Disentangled representations: towards interpretation of sex determination from hip bone

Final version available in *The Visual Computer*
https://doi.org/10.1007/s00371-022-02755-0

Kaifeng Zou[1], Sylvain Faisan[1], Fabrice Heitz[1], Marie Epain[2], Pierre Croisille[3,4], Laurent Fanton[2,4] and Sébastien Valette[4*]

[1]ICube, University of Strasbourg, CNRS, 300 Bd Sébastien Brant, BP 10413, 67412 Illkirch CEDEX - France .
[2]Hospices Civils de Lyon, 3 Quai des Célestins, 69002 Lyon - France.
[3]University Hospital of Saint-Etienne, 25 bd Pasteur, 42100 Saint-Etienne - France.
[4]CREATIS, Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, CNRS, Inserm, UMR 5220, U1294, F-69621, LYON, France.

*Corresponding author(s). E-mail(s): sebastien.valette@creatis.insa-lyon.fr;

Abstract

Neural network-based classification methods are often criticized for their lack of interpretability and explainability. By highlighting the regions of the input image that contribute the most to the decision, saliency maps have become a popular method to make neural networks interpretable. In medical imaging, they are particularly well-suited for explaining neural networks in the context of abnormality localization. Nevertheless, they seem less suitable for classification problems in which the features that allow distinguishing classes are spatially correlated and scattered. We propose here a novel paradigm based on Disentangled Variational Auto-Encoders. Instead of seeking to understand what the neural network has learned or how prediction is done, we seek to reveal class differences. This is achieved by transforming the sample from a given class into the "same" sample but belonging to another class, thus paving the way to easier interpretation of class differences. Our experiments in the context of automatic sex determination from hip bones show that the obtained results are consistent with expert knowledge. Moreover, the proposed approach enables us to confirm or question the choice of the classifier, or eventually to doubt it.

Keywords: Pattern recognition and classification, Shape analysis, Neural Network, Bone

1 Introduction

In forensic medicine and anthropology, sex determination is generally carried out by manually assessing hip bone features [1]. Automatic classification algorithms are mainly guided by the knowledge of anthropologists, taking into account distances or angles measured from a few anatomical landmarks [2–5]. Currently there exists a

crucial need for practitioners in forensic science to understand classification results and such approaches have the advantage of providing easily interpretable results. But they are specifically tailored for hip bones, and are not well suited to sex determination from other bones or bone fragments, which may be necessary in forensic science.

We propose here an (automatic) deep learning-based classification approach that is completely data-driven, is free of expert knowledge, and is suited to sex determination from other bones or bone fragments. Regardless of these advantages, the proposed method will not be used by practitioners if they cannot interpret the classification results. However, meeting the need for understanding and explainability is far from easy with deep learning classification methods.

Neural networks-based classification methods are often criticized for their lack of interpretability and explainability. Even if there is not a clear consensus on the definition of interpretability and explainability, most methods dealing with interpretability and explainability aim to understand what the neural network has learned or how prediction is done. One common method to interpret the predictions of neural networks is to compute saliency maps (SMs) [6]. However, in the context of this application, the information extracted with SMs was difficult to interpret (examples of SMs are presented in Fig. 7).

To overcome this limitation, we consider here a different paradigm, based on disentangled generative representations. The main novelty of this paper is to show that disentanglement may bring a better understanding of classification results, highlighting the differences between the possible classes.

Disentangled representations allow us to reveal the effects of the factors of interest through the generation of new data obtained by changing the labels related to these factors [7]. As an example, [8] samples the latent space so as to provide insights from brain structure representations. Another model proposed in [9] can simulate brain images at different ages, providing an alternative way of interpreting the aging pattern.

We introduce a disentangled Variational Auto-Encoder (DVAE) to obtain a hip bone mesh representation, in which the sex label is disentangled from the other latent variables. In addition to providing the class of a given sample to analyze, a DVAE can also provide a reconstruction for each class, which provides supplementary information to the user. As an example, if the input mesh is a male one, its reconstruction as a man should be similar to the input mesh and its reconstruction as a woman, on the other hand, should display interpretable differences in sex-specific regions. Moreover, by comparing the two reconstructions with the original mesh for several subjects, the user can get an insight into the morphological differences between male and female hip bones.

Although SMs and the proposed approaches provide understanding and explainability, they do not act at the same level. The SMs facilitate understanding of the decision process (related to a classification method): the purpose is to understand what the neural network has learned or how prediction is performed. An SM therefore reveals information about the classifier itself and not about the classification task. On the contrary, the proposed approach makes it possible to highlight the differences between the classes and thus provides information on the classification problem to be solved.

Finally, in addition to showing that disentanglement can bring a better understanding of classification results, we also show in this paper that feeding a binary classifier with the reconstructions provided by DVAE allows to obtain a classification method that is robust to missing data and therefore well-suited to bone fragments, which is a major advantage (compared to other existing methods) for applications in forensic medicine and anthropology.

Note that the classification approach as such is not the main contribution of this article. Indeed, sex determination from the hip bone may not be considered as challenging in terms of the classification task: the hip bone exhibits significant sexual dimorphism (note that the classification accuracy is very high (Tab. 2)). There are indeed strong anatomical differences between the male and female hip bones, such as the subpubic angle and the shapes of the obturator foramen, of the greater sciatic notch, of the pelvic inlet and of the symphysis.

The main contribution is the proposition that disentanglement can contribute to a better understanding of classification results. In particular, the proposed method allows the users to form their

own opinions. As an example, we will see in Sec. 6 that the reconstructions provided by the proposed approach can sometimes allow us to confirm the choice made by the classifier, or it can also allow us to doubt its choice or even question it.

The remainder of this paper is organized as follows: after the presentation of the related works (Sec. 2), we briefly explain in Sec. 3 how hip bone meshes are obtained from CT scans. Sec. 4 presents the DVAE. Sec. 5 describes the experiments and the results and Sec. 6 proposes a discussion. Since the two reconstructions provided by DVAE enable the users to form their own opinions, Sec. 7 shows that the two reconstructions may also be useful to improve the accuracy of an independent classifier. This section also addresses the case of missing data. In Sec. 8, we illustrate SMs for the proposed networks for comparison. Finally, Sec. 9 concludes the paper.

2 Related works

Interpretability and explainability of deep neural networks may be achieved in two ways.

The first paradigm, known as activation maximization or feature visualization via optimization, consists of producing intuitive visualizations that reveal the meaning of hidden layers. This is mainly achieved by finding a representative input that can maximize the activation of a layer [10, 11].

The second paradigm, known as attribution methods, looks for the network inputs with the highest impact on the network response. In the case of image models, this leads to the estimation of SM, which highlights the regions of the input image that contribute the most to the decision. Many attribution techniques are based on back-propagation. An SM is, for instance, computed in [6] by computing the derivative of the output with respect to the image. Several methods such as SmoothGrad [12] have been proposed to reduce the noise that is present in the gradient. Methods such as CAM [13] and Grad-CAM [14] combine gradients, network weights and/or activations at a specific layer. Other attribution techniques analyze how a perturbation in the input affects the output [15]. Finally, attribution techniques can also be achieved via local model approximation [16].

In medical imaging, SMs are becoming a popular approach that provides interpretability, especially when it comes to localization of abnormalities. Different sanity checks [17], such as intra-architecture repeatability, inter-architecture reproductibility, sensitivity to weight randomization [18] and localization accuracy can be used to assess the relevance of SMs. These criteria helped to justify the use of SMs in some studies such as in [17], but have also led to questions about the relevance of SMs [19, 20]. This indicates that SMs are not suited to all situations.

In our experiments, the information extracted with SMs was difficult to interpret (examples of SMs are presented in Fig. 7). Our hypothesis is that SMs are not easily interpretable on medical imaging classification problems in which the underlying features used by the neural network are spatially correlated, scattered and non-trivial.

