Continuous Emotion Transfer Using Kernels

Alex Lambert[*]
Télécom Paris, IP Paris & KU Leuven
alex.lambert@kuleuven.be

Sanjeel Parekh[*]
Télécom Paris, IP Paris
sanjeel.parekh@telecom-paris.fr

Zoltán Szabó
London School of Economics
z.szabo@lse.ac.uk

Florence d'Alché-Buc
Télécom Paris, IP Paris
florence.dalche@telecom-paris.fr

Abstract

Style transfer is a central problem of machine learning with numerous successful applications. In this work, we present a novel style transfer framework building upon infinite task learning and vector-valued reproducing kernel Hilbert spaces. We consider style transfer as a functional output regression task where the goal is to transform the input objects to a *continuum* of styles. The learnt mapping is governed by the choice of two kernels, one on the object space and one on the style space, providing flexibility to the approach. We instantiate the idea in emotion transfer where facial landmarks play the role of objects and styles correspond to emotions. The proposed approach provides a principled way to gain explicit control over the continuous style space, allowing to transform landmarks to emotions not seen during the training phase. We demonstrate the efficiency of the technique on popular facial emotion benchmarks, achieving low reconstruction cost.

1 Introduction

Recent years have witnessed an increasing attention around style transfer problems [21, 67, 27] in machine learning. In a nutshell, style transfer refers to the transformation of an object according to a target style. It has found numerous applications in computer vision [62, 11, 47, 68], natural language processing [20] as well as audio signal processing [25] where objects at hand are contents in which style is inherently part of their perception. In computer graphics, need for efficiently animating digital characters and avatars has led to several approaches for automatic style transfer on body motion capture sequences [3, 1]. Style transfer is one of the key components of data augmentation [43] as a means to artificially generate meaningful additional data for the training of deep neural networks. Besides, it has also been shown to be useful for counterbalancing bias in data by producing stylized contents with a well-chosen style (see for instance [22]) in image recognition. More broadly, style transfer fits into the wide paradigm of parametric modeling, where a system, a process or a signal can be controlled by its parameter value. Adopting this perspective, style transfer-like applications can also be found in digital twinning [60, 5, 35], a field of growing interest in health and industry.

This work introduces a novel principled framework for style transfer with the aim of transforming an input object to a *continuum* of styles. The framework is exemplified in the context of emotion transfer. Given a set of emotions, the general task of emotion transfer refers to transforming the representation of objects such as faces [11], hands [26], body movement [3] *etc.* according to these target emotions. Object representations typically include 2D images, 3D meshes, body skeletons, motion capture sequences. To highlight the relevance of our approach, we choose to instantiate it on emotion transfer for faces and to use facial landmark locations as the object representations and

[*] The two authors contributed equally.

1st CtrlGen: Controllable Generative Modeling in Language and Vision Workshop at NeurIPS 2021.

emotions as the style. Landmarks [28] have proven to be a useful, interpretable low-dimensional representation for capturing face expressions and dynamics in various applications such as facial identification [39], expression analysis [14] and medical diagnosis [4]. They provide objects of reasonable complexity (less than a hundred pairs of 2D coordinates for a single face) to demonstrate the strength of our approach. To our best knowledge, this is a novel task with no existing studies. Nevertheless, we briefly discuss below recent works from the related problem of emotion transfer on facial images and discuss some of their shortcomings.

Pioneering works in emotion transfer for faces include that of Blanz and Vetter [6] who proposed a morphable 3D face model whose parameters could be modified for facial attribute editing. Susskind et al. [58] designed a deep belief net for facial expression generation using action unit (AU) annotations. More recently, extensions of generative adversarial networks (GANs, [24]) have proven to be particularly powerful for tackling image-to-image translation problems [70]. Several works have addressed emotion transfer for facial images by conditioning GANs on a variety of guiding information ranging from discrete emotion labels to photos and videos. In particular, StarGAN [11] is conditioned on discrete expression labels for face synthesis. ExprGAN [15] proposes synthesis with the ability to control expression intensity through a controller module conditioned on discrete labels. TAAN [66] trains two GANs to generate more realistic face images and allow mixing and control of discrete emotion intensities. Other GAN-based approaches make use of additional information such as AU labels [46, 65], target landmarks [48], fiducial points [57] and photos/videos [23, 64].

While GANs have achieved high quality image synthesis, they come with some pitfalls: they are particularly difficult to train and require large amounts of training data. While prior works only synthesize outputs for a set of discrete labels or styles, we propose a framework to allow generation by continuously varying style space parameters. Illustrated on facial emotion transfer, this novel functional point of view translates into the following assumption: for a given person, the full range of the emotional faces is modelled as a continuous function from emotions to faces. This view exploits the geometry of the representation of emotions [50], assuming that one can pass a face “continuously" from one emotion to another. We then propose to address the problem of emotion transfer by learning a landmarks-to-function model able to predict for a given facial input image represented by its landmarks [61], the continuous function that maps an emotion to the landmarks transformed by this emotion.

This function-valued regression approach relies on a technique recently introduced by Brault et al. [8] called infinite task learning (ITL). ITL enlarges the scope of multi-task learning [18, 19] by learning to solve simultaneously a set of tasks parametrized by a continuous parameter. While strongly linked to other parametric learning methods such the one proposed by Takeuchi et al. [59], the approach differs from previous works by leveraging the use of operator-valued kernels and vector-valued reproducing kernel Hilbert spaces (vv-RKHS; [45, 42, 9]). vv-RKHSs have proven to be relevant in solving supervised learning tasks such as multiple quantile regression [51] or unsupervised problems like anomaly detection [55]. A common property of these works is that the output to be predicted is a real-valued function of a real parameter.

To solve the style transfer problem, we present an extension of ITL, vector ITL (or shortly vITL) which involves functional outputs with vectorial representation of the objects and the styles, showing that the approach is easily controllable by the choice of appropriate kernels guaranteeing continuity and smoothness. In particular, the functional point of view by the inherent regularization induced by the kernel makes the approach suitable even for limited and partial observation. We demonstrate the efficiency of the vITL approach in a series of numerical experiments in emotion transfer on two popular facial benchmarks.

The paper is structured as follows. In Section 2 we formulate the proposed general style transfer problem and exemplify it in the context of emotion transfer. The solution of the resulting vITL approach is elaborated in Section 3. Numerical experiments conducted on two benchmarks of the domain are presented in Section 4. Discussion and future work conclude the paper in Section 5. Proofs of auxiliary lemmas, additional details regarding implementation and experiments are collected in the appendices.

2 Problem Formulation

In this section we first present the general style transfer problem and then instantiate it in the context of emotion transfer of facial landmarks. Our aim is to design a system capable of transferring styles: having access to the representation of an object, our goal is to convert this object to a specified target style. In other words, the system should implement a mapping of the form

$$\text{(object, style)} \mapsto \text{object}. \tag{1}$$

In order to capture this relation, we propose to learn a function h that takes the following form

$$h :\mathcal{X} \times \Theta \mapsto \mathcal{X}, \text{ or equivalently } h : \mathcal{X} \mapsto (\Theta \mapsto \mathcal{X}), \tag{2}$$

where Θ is the style space and $\mathcal{X}$ is the object space. The learning takes place based on training samples $(x_{i,j}, \theta^{\text{out}}_{i,j}, y_{i,j})_{i\in[n],j\in S_i}$ where $x_{i,j} \in \mathcal{X}$ corresponds to an object with input style $\theta^{\text{in}}_{i,j} \in \Theta$, $y_{i,j} \in \mathcal{X}$ is the same object with output style $\theta^{\text{out}}_{i,j} \in \Theta$, and for each object $i \in [n]$ we have access to $|S_i|$ style transition pairs $\left\{\left(\theta^{\text{in}}_{i,j}, \theta^{\text{out}}_{i,j}\right)\right\}_{j\in S_i}$. To measure the quality of the reconstruction using a function h, one can consider a loss $\ell\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on the object space, and then formulate the task of style transfer as the minimization of the cost function

$$\mathcal{R}_{\mathcal{S}}(h) := \frac{1}{n}\sum_{i\in[n]}\frac{1}{|S_i|}\sum_{j\in S_i}\ell\Big(\overbrace{h\big(\underbrace{x_{i,j}}_{\substack{\text{input}\\ \text{object}}}\big)\big(\underbrace{\theta^{\text{out}}_{i,j}}_{\substack{\text{output}\\ \text{style}}}\big)}^{\text{predicted output object}}\ \underbrace{y_{i,j}}_{\substack{\text{output}\\ \text{object}}}\Big). \tag{3}$$

