GANmut: Learning Interpretable Conditional Space for Gamut of Emotions

Stefano d'Apolito[1], Danda Pani Paudel[1], Zhiwu Huang[1], Andrés Romero[1], Luc Van Gool[1,2]
[1]Computer Vision Lab, ETH Zürich, Switzerland [2]PSI, KU Leuven, Belgium
dstefano@alumni.ethz.ch {paudel, zhiwu.huang, roandres, vangool}@vision.ee.ethz.ch

Abstract

Humans can communicate emotions through a plethora of facial expressions, each with its own intensity, nuances and ambiguities. The generation of such variety by means of conditional GANs is limited to the expressions encoded in the used label system. These limitations are caused either due to burdensome labelling demand or the confounded label space. On the other hand, learning from inexpensive and intuitive basic categorical emotion labels leads to limited emotion variability. In this paper, we propose a novel GAN-based framework that learns an expressive and interpretable conditional space (usable as a label space) of emotions, instead of conditioning on handcrafted labels. Our framework only uses the categorical labels of basic emotions to learn jointly the conditional space as well as emotion manipulation. Such learning can benefit from the image variability within discrete labels, especially when the intrinsic labels reside beyond the discrete space of the defined. Our experiments demonstrate the effectiveness of the proposed framework, by allowing us to control and generate a gamut of complex and compound emotions while using only the basic categorical emotion labels during training. Our source code is available at https://github.com/stefanodapolito/GANmut.

1. Introduction

Facial expressions undoubtedly play a major role in the non-verbal communication of human emotions. However, the relationship between what is felt and the corresponding expression is complex, and not yet fully understood. When an emotion is externalized by facial muscle movements, sometimes even different emotions lead to the same expression [1, 11]. This is merely the beginning of the many issues in understanding human emotions. Therefore, modelling human emotions is a century-long ongoing topic of study [30, 28, 11, 31, 17, 16, 18, 29]. In this process, several psychological models for emotion representation have been proposed, with no clear consensus among psychologists. At this point, one may wonder, *what if we could leverage machine learning for emotion modelling instead?*

Existing emotion models rely upon psychologists, who individually have limited observations and personal biases. Machines, on the other hand, can potentially observe many more images of the diverse emotions. The key question that motivates us, maybe a little beyond the scope of this paper, is, *how can we make machines model emotions in a way that is also interpretable to humans?*

Before proceeding further, we first discuss the current issues we perceive. If a categorical model, such as basic emotions [10, 20, 27], is used to understand the lab controlled posed expressions, there is almost no need for a new emotion model[1]. On the contrary, understanding spontaneous expressions challenges even the human experts[2]. One can only expect a further deterioration of agreement if spontaneous expressions are annotated by non-experts. Therefore, it seems hopeless to collect large-scale data relying upon the categorical emotion model and its corresponding labels.

Several other emotion models also do exist [32]. Among these, the most commonly used are compound emotions [8], Valance-Arousal (VA) [31], and Action Units (AUs) [9]. These models, although not thoroughly studied, can be only suspected to pose more problems due to not being intuitive to non-experts. In this regard, the categorical emotion model is arguably the most intuitive one, as well as the most inexpensive means of collecting the data [3]. This leaves us with the necessity of developing a learning algorithm that can learn from imperfect emotion labels. With a slight twist, we ponder whether it is possible to *learn the emotional label space* itself, which can be tractably (or with meaningful interpretation) mapped to the available imperfect labels.

A major source of confusion in labeling is the discretization of labels, while emotions themselves are continuous (basic emotions as a psychological model are outdated [1]). Such discretization lacks two main aspects: intensity and fusion. For example, anger can be of different intensities, which may be expressed in sadness or fear [13]. More specifically, the hidden intensity and the intermediate emotions must be discovered to make the best use of the categor-

[1]Recognition accuracy on constrained CK+ [21] datasets is $\approx 100\%$.
[2]AffectNet [22] reports an agreement of only 60.7% between experts.

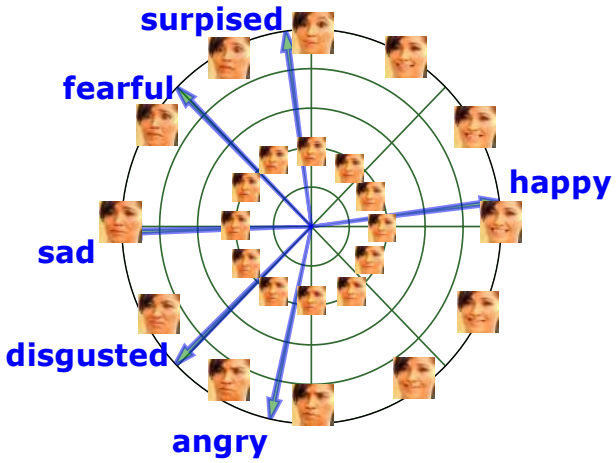

Figure 1: Conceptual illustration of a Gamut of emotions. We learn to generate diverse/complex emotions merely using the labels of basic categorical emotions (blue arrows).

ical labels. This is by no means surprising if the motivation of VA and compound emotions is considered. Therefore, the categorical labels for spontaneous expressions must be treated merely as proxies to the emotional states.

