Deforming Autoencoders: Unsupervised Disentangling of Shape and Appearance

Zhixin Shu[1(✉)], Mihir Sahasrabudhe[2], Rıza Alp Güler[2,3], Dimitris Samaras[1], Nikos Paragios[2,4], and Iasonas Kokkinos[5,6]

[1] Stony Brook University, Stony Brook, NY, USA
zhshu@cs.stonybrook.edu
[2] CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France
[3] INRIA, Rocquencourt, France
[4] TheraPanacea, Paris, France
[5] Univeristy College London, London, UK
[6] Facebook AI Research, Paris, France

Abstract. In this work we introduce Deforming Autoencoders, a generative model for images that disentangles shape from appearance in an unsupervised manner. As in the deformable template paradigm, shape is represented as a deformation between a canonical coordinate system ('template') and an observed image, while appearance is modeled in deformation-invariant, template coordinates. We introduce novel techniques that allow this approach to be deployed in the setting of autoencoders and show that this method can be used for unsupervised groupwise image alignment. We show experiments with expression morphing in humans, hands, and digits, face manipulation, such as shape and appearance interpolation, as well as unsupervised landmark localization. We also achieve a more powerful form of unsupervised disentangling in template coordinates, that successfully decomposes face images into shading and albedo, allowing us to further manipulate face images.

1 Introduction

Disentangling factors of variation is important for the broader goal of controlling and understanding deep networks, but also for applications such as image manipulation through interpretable operations. Progress in the direction of disentangling the latent space of deep generative models has facilitated the separation of latent image representations into dimensions that account for independent factors of variation, such as identity, illumination, normals, and spatial support [1–4], low-dimensional transformations, such as rotations, translation, or scaling [5–7] or finer-levels of variation, including age, gender, wearing glasses, or other attributes e.g. [2,8] for particular classes, such as faces.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-01249-6_40) contains supplementary material, which is available to authorized users.

V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11214, pp. 664–680, 2018.
https://doi.org/10.1007/978-3-030-01249-6_40

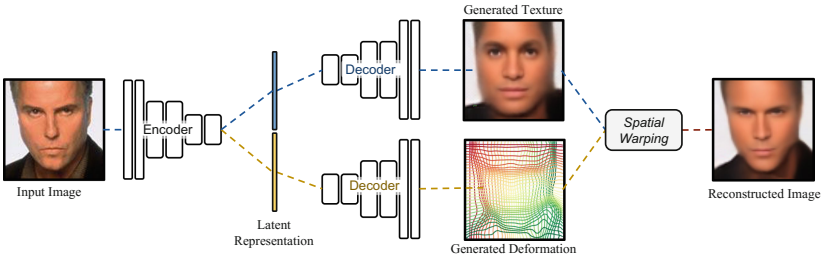

Fig. 1. Deforming Autoencoders follow the deformable template paradigm and model image generation through a cascade of appearance (or, 'texture') synthesis in a canonical coordinate system and a spatial deformation that warps the texture to the observed image coordinates. By keeping the latent vector for texture short, we force the network to model shape variability through the deformation branch. This allows us to train a deep generative image model that disentangles shape and appearance in an entirely unsupervised manner, using solely an image reconstruction loss for training.

Shape variation is more challenging as it is a transformation of a function's domain, rather than its values. Even simple, supervised additive shape models result in complex nonlinear optimization problems [9,10]. Nonetheless, several works in the previous decade aimed at learning shape/appearance factorizations in an unsupervised manner, exploring groupwise image alignment, [11–14].

In a deep learning context, several works incorporated deformations and alignment in supervised settings, including Spatial Transformers [15], Deep Epitomic Networks [16], Deformable CNNs [17], Mass Displacement Networks [18], Mnemonic Descent [19], Densereg [20] or more recently, works that use surface-based 3D face models for accurate face analysis [21,22]. These works have shown that one can improve the accuracy of both classification and localization tasks by injecting deformations and alignment within traditional CNN architectures.

Turning to unsupervised deep learning, even though most works focus on rigid or low-dimensional parametric deformations, e.g. [5,6], several works have attempted to incorporate richer non-rigid deformations within learning. A thread of work aims at dynamically rerouting the processing of information within the network's graph based on the input, starting from neural computation arguments [23–25] and eventually translating into concrete algorithms, such as the 'capsule' works [26,27] that bind neurons on-the-fly. Still, these works lack a transparent, parametric handling of non-rigid deformations. On a more geometric direction, recent work aims at recovering dense correspondences between pairs [28] or sets of RGB images, e.g. [29,30]. These works however do not have the notion of a reference coordinate system ('template') to which images can get mapped - this makes image generation and manipulation harder. More recently, [31] use the equivariance principle to align sets of images to a common coordinate system, but do not develop this into a full-blown generative model of images.