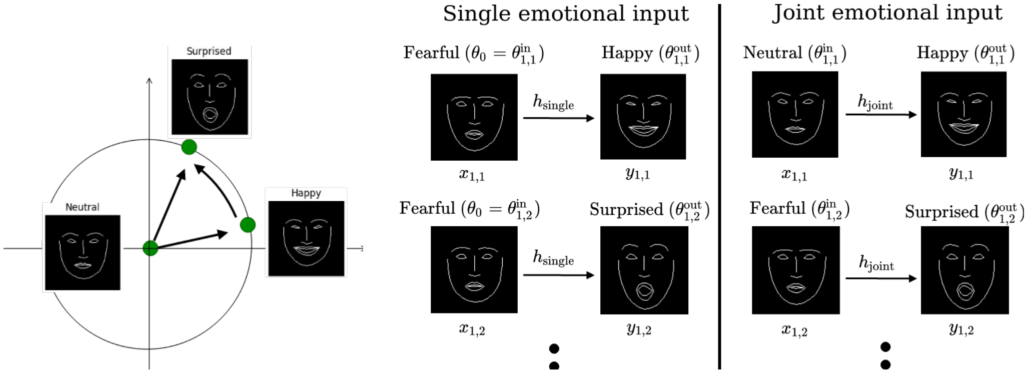
The risk $\mathcal{R}_{\mathcal{S}}(h)$ captures how well the function h reconstructs on average the output objects $y_{i,j}$ when applied to the input objects $x_{i,j}$ with target output style $\theta^{\text{out}}_{i,j}$. The minimization is performed in a hypothesis space $\mathcal{H}$ whose elements h model a relation from the input space $\mathcal{X}$ to a functional output space $\mathcal{F} : \Theta \to \mathcal{X}$ as it appears in the second part of (2). If the object space $\mathcal{X}$ is a Hilbert space, one can leverage vector-valued reproducing Hilbert spaces (vv-RKHSs, [45]) hypothesis classes to model $\mathcal{F}$ by using a suitable kernel on the style space Θ. Moreover, having access to a kernel on the object space $\mathcal{X}$ allows to define $\mathcal{H}$ as a vv-RKHS; this is the choice we make for $\mathcal{F}$ and $\mathcal{H}$ throughout the manuscript. The motivation of working with vv-RKHSs is three-fold: (i) vv-RKHSs provide rich function classes under mild conditions [10], (ii) they are capable of encoding output similarities in a principled way independently whether the outputs are finite-dimensional or not, (iii) despite their flexibility vv-RKHSs are computationally tractable. These are the points we detail in the sequel.

We now instantiate the style transfer problem in the specific case of emotion transfer for facial landmarks. In order to tackle this task, one requires a representation of the emotions, and similarly that of the faces. The classical categorical description of emotions deals with the classes 'happy', 'sad', 'angry', 'surprised', 'disgusted', 'fearful'. The valence-arousal model [50] embeds these categories into the 2-dimensional Euclidean space. The resulting representation of the emotions are points $\theta \in \mathbb{R}^2$, each coordinate of these vectors encoding the valence (pleasure to displeasure) and arousal (high to low) associated to the emotions. This is the emotion representation we use while noting that there are alternative encodings in higher dimension ($\Theta \subset \mathbb{R}^p$, $p \geq 2$; see for instance [63]) to which the presented framework can be naturally adapted. Throughout this work faces are represented by landmark points. Landmarks are specific locations pinpointed on the face, like the corner of the eyes, that of the mouth, and so on. They have proven to be a useful representation in facial recognition [69, 54, 53], 3D facial reconstruction [13] and sentiment analysis [56]. Tautkute et al. [61] have shown that emotions can be accurately recognized by detecting changes in the localization of the landmarks. Given M number of landmarks on the face, this means a description $x \in \mathcal{X} := \mathbb{R}^{2M}$; let the corresponding dimension be denoted by $d := 2M$.

The resulting mapping (1) is illustrated in Fig. 1a: starting from a neutral face and the target emotion happy one can traverse to the happy face; from the happy face, given the target emotion surprise one can get to the surprised face.

The observations $x_{i,j}, y_{i,j}$ in (3) can be interpreted as follows. Each person $i \in [n]$ is captured by a trajectory $z_i \in \mathcal{F}$, where $z_i(\theta) \in \mathcal{X}$ describes the landmarks associated to the emotion $\theta \in \Theta$. These trajectories z_i are observed at emotion transition pairs $\left\{\left(\theta^{\text{in}}_{i,j}, \theta^{\text{out}}_{i,j}\right)\right\}_{j\in S_i}$, giving rise to the landmark representation of these emotions:

$$x_{i,j} := z_i(\theta^{\text{in}}_{i,j}), \qquad\qquad y_{i,j} := z_i(\theta^{\text{out}}_{i,j}),\ \ i \in [n], j \in S_i. \tag{4}$$


Surprised
Neutral
Happy
Single emotional input
Joint emotional input
Fearful ($\theta_0 = \theta^{\text{in}}_{1,1}$)
Happy ($\theta^{\text{out}}_{1,1}$)
Neutral ($\theta^{\text{in}}_{1,1}$)
Happy ($\theta^{\text{out}}_{1,1}$)
h_{single}
h_{joint}
$x_{1,1}$
$y_{1,1}$
$x_{1,1}$
$y_{1,1}$
Fearful ($\theta_0 = \theta^{\text{in}}_{1,2}$)
Surprised ($\theta^{\text{out}}_{1,2}$)
Fearful ($\theta^{\text{in}}_{1,2}$)
Surprised ($\theta^{\text{out}}_{1,2}$)
h_{single}
h_{joint}
$x_{1,2}$
$y_{1,2}$
$x_{1,2}$
$y_{1,2}$

(a) Illustration of emotion transfer.

(b) Example triplets $(x_{i,j}, \theta^{\text{out}}_{i,j}, y_{i,j})$ for the single emotional (left) & joint emotional (right) input settings.

Figure 1: Visual illustration of (a) emotion transfer and (b) our tasks.

In this context, the minimizer of (3) allows to predict the landmarks from face x and target emotion θ as $h(x)(\theta)$. We focus on two specific cases of the problem family:

- **Single emotional input**: In this case, the input emotion is assumed to be identical and fixed for each person (θ_0) and the same m number of output emotions are considered. Hence the input-output emotion pairs used for learning take the form $\{(\theta_0, \theta_{i,j})\}_{j\in[m]}$, and $|S_i| = m$ for all $i \in [n]$.
- **Joint emotional input**: Here, for each individual m emotions are considered ($\{\theta_{i,a}\}_{a\in[m]}$), and the input-output emotion pairs in learning are taken in all possible combinations as $\{(\theta_{i,a}, \theta_{i,b})\}_{a,b\in[m]}$, and $|S_i| = m^2$ for all $i \in [n]$.

For a visual illustration of these tasks, see Fig. 1b. Throughout the manuscript we will use the squared loss $\ell(x, x') = \frac{1}{2}\|x - x'\|_2^2$.

3 Learning in vv-RKHSs

In this section, we propose a principled way to solve the task introduced in Section 2. As mentioned, we leverage the flexible class of vector-valued reproducing kernel Hilbert spaces (vv-RKHS; [10]) for both the functional output space $\mathcal{F}$ and the hypothesis class $\mathcal{H}$. The notion of vv-RKHS extends those of RKHS, allowing to generalize kernel methods to vector-valued regression tasks. Instead of relying on scalar-valued kernels, vv-RKHS are based on kernels whose values are operators on the output space. As emphasized by the seminal work of Micchelli and Pontil [42], multi-task learning is the most emblematic use of vv-RKHS when kernels are matrix-valued, allowing to cope with dependency among a finite number of tasks. Similarly, learning within vv-RKHS has been shown to be relevant for tackling function-valued regression [29, 31] when using kernels whose values are operators on output functional space. The organization of the section is as follows: in Section 3.1 we introduce formally vv-RKHSs and phrase the task as a regularized empirical risk minimization problem, before delving into the numerical solution in Section 3.2.

