Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Sister Conferences Best Papers Track

Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild (Extended Abstract)*

Shangzhe Wu, **Christian Rupprecht** and **Andrea Vedaldi**
University of Oxford
{szwu, chrisr, vedaldi}@@robots.ox.ac.uk

Abstract

We propose a method to learn 3D deformable object categories from raw single-view images, without external supervision. The method is based on an autoencoder that factors each input image into depth, albedo, viewpoint and illumination. In order to disentangle these components without supervision, we use the fact that many object categories have, at least approximately, a symmetric structure. We show that reasoning about illumination allows us to exploit the underlying object symmetry even if the appearance is not symmetric due to shading. Furthermore, we model objects that are probably, but not certainly, symmetric by predicting a symmetry probability map, learned end-to-end with the other components of the model. Our experiments show that this method can recover very accurately the 3D shape of human faces, cat faces and cars from single-view images, without any supervision or a prior shape model. Code and demo available at https://github.com/elliottwu/unsup3d.

1 Introduction

In this paper, we consider the problem of learning 3D models for deformable object categories. In particular, we study this problem under two challenging conditions. The *first* condition is that no 2D or 3D ground truth information (such as keypoints, segmentation, depth maps, or prior knowledge of a 3D model) is available. Learning without external supervisions removes the bottleneck of collecting image annotations, which is often a major obstacle to deploying deep learning for new applications. The *second* condition is that the algorithm must use an unconstrained collection of single-view images — in particular, it should not require multiple views of the same instance. Learning from single-view images is useful because in many applications, and especially for deformable objects, we only have a source of still images to work with. Consequently, our learning algorithm ingests a number of single-view images of a deformable object category and produces as output a deep network that can estimate the 3D shape of any instance given a single image of it, as illustrated in Fig. 1.

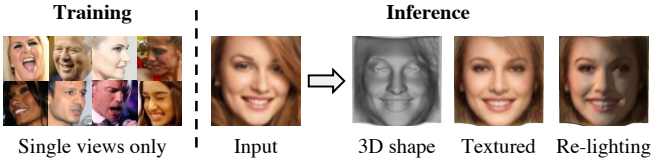

Figure 1: Left: Training requires *only* single views of the object category with *no* additional supervision at all. Right: Once trained, our model reconstructs the 3D pose, shape, albedo and illumination of a deformable object instance from a single image.

We formulate this as an autoencoder that internally decomposes the image into albedo, depth, illumination and viewpoint, *without direct supervision for any of these factors*. However, without further assumptions, decomposing images into these four factors is ill-posed. In search of minimal assumptions to achieve this, we note that many object categories are *symmetric* (*e.g.* almost all animals and many handcrafted objects). Assuming an object is perfectly symmetric, one can obtain a virtual second view of it by simply mirroring the image. In fact, if correspondences between the pair of mirrored images were available, 3D reconstruction could be achieved by stereo reconstruction [Mukherjee *et al.*, 1995; François *et al.*, 2003]. Motivated by this, we seek to leverage symmetry as a geometric cue to constrain the decomposition.

However, specific object instances are in practice never fully symmetric, neither in shape nor appearance. Shape is non-symmetric due to variations in pose or other details (*e.g.* hair style or expressions on a human face), and albedo can also be non-symmetric (*e.g.* asymmetric texture of cat faces). Even when both shape and albedo are symmetric, the appearance may still not be, due to asymmetric illumination.

We address this issue in two ways. First, we explicitly model illumination to exploit the underlying symmetry. Furthermore, we show that, by doing so, the model can exploit illumination as an additional cue for recovering the shape. Second, we augment the model to reason about potential lack of symmetry in the objects. To do this, the model predicts, along with the other factors, a probability map that each given pixel has a symmetric counterpart in the image.

We combine these elements in an end-to-end learning formulation, where all components, including the confidence

*This is an extended abstract of [Wu *et al.*, 2020] published at CVPR 2020. Please refer to full paper for more details.

