method is a step towards tackling the black box nature of neural networks which restricts the entry of AI into key fields such as medicine and security. The flow-based visualization can give us key insight into the subjective nature of expression [7]. We are able to see that the flow-based latent approach to represent expression was able to extract meaningful information such as each subject and their corresponding expressions are separable, therefore they are largely unique and can be easily classified. On the other hand, the visualization of the deep features does not offer this same insight as the clusters contain both subjects, and similar and different expressions.

To further explore this phenomenon of the latent space visualization we asked the question, will this specific clustering hold when subject independent experiments are conducted? To answer this question, we trained on 28 subjects of the BU-EEG dataset, and left one out (subject 29) for testing. The t-SNE plot for the latent vector of this test subject can be seen in Fig. 3. We see that the patterns veers away from the deterministic nature (Fig. 2a) to a more stochastic behaviour. This again aligns with work that details the biased and subjective nature of human emotions (e.g. expression) [45]. This visualization allows insight into the difficulties

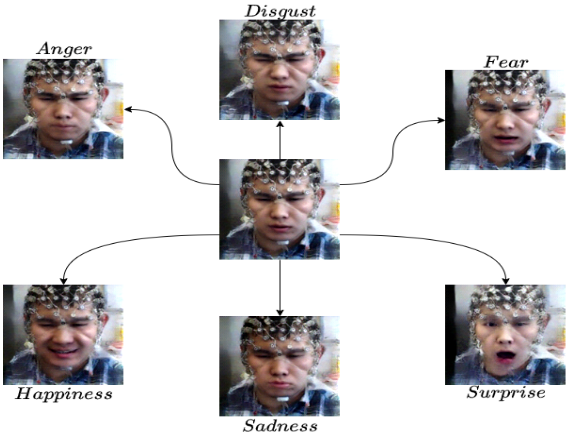


Fig. 4: Interpolation of 6 prototypical expressions from neutral.

associated with generalization of facial expressions, such as different expressions of emotion are fuzzy, and can overlap (i.e. they are similar) [46].

In addition to this, we use an extended application of Glow [16] to visualize the interpolation of a test face image to other prototypical expressions. To generate high quality facial expressions, we zero out the z_1 of the latent vector Z from our model and run it for a total of 6000 epochs. This has to do with the fact that retaining the z_1' sub-vector in Z_{cond} deters the reconstructed image. Then, for each expression class c, we calculate pairs of averages in the latent space $(z^c_{avg}, z^{other}_{avg})$ for all the training data, where z^{other}_{avg} is the average of all the other classes combined except c. To interpolate to a specific expression for an arbitrary image x_{test}, we add the corresponding z_{test} with the difference of the z^c_{avg} and z^{other}_{avg}. See Fig. 4 for generation of the 6 prototypical expression from a neutral expression. This experiment was conducted to highlight the potential use cases of this approach in generating new facial expression to accommodate for class imbalance problems. As our model is able to distinguish between different expressions subjectively, the generated expressions for subjects will be in close proximity of the subject's original expression.

V. Discussion

We detail some of the potential use-cases of the proposed method in the context of ethics and privacy for applications in affective computing. To be able to enhance a machine's ability to decode and respond to the affective states of a human, there is potential breach of privacy involved [47]. With the high quality generative ability of flow-based models, it is possible to address some of these ethical concerns and simultaneously leverage the full extent of meaningful classification that it performs. Federated learning [48] can be incorporated in the current method to have a more personalized and secure system in place. Only the deepest latent features z can be processed centrally and rest of the modules, f and h can be trained locally avoiding any bottlenecks. These latent features z, although play a significant role in modelling the true data, are usually considered *Gaussian noise* in the outside world. Moreover, the encoder-decoder nature of $f(x)$ and $f^{-1}(z)$ can be employed as encryption and decryption keys [49] further bolstering the security aspects.

