# Unsupervised and weakly supervised deep learning methods for computer vision and medical imaging

Mihir Sahasrabudhe

université
PARIS-SACLAY

# Unsupervised and Weakly Supervised Deep Learning Methods for Computer Vision and Medical Imaging

**Thèse de doctorat de l'Université Paris-Saclay**

École doctorale n° 580, sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat: Mathématiques et Informatiques
Unité de recherche: Université Paris-Saclay, CentraleSupélec, Centre de Vision Numérique, 91190, Gif-sur-Yvette, France
Référent: CentraleSupélec

**Thèse présentée et soutenue à Gif-sur-Yvette le 6 mars 2020, par**

**Mihir SAHASRABUDHE**

**Composition du jury:**

| | |
|---|---|
| **Maja PANTIC**<br>Professeure, Imperial College London | Présidente |
| **Rafeef GARBI**<br>Professeure, University of British Columbia | Rapporteur |
| **Stefanos ZAFEIRIOU**<br>Reader, Imperial College London, HDR | Rapporteur et Examinateur |
| **Mathieu AUBRY**<br>Professeur Associé, École des Ponts ParisTech, HDR | Examinateur |
| **Liming CHEN**<br>Professeur, École Centrale Lyon | Examinateur |
| **Céline HUDELOT**<br>Professeure, CentraleSupélec | Examinatrice |
| **Nikos PARAGIOS**<br>Professeur, CentraleSupélec | Directeur |
| **Maria VAKALOPOULOU**<br>Maîtresse de Conférences, CentraleSupélec | Invitée |

Thèse de doctorat

NNT: 2020UPASC010

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

**Titre:** Méthodes Non-Supervisées et Faiblement Supervisées d'Apprentissage Profond pour la Vision par Ordinateur et l'Imagerie Médicale

**Mots clés:** Apprentissage profond, apprentissage non-supervisé, apprentissage faiblement supervisé, vision par ordinateur, imagerie médicale

**Résumé:** Les premières contributions de cette thèse (Chapter 2 et Chapitre 3) sont des modèles appelés Deforming Autoencoder (DAE) et Lifting Autoencoder (LAE), utilisés pour l'apprentissage non-supervisé de l'alignement 2-D dense d'images d'une classe donnée, et à partir de cela, pour apprendre un modèle tridimensionnel de l'objet. Ces modèles sont capable d'identifier des espaces canoniques pour représenter de différent caractéristiques de l'objet, à savoir, l'apparence des objets dans l'espace canonique, la déformation dense associée permettant de retrouver l'image réelle à partir de cette apparence, et pour le cas des visages humains, le modèle 3-D propre au visage de la personne considérée, son expression faciale, et l'angle de vue de la caméra. De plus, nous illustrons l'application de DAE à d'autres domaines, à savoir, l'alignement d'IRM de poumons et d'images satellites. Dans le Chapitre 4, nous nous concentrons sur une problématique lié au cancer du sang—diagnostique d'hyperlymphocytosis. Nous proposons un modèle convolutif pour encoder les images appartenant à un patient, suivi par la concaténation de l'information contenue dans toutes les images. Nos résultats montrent que les modèles proposés sont de performances comparables à celles des biologistes, et peuvent dont les aider dans l'élaboration de leur diagnostique.

**Title:** Unsupervised and Weakly Supervised Deep Learning Methods for Computer Vision and Medical Imaging

**Keywords:** Deep Learning, Unsupervised Learning, Weakly Supervised Learning, Computer Vision, Medical Imaging

**Abstract:** The first two contributions of this thesis (Chapter 2 and 3) are models for unsupervised 2D alignment and learning 3D object surfaces, called Deforming Autoencoders (DAE) and Lifting Autoencoders (LAE). These models are capable of identifying canonical space in order to represent different object properties, for example, appearance in a canonical space, deformation associated with this appearance that maps it to the image space, and for human faces, a 3D model for a face, its facial expression, and the angle of the camera. We further illustrate applications of models to other domains—alignment of lung MRI images in medical image analysis, and alignment of satellite images for remote sensing imagery. In Chapter 4, we concentrate on a problem in medical image analysis—diagnosis of lymphocytosis. We propose a convolutional network to encode images of blood smears obtained from a patient, followed by an aggregation operation to gather information from all images in order to represent them in one feature vector which is used to determine the diagnosis. Our results show that the performance of the proposed models is at-par with biologists and can therefore augment their diagnosis.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

# Acknowledgements

I would first of all like to thank my supervisor Prof. Nikos Paragios for giving me the opportunity to work towards a doctorate under his guidance. I have learnt a lot from him over the years, the most important being to always concentrate and remain on the goal of the task at hand and to separate it from what is less important. I would also like to thank Prof. Maria Vakalopoulou who has helped me immensely during my thesis work. I want to extend my thanks to my thesis examiners Prof. Stefanos Zafeiriou and Prof. Rafeef Garbi for going through my thesis manuscript and giving valuable feedback, as well as my jury members Prof. Maja Pantic, Prof. Liming Chen, Prof. Céline Hudelot, and Prof. Mathieu Aubry for their time and effort towards evaluating my thesis work.

I was fortunate to work with some exceptional people in the fields of computer vision, medicine, and biology. I would like to thank Prof. Iasonas Kokkinos and Prof. Dimitris Samaras for their teaching and support. I would like to thank Mr. Alain Rivéro and his colleagues from the SNCF for allowing us the opportunity to work with them, as well as the experience of travelling in a train engine. I would also like to thank Dr. Pierre Sujobert and his team at CHU Lyon as well as Dr. Evangelia I. Zacharaki for their collaboration. I express my thanks also to Prof. Tomas Kirchhausen at Harvard Medical School where I spent two valuable months working at the intersection of ML and biology.

I would like to thank all the faculty at CVN for building a research atmosphere in the lab—Prof. Pesquet for all his help during my final years of thesis, as well as Prof. Malliaros, Prof. Chouzenoux, Prof. Castella, and Prof. Talbot. I would also like to thank Natalia and Jana for helping me immensely with administrative tasks and making life easier. Thanks to Anthony Guindon for the interesting chats as well as all his help with the servers. My thanks also to Prof. Duc and Mme. Batalie from the École Doctorale STIC. I thank Fondation CentraleSupélec for their generous support during my last year.

I would like to thank my collaborators, fellow doctoral students, Rıza Alp Güler, Zhixin Shu, Stergios Christodoulidis, and Edward Bartrum, for making the deadlines much less stressful.

I would like to thank Siddhartha and Puneet, whose presence and help with adjusting to life in France and the lab as well as with research cannot be stressed enough. I thank Marie-Caroline for teaching me French among other valuable things, Maria V for being a friend as well as a guide, Stergios for the fun times in Cité U, and Alp for the gatherings he hosted. I would also like to thank the "old" CVN people—Hari and Suganya, Evgenios, Stefan, Khuê, Guillaume, Enzo, Eugene, Rafael, Norbert, Simon, Stavros, Jiaqian, and Wacha—as well as the "new" ones—Maria P for the movies, Arthur for his stories, Matthieu, Théo, Enzo, Kadir, Marvin, Maïssa, Sagar, Giorgos, Roger, Kavya, Ana, Yingping, and Yunshi, and all the others I might have missed. I would also like to thank all the people from Kirchhausen Lab for their warm welcome—Rasmus, Giuseppe, Mootaz, Ilja, Tegy, George, and Cat. My thanks go out to all my friends at the Maison de l'Inde at Cité U—Ayush, Akhil, Amulya, Hemlata, Manohar, Marie-Pascal, Sumit, Komal, Desmond, Sanket, Disha, Arthur, and most importantly to Saurabh and Irene for the many dinners and discussions on the balcony. Many thanks also to Fanny, Marie-Claude, and Grégoire for introducing me to French culture.

Finally, I would like to express my gratitude to my family, and especially to my grandmother, my parents, my brother Harshad, and my sister-in-law Leena for their constant love, support, and encouragement during my thesis years. Our trips to France and the US, and more recently, the time I spent in Boston with Harshad and Leena are part of the memories from my PhD years that I cherish the most.

# Abstract

Data are an indispensable component of any machine learning system. For specialised domains, labelled data can be very hard and expensive to obtain. For example, gathering labelled medical data requires significant time and effort of doctors and biologists. This thesis proposes methods for unlabelled and weakly labelled data for two problems in computer vision and medical imaging. The first problem centres around morphable model learning for computer vision, where a framework for learning a morphable model without an intensive image acquisition process is introduced. The second problem is focussed on automatic diagnosis of lymphoproliferative disorders in the presence of lymphocytosis with weak, patient-level ground truth labels.

The first contribution of this thesis (Chapter 2) is a model for unsupervised dense alignment of 2D images of an object category. A model is proposed to learn a canonical space for the category from a set of unconstrained images and simultaneously to infer dense correspondences between an image of the category and the canonical template. This is achieved using a deep autoencoder, which disentangles the appearance and shape of the object in its latent space. The appearance latent vector is used to decode the appearance of the object in the template space, while the shape latent vector is used to decode a dense deformation between the template space and the image space. A technique to regress the dense deformation grid using a convolutional decoder is introduced and it is shown that this technique outperforms direct regression of the grid or the residual grid (offsets). The alignment in the canonical space is evaluated using landmark localisation and the proposed method is shown to outperform the state-of-the-art. It is also demonstrated that, for face images, we can further disentangle the appearance latent vector into albedo and shading, and that it becomes an easier problem because of the aligned nature of the template space. Finally, applications of the alignment method to registration of lung MRIs and satellite imagery are demonstrated.

Chapter 3 extends the ideas of unsupervised dense alignment in 2D to learn 3D shapes for faces. While treating the alignment in 2D as ground truth, it is shown that it is possible to recover 3D shape using non-rigid structure-from-motion (NRS$f$M). In similar fashion to Chapter 2, an autoencoder is employed to regress image-specific parameters for deformation as well as pose. A mean (or *base*) shape in the form of a surface mesh in 3D is learnt simultaneously. The regressed parameters and the base shape determine the image-specific mesh, which is then rendered during image formation. A significant addition to the model is achieved using weak supervision for pose, identity, and expression, using which the proposed framework is able to learn a highly controllable 3D model for the human face. The resulting model is evaluated using landmark localisation, and a method is proposed for the validation of the inferred shape and pose using Procrustes analysis.

The use of weakly supervised learning for medical imaging is investigated in Chapter 4. An important part of the diagnosis of a patient exhibiting lymphocytosis, i.e., absolute lymphocyte count above $4\times10^9$/L, is determining whether its cause is reactive or the manifestation of a lymphoproliferative disorder. Due to large inter- and intra-operator variability in assessing individual lymphocytes, ground truth is available only at the patient-level, not at the image level. A multi-instance classification framework based on embedding-level pooling and a mixture-of-experts model is proposed to reliably classify cases as reactive or tumoral. A further comparison with the average prediction of 12 experienced biologists is made, in which the proposed method is shown to perform better. Promising results show that it is possible to reliably delegate this diagnosis to a machine.

# Contents

# List of Figures

Chapter 1

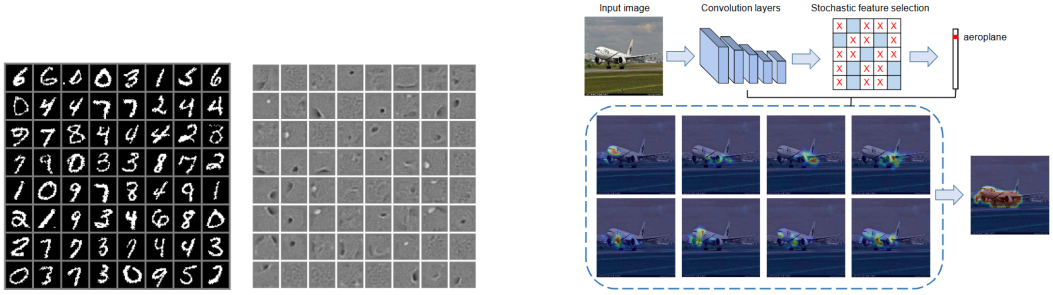
# Introduction

Deep learning has revolutionised the applications of machine learning and computer vision to our daily lives. The past decade has seen staggering growth in deep learning research which has eventually led to unprecedented rise in the number of companies and start-up companies based on the deep learning model, be it through social media, urban organisation, transportation, medicine, or weather. This has been made possible by the vast amounts of data available for computational use. Even though a lot of effort has been invested in order to minimise the burden of annotation, it still takes a considerable amount of effort to gather, annotate, compile, and release data. On top of that, annotators can have high variance because different people can have different interpretations of the data and the rules of annotation. For specialised domains, for example, medical imaging, data can be hard to obtain and even harder to annotate as it requires the time and effort of specialised personnel.

Humans are capable of learning from sparsely annotated data by generalising efficiently to unseen data. To replicate the learning process of the human brain, we must understand and replicate its inner workings. Most contemporary artificial intelligence research is based on supervised methods [He 2017, Gkioxari 2019], and requires large amounts of human-annotated training data. This is contrary to the way we as humans learn, which is simply by observing and making connections and drawing inferences based on what we have observed so far [LeCun 2015].

Early research in unsupervised deep learning focussed on learning discriminative feature representations from data, that were later validated by their performance on related classification tasks [Lee 2009]. This is important for scenarios where annotated examples are not available or not as straightforward to obtain, and unsupervised feature learning can allow us to facilitate learning. For example, Figure 1.1a depicts unsupervised feature extraction using MNIST digits. In the

closely related weakly supervised learning paradigm, the goal is to learn associations between data $x_i$ and targets $y_i$ when $(x_i, y_i)$ pairs are not exactly known. For example, in the weakly supervised semantic segmentation problem, the goal is to segment an object in an image when only the image-level object category label is available (Figure 1.1b). Unsupervised and weakly supervised methods remain an important area of research in machine learning. In this thesis, we examine such deep learning methods for two problems in computer vision and medical imaging. These are elaborated further.




(a) (b)

Figure 1.1: (a) Feature extraction with unsupervised feature learning; and (b) weakly supervised semantic segmentation using FickleNet using image-level labels. Figures taken from [Tang 2011] and [Lee 2019].