Our work advances this line of research by following the deformable template paradigm [9,10,32–34]. In particular, we consider that object instances are

obtained by deforming a prototypical object, or 'template', through dense, diffeomorphic deformation fields. This makes it possible to factor object variability within a category into variations that are associated to spatial transformations, generally linked to the object's 2D/3D shape, and variations that are associated to appearance (or, 'texture' in graphics), e.g. due to facial hair, skin color, or illumination. In particular we model both sources of variation in terms of a low-dimensional latent code that is learnable in an unsupervised manner from images. We achieve disentangling by breaking this latent code into separate parts that are fed into separate decoder networks that deliver appearance and deformation estimates. Even though one could hope that a generic convolutional architecture will learn to represent such effects, we argue that explicitly injecting this inductive bias in a network can help with training, while also yielding control over the generative process. Our main contributions in this work are:

First, we introduce the *Deforming Autoencoder* architecture, bringing together the deformable modeling paradigm with unsupervised deep learning. We treat the template-to-image correspondence task as that of predicting a smooth and invertible transformation. As shown in Fig. 1, our network first predicts a transformation field in tandem with a template-aligned appearance field. It subsequently deforms the synthesized appearance to generate an image similar to its input. This allows us to disentangle shape and appearance by explicitly modelling the effects of image deformation during decoding.

Second, we explore different ways in which deformations can be represented and predicted by the decoder. Instead of building a generic deformation model, we compose a global, affine deformation field, with a non-rigid field that is synthesized as a convolutional decoder network. We develop a method that prevents self-crossings in the synthesized deformation field and show that it simplifies training and improves accuracy. We also show that class-related information can be exploited, when available, to learn better deformation models: this yields sharper images and can be used to learn models that jointly account for multiple classes - e.g. all MNIST digits.

Third, we show that disentangling appearance from deformation has several advantages for modeling and manipulating images. Disentangling leads to clearly better synthesis results for tasks such as expression, pose or identity interpolation, compared to standard autoencoder architectures. Similarly, we show that accounting for deformations facilitates further disentangling of appearance components into intrinsic, shading-albedo decompositions, which allow us to re-shade through simple operations on the latent shading coordinates.

We complement these qualitative results with a quantitative analysis of the learned model in terms of landmark localization accuracy. We show that our method is not too far below supervised methods and outperforms with a margin the latest state-of-the-art works on self-supervised correspondence estimation [31], even though we never explicitly trained our network for correspondence estimation, but rather only aimed at reconstructing pixel intensities.

2 Deforming Autoencoders

Our architecture embodies the deformable template paradigm in an autoencoder architecture. Our premise is that image generation can be interpreted as the combination of two processes: a synthesis of appearance on a deformation-free coordinate system ('template'), followed by a subsequent deformation that introduces shape variability. Denoting by $T(\mathbf{p})$ the value of the synthesized appearance (or, texture) at coordinate $\mathbf{p} = (x, y)$ and by $W(\mathbf{p})$ the estimated deformation field, we reconstruct the observed image, $I(\mathbf{p})$ as follows:

$$I(\mathbf{p}) \simeq T(W(\mathbf{p})), \tag{1}$$

namely the image appearance at position $\mathbf{p}$ is obtained by looking up the synthesized appearance at position $W(\mathbf{p})$. This is implemented in terms of a bilinear sampling layer [15] that allows us to pass gradients through the warping process.

The appearance and deformation functions are synthesized by independent decoder networks. The inputs to the decoders are delivered by a joint encoder network that takes as input the observed image and delivers a low-dimensional latent representation, Z, of shape and appearance. This is split into two parts, $Z = [Z_T, Z_S]$ which feed into the appearance and shape networks respectively, providing us with a clear separation of shape and appearance.

2.1 Deformation Field Modeling

Rather than leave deformation modeling entirely to back-propagation, we use some domain knowledge to simplify and accelerate learning. The first observation is that global aspects can be expressed using low-dimensional linear models. We account for global deformations by an affine Spatial Transformer layer, that uses a six-dimensional input to synthesize a deformation field as an expansion on a fixed basis [15]. This means that the shape representation, Z_S described above is decomposed into two parts, Z_W, Z_A, where Z_A accounts for the affine, and Z_W for the non-rigid, learned part of the deformation field. As is common practice in deformable modeling [9,10], these deformation fields are generated by separate decoders and are composed so that the affine transformation warps the detailed non-rigid warps to the image positions where they should apply.

We note that not every non-rigid deformation field is plausible. Without appropriate regularization the deformation field can amount to a generic permutation matrix. As observed in Fig. 2(f), a non-regularized deformation can spread a connected texture pattern to a disconnected image area.

To prevent this problem, instead of the shape decoder CNN directly predicting the local warping field $W(\mathbf{p}) = (W_x(x, y), W_y(x, y))$, we consider a 'differential decoder' that generates the spatial gradient of the warping field: $\nabla_x W_x$ and $\nabla_y W_y$, where ∇_c denotes the $c - th$ component of the spatial gradient vector. These two quantities measure the displacement of consecutive pixels - for instance $\nabla_x W_x = 2$ amounts to horizontal scaling by a size of 2, while $\nabla_x W_x = -1$ amounts to left-right flipping; a similar behavior is associated with