Generative models are proposed here as a way of better understanding classification results. These models play a crucial role in many applications and in many common tasks of data science [21–28]. Moreover, there is a key challenge to learn disentangled (generative) representations where some variables of interest (such as acquisition parameters, age, sex or pathology in medical applications) would be independently and explicitly encoded [29]. These representations can either be obtained with Variational Auto-Encoders (VAEs) [30] or with generative adversarial networks (GANs) [31].

Probabilistic generative models, such as VAEs [30], define a joint probability distribution over the data and over latent random variables. Very few assumptions are generally made about the latent variables of deep generative models, leading to entangled representations.

Disentanglement can be achieved with VAEs in the unsupervised case [8, 32], in the (semi)-supervised case [9, 33, 34], and in the weakly-supervised case [35]. In the supervised or semi-supervised case, the factors of interest are explicitly labelled in all or in a part of the training set. In the weakly-supervised case, only implicit information about factors of interest is provided during learning.

The semi-supervised case is of primary importance because better disentangled models can be obtained under supervision [36]. In this case, the latent representation is generally divided into two parts: the non-interpretable part and the disentangled part corresponding to variables that explicitly

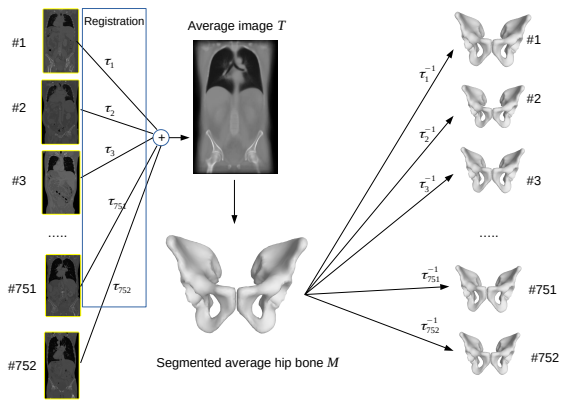

Fig. 1 From CT scans to hip bone meshes

model the factors of interest. In this context, several patterns of conditional dependency structures have been proposed [33, 34, 37].

In addition to VAE approaches, there is a substantial literature on image-to-image translation between unpaired image data using GAN [38–41]. First, some methods try to map an image from one domain (e.g. smiling) to another one (e.g. neutral face). Among these methods, the best known is CycleGAN [42]. This approach is able to preserve key attributes of two different domains and allows to transform an image from one domain to another. Note that StarGAN [43] can perform image-to-image translations for multiple domains. Similar methods, inspired by dual learning, can also be used [44–46] to map the domains. Other GAN based approaches use architectures that are more similar to the VAEs [47, 48]. As an example, conditional GAN [48] allows to disentangle the high level factors from the intrinsic features of the face using two different encoders that compute the latent representation and the attribute information from the image.

3 From CT scans to meshes

In this section we assume that we have one 3D CT scan I_k for each individual k. Computing a mesh of the hip bone from a CT image (Fig. 1) is carried out in six steps:

- *(i)* The scans are registered to a common space using the groupwise registration algorithm FROG [49], that provides a transformation field t_k (for each k) that relates the common space to the I_k's image space.
- *(ii)* Each scan I_k is warped according to t_k (so as to obtain I_k in the common space), and a template T is obtained by averaging the warped images.
- *(iii)* The coxal bone is segmented and meshed in T, thus providing a mesh M. The mesh is composed of about 5000 vertices (we denote by P the 3-D points associated with the mesh M).
- *(iv)* The points P are back-transformed in the native space of each scan I_k using the inverse transform t_k^{-1}, providing for each scan I_k a matrix X_k of size $N_p \times 3$ (N_p is the number of points). Each row of X_k is the 3-D coordinate of one point. Note that the points are ordered since the i-th row of each matrix is associated with the same "anatomical" point.
- *(v)* A shape description invariant to position, size and orientation denoted P_k is obtained using a Procrustes alignment of X_k onto P (for each X_k, we estimate a similarity transformation, namely the combination of a rigid transformation with an isotropic scaling transform). A shape description invariant to position and orientation is required since all subjects do not have the same position during acquisition. However, a description invariant to size is more debatable.
- *(vi)* Since the point sets P_k and P are ordered, the mesh M_k is straightforwardly derived from M and P_k.

4 Disentangled Variational Auto-Encoders for classification and reconstruction

4.1 Conditional dependency structure

The proposed model is part of the family of partially-specified models because an explicit latent variable is defined (the sex of the subject) whereas the semantics of the other latent variables is undefined. Several conditional dependency structures can be defined. As an example, [9] explicitly conditions the latent variables z on age c, such that the conditional distribution $p(z|c)$ captures an age-specific prior on latent representations. We propose here to use a conditional

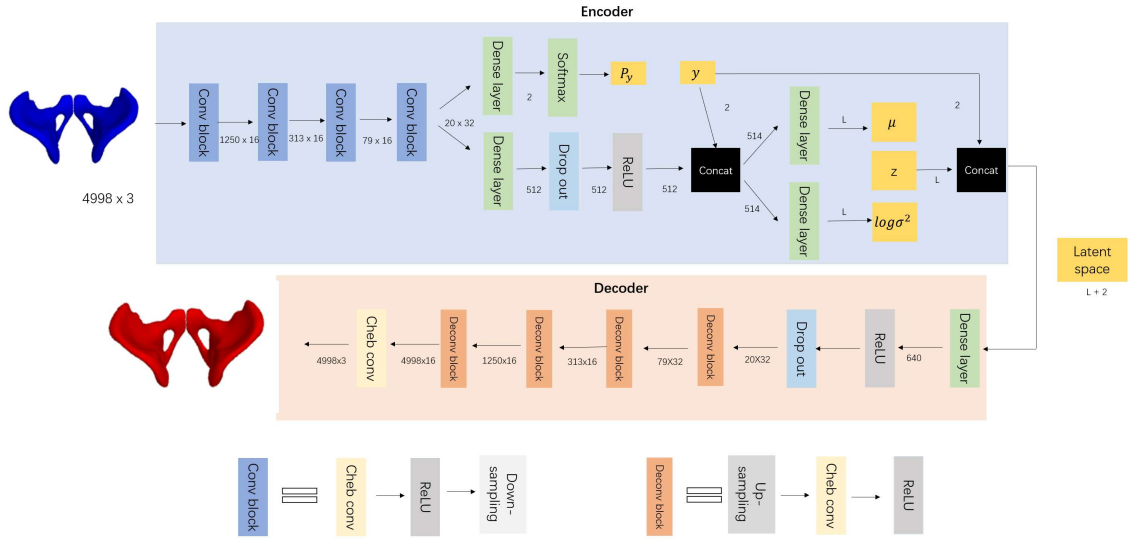

Fig. 2 DVAE for sex determination. There are four main steps. **1.** The distribution $q_\phi(y|x)$ (Eq. 5) is computed using the neural network q_0 (Eq. 6) that outputs the vector P_y whose i-th element is equal to $q_\phi(y = i|x)$ ($i = 1$ or 2). Then, y is set to the most likely label for testing, and is assumed to be known for training. **2.** The parameters μ and $\log \sigma^2$ (both are vectors of size L) of the distribution $q_\phi(z|x, y)$ (Eq. 4) are estimated using the neural networks q_1 and q_2 (Eq. 7). The networks q_1 and q_2 share all their layers except the last one. Moreover, q_0 shares with q_1 (and q_2) the four first convolution blocks of the encoder. Note also that y is injected into the networks q_1 and q_2 through a concatenation layer located before the two dense layers. Since one-hot encoding is used to model y, y is of dimension 2. This explains why the concatenation layer takes as input a vector of dimension 512 and outputs a vector of size 514. **3.** For learning, z is sampled from the distribution $q_\phi(z|x, y)$ using the reparameterization trick (Eq. 8). For testing, z is set to μ. The latent representation of the input data is composed of y and z and is of dimension $L+2$. **4.** The reconstruction can be performed from the latent representation using the decoder (Eq. 9). Note that the two latent representations ($z, y =$ "man"), ($z, y =$ "woman") correspond to the "same" individual but of opposite sex. Consequently, by setting y to the man (resp. woman) label in the latent representation, we can reconstruct the original data as a man (resp. woman). This will enable us to transform a sample from a given class into the "same" sample but of another class (see Sec. 4.3).

dependency structure, as presented in [33, 34], which is suited to our problem.