The main idea of this paper is to use categorical proxies to discover the hidden intensity and fusion. As shown in Figure 1, we represent every discrete emotion as an outgoing vector from the neutral emotion. The length of the vector provides the intensity of that emotion. We learn these vectors within the proposed framework. The fusion, on the other hand, is represented using the samples in the intermediate space between the vectors of basic emotions. Two sets of conditions are sampled from the conditional space: (i) those along the vectors, and (ii) the others. When the samples are selected along the vectors, we ensure that the generated images can be classified to the corresponding categorical emotion (represented by that vector), with confidence in accordance with the intensity. Moreover, the generated images must still be realistic, as well as recognizable in the learned conditional space. In other words, such images are used to regress the input conditions themselves. Such strategy encourages a smooth transition of emotions in the conditional space, as every generated image must be realistic and inversely mappable – thereby generating faces with various expressions mimicking the distribution of the real images, via adversarial training.

In summary, our learned conditional space is continuous and interpretable, which when used to generate images offer the gamut of emotions [7]. Such learning – only from categorical labels (*e.g.*, "happy") – is possible thanks to the proposed method. Note that the same is not possible by interpolating between emotions using existing methods, as they do not consider the issue addressed in this paper. More specifically, conventional conditional GANs fail to do so, as they use classification loss which encourages the generator to generate only easily recognizable emotions. In contrast, we introduce the learned labels of complex expressions and reproduce them. The major contributions of this paper are:

- We introduce the problem of learning conditional space for GANs, suitable for imperfect conditional labels. This problem is shown to be well suited for the task of emotion modelling and manipulation.
- A novel scheme for training conditional GANs, to search for condition interpretability, has been proposed. Our method enables us to generate a gamut of emotions, using only the categorical emotion labels.
- Our experiments on the benchmark dataset demonstrate the superiority of the proposed method, in terms of both qualitative and quantitative measures.

In fact, we also introduces an another conditional space, defined as a mixture of Gaussians, where each mode represents a basic emotion. This paper, however, will progress with the linear representation of basic emotions, for the sake of better readability. Once the linear model is well introduced and established, we shall proceed with the second approach. In the theoretical aspect, the Gaussian model differs only in terms of parameterization. From the experimental point of view, such modelling yields a better mixture (in some sense compound) of emotions [8].

2. Related Works

Facial expression conditioning. Automatically understanding facial emotions has been a very active and controversial topic in computer vision. Many theories have led to different machine learning approaches, of which three major affective computing models are: 1) Basic emotions, first described by *Ekman and Friesen* [10] uses seven discrete emotions such as "happy", "sad" or "surprised". 2) Valence-Arousal (VA) model [31] describes facial emotions in a continuous 2D space with Valence (*i.e.*, how negative/positive is the emotion) and Arousal (*i.e.*, how intense is the emotion) parameters. 3) Action Units (AUs) are fine-grained facial muscle movements [9], modelling expressions as an ensemble of several and distinctive facial muscles' contractions or relaxations.

For emotion annotation, categorical models are often preferred because of its simplistic nature, which leads to no strict demand on sharp skills. Unfortunately, categorical labels cannot describe mixed and more complex emotions. Although VA and AUs based models are better representative, VA is limited only in two qualities of valance and arousal. For example, both scared and angry expressions have high arousal and low valence, which makes them

largely ambiguous within VA-based modelling. On the other hand, AUs demand an expensive labelling process and yet do not reveal the emotion states directly.

Most existing methods [5, 24, 25] that use GANs [12] to produce impressive results in manipulating facial images, consider categorical emotions. In this regard, StarGAN [5] uses conditional GANs to inject categorical labels in the generator to produce domain-targeted images. GANimation [24], which also proposes AU conditioning, exploits a similar scheme while focusing on local transformations using an attention mechanism. Another method, SMIT [25] switches the StarGAN's deterministic output into a stochastic noise-driven manipulation. This allows SMIT to produce many outputs from a single input. A major shortcoming of [5, 24, 25] is their inability to go beyond labelled emotion definition. At most, they can interpolate emotions with often non-interpretable outcomes. Recently proposed StarGAN-v2 [6] is an alternative to produce photo-realistic conditioned image manipulation with impressive realism in the interpolations, which may also be suitable for facial expression manipulation. Nonetheless, as *Romero et al*. [26] suggested, StarGAN-v2 fails when involved domains are visually close, as it is the case of facial attributes or emotions.

For prior methods using conditional GANs [23], it is common to assume an auxiliary classifier besides the discriminator, which encourages the generator to reproduce existing yet easily recognizable labels. Thus, categorical labels cannot describe spontaneous expressions which usually are ambiguous and complex. Examples include, "happily surprised", "sadly fearful", their intensities, and other nuances. Consequently, conditioning upon the basic emotions will not allow producing a good variability of expressions.

In contrast to previous approaches, rather than conditioning on human-designed labels, we learn an expressive conditional space Z (see Figure 1), in which we can produce a gamut of emotions that are not explicitly labelled in the dataset, hence creating new *learned labels*.

Semantic structure of GANs' latent space. We aim to capture the Gamut of human emotions in a conditional space that allows for intuitive and seamless facial manipulation. Recently, *Voynov and Babenko* [36] proposed a method for the unsupervised discovery of human interpretable directions within the latent space of a trained generator G. A reconstructor R and a number n of learnable directions $d_i, i \in \{1, 2, ..., n\}$ are introduced. Given $G(x)$ and $G(x + \epsilon d_i)$, with x latent code and $\epsilon \sim \mathcal{U}([-c, c])$, d_i and R are learned so that the latter can regress i and ϵ. In this setup, d_i should acquire a precise semantic disentangled by others. Other interesting works on the inseparability of latent space come from *Härkönen et al*. [14] and *Schen et al*. [33]. Furthermore, *Laine* [19] observed that linearly interpolating in the latent space of a trained generator is not guaranteed to produce the smoothest transition by a visual, nor does it by a semantic point of view. Therefore, a method for searching by minimizing some feature-based loss was suggested. Since we aim to learn an expressive conditional space, we adapt a similar approach to make straight paths semantically meaningful.