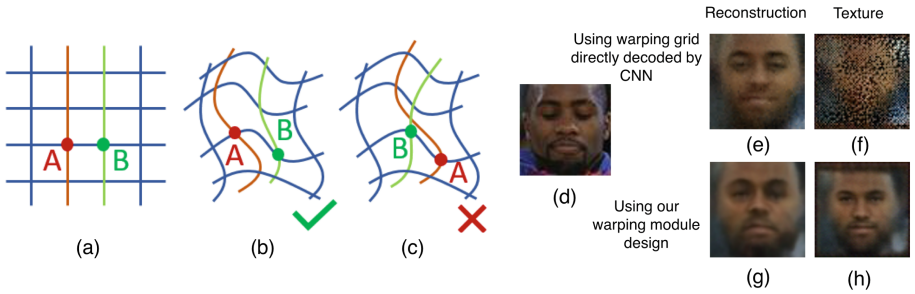

Fig. 2. Our warping module design only permits locally consistent warping, as shown in (b), while the flipping of relative pixel positions, as shown in (c), is not allowed by design. To achieve this, we let the deformation decoder predict the horizontal and vertical increments of the deformation ($\nabla_x W$ and $\nabla_y W$, respectively) and use a ReLU transfer function to remove local flips, caused by going back in the vertical or horizontal direction. A spatial integral module is subsequently applied to generate the grid. This simple mechanism serves as an effective constraint for the deformation generation process, while allowing us to model free-form/non-rigid local deformation.

$\nabla_y W_y$ in the vertical axis. We note that global rotations are handled by the affine warping field, and the $\nabla_x W_y, \nabla_y W_x$ are associated with small local rotations of minor importance - we therefore focus on $\nabla_x W_x, \nabla_y W_y$. Having access to these two values gives us a handle on the deformation field, since we can prevent folding/excessive stretching by controlling $\nabla_x W_x, \nabla_y W_y$.

In particular, we pass the output of our differential decoder through a Rectified Linear Unit (ReLU) layer, which enforces positive horizontal offsets on horizontally adjacent pixels, and positive vertical offsets on vertically adjacent pixels. We subsequently apply a spatial integration layer, implemented as a fixed network layer, on top of the output of the ReLU layer to reconstruct the warping field from its spatial gradient. Thus, the new deformation module enforces the generation of smooth and regular warping fields that avoid self-crossings. In practice we found that clipping the decoded offsets by a maximal value significantly eases training, which amounts to replacing the ReLU layer, $\mathrm{ReLU}(x) = \max(x, 0)$ with a $\mathrm{HardTanh}_{0,\delta}(x) = \min(\max(x, 0), \delta)$ layer. In our experiments we set $\delta = 5/w$, where w denotes the number of pixels along an image dimension.

2.2 Class-Aware Deforming Autoencoder

We can require our network's latent representation to predict not only shape and appearance, but also instance class, if that is available during training. This discrete information may be easier to acquire than the actual deformation field, which requires manual landmark annotation. For instance, for faces such discrete information could represent the expression or a person's identity.

In particular we consider that the latent representation can be decomposed as follows: $Z = [Z_T, Z_C, Z_S]$, where Z_T, Z_S are as previously the appearance-

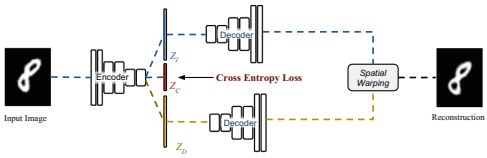

Fig. 3. A *class-aware* model can account for multi-modal deformation distributions by utilizing class information. Introducing a classification loss into latent space helps the model learn a better representation of the input as demonstrated on MNIST.

and shape- related parts of the representation, respectively, while Z_C is fed as input to a sub-network trained to predict the class associated with the input image. Apart from assisting the classification task, the latent vector Z_C is fed into both the appearance and shape decoders, as shown in Fig. 3. Intuitively this allows our decoder network to learn a mixture model that is conditioned on class information, rather than treating the joint, multi-modal distribution through a monolithic model. Even though the class label is only used during training, and not for reconstruction, our experimental results show that a network trained with class supervision can deliver more accurate synthesis results.

2.3 Intrinsic Deforming Autoencoder: Deformation, Albedo and Shading Decomposition

Having outlined Deforming Autoencoders, we now use a Deforming Autoencoder to model complex physical image signals, such as illumination effects, without a supervision signal. For this we design the Intrinsic Deforming-Autoencoder (Intrinsic-DAE) to model shading and albedo for in-the-wild face images. As shown in Fig. 4(a), we introduce two separate decoders for shading S and albedo A, each of which has the same structure as the original texture decoder. The texture is computed by $T = S \circ A$ where $\circ$ denotes the Hadamard product.

In order to model the physical properties of shading and albedo, we follow the intrinsic decomposition regularization loss used in [2]: we apply the L2 smoothness loss on ∇S, meaning that shading is expected to be smooth, while leaving albedo unconstrained. As shown in Fig. 4 and more extensively in the experimental results section, when used in tandem with a Deforming Autoencoder, we can successfully decompose a face image into shape, albedo, and shading components, while a standard Autoencoder completely fails at decomposing unaligned images into shading and albedo. We note that unlike [22], our decomposition is obtained in an entirely unsupervised manner.