3.1 Operator-valued Kernels and vv-RKHSs

We provide a short overview of vv-RKHSs, instantiated to our problem. First we focus on the definition of $\mathcal{F}$, the emotion-to-landmark function space that we choose as the vv-RKHS $\mathcal{H}_G$ associated to a matrix-valued kernel G ($\mathcal{F} := \mathcal{H}_G$). Next, we define the hypothesis space $\mathcal{H}$ as the vv-RKHS $\mathcal{H}_K$ associated to an operator-valued kernel K ($\mathcal{H} := \mathcal{H}_K$). The construction follows the scheme:

$$h : \underbrace{\mathcal{X} \mapsto \underbrace{(\Theta \mapsto \mathcal{X})}_{\in \mathcal{F} := \mathcal{H}_G}}_{\in \mathcal{H} := \mathcal{H}_K} . \tag{5}$$

Definition 1. *A* matrix-valued kernel *on Θ is a function $G\colon \Theta \times \Theta \to \mathcal{L}(\mathbb{R}^d)$ such that the following two conditions hold: (i) $G(\theta, \theta') = G(\theta', \theta)^\top$ for all $(\theta, \theta') \in \Theta^2$, (ii) $\sum_{i,j \in [n]} v_i^\top G(\theta_i, \theta_j) v_j \geq 0$ for all $n \in \mathbb{N}^*$, $(\theta_i)_{i\in[n]} \in \Theta^n$, $(v_i)_{i\in[n]} \subset \mathbb{R}^d$, where $\mathbb{N}^* := \{1, 2, \ldots\}$ denotes the set of positive integers, $(\cdot)^\top$ stands for transposition, and $\mathcal{L}(\mathbb{R}^d)$ is the space of bounded linear operators on $\mathcal{X}$ (i.e. the set of $d \times d$-sized matrices).*

Such kernels can be associated to functional spaces whose elements model the $\Theta \to \mathcal{X}$ relation. In particular, any G satisfying Def. 1 gives rise to a unique Hilbert space of functions $\mathcal{H}_G$, so-called *vector-valued RKHS* associated to G: $\mathcal{H}_G = \overline{\operatorname{Span}}\left\{G(\cdot, \theta)x : (\theta, x) \in \Theta \times \mathbb{R}^d\right\}$, where $\operatorname{Span}(\cdot)$ denotes the linear hull of its argument and the closure is taken with respect to the scalar product induced by the positive quadratic form in Definition 1. While these objects can be fairly complex, a simple and popular choice of kernel is given by the so-called *separable* kernels, which can be written as

$$G(\theta, \theta') = k_\Theta(\theta, \theta')\mathbf{A}, \tag{6}$$

for a (scalar-valued) kernel $k_\Theta : \Theta \times \Theta \to \mathbb{R}$ and a symmetric, positive definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$.

The choice of kernel G governs the inner product in the Hilbert space $\mathcal{H}_G$, and thus the corresponding norm which is often a key component of regularization. More precisely, smoothness (analytical property) of an emotion-to-landmark output function $f \in \mathcal{H}_G$ can be induced for instance by choosing a Gaussian kernel $k_\Theta(\theta, \theta') = \exp\left(-\gamma \|\theta - \theta'\|_2^2\right)$ with $\gamma > 0$. The matrix $\mathbf{A}$ is known to capture inter-dependency between the output landmarks [2], and suitable choices of $\mathbf{A}$ can encode prior knowledge about this. In particular, choosing $\mathbf{A} = \mathbf{I}_d$ corresponds to independent landmarks coordinates. We use analogous tools for modeling the function $h : \mathcal{X} \to \mathcal{H}_G$ illustrated in (5).

Definition 2. *An* operator-valued kernel *on $\mathcal{X}$ is a function $K\colon \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H}_G)$ such that the following two conditions hold: (i) $K(x, x') = K(x', x)^*$ for all $(x, x') \in \mathcal{X}^2$, (ii) $\sum_{i,j\in[n]} \langle f_i, K(x_i, x_j) f_j \rangle_{\mathcal{H}_G} \geq 0$ for all $n \in \mathbb{N}^*$, $(x_i)_{i\in[n]} \in \mathcal{X}^n$, $(f_i)_{i\in[n]} \in \mathcal{H}_G^n$, where $(\cdot)^*$ means the adjoint operator.*

In this work, we choose K as a separable kernel with identity operator, defined as

$$K(x, x') = k_{\mathcal{X}}(x, x')\mathrm{Id}_{\mathcal{H}_G}, \tag{7}$$

where $\mathrm{Id}_{\mathcal{H}_G}$ is the identity operator on $\mathcal{H}_G$ and $k_{\mathcal{X}}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a (scalar-valued) kernel. Similarly to the matrix-valued case, the kernel K gives rise to a vv-RKHS $\mathcal{H}_K = \overline{\operatorname{Span}}\left\{K(\cdot, x)f : (x, f) \in \mathcal{X} \times \mathcal{H}_G\right\}$ used for modeling the $\mathcal{X} \to \mathcal{H}_G$ relation. The smoothness of functions $h \in \mathcal{H}_K$ can be driven by the choice of a Gaussian kernel over $\mathcal{X}$. The identity operator on $\mathcal{H}_G$ is the simplest choice to cope with operator-valued kernel, leading to computable scalar products on simple elements: $\langle K(\cdot, x)f_1, K(\cdot, x')f_2\rangle_{\mathcal{H}_K} = k_{\mathcal{X}}(x, x')\langle f_1, f_2\rangle_{\mathcal{H}_G}$ for all $(x, x') \in \mathcal{X}^2, (f_1, f_2) \in \mathcal{H}_G^2$.

One advantage of working with vv-RKHSs is to make use of a natural regularization given by the associated norm, denoted by $\|\cdot\|_{\mathcal{H}_K}$. Thus, we propose to solve the regularized empirical risk minimization problem

$$\min_{h\in\mathcal{H}_K} \mathcal{R}_\lambda(h) := \mathcal{R}_{\mathcal{S}}(h) + \frac{\lambda}{2} \|h\|_{\mathcal{H}_K}^2, \tag{8}$$

with a regularization parameter $\lambda > 0$ which balances between the data-fitting term ($\mathcal{R}_{\mathcal{S}}(h)$) and smoothness ($\|h\|_{\mathcal{H}_K}^2$). We refer to (8) as vector-valued infinite task learning (vITL).

Remark: This problem is a natural adaptation of the ITL framework [8] learning with operator-valued kernels mappings of the form $\mathcal{X} \mapsto (\Theta \mapsto \mathcal{Y})$ where $\mathcal{Y}$ is a subset of $\mathbb{R}$; here $\mathcal{Y} = \mathcal{X}$.

3.2 Optimization Task

This section is dedicated to the solution of (8) which is an optimization problem over functions ($h \in \mathcal{H}_K$). To get a unified solution for both the single and the joint emotional input task, we briefly introduce two dataset notations that allow to rephrase the empirical risk from (3) as

$$\mathcal{R}_{\mathcal{S}}(h) := \frac{1}{tm} \sum_{i\in[t]} \sum_{j\in[m]} \ell(h(x_i)(\theta_{i,j}), y_{i,j}). \tag{9}$$

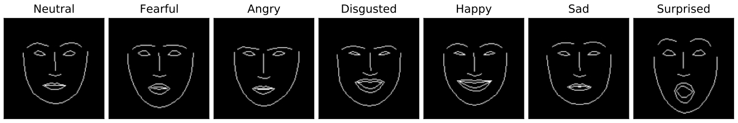


Figure 2: Illustration of the landmark edge maps for different emotions on KDEF

Intuitively, this corresponds to a reordering of the datasets so that the $(x_{i,j})_{i\in[n],j\in S_i}$ are only indexed as $(x_i)_{i\in[t]}$. The size t of the datasets captures the complexity of both tasks: we have $t := \frac{n|S_i|}{m}$, with $t = n$ for the single emotional input case and with $t = nm$ for the joint setting. The corresponding schemes allowing to get these datasets are summarized in Appendix B.

The following representer lemma provides a finite-dimensional parameterization of the optimal solution.

Lemma 1 (representer)**.** *Problem* (8) *has a unique solution* $\hat{h}$ *and it takes the form* $\hat{h}(x)(\theta) = \sum_{i=1}^{t}\sum_{j=1}^{m} k_{\mathcal{X}}(x, x_i)k_{\Theta}(\theta, \theta_{i,j})\mathbf{A}\hat{c}_{i,j}$*, for all* $\forall(x,\theta) \in \mathcal{X}\times\Theta$ *with some* $\{\hat{c}_{i,j}\}_{i\in[t],j\in[m]} \subset \mathbb{R}^d$.

Based on this lemma finding $\hat{h}$ is equivalent to determining the coefficients $\{\hat{c}_{i,j}\}_{i\in[t],j\in[m]}$. Recall that we consider the squared loss $\ell(x,x') = \frac{1}{2}\|x - x'\|_2^2$; in this case the task boils down to the solution of a linear equation as detailed in the following result.

Lemma 2 (optimization task for $\mathbf{C}$)**.** *Define the matrix* $\hat{\mathbf{C}} = [\hat{\mathbf{C}}_i]_{i\in[tm]} \in \mathbb{R}^{(tm)\times d}$ *containing all the coefficients, the Gram matrix* $\mathbf{K} = [k_{i,j}]_{i,j\in[tm]} \in \mathbb{R}^{(tm)\times(tm)}$*, and the matrix consisting of all the observations* $\mathbf{Y} = [\mathbf{Y}_i]_{i\in[tm]} \in \mathbb{R}^{(tm)\times d}$ *as* $\hat{\mathbf{C}}_{m(i-1)+j} := \hat{c}_{i,j}^{\top}$, $(i,j) \in [t]\times[m]$, $k_{m(i_1-1)+j_1,m(i_2-1)+j_2} := k_{\mathcal{X}}(x_{i_1}, x_{i_2})k_{\Theta}(\theta_{i_1,j_1}, \theta_{i_2,j_2})$, $(i_1,j_1),(i_2,j_2) \in [t]\times[m]$, $\mathbf{Y}_{m(i-1)+j} := y_{i,j}^{\top}$, $(i,j) \in [t]\times[m]$*. Assume moreover that* $\mathbf{K}$ *is invertible. Then* $\hat{\mathbf{C}}$ *is the solution of the following linear equation*

$$\mathbf{K}\hat{\mathbf{C}}\mathbf{A} + tm\lambda\hat{\mathbf{C}} = \mathbf{Y}. \tag{10}$$

When $\mathbf{A} = \mathbf{I}_d$ *(identity matrix of size* $d\times d$*), the solution is analytic:* $\hat{\mathbf{C}} = (\mathbf{K} + tm\lambda\mathbf{I}_{tm})^{-1}\mathbf{Y}$.