3. Problem Formulation

We focus on the problem of producing arbitrary emotional facial expressions for a Gamut of emotions, given a dataset of real facial expression images $X = \{x_1, \ldots, x_N\}$ as well as their corresponding categorical labels on the basic emotions $C = \{c_1, \ldots, c_N\}$, $c_i \in \{1, \ldots, M\}$, where M is usually 7. One natural solution is to exploit the methodology of conditional generative adversarial networks (GANs), which play an adversarial game between a generator G and a discriminator D, to approximate the distribution of the real emotional facial expressions, conditioning on the given emotion labels. However, the major issue is that the given basic emotion labels cannot faithfully cover the Gamut of emotions of the real facial expressions due to the general ambiguity of these. For example, a happily surprised face might be only labelled as happy. On the other hand, as the conventional GAN methods generally rely on a static condition space Z that only encodes the given labels, they are merely expected to approximate the distributions over the pure basic emotions. Motivated by this, we introduce the problem of making the GAN conditional space learnable. In this way, it can deal with the imperfect emotion labelling issue so that the conditional GAN methods can be enhanced to discover the complete distribution of the given arbitrary emotional facial expressions.

Approach I: For the new problem, we should exploit parameterization techniques on the GAN conditional space, so that it can learn to encode the Gamut of emotions. One feasible strategy is to parametrise a 2D conditional space Z with polar coordinates (θ, ρ)[3]. The conditional latent code is interpreted as a random variable $z = (\theta, \rho)$, with its coordinates coming from a uniform distribution: $\theta \sim \mathcal{U}([0, 2\pi]), \rho \sim \mathcal{U}([0, 1])$. As illustrated in Figure 2 (left), θ is related to the quality/category of the emotion, whereas ρ represents its purity/intensity. For each basic emotion c_i, we parametrise its condition with $z = (\theta_{c_i}, \cdot)$, where θ_{c_i} is the learned direction for emotion c_i.

Approach II: As spontaneous facial expressions are often ambiguous, they could be better described by a label distribution rather than a single categorical emotion. This motivates us to propose also a second parameterization. The relative conditional latent code z is a 2D random variable uniformly distributed: $z \sim \mathcal{U}(Z)$, $Z = [-1, 1]^2$. In

[3]The shape and the dimension of the space is inspired by the circumplex model of affect of Russel [28]

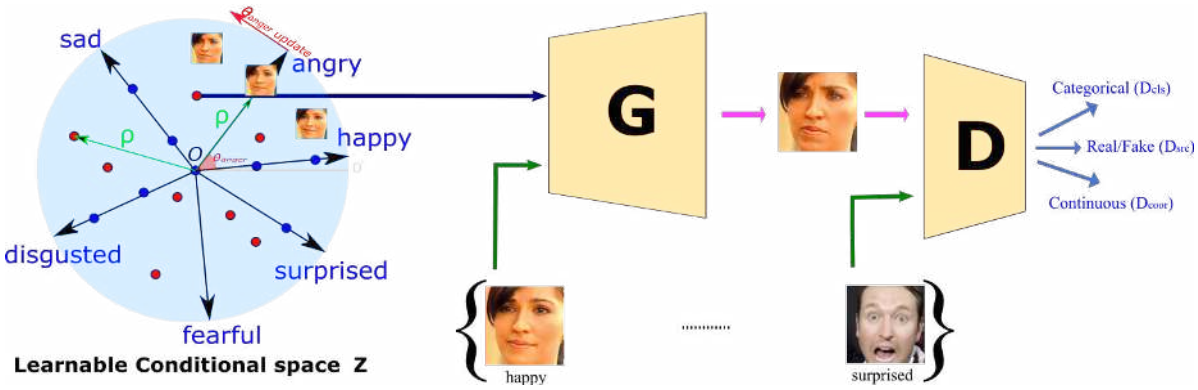

Figure 2: Overview of the proposed method. During the training, one part of the batch is conditioned with codes randomly sampled from $\mathcal{Z}$ (red points), and the other (blue points) sampled in the proximity of one of the learnable vectors (representing basic emotion). Only the second part of the batch undergoes the classification loss (for categorical labels), so that red points are free to encode any expression difficult to describe by basic emotions. All points are expected to generate realistic faces.

particular, the basic emotion c_i is represented by a mode $z = (\mu_{c_i}, \Sigma_{c_i})$. This is characterized by a mean $\mu_{c_i} = \tanh(w_{c_i}), w_{c_i} \in \mathbb{R}^2$ being a learnable parameter (the activation $\tanh()$ is used to constrain μ_{c_i} in Z). On the other hand, the corresponding covariance matrix $\Sigma_{c_i} \in \mathbb{R}^{2\times 2}$ can be parametrised by its eigenvalues $\sigma^2_{1,c_i}, \sigma^2_{2,c_i}$, and eigenvector orientation θ_{c_i}, which define the alignment and the length of the covariance ellipses axes respectively. Please refer Figure 3 for a visual illustration.