2.4 Training

Our objective function is formed as the sum of three losses, combining the reconstruction error with the regularization terms required for the modules described

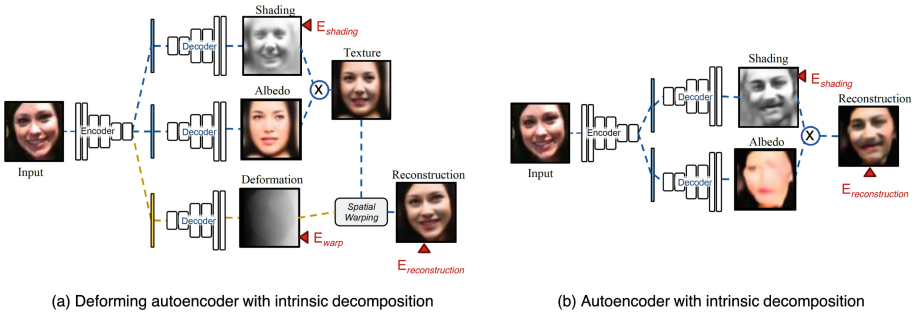

Fig. 4. Autoencoders with intrinsic decomposition. (a) Deforming Autoencoder with intrinsic decomposition (Intrinsic-DAE): we model the texture by the product of shading and albedo components, each of which is decoded by an individual decoder. The texture is subsequently warped by the predicted deformation field. (b) A plain autoencoder with intrinsic decomposition. Both networks are trained with a reconstruction loss ($E_{\text{Reconstruction}}$) for the final output and a regularization loss on shading (E_{Shading}).

above. Concretely, the loss of the deforming autoencoder can be written as

$$E_{\text{DAE}} = E_{\text{Reconstruction}} + E_{\text{Warp}}, \tag{2}$$

where the reconstruction loss is defined as the standard ℓ_2 loss

$$E_{\text{Reconstruction}} = \|I_{\text{Output}} - I_{\text{Input}}\|^2, \tag{3}$$

and the warping loss is decomposed as follows:

$$E_{\text{Warp}} = E_{\text{Smooth}} + E_{\text{BiasReduce}}. \tag{4}$$

The smoothness cost, E_{smooth}, penalizes quickly-changing deformations encoded by the local warping field. It is measured in terms of the total variation norm of the horizontal and vertical differential warping fields, and is given by:

$$E_{\text{Smooth}} = \lambda_1 \left(\|\nabla W_x(x,y)\|_1 + \|\nabla W_y(x,y)\|_1 \right), \tag{5}$$

where $\lambda_1 = 1e-6$. Finally, $E_{\text{BiasReduce}}$ is a regularization on (1) the affine parameters defined as the L2-distance between S_A and S_0, S_0 being the identity affine transform and (2) the average of the deformation grid for a random batch of training data being close to identity mapping grid:

$$E_{\text{BiasReduce}} = \lambda_2 \|S_A - S_0\|^2 + \lambda_2' \|\bar{W} - W_0\|^2, \tag{6}$$

where $\lambda_2 = \lambda_2' = 0.01$. $\bar{W}$ denotes the average deformation grid of a mini-batch of training data and W_0 denotes an identity mapping grid. In the class-aware variant described in Sect. 2.2 we augment the loss above with the cross-entropy loss evaluated on the classification network's outputs. We add the following objective function in the training of the Intrinsic-DAE: $E_{\text{Shading}} = \lambda_3 \|\nabla S\|^2$ where $\lambda_3 = 1e-6$.

We experiment with two architecture types: (1) DAE with a standard convolutional auto-encoder, where both encoder and decoders are CNNs with standard convolution-BatchNorm-ReLU blocks. The number of filters and the texture bottleneck capacity can vary per experiment, image resolution, and dataset, as detailed in the supplemental material; (2) Dense-DAE with a densely connected convolutional network [35] for encoder and decoders respectively (no skip connections over the bottleneck layers). In particular, we follow the architecture of DenseNet-121, but without the 1×1 convolutional layers inside each dense block.

3 Experiments

To demonstrate the properties of our deformation disentangling network, we conduct experiments on MNIST, 11k Hands [36] and Faces-in-the-wild datasets [37, 38]. Our experiments include (1) unsupervised image alignment/appearance inference; (2) learning semantically meaningful manifolds for shape and appearance; (3) unsupervised intrinsic decomposition and (4) unsupervised landmarks detection.

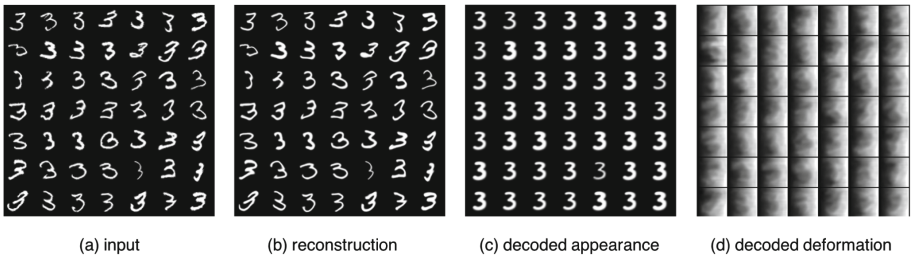

Fig. 5. Unsupervised deformation-appearance disentangling on a single MNIST digit. Our network learns to reconstruct the input image while automatically deriving a canonical appearance for the input image class. In this experiment, the dimension of the latent representation for appearance Z_T is 1.