We denote by x a sample (a mesh), by y its class (male or female), and by $z \in \mathbb{R}^L$ the other latent variables. Note that the latent representation of x is the pair (y, z). We use the following factorization for the generative process:

$$p_\theta(x, y, z) = p_\theta(x|y, z)p(y)p(z), \tag{1}$$

where a weak prior is defined over z and $y : p(z) = \mathcal{N}(z|0, I)$ and $p(y) = \frac{1}{2}$. $p_\theta(x|y, z)$ is modelled as a Gaussian distribution whose mean is given by a neural network f with parameter θ that takes as input y and z. We have:

$$\begin{aligned} p_\theta(x|y, z;\ \theta) &= \mathcal{N}\left(x \mid f(y, z;\ \theta), vI\right), \\ &= \mathcal{N}\left(x \mid \widehat{x}, vI\right), \end{aligned} \tag{2}$$

where $v > 0$ is a hyperparameter and $\widehat{x}$ is the reconstruction computed from y and z.

As usual in variational inference, the posterior $p_\theta(y, z|x)$ is approximated by $q_\phi(y, z|x)$. In order to disentangle the label y from the other latent variables z, we use the following factorization:

$$q_\phi(y, z|x) = q_\phi(y|x)q_\phi(z|x, y). \tag{3}$$

The distribution $q_\phi(z|x, y)$ shows that the estimation of z requires the data x, but also the label y. To understand why this is relevant, let us consider a toy example where z is supposed to represent the size of the subject. If the sex label y is well disentangled from z, z ought to be an intrinsic measure of a subject's size. This means that its estimation needs to regress out the influence of the label y: indeed, a woman who is 160 centimeters tall can be considered as average height while a man of the same height can be considered as short, so that the value of z associated with this woman has to be larger than the one related to this man (even if they have both the

same height). Consequently, in order to obtain a disentangled representation, it seems appropriate that z depends both on x and y.

The distribution $q_\phi(z|x,y)$ in Eq. 3 is defined as a Gaussian distribution whose mean (resp. covariance matrix) is given by a neural network q_1 (resp. q_2) with parameter ϕ_1 (resp. ϕ_2) that both take as input x and y:

$$q_\phi(z|x,y) = \mathcal{N}(z;\ \mu, \sigma^2), \tag{4}$$

where μ and $\log \sigma^2$ are vectors of size L (see Eq. 7 for details). Finally, the distribution $q_\phi(y|x)$ that also appears in Eq. 3 is simply defined as:

$$q_\phi(y|x) = Discrete(y|q_0(x;\ \phi_0)), \tag{5}$$

where q_0 is a neural network with parameter ϕ_0 that takes x as input. The output of this network is a positive vector P_y (Eq. 6) of size 2 summing to 1: the probability $q_\phi(y = i|x)$ is the i-th element of $q_0(x;\ \phi_0)$ ($i = 1$ or 2).

The proposed approach can be summarized as follows:

- If y is known, the neural network q_0 is not required. Otherwise, it acts like a classifier such that the distribution $q_\phi(y|x)$ (Eq. 5) is computed as follows:

$$P_y = q_0(x;\ \phi_0), \tag{6}$$

  and y is set to the most likely label.
- The latent variable z is computed from x and y. Firstly, μ and $\log \sigma^2$ that appear in Eq. 4 are computed such as:

$$\mu = q_1(x, y;\ \phi_1), \log \sigma^2 = q_2(x, y;\ \phi_2). \tag{7}$$

  Then, the latent variable z is set to μ for testing new data whereas Eq. 8:

$$z = \mu + \sigma \odot \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, I), \tag{8}$$

  represents the reparameterization trick that is used for learning (please see the next section). Note that the latent representation of x contains both the variables y and z.
- The reconstruction $\widehat{x}$ can be obtained from y and z as follows:

$$\widehat{x} = f(z, y;\ \theta). \tag{9}$$

The neural networks q_0 (Eq. 6), $q1$ and $q2$ (Eq. 7) represent the encoder and f is the decoder (Eq. 9).

The proposed architecture is depicted in Fig. 2. Networks q_0, q_1, q_2 and f (Fig. 2) are defined using a combination of the convolutions, max-pooling (downsampling) and upsampling operators presented in [50]. Note that mesh convolution is performed in the spectral domain with a kernel parametrized as a Chebyshev polynomial of order K (K is set to 6).

4.2 Parameter optimization

As usual for learning a VAE, the parameters of the DVAE are set to maximize the Evidence Lower BOund (ELBO) [30]. We can show that the term $q_\phi(y|x)$ does not contribute to the loss function because all labels y are known during training. Thus, maximizing the ELBO does not allow the estimation of ϕ_0 (Eq. 6). Consequently, following [33, 34], we add a classification loss $\alpha \log q_\phi(y|x)$ to the ELBO term. The criterion writes:

$$E_{z\sim q_\phi(z|x,y)}\left[\log \frac{p_\theta\,(x,y,z)}{q_\phi(z \mid x,y)}\right] + \alpha \log q_\phi(y|x). \tag{10}$$

Based on the conditional dependency structure of the model, Eq. 10 can be simplified as:

$$\begin{aligned} &E_{z\sim q_\phi(z|x,y)}\left[log(p(z)) - log(q_\phi(z|x,y))\right] \quad + \\ &E_{z\sim q_\phi(z|x,y)}\left[\log(p_\theta(x|y,z))\right] \quad + \\ &\log(p(y)) + \alpha \log q_\phi(y|x). \end{aligned} \tag{11}$$

The first term may be expressed as a Kullback–Leibler divergence ($-KL((q_\phi(z|x,y)||p(z)))$ which can be computed analytically since the encoder model and prior are Gaussian. The second term is approximated by a Monte Carlo estimate: we use the SGVB estimator and the reparameterization trick [30] (Eq. 8). The third term corresponds to the prior of the label y, that has been set to $1/2$. Finally, the last term is computed by the neural network q_0.

The loss function contains two hyperparameters: α that weights the contribution of the classification loss, and the variance v (Eq. 2), which is used to compute the second term of Eq. 11. As in the VAE case, the variance v weights the contribution of the mean squared error reconstruction and special care is needed to set v. In the

following, the two hyperparameters v and α are estimated using cross-validation strategies (note that the influence of the parameter α is limited and could simply be set to 1).

4.3 DVAE for classification and reconstruction

The proposed generative model can be used for classification but it also offers the opportunity to transform a sample from a given class to the "same" sample but belonging to another class, by modifying the value of the categorical variables y in the latent representation. The reconstruction of a male mesh (resp. female) as a female mesh (resp. male) is carried out according to the following "sex change" procedure:

- Step 1: The latent variable z is computed from the input data x and its true label y using Eq. 7 (z is set to μ). The latent representation corresponds to variables z and y.
- Step 2: We change the value of y in the latent representation, so that we obtain the latent representation of the "same" individual but of the opposite sex.
- Step 3: The reconstruction can be performed with Eq. 9 (using the modified latent representation).

In order to test the consistency of the results, we also developed a sex preservation procedure. This is the same procedure as the sex change procedure except that the value of y is not modified in the latent representation (Step 2 is not performed).

Note that the computation of the latent variable z requires knowledge of the sex of the mesh under analysis since the true label y is required to compute μ (Eq. 7). For testing, since the sex of the mesh under analysis is not known, we have to replace the true label by its most likely estimate computed with q_0.