4. Proposed Method

In this section, we describe our approach to learning the GANmut conditional space. The key idea is to replace the static conditional space with our suggested parametrised one by employing an off-the-shelf conditional multi-domain GAN model (*e.g.*, StarGAN [5]), so the parametrised GAN model and the parametrised conditional space can be jointly optimized in an adversarial manner. The suggested polar parameterization on the conditional space enables us to sample not only from the labelled basic emotions but also to explore the space representation. The more complete and dynamic sampling on the conditional space further allows us to discover the distributions of the Gamut of emotions using a mixture of adversarial loss, classification loss, and regression loss, which are employed to enhance the generator and discriminator to produce realistic-looking emotional facial expressions, predict the basic emotion labels, and regress the continuous latent variables, respectively. The overview of our proposed model is illustrated in Figure 2. Additionally, we also apply a Gaussian parameterization of the conditional space (that we will call GGANmut). To this end, we employ the Kullback-Leibler (KL) divergence to replace the regression loss so that the model can discover the Gaussian modes of emotions. In the following sections, we explain in detail our two proposals: the linear model and the Gaussian one.

4.1. Linear Model

The primary purpose of this model is to exploit a more complete and dynamic sampling strategy on the polar characterization of the conditional space. On the one hand, the conditions associated with the basic emotions can be progressively optimized with the new strategy. On the other hand, the conditions associated with more complex emotions are sampled and updated simultaneously.

In particular, for each basic emotion c_i, by sampling a series of correlated latent codes $\hat{z}_j = (\theta_{c_i}, \hat{\rho}_j)$ where $\hat{\rho}_j$ is progressively increased (*i.e.*, $\hat{z}_j$ moves outward from the origin in direction θ_{c_i}), θ_{c_i} will be updated so that the generated images $y_{\hat{z}_j}$ should be classified as c_i with increasing confidence. Precisely, the discriminator D should classify $y_{\hat{z}}$ as c_i with the confidence being proportional to the distance $\hat{\rho}_{c_i}$ from the origin. As long as $\hat{\rho}_{c_i}$ is lower than a certain threshold τ, *i.e.*, $\hat{\rho}_j < \tau$ (we empirically set $\tau = 0.2$ throughout the paper), D should classify $y_{\hat{z}}$ with the neutral expression. A similar strategy applies for the remaining emotions with the parametrised condition being $z = (\theta, \rho) \in \mathcal{Z}$. The main difference is that ρ becomes now proportional to the highest emotion classification Softmax score attributed by D to the generated image y_z.

Following the new sampling strategy, we suggest a mini-batch sampling scheme. For each step of the gradient descend, the mini batch S of size n is split into 2 subsets (see Figure 2): 1) S_c for basic emotions containing n_c samples, and 2) S_r for the other emotions containing n_r samples. Based on this new sampling strategy, and as depicted in Figure 2, we plug the learnable conditional space into a regular conditional multi-domain GAN model[4], which generally consists of two components: one is generator G and

[4]We use the well established StarGAN [5] as our backbone.

the other is discriminator D. The latter is trained to distinguish between real and fake facial expressions, whereas G to fool D. Additionally, D serves as an emotion classifier, with G trying to produce correctly classifiable expressions. Rather than just synthesizing basic facial emotions, we need to learn also the conditional space to achieve a Gamut of emotions. In general, the problem of jointly learning a GAN model and a GAN conditional space learning is a bi-level problem. One feasible solution could be to optimize them separately. However, this is likely to lead to a bad optimization as they are highly dependent: a more expressive conditional space will increase the capacity of the GAN model to approximate the real data distribution more accurately, and a more powerful GAN model encourages the conditional space to become more expressive. Therefore, we suggest a joint training scheme for them.

For the training of the conditional GAN models, we apply the regular GAN loss $\mathcal{L}_{adv}$ (Equation 3) with Wasserstein [2] loss. Regarding the conditional GAN space, we enforce an emotion classification loss $\mathcal{L}_{cls}$ (Equation 4 and 5) to optimize progressively, besides G and D, the parametrised directions θ_{c_i} so that they can be aligned with the human labels. This will increase the interpretability of the learned basic emotions directions over the conditional space. Furthermore, we exploit a condition regression loss $\mathcal{L}_{info}$ (Equation 6), whose goal is to make D correctly estimate the expression coordinates $\hat{z} \in Z$ of real and generated images. This comes with increased mutual information between z and $G(x, z)$ [4]. The mixed-use of the three major losses is expected to optimize jointly both the GAN model and the conditional space. Moreover, we introduce $\mathcal{L}_\rho$ (Equation 7) to ensure that the interpolation along radii is semantically meaningful. More precisely, moving outward from the origin, emotion should be expressed more clearly, which often aligns with a stronger intensity. Finally, inspired by [5, 37], we also apply the cyclic loss $\mathcal{L}_{rec}$ (Equation 8) which performs a fundamental regularization action. The full objective function, respectively, for D and G is formulated as:

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^r_{cls} + \lambda_{info_D}\mathcal{L}_{info}, \tag{1}$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}^f_{cls} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{info_G}\mathcal{L}_{info} + \lambda_\rho\mathcal{L}_\rho, \tag{2}$$

where all the involved hyperparameters are used to make a trade-off among the correlated losses, and each loss is formulated as follows:

