Latent Encodings for Valence-Arousal Structure Alignment

Surabhi S. Nath
IIIT Delhi, India
surabhi16271@iiitd.ac.in

Vishaal Udandarao
IIIT Delhi, India
vishaal16119@iiitd.ac.in

Jainendra Shukla
IIIT Delhi, India
jainendra@iiitd.ac.in

Abstract
In recent years, great strides have been made in the field of affective computing. Several models have been developed to represent and quantify emotions. Two popular ones include (i) categorical models which represent emotions as discrete labels, and (ii) dimensional models which represent emotions in a Valence-Arousal (VA) circumplex domain. However, there is no standard for annotation mapping between the two labelling methods. We build a novel algorithm for mapping categorical and dimensional model labels using annotation transfer across affective facial image datasets. Further, we utilize the transferred annotations to learn rich and interpretable data representations using a variational autoencoder (VAE). We present “LeVAsa”, a VAE model that learns implicit structure by aligning the latent space with the VA space. We evaluate the efficacy of LeVAsa by comparing performance with the Vanilla VAE using quantitative and qualitative analysis on two benchmark affective image datasets. Our results reveal that LeVAsa achieves high latent-circumplex alignment which leads to improved downstream categorical emotion prediction. The work also demonstrates the trade-off between degree of alignment and quality of reconstructions.

Keywords
Affective computing, representation learning, variational autoencoder, valence, arousal, circumplex model

1 Introduction
Emotions are intrinsic characteristics of most living species, particularly overt in human behaviour [5, 18, 29]. Intelligent systems must employ means to incorporate emotions for a more natural interaction [31]. This surge for “emotional intelligence” has evolved into the field of affective computing, which by definition encompasses the creation of and interaction with machines that can sense, recognize, respond to, and influence emotions [32]. Several models of emotion have been developed over the years, which are considered as the backbone of affective computing [12, 13, 25, 40]. Among these models, a popular choice is the Categorical Model which describes six basic discrete emotions, namely, happiness, anger, disgust, sadness, fear, and surprise [9]. However, this model failed to capture relations between the discrete emotions. Moreover, there is a lack of consistency in the choice of these fundamental emotions [8]. As a result, Russell & Mehrabian [35] developed the Dimensional Model which suggests that each emotional state can be defined in terms of Valence (pleasure of an emotion), Arousal (energy of an emotion) and Dominance (controlling nature of an emotion). The Dominance dimension is commonly ignored since the valence-arousal (VA) dimensional model was shown to possess adequate reliability, convergent validity, and discriminant validity [36]. This led to the conceptualization of the Circumplex Model to represent affective states as a circle in a 2D bipolar VA space [34]. The VA variables are typically considered independent [11].

The existence of different models of emotions result in a range of possible annotation strategies for affective data [7, 10, 24, 27, 28]. This poses two challenges: (i) building deep learning models on affective data, and (ii) drawing collective insights from multiple datasets having potentially different formats of annotations [6]. In this paper, we present a novel algorithm for mapping annotations of the Categorical Model to those of the Dimensional Model through annotation transfer across affective facial image datasets. The subsequent task following annotation mapping is to obtain meaningful data representations. With the increased use of deep neural networks and generative models, there have been significant advances in emotion modelling and affective computing [14, 20, 33]. Variational Autoencoders (VAEs) [21] are known to yield disentangled latent representations and generate new data samples [16, 17, 39]. They have been used extensively in affective computing to represent text, audio, image and electroencephalography (EEG) data [22, 41]. Applying VAEs on affective facial images to obtain disentangled image representations can (i) provide high quality feature representations for downstream tasks [2, 30], and (ii) serve applications like facial editing and data augmentation [23]. In our study, we obtain interpretable features by aligning the latent space of a VAE with the VA space. This enables improved affect classification and regression as demonstrated on two benchmark affective image datasets using a series of evaluation tasks.