4 Numerical Experiments

In this section we demonstrate the efficiency of the proposed vITL-based style transfer framework in emotion transfer. We first introduce the two benchmark datasets used in our experiments and give details about data representation and choice of the hypothesis space in Section 4.1. In Section 4.2, we provide a quantitative performance assessment of the vITL approach (in mean squared error and classification accuracy sense) with a comparison to the state-of-the-art StarGAN method, modified to work on landmarks. These results are augmented with a qualitative analysis in Section 4.3. Additional properties of vITL are elaborated in Appendix C.4. The code used to run the experiments presented in this paper is available on github.

4.1 Experimental Setup

We used two popular face datasets for evaluation, namely Karolinska Directed Emotional Faces (KDEF; [37]) and Radboud Faces Database (RaFD; [34]). In our experiments, we used frontal images and seven emotions from each of these datasets. An edge map illustration of landmarks for different emotions is shown in Fig. 2; detailed description of the datasets is provided in Appendix C.1.

At this point, it is worth recalling that we are learning a function-valued function, $h : \mathcal{X} \mapsto (\Theta \mapsto \mathcal{X})$ using a vv-RKHS as our hypothesis class (see Section 3). We aligned the facial images and extracted 68 2D landmark points; this gave rise to the vectorized landmarks $\mathbf{x} \in \mathbb{R}^{136=2\times 68}$. The emotion labels were represented as points in the 2D valence-arousal (VA) space [50]. We took the kernels $k_{\mathcal{X}}$, k_{Θ} to be Gaussian on the landmark representation space and the emotion representation space, with respective bandwidth $\gamma_{\mathcal{X}}$ and γ_{Θ}. These kernels are known to induce smooth spaces of functions which aligns well with our goal. $\mathbf{A}$ was assumed to be $\mathbf{I}_d$ unless specified otherwise. Further implementation details are provided in Appendix C.2.

4.2 Quantitative Performance Assessment

Performance measures: We applied two metrics to quantify the performance of the compared systems, namely the test mean squared error (MSE) and emotion classification accuracy, which are commonly used for landmarks prediction [28]. The MSE provides a direct measure of the accuracy of the landmarks predictions and can be considered as the primary way to assess performance. The classification accuracy can be thought of as an indirect evaluation and reflects the ability of the functional model to produce the correct emotion on the predicted landmarks [14]. To compute this measure, for each dataset we trained a ResNet-18 classifier to recognize emotions from ground-truth landmark edge maps (as depicted in Fig. 2). The trained network was then used to compute classification accuracy over the predictions at test time. To rigorously evaluate outputs for each split of the data, we used a classifier trained on RaFD to evaluate KDEF predictions and vice-versa; this also allowed us to make the problem more challenging. The ResNet-18 network was appropriately modified to take grayscale images as input. During training, we used random horizontal flipping and cropping between 90-100% of the original image size to augment the data. All the images were finally resized to 224×224 and fed to the network. The network was trained from scratch using the stochastic gradient descent optimizer with learning rate and momentum set to 0.001 and 0.9, respectively. The training was carried out for 10 epochs with a batch size of 16.

We report the mean and standard deviation of the aforementioned metrics over ten 90%-10% train-test splits of the data. The test set for each split is constructed by removing 10% of the identities from the data. For each split, the best $\gamma_{\mathcal{X}}, \gamma_{\Theta}$ and λ values were determined by 6-fold and 10-fold cross-validation on KDEF and RaFD, respectively.

Baseline: We designed the baseline with two objectives in mind: (i) a generative adversarial network with state-of-the-art properties as imposed through the loss functions (adversarial, cycle consistency and domain classification loss), (ii) a model that takes exactly the same inputs as the proposed system. To this end, we used the popular StarGAN [11] system as our baseline with appropriate modifications to work with landmarks. Other GAN-based studies that operate on image representations use additional information and are not directly comparable to our setting. For fair comparison, the generator G and discriminator D were modified to be fully-connected networks that take vectorized landmarks as input. In particular, G was an encoder-decoder architecture where the target emotion, represented as a 2D emotion encoding as for our case, was appended at the bottleneck layer. It contained approximately one million parameters, which was chosen to be comparable with the number of coefficients in vITL ($839,664 = 126 \times 7 \times 7 \times 136$ for KDEF). ReLU activation function was used in all layers except before bottleneck in G and before penultimate layers of both G and D. We used their default parameter values in the code.[2] Experiments over each split of KDEF and RaFD were run for 50K and 25K iterations, respectively. Henceforth, we refer to this modified system as “Landmark-StarGAN".

MSE results: The test MSE for the compared systems is summarized in Table 1. As the table shows, the vITL technique outperforms Landmark-StarGAN on both datasets. One can observe low reconstruction cost for vITL in both the single and the joint emotional input case. Interestingly, a performance gain is obtained with vITL joint on the RaFD data in MSE sense. We hypothesize that this is due to the joint model benefiting from input landmarks for other emotions in the small data regime (only 67 samples per emotion for RaFD). Despite our best efforts, we found it quite difficult to train Landmark-StarGAN reliably. Its outputs had low diversity and were visibly more distorted.

Classification results: The emotion classification accuracies are available in Table 1. The classification results clearly demonstrate the improved performance and the higher quality of the generated emotion of vITL over Landmark-StarGAN; the latter also produces predictions with visible face distortions as it is illustrated in Section 4.3. Confusion matrices are presented in Appendix C.3 for further insight into the classification performance.

4.3 Qualitative Analysis

Here we show example outputs produced by vITL in the context of discrete and continuous emotion generation. While the former is the classical task of synthesis given input landmarks and target emotion label, the latter serves to demonstrate a key benefit of our approach, which is the ability to synthesize meaningful outputs while continuously traversing the emotion embedding space.

[2] The code is available at https://github.com/yunjey/stargan.

Methods	MSE Error ↓		Emotion Classification Acc. ↑	
	KDEF frontal	RaFD frontal	KDEF frontal	RaFD frontal
vITL: θ_0 = neutral	0.010 ± 0.001	0.009 ± 0.004	76.12 ± 4.57	79.76 ± 7.88
vITL: θ_0 = fearful	0.010 ± 0.001	0.010 ± 0.005	76.22 ± 4.91	78.81 ± 8.36
vITL: θ_0 = angry	0.012 ± 0.002	0.010 ± 0.005	74.49 ± 2.31	78.10 ± 7.51
vITL: θ_0 = disgusted	0.012 ± 0.001	0.010 ± 0.004	74.18 ± 4.22	78.33 ± 4.12
vITL: θ_0 = happy	0.011 ± 0.001	0.010 ± 0.004	73.57 ± 2.74	80.48 ± 5.70
vITL: θ_0 = sad	0.011 ± 0.001	0.009 ± 0.004	75.82 ± 4.11	77.62 ± 5.17
vITL: θ_0 = surprised	0.010 ± 0.001	0.011 ± 0.006	74.69 ± 2.25	80.71 ± 5.99
vITL: Joint	$\mathbf{0.011} \pm 0.001$	$\mathbf{0.007} \pm 0.001$	$\mathbf{74.81} \pm 3.10$	$\mathbf{77.11} \pm 3.97$
Landmark-StarGAN	0.029 ± 0.003	0.024 ± 0.007	70.69 ± 8.46	65.88 ± 8.92

Table 1: MSE error and emotion classification accuracy (mean ± std) on test data for the vITL single (top), the vITL joint and the Landmark-StarGAN system (bottom). The best results are in bold when relevant.

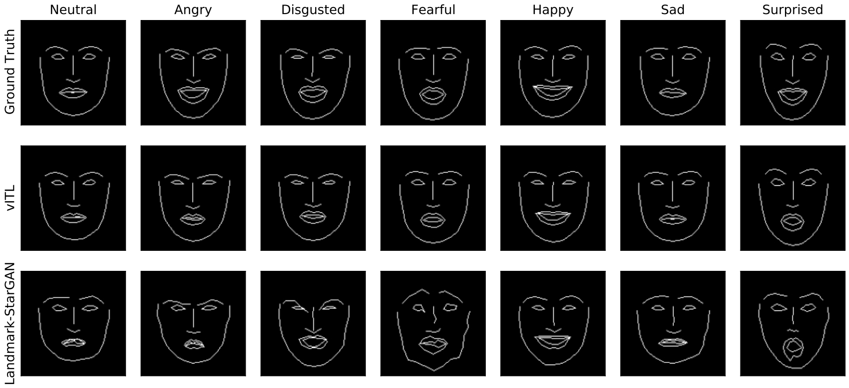


Figure 3: Discrete expression synthesis results on the KDEF dataset with ground-truth neutral landmarks as input.

