Lifting AutoEncoders: Unsupervised Learning of a Fully-Disentangled 3D Morphable Model Using Deep Non-Rigid Structure From Motion

Mihir Sahasrabudhe*
CentraleSupélec
mihir.sahasrabudhe@ecp.fr

Zhixin Shu*
Stony Brook University
zhshu@cs.stonybrook.edu

Edward Bartrum
The Alan Turing Institute, UCL
ebartrum@turing.ac.uk

Rıza Alp Güler
Ariel AI, Imperial College
r.guler@imperial.ac.uk

Dimitris Samaras
Stony Brook University
samaras@cs.stonybrook.edu

Iasonas Kokkinos
Ariel AI, UCL
ucaciko@ucl.ac.uk

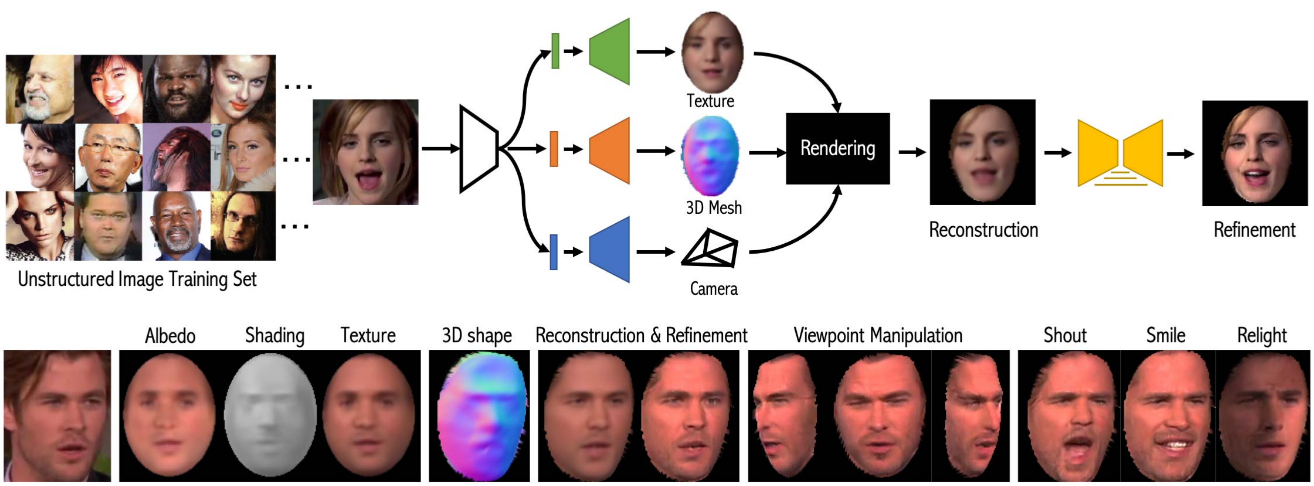

Figure 1: We introduce Lifting AutoEncoders, a deep generative model of 3D shape variability that is learned from an unstructured photo collection without supervision. Having access to 3D allows us to disentangle the effects of viewpoint, non-rigid shape (due to identity/expression), illumination and albedo and perform entirely controllable image synthesis.

Abstract

In this work we introduce Lifting Autoencoders, a generative 3D surface-based model of object categories. We bring together ideas from non-rigid structure from motion, image formation, and morphable models to learn a controllable, geometric model of 3D categories in an entirely unsupervised manner from an unstructured set of images. We exploit the 3D geometric nature of our model and use normal information to disentangle appearance into illumination, shading, and albedo. We further use weak supervision to disentangle the non-rigid shape variability of human faces into identity and expression. We combine the 3D representation with a differentiable renderer to generate RGB images and append an adversarially trained refinement network to obtain sharp, photorealistic image reconstruction results. The learned generative model can be controlled in terms of interpretable geometry and appearance factors, allowing us to perform photorealistic image manipulation of identity, expression, 3D pose, and illumination properties.

1. Introduction

Computer vision can be understood as the task of inverse graphics, namely the recovery of the scene that underlies an observed image. The scene factors that govern image formation primarily include surface geometry, camera position, material properties, and illumination. These are independent of each other but jointly determine the observed image intensities.

* indicates equal contribution.

In this work we incorporate these factors as disentangled variables in a deep generative model of an object category and tackle the problem of recovering all of them in an entirely unsupervised manner. We integrate in our network design ideas from classical computer vision, including structure-from-motion, spherical harmonic models of illumination and deformable models, and recover the three-dimensional geometry of a deformable object category in an entirely unsupervised manner from an unstructured collection of RGB images. We focus in particular on human faces and show that we can learn a three-dimensional morphable model of face geometry and appearance without access to any 3D training data, or manual labels. We further show that by using weak supervision we can further disentangle identity and expression, leading to even more controllable 3D generative models.

We first introduce Lifting AutoEncoders (LAEs) to recover, and then exploit the underlying 3D geometry of an object category by interpreting the outputs of a Deforming AutoEncoder (DAE) [36] in terms of a 3D representation. For this, we train a network to minimize a non-rigid SfM objective, which results is a low-dimensional morphable model of 3D shape, coupled with an estimate of the camera parameters. The resulting 3D reconstruction is coupled with a differentiable renderer [21] that propagates information from a 3D mesh to a 2D image, yielding a generative model for images that can be used for both image reconstruction and manipulation.

Our second contribution consists in exploiting the 3D nature of our novel generative model to further disentangle the image formation process. This is done in two complementary ways. For illumination modeling we use the 3D model to render normal maps and then shading images, which are combined with albedo maps to synthesize appearance. The resulting generative model incorporates spherical-harmonics-based [54, 47, 48] modeling of image formation, while still being end-to-end differentiable and controllable. For shape modeling we use sources of weak supervision to factor the shape variability into 3D pose, and non-rigid identity and expression, allowing us to control the expression or identity of a face by working with the appropriate latent variable code. Finally, we combine our reconstruction-driven architecture with an adversarially trained refinement network which allows us to generate photo-realistic images as its output.