**Morphable models.** Morphable models for object categories define a shape and/or texture basis to explain or generate images of the object. Building morphable models is a rather cumbersome process. Early work in morphable model building required carefully calibrated 3D scans of human faces [Blanz 1999]. While more automated processes have been proposed recently [Booth 2018], they still require highly sophisticated pipelines. Further, while there have been considerable advances in unsupervised morphable model fitting using deep learning [Richardson 2016, Tewari 2017, Tewari 2018], few works have targetted jointly learning and fitting a morphable model. In this thesis, we propose advances in order to answer the following question—

> *Is it possible to learn a morphable model for an object category using a set of unlabelled images of the category?*

**Automated diagnosis of lymphocytosis.** Lymphocytosis is a widely observed medical condition in which the absolute lymphocyte count crosses $4 \times 10^9$/L. This is symptomatic of either reactive lymphocytosis, or the manifestation of lymphoproliferative disorder. The former is typically caused by infection, stress, and viral illnesses. The latter is an indication of tumoral behaviour and hence, the patient requires further care. Doctors and biologists spend a considerable amount of time analysing blood smears to diagnose patients exhibiting lymphocytosis as either reactive or tumoral and prescribing further examination based on it. For them, it is

preferable to delegate this task to an automated system due to the analytical nature of the problem. The second problem that we target in this thesis is as follows—

> *Is it possible to learn correlations between blood smears and tumoral lymphocytosis using weakly-labelled examples?*

In the following section, some published works related to the problems discussed above are reviewed.

## 1.1 Previous Work

A short review of unsupervised deep learning methods is followed by a recall of works on deformable models in computer vision. Finally, applications of deep learning to medical imaging problems are discussed.

### 1.1.1 Unsupervised and Weakly Supervised Deep Learning

Unsupervised learning using deep learning methods dates back to the introduction of restricted Boltzman machines (RBMs) [Smolensky 1986]. An RBM (Figure 1.2) is a simple model represented by a bipartite graph, with the two groups of node represents visible units and hidden units. [Hinton 2002, Carreira-Perpinan 2005] introduced and studied a learning algorithm for RBMs, called contrastive divergence (CD). This learning algorithm uses MCMC sampling to compute weight updates, resulting in meaningful hidden representations of input examples. The resulting representations can be used for dimensionality reduction [Makhzani 2015], clustering [Chandra 2013, Xie 2016], feature learning [Coates 2011], and classification [Hinton 2002] among other applications.

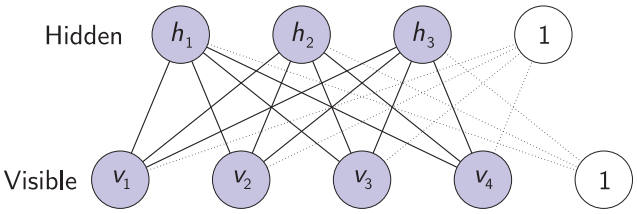



Figure 1.2: A restricted Boltzmann machine (RBM) with four visible and three hidden units.

RBMs can be stacked to produce *deep belief networks* [Hinton 2006], which have been shown to demonstrate higher learning capacity. [Bengio 2007] proposed a method of training deep belief with greedy layer-wise pretraining of each layer using contrastive divergence, followed by supervised fine-tuning. [Hinton 2006] also showed stacking RBMs with greedy layer-wise pretraining gives in a better low-dimensional codes for data. In 1998, the seminal work of [LeCun 1998] showed the application of convolutional neural networks to handwritten character recognition. These convolutional nets were able to learn shift-invariant features for accurate

classification, with the classification becoming more robust with the addition of random deformations to training examples. Applying these notions to RBMs, [Desjardins 2008] were the first to explore convolutional kernels in RBMs for feature extraction, quickly followed by a more extensive study [Lee 2009, Lee 2011]. In the latter, the authors introduced hierarchical probabilistic inference which can be used to reconstruct masked portions of an image based on the regions around by sampling from the joint probability distribution. Their hierarchical model, called the convolutional deep belief network, was able to learn hierarchies of features, from edges, to object parts, to entire objects. Figure 1.3 shows and example of the hierarchy learnt on images of faces.


Figure 1.3: Hierarchical features learnt using a convolutional deep belief network.

RBMs, though effective for dimensionality reduction, are only one layer "deep". Autoencoders are another popular dimensionality technique, built as an encoder-decoder pipeline, where each of the two consists of several stacked non-linearities (Figure 1.4). Autoencoders are trained to minimise reconstruction loss, and their modelling typically sees progressive reduction in the size of the latent code, followed by progressive expansion back to the original dimension. The central layer, called the *bottleneck* layer [Hinton 2016], is set according to the desired compression. Some variants of the autoencoder include the sparse autoencoder [Ng 2011], which constrains the latent representation to be sparse by minimising a KL-divergence term for every latent unit; the $k$-sparse autoencoder [Makhzani 2013], which enforces sparsity by reconstructing using only the latent units with the top-$k$ activations, the variational autoencoder [Kingma 2013], which is able to learn a more controllable model by forcing latent units to follow certain distribution; and convolutional autoencoders [Masci 2011], which use convolutional layers instead of fully-connected layers in the encoder and the decoder. Figure 1.5 visualises some features learnt using a sparse autoencoder.

#### 1.1.1.1 Adversarial learning

So far, RBMs and autoencoders have tried to capture data, but using reconstruction loss, which is usually the mean-squared error (MSE). Minimising the MSE is equivalent to maximising the log-likelihood of a Gaussian over the pixel values. This, however, inherently introduces blur into reconstructed images [Mathieu 2015]. Furthermore, instead of capturing the real data distribution actively, the reconstruction loss tries to achieve it passively by fitting a model to the shown training examples.

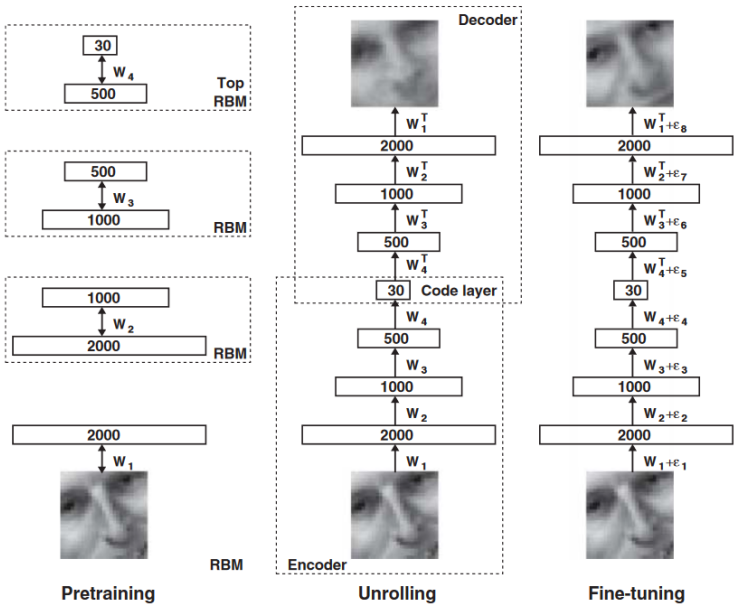



Figure 1.4: A schematic diagramme of the autoencoder. Figure taken from [Hinton 2006].

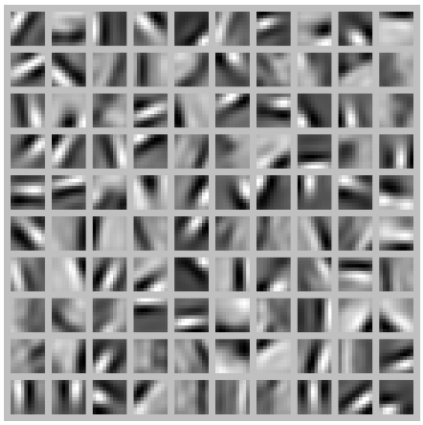


Figure 1.5: Filters learnt using a sparse autoencoder. Image taken from [Ng 2011].

To better capture the true data distribution, [Goodfellow 2014] proposed a new model called the Generative Adversarial Net (GAN). A GAN consists of a generator $G$ and a discriminator $D$. The goal of the generator is to capture the true data distribution and drawing examples from it, while the goal of the discriminator is to determine whether an example was drawn from the true distribution or from the generator's distribution. [Goodfellow 2014] show that in such a competing scenario, it is possible to reach a solution where $G$ captures the true distribution exactly, while $D$ fails to discriminate between the two. [Radford 2015] and [Salimans 2016] introduced new insights into training convolutional architectures with adversarial losses for learning interpretable representations in an unsupervised manner. As a result,

adversarial nets have been shown capable of generating hyper-realistic and controllable images [Pumarola 2018, Karras 2019], image-to-image transfer [Isola 2016], super resolution [Ledig 2017], and face aging [Antipov 2017]. An application of image-to-image transfer to 3D face reconstruction was shown in [Sela 2017] in which the authors train an image-to-image translation network to predict depth and dense UV maps from an image of the human face. The inferred maps are then used for unrestricted 3D reconstruction. For training, the authors use synthetic images created using a 3D morphable model (1.1.2.1) and the corresponding depth and correspondence maps. [Zhu 2017a] proposed a model for image-to-image transfer that does not require paired examples. It introduced a new cycle-consistency loss for GANs (Figure 1.6) which was shown to allow meaningful image-to-image translation without requiring paired examples in the two domains.

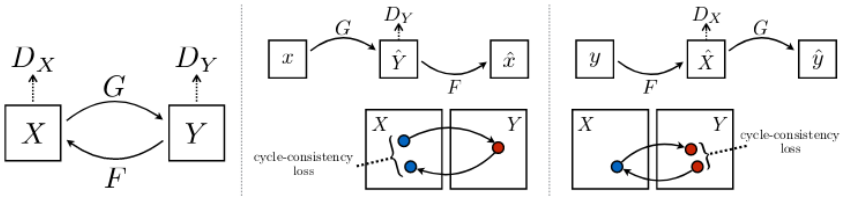



Figure 1.6: GANs with cycle consistency loss for unpaired image-to-image translation.

#### 1.1.1.2 Unsupervised and weakly supervised disentanglement

Disentanglement can be understood as dividing the latent space of an encoder-decoder network into independent parts that represent different sources of variation, and can also possibly be manipulated independently. Disentanglement is a powerful tool for unsupervised learning as focussing on prominent image characteristics, often suited to the task at hand, can help achieve better object understanding. For instance, [Chen 2016b] use variational mutual information maximisation [Barber 2003] to learn digit type, width, and orientation in MNIST digits, azimuth, elevation, and lighting for 3D faces, and facial characteristics and emotions for CelebA images. [Worrall 2017] achieve disentanglement in an unsupervised setting by forcing equivariance through additional losses. [Shu 2017] disentangle a face image into albedo, shading, and normals using an encoder-decoder architecture. They also demonstrate that the learnt model is able to manipulate facial attributes like age, smile, spectacles, etc. by moving on the learnt latent manifold. [Sengupta 2018b] learn to disentangle face images into shape, reflectance, and illumination, in a semi-supervised setting. They use synthetic labelled data to aide the disentanglement. Unlike [Shu 2017], this work uses a 3DMM to generate training examples which guides the learning process. Another work [Tewari 2017] also disentangles shape and appearance using a 3DMM, with the added benefit of directly encoding images into interpretable pose, shape, expression, and illumination parameters. [Sun-

dermeyer 2018] use an autoencoder for 6D object detection. They show that by rendering synthetic views at various rotations and using reconstruction loss, it is possible to predict the rotation from the latent representation in new examples. More recently, [Wiles 2018b] disentangle pose and expression from identity using a video of a person, and show that it is possible to synthesise new videos of the person using a target video.

#### 1.1.1.3 Unsupervised alignment

Significant work has been done recently in achieving unsupervised alignment using deep learning. One approach taken quite often is to locate a given number of discriminative landmark locations in images and to find a transformation of the image to canonical space using these landmarks. [Thewlis 2017b] propose an unsupervised method for discovering a canonical shape $S \subset \mathbb{R}^3$ for a category of objects from observed images $\mathbf{x} \in \Lambda \subset \mathbb{R}^3$. Let $\Phi_S(\cdot, \mathbf{x}) : \mathbb{R}^3 \to \mathbb{R}^3$ be a function that maps a point in the image space to a point in the canonical space. At the core of this method lies the equivariance constraint, which says that if the image $\mathbf{x}$ is visualised from another viewpoint to give an image $\mathbf{x}'$ under a 2D warp $g$, then

$$\forall\, p \in S, \Phi_S(p, \mathbf{x} \circ g) = g \circ \Phi_S(p, \mathbf{x})\,. \tag{1.1}$$

This paper demonstrates unsupervised landmark discovery on cats, shoes, and human faces. Landmark detection using the learnt models is also shown on the MAFL [Zhang 2014a] AFLW [Zhu 2016] test sets. The authors further improve landmark learning in [Thewlis 2017b] by adding a constraint for the detected landmarks to be distinctive. They also include dilations in the neural network architecture used and demonstrate an increase in landmark accuracy of 1 percentage point. [Zhang 2018] add separation and concentration constraints on top of the equivariance constraints, which correspond respectively to inter-landmark variance and concentration of one landmark to a region. They also use landmark-based decoders to train the network end-to-end using reconstruction loss. [Jakab 2018a] use a cycle-consistency constraint on the detected landmarks by learning two functions $\Phi$ and $\Psi$, $\Phi : \Lambda \to \mathcal{Y}$ and $\Psi : \Lambda \times \mathcal{Y} \to \Lambda$, so that

$$\mathbf{x}' = \Psi(\mathbf{x}, \Phi(\boldsymbol{x}'))\,, \tag{1.2}$$

where $\boldsymbol{x}, \boldsymbol{x} \in \Lambda$ are images of two objects viewed from different viewpoints, and $\mathcal{Y} \subset \mathbb{R}^2$ is the set of desired landmarks. To prevent learning identity mappings for $\Phi$ and $\Psi$, they enforce sparse latent representations by concentrating each landmark to one location. [Suwajanakorn 2018] extend this notion to 3D keypoints, using multi-view consistency and introducing weak supervision in the form of rigid transformations between the views.

#### 1.1.1.4 Multiple instance and weakly supervised learning

Multiple instance learning (MIL) is a learning paradigm in which data points consists of bags and labels. Each bag consists of several instances. Akin to a classical classification problem, the goal is to predict the label given the bag. However, this problem is more difficult because the relationship between the instances and the bag-level label is not known [Keeler 1991, Dietterich 1997]. The standard MIL assumption models this relationship by the presence of at least one positive instance, i.e., the instance are assumed to take positive and negative labels, and the bag is given a positive label if at least one instance is positive [Foulds 2010a]. This assumption can however be strict for some problems, so smoother functions are sometimes used [Pinheiro 2015]. In essence, this problem can be said to lie in the subset of weakly supervised learning. We will hence discuss some previous work in these two paradigms together in this section.