Discrete emotion generation: In Fig. 3 we show qualitative results for generating landmarks using discrete emotion labels present in the datasets. For vITL, not only are the emotions recognizable, but landmarks on the face boundary are reasonably well synthesized and other parts of the face visibly less distorted when compared to Landmark-StarGAN. The identity in terms of the face shape is also better preserved.

Continuous emotion generation: We show in Fig. 4 the capability to generate intermediate emotions by changing the angular position in the emotion embedding space, in this case from 'happy' to 'surprised'. For a more fine-grained video illustration traversing from 'happy' to 'sad' along the circle, see the supplements. Additional qualitative results are presented in Appendix C.5.

5 Conclusion

We introduced a novel general framework for style transfer based on function-valued regression, and exemplified it on the problem of emotion transfer for facial landmarks. The proposed framework is quite general and can be applied as long as one is able to define kernels on the object and the

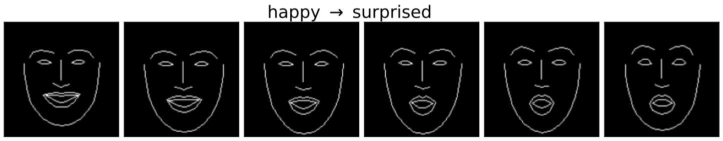


Figure 4: Continuous expression synthesis with vITL technique on the RaFD dataset, with ground-truth neutral landmarks. The generation is starting from 'happy' and proceeds by changing angular position towards 'surprised'. For a more fine-grained video illustration traversing from 'happy' to 'sad' along the circle, see the demo in supplements.

style spaces, allowing for a wide range of applications. The vector-valued infinite task learning (vITL) formulation relies on operator-valued kernels. vITL (i) is capable of encoding and controlling continuous style spaces, (ii) benefit from a representer theorem for efficient computation, and (iii) facilitates regularity control via the choice of the underlying kernels. The framework can be extended in several directions. The most natural one would be to apply this framework to more complex object and style spaces, for instance time series for motion transfer [1]. In this case, the choice of kernel would be critical and could be tackled by learning deep kernel architectures [41, 36] instead of designing the kernel prior to learning. Finally, other losses [52, 33] can be leveraged to produce outlier-robust or sparse models.

Acknowledgements A.L. and S.P. were funded by the research chair Data Science & Artificial Intelligence for Digitalized Industry and Services at Télécom Paris. ZSz benefited from the support of the Europlace Institute of Finance and that of the Chair Stress Test, RISK Management and Financial Steering, led by the French École Polytechnique and its Foundation and sponsored by BNP Paribas.

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4): 64, 2020. 1, 9

[2] M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012. 5

[3] Andreas Aristidou, Qiong Zeng, Efstathios Stavrakis, KangKang Yin, Daniel Cohen-Or, Yiorgos Chrysanthou, and Baoquan Chen. Emotion control of unstructured dance movements. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 1–10, 2017. 1

[4] Asghar Tabatabaei Balaei, Kate Sutherland, Peter A. Cistulli, and Philip de Chazal. Automatic detection of obstructive sleep apnea using facial images. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 215–218, 2017. 2

[5] Barbara Rita Barricelli, Elena Casiraghi, and Daniela Fogli. A survey on digital twin: Definitions, characteristics, applications, and design implications. *IEEE Access*, 7:167653–167671, 2019. 1

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 187–194, 1999. 2

[7] Karsten Borgwardt, Elisabetta Ghisu, Felipe Llinares-López, Leslie O'Bray, and Bastian Rieck. Graph kernels: State-of-the-art and future challenges. *Foundations and Trends in Machine Learning*, 13(5-6):531–712, 2020. 14

[8] Romain Brault, Alex Lambert, Zoltán Szabó, Maxime Sangnier, and Florence d'Alché-Buc. Infinite task learning in RKHSs. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1294–1302, 2019. 2, 5

[9] Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4: 377–408, 2006. 2

[10] Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(1):19–61, 2010. 3, 4

[11] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018. 1, 2, 7

[12] Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui. A kernel for time series based on global alignments. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 413–416, 2007. 14

[13] Michel Desvignes, Gerard Bailly, Yohan Payan, and Maxime Berar. 3D semi-landmarks based statistical face reconstruction. *Journal of computing and Information technology*, 14(1):31–43, 2006. 3

[14] Terrance Devries, Kumar Biswaranjan, and Graham W. Taylor. Multi-task learning of facial landmarks and expression. In *Canadian Conference on Computer and Robot Vision*, pages 98–103, 2014. 2, 7

[15] Hui Ding, Kumar Sricharan, and Rama Chellappa. ExprGAN: Facial expression editing with controllable expression intensity. In *Conference on Artificial Intelligence (AAAI)*, pages 6781–6788, 2018. 2

[16] Paul Ekman, Wallace Friesen, and Joseph Hager. Facial action coding system: The manual. *Salt LakeCity, UT: Research Nexus.*, 2002. 15

[17] A El Guennouni, Khalide Jbilou, and AJ Riquet. Block Krylov subspace methods for solving large Sylvester equations. *Numerical Algorithms*, 29(1):75–96, 2002. 15

[18] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 109–117, 2004. 2

[19] Theodoros Evgeniou, Charles Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005. 2

[20] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In *Conference on Artificial Intelligence (AAAI)*, pages 663–670, 2018. 1

[21] Justin J. Gatys, Alexandre A., and F.-F Li. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711, 2016. 1

[22] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. 1

[23] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided GANs for single-photo facial animation. *ACM Transactions on Graphics*, 37(6):1–12, 2018. 2

[24] Ian J. 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 (NIPS)*, pages 2672–2680, 2014. 2

[25] Eric Grinstein, Ngoc QK Duong, Alexey Ozerov, and Patrick Pérez. Audio style transfer. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 586–590, 2018. 1

[26] Ana-Sabina Irimia, Jacky CP Chan, Kamlesh Mistry, Wei Wei, and Edmond SL Ho. Emotion transfer for hand animation. In *Motion, Interaction and Games*, pages 1–2. 2019. 1

[27] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics*, 26(11): 3365–3385, 2020. 1

[28] Benjamin Johnston and Philip de Chazal. A review of image-based automatic facial landmark identification techniques. *EURASIP Journal on Image and Video Processing*, 2018(1):1–23, 2018. 2, 7

[29] Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, and Manuel Davy. Nonlinear functional regression: a functional RKHS approach. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 374–380, 2010. 4

[30] Hachem Kadri, Mohammad Ghavamzadeh, and Philippe Preux. A generalized kernel approach to structured output learning. In *International Conference on Machine Learning (ICML)*, pages 471–479, 2013. 16

[31] Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20):1–54, 2016. 4

[32] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874, 2014. 16

[33] Pierre Laforgue, Alex Lambert, Luc Brogat-Motte, and Florence d'Alché Buc. Duality in RKHSs with infinite dimensional outputs: Application to robust losses. In *International Conference on Machine Learning (ICML)*, pages 5598–5607, 2020. 9

[34] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. Presentation and validation of the Radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010. 6, 15

[35] Kendrik Yan Hong Lim, Pai Zheng, and Chun-Hsien Che. A state-of-the-art survey of digital twin: techniques, engineering product lifecycle management and business innovation perspectives. *Journal of Intelligent Manufacturing*, 31:1313–1337, 2020. 1

[36] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning (ICML)*, pages 6316–6326, 2020. 9

[37] Daniel Lundqvist, Anders Flykt, and Arne Öhman. The Karolinska directed emotional faces (KDEF). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91(630):2–2, 1998. 6, 15

[38] Pierre Mahé and Jean-Philippe Vert. Graph kernels based on tree patterns for molecules. *Machine learning*, 75(1):3–35, 2009. 14

[39] Shraddha Mane and Gauri Shah. Facial recognition, expression recognition, and gender identification. In *Data Management, Analytics and Innovation*, pages 275–290. Springer, 2019. 2

[40] Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. Kernel methods through the roof: handling billions of points efficiently. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 15

[41] Siamak Mehrkanoon and Johan A. K. Suykens. Deep hybrid neural-kernel networks using random Fourier features. *Neurocomputing*, 298:46–54, 2018. 9

[42] Charles Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005. 2, 4

[43] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *International Interdisciplinary PhD Workshop (IIPhDW)*, pages 117–122, 2018. 1

[44] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 16

[45] George Pedrick. *Theory of reproducing kernels for Hilbert spaces of vector valued functions*. PhD thesis, 1957. 2, 3