As a result of these advances, we have a deep generative model that uses 3D geometry to model shape variability and provides us with a clearly disentangled representation of 3D shape in terms of identity, expression and camera pose and appearance in terms of albedo and illumination/shading. We report quantitative results on a 3D landmark localization task and show multiple qualitative results of controllable photorealistic image generation.

2. Previous work

The task of disentangling deep models can be understood as splitting the latent space of a network into independent sources of variation. In the case of learning generative models for computer vision, this amounts to uncovering the independent factors that contribute to image formation. This can both simplify learning, by injecting inductive biases about the data generation process, and can also lead to interpretable models that can be controlled by humans in terms of a limited number of degrees of freedom. This would for instance allow computer graphics to benefit from the advances in the learning of generative models.

Over the past few years rapid progress has been made in the direction of disentangling the latent space of deep models into dimensions that account for generic factors of variation, such as identity and low-dimensional transformations [8, 52, 28, 51, 39], or even non-rigid, dense deformations from appearance [57, 12, 41, 36, 49]. Several of these techniques have made it into some of the most compelling photorealistic, controllable generative models of object categories [32, 20].

Despite these advances, the disentanglement of the three-dimensional world geometry from the remaining aspects of image formation still remains very recent in deep learning. Effectively all works addressing aspects related to 3D geometry rely on paired data for training, e.g. multiple views of the same object [45], videos [29] or some pre-existing 3D mesh representation that is the starting point for further disentanglement [13, 35, 53, 40] or self-supervision [56]. This, however, leaves open the question of how one can learn about the three-dimensional world simply by observing a set of unstructured images. Very recently, a few works have started tackling the problem of recovering the three-dimensional geometry of objects from more limited information—[18] use keypoints and masks to learn a 3D deformable model of birds, [44] use keypoints and semi-supervised pretraining using a 3DMM, while [50] use a 3D geometry-based reprojection loss and correspondences of object instances during training.

Even though these works present exciting progress in the direction of deep 3D reconstruction, they fall short of providing us with a model that operates like a full-blown rendering pipeline. By contrast in our work, we propose for the first time a deep learning-based method that recovers a three-dimensional, surface-based, deformable template of an object category from an unorganized set of images, leading to controllable photorealistic image synthesis.

We do so by relying on Non-Rigid Structure from Motion (NRSfM). NRSfM, developed originally to establish a 3D model of a deformable object by observing its motion [4], was developed to solve increasingly accurately the underlying mathematical optimization problems [43, 30, 1, 9], extending to dense reconstruction [11], lifting object cat-

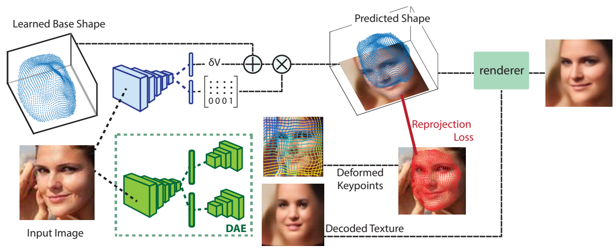

Figure 2: Lifting AutoEncoders (LAE) bring Non-Rigid Structure from Motion (NRSfM) into the problem of learning disentangled generative models for object categories. We start from a Deforming-AutoEncoder (DAE) that produces dense correspondences between an observed and a template image, we train an LAE by minimising an NRSfM-based reprojection objective.

egories from keypoints and masks [7, 18], incorporating spatio-temporal priors [38] and illumination models [27], while leading to impressively high-resolution 3D Reconstruction results [14, 27, 16, 22, 24, 23]. In [25] it has recently been proposed to represent non-rigid variability in terms of a deep architecture - but still the work relies on given point correspondences between instances of the same category. We now show this is no longer necessary - we delegate the task of establishing correspondences across image pixels of multiple images to a Deforming AutoEncoder [36] and proceed to lifting images through an end-to-end trainable deep network as we now describe.

3. Lifting AutoEncoders

We start by briefly describing Deforming AutoEncoders, as these are the starting point of our work. We then turn to our novel contributions of 3D lifting in Sec. 3.2 and shape disentanglement in Sec. 4.2.

3.1. DAEs: from image collections to deformations

Deforming Autoencoders, introduced in [36], and shown in Fig. 2, follow the deformable template paradigm and model image generation through a combination of appearance synthesis in a canonical coordinate system and a spatial deformation that warps the appearance (or, 'texture') to the observed image coordinates. The resulting model disentangles shape and appearance in an entirely unsupervised manner and also provides dense correspondes between images and the learnt template, while using solely an image reconstruction loss for training. Further details for training DAEs are provided in [36].

3.2. LAEs: 3D structure-from-deformations

We now turn to the problem of recovering the 3D geometry of an object category from an unstructured set of images. For this, we rely on DAEs to identify corresponding points across this image set and address our problem by training a network to minimize an objective function that is inspired from Non-Rigid Structure from Motion (NRSfM). Our central observation is that DAEs provide us with an image representation on which NRSfM optimization objectives can be easily applied. In particular, disentangling appearance and deformation labels all image positions that correspond to a single template point with a common, discovered UV value. LAEs take this a step further and interpret the DAE's UV decoding outputs as indicating the positions where an underlying 3D object surface position projects to the image plane. The task of an LAE is then to infer a 3D structure that can successfully project to all of the observed 2D points.