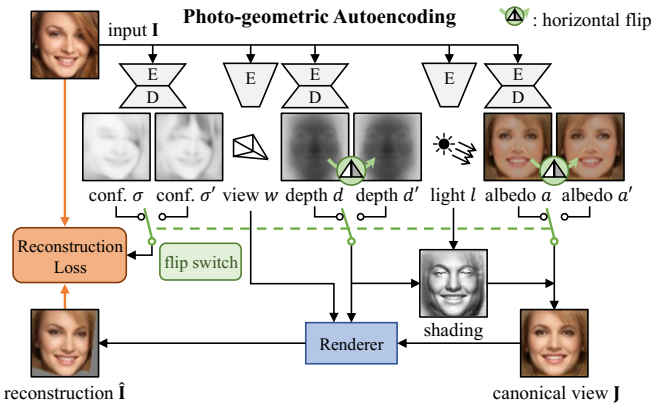

Figure 2: Our model decomposes an input image into depth, albedo, viewpoint and lighting, together with a pair of confidence maps. It is trained through reconstructing the input without external supervision.

maps, are learned from raw RGB data only. We also show that symmetry can be enforced by flipping internal representations, which is particularly useful for reasoning about symmetries probabilistically.

We demonstrate the quality of our method in several datasets, including human faces, cat faces and cars. We provide a thorough ablation study using a synthetic face dataset to obtain the necessary 3D ground truth. On real images, we achieve higher fidelity reconstruction results compared to other methods [Sahasrabudhe *et al.*, 2019; Szabó *et al.*, 2019] that do not rely on 2D or 3D ground truth information, nor prior knowledge of a 3D model of the instance or class. We also demonstrate that our trained face model generalizes to non-natural images such as face paintings and cartoon drawings without fine-tuning.

2 Method

Given an unconstrained collection of images of an object category, such as human faces, our goal is to learn a model Φ that receives as input an image of an object instance and produces as output a decomposition of it into 3D shape, albedo, illumination and viewpoint, as illustrated in Fig. 2.

As we have only raw images to learn from, the learning objective is reconstructive: namely, the model is trained so that the combination of the four factors gives back the input image. This results in an autoencoding pipeline where the factors have, due to the way they are recomposed, an explicit photo-geometric meaning.

2.1 Photo-Geometric Autoencoding

An image $\mathbf{I}$ is a function $\Omega \to \mathbb{R}^3$ defined on a grid $\Omega = \{0, \dots, W-1\} \times \{0, \dots, H-1\}$, or, equivalently, a tensor in $\mathbb{R}^{3\times W\times H}$. We assume that the image is roughly centered on an instance of the object of interest. The goal is to learn a function Φ, implemented as a neural network, that maps the image $\mathbf{I}$ to four factors (d, a, w, l) comprising a *depth map* $d : \Omega \to \mathbb{R}_+$, an *albedo image* $a : \Omega \to \mathbb{R}^3$, a global *light direction* $l \in \mathbb{S}^2$, and a *viewpoint* $w \in \mathbb{R}^6$ so that the image can be reconstructed from them.

The image $\mathbf{I}$ is reconstructed from the four factors in two steps, *lighting* Λ and *reprojection* Π, as follows:

$$\hat{\mathbf{I}} = \Pi\left(\Lambda(a, d, l), d, w\right). \tag{1}$$

The lighting function Λ generates a version of the object based on the depth map d, the light direction l and the albedo a as seen from a canonical viewpoint $w = 0$, assuming Lambertian shading with a directional light source. The viewpoint w represents the transformation between the canonical view and the viewpoint of the actual input image $\mathbf{I}$. Then, the reprojection function Π simulates the effect of a viewpoint change and generates the image $\hat{\mathbf{I}}$ given the canonical depth d and the shaded canonical image $\Lambda(a, d, l)$. We use a differentiable renderer from [Kato *et al.*, 2018]. Learning uses a reconstruction loss which encourages $\mathbf{I} \approx \hat{\mathbf{I}}$ (Section 2.2).

Discussion. The effect of lighting could be incorporated in the albedo a by interpreting the latter as a texture rather than as the object's albedo. However, there are two good reasons to avoid this. First, the albedo a is often symmetric even if the illumination causes the corresponding appearance to look asymmetric. Separating them allows us to more effectively incorporate the symmetry constraint described below. Second, shading provides an additional cue on the underlying 3D shape [Horn, 1975]. In particular, unlike the recent work of [Shu *et al.*, 2018] where a shading map is predicted independently from shape, our model computes the shading based on the predicted depth, mutually constraining each other.

2.2 Probably Symmetric Objects

Leveraging symmetry for 3D reconstruction requires identifying symmetric points in an image. Here we do so implicitly, assuming that depth and albedo, which are reconstructed in a canonical frame, are symmetric about a fixed vertical plane.