[46] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. GANimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision (ECCV)*, pages 818–833, 2018. 2

[47] Gilles Puy and Patrick Pérez. A flexible convolutional solver for fast style transfers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8963–8972, 2019. 1

[48] Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. Geometry-contrastive GAN for facial expression transfer. Technical report, 2018. (https://arxiv.org/abs/1802.01822). 2

[49] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. FALKON: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3891–3901, 2017. 15

[50] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. 2, 3, 6, 16

[51] Maxime Sangnier, Olivier Fercoq, and Florence d'Alché Buc. Joint quantile regression in vector-valued RKHSs. *Advances in Neural Information Processing Systems (NIPS)*, pages 3693–3701, 2016. 2

[52] Maxime Sangnier, Olivier Fercoq, and Florence d'Alché-Buc. Data sparse nonparametric regression with ϵ-insensitive losses. In *Asian Conference on Machine Learning (ACML)*, pages 192–207, 2017. 9

[53] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Face alignment through subspace constrained mean-shifts. In *International Conference on Computer Vision (ICCV)*, pages 1034–1041, 2009. 3

[54] Ulrich Scherhag, Dhanesh Budhrani, Marta Gomez-Barrero, and Christoph Busch. Detecting morphed face images using facial landmarks. In *International Conference on Image and Signal Processing (ICISP)*, pages 444–452, 2018. 3

[55] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J. Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001. 2

[56] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017. 3

[57] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *International Conference on Multimedia (MM)*, pages 627–635, 2018. 2

[58] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. Generating facial expressions with deep belief nets. In *Affective Computing*, chapter 23. IntechOpen, 2008. 2

[59] Ichiro Takeuchi, Quoc Le, Timothy Sears, and Alexander Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006. 2

[60] Fei Tao, He Zhang, Ang Liu, and A. Y. C. Nee. Digital twin in industry: State-of-the-art. *IEEE Transactions on Industrial Informatics*, 15(4):2405 – 2415, 2019. 1

[61] Ivona Tautkute, T. Trzciński, and Adam Bielski. I know how you feel: Emotion recognition with facial landmarks. *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1959–19592, 2018. 2, 3

[62] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *International Conference on Machine Learning (ICML)*, pages 1349–1357, 2016. 1

[63] Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5683–5692, 2019. 3

[64] Rongliang Wu and Shijian Lu. LEED: Label-free expression editing via disentanglement. In *European Conference on Computer Vision (ECCV)*, pages 781–798, 2020. 2

[65] Rongliang Wu, Gongjie Zhang, Shijian Lu, and Tao Chen. Cascade EF-GAN: Progressive facial expression editing with local focuses. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5021–5030, 2020. 2

[66] Xuehui Wu, Jie Shao, Dongyang Zhang, and Junming Chen. Unsupervised facial image synthesis using two-discriminator adversarial autoencoder network. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1162–1167, 2019. 2

[67] Daan Wynen, Cordelia Schmid, and Julien Mairal. Unsupervised learning of artistic styles with archetypal style analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6584–6593, 2018. 1

[68] Xu Yao, Gilles Puy, Alasdair Newson, Yann Gousseau, and Pierre Hellier. High resolution face age editing. In *International Conference on Pattern Recognition (ICPR)*, 2020. 1

[69] Zheng Zhang, Long Wang, Qi Zhu, Shu-Kai Chen, and Yan Chen. Pose-invariant face recognition using facial landmarks and Weber local descriptor. *Knowledge-Based Systems*, 84:78–88, 2015. 3

[70] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 2

Appendices

We provide proofs of auxiliary lemmas in Appendix A. Appendix B presents additional remarks on modeling and optimization in vvRKHSs. Implementation details, additional remarks and numerical illustrations are shown in Appendix C.

A Proofs

This section contains the proofs of our auxiliary lemmas.

Proof. (Lemma 1) For all $g \in \mathcal{H}_G$, let $K_x g$ denote the function defined by $(K_x g)(t) = K(t,x)g$ $\forall t \in \mathcal{X}$. Similarly, for all $c \in \mathcal{X}$, $G_\theta c$ stands for the function $t \mapsto G(t,\theta)c$ where $t \in \Theta$. Let us take the finite-dimensional subspace

$$E = \operatorname{span}\left(K_{x_i} G_{\theta_{ij}} c \,:\, i \in [t], j \in [m], c \in \mathbb{R}^d\right).$$

The space $\mathcal{H}_K$ can be decomposed as E and its orthogonal complement: $E \oplus E^\perp = \mathcal{H}_K$. The existence of $\hat{h}$ follows from the coercivity of $\mathcal{R}_\lambda$ (i.e. $\mathcal{R}_\lambda(h) \to +\infty$ as $\|h\|_{\mathcal{H}_K} \to +\infty$) which is the consequence of the quadratic regularizer and the lower boundedness of ℓ. Uniqueness comes from the strong convexity of the objective. Let us decompose $\hat{h} = \hat{h}_E + \hat{h}_{E^\perp}$, and take any $c \in \mathbb{R}^d$. Then for all $(i,j) \in [t] \times [m]$, one has

$$\left\langle \hat{h}_{E^\perp}(x_i)(\theta_{ij}), c \right\rangle_{\mathbb{R}^d} \overset{(a)}{=} \left\langle \hat{h}_{E^\perp}(x_i), G_{\theta_{ij}} c \right\rangle_{\mathcal{H}_G} \overset{(b)}{=} \langle \hat{h}_{E^\perp}, \underbrace{K_{x_i} G_{\theta_{ij}} c}_{\in E} \rangle_{\mathcal{H}_K} \overset{(c)}{=} 0.$$

(a) follows from the reproducing property in $\mathcal{H}_G$, (b) is a consequence of the reproducing property in $\mathcal{H}_K$, and (c) comes from the decomposition $E \oplus E^\perp = \mathcal{H}_K$. This means that $\hat{h}_{E^\top}(x_i)(\theta_{ij}) = 0$ $\forall (i,j) \in [t] \times [m]$, and hence $\mathcal{R}_{\mathcal{S}}(\hat{h}) = \mathcal{R}_{\mathcal{S}}(\hat{h}_E)$. Since $\lambda \|\hat{h}\|^2_{\mathcal{H}_K} = \lambda \left(\|\hat{h}_E\|^2_{\mathcal{H}_K} + \|\hat{h}_{E^\perp}\|^2_{\mathcal{H}_K} \right) \geq \lambda \|\hat{h}_E\|^2_{\mathcal{H}_K}$ we conclude that $\hat{h}_{E^\top} = 0$ and get that there exist coefficients $\hat{c}_{i,j} \in \mathbb{R}^d$ such that $\hat{h} = \sum_{i \in [t]} \sum_{j \in [m]} K_{x_i} G_{\theta_{i,j}} \hat{c}_{i,j}$. This evaluates for all $(x,\theta) \in \mathcal{X} \times \Theta$ to $\hat{h}(x)(\theta) = \sum_{i=1}^t \sum_{j=1}^m k_{\mathcal{X}}(x,x_i) k_\Theta(\theta,\theta_{i,j}) \mathbf{A} \hat{c}_{ij}$ as claimed in Lemma 1. □

Proof. (Lemma 2) Applying Lemma 1, problem (8) writes as

$$\min_{\mathbf{C} \in \mathbb{R}^{(tm) \times d}} \frac{1}{2tm} \|\mathbf{KCA} - \mathbf{Y}\|^2_{\mathrm{F}} + \frac{\lambda}{2} \operatorname{Tr}\left(\mathbf{KCAC}^\top\right),$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. By setting the gradient of this convex functional to zero, and using the symmetry of $\mathbf{K}$ and $\mathbf{A}$, one gets

$$\frac{1}{tm} \mathbf{K}(\mathbf{KCA} - \mathbf{Y})\mathbf{A} + \lambda \mathbf{KCA} = \mathbf{0}$$

which implies (10) by the invertibility of $\mathbf{K}$ and $\mathbf{A}$. □

B Remarks on the Problem Formulation and Optimization

Schemes to construct datasets for single and joint emotional input tasks are presented in Table 2. Following are some additional remarks on modeling and optimization in vv-RKHSs:

- The considered squared loss leads to closed-form solution; see Lemma 1 and Lemma 2. Alternative loss functions, for instance to encode robustness, would similarly be applicable at the price of using iterative optimization schemes.
- In (1), the space $\mathcal{X}$ plays the role of both the input and the output representation. One can imagine scenarii where the output space (denoted by $\mathcal{Y}$) is different from the input space, leading to a model

  $$h : \mathcal{X} \mapsto (\Theta \mapsto \mathcal{Y}). \tag{11}$$

  In this case, the developed methodology would still be applicable provided that $\mathcal{Y}$ is a Hilbert space, and $\mathcal{X}$ could be any representation set over which one can design a kernel, for instance time series [12], or graphs [38, 7].