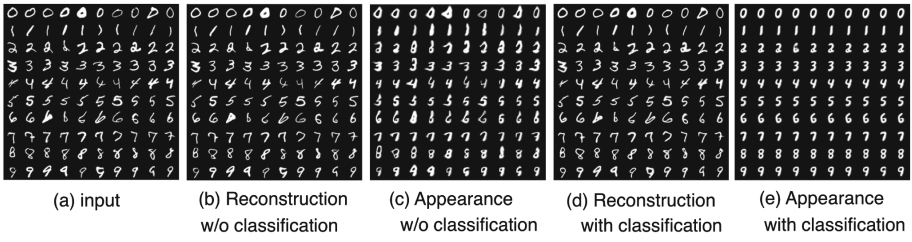

Fig. 6. Class-aware Deforming Autoencoders effectively model the appearance and deformation for multi-class data.

3.1 Unsupervised Appearance Inference

We model canonical appearance and deformation for single category objects. We demonstrate results in the MNIST dataset (Figs. 5 and 6). By limiting the size of Z_T (1 in Fig. 5), we can successfully infer a canonical appearance for a class. In Fig. 5, all different types of digit '3' are aligned to a simple canonical shape.

In cases where the data has a multi-modal distribution exhibiting multiple different canonical appearances, e.g., multi-class MNIST images, learning a single appearance is less meaningful and often challenging (Fig. 6(b)). In such cases, utilizing class information (Sect. 2.2) significantly improves the quality of multi-modal appearance learning (Fig. 6(d)). As the network learns to classify the images implicitly in its latent space, it learns to generate a single canonical appearance for each class. Misclassified data will be decoded into an incorrect class: the image at position (2, 4) in Fig. 6(c, d) is interpreted as a 6.

Moving to a more challenging modeling task, we consider modeling faces in-the-wild. Using the MAFL face dataset we show that our network is able to align the faces to a common texture space under various poses, illumination conditions, or facial expressions (Fig. 9(d)). The aligned textures retain the information of the input image such as lighting, gender, and facial hair, without using any relevant supervision. We further demonstrate the alignment on the 11k Hands dataset [36], where we align palmar images of the left hand of several subjects (Fig. 7). This property of our network is especially useful for applications such as computer graphics, where establishing correspondences (UV map) between a class of objects is important but usually difficult.

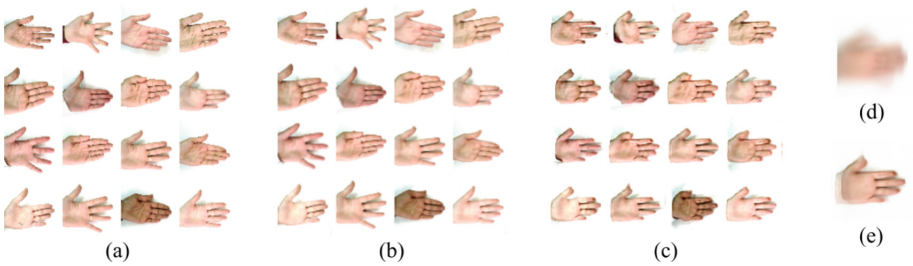

Fig. 7. Unsupervised alignment on images of palms of left hands. (a) The input images; (b) reconstructed images; (c) texture images warped with the average of the decoded deformation; (d) the average input image; and (e) the average texture.

3.2 Autoencoders Vs. Deforming Autoencoders

We now show the ability of our network to learn meaningful deformation representations without supervision. We compare our disentangling network with a plain auto-encoder (Fig. 8). Contrary to our network which disentangles an image into a template texture and a deformation field, the auto-encoder is trained to encode all of the image in a single latent representation.

Fig. 8. Latent representation interpolation: we embed a face image in the latent space provided by an encoder network. Our network disentangles the texture and deformation in the respective parts of the latent representation vector, allowing a meaningful interpolation between images. Interpolating the deformation-specific part of the latent representation changes the face shape and pose (1); interpolating the latent representation for texture will generate a pose-aligned texture transfer between the images (2); traversing both latent representations will generate smooth and sharp image deformations (3, 5, 7). In contrast, when using a standard auto-encoder (4, 6, 8) such an interpolation often yields artifacts.

We train both networks with the MAFL dataset. To evaluate the learned representation, we conduct manifold traversal (i.e., latent representation interpolation) between two randomly sampled face images: given a source face image I^s and a target image I^t, we first compute their latent representations Zs. We use $Z_T(I^s)$ and $Z_S(I^s)$ to denote the latent representations in our network for I^s, and $Z_{ae}(I^s)$ for the latent representation learned by a plain autoencoder. We then conduct linear interpolation on Z, between Z^s and Z^t: $Z^\lambda = \lambda Z^s + (1-\lambda)Z^t$. We subsequently reconstruct the image I^λ from Z^λ using the corresponding decoder(s), as shown in Fig. 8.

By traversing the learned deformation representation only, we can change the shape and pose of a face while maintaining its texture (Fig. 8(1)); interpolating the texture representation results in pose-aligned texture transfer (Fig. 8(2)); traversing on both representations will generate a smooth deformation from one image to another (Fig. 8(3, 5, 7)). Compared to the interpolation using the autoencoder (Fig. 8(4, 6, 8)), which often exhibits artifacts, our traversal stays on the semantic manifold of faces and generates sharp facial features.