Given that we want to handle a deformable, non-rigid object category, we introduce a loss function that is inspired from Non-Rigid Structure from Motion, and optimized with respect to it. The variables involved in the optimization include (a) the statistical 3D shape representation, represented in terms of a linear basis (b) the per-instance expansion coefficients on this basis and (c) the per-instance 3D camera parameters. We note that in standard NRSfM all of the observations come from a common instance that is observed in time - by constrast in our case every training sample stems from a different instance of the same category, and it is only thanks to the DAE-based preprocessing that these distinct instances become commensurate.

3.3. 3D Lifting Objective

Our 3D structure inference task amounts to the recovery of a surface model that maps an intrinsic coordinate space (u, v) to 3D coordinates: $S(u, v) \to \mathbb{R}^3$. Even though the underlying model is continuous, our implementation is discrete: we consider a set of 2D points sampled uniformly on a cartesian grid in intrinsic coordinates,

$$\mathcal{S}_i = S(u_i, v_i),\ (u_i, v_i) \in D \times D\,, \tag{1}$$

$$\text{with } D = \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right\},\ i = 1, \ldots, N = (n+1)^2\,, \tag{2}$$

where n determines the spatial resolution at which we discretize the surface. We parameterize the three-dimensional position of these vertices in terms of a low-dimensional linear model, that captures the dominant modes of variation around a mean shape $\mathbb{B}^0$,

$$\mathcal{S}_i = \mathbb{B}^0 + \sum_{s=1}^{S} s_s \mathbb{B}_i^s\,. \tag{3}$$

In morphable models [46, 3] the mean shape and deformation basis elements are learned by PCA on a set of aligned 3D shapes, but in our case, we discover them from 2D by solving an NRSfM minimization problem that involves the projection to an unknown camera viewpoint.

In particular, we consider scaled orthographic projection Π through a camera described by a rotation matrix R and translation vector t. Under this assumption, the 3D surface points project to the points $\mathbf{x}_i$, given by

$$\mathbf{x}_i = \Pi\left[R\mathcal{S}_i\right] + t\,, \quad \Pi = \left[\begin{array}{ccc} \sigma & 0 & 0 \\ 0 & \sigma & 0 \end{array}\right], \tag{4}$$

where σ defines a global scaling.

We measure the quality of a 3D reconstruction in terms of the Euclidean distance of the predicted projection of a 3D point and its actual position in the image. In our case a 3D point $\mathcal{S}_i$ is associated with surface coordinate (u_i, v_i), we therefore penalize its distance from the image position $\hat{x}_i$ that the DAE's deformation decoder labels as u_i, v_i—

$$\hat{\mathbf{x}}_i = \hat{\mathbf{x}} : \mathrm{argmin}_{\mathbf{x}} \|DAE(\mathbf{x}) - (u_i, v_i)\|_2\,. \tag{5}$$

In practice, we find $\mathbf{x}$ by warping (u_i, v_i) under the DAE's deformation grid, W, and locating the point in the image coordinates that it warps to. If we can find no such point in the image coordinates, we deem the point *invisible*, setting a visibility variable ν_i to zero. We treat $\hat{\mathbf{x}}$ and ν as data terms, which specify the constraints that our learned 3D model must meet: the 3D points $\mathcal{S}_i$ must project to points $\mathbf{x}_i$ that lie close to their visible 2D counterparts, $\hat{\mathbf{x}}_i$. We express this *reprojection objective* in terms of the remaining variables,

$$L(R, t, \sigma, s, \mathbb{B}) = \sum_{i=1}^{N} \nu_i \|\hat{\mathbf{x}}_i - \mathbf{x}_i(R, t, \sigma, \mathcal{S}, s)\|_2\,, \tag{6}$$

where we have expressed $\mathbf{x}_i$ as a differentiable function of $R, t, \mathcal{S}, s$ through Eq. 4 and Eq. 3.

For a set of K images we have different camera and shape parameters $(R^k, t^k), s^k, k = 1, \ldots, K$ since we consider a non-rigid object seen from different viewpoints. The basis elements $\mathbb{B}$ are however considered to be common across all images, since they describe the inherent shape variability of the whole category. Our 3D non-rigid reconstruction problem thus becomes

$$\mathcal{L}_{3D} = \sum_{k=1}^{K} L(R^k, t^k, \sigma^k, s^k, \mathbb{B})\,. \tag{7}$$

3.4. LAE learning via Deep NRSfM

Minimizing the objective of Eq. 7 amounts to the common Non-Rigid Structure-from-Motion objective [4, 43, 30, 1, 9]. Even though highly efficient and scalable algorithms have been proposed for its minimization, we would only consider them for initialization, since we want 3D Lifting to be a component of a larger deep generative model of images. We do not use any such technique, in order to simplify our model's training, implementing it as a single deep network training process.

The approach we take is to handle the shape basis $\mathbb{B}$ as the parameters of a linear 'morphable' layer, tasked with learning the shape model for our object category. We train this layer in tandem with complementary, multi-layer network branches that regress from the image to (a) the expansion coefficients s^k, (b) the Euler angles/rotation matrix R^k, and (c) the displacement vector t^k describing the camera position. These are components of a larger deep network that can learn to reconstruct an image in 3D - a task we refer to as Deep NRSfM.

If we only train a network to optimize this objective, we obtain a network that can interpret a given image in terms of its 3D geometry, as expressed by the 3D camera position (rigid pose) and the instance-specific expansion coefficients (non-rigid shape). Having established this, we can conclude the task of image synthesis by projecting the 3D surface back to 2D. For this, we combine the 3D lifting network with a differentiable renderer [21], and bring the synthesized texture image in correspondence with the image coordinates. The resulting network is an end-to-end trainable pipeline for image generation that passes through a full-blown, 3D reconstruction process.