Table 2: Dataset generation corresponding to the single and joint emotional models.

```
Dataset : Single Emotional Input
init    : $t := n$
for $i \in [t]$ do
    $x_i := z_i(\theta_0)$
    for $j \in [m]$ do
        $\theta_{i,j} := \theta_{i,j}^{\text{out}}$
        $y_{i,j} := z_i(\theta_{i,j}^{\text{out}})$
return $\left(x_i, (y_{i,j})_{j=1}^m\right)_{i=1}^t$, $(\theta_{i,j})_{i,j=1}^{t,m}$
```

```
Dataset : Joint Emotional Input
init    : $t := nm$
for $i \in [t]$ do
    Define $(l,k)$ such that
      $i = (l-1)m + (k-1)$ (Euclidean
      division of $i$ by $m$)
    $x_i := z_l(\theta_{l,k}^{\text{out}})$
    for $j \in [m]$ do
        $\theta_{i,j} := \theta_{l,j}^{\text{out}}$
        $y_{i,j} := z_l(\theta_{l,j}^{\text{out}})$
return $\left(x_i, (y_{i,j})_{j=1}^m\right)_{i=1}^t$, $(\theta_{i,j})_{i,j=1}^{t,m}$
```

- Computational complexity: In case of $\mathbf{A} = \mathbf{I}_d$, the complexity of the closed-form solution is $\mathcal{O}\left((tm)^3\right)$. If all the samples are observed at the same locations $(\theta_{i,j})_{i,j\in[t]\times[n]}$, i.e. $\theta_{i,j} = \theta_{l,j}$ for $\forall(i,l,j) \in [t] \times [t] \times [m]$, then the Gram matrix $\mathbf{K}$ has a tensorial structure $\mathbf{K} = \mathbf{K}_{\mathcal{X}} \otimes \mathbf{K}_{\Theta}$ with $\mathbf{K}_{\mathcal{X}} = [k_{\mathcal{X}}(x_i, x_j)]_{i,j\in[t]} \in \mathbb{R}^{t\times t}$ and $\mathbf{K}_{\Theta} = [k_{\Theta}(\theta_{1,i}, \theta_{1,j})]_{i,j\in[m]} \in \mathbb{R}^{m\times m}$. In this case, the computational complexity reduces to $\mathcal{O}\left(t^3 + m^3\right)$. If additional scaling is required one can leverage recent dedicated kernel ridge regression solvers [49, 40]. If $\mathbf{A}$ is not identity, then multiplying (10) with $\mathbf{A}^{-1}$ gives $\mathbf{K}\hat{\mathbf{C}} + tm\lambda\hat{\mathbf{C}}\mathbf{A}^{-1} = \mathbf{Y}\mathbf{A}^{-1}$ which is a Sylvester equation for which efficient custom solvers exist [17].
- Regularization in vv-RKHS: Using the notations above, for any $h \in \mathcal{H}_K$ parameterized by a matrix $\mathbf{C}$, it holds that $\|h\|_{\mathcal{H}_K}^2 = \operatorname{Tr}\left(\mathbf{KCAC}^\top\right)$. Given two matrices $\mathbf{A}_1, \mathbf{A}_2$ and associated vv-RKHSs $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$, if $\mathbf{A}_1$ and $\mathbf{A}_2$ are invertible then any function in $\mathcal{H}_{K_1}$ parameterized by $\mathbf{C}$ also belongs to $\mathcal{H}_{K_2}$ (and vice versa), within which it is parameterized by $\mathbf{CA}_2^{-1}\mathbf{A}_1$. This means that the two spaces contain the same functions, but their norms are different.

C Experiment Details and Additional Results

This section is arranged as follows:

- Section C.1 gives details about the two benchmark datasets KDEF and RaFD considered in our experiments.
- Section C.2 provides exact implementation details for the applied landmark and emotion representation pre-processing steps.
- Section C.3 shows the vITL joint model confusion matrix for emotion classification results.
- Section C.4 presents quantitative analysis of properties of vITL pertaining to the choice of $\mathbf{A}$ (in kernel G) and the robustness w.r.t. partial observation.
- Section C.5 illustrates additional results for discrete and continuous generation capabilities of vITL. In particular, discrete generation results on RaFD and continuous generation in the radial direction are presented.

C.1 Details on the Used Benchmarks

We used the following two popular face datasets for evaluation.

- Karolinska Directed Emotional Faces (KDEF; [37]): This dataset contains facial emotion pictures from 70 actors (35 females and 35 males) recorded over two sessions which give rise to a total of 140 samples per emotion. In addition to neutral, the captured facial emotions include afraid, angry, disgusted, happy, sad and surprised.
- Radboud Faces Database (RaFD; [34]): This benchmark contains emotional pictures of 67 unique identities (including Caucasian males and females, Caucasian children, and Moroccan Dutch males). Each subject was trained to show the following expressions: anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral according to the facial action coding system (FACS; [16]).

C.2 Implementation Details

This section is dedicated to implementation details.

Landmark representation, pre-processing: We applied the following pre-processing steps to get landmark representations which form the input of the algorithms. To extract 68 landmark points for all facial images, we used the standard `dlib` library. The estimator is based on `dlib`'s implementation of [32], trained on the iBUG 300-W face landmark dataset. Each landmark is represented by its 2D location. The alignment of the faces was carried out by the Python library `imutils`. The method ensures that faces across all identities and emotions are vertical, centered and of similar sizes. In essence, this is implemented through an affine transformation computed after drawing a line segment between the estimated eye centers. Each image was resized to the size 128×128. The landmark points computed in the step above were transformed through the same affine transformation. These two preprocessing steps gave rise to the aligned, scaled and vectorized landmarks $\mathbf{x} \in \mathbb{R}^{136=2\times 68}$.

Emotion representation: We represented emotion labels as points in the 2D valence-arousal (VA) space [50]. Particularly, we used a manually annotated part of the large-scale AffectNet database [44]. For all samples of a particular emotion in the AffectNet data, we computed the centroid (data mean) of the valence and arousal values. The resulting ℓ_2-normalized 2D vectors constituted our emotion representation (see Fig. 5). The normalization is akin to assuming that the modeled emotions are of the same intensity. In our experiments, the emotion 'neutral' was represented by the origin. Such an emotion embedding allowed us to take into account prior knowledge about the angular proximity of emotions in the VA space, while keeping the representation simple and interpretable for post-hoc manipulations.

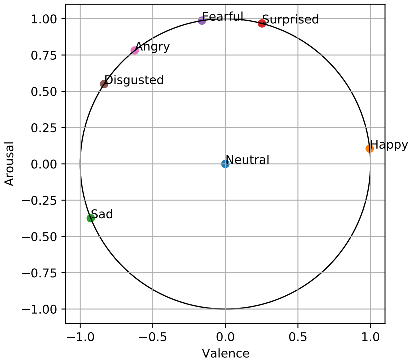


Figure 5: Extracted ℓ_2-normalized valence-arousal centroids for each emotion from the manually annotated train set of the AffectNet database.

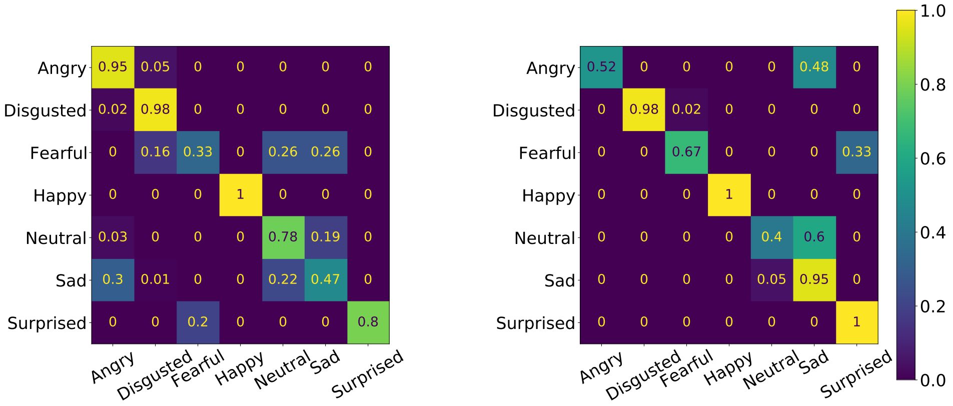
C.3 Quantitative Results: Confusion Matrix for Emotion Classification

To provide further insight into the classification performance we also show the confusion matrices for the joint vITL model on a particular split of KDEF and RaFD datasets in Fig. 6. For both the datasets, the classes 'happy' and 'surprised' are easiest to detect. Some confusions arise between the classes 'neutral' vs 'sad' and 'fearful' vs 'surprised'.

C.4 Additional Properties of vITL

This section is dedicated to the effect of the choice of $\mathbf{A}$ (in kernel G) and to the robustness of vITL w.r.t. partial observation.