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbb{E}_x\left[D_{src}(x)\right] - \mathbb{E}_{x,z}\left[D_{src}(G(x,z))\right] \\ &- \lambda_{gp}\mathbb{E}_{\hat{x}}\left[\left(\|\nabla_{\hat{x}} D_{src}(\hat{x})\|_2 - 1\right)^2\right],\end{aligned} \tag{3}$$

$$\mathcal{L}^f_{cls} = \mathbb{E}_{x,c,\rho}\left[-\log D_{cls}(c \mid G(x, z_{c,\rho}(\theta_c)))\right], \tag{4}$$

$$\mathcal{L}^r_{cls} = \mathbb{E}_{x,c'}\left[-\log D_{cls}\left(c' \mid x\right)\right], \tag{5}$$

$$\mathcal{L}_{info} = \mathbb{E}_{x,z}\left[\|D_{coor}\left(G(x,z)\right) - z\|^2_2\right], \tag{6}$$

$$\mathcal{L}_\rho = \mathbb{E}_{x,z}\left[\|\hat{\rho}(G(x, z_{\cdot,\rho})) - \rho\|^2_2\, \mathbb{1}_{\rho>0.2}\right], \tag{7}$$

$$\mathcal{L}_{rec} = \mathbb{E}_{x,z}\left[\|x - G\left(G(x,z), D_{coor}(x)\right)\|_1\right], \tag{8}$$

where z is the conditional code passed to G during training, c' is the dataset emotion label, c a randomly sampled label ($c \overset{d}{=} c'$), $D_{src}(x)$ denotes the probability, given by D, of the input image x to be real, $D_{cls}\left(c' \mid x\right)$ the probability assigned to the correct label, and $D_{coor}(x)$ the estimated expression coordinates $\hat{z} \in Z$ of x. $\hat{\rho}(x) = \max_C(D_{cls}(c_i \mid x)))$. Finally, $\hat{x}$ is sampled uniformly along a straight line between a pair of a real and a generated images.

4.2. Gaussian Model (GGANmut)

With our suggested Gaussian parameterization of the conditional space, an emotion distribution $p_z(c_i)$ can be associated with each conditional code $z \in Z$ based on the Mahalanobis distances from the modes. More precisely, the Mahalanobis distance $d_{c_i,z}$ of z from the mode $(\mu_{c_i}, \Sigma_{c_i})$ is computed by $d_{c_i,z} = \sqrt{(z-\mu_{c_i})^T\Sigma^{-1}_{c_i}(z-\mu_{c_i})}$. We then associate an emotion distribution $p_z(c_i)$ with z such that $p_z(c_i) = \frac{e^{-d^2_{c_i,z}}}{\sum_j e^{-d^2_{c_j,z}}}$. This association is based on the rationale that if mode c_i has a Gaussian distribution, then $d_{c_i,z}$ would be proportional to the square root of the negative log likelihood that z is coming from mode c_i.

The final goal of the training is, given a certain z, to condition an image so that the new expression reflects the emotion distribution associated with the latent code. If we assume that the evaluation of D_{cls} follows human expression judgement, then we can quantitatively describe the expression of a generated images $x_z = G(x, z)$ with the distribution label $q_{x_z}(c_i) = D_{cls}(c_i \mid x_z)$. Therefore, the goal would be to minimise the expected divergence between $p_z(c_i)$ and $q_{x_z}(c_i)$. To this end, we consider the loss:

$$\mathcal{L}_{div} = \mathbb{E}_{x,z}\left[KL(p_z||q_{x_z}) + KL(q_{x_z}||p_z)\right], \tag{9}$$

where $KL(\cdot||\cdot)$ indicates the Kullback-Leibler (KL) divergence. To encourage expression variability, and similar to the previous model, for each iteration of the gradient decent algorithm, we divided the mini-batch S into two subsets S_r and S_c. The first split is conditioned with code $z \sim \mathcal{U}([-1, 1]^2)$, the second with μ_{c_i}. In this case we apply:

$$\mathcal{L}^f_{cls} = \mathbb{E}_{x,c}\left[-\log D_{cls}(c \mid G(x, \mu_c))\right] \tag{10}$$

Please note that the other loss terms including the adversarial loss and the reconstruction loss are similar to those (Equation 3 and 8) in the linear model, and they are reported in the supplementary material.

5. Experimental Evaluation

Implementation Details. Our implementation is based on StarGAN [5]. We set the number of training iterations to 1M for the proposed method and baselines, except for the Gaussian variant, for which we set 1.8M. We adopted the same training strategy of StarGAN, and the same hyper-parameters, if shared. The others were chosen with a naive qualitative approach. We observed a low sensitivity to λ_{info} (as in [4]). $\mathcal{L}_{info}$ can not be ablated as it is necessary to train D_{src}, whereas a study of the impact of $\mathcal{L}_{\rho}$ is reported in Section 5.2. The chosen hyper-parameters are: $\lambda_{cls} = 1, (0.5)$ (GGANmut), $\lambda_{gp} = 10$, $\lambda_{info} = 1$, $\lambda_{\rho} = 4$, $\lambda_{div} = 2$, $n_c = 9, (7)$ (GGANmut). The parameters of emotion vectors and Gaussian modes were randomly initialized. Please, refer to our supplementary material for more implementation details.

Datatest. We trained our models and baselines from scratch on AffectNet [22]. It is the largest dataset for affect computing, collecting $\approx$ 1M of images retrieved from the Internet. Queries were performed on major search engines (Google, Bing, Yahoo), using 1250 emotion-related keywords in six different languages. Notably, 450K of these images were annotated manually by 12 experts with basic emotions. As we seek to learn complex and challenging emotions in a very diverse setup, this dataset has become our ideal choice. We selected images labelled as Neutral, Happy, Sad, Surprised, Fearful, Scared, and Angry. For *all* considered methods only these categorical labels were used.