To do this, we consider the operator that flips a map $a \in \mathbb{R}^{C\times W\times H}$ along the horizontal axis[1]: $[\operatorname{flip} a]_{c,u,v} = a_{c,W-1-u,v}$. We then require $d \approx \operatorname{flip} d'$ and $a \approx \operatorname{flip} a'$. While these constraints could be enforced by adding corresponding loss terms to the learning objective, they would be difficult to balance. Instead, we achieve the same effect indirectly, by obtaining a second reconstruction $\hat{\mathbf{I}}'$ from the flipped depth and albedo:

$$\hat{\mathbf{I}}' = \Pi\left(\Lambda(a', d', l), d', w\right), \quad a' = \operatorname{flip} a, \quad d' = \operatorname{flip} d. \tag{2}$$

Then, we consider two reconstruction losses encouraging $\mathbf{I} \approx \hat{\mathbf{I}}$ and $\mathbf{I} \approx \hat{\mathbf{I}}'$. Since the two losses are commensurate, they are easy to balance and train jointly. Most importantly, this approach allows us to easily reason about symmetry probabilistically, as explained next.

The source image $\mathbf{I}$ and the reconstruction $\hat{\mathbf{I}}$ are compared via the loss:

$$\mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}, \sigma) = -\frac{1}{|\Omega|}\sum_{uv\in\Omega} \ln \frac{1}{\sqrt{2}\sigma_{uv}} \exp -\frac{\sqrt{2}\ell_{1,uv}}{\sigma_{uv}}, \tag{3}$$

where $\ell_{1,uv} = |\hat{\mathbf{I}}_{uv} - \mathbf{I}_{uv}|$ is the L_1 distance between the intensity of pixels at location uv, and $\sigma \in \mathbb{R}_+^{W\times H}$ is a *confidence*

[1]The choice of axis is arbitrary as long as it is fixed.

map, also estimated by the network Φ from the image $\mathbf{I}$, which expresses the *aleatoric uncertainty* of the model. The loss can be interpreted as the negative log-likelihood of a factorized Laplacian distribution on the reconstruction residuals. Optimizing likelihood causes the model to self-calibrate, learning a meaningful confidence map [Kendall and Gal, 2017].

Modelling uncertainty is generally useful, but in our case is particularly important when we consider the "symmetric" reconstruction $\hat{\mathbf{I}}'$, for which we use the same loss $\mathcal{L}(\hat{\mathbf{I}}', \mathbf{I}, \sigma')$. Crucially, we use the network to estimate, also from the same input image $\mathbf{I}$, a *second* confidence map σ'. This confidence map allows the model to learn which portions of the input image might *not* be symmetric. For instance, in some cases hair on a human face is not symmetric as shown in Fig. 2, and σ' can assign a higher reconstruction uncertainty to the hair region where the symmetry assumption is not satisfied. Note that this depends on the *specific* instance under consideration, and is learned by the model itself.

Overall, the learning objective is given by the combination of the two reconstruction errors:

$$\mathcal{E}(\Phi; \mathbf{I}) = \mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}, \sigma) + \lambda_{\text{f}} \mathcal{L}(\hat{\mathbf{I}}', \mathbf{I}, \sigma'), \tag{4}$$

where $\lambda_{\text{f}} = 0.5$ is a weighing factor, $(d, a, w, l, \sigma, \sigma') = \Phi(\mathbf{I})$ is the output of the neural network, and $\hat{\mathbf{I}}$ and $\hat{\mathbf{I}}'$ are obtained according to Eqs. (1) and (2).

2.3 Image Formation Model

We now describe the functions Π and Λ in Eq. (1) in more detail. The image is formed by a camera looking at a 3D object. If we denote with $P = (P_x, P_y, P_z) \in \mathbb{R}^3$ a 3D point expressed in the reference frame of the camera, this is mapped to pixel $p = (u, v, 1)$ by the following projection:

$$p \propto KP, \quad K = \begin{bmatrix} f & 0 & c_u \\ 0 & f & c_v \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{cases} c_u = \frac{W-1}{2}, \\ c_v = \frac{H-1}{2}, \\ f = \frac{W-1}{2 \tan \frac{\theta_{\text{FOV}}}{2}}. \end{cases} \tag{5}$$

Given that the images are cropped around the objects, we assume a relatively narrow *field of view* (FOV) of $\theta_{\text{FOV}} \approx 10°$.