Weakly supervised learning has seen considerable application in the problem of object detection. In the problem of object classification the goal is to predict a class label for every image. Object detection, on the other hand, differs from classification problems in that it entails prediction of a bounding box around objects along with the class label, which renders the problem more difficult, and at the same time, requires more annotations for the training data. Weakly supervised object detection tries to circumvent this problem by using image-level labels coupled with other techniques to predict bounding boxes for objects. [Pinheiro 2015] proposed using a CNN to predict pixel-wise heatmaps for object classes coupled with an aggregation function for class scores to give an image-level label. The image-level classifier can be trained with negative log-likelihood. They show that using a smooth maximum function over the heatmaps to aggregate the scores enables learning localised detections using the heatmaps, which can be smoothed to simultaneously predict bounding boxes. [Oquab 2014] localise objects using a CNN pre-trained on ImageNet on top of sliding windows on images to generate training examples for an object detector for Pascal VOC object detection problem. In a follow-up work [Oquab 2015], they use a modified loss function which transfers labels to sliding windows based on the image-level label. [Zhou 2016a] achieve object localisation by projectecting class activation maps onto the image using global averagen pooling. [Bilen 2016] use a pre-trained CNN and a region-proposal network followed by dedicated classification and detection branches to learn from image-level labels. [Cinbis 2016] use an iterative method of training using the MIL strategy, by splitting the training data into folds, in order to escape local minima. [Kantorov 2016] score regions of interest (ROIs) using surrounding context under additive and constrastive strategies. They answer the object detection question in two different forms, respectively-(a)whether the object is in the context as well as the ROI, and (b) whether the object is in the context but not in the ROI. Moving on to segmentation and saliency maps, some recent works also propose approaches for weakly supervised semantic segmentation. [Chaudhry 2017] propose a network for joint classification and semantic segmentation using class labels. [Huang 2018] use region growing using

initially predicted discriminative regions in the image to grow the segmentation. Figure 1.7 shows an example of the iterative refinement of semantic segmentation of their method. Their framework updates the ground truth labels at each iteration. [Lee 2019] use stochastically sampled hidden units to obtain localisation maps that identify discriminative and non-discriminative parts in the image.

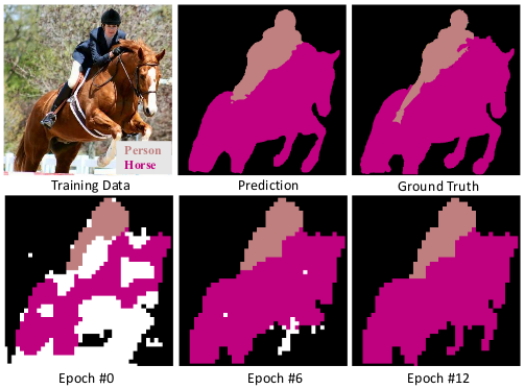



Figure 1.7: Iterative refinement of the prediction pixel-level labels. Image taken from the paper.

MIL has also seen applications to in other domains. [Kraus 2016] propose the Noisy-AND aggregation function for classification and segmentation of microscopy images. [Wang 2018] show applications of embedding-level pooling in MIL coupled with deep supervision to various datasets. [Ilse 2018] use an attention mechanism in a deep-learning scheme to discriminate a certain type of nucleus from others in histopathology images. [Hou 2016] use MIL in an EM framework to find discriminative regions in whole slide histopathology images. [Li 2019a] propose a two-stage framework for prostrate whole slide image classification. Each stage is MIL-based using an attention module to locate discriminative regions in the slide. [Papadopoulos 2019] propose an architecture inspired by [Ilse 2018] for automatic detection of tremorous episodes related to Parkinson's disease. [Li 2019b] show the benefits of a multi-scale model with MIL and top-$k$ pooling for clasification of medical images.

### 1.1.2 Computer Vision

We will now recall previous work in computer vision that is related to the topics discussed in this thesis.

#### 1.1.2.1 Deformable Models

Deformable models represent the shape of an object category using a mean shape along with a set of principal deformation vectors. The texture is similary modelled, with illumination being modelled separately sometimes as well. In this section we review advances in deformable template and morphable model generation, learning, and fitting.

Deformable templates have been used for object understanding since long. Furthermore, point-to-point correspondences have been used to compute and register deformable templates to observed objects. Albrecht Dürer, in his work *Four Books on Human Proportion* [Dürer 1534], published his research on using specifically marked points on the human body and comparing then accross subjects to develop a parts-based understanding of the body. These are used to develop *canonical* appearance images which are then transformed to model observed subjects. Figure 1.8 shows a visualisation of such a transformation.

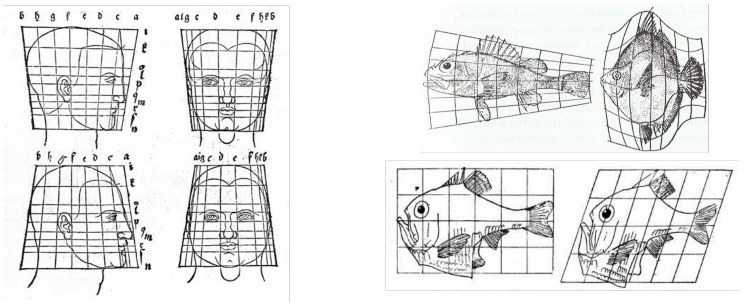


Figure 1.8: *left*: generating human heads using a template [Dürer 1534]; and *right*: deformations between different species of fish [Thompson 1917]. Images takes from respective works.

The deformable template paragidm has proved successful at targetting several computer vision tasks, for example, object localisation [Jain 1996], object matching and retrieval [Funkhouser 2003]

One of the principal works to which unsupervised dense alignment is closely related is continuous joint alignment, or *congealing* [Learned-Miller 2006]. Given a set of binary images of an object category, this method minimises the empirical entropy $\hat{H}(x_i)$ of binary pixel values $x_i^j$ at pixel locations $i$ over all images $j$ in the dataset. This entropy is defined as

$$\hat{H}(x_i) = -\left(\frac{N_0}{N}\log_2\frac{N_0}{N} + \frac{N_1}{N}\log_2\frac{N_1}{N}\right), \tag{1.3}$$

where $N = N_0 + N_1$, and $N_0$ and $N_1$ represent, respectively, the number of images with a value 0 and 1 at pixel location $i$. This value is minimised when all pixel values are either 0 or 1, which in line with the alignment of images. The objective function for overall alignment is then

$$\mathcal{L}_{\text{CONGEALING}} = R + \sum_{i=1}^{i=P} \hat{H}(x_i)\,, \tag{1.4}$$

where $R$ is a regularisation term to keep deformations small, usually the $\ell_2$ norm of the deformation vectors.

This work further demonstrates the potential applications of congealing to character recognition and registration of MRI images [Learned-Miller 2005]. The authors


Figure 1.9: The effect of congealing. (a) mean image from a set of handwritten 0s; (b) mean image after congealing; and (c-d) the same effect on a set of handwritten 2s. Image taken from [Learned-Miller 2006].

have shown further extensions to 3-dimensional data [Zöllei 2005] and more complex images [Huang 2007]. The latter work [Huang 2007] shows alignment results on images of cars and objects—object caterogies that much more complicated than handwritten characters—without using additional annotations. To exploit congealing algorithm's alignment capacity, this alignment uses SIFT [Lowe 2004] feature descriptors.

A closely-related work is Collection Flow [Kemelmacher-Shlizerman 2012] which proposes learning a deformable template using a collection of face photos taken *in-the-wild*, downloaded from the Internet. This powerful approach manages to learn deformable templates using repeated applications of PCA and optical flow. They show that normally about $k = 4$ first principal components of the design matrix (formed using pixel values of images in the collection) are good enough to generate an expression-neutral template and capture illumination changes. However, using higher principal components captures expressions as well as expression changes are less dominant than illumination changes. Then, optical flow and projection onto the principal components are used iteratively to remove variability due to expressions and retain sharper images at the same time. Their results Figure 1.10 show that they are able to discover a high-resolution template, as well as synthesize images by interpolating between representations of two images.

Figure 1.10: Discovering a template face from a collection of faces using Collection Flow. *left*: A set of input images; *centre*: input images warped to align with the template; and *right*: interpolating between two photos (extreme left and extreme right). Images taken from the paper.

#### 1.1.2.2 Active Shape Models

A method of discovering the mean shape and modes of variation was introduced in [Cootes 1992], called Point Distribution Models (PDMs). PDMs work by consuming

accurate human annotations of keypoints on objects, and discovering a mean shape and primary deformations. The method first aligns the images using an iterative process, and then computes the modes of variation using the covariance matrix of per-instance deformation from the mean shape. Active Shape Models (ASMs) [Cootes 1995] take this idea further by finding instances of objects in images which conform with the learnt shape using the training set. This is done by iteratively improving pose, shape, and scale parameters of the underlying PDM by examining the region around each point of an initial estimate of the object in the image. ASMs, however, do not model the appearance explicitly.

#### 1.1.2.3 Active Appearance Models

Active apperance models (AAMs) [Cootes 1998, Matthews 2004] model a set of images in the shape, as well as the apperance domains. *Independent* AAM*s* model the shape and appearance independently by parametrising them with different sets of parameters. In this parametrisation, shape and appearance are modelled as a base vector plus a linear combination of several vectors, each determining one axis of variation. More concretely, the appearance, $\mathcal{A}$, and shape, $\mathcal{S}$ are written as

$$\mathcal{A}(\mathbf{p}) = \mathcal{A}_0(\mathbf{p}) + \sum_{i=1}^{N_A} a_i \mathcal{A}_i(\mathbf{p}) \text{ , and} \tag{1.5}$$

$$\mathcal{S} = \mathcal{S}_0 + \sum_{j=1}^{N_S} s_j \mathcal{S}_j \text{ ,} \tag{1.6}$$

where $\mathbf{p}$ represents a pixel location, and $a_i$ and $s_j$ represent the appearance and shape parameters, respectively. In this thesis, we will also refer to them as *mixing coefficients*. Fitting an AAM to an image boils down to optimising these parameters with respect to the image. The vectors $\mathcal{A}_i$ and $\mathcal{S}_j$ are inferred from a set of hand-labelled training images. The annotations highlight a set of landmark locations in the images, and hence represent a set of corresponding points in any subset of training images. We will denote by $\mathcal{V}_i^j \in \mathbb{R}^2$, the $i$-th landmark in the $j$-th training image. Further, let $\mathcal{V}_i = \{\mathcal{V}_i^j \mid j = 1, 2, \ldots, M\}$.

The shape vectors $\mathcal{S}_j$ are obtained first. We begin by aligning $\mathcal{V}_i$ over the training set using generalised Procrustes analysis [Gower 1975] in order to remove large variations arising from global translations, scalings, and rotations of shape. We wish to constrain the shape vectors to learn local non-rigid deformations only. We can then apply Principal Component Analysis (PCA) to the *normalised* points, thus recovering principal axes of variations that ultimately become $\mathcal{S}_j$, and a *base shape* $\mathcal{S}_0$. To compute appearance vectors, all images are warped to align with the base shape, thus removing variations due to shape, followed by principal component analysis to discover vectors $\mathcal{A}_i$ and a *base appearance* $\mathcal{A}_0$.

*Combined* AAM*s* use the same set of parameters to model both, shape and

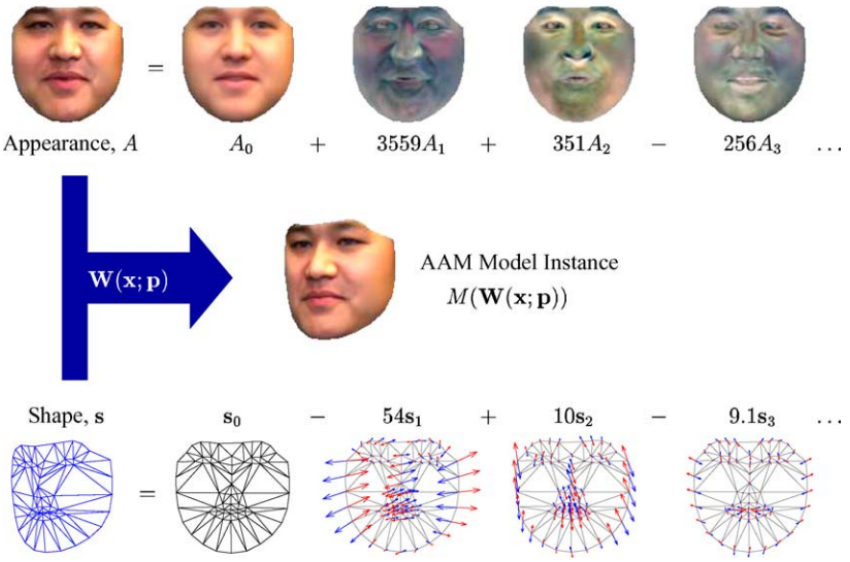



Figure 1.11: Modelling a face using an AAM. Appearance and shape are represented as linear combinations of learnt appearance and shape vectors. The final image is obtained by warping the appearnce with a deformation grid resulting from the shape parameters. Image taken from respective papers.

appearance. In the combined AAM model,

$$\mathcal{A}(\mathbf{p}) = \mathcal{A}_0(\mathbf{p}) + \sum_{i=1}^{N} c_i \mathcal{A}_i(\mathbf{p}) \,, \text{ and} \tag{1.7}$$

$$\mathcal{S} = \mathcal{S}_0 + \sum_{i=1}^{N} c_i \mathcal{S}_i \,. \tag{1.8}$$

**Learning AAMs.** Better AAM learning algorithms have been explored in literature. [Walker 2002] use salient points on the object (for example, eyes, nose, etc. for human faces) as landmarks to discover the shape basis, in order to side-step the labelling of images. However, the feature extraction process is less robust than human annotations. The work of [Kokkinos 2007] proposes a EM-based approach to learn objects as a combination of deformable parts with their relationship being modelled using a Markov random field. [Baker 2004] propose an encoding-decoding process, where the decoding corresponds to the image generation process, and the encoding to the fitting process. They formulate the objective so as to optimise the AAM appearance and shape bases, as well as fitting parameters jointly. Our work in Chapter 2 follows a similar approach where we use deep autoencoders in our encoding-decoding formulation.

**Fitting AAMs.** The fitting of AAMs comprises optimising the parameters with respect to a goodness-of-fit criterion, which is normally the squared error in pixel intensities [Cootes 1998, Matthews 2004]. This is a non-linear optimisation problem. Solutions have been tried with gradient descent and its varaints [Sclaroff 1998,

Blanz 2003b, Jones 1998]. These, however, are usually because of extensive computations involved.

Improvements to these have been proposed which model the parameter increments as a linear function of the error, so that computing further iterates in the optimisation algorithm does not involve computing high dimensional hessian matrices [Cootes 2004]. They assume that these linear operators are not dependent on model parameters and hence can be pre-computed. It is however, difficult to determine these operators, and hence are computed by fitting a linear regression model to data points generated by perturbing model parameters and observing the error [Edwards 1998].