Having established a controllable, 3D-based rendering pipeline, we turn to photorealistic synthesis. For this, we further refine the rendered image by a U-Net [33] architecture that takes as input the reconstructed image and augments the visual plausibility. This refinement module is trained using two losses, firstly an L_2 loss to reconstruct the input image and secondly an adversarial loss to provide photorealism. The results of this module are demonstrated in Figure 5 - we see that while keeping intact the image generation process, we achieve a substantially more realistic synthesis.

4. Geometry-Based Disentanglement

In this section we show that having access to the underlying 3D scene behind an image allows to further decompose the image generation into distinct, controllable sub-models, in the same way that one would do within a graphics engine. We first describe in Sec. 4.1 how surface-based normal estimation allows us to disentangle appearance into albedo and shading using a physics-based model of illumination. In Sec. 4.2 we then turn to learning a more fine-grained model of 3D shape and use weak supervision to disentangle per-instance non-rigid shape into expression and identity.

4.1. LAE-lux: Disentangling Shading and Albedo

As in several recent works [37, 35] we consider a Lambertian reflectance model for human faces and adopt the Spherical Harmonic model to model the effects of illumination on appearance [54, 47, 48]. We pursue the intrinsic

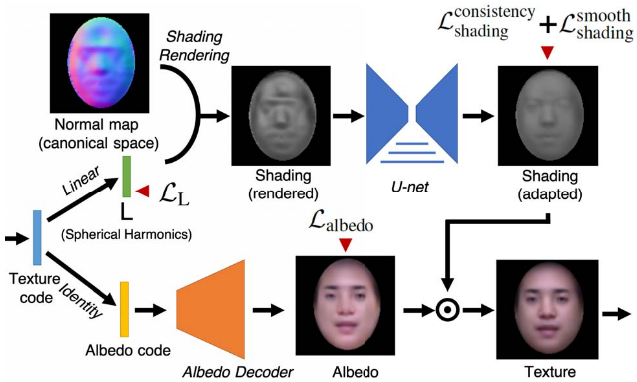

Figure 3: Texture decoder for LAE-lux: disentangling albedo and illumination with 3D shape and Spherical Harmonics representation for illumination.

decomposition [2] of the canonical texture T into albedo, A and shading, S:

$$T = S \odot A \tag{8}$$

where $\odot$ denotes Hadamard product, by constraining the shading image to be connected to the normals delivered by the LAE surface. In particular, denoting by L the representation of the scene-specific spherical harmonic illumination vector, and by $H(N(x))$ the representation of the local normal field $N(x)$ on the first 9 spherical harmonic coefficients, we consider that the local shading, $S(x)$ is expressed as an inner product, $S(x) = \langle L, H(N(x)) \rangle$. As such the shading field can be obtained by a linear layer that is driven by regressed illumination coefficients L and the surface-based harmonic field, $H(N(x))$. Given $S(x)$, the texture can then be obtained from albedo and shading images according to Eq. 8.

In practice, the normal field we estimate is not as detailed as would be needed, e.g. to capture sharp corners, while the illumination coefficients can be inaccurate. To compensate for this, we first render an estimate of the shading S^{render} with spherical harmonics parameters L and normal maps $N$and then use a U-Net to refine it, obtaining S^{adapted}.

Given that the shading-albedo decomposition is an ill-posed problem, we further use a combination of losses that capture increasingly detailed prior knowledge about the desired solution. First, as in [37] we employ intrinsic image-based smoothness losses on albedo and shading, $\mathcal{L}^{\text{smooth}}_{\text{shading}} = \lambda_{\text{shade}} \left\| \nabla S^{\text{adapted}} \right\|_2$, and $\mathcal{L}_{\text{albedo}} = \lambda_{\text{albedo}} \left\| \nabla A \right\|_1$, where ∇ represents the spatial gradient, which means that we allow the albedo to have sharp discontinuities, while the shading image should have mostly smooth variations [34]. In our experiment, we set $\lambda_{\text{shade}} = 1{\times}10^{-4}$ and $\lambda_{\text{albedo}} = 2{\times}10^{-6}$. Second, we compute a deterministic estimate $\hat{L}$ of the illumination parameters and penalize its distance to the regressed illumination values, $\mathcal{L}_{\text{L}} = \left\| L - \hat{L} \right\|_2$. More specifically, $\hat{L}$ is based on the crude assumption that the face's albedo is constant, $\hat{A}(x) = 0.5$, where we treat albedo as a grayscale. Even though clearly very rough, this assumption captures the fact that a face is largely uniform, and allows us to compute a proxy to the shading in terms of $\hat{S} = T \oslash \hat{A}$ where $\oslash$ denotes Hadamard division. We subsequently compute the approximation $\hat{L}$ from $\hat{S}$ and the harmonic field $H(N)$ using least squares. For face images, similar to [37], $\hat{L}$ serves as a reasonably rough approximation of the illumination coefficient and is used for weak supervision via $\mathcal{L}_{\text{L}}$.

Finally, the shading consistency loss regularizes the U-Net, and is designed to encourage the U-Net based adapted shading S^{adapted} to be consistent with the shading rendered from the spherical harmonics representation S^{rendered}—

$$\mathcal{L}^{\text{consistency}}_{\text{shading}} = \text{Huber}(S^{\text{adapted}}, S^{\text{rendered}}), \tag{9}$$

where we use Huber loss for a robust regression since S^{rendered} can contain some outlier pixels due to an imperfect 3D shape.

4.2. Disentangling Expression, Identity and Pose

We consider that a face shape as observed in an image is the composite effect of camera pose, identity, and expression. Without some guidance, the parameters controlling shape can be mixed - for instance accounting for the effects of camera rotation through non-rigid deformations of the face.