However, for the reconstruction step (Eq. 9), note that we can choose to reconstruct a subject either as a man or as a woman by setting y in the latent representation appropriately.

5 Experiments

Our database consists of 752 CT scans from the University Hospital of Saint-Etienne, France, of which 470 subjects are men and 282 subjects are women. The men are on average 65.8 years old with a standard deviation of 14.2 years and the women are on average 65.6 years old with a standard deviation of 14.6 years.

For each scan, a hip bone mesh is extracted as explained in Sec. 3. Each point coordinate is normalized so as to have zero-mean and unit-variance. The means and standard deviations are computed using the training dataset (see Section 5.1.2).

In addition to training a DVAE, we also train a vanilla VAE whose architecture is the same as that represented in Fig. 2 except that the label y and the computation of P_y (Eq. 6) are removed. The usual criterion [30] is used for training the VAE.

We also learn a classifier (denoted C) whose architecture is derived from the one in Fig. 2 by keeping only the layers that are useful for the computation of P_y (Eq. 6). C and q_0 have the same architecture but q_0 is only a subpart of the DVAE (q_0 shares some layers with q_1 and q_2) whereas C is an independent classifier. The binary cross entropy loss is used for training C.

Finally, we use PyTorch for implementation.

5.1 Evaluation protocol

5.1.1 Hyperparameter setting

In the VAE case, the variance v is estimated automatically during the training process with the method proposed in [51]: v is computed for each batch as the MSE loss.

Regarding the DVAE, several methods have been tested without success to estimate v automatically. This is why the parameter v as well as the parameter α (Eq. 11) are set using cross-validation strategies.

It has been observed that the size of the latent space has limited influence on classification accuracy and on the disentanglement properties for a large range of values of L (for $L = 1$ to 64). However, using too small values of L leads to an increase in the reconstruction error. L has been set to 16 in all experiments. For a fair comparison, the size of the latent space of the VAE has been set to L+2=18.

Optimization of the parameters was done using the Adam optimization algorithm with a batch size of 16. During training, all models are trained

for 600 epochs. We keep the same learning rate of 0.0006 for the first 200 epochs and then decay the learning rate to 0.0003 for the next 200 epochs. For the last 200 epochs, we set the learning rate to 0.0001. Training time for DVAE is about 7.2 sec per epoch with a 2080 Ti graphics card. The DVAE needs about 0.2 seconds to generate both male and female hip bones during testing.

5.1.2 Nested-cross validation strategy

In order to estimate the ability of the models to handle unseen data and to set the hyperparameters α and v for the DVAE, we follow the nested cross-validation strategy.

First, an (outer) stratified 5-fold cross-validation strategy is used to assess the performance of the models. At each iteration, all folds except one are used as training data (it will be denoted TR) and the remaining one is used as testing data (TE). The three models (DVAE, C, and VAE) are trained from TR and their performances are evaluated on TE. Note that a score can be computed for each fold. We can then derive an average score and its standard deviation.

However, the DVAE learning process requires the hyperparameters α and v to be defined. An inner K-fold cross-validation could be applied at each iteration of the outer cross-validation. However, this would require training a very large number of models. To make the problem tractable, we instead randomly divide the training set TR into a validation set denoted V and a training set T (20% and 80 %). Afterwards, several models are trained from T based on different values for the hyperparameters: a grid search is performed for α and v (α and $\sqrt{v}$ take resp. their value in $\{0.5, 1, 2, 3, 4, 5\}$ and in $\{0.7, 1, 1.3, 1.6, 1.9\}$). Once all models have been trained, the set V is used to select the model that provides the highest disentanglement, that is, the one that leads to the highest success rate for the sex change procedure (see sec. 5.1.3). Then, a final model is trained from TR based on the hyperparameters that have led to obtain the selected model (note that TE is used neither to estimate the parameters of the model nor to estimate the hyperparameters).

5.1.3 Evaluation metrics

In the (semi)-supervised case, evaluating disentanglement is often achieved by visualising the reconstructions while modifying the value of a latent variable of interest. In our specific case, this can be easily achieved since the latent variable of interest y is binary (a hip bone is either associated with a man or a woman). Consequently, the model is tested on its ability to perform conditional generation according to the sex label (Sec. 5.2.1 proposes quantitative results while Sec. 5.2.2 presents some visual examples). The model is also tested for its ability to classify hip bones and to reconstruct the original data.

For each fold, we compute four different metrics to evaluate the performance of the model:

- The classification accuracy (CA) obtained with q_0 (DVAE) or with classifier C.
- The opposite sex reconstruction success rate (OSRSR): we reconstruct a male (resp. female) as a female (resp. male) mesh using the sex change procedure (Sec. 4.3). This procedure is considered as successful if the transformed mesh is classified as female (resp. male) using C. This rate should be high if the sex label y has been been properly disentangled from z.
- The same sex reconstruction success rate (SSRSR): we reconstruct a male (resp. female) as a male (resp. female) mesh using the sex preservation procedure (Sec. 4.3). This procedure is deemed as successful if the transformed mesh is classified by classifier C as male (resp. female).
- The reconstruction error (RE) in millimeters. The reconstruction obtained with the sex preservation procedure is compared with the initial mesh in the native space of the image I_k (see Sec. 3). The mean of the euclidean distances between each associated point is computed leading to a score for a given subject. This score is then averaged over all subjects in the fold. Note that obtaining the reconstruction in the space of I_k requires the inversion of the normalization step applied to each point coordinate (second paragraph of Sec. 5) as well as the similarity transformation (point (v) in Section 3).

Note that all metrics except CA are computed using different reconstructions of the mesh under analysis. In order to distinguish between classification errors and reconstruction/disentanglement errors, the true label is used to compute the latent representation.

Table 1 Results (mean and standard deviation) obtained with the DVAE approach. CA, OSRSR, SSRSR, and RE stand resp. for classification accuracy, opposite sex reconstruction success rate, same sex reconstruction success rate, and reconstruction error.

CA	OSRSR	SSRSR	RE
$99.59 \pm 0.34\%$	$99.10 \pm 0.92\%$	100%	$1.647mm \pm 0.098mm$

Table 2 Comparison with previous works on sex determination. Note that previous works rely on manual estimation (such as lengths, angles or landmark positions) while our approach is fully automatic.

Method	individuals	variables	accuracy
CADOES [2]	256	40 (manual)	97 %
DSP [3, 4]	2040	17 (manual)	$>$ 99 %
Nikita et al. [5]	132	3 (manual)	97 %
Ours	752	5000 (autom.)	$>$ 99 %

5.2 Experimental performance analysis

5.2.1 Quantitative results

The results obtained with the DVAE approach are shown in Tab. 1. Regarding the classification accuracy, the DVAE classifier achieves a very high prediction accuracy ($99.59 \pm 0.34\%$). This corresponds to a total of 3 misclassifications out of 752 (one misclassification in 3 folds and zero in 2 folds). The independent classifier C achieves similar results since only three subjects are misclassified (these are not the same subjects).

As a comparison, Tab. 2 gives sex prediction accuracy for recent works that are based on the manual positioning of a few landmarks. We cannot claim that the proposed method provides better results since all the methods should be compared on the same database (which unfortunately is not available). However, the proposed method yields state-of-the-art classification results while being free of any manual positioning of landmarks. Moreover, the method is data-driven and not guided by expert knowledge. It is also suited to sex determination from other bones and, as shown in Sec. 7.2, from bone fragments.

In terms of reconstruction error, the DVAE performs similarly to a vanilla VAE, which obtains a mean reconstruction error of 1.728 mm, even if the selected values of v at each fold (DVAE) are always larger than those estimated (for each batch) with the method of [51] (VAE). The selected values of v in the DVAE case are relatively large because it has been observed that small values of v lead to poor disentanglement properties. However, an increase in v did not increase reconstruction error.