Our major contributions are as follows:
(1) an annotation transfer algorithm for label transfer between Categorical and Dimensional models of emotion
(2) a regularised VAE model “LeVAsa” (Latent Encodings for Valence-Arousal Structure Alignment) that yields an interpretable latent space with an implicit structure aligned with the VA space

The rest of the paper is organized as follows. Section 2 presents our annotation transfer algorithm, the VAE model architectures and the datasets used in our experiments. Section 3 outlines the evaluation tasks conducted along with the obtained results. Section 4 concludes the paper and motivates future work.

2 Methods
In this section, we present our annotation transfer algorithm and describe our VAE model architectures. Our code and models are publically available 1.

1https://github.com/vishaal27/LeVAsa
2.1 Annotation Transfer Algorithm

For the task of annotation transfer between Categorical and Dimensional emotion models, we use an external reference dataset ($D_r$) containing both discrete categorical emotion labels ($I_l \in \{e_1, e_2, ..., e_n\}$, where $e_1, e_2, ..., e_n$ are the $n$ discrete emotional labels) and valence, arousal values ($v_i \in [limv_{\min}, ulimv_{\min}], a_i \in [lima_{\min}, ulima_{\min}]$, where $limv_{\min}, ulimv_{\min}, lima_{\min}, ulima_{\min}$ are the lower and upper limits for valence and arousal values respectively). Each data sample $x_i \in D_r$ thus has an emotion label $l_i$, a valence value $v_i$, and an arousal value $a_i$. $D_r$ serves as the standard based on which continuous or discrete VA values can be sampled for data points in a working dataset ($D$) with only emotion labels, or conversely, the most likely emotion labels can be obtained for data points in dataset $D_h$ with only VA tuples (Figure 2). Algorithm 1 is detailed as follows.

**Algorithm 1: Annotation transfer algorithm**

**Input:** reference dataset $D_r$, discrete categorical emotion labels $\{I_1, I_2, ..., I_{|D_r|}\}$, VA values $\{v_1, v_2, ..., v_{|D_r|}\}$ and $\{a_1, a_2, ..., a_{|D_r|}\}$, working dataset $D$ with discrete emotion labels, working dataset $D'$ with VA tuples

**Output:** VA values for the working dataset $D$, discrete emotion labels for the working dataset $D'$

1. Partition each sample $x_i \in D_r$ into $n$ groups ($g_1, g_2, ..., g_n$) based on discrete emotion labels $I_l \in \{e_1, e_2, ..., e_n\}$
2. For each group $g_j$, $j \in \{1, 2, ..., n\}$, obtain the mean valence $\mu_{v_j}$, standard deviation valence $\sigma_{v_j}$, mean arousal $\mu_{a_j}$, and standard deviation arousal $\sigma_{a_j}$ values
3. Generate ellipses $e_j$, $j \in \{1, 2, ..., n\}$ for group $g_j$ representing emotion $e_j$ with centre $(\mu_{v_j}, \mu_{a_j})$, semi major axis $\sigma_{v_j}$ and semi minor axis $\sigma_{a_j}$
4. To obtain VA values for data point $x_k$ in $D$ with label $I_{k}$, sample $(\nu_k, \alpha_k)$ from ellipse $e_k$ as: $\nu_k = x \cos \theta + \mu_{v_k}$, $\alpha_k = y \sin \theta + \mu_{a_k}$, where $x = \sqrt{\nu^2}, y = \sqrt{\alpha^2}$ and $\theta \sim [0, \pi], \sigma \sim [0, 2\pi]$
5. Convert $(\nu_k, \alpha_k)$ to discrete values by scaling and rounding-off if desired
6. To obtain emotion label for sample $x_k$ in $D_h$ with VA $(\nu_k, \alpha_k)$, find ellipse $e_k$ with centroid at least Euclidean distance from $(\nu_k, \alpha_k)$, and assign $e_k$ as most likely emotion

2.2 VAE model architectures

We train a generative model with an interpretable latent space with an implicit structure given a raw distribution of affective face images. We employ variational autoencoder based models because of their simple training protocols and structured inductive priors. We compare two VAE models, Vanilla VAE and LeVAsa. The latent space for both models was constructed to comprise three chunks. Figure 1 depicts our model architectures.