Baselines. We compared both of our methods with three state-of-the-art methods namely, StarGAN [5], GANimation [24], SMIT [25]. StarGAN is the backbone of our framework, GANimation adds an attention mask to focus changes only where needed. SMIT leverages random noise to produce multiple expressions for the same basic emotion. We used the default hyper-parameters and training strategy suggested by the authors of the competing methods. We needed, however, to increase the value λ_A of GANimation up to 1, to make $\mathcal{L}_A$ effective on AffecNet.

Evaluation metric. Since GANmut is not constrained to generate only categorical expressions, like StarGAN or SMIT, it can better reflect the variability of human emotion. To evaluate this quality, we adapt the popular Fréchet Inception Distance (FID) [15] score, by making it more focused on Emotion. We will call the new metric **Fréchet Emotion Distance (FED)**. To obtain FED, we trained a VGGNet [34] for emotion classification on AffectNet. We then feed the VGGNet with real and fake images and extract the features closest to the final classifier. Finally, assuming that the distributions of the two groups of features are Gaussian, we compute their Fréchet distance. The hope is that FED is as low as possible. The scores were computed on the test-set of AffectNet, containing 500 samples for each categorical emotion. For GANmut we conditioned images sampling uniformly from the conditional space, whereas for SMIT, GANimation, and StarGAN we sampled the categorical label from the test-set label distribution. For GGANmut, we used the mean of the modes relative to those labels and added some noise. We also provide other evaluation protocols, which will be introduced in the following subsection. As a proxy for human emotion evaluation, we will use the softmax score of the trained VGGNet.

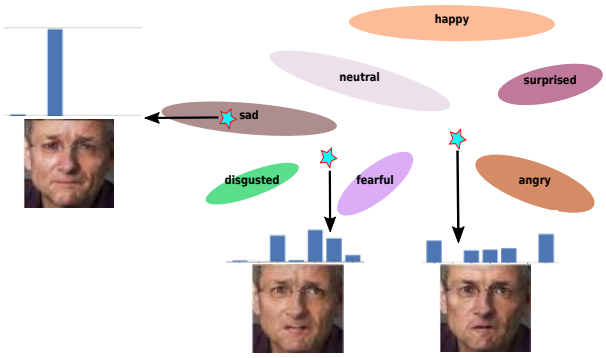

Figure 3: Learned conditional space for the Gaussian model (GGANmut). Each point of the space represents a label distribution whose expression can be generated. The distributions represent the confidence of neutral, happy, sad, surprised, fear, disgust, and angry (from left to right).

5.1. Learned Conditional Space

We visualise the learned conditional space of our methods in Figure 1 and Figure 3. Since the initialization of parameters of the directions and the modes is random and SDG is used for training, the outcome differs between training runs. However, across different experiments, it has been noted that the final output tends to be very similar to the one reported in the two figures. For GANmut, about half of the conditional space is covered by the happy expressions (reflecting the unbalanced label distribution of the training set) and they are opposed to negative valence ones. Note that the emotion $surprise$ is in between. Similar observations can be done for the Gaussian variant of the model. One can draw a striking similarity between human emotion understanding and the conditional space learned by our methods.

5.2. Semantic Interpolation and Smoothness

In this section, we evaluate the ability of our method to express a basic emotion with different intensities. We make a comparison with StarGAN as we want to show that this virtue does not stem from the backbone of the method, rather from the proposed framework. We also compare GANmut trained with $\lambda_{\rho} = 4$ and $\lambda_{\rho} = 1$, to highlights the effect of $\mathcal{L}_{\rho}$, see Figure 4 . In this regard, we compute an interpolation in 10 steps between the $neutral$ expression and each basic emotion c_i. For our method, we simply start

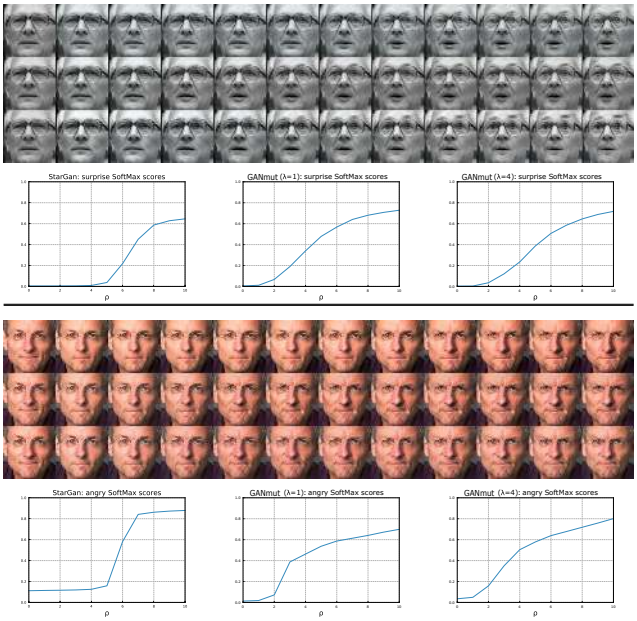

Figure 4: Two examples of our method compared to StarGAN. StarGAN is our natural baseline also because of being our backbone. Top to bottom rows corresponds to StarGAN, proposed $\lambda_\rho = 1$ and $\lambda_\rho = 4$, respectively. For the reference, the original image is provided in the first column. The plots reflect the corresponding scores (details provided in Table 1). Smoother transition are considered better. Interpolation is performed between neutral to extreme emotions (*upper*: neutral to surprised, *lower*: neutral to angry).

from the origin of $\mathcal{Z}$ and then move in direction θ_{c_i}, (*i.e.*, for each emotion c_i, we compute $x_{c_i,\rho_j} = G(x, z_{c_i,\rho_j})$ with $z_{c_i,\rho_j} = (\theta_{c_i}, \rho_j)$, $\rho_j = 0.1 \times j$, $j \in \{0, 1, ..., 10\}$). For StarGAN we use a linear interpolation between the one-hot vector encoding $neutral$ and c_i. To finally obtain a quantitative summary of the semantic smoothness over the entire dataset, we pass the series to the VGGNet classifier C. For each interpolation, we then compute the ratio between the maximum increase, in one interval, of the classification softmax score $C(c_i \mid x_{c_i,\rho_j})$ and its total variation range. The results averaged over the dataset are reported in Table 1. We have chosen this metric as we observed that StarGAN switches suddenly from one expression to the other, without expressing the emotion in a vague and mitigated way.