3.3 Intrinsic Deforming Autoencoders

Having demonstrated the disentanglement abilities of Deforming Autoencoders, we now explore the disentanglement capabilities of the Intrinsic-DAE described in Sect. 2.3. Using only the E_{DA} and regularization losses, the Intrinsic-DAE is able to generate convincing shading and albedo estimates without direct supervision (Fig. 9(b) to (g)). Without the "learning-to-align" property, a baseline autoencoder with an intrinsic decomposition design (Fig. 4(b)) cannot decompose the image into plausible shading and albedo(Fig. 9(h), (i), (j)).

In addition, we show that by manipulating the learned latent representation of S, Intrinsic-DAE allows us to simulate illumination effects for face images, such as interpolating lighting directions (Fig. 10).

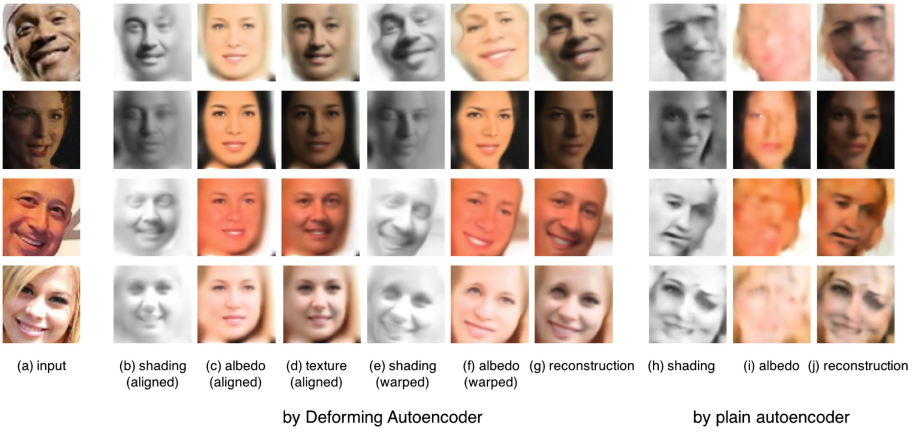

Fig. 9. Unsupervised intrinsic decomposition with an Intrinsic-DAE. Thanks to the "automatic dense alignment" property of DAE, shading and albedo are faithfully separated (e, f) by the intrinsic decomposition loss. Shading (b) and albedo (c) are learned in an unsupervised manner in the densely aligned canonical space. With the deformation field also learned without supervision, we can recover the intrinsic image components for the original shape and viewpoint (e, f). Without dense alignment, the intrinsic decomposition loss fails to decompose shading and albedo (h, i, j).

As a final demonstration of the potential of the learned models for image synthesis, we note that with $L2$ or $L1$ reconstruction losses, autoencoder-like

Fig. 10. Lighting interpolation with Intrinsic-DAE. With latent representations learned in an unsupervised manner for shading, albedo, and deformation, the DAE allows us to simulate smooth transitions of the lighting direction. In this example, we interpolate the latent representation of the shading from source (lit from the left) to target (mirrored source, hence lit from the right). The network generates smooth lighting transitions, without explicitly learning geometry, as shown in shading (1) and texture (2). Together with the learned deformation of the source image, DAE enables the relighting of the face in its original pose (3).

architectures are prone to generating smooth images which lack visual realism (Fig. 9). Inspired by generative adversarial networks (GANs) [39], we follow [2] and use an adversarial loss to generate visually realistic images. We train the Intrinsic-DAE with an extra adversarial loss term $E_{\text{Adversarial}}$ applied on the final output, yielding:

$$E_{\text{Intinsic-DAE}} = E_{\text{Reconstruction}} + E_{\text{Warp}} + \lambda_4 E_{\text{Adversarial}}. \tag{7}$$

In practice, we apply a PatchGAN [40,41] as the discriminator and set $\lambda_4 = 0.1$. As shown in Fig. 11, the adversarial loss improves the visual sharpness of the reconstruction while the deformation, shading are still successfully disentangled.

3.4 Unsupervised Alignment Evaluation

Having qualitatively analyzed the disentanglement capabilities of our networks, we now turn to quantifying their performance on the task of unsupervised face landmark localization. We report performance on the MAFL dataset, which contains manually annotated landmark locations (eyes, nose, and mouth corners) for 19,000 training and 1,000 test images. In our experiments, we use a model trained on the CelebA dataset without any form of supervision. Following the evaluation protocol of previous work [31], we train a landmark regressor post-hoc on these deformation fields using the provided training annotations in MAFL. The annotation from the MAFL training set is only used to train the regressor while the DAE is fixed after pre-training. The regressor is a 2-layer MLP. Its inputs are flattened deformation fields (vectors of size $64 \times 64 \times 2$), which are provided as input to a 100-dimensional hidden layer, followed by a ReLU and a