For the Vanilla VAE, no explicit alignment was imposed on the latent space, whereas for LeVAsa, we take inspiration from recent work [4, 19] and model the latent space as follows:

- $Z_a$ - subspace consisting of valence attributes $z_a$ that learn to encode the valence features of image samples
- $Z_v$ - subspace consisting of arousal attributes $z_v$ that learn to encode the arousal features of image samples

![Figure 1: Model Architecture](image)

Given a dataset of $N$ affective images $X = \{x_1, x_2, ..., x_N\}$, our VAE backbone consists of an encoder $f_\theta$ and a decoder $g_\phi$ given by:

$$z_{v_i}, z_{a_i}, z_{z_i} = f_\theta(x_i), x_i \in X$$

$$\hat{x}_i = g_\phi(z_{v_i}, z_{a_i}, z_{z_i})$$

$$z_{v_i} \in Z_v, z_{a_i} \in Z_a, z_{z_i} \in Z_z$$

We train the Vanilla VAE with a simple reconstruction loss along with a modified Kullback-Leibler (KL) loss (Eq. 1). We induce a $\text{N}(0, I)$ prior on all three attributes $z_v, z_a$ and $z_z$.

$$L_{KL} = \sum_{z \in \{z_v, z_a, z_z\}} KL\left(f_\theta(z|X)||\text{N}(0, I)\right)^2 \quad (1)$$

We employ the same backbone Vanilla VAE architecture for the LeVAsa model with two major modifications:

1. **Projection Heads:** We make use of two non-linear projection heads $h_v$ and $h_a$ which map the encoded valence and arousal representations $z_v$ and $z_a$ to the valence and arousal latent space (giving label representations $r_v$ and $r_a$) where VA-regularisation loss is applied. The projections obtained are represented as follows:

   $$r_{v_i} = h_v(z_{v_i}), r_{a_i} = h_a(z_{a_i})$$

2. **VA-regularization loss:** To impose an explicit alignment of the $z_v$ and $z_a$ attributes with the VA ground truth factors, we introduce a VA-regularization loss as follows:

   $$L_C = \sum_{i=0}^N (L(r_{v_i}, v_i) + L(r_{a_i}, a_i)) \quad (2)$$

where $L$ takes the form of MSE for continuous and BCE for discrete annotation type.

$$KL(P||Q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{\text{exp}(P(x))}{\text{exp}(Q(x))}\right)$$

$p$ and $q$ are the densities of arbitrary continuous distributions $P$ and $Q$ respectively.
The overall optimization objective for the LeVAsa model is:
\[
L_{\text{total}} = L_R + \lambda_{KL} L_{KL} + \lambda_C L_C
\]
where \(\lambda_{KL}\) and \(\lambda_C\) are hyperparameters.

2.3 Datasets

We use the following datasets in our experiments.

Annotation Transfer: AffectNet
- AffectNet [26] is the largest facial expression dataset, with over 420,000 annotated images and contains both continuous VA annotations in \([-1, 1]\) and discrete emotional labels in Neutral, Anger, Happiness, Sadness, Surprise, Fear, Disgust, Contempt, None, Uncertain and Non-face. The dataset also incorporates a wide diversity in gender, age and ethnicity, hence is an ideal choice for the reference dataset in the annotation transfer algorithm (Algorithm 1). The generated ellipses are shown in Figure 2.

Model Training: IMFDB, AFEW
- IMFDB [38] contains around 34,000 annotated zoomed-in facial images of 100 Indian actors, with only emotional labels Neutral, Anger, Happiness, Sadness, Surprise, Fear, Disgust and no VA supervision. Continuous and discrete VA supervision for IMFDB is obtained from annotation transfer using AffectNet. This is particularly well suited due to the similar nature of images in IMFDB and AffectNet datasets.