The depth map $d : \Omega \to \mathbb{R}_+$ associates a depth value d_{uv} to each pixel $(u, v) \in \Omega$ in the canonical view. By inverting the camera model Eq. (5), we find this corresponds to the 3D point $P = d_{uv} \cdot K^{-1} p$. The viewpoint $w \in \mathbb{R}^6$ represents an Euclidean transformation $(R, T) \in SE(3)$, where $w_{1:3}$ and $w_{4:6}$ are rotation angles and translations in xyz axes respectively. The map (R, T) transforms 3D points from the canonical view to the actual view. Thus a pixel (u, v) in the canonical view is mapped to the pixel (u', v') in the actual view by the warping function $\eta_{d,w} : (u, v) \mapsto (u', v')$ given by:

$$p' \propto K(d_{uv} \cdot RK^{-1} p + T), \tag{6}$$

where $p' = (u', v', 1)$.

Finally, the reprojection function Π takes as input the depth d and the viewpoint change w and applies the resulting warp to the canonical image $\mathbf{J}$ to obtain the actual image $\hat{\mathbf{I}} = \Pi(\mathbf{J}, d, w)$ as $\hat{\mathbf{I}}_{u'v'} = \mathbf{J}_{uv}$, where $(u, v) = \eta^{-1}_{d,w}(u', v')$.

The canonical image $\mathbf{J} = \Lambda(a, d, l)$ is in turn generated as a combination of albedo, normal map and light direction.

No	Baseline	SIDE ($\times 10^{-2}$) ↓	MAD (deg.) ↓
(1)	Supervised	0.410 ± 0.103	10.78 ± 1.01
(2)	Const. null depth	2.723 ± 0.371	43.34 ± 2.25
(3)	Average g.t. depth	1.990 ± 0.556	23.26 ± 2.85
(4)	Ours (unsupervised)	0.793 ± 0.140	16.51 ± 1.56

Table 1: Comparison with baselines. SIDE and MAD errors for 3D reconstruction in the BFM dataset of our unsupervised reconstruction method against a fully-supervised and trivial baselines.

To do so, given the depth map d, we derive the normal map $n : \Omega \to \mathbb{S}^2$ by associating to each pixel (u, v) a vector normal to the underlying 3D surface. In order to find this vector, we compute the vectors t^u_{uv} and t^v_{uv} tangent to the surface along the u and v directions. For example, the first one is: $t^u_{uv} = d_{u+1,v} \cdot K^{-1}(p + e_x) - d_{u-1,v} \cdot K^{-1}(p - e_x)$ where p is defined above and $e_x = (1, 0, 0)$. Then the normal is obtained by taking the vector product $n_{uv} \propto t^u_{uv} \times t^v_{uv}$.

The normal n_{uv} is multiplied by the light direction l to obtain a value for the directional illumination and the latter is added to the ambient light. Finally, the result is multiplied by the albedo to obtain the illuminated texture, as follows: $\mathbf{J}_{uv} = (k_s + k_d \max\{0, \langle l, n_{uv} \rangle\}) \cdot a_{uv}$. Here k_s and k_d are the scalar coefficients weighting the ambient and diffuse terms, and are predicted by the model with range between 0 and 1 via rescaling a `tanh` output. The light direction $l = (l_x, l_y, 1)^T / (l_x^2 + l_y^2 + 1)^{0.5}$ is modeled as a spherical sector by predicting l_x and l_y with `tanh`.

3 Experiments

Datasets. We test our method on human faces and cat faces, using the public datasets, *CelebA* [Liu *et al.*, 2015], *3DFAW* [Gross *et al.*, 2010; Jeni *et al.*, 2015; Zhang *et al.*, 2014; Yin *et al.*, 2008], and *cat datasets* [Zhang *et al.*, 2008; Parkhi *et al.*, 2012]. We roughly crop the images around the head region and use the official train/val/test splits. In order to assess the quality of the 3D reconstructions (since the in-the-wild datasets lack ground-truth), we generate a synthetic face dataset (*BFM*) with variation in shape, pose, texture and illumination using the Basel Face Model [Paysan *et al.*, 2009], following the protocol of [Sengupta *et al.*, 2018].