One could remark that the comparison of the reconstruction errors may be unfair since the true sex label is employed to perform the reconstruction in the DVAE case. However, the same result is obtained when using the estimated label: there are only 3 misclassified cases and using the true label or the false one leads to reconstructions that are mostly similar, except in some specific regions.

Finally, excellent results are obtained for the opposite sex reconstruction success rate, and for the same sex reconstruction success rate. The reconstruction as a female (resp. male) mesh of a male (resp. female) mesh is well-classified by C in more than 99% of the cases (OSRSR). Moreover, the reconstruction as a male (resp. female) mesh of a male (resp. female) mesh is always well-classified by C in our experiments (SSRSR). Note that the accuracy of the classifier C reaches only $97.17 \pm 1.05\%$ when classifying data reconstructed with the vanilla VAE (instead of 100% in the DVAE case).

As noted previously, the comparison with the VAE approach may be unfair since the true label is used for reconstruction in the DVAE case. However, we can use a sex preservation procedure that does not rely on the true label (the label can be estimated by q_0). In this case, when classifying the reconstructions obtained by DVAE, the classifier C reaches an accuracy of $99.59 \pm 0.34\%$, which is exactly the accuracy of q_0 (see Tab. 1). Indeed, classifying with C the reconstruction obtained with the DVAE provides exactly the same results than classifying the original mesh with q_0. This

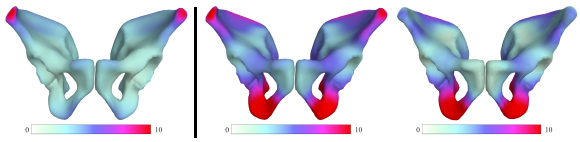

Fig. 3 Local average distances. From left to right: original meshes vs reconstructed meshes (lower distance is better), original meshes vs reconstructed opposite sex meshes, reconstructed meshes vs reconstructed opposite sex meshes. Distances are in mm. See text for details.

clearly shows the consistency of the method. As an example, if a male mesh is considered as a female one by q_0, the DVAE will reconstruct this male mesh as a female one so that the classifier C will be also wrong.

5.2.2 Qualitative results

In order to evaluate more precisely the disentanglement properties of the model, each original mesh M_k is compared with its reconstructed (same sex) mesh or with its reconstructed opposite sex mesh. Furthermore, the two reconstructions are also compared together. Note that the two reconstructed meshes are those computed in the previous section (the true label y is used to compute z).

We start by analyzing average results. As in Sec. 5.1.3 (please see the definition of RE), the reconstructions (associated with M_k) are computed in the native space of I_k. To compare two (out of the three) meshes, we associate at each vertex v of the template mesh M a real value representing the distance between the two vertices v of the meshes under analysis. These distances are averaged across the different subjects of the testing set. Each vertex of the template mesh therefore receives a color representing the (local) average distance.

These local average distances are represented in Fig. 3 (left) when the original meshes are compared with the same sex reconstructed meshes. This comparison shows that the iliac crest is not well reconstructed. This is mainly due to large registration errors that can be observed for some subjects in this region. This makes the problem more difficult because the variability of the data is increased.

As illustrated in Fig. 3 (middle) that represents the local average differences between the original meshes and the opposite sex meshes, the

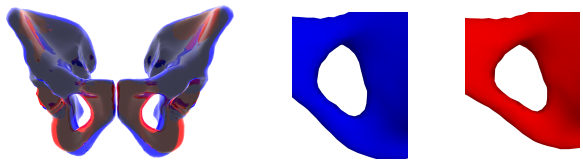

Fig. 4 Example of changing a male hip bone (blue) to a female hip bone (red). Left: angle comparison: the subpubic angle is larger for the female bone than for the male bone. Right: the male obturator foramen (left) exhibits an oval shape, while the female obturator foramen (right) exhibits a triangular shape.

opposite sex reconstruction changes the geometry as expected. Moreover, the differences that can be observed are consistent with expert knowledge. As an example, the subpubic angle is known to be larger for women, leading to the difference observed in the pubic arch.

The two reconstructed meshes can be compared (Fig. 3 (right)) in order to gain a deeper understanding of the results. This is particularly true for the iliac crest, which is not well reconstructed in both cases. In the case of complete disentanglement of the sex label, we expect this area to be reconstructed similarly for both reconstructions. This is because the iliac crest is known to show little sexual dimorphism compared to other areas of the hip bone. Even if Fig. 3 (right) still exhibits differences in the iliac crest between the two reconstructions, they remain low compared to the original reconstruction errors (Fig. 3 (left)).

Finally, these results reinforce the idea that the sex variable has been properly disentangled.

We can explore further by analyzing individual results. The analysis of the differences between two meshes was carried out using "cine mode" (rapidly switching between them) because the eye is sensitive to movement. For the sake of simplicity, the two meshes are here directly superimposed to compare them (see Fig. 4 and 5).

When opposite sex reconstruction is successful, the comparison of the opposite sex mesh with the original mesh (or the reconstructed one) reveals the significant anatomical differences between the male and female hip bones, such as the subpubic angle (Fig. 4, left) as well as the shape of the obturator foramen (Fig. 4, right), of the greater sciatic notch, of the pelvic inlet and of the symphysis. Note that it may sometimes happen that the two

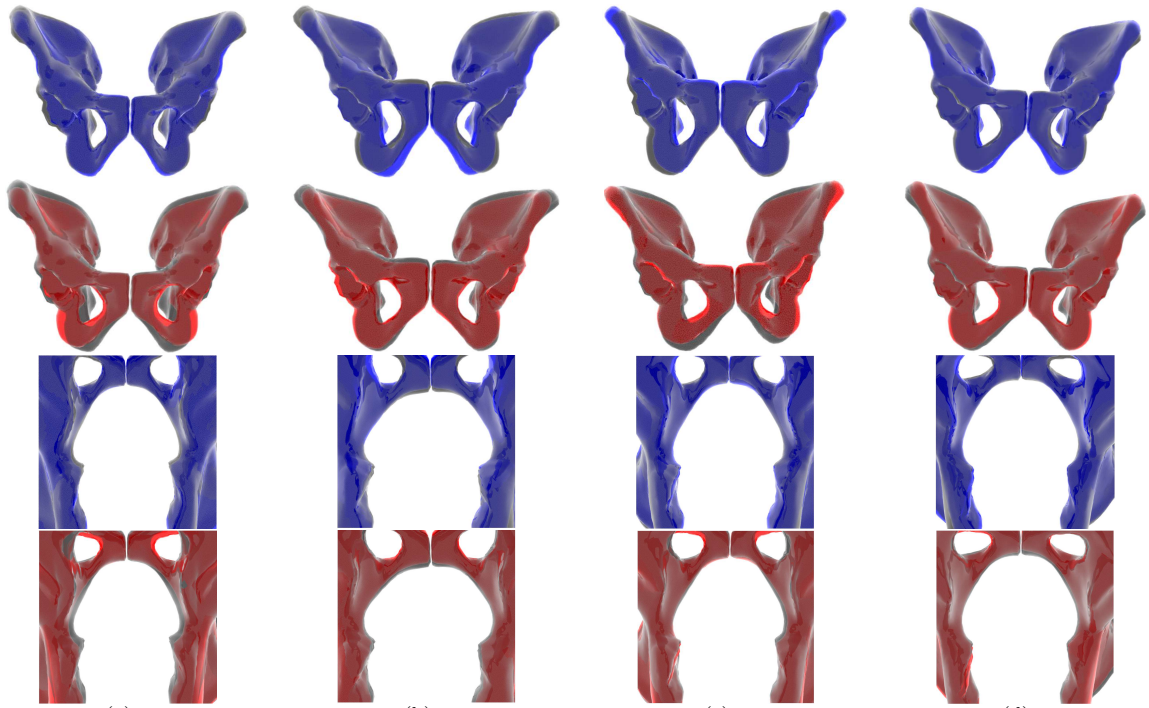

Fig. 5 Examples of DVAE results. Original mesh (grey) *vs* mesh reconstructed as a female one (red). Original mesh (grey) *vs* mesh reconstructed as a male one (blue). The original mesh of (b) is a female one while those of (a,c,d) are male meshes.

meshes do not exhibit all the expected differences, but most of them are generally easily observable.