Model Training: IMFDB, AFEW
- AFEW [7] on the other hand, contains around 24,000 annotated images from videos of real world scenes of approximately 600 actors with only discrete VA values in \([-10, -9, \ldots, 9, 10]\) and no discrete emotional labels.

The different nature of IMFDB and AFEW datasets allow us to analyse and compare model performance based on different factors including image type (zoomed in faces/video scenes) and annotation type (discrete VA supervision/continuous VA supervision).

3 Experiments

We perform our analyses and evaluations through a series of qualitative and quantitative experiments. This enables comparisons based on three aspects: (i) architecture (Vanilla VAE vs LeVAsa), (ii) dataset (IMFDB vs AFEW), and (iii) nature of annotations (Continuous VA vs Discrete VA). Altogether, we train five models: (i) Vanilla VAE on IMFDB, (ii) LeVAsa on IMFDB with continuous VA annotations, (iii) LeVAsa on IMFDB with discrete VA annotations, (iv) Vanilla VAE on AFEW, and (v) LeVAsa on AFEW with discrete VA annotations.

3.1 Latent-Circumplex Alignment

We measure the alignment of LeVAsa’s \(Z_v \cup Z_a\) latent space with the VA ground truths using normalized Euclidean and Manhattan distance metrics for continuous annotations, and Cross Entropy measure for discrete annotations. This helps quantify the degree of latent-circumplex alignment. For the Vanilla VAE, we determine the \(z_v\) and \(z_a\) chunks heuristically by considering the two latent chunks which aligned best with the corresponding valence and arousal ground truths. Further, we reduce the dimensionality of the \(z_v\) and \(z_a\) latent chunks and plot them alongside the ground truth to replicate the circumplex representation.

It is found that LeVAsa outperformed Vanilla VAE for both continuous and discrete annotations (Table 1). This clearly exhibits the superior latent-circumplex alignment achieved by LeVAsa. For discrete annotations, in case of AFEW, the difference between the cross entropy measures of the Vanilla VAE and LeVAsa is greater than in case of IMFDB. This could be attributed to the different image types in both datasets. The circumplex plots (Figure 3) for LeVAsa reveal reduced variance and increased alignment with true labels. This validates the quantitative results in Table 1.

| IMFDB | Valence MSE | MAE  | Arousal MSE | MAE | Combined MSE | MAE |
|-------|-------------|------|-------------|-----|--------------|-----|
| Vanilla VAE | 1.83 | 0.29 | 1.49 | 0.26 | 3.31 | 0.55 |
| LeVAsa | 0.14 | 0.14 | 0.06 | 0.09 | 0.2 | 0.23 |
| [a] Continuous | | | | | | |
| Model | IMFDB | AFEW |
| Vanilla VAE | 8.9 | 8.9 |
| LeVAsa | 6.63 | 2.54 |
| [b] Discrete | | |

Figure 2: Ellipses from AffectNet for annotation transfer

Figure 3: Circumplex Representation

To gain further insights, we assess the regressive power of the latent chunks \(z_v\) and \(z_a\) by their ability to predict the corresponding VA ground truths. We used Multi Layer Perceptron (MLP) Regression for this task. This analysis applies to continuous annotations hence it was conducted only on the LeVAsa and Vanilla VAE models trained on IMFDB dataset with continuous VA values.
It is observed that the MSE and MAE values computed for LeVAsa were lower by 2.25% and 1.42% as compared to Vanilla VAE for valence, and lower by 19.13% and 7.18% as compared to Vanilla VAE for arousal (Table 2). Furthermore, the goodness of fit metrics (explained variance and $R^2$) showed better performance in the case of LeVAsa. These results further strengthen our hypothesis.