Another fitting algorithm is Lucas-Kanade [Lucas 1981] applied to AAMs. It uses a Taylor series expansion of the error function to approximate it by a linear function in the parameters of the model. Other fitting algorithms include forward compositional alignment, inverse compositional alignment [Matthews 2004]. Extensions of this algorithm were also proposed in [Gross 2005, Papandreou 2008, Tzimiropoulos 2017]. [Donner 2006] proposed discovering correlations between model parameters and residuals of the synthesised texture image using canonical correlations analysis (CCA) to speed up the convergence of the parameter search by a factor of four.

#### 1.1.2.4 3D Morphable Models

The seminal work of [Blanz 1999] was the first to introduce high resolution models for human face understanding. In this work, the authors proposed building a 3D surface-based model which is obtained by observing densely annotated keypoints on high-resolution laser scans of 200 human faces. The dense sampling gives a much more refined surface geomtry than sparsely sampled keypoints, as is the case in AAMs. As a result, the authors are able to build a triangulated 3D surface to fit human faces. Further, applying PCA similarly as in AAMs, we can also obtain a shape and texture basis from the 200 scans, and thus express the morphable model as

$$\mathbf{S}_{\text{model}} = \bar{\mathbf{S}} + \sum_{i=1}^{m-1} \alpha_i \mathbf{s}_i \,; \text{ and} \tag{1.9}$$

$$\mathbf{T}_{\text{model}} = \bar{\mathbf{T}} + \sum_{i=1}^{m-1} \beta_i \mathbf{t}_i \,. \tag{1.10}$$

Here, $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ are the mean shape and texture vectors, $\mathbf{s}_i$ and $\mathbf{t}_i$ are the shape and texture bases, and $\alpha_i$ and $\beta_i$ are the weighting coefficients for the $i$-th shape and texture basis vectors. This results in a generative model that can (a) express new shapes and views as a linear combination of certain basis vectors; and (b) disentangles shape and texture information into separate bases so that they can be manipulated independently. Extensions of the morphable models have further disentangled the pose and expression components of shape, as well as illumination

and colour components of texture.

In the work of [Blanz 1999], model fitting is done by formulating a similarly function between the pixel values of the rendered image and the observed image. The rendered image is obtained by a given set of values for the parameters $\alpha_i, \beta_i$, as well as pose and illumination parameters. Perspective projection is used in the rendering pipeline, and the optimisation process starts with an initial estimate of the projection parameters from the user, and is done using stochastic gradient descent. The authors show in a subsequent work [Blanz 2003b] how the 3D morphable model can be used for face recognition by representing a face with the fitted models' parameters, while further works showed other possible uses [Romdhani 2005, Heisele 2007, Allen 2003, Blanz 2003a, Leopold 2001].

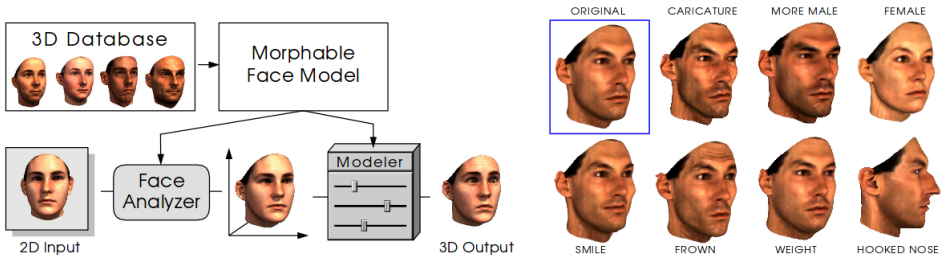



Figure 1.12: *left*: Construction of a 3D morphable model from a database, and fitting it to a 2D image; *right*: face manipulation using the fitted morphable model. Images taken from [Blanz 1999].

There have been further works proposing other morphable models, typically varying in the way the group-wise alignment is performed, the number of people scanned, the diversity in the scanned population, and method of collecting the annotations. [Blanz 1999] performs the alignment using few feature points marked on the surfaces. [Patel 2009] propose to instead use generalised Procrustes analysis to find a more robust estimate of the mean shape, which results in a better deformation basis. Further, to find dense correspondences between faces, they use few manually-annotated reliable landmarks followed by thin-plate splines to warp each scan to the mean landmarks. [Paysan 2009] construct a new morphable model, called the Basel face model (BFM) by aligning the scans with a template using the nonrigid iterative closest point algorithm [Amberg 2007]. This method has also been shown to work for registration of 3D medical volumes [Liang 2018].

Several works use BFM as a 3DMM for varied tasks in human face understanding. [Genova 2018] use CNNs to regress fitting parameters using auxiliary guiding losses. [Higgins 2017, Chen 2016b] learn disentangled representations using BFM. [Zhu 2016, Zhu 2017b] improve morphable model fitting over a large range of poses, with profile views up to 90°, and use the BFM's identity basis.

Contemporary morphable model offer powerful controllable representations of faces, by separating shape and expression [Cao 2013], shape, expression, and appearance [Gerig 2018], shape, expression, and pose [Li 2017]. The Surrey face model [Huber 2016] offers a multi-resolution mesh of either $3,448$, $16,759$, or $29,587$ vertices,

with annotated landmarks. A more complete list of contemporary 3D morphable models can be found at [Community 2019, Egger 2019].

**3DMM fitting.** Fitting models based on pixel values alone can be unrealiable as the texture does not always convey information about the location. For instance, the texture on the forehead comprises a wide area is mostly uniform, while that on the cheeks can be uniform too, under certain lighting conditions. Methods avoiding this problem focus on feature-based error functions instead of pixel intensity-based ones. [Sanyal 2019] detect landmark points detected using a 2D landmark detector and use the reprojection loss. [Romdhani 2005] use edges and specular highlights as further cues in the energy function.

Deep networks have been shown to learn powerful representations from data using non-linear encoders. As a straightforward application, several works have shown that deep networks can be used to regress 3DMM parameters directly, without the need for iterative optimisation, in supervised, weakly-supervised and self-supervised settings. [Richardson 2016] use synthetic examples with known geomtry to train a CNN to predict geomtry from a single image. [Tewari 2017] remove the synthetic images constraint and also introduce a self-supervised approach encode semantically significant image features like pose, shape, expression, reflectance, and illumination. They extend this notion to videos [Tewari 2018] and show that adding temporal information on identity greatly improves reconstruction quality. [Genova 2018], on the other hand, add the identity constraint to disentanglement by rendering synthetic examples from different poses using the regressed morphable model parameters. [Güler 2017] take a new approach to model fitting by regressing not the morphable model parameters, but dense UV coordinates using a quantised regression technique. [Bulat 2017b] construct a 2D-to-3D network to predict landmarks in 3D using a novel 3D Face Alignment Network (FAN) and propose a new dataset for 3D facial landmarks [Bulat 2017a].

**Shape from Non-Rigid Structure from Motion .** Non-rigid structure-from-motion (NRS$f$M) models shape as a linear combination of a set of basis vectors. This relaxes the rigid constraint imposed by classical structure-from-motion. One of the first works to use NRS$f$M for shape prediction, [Bregler 2000], formulates the system as a factorisation problem to separate motion into pose and shape. They write the shape $\mathbf{S}$ at an instant of time $t$ as a linear combination of $K$ basis vectors,

$$\mathbf{S}^{(t)} = \sum_{i=1}^{K} l_i^{(t)} \mathbf{s}_i \,, \tag{1.11}$$

where $\{\mathbf{s}_i\}$ is the shape basis, and $l_i^{(t)}$ are shape coefficients. The points in the basis shapes, at time $t$, under a scaled orthographic projection defined by rotation matrix

$\mathbf{R}^t = \begin{bmatrix} r_1^{(t)} & r_2^{(t)} & r_3^{(t)} \\ r_4^{(t)} & r_5^{(t)} & r_6^{(t)} \end{bmatrix}$ can be written as

$$\begin{bmatrix} \mathbf{u}^{(t)} & \mathbf{v}^{(t)} \end{bmatrix}^\top = \begin{bmatrix} l_1^{(t)}\mathbf{R}^{(t)} & l_2^{(t)}\mathbf{R}^{(t)} & \dots & l_k^{(t)}\mathbf{R}^{(t)} \end{bmatrix} \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \dots \\ \mathbf{s}_K \end{bmatrix}. \tag{1.12}$$

By concatenating these matrices over all time instances $t$, we get

$$\mathbf{W} = \begin{bmatrix} \mathbf{u}^{(1)} & \mathbf{v}^{(1)} & \mathbf{u}^{(2)} & \mathbf{v}^{(2)} & \dots & \mathbf{u}^{(t)} & \mathbf{v}^{(t)} \end{bmatrix}^\top \tag{1.13}$$

$$= \begin{bmatrix} l_1^{(1)}\mathbf{R}^{(1)} & l_2^{(1)}\mathbf{R}^{(1)} & \dots & l_k^{(1)}\mathbf{R}^{(1)} \\ l_1^{(2)}\mathbf{R}^{(2)} & l_2^{(2)}\mathbf{R}^{(2)} & \dots & l_k^{(2)}\mathbf{R}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ l_1^{(t)}\mathbf{R}^{(t)} & l_2^{(t)}\mathbf{R}^{(t)} & \dots & l_k^{(t)}\mathbf{R}^{(t)} \end{bmatrix} \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \dots \\ \mathbf{s}_K \end{bmatrix} \tag{1.14}$$

$$= \mathbf{Q}\mathbf{B} = \hat{\mathbf{Q}}\mathbf{G}\mathbf{G}^{-1}\hat{\mathbf{B}}. \tag{1.15}$$

where $\mathbf{G}$ is used to enforce orthonormality in the rotation by solving the least squares problems

$$\begin{bmatrix} r_1^{(t)} & r_2^{(t)} & r_3^{(t)} \end{bmatrix} \mathbf{G}\mathbf{G}^\top \begin{bmatrix} r_1^{(t)} & r_2^{(t)} & r_3^{(t)} \end{bmatrix}^\top = 1, \tag{1.16}$$

$$\begin{bmatrix} r_4^{(t)} & r_5^{(t)} & r_6^{(t)} \end{bmatrix} \mathbf{G}\mathbf{G}^\top \begin{bmatrix} r_4^{(t)} & r_5^{(t)} & r_6^{(t)} \end{bmatrix}^\top = 1, \tag{1.17}$$

$$\begin{bmatrix} r_1^{(t)} & r_2^{(t)} & r_3^{(t)} \end{bmatrix} \mathbf{G}\mathbf{G}^\top \begin{bmatrix} r_4^{(t)} & r_5^{(t)} & r_6^{(t)} \end{bmatrix}^\top = 0. \tag{1.18}$$

$\mathbf{W}$ is then factorised to obtain pose and shape by observing that the part of the matrix $\mathbf{Q}$ corresponding to each time instance is of rank 1. [Xiao 2004] later showed that this factorisation is not necessarily unambiguous, as one cannot recover basis shapes and shape coefficients uniquely. They further propose to add orthonormality constraints on the basis shapes to overcome this ambiguity. [Akhter 2009] later showed that this does not necessarily lead to an ambiguous shape, but suggested that the difficulty in achieving good 3D reconstruction is rather due to the optimisation and not the orthogonality constraints. [Dai 2014] proposed an optimisation algorithm to address this difficulty without using additional priors like orthonormality constraints on the basis and rotation. Other notable works focussing on improving the optimisation process can be found in [Paladini 2009, Garg 2013].In other contemporary works, to cope with the ill-posed nature of the 3D reconstruction problem as well as enforce restrictions on the shape, [Torresani 2008] use probabilistic PCA to estimate the shape basis. They take advantage of the robustness of probabilistic PCA towards missing data [Tipping 1999]. They further propose an EM algorithm to estimate the underlying probabilistic model to estimate motion and shape. [Russell 2011] formulate the NRS$f$M task as a labelling problem. In their formulation,

deformations are modelled by piece-wise models, with points being explained by them, where the labelling determines which model explains a point. They further allow labellings where one point might be explained by several models, so as to allow overlap between adjacent models and enforce global uniformity. [Garg 2013] also model the shape with a low-rank representation, but instead of fixing the number of basis shapes, they learn a rank-minimised matrix. They also introduce total variation smoothness constraints. [Yu 2015] further incorporate strong cues using optical flow.

Other prominent works extending NRS$f$M ideas have since been proposed. [Carreira 2016] proposed lifting object categories from 2D to 3D for object detection. They use ground-truth segmentations with annotated keypoints to retrieve category-specific 3D reconstructions. [Kanazawa 2018b] proposed a system that decodes 3D structure of an object category from an image, using a photometric loss coupled with a keypoints projection loss. They demonstrate reconstruction results on the birds and show that their system is able to further learn meaningful deformation components. [Garrido 2016] combine photometric stereo, optical flow, and multi-view stereo and solves them together to generate detailed 3D reconstructions. [Liu-Yin 2017] use a similar strategy, where they use a non-Lambertian model to predict shape jointy with the reconstruction objective.

#### 1.1.2.5 Modelling deformations with deep neural nets

Over the last decade, several works have used deep learning to model deformations and alignment for computer vision and medical imaging problems. The pioneering work of [Jaderberg 2015] on spatial transformer networks was one of the first to explore the idea of introducing deformations in convolutional networks to improve classification and detection. Figure 1.13 shows a spatial transformer module that can be inserted into a feed-forward convolutional network. The module has a learnable part, denoted by $f_{\text{loc}}$, which regresses a set of deformation parameters $\theta$ from a feature map ($U$). $U$, being an intermediate feature map in the CNN, can be a multi-channel image, in which case the deformation is applied to all channels equally (the deformation is always 2D). The parameters $\theta$ are then used to generate a dense sampling grid, which is in turn used to warp the feature map using bi-linear sampling. As all of these operations are differentiable, the spatial transformer can be inserted between two layers of a CNN.

Instead of warping the feature map, some works propose introducing offsets into the following convolution operation. While this does mean the overhead of warping the feature map is reduced, it introduces an extra bi-linear sampling operation per offset just before the convolution operation, and can be difficult to implement in standard deep learning frameworks. [Jeon 2017] proposed learning a convolution-specific offset that is applied universally to all pixels. Deformable convolutional neural networks [Dai 2017] go a step further by relaxing the global constraint, as well by predicting the offsets dynamically depending on the feature map.

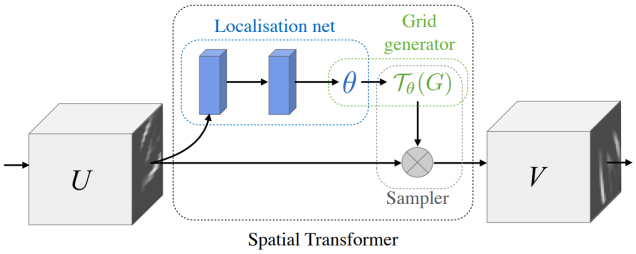



Figure 1.13: A spatial transformer module. The localisation net determines an appropriate deformation, which is then used to warp the feature map $U$ to produce a warped feature map $V$. Image taken from [Jaderberg 2015].