For a given identity we can understand expression-based shape variability in terms of deviation from a neutral pose. We can consider that a reasonable approximation to this consists in using a separate linear basis $\mathbb{B}^I$ for identity and another for expression $\mathbb{B}^E$, which amounts to following model—

$$\mathcal{S}_i(s^I, s^E) = \mathbb{B}^0_i + \sum_{s=1}^{I} s^I_s \mathbb{B}^{I,s}_i + \sum_{s=1}^{E} s^E_s \mathbb{B}^{E,s}_i \,. \tag{10}$$

Even though the model is still linear and is at first sight equivalent, clearly separating the two subspaces means that we can control them through side information—for instance, by imposing specific losses on these subspaces depending on the property. Here we use the MultiPIE[15] dataset to help disentangle the latent representation of a person's identity, facial expression, and pose (camera). MultiPIE is captured under a controlled environment and contains image pairs acquired under identical conditions with differences only in (1) facial expression, (2) camera position, and (3) illumination conditions.

We use facial expression disentangling as an example, and follow a similar procedure for pose and camera disentangling. Given an image I_{exp} with known expression exp, we sample two more images from the dataset. The first, I^{+}_{exp}

has the same facial expression but different identity, pose, and illumination conditions. The second, I^-_{exp}, has a different facial expression but the same identity, pose and illumination condition as I_{exp}. We use siamese training to encourage I_{exp} and I^+_{exp} to have similar latent representations for facial expression, and a triplet loss to ensure that I_{exp} and I^+_{exp} are closer in expression space than I_{exp} and I^-_{exp}:

$$\begin{aligned}
\mathcal{L}_{\text{expression}} &= \mathcal{L}^{\text{similarity}}_{\text{expression}} + \mathcal{L}^{\text{triplet}}_{\text{expression}}, \text{ where} && (11)\\
\mathcal{L}^{\text{similarity}}_{\text{expression}} &= \left\| f_{\text{exp}}(I_{\text{exp}}) - f_{\text{exp}}(I^+_{\text{exp}}) \right\|_2, && (12)\\
\mathcal{L}^{\text{triplet}}_{\text{expression}} &= \max(0, 1 + \left\| f_{\text{exp}}(I_{\text{exp}}) - f_{\text{exp}}(I^+_{\text{exp}}) \right\|_2 \\
&\quad - \left\| f_{\text{exp}}(I_{\text{exp}}) - f_{\text{exp}}(I^-_{\text{exp}}) \right\|_2). && (13)
\end{aligned}$$

Following a similar collection of triplets for the remaining sources of variability, we disentangle the latent code for shape in terms of camera pose, identity, and expression. With MultiPIE, the overall disentanglement objective for shape is hence

$$\mathcal{L}_{\text{disentangle}} = \mathcal{L}_{\text{expression}} + \mathcal{L}_{\text{identity}} + \mathcal{L}_{\text{pose}}, \tag{14}$$

where $\mathcal{L}_{\text{identity}}$ and $\mathcal{L}_{\text{pose}}$ are defined similarly to $\mathcal{L}_{\text{expression}}$. In our experiments, we used the scaling parameter for this loss, $\lambda_{\text{disentangle}} = 1$.

4.3. Complete Objective

We further control the model learning with a regularization loss given by

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{scale}} \sum_{k=1}^{K} \|\sigma_k\|_2 + \lambda_{\text{shape}} \sum_{k=1}^{K} \left\| \sum_{s=1}^{S} s^k_s \mathbb{B}^s \right\|_2, \tag{15}$$

where σ_k is the scaling parameter in Eq. 4 and $\sum_{s=1}^{S} s_s \mathbb{B}^s_i$ is the non-rigid deviation from the mean shape, $\mathbb{B}^0$. We use $\lambda_{\text{scale}} = 0.01$, and $\lambda_{\text{shape}} = 0.1$ in all our experiments.

Combining this with the reprojection loss, $\mathcal{L}_{3\text{D}}$, defined in Eq. 7, we can write the complete objective function, which is trained end-to-end:

$$\begin{aligned}
\mathcal{L}_{\text{total}} = \; &\lambda_{3\text{D}} \cdot \mathcal{L}_{3\text{D}} + \lambda_{\text{disentangle}} \cdot \mathcal{L}_{\text{disentangle}} + \\
&\lambda_{\text{scale}} \cdot \mathcal{L}_{\text{scale}} + \lambda_{\text{shape}} \cdot \mathcal{L}_{\text{shape}}.
\end{aligned} \tag{16}$$

In our experiments, we used the scaling factor for the 3D reprojection loss, $\lambda_{3\text{D}} = 50$. This relatively high scaling factor was chosen so that the reprojection loss is not overpowered by other losses at later training iterations. A similar hyperparameter setting was also used by the authors of [18].

For training the LAE-Lux, we also add the albedo-shading disentanglement losses, summarised by

$$\mathcal{L}_{\text{lux}} = \mathcal{L}^{\text{smooth}}_{\text{shading}} + \mathcal{L}^{\text{consistensy}}_{\text{shading}} + \mathcal{L}_{\text{albedo}} + \mathcal{L}_{\text{L}}. \tag{17}$$

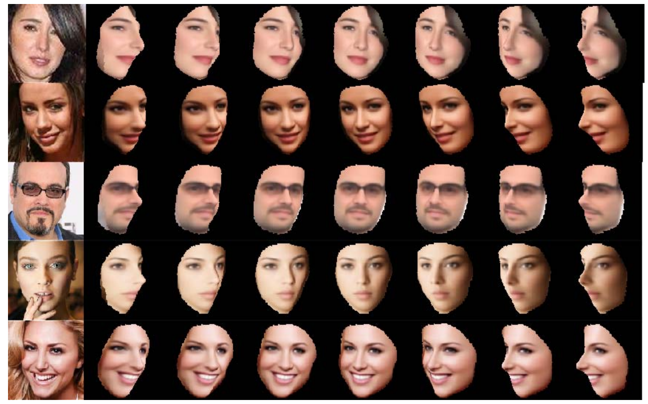

Figure 4: Visualizations from various yaw angles of the learned 3D shapes *without weak supervision*. Our reconstructions respect prominent face features, such as the nose, forehead and checks, allowing us to rotate an object reconstruction in 3D.