Influence of A in the matrix-valued kernel G**:** Here, we illustrate the effect of matrix $\mathbf{A}$ (see (6)) on the vITL estimator and show that a good choice of $\mathbf{A}$ can lead to lower dimensional models, while preserving the quality of the prediction. The choice of $\mathbf{A}$ is built on the knowledge that the empirical covariance matrices of the output training data contains structural information that can be exploited with vv-RKHS [30]. In order to investigate this possibility, we performed the singular value decomposition of $\mathbf{Y}^\top\mathbf{Y}$ which gives the eigenvectors collected in matrix $\mathbf{V} \in \mathbb{R}^{d\times d}$. For a



(a) KDEF

(b) RaFD

Figure 6: Confusion matrices for classification accuracy of vITL joint model. Left: dataset KDEF. Right: dataset RaFD. The y axis represents the true labels, the x axis stands for the predicted labels. More diagonal is better.

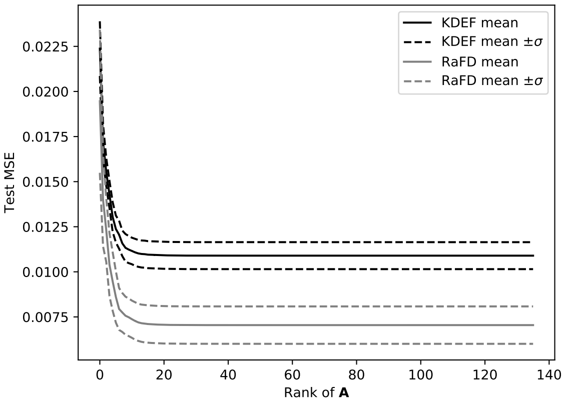


Figure 7: Test MSE (mean ± std) as a function of the rank of the matrix $\mathbf{A}$. Smaller MSE is better.

fixed rank $r \leq d$, define $\mathbf{J}_r = \operatorname{diag}(\underbrace{1,\cdots,1}_{r},\underbrace{0,\cdots,0}_{d-r})$, set $\mathbf{A} = \mathbf{V}\,\mathbf{J}_r\,\mathbf{V}^\top$ and train a vITL system with the resulting $\mathbf{A}$. While in this case $\mathbf{A}$ is no longer invertible, each coefficient $\hat{c}_{i,j}$ from Lemma 1 belongs to the r-dimensional subspace of $\mathbb{R}^d$ generated by the eigenvectors associated to the r largest eigenvalues of $\mathbf{Y}^\top\mathbf{Y}$. This makes a reparameterization possible and leads to a decrease in the size of the model, going from $t \times m \times d$ parameters to $t \times m \times r$. We tested this idea in the joint emotional input setting, and report in Fig. 7 the resulting test MSE performance (mean ± standard deviation) obtained from 10 different splits, and empirically observe that $r = 20$ suffices to preserve the optimal performances of the model.

Learning under a partial observation regime: To assess the robustness of vITL w.r.t. missing data, we considered the joint emotional setting, and a random mask $(\eta_{i,j})_{i\in[n],j\in[m]} \in \{0,1\}^{n\times m}$; a sample $z_i(\theta_{i,j})$ was used for learning only when $\eta_{i,j} = 1$. Thus, the percentage of missing data was $p := \frac{1}{nm}\sum_{i,j\in[n]\times[m]} \eta_{i,j}$. The experiment was repeated for 10 splits of the dataset, and on each split we averaged the results using 4 different random masks $(\eta_{i,j})_{i\in[n],j\in[m]}$. The resulting test MSE of the predictor as a function of p is summarized in Fig. 8. As it can be seen, the vITL approach is quite stable in the presence of missing data on both datasets.

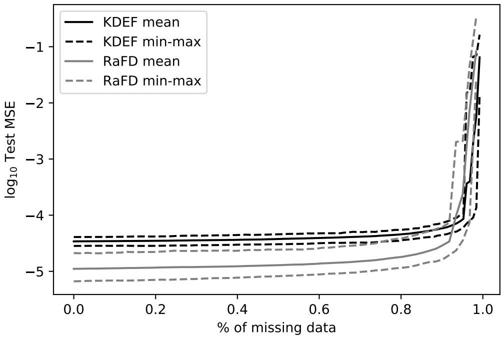


Figure 8: Logarithm of the test MSE (min-mean-max) as a function of the percentage of missing data. Solid line: mean; dashed line: min-max. Smaller MSE is better.

C.5 Additional Qualitative Results

In this section additional quantitative results are shown.

Discrete generation: For discrete expression synthesis results on the RaFD benchmarks, see Fig. 9.

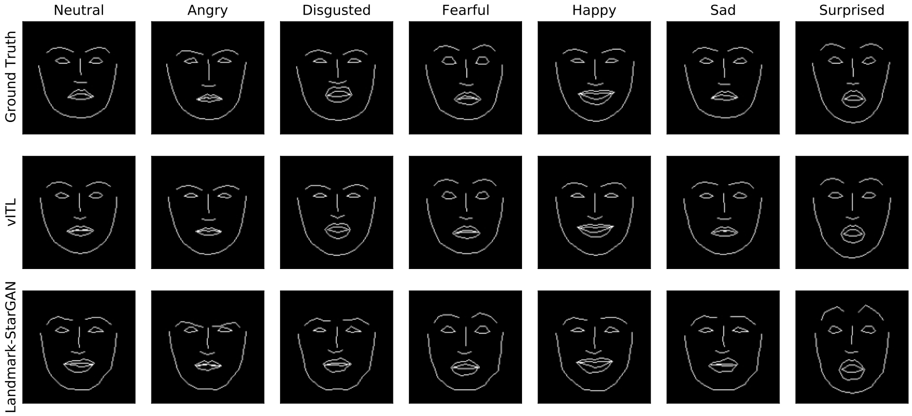


Figure 9: Discrete expression synthesis results on the RaFD dataset with ground-truth neutral landmarks as input.

Continuous generation: Starting from neutral emotion, continuous generation in the radial direction is illustrated in Fig. 10. The landmarks vary smoothly and conform to the expected intensity variation in each emotion on increasing the radius of the vector in VA space.

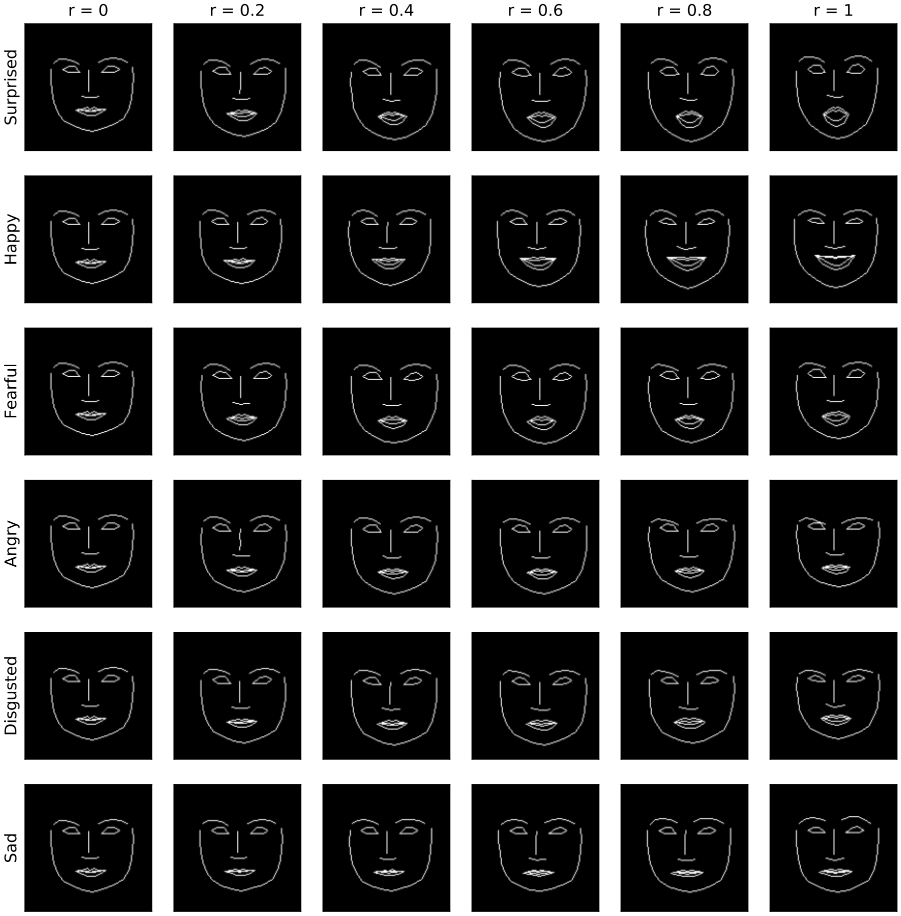


Figure 10: Continuous expression synthesis results with vITL on the KDEF dataset, with ground-truth neutral landmarks. The generation is starting from neutral and proceeds in the radial direction towards an emotion with increasing radii r.