### 1.1.3 Deep Learning for Medical Imaging

Since the break-through paper from Krizhevsky and Hinton [Krizhevsky 2012] demonstrating the potential of deep learning for computer vision tasks, there has been keen interest on targetting medical imaging problems with deep learning [Ronneberger 2015, Seyedhosseini 2013, Badrinarayanan 2017, Milletari 2016, Çiçek 2016, Kamnitsas 2017]. [Ronneberger 2015] proposed the fully-convolutional semantic segmentation network called U-Net (Figure 1.14) which achieved state-of-the-art results on cell segmentation and cell tracking problems. They use skip connections between the encoder and decoder. There are several demonstrated applications of U-Net to other medical image segmentation problems. Some of these include [Xiancheng 2018] for blood vessel segmentation in retina scans, [Skourt 2018] for lung CT segmentation, and [Christ 2017] for liver and tumor segmentation. Further improvements on the U-net have also been proposed. [Milletari 2016] extend the U-Net approach to 3-dimensional data by applying the resulting V-net to the problem of prostate segmentation. They also introduce a soft Dice loss for training which works better than cross entropy, particularly for unbalanced regions. [Yu 2017] improve over their method using a similar model, but trained with deep supervision. In a further work, [Zhou 2018] propose augmenting the skip connections of the U-Net using dense convolutional connections, in the style of DenseNets [Huang 2017]. They further incorporate both, cross entropy and the soft Dice loss in their network, along with deep supervision.

Deep learning has seen a consistent rise in histopathology as well. [Chen 2014] propose cell detection on H&E images using an extension of the colour un-mixing technique of [Ruifrok 2001]. [Chen 2016a] propose lymphocyte detection using a deep neural network. However it does not generalise to tumor cells. Weakly-supervised approaches are widely used in analysis of histopathology images. Whole slide images with image-level annotations demand patch-based processing of the slide, as the entire slide is too big to fit into memory. However, being that the number of patches can vary and that patch-level annotations are usually not available, a weakly-supervised approach is best suited. For such a setting, [Hou 2016] propose an approach to find discriminative tiles in a whole-slide histopathology image. They

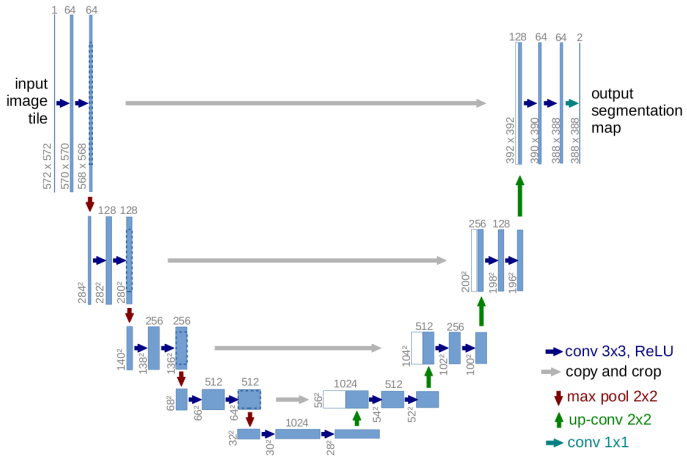



Figure 1.14: The fully-convolutional U-Net architecture for semantic segmentation. Image taken from paper.

model this as a hidden variable in an EM method coupled with a CNN. Another recent method from [Jia 2017] predicts cancerous regions in histopathology images in a multi-instance learning framework. They also employ weak supervision signals based on the area of the cancerous regions, and also a multi-stage cross entropy loss. However, for extremely large slides, existing methods are not enough because they cannot fit on a GPU or in memory. For such scenarios, [Xu 2017] develop a distributed computing approach that targets image classification, segmentation, and clustering problems for images with $\sim 10$ billion pixels.

Deep learning has also extensively been used with MRI and CT images. [Lu 2016] use CNNs for kidney localisation by aggregating inference from several models and local context obtained at three orthogonal orientations. [Thong 2018] use sliding window operations to segment kidneys using CNNs. Their CNN predicts the class scores for the central patch in a window. [Hussain 2017b] use multiple CNNs which observe the 3D volume from three directions and detect kidneys by aggregating inferences from the three models. They further use this detection to estimate the size of the renal volume. [Hussain 2017a] detect renal cell carcinoma from CT scans by arranging all slices from the 3D volume into a regular 2D grid. This enables transferring the volume-level label to the patch-level and hence, enables meaningful feature extraction using a CNN. [Alansary 2016] use a 3D multi-scale CNN followed by a dense CRF to segment placenta from MRI scans. They also propose a visualisation framework which converts the segmented placenta into a mesh-based textured surface representation. [Moeskops 2016] use a single CNN to segment MR brain images, MR breast images, and coronary arteries in cardiac CTA. Their multi-task CNN is able to achieve equal performance to task-specific CNNs.

#### 1.1.3.1 Unsupervised and weakly supervised approaches

One of the main difficulties encountered in medical imaging tasks is that there are few data available. This is because (a) the data are real specimens showing normal/abnormal organs and are not cheap or easy to acquire; and (b) annotations on the data need to be performed by experts working in the field [Hou 2016], unlike how crowd-sourced annotations are possible for other computer vision tasks [Güler 2018]. When coupled with the (usually) high dimensionality of the data (histopathology images, CT scans), researchers are faced with problems comprising generalisation, memory usage, and overfitting. To reduce the annotation task, few-shot learning [Li 2006, Larochelle 2008, Gidaris 2018] has also seen applicability to dataset generation [Pierrard 2019]. Currently, few methods in medical image analysis using deep learning focus on unsupervised learning and weakly-supervised learning.

In histopathology images analysis, dataset and annotation issues are particularly apparent. Whole-slide tissue images, when viewed at maximum magnification, can range from several ten to hundred thousand pixels per side [Hou 2016, Xu 2017]. To circumvent the difficulty in obtaining new samples as well as valuable time spent by pathologists annotating the collected samples, several recent works propose generating new datasets of training images using existing annotations, thereby increasing the training data size manyfold.

In mammography, [Kallenberg 2016] proposed an weakly supervised approach to estimate risk of breast cancer. Their method first extracts multi-scale sparse representations of mammograms using an autoencoder, followed by segmentation of breast density and scoring of mammographic texture. [Hwang 2016] propose a self-supervised CNN to classify and localise of abnormalities in mammograms. They do so without using pre-trained networks of any sort. In this architecture, the localisation and classification branches share a common feature extractor. The network is trained in an alternating fashion, with one branch fixed while the other is trained, thus "transfering" the learnt feature extractor between tasks.

Attention-based models, which operate in the setting of weak labels, try to solve the multi-instance learning problem of learning from weak, bag-level labels using attention mechanisms. A recent work [Ilse 2018] shows the application of an attention network to detection of epithelial cells in histopathology images. [Katharopoulos 2019] improve over their method by approximating the attention distribution and sampling it to improve classification.

## 1.2 Contributions of the Thesis

A list of contributions of this thesis follows. The first three entries below are contributions towards answering the first question on morphable model learning, followed by our contibutions on automated diagnosis of lymphocytosis.

**Unsupervised dense alignment.** A principal contributions of this thesis is unsupervised dense alignment on images of an object category. Dense correspondences

are essential for learning shape [Matthews 2004, Torresani 2008] and is also one of the stages in morphable model building [Blanz 1999, Paysan 2009]. However, generating dense correspondences using optical flow can be computationally expensive and can also limit the number of images we can use [Kemelmacher-Shlizerman 2013]. Generating dense correspondences using deep learning in an unsupervised manner can alleviate these issues. To this end, we propose deforming autoencoders (DAEs) to learn jointly, a canonical template space for an object category, as well dense correspondences between an object in the image space with the canonical space. The unsupervised alignment is evaluated against other state-of-the-art methods on landmark detection accuracy using the disentangled dense deformation grid.

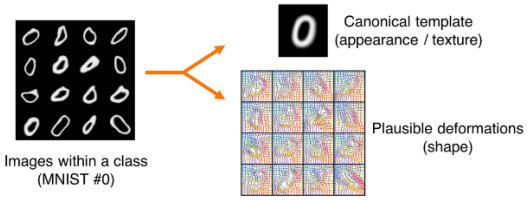



Figure 1.15: Discovering a canonical space with a DAE.

**3D shape using NRS*f*M and DAEs.** By treating the dense alignment obtained using DAEs as ground truth, we show that we can learn a 3D morphable model which learns a mean shape, a camera basis, and a deformation basis separately. This is also achieved in an autoencoder framework where the latent space represents the shape and camera coefficients, while the mean shape is represented by a mesh. The resulting model, which we call a lifting autoencoder (LAE) learns 3D shape from 2D dense correspondences using deep non-rigid structure from motion. We evaluate our shape prediction and pose prediction using a method that utilises Procrustes analysis for landmark alignment.

**Disentangled representations.** For both, the DAE and the LAE, disentanglement of object properties in the latent space is demonstrated. Specifically, shape, albedo, and shading are disentangled in the DAE's latent space, while identity, pose, expression, albedo, and shading are disentangled in the LAE's latent space. Controllable image synthesis can then be performed by manipulating one or more of the latent vectors. For the case of the LAE, this means visualising a shape from different poses, interpolating between identities and expressions of different faces, and transferring illumination from a source face image to a target face image.

**Deep multi-instance learning for automated diagnosis.** We introduce a deep convolutional model for automated prediction of the chance of tumoral lymphocytosis in a patient from blood smears only. A mixture-of-experts model is also evaluated to combine predictions from images and clinical data. Comparisons are performed with state-of-the-art deep multi-instance learning approaches and predictions of 12 biologists. We show that our method beats attention-based mechanisms,

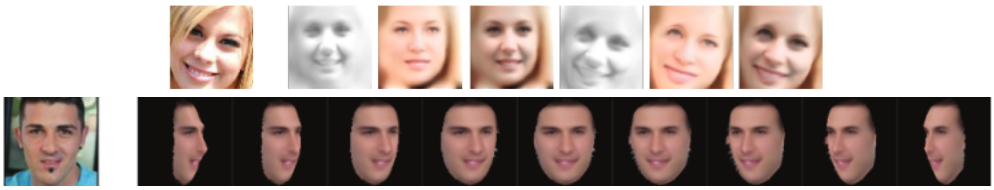


Figure 1.16: *top*: disentanglement of shape, albedo, and shading using DAE. From left to right: input image, shading in canonical space, albedo in canonical space, texture in canonical space, shading in image space, albedo in image space, texture in image space; emphbottom: disentanglement of expression and pose with LAE. For the input image on the left, visualisations of learnt 3D shape from different poses while keeping the expression intact.

as well as the average prediction of the biologists, thus making a strong case for deployment in real-world scenarios.

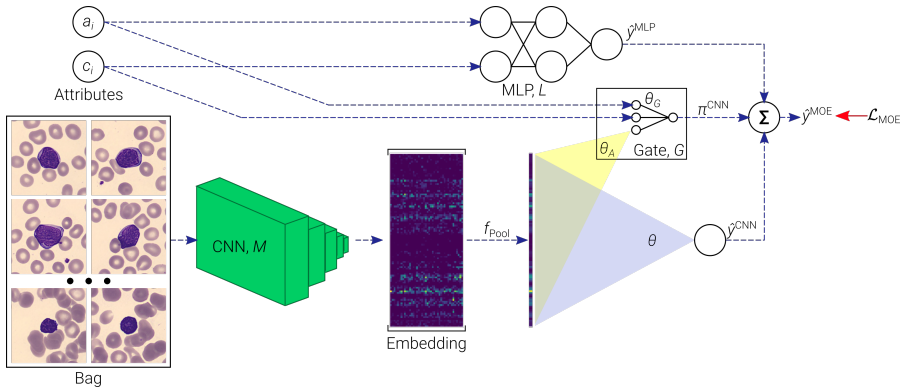



Figure 1.17: A mixture-of-experts model for the diagnosis of lymphocytosis.

## 1.3 Organisation of the Thesis

This thesis is organised as follows.

1. In Chapter 2, we introduce deforming autoencoders as a means of achieving unsupervised dense alignment of objects of a category in 2D. We also demonstrate shape, albedo, and shading disentanglement with DAEs. We introduce a simple and effective method of regressing the deformation grid using a convolutional neural network. We further show that this method performs well at registration tasks in medical imaging as well as remote sensing.

2. In Chapter 3, we introduce lifting autoencoders that build on DAEs and use the unsupervised dense alignment to lift objects from 2D to 3D in an unsupervised manner. We also show that by introducing pose, identity, and expression cues, we can learn a fully controllable 3D morphable model using minimal supervision.

3. In Chapter 4, we move our focus to medical imaging, and introduce a model for automatic prediction of tumoral lymphocytosis. We combine inference from images and clinical data in an end-to-end trainable deep neural network to capture the probability of the presence of cancer.

4. Finally, we conclude in Chapter 5 and present possible future work based on the contributions of this thesis.

## 1.4 List of Publications

1. Z. Shu, **M. Sahasrabudhe**, R. A. Güler, D. Samaras, N. Paragios, I. Kokkinos. *Deforming Autoencoders: Unsupervised Disentangling of Shape and Appearance*, ECCV 2018.

2. **M. Sahasrabudhe***, Z. Shu*, E. Bartrum, R. A. Güler, D. Samaras, I. Kokkinos. *Lifting Autoencoders: Unsupervised Learning of a Fully-Disentangled 3D Morphable Model using Deep Non-Rigid Structure from Motion* (oral), Geometry Meets Deep Learning Workshop, ICCV 2019.

3. **M.Sahasrabudhe**, P. Sujobert, E. Zacharaki, E. Maurin, B. Grange, L. Jallades, N. Paragios, M. Vakalopoulou, *Deep Multi-Instance Learning for Diagnosis of Lymphocytosis* (under submission). IEEE Journal of Biomedical And Health Informatics.

4. S. Christodoulidis, **M. Sahasrabudhe**, M. Vakalopoulou, G. Chassagnon, M.-P. Revel, S. Mougiakakou, N. Paragios. *Linear and Deformable Image Registration with 3D Convolutional Neural Networks* (oral), RAMBO workshop, MICCAI 2018.

5. M. Vakalopoulou, S. Christodoulidis, **M. Sahasrabudhe**, S. Mougiakakou, N. Paragios, *Image Registration of Satellite Imagery with Deep Convolutional Neural Networks* (oral), IGARSS 2019.