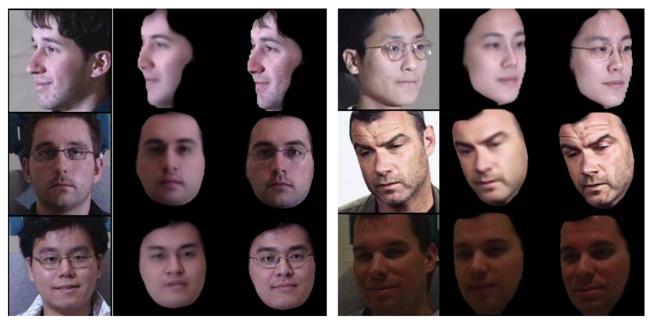

Figure 5: Reconstruction from learnt shape *trained with weak supervision* and photorealistic refinement. Starting from an image reconstruction by an LAE (centre), weak supervision learns a much better shape, and an adversarially-trained refinement network adds details to increase the photorealism of a face (right).

5. Experiments

5.1. Architectural Choices

Our encoder and decoder architectures are similar to the ones employed in [36], but working on images of size 128×128 pixels instead of 64×64. We use convolutional neural networks with five stridedConv-batchNorm-leakyReLU layers in image encoders, which regress the expansion coefficients ss. Image decoders consist similarly of five stridedDeconv-batchNorm-ReLU layers.

5.2. Datasets

We now note the face datasets that we used for our experiments. Certain among them contain side information, for instance, multiple views of the same person, or videos of the same person. This side information was used for expression-identity disentanglement experiments, but not for the 3D lifting part. For the reconstruction results, our algorithms were only provided with unstructured datasets unless otherwise noted.

1. **CelebA** [26]: This dataset contains about 200,000 in-the-wild images, and is one of the datasets we use to train our DAE. A subset of this dataset, MAFL [55], was also released which contains annotations for five facial landmarks. We use the training set of MAFL in our evaluation experiments, and report results on the test set. Further, as MAFL is a subset of CelebA, we removed the images in the MAFL test set from the CelebA training set before training the DAE.
2. **Multi-PIE** [15]: Multi-PIE contains images of 337 subjects of 7 facial expressions, each of which is captured under 15 viewpoints and 19 illumination conditions simultaneously.
3. **AFLW2000-3D** [58]: This dataset consists of 3D fitted faces for the first 2000 images of the AFLW dataset. In this paper, we employ it for evaluation of our learned shapes using 3D landmark localization errors.

5.3. Qualitative Results

We show examples of the learned 3D shapes in Figure 4. The figure also visualises reconstructed faces from various yaw angles using a model that was trained only on CelebA images. We see that the model learns a shape that expresses the input well, as well as captures prominent facial features. The reconstructions are, however, weak for side poses. This is further refined using weak supervision from the Multi-PIE dataset.

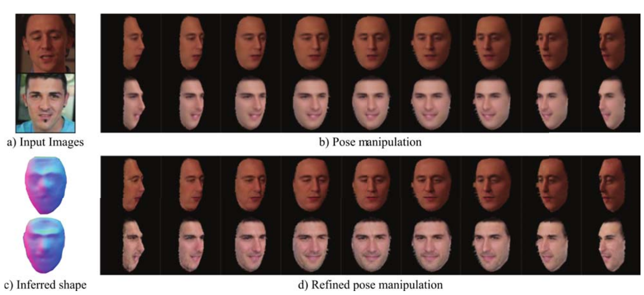

Figure 6: Changing Pose with LAE. Given input face image (a), LAE learns to recover the 3D shape (c), with which we can manipulate the pose of the faces (b). With the additional refinement network, we can enhance the manipulated face image by adding facial details (d) that better preserve the characteristic features of the input faces.

5.4. Face manipulation results

In this section, we show some results of manipulating the expression and pose latent spaces. In Figure 6 (b), we visualize the decoded 3D shape from input images in 6 (a) from various camera angles. Furthermore, in Figure 6 (d), we show results after passing the visualizations in Figure 6 (b) through the refinement network.

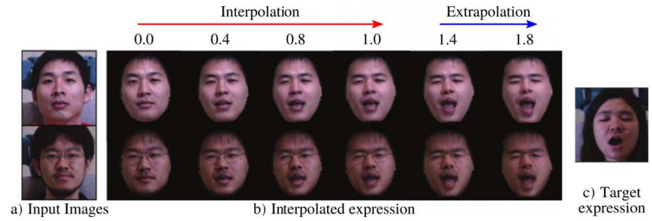

Figure 7: Changing expression with LAE. We interpolate between source (a) and target (c) images and show results of synthesizing new faces (b). We also show the result of extrapolation.

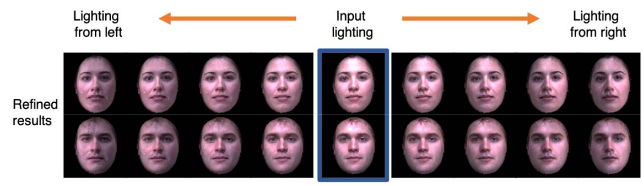

Figure 8: Lighting manipulation using LAE-lux. We demonstrate the result of synthesizing new images by moving the lighting direction.