A difficult task in affective computing is the collection of robust, accurate data [50]. Collecting data with true annotations, be it for expression or emotion, involves immense amounts of labor. Due to the flow-based model's ability to interpolate between the latent space to produce meaningful facial images [16] (as shown in Section IV-D), we can potentially impart subjective attributes of the facial features without the need of the original data. This reduces the need for copious amounts of data where relevant tasks can then be managed only through sample data. This also gives the opportunity to steer clear from some of the ethical concerns encountered along the way, such as collecting data of painful expressions [51]. This has the potential to support these future applications (e.g., pain recognition) of affective computing.

The proposed approach has broad impacts in fields such as medicine, security, and defense. We hypothesize that the latent space representations are useful for medical applications such as recognition of disorders including, but not limited to, Autism Spectrum Disorder and Post Traumatic Stress Disorder. As previously mentioned, using Federated Learning along with the proposed method can significantly improve security and user privacy. We also hypothesize that the latent space can be used to represent different signals aside from images. This can include physiological signals such as heart rate and EEG, thermal images, and 3D and 4D facial models. Considering this, our future work includes investigating the flow-based latent representation of EEG data, which may allow for a less noisy representation of the signal [52]. Along with this, we will also investigate adding more generalizability to the model, however, the subjective nature of facial expressions and emotions have been explored before. Our findings align with that of Hinduja et. al. [53] which statistically showed, by evaluating facial expressions, that self-reported emotions are different and subjective compared to expected emotions.

VI. Conclusion

This work investigates the idea of using generative flow-based models for performing interpretable and comprehensible modeling of latent representations in the domain of affective computing. We show that these models are able to transform the image signals into clear, segregated clusters in the latent space. Our results suggest the subjective nature of expression giving insight into how expression clusters by subject facilitating accurate recognition. We also explore the applicability of this work to perform supervised expression recognition on a posed facial expression dataset (BU-EEG). Finally, we detail potential use cases and broader impacts, to establish the proposed method in real-world applications by addressing some of the ethical and privacy concerns.

VII. Acknowledgement

This material is based upon the work supported in part by the National Science Foundation under grant CNS-2039373. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] G. Muhammad, M. Alsulaiman, S. U. Amin, A. Ghoneim, and M. F. Alhamid, "A facial-expression monitoring system for improved healthcare in smart cities," *IEEE Access*, vol. 5, pp. 10 871–10 881, 2017.

[2] A. A. M. Al-modwahi *et al.*, "Facial expression recognition intelligent security system for real time surveillance," in *World Congress in Computer Science, Computer Engineering, and Applied Computing*, 2012.

[3] J. Khalfallah and J. B. H. Slama, "Facial expression recognition for intelligent tutoring systems in remote laboratories platform," *Procedia Computer Science*, vol. 73, pp. 274–281, 2015.

[4] M. Liu *et al.*, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Computer Vision – ACCV 2014*. Cham: Springer International Publishing, 2015, pp. 143–157.

[5] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[6] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[7] L. F. Barrett *et al.*, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[8] I. O. Ertugrul *et al.*, "Cross-domain au detection: Domains, learning approaches, and measures," in *FG)*. IEEE, 2019, pp. 1–8.

[9] P. Lucey, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, and K. M. Prkachin, "Automatically detecting pain using facial actions," in *ACIIW*, 2009, pp. 1–8.

[10] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009, visual and multimodal analysis of human spontaneous behaviour:.

[11] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," in *CVPRW*, 2017.

[12] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *arXiv preprint arXiv:1902.01019*, 2019.

[13] M. A. Takalkar and M. Xu, "Image based facial micro-expression recognition using deep learning on small datasets," in *2017 international conference on digital image computing: techniques and applications (DICTA)*. IEEE, 2017, pp. 1–7.

[14] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *Journal of Electronic Imaging*, vol. 25, no. 6, pp. 1 – 8, 2016.

[15] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, pp. 82–92, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219306137

[16] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," 2018.

[17] D. Reynolds, *Gaussian Mixture Models*. Boston, MA: Springer US, 2009, pp. 659–663. [Online]. Available: https://doi.org/10.1007/978-0-387-73003-5_196

[18] K. McGarigal, S. Stafford, and S. Cushman, *Discriminant Analysis*. New York, NY: Springer New York, 2000, pp. 129–187. [Online]. Available: https://doi.org/10.1007/978-1-4612-1288-1_4

[19] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE Transactions on Affective Computing*, 2020.