## 1.5 Dissemination Activities

- Code, data, and models available online at https://msahasrabudhe.github.io.

- Poster presentation at ECCV 2018.

- Oral and poster presentations at Geomtry Meets Deep Learning workshop, ICCV 2019.

## 1.6 Other Academic Activities

- Visiting researcher at Harvard Medical School from Oct to Dec 2019 (PI: Dr. Tomas Kirchhausen).
- Reviewer for the conferences JURSE 2019, ICANN 2019, and the journal CVIU.
- Course TA for Foundations of Machine Learning (2016-18), and Introduction to Visual Computing (2019) at CentraleSupélec.
- Course instructor for Programming and Languages at ESSEC and CentraleSupélec (2018-19).

Chapter 2

# Deforming Autoencoders: Unsupervised Dense Alignment in 2D

We first introduce the deforming autoencoder (DAE), a generative model for images that infers dense alignment between object categories by disentangling shape from appearance in an unsupervised setting.

## 2.1 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. This pushes this line of research by following the deformable template paradigm [Amit 1991, Yuille 1991, Cootes 1998, Blanz 2003b, Matthews 2004]. In particular, we consider that object instances are obtained by deforming a prototypical object, or 'template', through dense, 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 consider that both sources of variation can be modelled 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 the training, while also yielding control over the generative process.

Our main contributions in this work can be summarized as follows. First, we introduce the *deforming autoencoder* architecture, bringing together the deformable modelling 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 Figure 2.1, our network predicts this transformation field alongside with the template-aligned appearance and subsequently deforms the synthesized appearance to generate an image similar to its input. This allows for a disentanglement of the shape and appearance parts of image generation by explicitly modelling the effects of image deformation during the decoding stage.

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 allows us to constrain the synthesized field and template to be semantically meaningful, 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 comes with several advantages when it comes to modelling and manipulating images. By using disentangling we obtain clearly better synthesis results when manipulating images for tasks such as expression, pose or identity interpolation when compared to standard autoencoder architectures. Along the same lines, we show that accounting for deformations facilitates a further disentangling of the appearance components into an intrinsic, shading-albedo decomposition which completely fails when naively performed in the original image coordinates. This allows us to perform re-shading through simple operations on the latent shading coordinate space.

We complement these qualitative results with a quantitative analysis of the learnt 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 [Thewlis 2017c], even though we never explicitly trained our network for correspondence estimation, but rather only aimed at reconstructing pixel intensities.

## 2.2 Related Work

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 [Chen 2016b, Shu 2017, Worrall 2017, Sengupta 2017], low-dimensional transformations, such as rotations, translation, or scaling, [Memisevic 2010, Worrall 2016, Park 2017] or finer-levels of variation, including age, gender, wearing glasses, or other attributes e.g. [Shu 2017, Lample 2017] for particular

classes, such as faces.

Shape variation is more challenging as it amounts to a transformation of a function's domain, rather than its values. Even simple, supervised additive models of shape result in complex nonlinear optimization problems [Cootes 1998, Matthews 2004]. Despite this challenge several works in the previous decade aimed at learning shape/appearance factorizations in an unsupervised manner, exploring groupwise image alignment, [Learned-Miller 2006, Kokkinos 2007, Frey 2003, Jojic 2003]. In the context of deep learning several works have aimed at incorporating deformations and alignment in a supervised setting, including Spatial Transformers [Jaderberg 2015], Deep Epitomic Networks [Papandreou 2015], Deformable CNNs [Dai 2017], Mass Displacement Networks [Neverova 2018], Mnemonic Descent [Trigeorgis 2016], or Densereg [Güler 2017]. 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. [Memisevic 2010, Worrall 2016], several works have attempted to incorporate richer non-rigid deformations within learning. A thread of works has been aimed at dynamically rerouting the processing of information within the network's graph based on the input, starting from neural computation arguments [Hinton 1981, Olshausen 1995, Malsburg 1981] and eventually translating into concrete algorithms, such as the 'capsule' works of [Hinton 2011, Sabour 2017] that bind neurons on-the-fly. Still, these works lack a transparent, parametric handling of non-rigid deformations. Working on a more geometric direction, several works have recently aimed at recovering dense correspondences between pairs [Bristow 2015] or sets of RGB images, as e.g. in the recent works of [Zhou 2016b, Gaur 2017]. These works however do not have the notion of a reference coordinate system ('template') to which images can get mapped - this makes the image generation and manipulation harder. More recently, [Thewlis 2017c] use the equivariance principle in order to align sets of images to a common coordinate system, but do not develop this into a full-blown generative model of images.

## 2.3 Deforming Autoencoders

Our architecture embodies the deformable template paradigm in an autoencoder architecture. The premise of our work 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 consider that the observed image, $I(\mathbf{p})$ can be reconstructed as follows:

$$I(\mathbf{p}) \simeq T(W(\mathbf{p}), \tag{2.1}$$

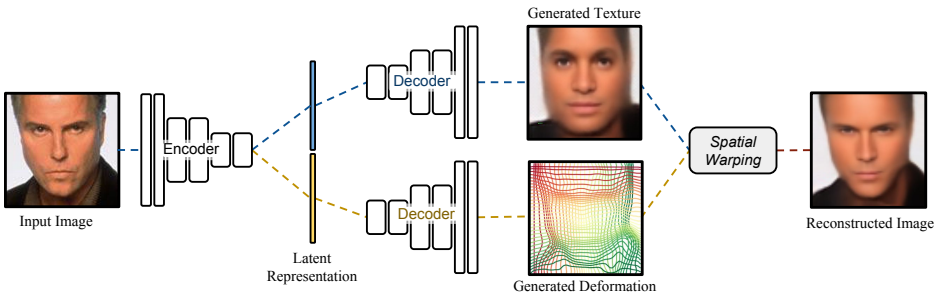



Figure 2.1: Deforming Autoencoders follow the deformable template paradigm and model image generation through a cascade of appearance (or, *texture*) synthesis in a canonical coordinate system and a spatial deformation that warps the texture to the observed image coordinates. By keeping the latent vector for texture short the network is forced to model shape variability through the deformation branch, so as to minimize a reconstruction loss. This allows us to train a deep generative image model that disentangles shape and appearance in an entirely unsupervised manner.

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 spatial transformer layer [Jaderberg 2015] 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, $\mathbf{Z}$, of shape and appearance. This is split into two parts, $\mathbf{Z} = [\mathbf{Z}_T, \mathbf{Z}_S]$ which feed into the appearance and shape networks respectively, providing us with a clear separation of shape and appearance.

### 2.3.1 Deformation Field Modelling

Rather than leave deformation modelling 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 [Jaderberg 2015]. This means that the shape representation, $\mathbf{Z}_S$ described above is decomposed into two parts, $\mathbf{Z}_W, \mathbf{Z}_A$, where $\mathbf{Z}_W$ accounts for the non-rigid part of the deformation field, and $\mathbf{Z}_A$ for the affine. 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. This is also a common decomposition in deformable models for faces [Cootes 1998, Matthews 2004].

Turning to local deformation effects, we quickly realized that not every deformation field is plausible. Without appropriate regularization we would often obtain deformation fields that could expand small areas to occupy whole regions,

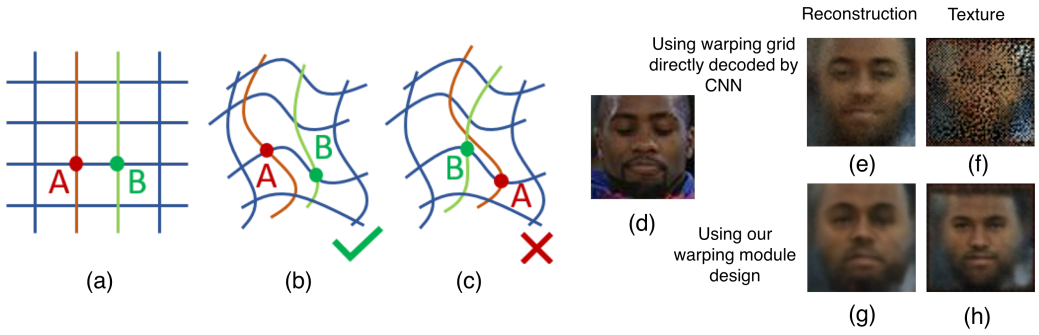



Figure 2.2: Our warping module design only permits locally consistent warping, as shown in (b), while the flipping of relative pixel positions, as shown in (c), is not allowed by design. To achieve this, we let the deformation decoder predict the horizontal and vertical increments of the deformation ($\nabla_x W$ 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.

and/or would be non-diffeomorphic, meaning that the deformation could spread a connected texture pattern to a disconnected image area (Figure 2.2-(f)).

To prevent this problem, instead of making the shape decoder CNN directly predict 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 $\nabla_c$ denotes the $c$-th component of the spatial gradient vector. These two quantities measure the displacement of consecutive pixels - for instance $\nabla_x W_x = 1$ amounts to translation in the horizontal axis, $\nabla_x W_x = 2$ amounts to horizontal shifting by a size of 2, while $\nabla_x W_x = -1$ amounts to left-right flipping; a similar behavior is associated with $\nabla_y W_y$ 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 outputs of our differential decoder through a Rectified Linear Unit (ReLU) module, 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 in terms of a fixed network layer, on top of the output of the ReLU layer to reconstruct the warping field from its spatial gradient. By doing so, the new deformation module enforces the generation of smooth and regular warping fields that avoid self-crossings. In practice we found that also clipping the decoded offsets by a maximal value significantly eases the training, which amounts to replacing the ReLU layer, $\text{ReLU}(x) = \max(x, 0)$ with a $\text{HardTanh}_{0,\delta}(x) = \min(\max(x, 0), \delta)$ layer. In our

experiments, we set $\delta = 5/W$ where $W$ denotes the number of pixels along one dimension of the image.

### 2.3.2 Class-aware Deforming Autoencoder

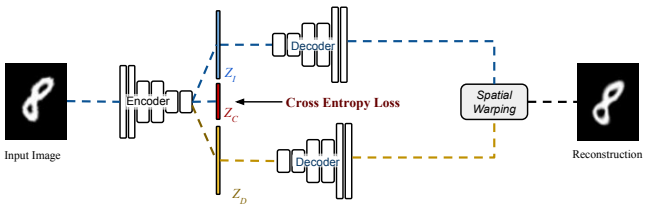



Figure 2.3: A *class-aware* model can account for multi-modal deformation distributions by utilizing class information. Introducing a classification loss into latent space helps the model learn a better representation of the input as demonstrated on MNIST.

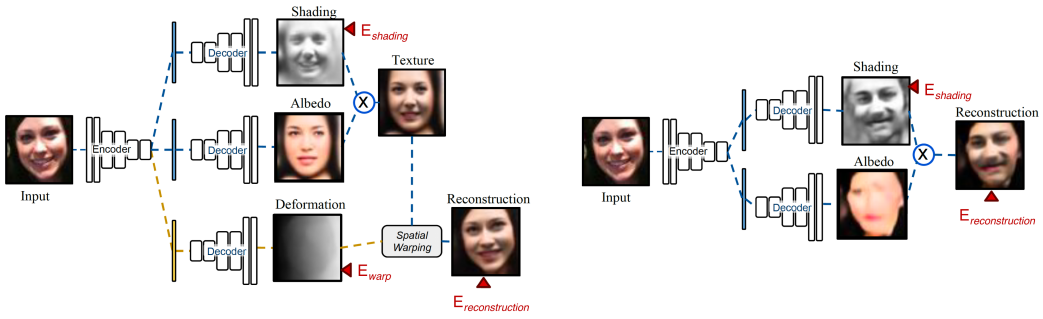
We can require our network's latent representation to be predictive of not only shape and appearance, but also of instance class, if that is available during training. We note that this information, being discrete may be easier to acquire than the actual deformation field, which would require manual 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: $\mathbf{Z} = [\mathbf{Z}_T, \mathbf{Z}_C, \mathbf{Z}_S]$, where $\mathbf{Z}_T, \mathbf{Z}_S$ are as previously the appearance- and shape- related parts of the representation, respectively, while $\mathbf{Z}_C$ 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 $\mathbf{Z}_C$ is fed into both the appearance and shape decoders. 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.

### 2.3.3 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, named Intrinsic-DAE to model shading and albedo for in-the-wild face images. As shown in Fig. 2.4-(a), we introduce two separate decoders for shading $S$ and albedo $A$, each of with 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 [Shu 2017]: we apply the L2 smoothness loss on $\nabla S$, meaning that shading is expected to be smooth, while leaving albedo unconstrained. As shown in Fig. 2.4 and more extensively in the experimental results section, when used in tandem with an Deforming Autoencoder this allows us to successfully decompose of face image into shape, albedo, and shading components, while a standard Autoencoder completely fails at decomposing unaligned images into shading and albedo.




(a) Deforming autoencoder with intrinsic decomposition (b) Autoencoder with intrinsic decomposition

Figure 2.4: Autoencoders with intrinsic decomposition. (a) Deforming Autoencoder with intrinsic decomposition (Intrinsic-DAE): we model the texture by the Hadamard 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 reconstruction loss ($\mathcal{L}_{\text{RECONSTRUCTION}}$) on the final output and regularization losses on shading ($\mathcal{L}_{\text{SHADE}}$) and deformation ($\mathcal{L}_{\text{WARP}}$), if it exists.

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

$$\mathcal{L}_{\text{DAE}} = \mathcal{L}_{\text{RECONSTRUCTION}} + \mathcal{L}_{\text{WARP}}\,, \tag{2.2}$$

where the reconstruction loss is defined as the standard $\ell_2$ loss

$$\mathcal{L}_{\text{RECONSTRUCTION}} = \|\hat{I} - I\|^2, \tag{2.3}$$

and the warping loss is decomposed as follows:

$$\mathcal{L}_{\text{WARP}} = \mathcal{L}_{\text{SMOOTH}} + \mathcal{L}_{\text{BIASREDUCE}} \tag{2.4}$$

In particular the smoothness cost, $\mathcal{L}_{\text{SMOOTH}}$, 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

$$\mathcal{L}_{\text{SMOOTH}} = \lambda_1 \left( \|\nabla W_x(x,y)\|_1 + \|\nabla W_y(x,y)\|_1 \right) , \tag{2.5}$$

where $\lambda_1 = 10^{-6}$. Finally, $\mathcal{L}_{\text{BIASREDUCE}}$ 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, given by

$$\mathcal{L}_{\text{BIASREDUCE}} = \lambda_2 \|S_A - S_0\|^2 + \lambda_2' \|\bar{W} - W_0\|^2 . \tag{2.6}$$

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 Sec. 2.3.2 we augment the loss above with the cross-entropy loss evaluated on the classification network's outputs.