Comparison with baselines. Table 1 uses the BFM dataset to compare the depth reconstruction quality obtained by our method, a fully-supervised baseline and two baselines. We discount the inherent scale ambiguity of the 3D reconstruction using the *scale-invariant depth error* (SIDE) [Eigen *et al.*, 2014] $E_{\text{SIDE}}(\bar{d}, d^*) = (\frac{1}{WH} \sum_{uv} \Delta^2_{uv} - (\frac{1}{WH} \sum_{uv} \Delta_{uv})^2)^{\frac{1}{2}}$ where $\Delta_{uv} = \log \bar{d}_{uv} - \log d^*_{uv}$. Additionally, we report the *mean angle deviation* (MAD) between normals computed from ground truth depth and from the predicted depth, measuring how well the surface is captured by the prediction. The supervised baseline is a version of our model trained to regress the ground-truth depth maps using an L_1 loss. The trivial baseline predicts a constant uniform depth map, which provides a performance lower-bound. The third baseline is a constant depth map obtained by averaging all ground-truth

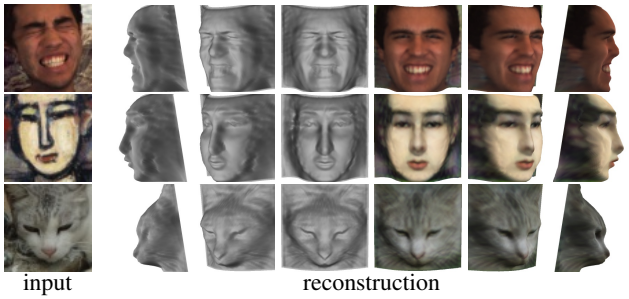

Figure 3: Reconstruction of human faces, paintings and cat faces.

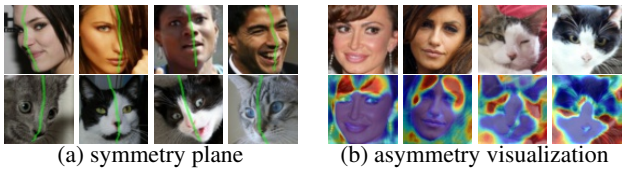

Figure 4: Symmetry plane and asymmetry detection. (a): our model uncovers the "intrinsic" symmetry plane of an in-the-wild object even though the appearance is highly asymmetric. (b): asymmetries (highlighted in red) are detected with the confidence maps.

depth maps in the test set. Our method largely outperforms the two constant baselines and approaches the results of supervised training. Improving over the third baseline (which has access to GT information) confirms that the model learns an *instance specific* 3D representation.

Qualitative results. In Fig. 3 we show reconstruction results of human faces and cat faces as well abstract face paintings from the Internet. The 3D shapes are recovered with high fidelity. The reconstructed 3D face, for instance, contain fine details of the nose, eyes and mouth even with extreme facial expression. Our method also generalizes well on paintings, even though it has never seen such images during training.

Symmetry and asymmetry detection. Since our model predicts a canonical view of the objects that is symmetric about the vertical center-line of the image, we can easily visualize the symmetry plane, which is otherwise non-trivial to detect from in-the-wild images. In Fig. 4, we render the center-line of the canonical image and warp it to the input viewpoint. The symmetry planes detected by our method are accurate despite the presence of extreme asymmetric texture and lighting effects. We also overlay the predicted confidence map σ' onto the image, confirming that the model assigns low confidence to asymmetric regions in a sample-specific way.

Comparison with SOTAs. We compare the reconstruction quality of our method with two recently proposed unsupervised reconstruction methods, LAE [Sahasrabudhe *et al.*, 2019] and [Szabó *et al.*, 2019]. Our method produces much higher quality reconstructions than both methods, with fine details of the facial expression, whereas LAE recovers 3D shapes poorly and [Szabó *et al.*, 2019] generates unnatural shapes. Note that [Szabó *et al.*, 2019] uses an unconditional GAN that generates 3D faces from random noise, and cannot recover 3D shapes from images. The input images for [Szabó *et al.*, 2019] in Fig. 5 were generated by their method.

Figure 5: Compared to state-of-the-art methods, our method recovers much higher quality shapes.

4 Related Work

Traditional Structure from Motion (SfM) [Faugeras and Luong, 2001] can reconstruct the 3D structure of individual rigid scenes given multiple views of each scene as well as 2D keypoint matches between the views. Learning-based methods have recently been leveraged to reconstruct objects from a single view. A variety of supervisory signals apart from direct 3D ground-truth have been explored, including videos [Zhou *et al.*, 2017], keypoint annotations [Kanazawa *et al.*, 2018b], object masks [Chen *et al.*, 2019], predefined shape models [Kanazawa *et al.*, 2018a; Gerig *et al.*, 2018]. These prior models are constructed using specialized hardware and/or other forms of supervision, which are often difficult to obtain for deformable objects in the wild, such as animals, and also limited in shape details.