Similarly, in Figure 7, we interpolate over the expression latent space from each of the images in (a) to the image in (c) and visualize the shape at each intermediate step in Figure (b).

Method	NME
Thewlis *et al*. (2017) [42]	6.67
Thewlis *et al*. (2018) [41]	5.83
Jakub *et al*. (2018) [17]	2.54
Shu *et al*. (2018), DAE, no regressor [36]	7.54
Shu *et al*. (2018), DAE, with regressor [36]	5.45
LAE, CelebA (no regressor)	7.96
LAE, CelebA (with regressor)	6.01

Table 1: 2D landmark localization results for the proposed LAEs compared with other state-of-the-art approaches. All numbers signify the average error per landmark normalized by the inter-ocular distance, over the entire dataset.

5.5. Quantitative Analysis: Landmark Localization

We evaluate our approach quantitatively in terms of landmark localization. Specifically, we evaluate on two datasets—the MAFL test set for 2D landmarks, and the AFLW2000-3D for 3D shape. In both cases, as we do not train with ground-truth landmarks, we manually annotate,

Method	Rotation	Yaw angle			
		$[0, 30]$	$(30, 60]$	$(60, 90]$	All
3DDFA [58] (supervised)	Y	4.25 ± 0.95	4.34 ± 1.04	4.39 ± 1.35	4.28 ± 1.03
	N	12.51 ± 6.40	23.20 ± 5.92	32.55 ± 3.85	17.31 ± 9.30
PRNet [10] (supervised)	Y	4.88 ± 1.24	6.94 ± 2.83	10.51 ± 5.31	6.01 ± 3.08
	N	7.17 ± 3.45	10.96 ± 5.00	16.34 ± 8.91	9.11 ± 5.66
3D-FAN [6] (supervised)	Y	2.73 ± 1.38	2.48 ± 2.24	3.74 ± 2.95	2.84 ± 1.92
	N	7.51 ± 2.21	7.06 ± 3.94	8.75 ± 4.53	7.61 ± 3.10
LAE (64) CelebA	Y	6.86 ± 1.07	9.01 ± 1.07	10.91 ± 1.37	7.89 ± 1.89
	N	9.29 ± 4.90	20.98 ± 7.74	37.62 ± 7.50	15.85 ± 11.89
LAE (128) CelebA	Y	6.02 ± 1.04	7.91 ± 1.04	9.58 ± 1.32	6.92 ± 1.73
	N	8.41 ± 4.96	19.56 ± 7.97	36.31 ± 7.78	14.80 ± 11.80
LAE (128) MultiPIE	Y	6.85 ± 0.85	7.94 ± 0.97	9.02 ± 1.26	7.39 ± 1.25
	N	9.80 ± 4.88	13.87 ± 6.51	24.19 ± 8.72	12.78 ± 7.83
LAE (128) CelebA+MultiPIE	Y	6.83 ± 0.96	8.41 ± 1.15	9.83 ± 1.65	7.59 ± 1.60
	N	9.11 ± 4.54	13.60 ± 6.08	24.62 ± 8.37	12.33 ± 7.84

Table 2: Mean 3D landmark localization errors, after Procrustes analysis, normalized by bounding box size and averaged over the entire AFLW2000-3D test set. The number in brackets for the LAEs refers to the dimension of the latent space for the rigid and non-rigid components of the deformable model. The second column specifies whether rotation is included in the Procrustes analysis. We also note the training dataset used for training each LAE.

only once, the necessary landmarks on the base shape as linear combinations of one or more mesh vertices. That is to say, each landmark location corresponds to a linear combination of the locations of several vertices.

We use five landmarks for the MAFL test set, namely the two eyes, the tip of the nose, and the ends of the mouth. Similarly to [42, 41, 36], we evaluate the extent to which landmarks are captured by our 3D shape model by training a linear regressor to predict them given the locations of the mesh vertices in 3D.

We observe from Table 1 that our system is able to perform at-par with the DAE, which is our starting model - and as such serves as the upper bound on the performance that we can attain. This shows that while being able to successfully perform the lifting operation, we do not sacrifice localization accuracy. The small increase in error can be attributed to the fact that perfect reconstruction of a system is nearly impossible with a low-dimensional shape model. Furthermore we use a feedforward, single-shot camera and shape regression network, while in principle this is a problem that could require iterative model fitting techniques to align a 3D deformable model to 2D landmarks [31].

We report localization results in 3D on 21 landmarks that feature in the AFLW2000-3D dataset. As our unsupervised system is often unable to locate human ears, the learned face model does not account for them in the UV space. This makes it impossible to evaluate landmark localization for points that lie on or near the ears, which is the case for two of these landmarks. Hence, for the AFLW2000-3D dataset, we report localization accuracies only for 19 landmarks. Furthermore, as an evaluation of the discovered shape, we also show landmark localization results after rigid alignment (without reflection) of the predicted landmarks with the ground truth. We perform Procrustes analysis, with and without adding rotation to the alignment, the latter giving us an evaluation of the accuracy of pose estimation as well.

Table 2 also demonstrates the gain achieved by adding weak supervision via the Multi-PIE dataset. We see that the mean NMEs for LAEs trained with and without the Multi-PIE dataset increase as the yaw angle increases. This is also visible in our qualitative results shown in Fig. 5, where we visualize the discovered shapes for both of these cases.

6. Conclusion

In this work we have introduced an unsupervised method for lifting an object category into a 3D representation, allowing us to learn a 3D morphable model of faces from an unorganized photo collection. We have shown that we can use the resulting model for controllable manipulation and editing of observed images.

Deep image-based generative models have shown the ability to deliver photorealistic synthesis results with substantially more diverse categories than faces [5, 19] - we anticipate that their combination with 3D representations like LAEs will further unleash their potential for controllable image synthesis.