For Intrinsic-DAE, we add the following objective function in training:

$$\mathcal{L}_{\text{SHADE}} = \lambda_3 \|\nabla S\|^2, \text{ where } \lambda_3 = 10^{-6} . \tag{2.7}$$

We experiment with two types of architectures; the majority of our results are obtained with a standard auto-encoder architecture, 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 Appendix A.1.2.

Follow the recent work on densely connected convolutional networks [Huang 2017], we have also experimented with incorporating dense connections into our encoder and decoders architectures respectively (no skip connections over the bottleneck layer for latent representations). In particular, we follow the architecture of DenseNet-121, but without the $1 \times 1$ convolutional layers inside each dense block. These have been shown to better exploit larger datasets, as indicated in the quantitative analysis of unsupervised face alignment. We call this version of the deforming autoencoder Dense-DAE.

## 2.4 Experiments

To demonstrate the properties of our deformation disentangling network, we conduct experiments on the following three datasets:

- Deformed MNIST. A synthetic dataset designed specifically to explore the deformation modelling power of our network. Deformed MNIST consists of handwritten MNIST images randomly distorted using a mixture of sinusoidal waveforms.

- MUG facial expression dataset [Aifanti 2010]. This dataset consists of videos of individuals performing facial expressions, with simple blue background and


Figure 2.5: 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 $\mathbf{Z}_T$ is 1.


Figure 2.6: Class-aware deforming autoencoders effectively model the appearance and deformation for multi-class data.

minor translation. The dataset also offers frames from the videos, classified according to the facial expression, as well as the subject.

- Faces-in-the-wild dataset: MAFL [Zhang 2014b] and CelebA [Liu 2015]. These datasets consist of uncontrolled "in-the-wild" faces with variability in pose, illumination, expression, age, etc.

Using these datasets we experimentally explored the ability of the unsupervised appearance-shape (or texture-deformation) disentangling network on 1) unsupervised image alignment/appearance inference; 2) learning semantically meaningful manifolds for shape and appearance; 3) decomposition into illumination intrinsics (shading, albedo); 4) unsupervised landmark detection, as detailed below. We intend to make all of the code of our system publicly available in order to facilitate the reproduction of our results.

### 2.4.1 Unsupervised Appearance Inference

We first use our network to model canonical appearance and deformation for single category objects. For this purpose, we demonstrate the results in the MNIST and MUG facial expression datasets (Fig. 2.5, 2.6, 2.7).

(a) Learning single template from image sequence

(b) Expression interpolation

Figure 2.7: Experiment on MUG dataset of face expressions: (a) With 0-length $\mathbf{Z}_T$, Deforming Autoencoders learn a single texture (row 3) from a subject in the MUG facial expression dataset. By doing so, the subject's facial expression is encoded only in the deformation domain. (b): Our network is able to disentangle the facial expression deformation and encode this information in a meaningful latent representation. By interpolating the latent deformation representation from the source (in orange) to the target (in blue), it generates sharp images and a smooth deformation interpolation between expressions as shown in each row.

We observe that by heavily limiting the size of $\mathbf{Z}_T$ (1 in Fig. 2.5 and 0 in Fig. 2.7), we can successfully infer a canonical appearance for such a class. In Fig. 2.5, all different types of handwritten digits '3' are aligned to a simple canonical shape. In Fig. 2.7, by limiting the dimension of $\mathbf{Z}_T$ to 0, the network learns to encode a single texture image for all expressions, and successfully distills expression-related information exclusively in the shape space. In Fig. 2.7-(b) we show that by interpolating the learnt latent representations, we can generate meaningful shape interpolations that mimic facial expressions.

In cases where data has a multi-modal distribution exhibiting multiple different canonical appearances, e.g., multi-class MNIST digit images, learning a single appearance is less meaningful and often challenging (Fig. 2.6-(b)). In such cases, utilizing class information (Sec. 2.3.2) significantly improves the quality of multi-modal appearance learning (Fig. 2.6-(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. 2.6-(c,d) is interpreted as a 6.

We now demonstrate the effectiveness of texture inference using our network on in-the-wild human faces. 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. 2.10)-(d). The aligned textures retain the information of the input image such as lighting, gender, and facial hair, without a relevant supervision training signal. We further demonstrate the alignment on the 11k Hands dataset [Afifi 2017], where we align palmar images of the left hand of several subjects 2.8. 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.

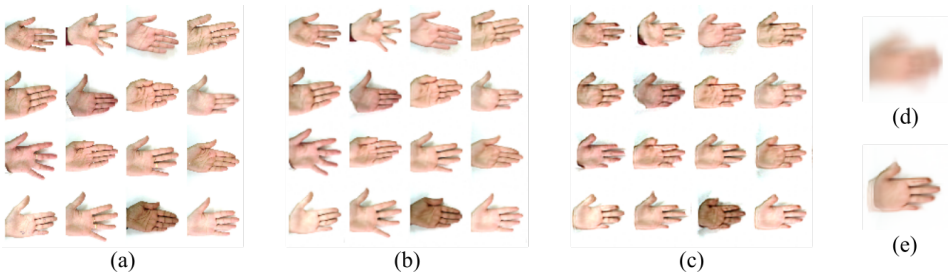



Figure 2.8: Unsupervised alignment on images of palms of left hands. (a) The input images; (b) reconstructed images; (c) texture images warped with the average of the decoded deformation; (d) the average input image; and (e) the average texture.

### 2.4.2 Autoencoders vs. Deforming Autoencoders

We show the ability of our network to learn meaningful deformation representations without supervision. We compare our disentangling network with a plain auto-encoder (Fig. 2.9). 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, i.e., the bottleneck.

We train both networks in the MAFL faces-in-the-wild dataset. To evaluate the learnt 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 $\mathbf{Z}$s. We use $\mathbf{Z}_T(I^s)$ and $\mathbf{Z}_S(I^s)$ to denote the latent representations in our network for $I^s$, and $\mathbf{Z}_{\text{AE}}(I^s)$ for the latent representation learnt by a plain autoencoder. We then conduct linear interpolation on $\mathbf{Z}$, between $\mathbf{Z}^s$ and $\mathbf{Z}^t$,

$$\mathbf{Z}^\lambda = \lambda \mathbf{Z}^s + (1-\lambda)\mathbf{Z}^t\,. \tag{2.8}$$

We subsequently reconstruct the image $I^\lambda$ from $\mathbf{Z}^\lambda$ using the corresponding decoder(s), as shown in Figure 2.9.

By traversing the learnt deformation representation only, we can change the shape and pose of a face while maintaining its texture (Figure 2.9-(1)); interpolating the texture representation results in pose-aligned texture transfer (Figure 2.9-(2)); traversing on both representations will generate a smooth deformation from one image to another (Figure 2.9-(3,5,7)). Compared to the interpolation using the autoencoder (Figure 2.9-(4,6,8)), which often exhibits artifacts, our traversal stays on the semantic manifold of faces and generates sharp facial features.

### 2.4.3 Intrinsic Deforming Autoencoders

Having demonstrated the disentanglement abilities of Deforming Autoencoders, we now explore the disentanglement capabilities of Intrinsic-DAE described in Sec.

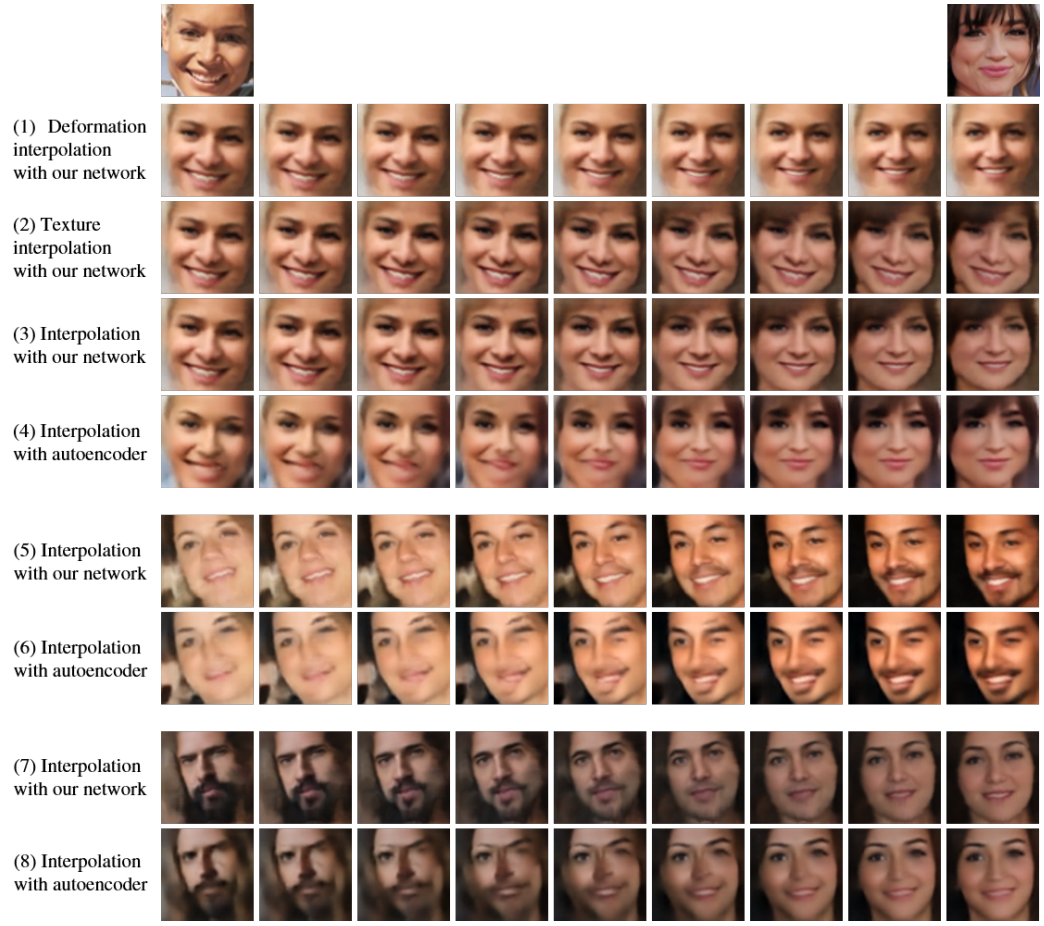



Figure 2.9: Latent representation interpolation: we embed a face image in the latent space provided by an encoder network trained on the MAFL dataset. Our network disentangles the texture and deformation in the respective parts of the latent representation vector, allowing a meaningful interpolation 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. For more results, please see Figure A.5,A.6 in Appendix.

2.3.3. Using only the $\mathcal{L}_{\text{DAE}}$ and regularization losses, the Intrinsic-DAE is able to generate convincing shading and albedo estimates without direct supervision (Fig. 2.10-(b) to (g)). Without the "learning-to-align" property, a baseline autoencoder structure with an intrinsic decomposition design (Fig. 2.4-(b)) cannot decompose the image into plausible shading and albedo components (Fig. 2.10-(h),(i),(j)).

In addition, we show that by manipulating the learnt latent representation of $S$, Intrinsic-DAE allows us to simulate illumination effects for face images, such as interpolating lighting directions (Fig. 2.11).

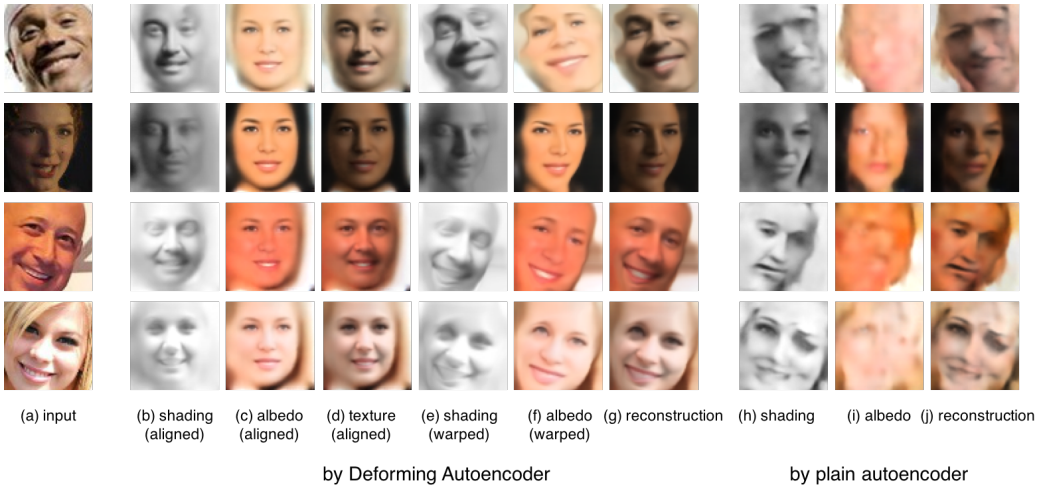



Figure 2.10: Unsupervised intrinsic decompostion with Deforming Autoencoders (Intrinsic-DAE). Thanks to the "automatic dense aligment" property of DAE, shading and albedo are faithfully separated (e,f) by the intrinsic decomposition loss. Shading (b) and albedo (c) are learnt in an unsupervised manner in the densely aligned canonical space. With the deformation field also learnt 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).


Figure 2.11: Lighting interpolation with Intrinsic-DAE. With latent representations learnt in an unsupervised manner for shading, albedo, and defomation, 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 learnt deformation of the source image, DAE enables the relighting of the face in its original pose (3).

Training with $L2$ reconstruction losses, autoencoder-like architectures are prone to generating smooth images which lack visual realism (Fig. 2.10). Inspired by the success of generative adversarial networks (GANs) [Goodfellow 2014], we follow previous work [Shu 2017] where an adversarial loss is adopted to generate visually

realistic images: we train the Intrinsic-DAE with an extra adversarial loss term $\mathcal{L}_{\text{ADVERSARIAL}}$ applied on the final output. The loss function becomes:

$$\mathcal{L}_{\text{INTRINSICDAE}} = \mathcal{L}_{\text{RECONSTRUCTION}} + \mathcal{L}_{\text{WARP}} + \lambda_4 \mathcal{L}_{\text{ADVERSARIAL}} \,. \tag{2.9}$$

In practice, we apply a PatchGAN [Li 2016, Isola 2016] as the discriminator and set $\lambda_4 = 0.1$. We found that the adversarial loss improves the visual sharpness of the reconstruction while the deformation, shading are still successfully disentangled (Fig. 2.12).