When opposite sex reconstruction is not successful, the modification is globally consistent, as some significant anatomical differences can be observed, but some of them are sometimes hard to see, or event not present.

6 Discussion: In what sense does the method provide understanding?

Predicting sex from a hip mesh is not an easy task for a non-expert and the classification results can be difficult to understand. In the proposed approach, in addition to providing the class of the mesh, its reconstructions as a man and as a woman are also provided. When the original mesh is that of a man (resp. woman), its reconstruction as a man (resp. woman) is very similar to the original mesh. Conversely, the comparison between the original mesh and its reconstruction with opposite sex exhibits differences in some specific areas (while others remain unchanged). The comparison of these reconstructions with the original mesh enables a non-expert to understand the choice of the classifier, or at the very least to make their own choice.

Fig.5(a) gives an illustrative example of the results provided by DVAE. The reconstruction of the mesh as a man is very similar to the original mesh. On the contrary, the reconstruction as a woman exhibits a wider subpic angle and a wider pelvic inlet. Consequently, a non-expert can easily classify the mesh as a male (without using the result of the classifier), or at least, understand why this mesh can be considered as a male one.

It is then legitimate to ask what happens if the label is not correctly estimated by q_0: will the proposed method justify a misclassification or will it detect the mistake? This part should not be considered as a failure case analysis. The purpose of the proposed method is to provide relevant and easily interpretable information so that the users can form their own opinions. Consequently, if the classifier is wrong but the information given by

DVAE enables the user to question its decision, this can certainly be considered a positive result.

Both DVAE and C misclassified 3 subjects, we analyze them in detail here. The different reconstructions relative to the misclassified meshes are shown in Fig. 5(b,c,d) (note that y is provided by q_0 for the computation of z so as not to bias the results). The 6 misclassified cases can be split into three groups.

The first group is composed of 3 misclassified subjects (one for C and two for DVAE). Fig. 5(c) is an illustrative example of this group. It is a man that has been misclassified by C. The reconstruction as a man is very similar to the original mesh in the sex-specific regions, whereas the reconstruction as a woman exhibits some differences in these regions. Consequently, the original mesh seems to be a male mesh and the user may question the choice of the classifier. Moreover, the iliac crest is particularly poorly reconstructed in these 3 subjects. The shape of this region may be responsible for the misclassification.

The second group is composed of 2 misclassified subjects (one for C and DVAE). Fig.5(b) is an illustrative example of this group. It represents a woman that has been misclassified by DVAE. When looking at the subpic angles, it seems to be consistent: the reconstruction as a woman is very similar to the original mesh in this area. However, the reconstruction of the pelvic inlet suggests that this is a male mesh (the reconstruction as a man is very similar to the original mesh in this area). Thus, this mesh has both male and female characteristics. This may explain why this subject is difficult to classify. In this case, the two reconstructions enable the user to doubt the result obtained by the classifier.

The last group is composed of one misclassified subject: this is a man (Fig. 5(d)) that has been misclassified by DVAE. When it is reconstructed as a woman, the subpubic angle is slightly increased and the pelvic inlet is made wider, as expected. When it is reconstructed as a man, we expect the reconstruction to be similar to the original mesh but the subpubic angle is slightly decreased. Consequently, the subpubic angle of this man seems to be larger than it should be. This may explain why this subject has been misclassified. However, a user could easily question the results obtained by the classifier, because it seems that the mesh exhibits more male characteristics than female ones.

Finally, the comparison of the two reconstructions with the original mesh is a simple way to understand the choice that was made by the classifier, or to doubt its choice (for the second group) or to question it (for the first and last groups).

7 Reconstruction-based classification: application to missing data

7.1 Reconstruction-based classification

As written in Sec. 6, the comparison of the two reconstructed meshes provided by the DVAE approach with the original mesh enables a non-expert to form an informed opinion. In the same way, one can wonder if the performance of an independent classifier can be improved by feeding the two reconstructed meshes obtained with DVAE to the classifier.

To this end, the following paradigm has been used: after having trained the DVAE, we train an independent classifier denoted C_{recon} whose input data are composed of two meshes: the first one is the original mesh from which we subtract its reconstruction as a man (provided by DVAE, z is computed using the label estimated by q_0) and the second one is the original mesh from which we subtract its reconstruction as a woman. The classifier C_{recon} is identical to C except the first layer that takes an input of size 4998x6 (we have points in R^6 because we model two meshes). In the following, we denote this method DVAE+C_{recon}.

DVAE+C_{recon} achieves an accuracy of 100% for each fold, even with meshes having both female and male characteristics (Sec. 6). One possible reason for these results is that the work of C_{recon} is much simpler than the one of C. As an example, let us consider the case of a male mesh. Its reconstruction as a man is very similar to the original mesh so that the first three components of the mesh (we have points in R^6) are close to zero. On the contrary, the reconstruction as a woman exhibits differences in some sex-specific regions so that the last three components of the mesh are close to zero except in the sex-specific regions. Consequently, for a male mesh, all components are

expected to be close to zero except the last three components that lie in the sex-specific regions. For a female mesh, all components are expected to be close to zero except the three first components that lie in the sex-specific regions. By highlighting the regions that allow to distinguish male from female hip bone, the input of C_{recon} is much easier to analyze than the original mesh.

7.2 Application to missing data

Since all the classifiers C, C_{recon} and DVAE have already achieved high accuracies, we propose here to make the problem more difficult by introducing missing data: vertices are deleted either on the left-hand, right-hand, lower, upper, front or rear side. The percentage of missing data is expressed in terms of the percentage of the mesh size (in the dimension where the data is removed). As an example, when deleting data on the lower side, the percentage of missing data is expressed in terms of the percentage of the height of the mesh. A very simple imputation strategy is used: missing values are set to the value 0 (which is the mean at each vertex).

Data augmentation is required during training to achieve acceptable results: with a probability of 0.6, the mesh is not modified. Otherwise, it is augmented as follows. The side where the vertices are set to 0 is chosen with a uniform distribution, and the percentage of missing data is selected with a uniform distribution in $0 - 40\%$.

Four different methods are used for classification:

1 The classifier C.
2 DVAE: note that the second term of the loss function (Eq. 11) uses the original mesh (and not the augmented one) since we want the reconstruction to be similar to the original mesh.
3 DVAE+ C_{recon}. DVAE is first trained as in the second point. Then, during the learning of C_{recon}, the two reconstructions of an augmented mesh are computed using the DVAE (z is computed using the label estimated by q_0) and the input of C_{recon} corresponds to the augmented mesh from which we subtract its reconstructions. This means that C_{recon} is somehow fed indirectly with augmented meshes during the learning.

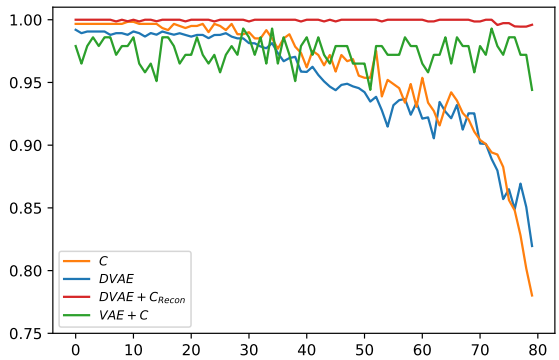

Fig. 6 Classification accuracy obtained with different methods in the presence of missing data. The x-axis corresponds to the percentage of missing data (×100).

4 the last method denoted VAE+C consists in classifying the reconstruction provided by the VAE with C. The VAE is trained in a similar way as the DVAE. Since the VAE provides a reconstruction without any missing data, the classifier C is trained with non-augmented meshes.