Acknowledgement. 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.

References

[1] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Non-rigid structure from motion in trajectory space. In *Advances in neural information processing systems*, pages 41–48, 2009. 2, 4

[2] H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2:3–26, 1978. 5

[3] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 2018. 3

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *cvpr*, volume 2, page 2690. Citeseer, 2000. 2, 4

[5] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 8

[6] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 8

[7] J. Carreira, S. Vicente, L. Agapito, and J. Batista. Lifting object detection datasets into 3d. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1342–1355, 2016. 3

[8] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Neural Information Processing Systems*, 2016. 2

[9] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014. 2, 4

[10] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 8

[11] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1272–1279, 2013. 2

[12] U. Gaur and B. S. Manjunath. Weakly supervised manifold learning for dense semantic object correspondence. In *ICCV*, 2017. 2

[13] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, 2018. 2

[14] P. F. Gotardo, T. Simon, Y. Sheikh, and I. Matthews. Photogeometric scene flow for high-detail dynamic 3d reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 846–854, 2015. 3

[15] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image Vision Comput.*, 28(5):807–813, May 2010. 5, 7

[16] M. Hernandez, T. Hassner, J. Choi, and G. G. Medioni. Accurate 3d face reconstruction via prior constrained structure from motion. *Computers & Graphics*, 66:14–22, 2017. 3

[17] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4016–4027. Curran Associates, Inc., 2018. 7

[18] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2, 3, 6

[19] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 8

[20] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018. 2

[21] H. Kato, Y. Ushiku, and T. Harada. Neural 3d mesh renderer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4

[22] I. Kemelmacher-Shlizerman. Internet based morphable model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3256–3263, 2013. 3

[23] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *2011 International Conference on Computer Vision*, pages 1746–1753. IEEE, 2011. 3

[24] I. Kemelmacher-Shlizerman and S. M. Seitz. Collection flow. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1799. IEEE, 2012. 3

[25] C. Kong and S. Lucey. Deep interpretable non-rigid structure from motion. *CoRR*, abs/1902.10840, 2019. 3

[26] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 7

[27] Q. Liu-Yin, R. Yu, L. Agapito, A. Fitzgibbon, and C. Russell. Better together: Joint reasoning for non-rigid 3d reconstruction with specularities and shading. *arXiv preprint arXiv:1708.01654*, 2017. 3

[28] R. Memisevic and G. E. Hinton. Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural Computation*, 2010. 2

[29] D. Novotny, D. Larlus, and A. Vedaldi. Learning 3d object categories by looking around them. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5218–5227, 2017. 2

[30] M. Paladini, A. Del Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito. Factorization for non-rigid and articulated structure using metric projections. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2898–2905. IEEE, 2009. 2, 4

[31] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis. 6-dof object pose from semantic keypoints. In *ICRA*, 2017. 8

[32] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018. 2

[33] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[34] D. Samaras, D. Metaxas, P. Fua, and Y. G. Leclerc. Variable albedo surface reconstruction from stereo and shape from shading. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages I: 480–487, 2000. 5

[35] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet : Learning shape, reflectance and illuminance of faces in the wild. In *CVPR*, 2018. 2, 4

[36] Z. Shu, M. Sahasrabudhe, R. A. Güler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *European Conference on Computer Vision*, 2018. 2, 3, 6, 7, 8

[37] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 4, 5

[38] T. Simon, J. Valmadre, I. Matthews, and Y. Sheikh. Separable spatiotemporal priors for convex reconstruction of time-varying 3d point clouds. In *European Conference on Computer Vision*, pages 204–219. Springer, 2014. 3

[39] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018. 2

[40] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H. Seidel, P. Pérez, M. Zollhöfer, and C. Theobalt. FML: face model learning from videos. *CoRR*, abs/1812.07603, 2018. 2

[41] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NIPS*, 2017. 2, 7, 8

[42] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3229–3238, Oct 2017. 7, 8

[43] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):878–892, 2008. 2, 4

[44] L. Tran and X. Liu. Nonlinear 3d face morphable model. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Salt Lake City, UT, June 2018. 2

[45] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017. 2

[46] T. Vetter, M. J. Jones, and T. A. Poggio. A bootstrapping algorithm for learning linear models of object classes. In *1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), June 17-19, 1997, San Juan, Puerto Rico*, pages 40–46, 1997. 3

[47] Y. Wang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. Face re-lighting from a single image under harsh lighting conditions. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007. 2, 4

[48] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):1968 –1984, nov. 2009. 2, 4

[49] O. Wiles, A. Sophia Koepke, and A. Zisserman. X2face: A network for controlling face generation

using images, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–686, 2018. 2

[50] O. Wiles and A. Zisserman. 3d surface reconstruction by pointillism. In *ECCV Workshop on Geometry Meets Deep Learning*, 2018. 2

[51] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *CVPR*, 2016. 2

[52] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Interpretable transformations with encoder-decoder networks. In *CVPR*, 2017. 2

[53] S. Yao, T. M. Hsu, J.-Y. Zhu, J. Wu, A. Torralba, B. Freeman, and J. Tenenbaum. 3d-aware scene manipulation via inverse graphics. In *Advances in Neural Information Processing Systems*, pages 1891–1902, 2018. 2

[54] L. Zhang, S. Wang, and D. Samaras. Face synthesis and recognition under arbitrary unknown lighting using a spherical harmonic basis morphable model. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages II:209–216, 2005. 2, 4

[55] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision 2014*, 2014. 7

[56] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017. 2

[57] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *CVPR*, 2016. 2

[58] X. Zhu, Z. Lei, S. Z. Li, et al. Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 7, 8