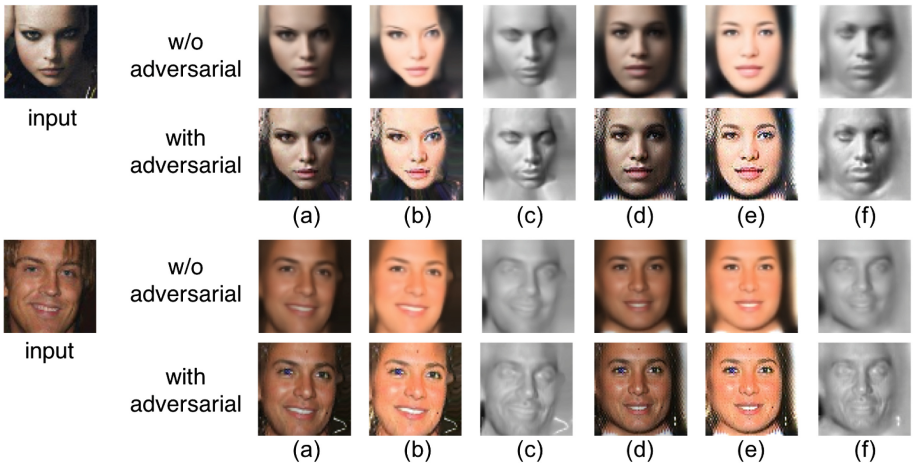

Fig. 11. Intrinsic-DAE with an adversarial loss: (a/d) reconstruction (b/e) albedo, (c/f) shading, in image and template coordinates, respectively. Adding an adversarial loss visually improves the image reconstruction quality of Intrinsic-DAE, while deformation, albedo, and shading can still be successfully disentangled.

Table 1. Landmark localization performance by different types of deformation modeling methods and different training corpus. A indicates affine transformation, I indicates non-rigid transformation by integration, whereas MAFL and CelebA denotes the training set. From columns 1 to 4, we manually annotate landmarks on the average texture, while for column 5, we train a regressor on the deformation fields to predict them. Latent vectors are 32D in these experiments.

A, MAFL	I, MAFL	$A + I$, MAFL	$A + I$, CelebA	$A + I$, CelebA, with regressor
14.13	9.89	8.50	7.54	5.96

10-D output layer to predict the spatial coordinates ((x, y)) for five landmarks. We use L1 loss as the objective function for regression.

We report the mean error in landmark localization as a percentage of the inter-ocular distance on the MAFL testing set (Tables 1 and 2). As the deformation field determines the alignment in the texture space, it serves as an effective mapping between landmark locations on the aligned texture and those on the original, unaligned faces. Hence, the mean error we report directly quantifies the quality of the (unsupervised) face alignment. In Table 2 we compare with previous state-of-the-art self-supervised image registration [31]. We observe that by better modeling of the deformation space we quickly bridge the gap in performance, even though we never explicitly trained to learn correspondences.

Table 2. Mean error on unsupervised landmark detection on the MAFL test set. Under DAE and Dense-DAE we specify the size of each latent vector. *NR* signifies training without regularization on the estimated deformations, while *Res* signifies training by estimating the residual deformation instead of the integral. Our results outperform the self-supervised method of [31] trained specifically for establishing correspondences.

DAE						Dense-DAE			TCDCN [42]	Thewlis et al. [31]
32-NR	32-Res	16	32	64	96	16	64	96		
10.24	9.93	5.71	5.96	5.70	6.46	6.85	5.50	**5.45**	7.95	5.83

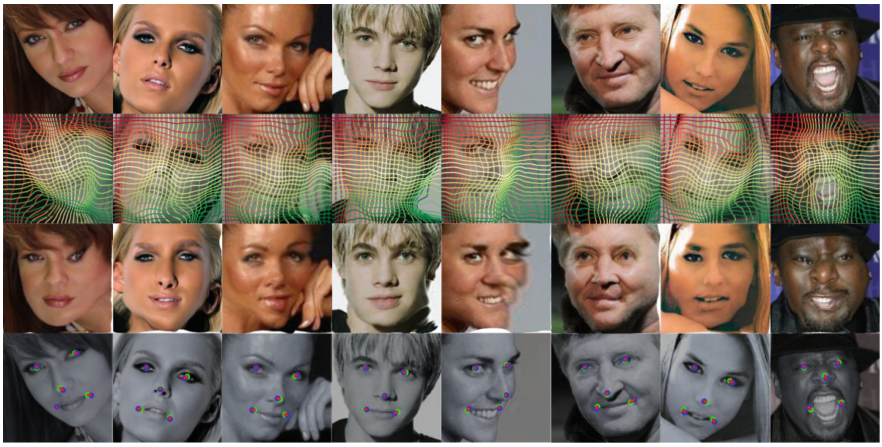

Fig. 12. Row 1: testing images; row 2: estimated deformation grid; row 3: image reverse-transformed to texture space; row 4: semantic landmark locations (green: ground truth, blue: estimation, red: error). (Color figure online)

4 Conclusion and Future Work

In this paper we have developed deforming autoencoders to disentangle shape and appearance in a learned latent representation space. We have shown that this method can be used for unsupervised groupwise image alignment. Our experiments with expression morphing in humans, image manipulation, such as shape and appearance interpolation, as well as unsupervised landmark localization, show the generality of our approach. We have also shown that bringing images in a canonical coordinate system allows for a more extensive form of image disentangling, facilitating the estimation of decompositions into shape, albedo and shading without any form of supervision. We expect that this will lead in the future to a full-fledged disentanglement into normals, illumination, and 3D geometry.