[20] Y. Song, L.-P. Morency, and R. Davis, "Distribution-sensitive learning for imbalanced datasets," in *FGW*, 2013.

[21] P. Bojanowski, A. Joulin, D. Lopez-Paz, and A. Szlam, "Optimizing the latent space of generative networks," 2019.

[22] A. Nguyen *et al.*, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *CVPR*, 2017.

[23] I. Kobyzev, S. Prince, and M. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[24] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *ICML*, 2015.

[25] I. J. Goodfellow *et al.*, "Generative adversarial networks," 2014.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.

[27] J. Lucas, G. Tucker, R. Grosse, and M. Norouzi, "Understanding posterior collapse in generative latent variable models," 2019.

[28] H. Yang *et al.*, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *FG*, 2018.

[29] S. Xie, H. Hu, and Y. Chen, "Facial expression recognition with two-branch disentangled generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[30] A. Atanov *et al.*, "Semi-conditional normalizing flows for semi-supervised learning," 2020.

[31] K. Sricharan *et al.*, "Semi-supervised conditional gans," *arXiv preprint arXiv:1708.05789*, 2017.

[32] P. Izmailov *et al.*, "Semi-supervised learning with normalizing flows," 2019.

[33] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," *CoRR*, vol. abs/1605.08803, 2016. [Online]. Available: http://arxiv.org/abs/1605.08803

[34] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[35] A. A. Kandeel *et al.*, "Explainable model selection of a cnn for driver's facial emotion identification," in *ICPRW*, 2021.

[36] K. Weitz *et al.*, "Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable ai methods," *tm-Technisches Messen*, vol. 86, no. 7-8, pp. 404–412, 2019.

[37] H. J. Escalante *et al.*, "Design of an explainable machine learning challenge for video interviews," in *IJCNN*, 2017.

[38] R. Goebel *et al.*, "Explainable ai: the new 42?" in *International cross-domain conference for machine learning and knowledge extraction*. Springer, 2018, pp. 295–303.

[39] R. Cowie, "Ethical issues in affective computing," in *The Oxford handbook of AC*. Oxford University Press, 2015, pp. 334–348.

[40] X. Li *et al.*, "An eeg-based multi-modal emotion database with both posed and authentic facial actions for emotion analysis," in *FG*, 2020.

[41] L. Rothkrantz *et al.*, "Facs-coding of facial expressions." Association for Computing Machinery, 2009.

[42] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang, "Deepfacelab: A simple, flexible and extensible face swapping framework," 2020.

[43] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[44] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[45] L. Nummenmaa, R. Hari, J. K. Hietanen, and E. Glerean, "Maps of subjective feelings," *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. 9198–9203, 2018. [Online]. Available: https://www.pnas.org/content/115/37/9198

[46] S. C. Widen *et al.*, "Anger and disgust: Discrete or overlapping categories," in *APS Annual Convention*, 2004.

[47] X. Hu *et al.*, "Ten challenges for eeg-based affective computing," *Brain Science Advances*, vol. 5, no. 1, pp. 1–20, 2019. [Online]. Available: https://doi.org/10.1177/2096595819896200

[48] O. Rudovic *et al.*, "Personalized federated deep learning for pain estimation from face images," *arXiv preprint arXiv:2101.04800*, 2021.

[49] E. Habler and A. Shabtai, "Using lstm encoder-decoder algorithm for detecting anomalous ads-b messages," *Computers & Security*, vol. 78, pp. 155–173, 2018.

[50] D. Melhart, A. Liapis, and G. N. Yannakakis, "The affect game annotation (again) dataset," *arXiv preprint arXiv:2104.02643*, 2021.

[51] N. Berthouze *et al.*, "Emopain challenge 2020: Multimodal pain evaluation from facial and bodily expressions," *arXiv preprint arXiv:2001.07739*, 2020.

[52] D. Fabiano and S. Canavan, “Emotion recognition using fused physiological signals,” in *ACII*. IEEE, 2019, pp. 42–48.

[53] S. Hinduja, S. Canavan, and L. Yin, “Recognizing perceived emotions from facial expressions,” in *FG*, 2020.