Classification accuracy is shown in Fig. 6 for a large range of missing data.

As previously, we can note that DVAE and C provide similar results. Even if 70% of the data is missing, C and DVAE can still achieve an accuracy of 90%.

We can also note that the two methods that use reconstructions (VAE+C and DVAE+C_{recon}) are quite robust to missing data but DVAE+C_{recon} performs always better than other classification methods. This clearly highlights the benefit of feeding the classifier indirectly with the two reconstructed meshes provided by DVAE.

Finally, the fact that the proposed method is able to achieve very good results in cases where there is a high proportion of missing data seems to indicate that it is able to take into account most of the differences that exist between female and male hip bones.

8 Comparison with saliency maps

To compare our approach for the interpretation of mesh classification with a standard method, we have computed SMs for the classifiers C and

C_{recon} (without missing data) with the method in [6]. For a given input mesh, the importance w_{ic} at each vertex v_i is computed as follows:

$$w_{ic} = |\frac{\partial p(y=0|x)}{\partial x_{ic}}| = |\frac{\partial p(y=1|x)}{\partial x_{ic}}|, \qquad (12)$$

where x_{ic} (c =1, 2 or 3) represents either the x, y or z coordinate.

Eq. 12 can be computed through back-propagation. For each vertex, the 3 computed importances (one for each coordinate c) are aggregated using the max function: the SM at vertex i is computed as $max_c(w_{ic})$. Instead of considering the derivative of $p(y|x)$, it is also possible to use the unnormalised score (the softmax layer is not considered for the computation of the derivative). In this case, Eq. 12 no longer holds and a SM is obtained for each class. Regardless of the methods used or the aggregation function used, the results were always very similar. Fig. 7 represents the mean of the SMs (across the subjects), computed with Eq. 12 and the max aggregation function.

It is difficult to understand how classifier C makes its decision (Fig. 7, left), as the most relevant vertices for the classification are distributed over the entire hip bone (we could expect them to lie specifically in regions that are known to differ between men and women, but this is not the case).

The individual SMs were also extremely different from one another, whereas one would expect that they would all highlight sex-specific regions. Finally, the results were neither intra-architecture repeatable nor inter-architecture repeatable. We suggest that SM may not be suitable for classification problems in which the features that allow distinguishing classes are spatially correlated and scattered. Under these conditions, two classifiers can achieve high accuracy results without having the same decision boundaries, hence their respective SMs will be different.

To illustrate this hypothesis, let us take a simplified problem in which the hip bone is modeled with four variables. To simulate the fact that the hip bone is symmetrical, suppose that x_1 is close to $-x_2$ and that x_3 is close to $-x_4$. The variables x_3 and x_4 represent sex-specific regions ($x_3 \geq 0$ for female hip bones and $x_3 \leq 0$ for male hip bones). Then, let us consider the two following neural networks whose boundary equations are $x_3 - x_4 = 0$ and $x_3\mathbb{1}_{x_1<0} - x_4\mathbb{1}_{x_1\geq 0} + x_1 + x_2 = 0$ (note that $x_1 + x_2$ is likely to be close to 0 for hip bones), where $\mathbb{1}$ is the indicator function. The two neural networks are expected to achieve high accuracy. However, only the SM of the first one is able to highlight the regions of interest x_3 and x_4. The SM of the last one is expected to highlight either x_3 or x_4 according to the value of x_1 as well as two regions that are not sex-specific (x_1 and x_2).

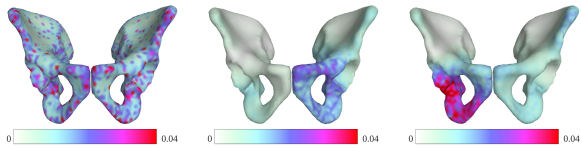

Fig. 7 Mean SMs for C (left) and C_{recon} (center and right). The SMs for C_{recon} are either averaged across the female hip bones (center) or the male ones (right).

For C_{recon}, the map is more consistent with our expectations (Fig. 7, center and right) except that a strong asymmetry is observed depending on whether the processed hip bone is a female one or a male one. That is why, the SMs are either averaged across the female hip bones (Fig. 7, center) or the male ones (Fig. 7, right). Moreover, contrary to the local average distances (Fig. 3), the mean SMs highlight the pubic left tubercle, whose shape is known to vary slightly according to the sex (this is clearly visible for the mean SM associated with women, a little less for that associated with men). It seems that the classifier focuses here on a subtle difference between female and male hip bones. Since the input of C_{recon} is partly fed with the output of the DVAE, it can be estimated that this small difference has been captured by DVAE.

Note also that similar mean SMs can be obtained when measuring intra-architecture repeatability and inter-architecture reproductibility. In all cases, the mean SMs associated with men and women highlight a different side of the hip bone and this asymmetry can be more or less pronounced. Moreover, the side of the regions of interest may be permuted: the mean SM of male hip bones highlights the regions that are on the right side (Fig. 7, right) but it can be the left side for other tests.

We can conclude that the SMs obtained with C_{recon} are more satisfactory than those obtained with C. Our interpretation is that the input of C_{recon} is much simpler to analyze since the sex-specific regions have been highlighted by DVAE:

all components that lie in regions that are not sex-specific are close to 0.

As a final point, it is noteworthy that the proposed method differs significantly from SMs.

First, as written in Sec. 1, they do not act at the same level. The SM facilitates understanding of the decision process related to a classification method whereas the proposed approach highlights the differences between the classes and thus provides information on the classification problem to be solved.

Then, an intrinsic limitation of SMs is that they do not provide any semantic meaning on the highlighted regions. In our application, the SMs can at best detect sex-specific regions, i.e. regions that allow to distinguish between male and female hip bones. In contrast, thanks to the conditional generation according to the sex, the proposed method not only provides a sex-specific region detection but also offers the user the opportunity to observe the difference in shape of regions: as an example, we clearly observe with the proposed method that the subpic angle is larger for women (Fig.4). Such an approach leads to a better understanding of the class differences.

Moreover, the proposed method provides the users with relevant information so that they can form their own opinions. As an example, we have seen in Sec. 6 that the comparison of the two reconstructions enables us to show that some meshes exhibit both male and female characteristics.

Finally, contrary to SMs which is a generic tool, the proposed approach is only suitable if the label to estimate is a variable corresponding to a source of variability (age, sex, outcomes of genomic-biological-cognitive tests, diagnosis, multicenter variability), which are common situations in medical imaging. As an example, it makes sense in the proposed application to reconstruct a male hip bone as a female one (or a diseased organ into a healthy one) because the latent space can be divided into two independent parts: the non-interpretable part represents the intrinsic (independent of sex) properties of the hip bone and the disentangled part represents the sex label.

9 Conclusion

This paper has presented a novel paradigm for the interpretation of classification by neural networks, based on Disentangled VAE representations. The approach provides reconstructions or data generation for each class, which paves the way for a better understanding of class differences. The approach has been illustrated through the interpretation of sex determination from meshed hip bones. It compares favorably with existing methods such as SMs.

The proposed paradigm is comprehensive and suited to the disentanglement and classification of other factors of general interest in medical imaging, such as age, pathology or acquisition parameters. Moreover, there are some cases where some features can be associated with high-level factors. As an example, features related to the disease label may be its severity, and more generally characteristics that model how the disease has transformed the disease-free sample. Note that studies [52, 53] have shown the benefit of modeling both the high-level factors and their related features to disentangle the high-level factors.

Future directions of this work include the modeling of these features and the comparison of the proposed approach with generative adversarial networks that also can achieve disentanglement in a supervised setting. Moreover, learning the significant differences between the classes (at the population level) during training is another perspective that would help to determine if the differences observed for a particular sample under classification are related to opposite sex reconstruction or if they stem from other reasons such as registration inaccuracy. This may further help the analysis of the results.