Acknowledgment. This work was supported by a gift from Adobe, NSF grants CNS-1718014 and DMS 1737876, the Partner University Fund, and the SUNY2020 Infrastructure Transportation Security Center. Rıza Alp Güler was supported by the European Horizons 2020 grant no 643666 (I-Support).

References

1. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In: NIPS (2016)
2. Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D.: Neural face editing with intrinsic image disentangling. In: CVPR (2017)
3. Worrall, D.E., Garbin, S.J., Turmukhambetov, D., Brostow, G.J.: Interpretable transformations with encoder-decoder networks. In: CVPR (2017)
4. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.: SfSNet: learning shape, reflectance and illuminance of faces in the wild. arXiv preprint arXiv:1712.01261 (2017)
5. Memisevic, R., Hinton, G.E.: Learning to represent spatial transformations with factored higher-order Boltzmann machines. Neural Comput. **22**, 1473–1492 (2010)
6. Worrall, D.E., Garbin, S.J., Turmukhambetov, D., Brostow, G.J.: Harmonic networks: deep translation and rotation equivariance (2016)
7. Park, E., Yang, J., Yumer, E., Ceylan, D., Berg, A.C.: Transformation-grounded image generation network for novel 3D view synthesis. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 702–711. IEEE (2017)
8. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.: Fader networks: manipulating images by sliding attributes. CoRR abs/1706.00409 (2017)
9. Edwards, G.J., Cootes, T.F., Taylor, C.J.: Face recognition using active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 581–595. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0054766
10. Matthews, I., Baker, S.: Active appearance models revisited. IJCV **60**, 135–164 (2004)
11. Learned-Miller, E.G.: Data driven image models through continuous joint alignment. PAMI **28**, 236–250 (2006)
12. Kokkinos, I., Yuille, A.L.: Unsupervised learning of object deformation models. In: ICCV (2007)
13. Frey, B.J., Jojic, N.: Transformation-invariant clustering using the EM algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **25**(1), 1–17 (2003)
14. Jojic, N., Frey, B.J., Kannan, A.: Epitomic analysis of appearance and shape. In: 9th IEEE International Conference on Computer Vision (ICCV 2003), 14–17 October 2003, Nice, France, pp. 34–43 (2003)
15. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. CoRR abs/1506.02025 (2015)
16. Papandreou, G., Kokkinos, I., Savalle, P.: Modeling local and global deformations in deep learning: epitomic convolution, multiple instance learning, and sliding window detection. In: CVPR (2015)
17. Dai, J., et al.: Deformable convolutional networks. In: ICCV (2017)
18. Neverova, N., Kokkinos, I.: Mass displacement networks. Arxiv (2017)

19. Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: a recurrent process applied for end-to-end face alignment. In: Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (2016)
20. Güler, R.A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I.: DenseReg: fully convolutional dense shape regression in-the-wild. In: CVPR (2017)
21. Cole, F., Belanger, D., Krishnan, D., Sarna, A., Mosseri, I., Freeman, W.T.: Face synthesis from facial identity features (2018)
22. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.W.: SfSNet : learning shape, reflectance and illuminance of faces in the wild. In: CVPR (2018)
23. Hinton, G.E.: A parallel computation that assigns canonical object-based frames of reference. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence, IJCAI 1981, 24–28 August 1981, Vancouver, BC, Canada, pp. 683–685(1981)
24. Olshausen, B.A., Anderson, C.H., Essen, D.C.V.: A multiscale dynamic routing circuit for forming size- and position-invariant object representations. J. Comput. Neurosci. **2**(1), 45–62 (1995)
25. Malsburg, C.: The correlation theory of brain function. Internal Report 81–2. Gottingen Max-Planck-Institute for Biophysical Chemistry (1981)
26. Hinton, G.E., Krizhevsky, A., Wang, S.D.: Transforming auto-encoders. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6791, pp. 44–51. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21735-7_6
27. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. CoRR abs/1710.09829 (2017)
28. Bristow, H., Valmadre, J., Lucey, S.: Dense semantic correspondence where every pixel is a classifier. In: ICCV (2015)
29. Zhou, T., Krähenbühl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3D-guided cycle consistency. In: CVPR (2016)
30. Gaur, U., Manjunath, B.S.: Weakly supervised manifold learning for dense semantic object correspondence. In: ICCV (2017)
31. Thewlis, J., Bilen, H., Vedaldi, A.: Unsupervised object learning from dense equivariant image labelling (2017)
32. Amit, Y., Grenander, U., Piccioni, M.: Structural image restoration through deformable templates. J. Am. Stat. Assoc. **86**(414), 376–387 (1991)
33. Yuille, A.L.: Deformable templates for face recognition. J. Cogn. Neurosci. **3**(1), 59–70 (1991)
34. Blanz, V.T., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1063–1074 (2003)
35. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
36. Afifi, M.: Gender recognition and biometric identification using a large dataset of hand images. CoRR abs/1711.04322 (2017)
37. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 94–108. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_7
38. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)

39. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
40. Li, C., Wand, M.: Precomputed Real-time texture synthesis with Markovian generative adversarial networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 702–716. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_43
41. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arxiv (2016)
42. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Trans. Pattern Anal. Mach. Intell. **38**(5), 918–930 (2016)