### 3.2 Categorical Emotion Predictive Power

We predict the discrete emotion labels using different combinations of latent representations obtained from Vanilla VAE and LeVAsa (Table 3). Due to lack of discrete emotion labels in the AFEW dataset, it was excluded from this analysis. We randomized the data splits across Continuous and Discrete experiments to ensure an unbiased setup. Model performance is evaluated using classification accuracy. We utilize a simple one-layered MLP to ensure that the accuracy is a direct measure of representation quality and not influenced by the complexity of the classifier.

### Table 3: Categorical emotion predictive power using vanilla VAE and LeVAsa models. All reported scores are accuracies (@ represents vector concatenation)

| Annotation Type | Chunk Combination | Vanilla VAE (V) | LeVAsa (L) | Difference L - V (in %) |
|-----------------|-------------------|-----------------|------------|-------------------------|
| Continuous      | $z_v$              | 0.29            | 0.36       | 7                       |
|                 | $z_a$              | 0.32            | 0.35       | 3                       |
|                 | $z_z$              | 0.32            | 0.36       | 4                       |
|                 | $z_v \oplus z_a$   | 0.32            | 0.38       | 6                       |
|                 | $z_v \oplus z_a \oplus z_z$ | 0.29 | 0.33 | 4 |
| Discrete        | $z_v$              | 0.30            | 0.35       | 5                       |
|                 | $z_a$              | 0.27            | 0.30       | 3                       |
|                 | $z_z$              | 0.24            | 0.30       | 6                       |
|                 | $z_v \oplus z_a$   | 0.25            | 0.33       | 8                       |
|                 | $z_v \oplus z_a \oplus z_z$ | 0.26 | 0.30 | 4 |

It is seen that LeVAsa has significantly better predictive power as compared to the Vanilla VAE. Moreover, for LeVAsa, the VA chunks alone are more informative in emotion prediction as compared to $z_v \oplus z_a \oplus z_z$ chunks altogether. Also, the improvement in classification accuracy by employing LeVAsa in place of Vanilla VAE can be compared under the continuous and discrete settings. This reveals that LeVAsa representations from the model trained with discrete annotations and BCE loss (Eq. 2) proves to be better at classifying emotion labels. This is due to the discrete nature of emotion labels which correlate well with the model representations.

### 3.3 Reconstruction Quality

VAE models are prone to posterior collapse and can produce unreliable reconstructions [15, 37]. Thus, along with analyses of the latent representations, we also study the quality of the reconstructed faces (Figure 4).

It is observed that the quality of the reconstructed faces is slightly compromised in the case of LeVAsa as compared to Vanilla VAE. This can be attributed to the slightly higher variance of the learnt LeVAsa decoding distribution [1, 16]. By Shannon’s rate-distortion theory [3], there is a trade-off between the distortion (reconstruction quality) and rate (representation quality). Since we are imposing an explicit compression bottleneck on the latent representations, it is expected that the reconstruction quality is slightly compromised in order to achieve better interpretability of latent representations.

### 4 Conclusion

In this paper, we have developed an annotation-transfer algorithm for mapping between Categorical and Dimensional emotion model annotations. Using them, we generated interpretable image features with a VA-regularized VAE model called LeVAsa. We conducted a series of evaluation tasks to verify and validate our experiments and compare performance based on three factors: (i) architecture (Vanilla VAE vs LeVAsa), (ii) dataset (IMFDB vs AFEW), and (iii) nature of annotations (Continuous VA vs Discrete VA). The results showed that the LeVAsa model obtains robust and interpretable representations enabling improved downstream affective task performance. In the future, we hope to (i) extend the annotation-transfer algorithm to action-unit annotations, and (ii) perform latent traversals for data augmentation and facial editing.

### Acknowledgements

This work was supported by the Infosys Center for Artificial Intelligence at IIIT Delhi, India.
References

[1] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Richars Saurous, and Kevin Murphy. 2018. Fixing a broken ELBO. In *International Conference on Machine Learning*. 159–168.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[3] Toby Berger. 2003. Rate-distortion theory. *Wiley Encyclopedia of Telecommunications* (2003).