5.3. Complex Emotion Reconstruction

The Gaussian variant of GANmut (GGANmut) is especially suited to create complex and ambiguous expressions, since it allows to combine many emotions at once. To assess this property, we select a "complex emotion" target and see which method manages to reproduce it better. To obtain a quantitative results, as a proxy for human evaluation, we

	StarGAN	**GANmut** (ours)
Happy	0.79	**0.38**
Sad	0.44	**0.38**
Surprised	0.43	**0.35**
Fearful	0.49	**0.37**
Disgusted	0.55	**0.36**
Angry	0.59	**0.33**

Table 1: **Smoothness score:** maximum increase of the classification softmax score in one "strength" interval normalized by its total variation range (see Section 5.2). We measure the smoothness during interpolation to quantify the ability to generate emotions with different levels of intensities. Our method is consistently better then the baseline.

Methods	ERE		FED
	Setting 1	Setting 2	
Real	–	–	0.21
StarGAN [5]	0.038	0.037	4.89
GANimation [24]	0.036	0.037	2.17
SMIT [25]	0.028	0.029	1.89
GGANmut (ours)	**0.020**	**0.018**	18.32
GANmut (ours)	0.025	0.025	**0.71**

Table 2: Comparison with the state-of-the-art methods. The **FED** score for "Real" is obtained comparing real images from two halves of the test-set. ERE represents the ability to generate a given target emotions. For the experiment we choose them among particularly complex ones. More precisely, in Setting 1, we chose a target label distribution with at least 3 high values, while in Setting 2 with at least 2.

use the softmax score of the VGGNet classifier. During this process, the test-set is divided into batches of 16 images. For each batch, the images are passed to the VGG classifier, and the softmax scores with the highest top 2 or top 3 values is selected as label distribution target (these should correspond to the most ambiguous expressions). Then, for each method (StarGAN, SMIT, GANimation, GANmut, GGANmut), we search on a grid the best conditional code, in terms of mean squared error (MSE), that allows reproducing the target label distribution when passed to the VGG. For StarGAN, GANimation, and SMIT we search over the 7 categorical emotions. Regarding the multimodalities in SMIT, we sample 63 different representations for each emotion. For GANmut, we search on a grid with steps of 0.05 radii and $2\pi/21$ angles, whereas we use steps of 0.1 for GGANmut. The scores are averaged over each sample of the test-set, which we call **Emotion Reconstruction Error (ERE)**, and are reported in Table 2 together with the FED scores.

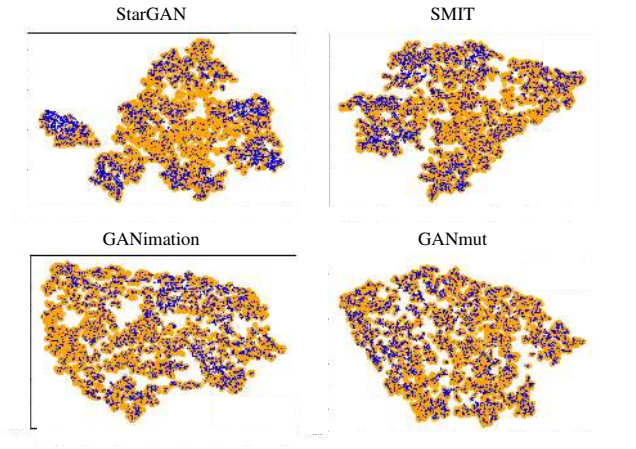

Figure 5: Low dimensional visualization (tSNE) of VGGNet features from real and generated images. Methods with FED scores: StarGAN (**4.89**); SMIT (**1.89**); GANimation (**2.17**); GANmut (**0.71**). The orange and blue points correspond to real and generated images, respectively.

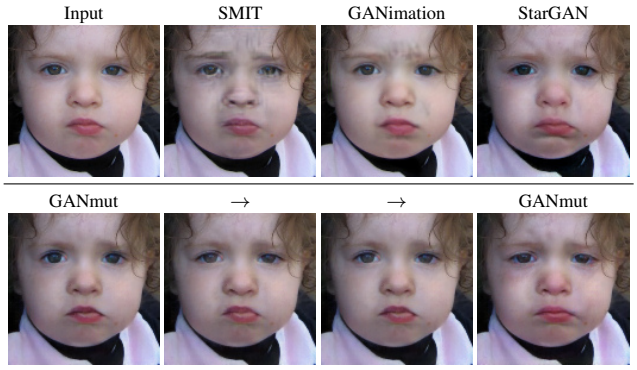

Figure 6: Visual results synthesized (*sad* conditioning) by different methods. Top left to right: Input image, SMIT, GANimation, StarGAN. At the bottom: Multiple images form the Gamut of emotions. More visual results are provided in the supplementary material.