Data availability

Due to ethical concerns, the original CT scans cannot be made openly available, but the hip bone mesh dataset is available from the corresponding author on reasonable request.

Statements and Declarations

This work was funded by the TOPACS ANR-19-CE45-0015 project of the French National Research Agency (ANR).

The authors have no financial or proprietary interests in any material discussed in this article.

References

[1] Komar, D., Buikstra, J.: Forensic Anthropology: Contemporary Theory And Practice. Oxford University Press, New York (2008)

[2] d'Oliveira Coelho, J., Curate, F.: Cadoes: An interactive machine-learning approach for sex estimation with the pelvis. Forensic Science International **302** (2019)

[3] Murail, P., Bruzek, J., Houët, F., Cunha, E.: DSP: A tool for probabilistic sex diagnosis using worldwide variability in hip-bone measurements. Bulletins et mémoires de la Société d'Anthropologie de Paris **17**(3-4), 167–176 (2005)

[4] Brůžek, J., Santos, F., Dutailly, B., Murail, P., Cunha, E.: Validation and reliability of the sex estimation of the human os coxae using freely available DSP2 software for bioarchaeology and forensic anthropology. American Journal of Physical Anthropology **164**(2), 440–449 (2017)

[5] Nikita, E., Nikitas, P.: Sex estimation: a comparison of techniques based on binary logistic, probit and cumulative probit regression, linear and quadratic discriminant analysis, neural networks, and naïve Bayes classification using ordinal variables. International Journal of Legal Medicine **134**(3), 1213–1225 (2020)

[6] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Workshop at International Conference on Learning Representations (2014)

[7] Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: ECCV (4), pp. 776–791 (2016)

[8] Liu, R., Subakan, C., Balwani, A.H., Whitesell, J., Harris, J., Koyejo, S., Dyer, E.L.: A generative modeling approach for interpreting population-level variability in brain structure. In: MICCAI, pp. 257–266 (2020)

[9] Zhao, Q., Adeli, E., Honnorat, N., Leng, T., Pohl, K.M.: Variational autoencoder for regression: Application to brain aging analysis. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pp. 823–831. Springer, ??? (2019)

[10] Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal (June 2009)

[11] Nguyen, A., Yosinski, J., Clune, J.: In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Understanding Neural Networks via Feature Visualization: A Survey, pp. 55–76. Springer, Cham (2019)

[12] Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise. In: Workshop on Visualization for Deep Learning, ICML (2017)

[13] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

[14] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**(2), 336–359 (2019)

[15] Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: ICCV, pp. 2950–2958 (2019)

[16] Ribeiro, M., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 97–101. Association for Computational Linguistics, San Diego, California (2016)

[17] Arun, N.T., Gaw, N., Singh, P., Chang,

K., Hoebel, K.V., Patel, J., Gidwani, M., Kalpathy-Cramer, J.: Assessing the validity of saliency maps for abnormality localization in medical imaging. In: Medical Imaging with Deep Learning (2020)

[18] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 9525–9536 (2018)

[19] Eitel, F., Ritter, K.: Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. Springer, Cham (2019)

[20] Young, K., Booth, G., Simpson, B., Dutton, R., Shrapnel, S.: Deep neural network or dermatologist? Lecture Notes in Computer Science (2019)

[21] Zhang, Y., Ong, C.C., Zheng, J., S.-T., L., Z., G.: Generative design of decorative architectural parts. The Visual Computer **38** (2022)

[22] Yoshikawa, T., Endo, Y., Kanamori, Y.: Diversifying detail and appearance in sketch-based face image synthesis. The Visual Computer **38** (2022)

[23] Li, Y., Wang, Z., Yin, L., Zhu, Z., Qi, G., Liu, Y.: X-net: a dual encoding–decoding method in medical image segmentation. The Visual Computer (2021)

[24] Azizi, V., Usman, M., Zhou, H., Faloutsos, P., Kapadia, M.: Graph-based generative representation learning of semantically and behaviorally augmented floorplans. The Visual Computer **38** (2022)

[25] Nozawa, N., Shum, H.P.H., Feng, Q., Ho, E.S.L., Morishima, S.: 3d car shape reconstruction from a contour sketch using gan and lazy learning. The Visual Computer **38** (2022)

[26] Wen, J., Ma, H., Luo, X.: Deep generative smoke simulator: connecting simulated and real data. The Visual Computer **36** (2020)

[27] Wang, S., Zou, Y., Min, W., Wu, J., Xiong, X.: Multi-view face generation via unpaired images. The Visual Computer **38** (2022)

[28] Phaphuangwittayakul, A., Ying, F., Guo, Y., Zhou, L., Chakpitak, N.: Few-shot image generation based on contrastive meta-learning generative adversarial network. The Visual Computer (2022)

[29] Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798–1828 (2013)

[30] Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: 2nd International Conference on Learning Representations, ICLR, Canada, Conference Track Proceedings (2014)

[31] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

[32] Chen, R.T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems, vol. 31 (2018)

[33] Siddharth, N., Paige, B., van de Meent, J.-W., Desmaison, A., Goodman, N.D., Kohli, P., Wood, F., Torr, P.H.S.: Learning disentangled representations with semi-supervised deep generative models. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

[34] Kingma, D.P., Rezende, D.J., Mohamed, S., Welling, M.: Semi-supervised learning with deep generative models. In: Advances in Neural Information Processing Systems (2014)

[35] Ruiz, A., Martinez, O., Binefa, X., Verbeek,

J.: Learning Disentangled Representations with Reference-Based Variational Autoencoders (2019)

[36] Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: International Conference on Machine Learning, pp. 4114–4124 (2019). PMLR

[37] Maaløe, L., Sønderby, C.K., Sønderby, S.K., Winther, O.: Auxiliary deep generative models. In: Proceedings of The 33rd International Conference on Machine Learning, pp. 1445–1453 (2016)

[38] Wang, Q., Artières, T., Chen, M., Denoyer, L.: Adversarial learning for modeling human motion. The Visual Computer **36** (2020)

[39] Liu, X., Huang, H., Wang, W., Zhou, J.: Multi-view 3d shape style transformation. The Visual Computer **38** (2022)

[40] Ju, Y., Zhang, J., Mao, X., Xu, J.: Adaptive semantic attribute decoupling for precise face image editing. The Visual Computer **37** (2021)

[41] Yin, Z., Xia, K., Wang, S., He, Z., Zhang, J., Zu, B.: Unpaired low-dose ct denoising via an improved cycle-consistent adversarial network with attention ensemble. The Visual Computer (2022)

[42] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

[43] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)

[44] Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

[45] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of International Conference on Computer Vision (ICCV) (2020)

[46] Yi, Z., Zhang, H., Tan, P., Gong, M.: DualGAN: Unsupervised Dual Learning for Image-to-Image Translation (2017)

[47] Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.: Fader networks: Manipulating images by sliding attributes. Advances in neural information processing systems **30** (2017)

[48] Perarnau, G., van de Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible Conditional GANs for image editing. In: NIPS Workshop on Adversarial Training (2016)

[49] Agier, R., Valette, S., Kéchichian, R., Fanton, L., Prost, R.: Hubless keypoint-based 3D deformable groupwise registration. Medical Image Analysis **59** (2020)

[50] Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3D faces using convolutional mesh autoencoders. In: European Conference on Computer Vision (ECCV), pp. 725–741 (2018)

[51] Rybkin, O., Daniilidis, K., Levine, S.: Simple and Effective VAE Training with Calibrated Decoders (2021)

[52] Joy, T., Schmon, S., Torr, P., Siddharth, N., Rainforth, T.: Capturing label characteristics in vaes. In: International Conference on Learning Representations (2020)

[53] Zou, K., Faisan, S., Heitz, F., Valette, S.: Joint disentanglement of labels and their features with VAE. In: IEEE International Conference on Image Processing (ICIP) (2022)