[4] Sarath Bhatag, Yishua Uddandaroa, and Shagoon Upal. 2020. *DisCont*: Self- Supervised Visual Attribute Disentanglement using Context Vectors. arXiv preprint arXiv:2006.05895 (2020).

[5] Charles Darwin and Phillip Proctor. 1998. The expression of the emotions in man and animals. *Oxford University Press*.

[6] Luna Da Bruyne, Pepa Atanasova, and Isabelle Augenstein. 2019. Joint Emotion Label Space Modelling for Affect Lexica. arXiv preprint arXiv:1911.08782 (2019).

[7] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Acted facial expressions in the wild database. *Australian National University, Canberra, Australia*. *Trans. Hum. Comput. Interact.* 2, 1 (2011), 1–13.

[8] Paul Ekman and Daniel Cordaro. 2011. What is meant by calling emotions basic. *Emotion review* 3, 4 (2011), 364–370.

[9] Paul Ekman and Wallace V Friesen. 1971. *Constants across cultures in the face and emotion*. *A sourcebook and manual to the facial action coding system*. 1971, 1–240.

[10] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. 2016. The extended cohn-kanade dataset (ck+): A complete dataset of spontaneous facial expressions in the wild. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5562–5570.

[11] Lisa Feldman Barrett and James A Russell. 1998. Independence and bipolarity in the structure of current affect. *Journal of personality and social psychology* 74, 4 (1998), 967–984.

[12] Jonathan Gratch, Stacy Marsella, Ning Wang, and Brooke Stankovic. 2009. Assessing the validity of appraisal-based models of emotion. In *Advances in the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–4.

[13] Stephan Hamann. 2012. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in cognitive sciences* 16, 9 (2012), 458–466.

[14] Jing Han, Zixing Zhang, and Bijorn Schuller. 2019. Adversarial training in affective VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*. 1–22.

[15] Qiuyang Hu, Attila Szabó, Tiarino Portenier, Paolo Favaro, and Matthias Zwicker. 2018. Disentangling factors of variation by mixing them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3399–3407.

[16] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VS N I Veeravasarapu. 2018. Disentangling factors of variation with cycle-consistent variational autoencoders. *In European Conference on Computer Vision*. Springer, 829–845.

[17] Na Lisa Feldman Barrett and James A Russell. 1998. Independence and bipolarity in the structure of current affect. *Journal of personality and social psychology* 74, 4 (1998), 967–984.

[18] Paul Ekman and Wallace V Friesen. 1971. *Constants across cultures in the face and emotion*. *A sourcebook and manual to the facial action coding system*. 1971, 1–240.

[19] Jonathan Gratch, Stacy Marsella, Ning Wang, and Brooke Stankovic. 2009. Assessing the validity of appraisal-based models of emotion. In *Advances in the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–4.

[20] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[21] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[22] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, 1–13.

[23] Raul Sanchez and Juan Raya. 2013. Product of Orthogonal Spheres Parameterization for Disentangled Representations. In *Computer Vision and Pattern Recognition*. 405–412.

[24] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2010. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *2010 20th International Conference on Pattern Recognition*. IEEE, 3695–3699.

[25] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2, 2 (2011), 92–105.

[26] Jaak Panksepp. 2004. *Affective neuroscience: The foundations of human and animal emotion*. Oxford University Press.

[27] Malte K Trivedi, Renuka Surya, and Raja Rajan. 2019. A preliminary study of Ekman and Cordaro, Izard, Levenson and Panksepp and Watt. In *Emotion Review*. 3, 4 (2011), 397–405.

[28] Chuan Wu, Fangzhao Wu, Siying Wu, Zhigang Yuan, Junxin Liu, and Yongfeng Huang. 2019. Semi-supervised dimensional sentiment analysis with variational autoencoder. *Knowledge-Based Systems* 165 (2019), 30–39.