Only recently have authors attempted to learn the geometry of object categories from raw, monocular views *only*. [Sahasrabudhe *et al.*, 2019] leverages deformation fields from DAE [Shu *et al.*, 2018] learned with a heavy bottleneck constraint and further extracts 3D shape and lighting. Others have considered adversarial learning. In particular, HoloGAN [Nguyen-Phuoc *et al.*, 2019] only uses raw images but does not obtain an explicit 3D reconstruction. [Szabó *et al.*, 2019] generates 3D meshes of faces using an unconditional GAN and cannot predict from images. [Henzler *et al.*, 2019] also learns from raw images, but only experiments with images with a white background, akin to supervision with masks.

Our work is also inspired from *shape from symmetry* [Mukherjee *et al.*, 1995; François *et al.*, 2003] and *shape from shading* [Horn and Brooks, 1989].

Acknowledgements

We thank Soumyadip Sengupta and Mihir Sahasrabudhe for sharing their code and results with us. We are also indebted to the members of Visual Geometry Group for insightful discussions and comments. This work is jointly supported by Facebook Research and ERC Horizon 2020 Research and Innovation Programme IDIU 638009.

References

[Chen *et al.*, 2019] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaako Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *NeurIPS*, 2019.

[Eigen *et al.*, 2014] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014.

[Faugeras and Luong, 2001] Olivier Faugeras and Quang-Tuan Luong. *The Geometry of Multiple Images*. MIT Press, 2001.

[François *et al.*, 2003] Alexandre R. J. François, Gérard G. Medioni, and Roman Waupotitsch. Mirror symmetry $\Rightarrow$ 2-view stereo geometry. *Image and Vision Computing*, 2003.

[Gerig *et al.*, 2018] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Lüthi, Sandro Schönborn, and Thomas Vetter. Morphable face models - an open framework. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018.

[Gross *et al.*, 2010] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 2010.

[Henzler *et al.*, 2019] Philipp Henzler, Niloy Mitra, and Tobias Ritschel. Escaping plato's cave using adversarial training: 3d shape from unstructured 2d image collections. In *Proc. ICCV*, 2019.

[Horn and Brooks, 1989] Berthold K. P. Horn and Michael J. Brooks. *Shape from Shading*. MIT Press, Cambridge Massachusetts, 1989.

[Horn, 1975] Berthold Horn. Obtaining shape from shading information. In *The Psychology of Computer Vision*, 1975.

[Jeni *et al.*, 2015] László A. Jeni, Jeffrey F. Cohn, and Takeo Kanade. Dense 3d face alignment from 2d videos in real-time. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2015.

[Kanazawa *et al.*, 2018a] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. CVPR*, 2018.

[Kanazawa *et al.*, 2018b] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proc. ECCV*, 2018.

[Kato *et al.*, 2018] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proc. CVPR*, 2018.

[Kendall and Gal, 2017] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. ICCV*, 2015.

[Mukherjee *et al.*, 1995] Dipti P. Mukherjee, Andrew Zisserman, and J. Michael Brady. Shape from symmetry – detecting and exploiting symmetry in affine images. *Philosophical Transactions of the Royal Society of London*, 351:77–106, 1995.

[Nguyen-Phuoc *et al.*, 2019] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proc. ICCV*, 2019.

[Parkhi *et al.*, 2012] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proc. CVPR*, 2012.

[Paysan *et al.*, 2009] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *Advanced video and signal based surveillance*, 2009.

[Sahasrabudhe *et al.*, 2019] Mihir Sahasrabudhe, Zhixin Shu, Edward Bartrum, Riza Alp Guler, Dimitris Samaras, and Iasonas Kokkinos. Lifting autoencoders: Unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion. In *ICCV Workshop on Geometry Meets Deep Learning*, 2019.

[Sengupta *et al.*, 2018] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David Jacobs. SfSNet: Learning shape, refectance and illuminance of faces in the wild. In *Proc. CVPR*, 2018.

[Shu *et al.*, 2018] Zhixin Shu, Mihir Sahasrabudhe, Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proc. ECCV*, 2018.

[Szabó *et al.*, 2019] Attila Szabó, Givi Meishvili, and Paolo Favaro. Unsupervised generative 3d shape learning from natural images. *arXiv*, abs/1910.00287, 2019.

[Wu *et al.*, 2020] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020.

[Yin *et al.*, 2008] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A high-resolution 3d dynamic facial expression database. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2008.

[Zhang *et al.*, 2008] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection - how to effectively exploit shape and texture features. In *Proc. ECCV*, 2008.

[Zhang *et al.*, 2014] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[Zhou *et al.*, 2017] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. CVPR*, 2017.