The visualization in a 2D space by means of tSNE [35] of the feature distribution is provided in Figure 5. Some visual comparisons are also provided in Figure 6. Finally, in Table 3 we report the classification error with the VGGnet over the generated images ($z = (\theta_c, 1)$, $z = \mu_c$ were used for (G)GANmut). As it can be seen, real images present expressions much more ambiguous then generated ones by the compared methods (details in Suppl. Sec. 4).

5.4. Discussion

As we can see from Figure 3, GGANmut can generate plausible expression mixing up to 4 different emotions. As such, we can not really know what the man is feeling, could be Anger, Fear, Surprise or Disgust. Similarly, in Figure 7, the lady expresses her anger in different ways, with a touch of impatience, disgust, and surprise. We then reach a level of complexity, of expression ambiguity, that is closer to ones of human emotion. One reason we suspect why GGANmut is better in ERE but not in FED is its ability to generate extremely ambiguous (hence rare) emotions when sampled randomly, and too recognisable if sampled around the mode's centre. We consider this a virtue, not a vice (refer Section 5 in the supplementary for further discussion).

	Neutral	Happy	Sad	Surprised	Fearful	Disgusted	Angry
Real	76.4	93.4	57.8	46.0	46.4	30.4	57.2
StarGAN	92.6	96.1	96.9	95.4	94.7	95.1	95.8
GANmut	87.8	99.2	96.5	91.6	87.5	77.4	97.9
GGANmut	**95.8**	**99.6**	**99.9**	**99.1**	**97.6**	**97.8**	**99.5**
GANimation	78.4	96.4	76.0	76.0	59.7	52.4	79.6
SMIT	87.4	98.7	87.5	85.5	80.0	78.8	84.6

Table 3: Classification accuracy of generated images.

Figure 7: The emotion interpolated from the gamut of emotions using the proposed Gaussian model. Please observer the compoundness of the emotions, and note that they are obtained using only the basic emotion labels. More visual results are in the provided in the supplementary material.

6. Conclusion

We proposed a new problem of searching the label space employing only the images and the proxy labels to them. Two parameterizations for emotion modelling were explored, each with its own benefit according to our experiments. We also observed that the label space learned by our method already resembles the existing continuous model of VA. We strongly believe that this approach opens up the possibility of learning new label spaces for emotion modelling, and beyond. On the side of image synthesis, learning a better conditional space also allowed us to both control and generate diverse/complex emotions in a very spontaneous fashion, using only the basic emotion labels during training. Our work also shows that one can leverage the data to re-define the label, in cases where the labels' definition is not very reliable. We showed via several experiments that the problem of emotion modelling falls into such category, thereby allowing us to uncover a better conditional space for spontaneous emotion synthesis.

References

[1] Cvpr 2020 invited speaker: Lisa feldman barrett. can machines perceive emotions? https://www.youtube.com/watch?v=yds2ANV8PAE. Accessed: 2020-06-29.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] Rafael A Calvo and Sunghwan Mac Kim. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527–543, 2013.

[4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *NIPS*, pages 2172–2180, 2016.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[7] Carolyn Dever. The gamut of emotions from a to b: Nickleby's" histrionic expedition". *Dickens Studies Annual*, pages 1–16, 2008.

[8] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[9] P Ekman. Facial action codind system (facs). *A human face*, 2002.

[10] P Ekman and W Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 1971.

[11] W. V. Ekman, P.and Friesen. The facial actin coding system: a technique for the measurements of facial movements. 1978.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[13] James A Green, Pamela G Whitney, and Michael Potegal. Screaming, yelling, whining, and crying: Categorical and intensity differences in vocal expressions of anger and sadness in children's tantrums. *Emotion*, 11(5):1124, 2011.

[14] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering Interpretable GAN Controls. *arXiv e-prints*, page arXiv:2004.02546, Apr. 2020.

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, pages 6626–6637, 2017.

[16] Schneider K and Josephs I. the expressive and communicative functions of preschool children's smiles in an achievement-situation. *J. Nonverb. Behav.*, (15):185–198, 1991.

[17] T. Kanade, J. F. Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53, 2000.

[18] Robert E. Kraut and Robert E. Johnston. Social and emotional messages of smiling: An ethological approach. 1979.

[19] Samuli Laine. Feature-based metrics for exploring the latent space of generative models. In *ICLR (Workshop)*. OpenReview.net, 2018.

[20] Randy J Larsen and Edward Diener. Promises and problems with the circumplex model of emotion. 1992.

[21] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason M. Saragih, Zara Ambadar, and Iain A. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, pages 94–101. IEEE Computer Society, 2010.

[22] Ali Mollahosseini, Behzad Hassani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.*, 10(1):18–31, 2019.

[23] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651, 2017.

[24] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018.

[25] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. Smit: Stochastic multi-label image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[26] Andrés Romero, Luc Van Gool, and Radu Timofte. Smile: Semantically-guided multi-attribute image and layout editing. *arXiv preprint arXiv:2010.02315*, 2020.

[27] Barbara H Rosenwein. Problems and methods in the history of emotions. *Passions in context*, 1(1):1–32, 2010.

[28] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980.

[29] James Russell, Jo-Anne Bachorowski, and José-Miguel Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Review of Psychology*, 54:329–349, 11 2003.

[30] James Russell, Maria Lewicka, and Toomas Niit. A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, 57:848–856, 11 1989.

[31] James Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 09 1977.

[32] Klaus R Scherer et al. Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162, 2000.

[33] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, pages 9240–9249. IEEE, 2020.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.

[35] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[36] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space, 2020. cite arxiv:2002.03754.

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.