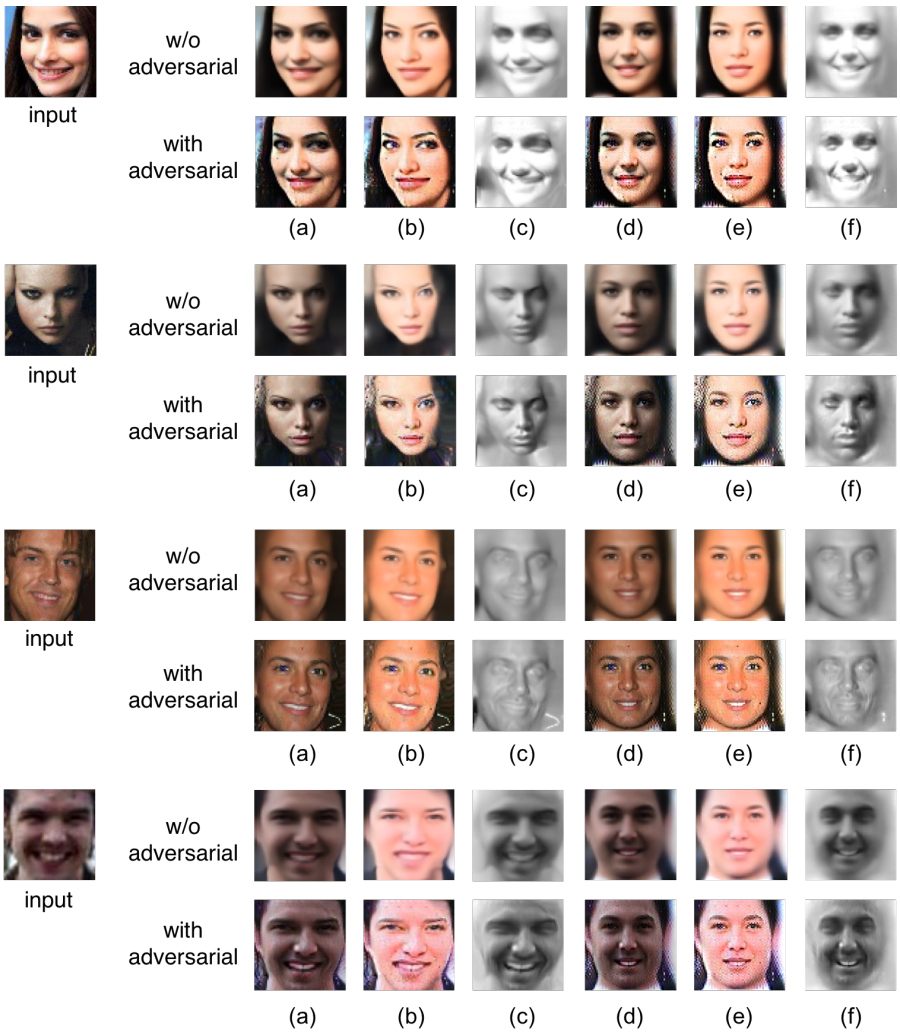



Figure 2.12: Intrinsic-DAE with an adversarial loss: (a/d) reconstruction (b/e) albedo, (c/f) shading, in image and template coordinates, respectively. Applying an adversarial loss to the final output results improves the visual quality of the image reconstruction (a) of Intrinsic-DAE, while the deformation, albedo, and shading can still be successfully disentangled.

### 2.4.4 Unsupervised alignment evaluation

Figure 2.13: *1st row*: Sample images from the MAFL test set; *2nd row*: Estimated deformation grid; and *3rd row*: Image reverse-transformed to texture space *4th row*: semantic landmark locations (green: ground truth landmark locations, blue: estimated landmark locations, red: error lines).

Having qualitatively analyzed the disentanglement capabilities of our networks, we now turn to quantifying their performance on the task of unsupervised image alignment. We report the performance of our face DAE's alignment on landmark detection on face images, specifically, the eyes, the nose, and corners of the mouth. We report performance on the MAFL dataset, which contains manually annotated landmark locations 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 to estimate deformation fields on the MAFL training set. Following the evaluation protocol of the work that we directly compare to [Thewlis 2017c], we train a landmark regressor post-hoc on these deformation fields using the provided annotations. We use landmark locations from the MAFL training set as training data for this regressor, but do not pass gradients to the Deforming Autoencoder, which thereby remains fixed to the model learnt without supervision. The regressor is a 2-layer fully-connected neural network. 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 10-D output layer to predict the spatial coordinates ($(x, y)$) for five landmarks corresponding to the eyes, nose, and mouth corner landmarks. We use L1 loss as the objective function for this regression task.

In testing, we predict landmark locations using the trained regressor and the deformation fields on the MAFL test set. In Table 1 we report the mean error in landmark localization as a percentage of the inter-ocular distance. 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.

| $A$, MAFL | $I$, MAFL | $A + I$, MAFL | $A + I$, CelebA | $A + I$, CelebA, with Regressor |
|---|---|---|---|---|
| 14.13 | 9.89 | 8.50 | 7.54 | 5.96 |

Table 2.1: Improvement in landmark localization errors on the MAFL test set as we add new types of deformation and new data. In the table, $A$ indicates a model which uses the affine transformation, $I$ indicates one with the integral transformation, whereas MAFL and CelebA denote which dataset the deforming autoencoder was trained on. For columns 1 to 4, we manually annotate landmarks on the average texture image, while for column 5, we train a regressor on the deformation fields to predict them. In all experiments, each latent vector in the DAE is of size 32.

In Table 2 we compare with the results of the best current method for semi-supervised image registration [Thewlis 2017c]. We observe that by better modelling of the deformation space we quickly bridge the gap in performance, even though we never explicitly trained to learn correspondences.

| **Method** | | **Normalised Mean Error (NME)** |
|---|---|---|
| [Zhang 2016] | | 7.95 |
| [Thewlis 2017c] | | 5.83 |
| DAE | 32-NR | 10.24 |
| | 32-Res | 9.93 |
| | 16 | 5.71 |
| | 32 | 5.96 |
| | 64 | 5.70 |
| | 96 | 6.46 |
| Dense-DAE | 16 | 6.85 |
| | 64 | 5.50 |
| | 96 | 5.45 |

Table 2.2: Mean error on unsupervised landmark detection on the MAFL test set, expressed as a percentage of the inter-ocular distance: modelling non-rigid deformations clearly reduces error more than just modelling affine ones. DAE and Dense-DAE denote two flavours of the deforming autoencoder - with and without dense convolutional connections, respectively. Under DAE and Dense-DAE we specify the size of each latent vector in the deforming autoencoder. *NR* signifies training without regularization on the estimated deformations, while *Res* signifies training by estimating the residual deformation grid instead of the integral. Our results clearly outperform the self-supervised method of [Thewlis 2017c] trained specifically for establishing correspondences.

## 2.5 Applications to Other Domains

We also test our method of modelling the deformation grid by applying it to other domains. More specifically, we look at registration problems in medical imaging

and remote sensing.

### 2.5.1 Deformable Lung Registration

In medical image analysis, an important problem is registering MRI scans of lungs to a common template, or to other scans. From a medical point of view, registration can assess spatio-temporal behaviour of organs, and can help in diagnosis and analysis of disease progression [Sotiras 2013].

We thus test our integral deformation modelling for applications to medical image registration problems. We use our module in a set-up to register 3-dimensional MRI scans of lungs. We employ a CNN to regress a dense deformation grid $W$ from a source image $S$ and a moving image $R$. The CNN is trained so that the grid $W$ is a dense warp from $S$ to $R$. The moved source image $D$ is defined as $D = \mathcal{W}(S, W)$, where $\mathcal{W}$ represents a trilinear interpolation sampling operation under the deformation grid $W$. Figure 2.14 shows the architecture of this CNN. The encoder adopts dilated convolutional kernels along with multi-resolution feature merging, while the decoder employs non-dilated convolutional layers and up-sampling operations. Specifically, a kernel size of $3 \times 3 \times 3$ was set for the convolutional layers while LeakyReLU activation was employed for all convolutional layers except the last two. Instance normalization was included before most of the activation functions. In total five layers are used in the encoder and their outputs are merged along with the input pair of image to form a feature map of 290 features with a total receptive field of $25 \times 25 \times 25$. In the decoder, two branches were implemented—one for the spatial deformation gradients and the other for the affine matrix. As far as the former is concerned, a squeeze-excitation block [Hu 2018] was added in order to weigh the most important features for the spatial gradients calculation while for the latter a simple global average operation was used to reduce the spatial dimensions to one.

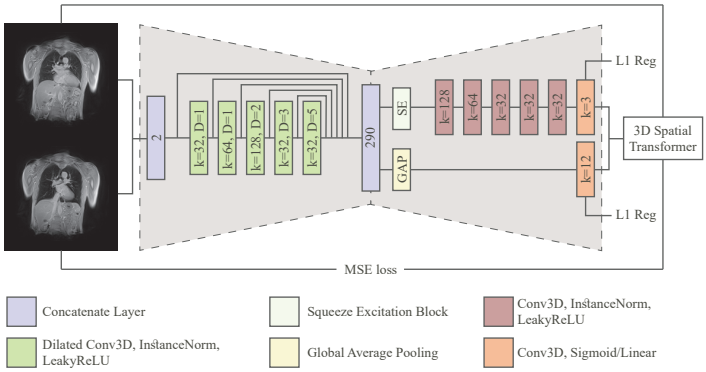



Figure 2.14: CNN architecture used to compute a dense deformation grid between a source image and a moving image.

The network was trained by minimizing the mean squared error (MSE) between the $R$ and $D$ image intensities as well as the regularization terms of the affine

transformation parameters and the spatial deformation gradients using the Adam optimizer [Kingma 2014a].

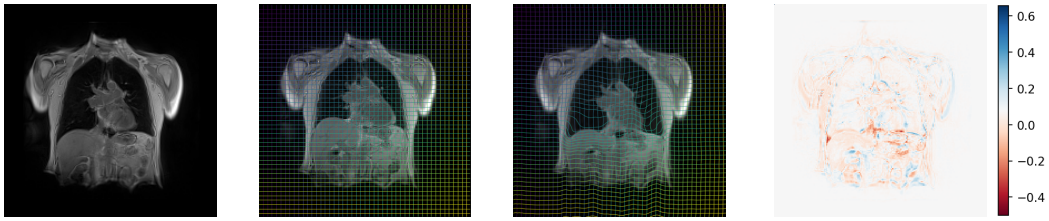
#### 2.5.1.1 Dataset

MRI exams were acquired as a part of a prospective study aiming to evaluate the feasibility of pulmonary fibrosis detection in systemic sclerosis patients by using magnetic resonance imaging (MRI) and an elastic registration-driven biomarker. This study received institutional review board approval and all patients gave their written consent. The study population consisted of 41 patients (29 patients with systemic sclerosis and 12 healthy volunteers). Experienced radiologists annotated the lung field for the total of the 82 images and provided information about the pathology of each patient (healthy or not). Additionally, eleven characteristic landmarks inside the lung area had been provided by two experienced radiologists.

As a pre-processing step, the image intensity values were cropped within the window $[0, 1300]$ and mapped to $[0, 1]$. Moreover, all the images were scaled down along all dimensions by a factor of $2/3$ with cubic interpolation resulting to an image size of $64 \times 192 \times 192$ to compensate for GPU memory constraints. A random split was performed and 28 patients (56 pairs of images) were selected for the training set, resulting in 3136 training pairs, while the rest 13 were used for validation.

#### 2.5.1.2 Results and Discussion

In Table 2.3 the mean Dice coefficient values along with their standard deviations are presented for different methods. We performed two different types of tests. In the first set of experiments (Table 2.3: Inhale-Exhale), we tested the performance of the different methods for the registration of the MRI images, between the inhale and exhale images, for the 13 validation patients. The SyN implementation reports the lowest Dice scores while at the same time, it is computationally quite expensive due to its CPU implementation. Moreover, we tested three different similarity metrics along with their combinations using the method proposed in [Ferrante 2017] as described earlier. In this specific setup, the MI metric seam to report the best Dice scores. However, the scores reported by the proposed architecture are superior by at least $\sim 2.5\%$ to the ones reported by the other methods. For the proposed method,




(a) Reference image (b) Moving image (c) Deformed image (d) Difference

Figure 2.15: A visualised registration of a pair of images, generated by the proposed architecture. The initial and deformed grids are superimposed on the images.

| Method | Inhale-Exhale | All Combinations | Time (s) |
|---|---|---|---|
| Unregistered | 75.62±10.89 | 57.22±12.90 | – |
| Deformable with NCC [Ferrante 2017] | 84.25±6.89 | 76.10±7.92 | ∼1 (GPU) |
| Deformable with DWM [Ferrante 2017] | 88.63±4.67 | 75.92±8.81 | ∼2 (GPU) |
| Deformable with MI [Ferrante 2017] | 88.86±5.13 | 76.33±8.74 | ∼2 (GPU) |
| Deformable with all above [Ferrante 2017] | 88.81±5.85 | 78.71±8.56 | ∼2 (GPU) |
| SyN [Avants 2008] | 83.86±6.04 | – | ∼2500 (CPU) |
| Proposed w/o Affine | 91.28±2.47 | 81.75±7.88 | ∼0.5 (GPU) |
| Proposed | **91.48±2.33** | **82.34±7.68** | ∼0.5 (GPU) |

Table 2.3: Dice coefficient scores (%) calculated over the deformed lung masks and the ground truth. The running time indicated is per patient.

the addition of a linear component to the transformation layer does not change the performance of the network significantly in this experiment.

### 2.5.2 Remote Sensing

Image registration in multimodal, multitemporal satellite imagery is one of the most important problems in remote sensing and essential for a number of other tasks such as change detection and image fusion [Karantzalos 2014b, Dawn 2010, Vakalopoulou 2016]. In this context, we employ a CNN to register a source satellite image to a moving one. The architecture we employ is the same as Figure 2.14, except that the input images are 2-dimensional instead of 3-dimensional. The network is again optimised to minimise the reconstruction error, and we use both affine and integral components of the deformation grid.

#### 2.5.2.1 Dataset

For our experiments, we used a pair of multispectral very high resolution images from the Quickbird satellite. The pair has been acquired in 2006 and 2007, covering a 14 $km^2$ region in the East Prefecture of Attica in Greece. This particular dataset was challenging due to the very large size of the high resolution satellite images, their complexity due to different acquisition angles, shadows, important height differences, numerous terrain objects, and the sparse multitemporal acquisitions. For evaluating the proposed architecture, patches of size $256 \times 256$ were created. In particular, 450 patches were selected randomly for training, 50 for validation and 50 for testing the proposed framework.

#### 2.5.2.2 Results and Discussion

To evaluate the performance of our method we perform different experiments using only the affine or the integral components, and also using their ensemble. We also compare the performance of our method with a state-of-the-art algorithm based on graphs as presented in [Karantzalos 2014a] that has been proven to work very

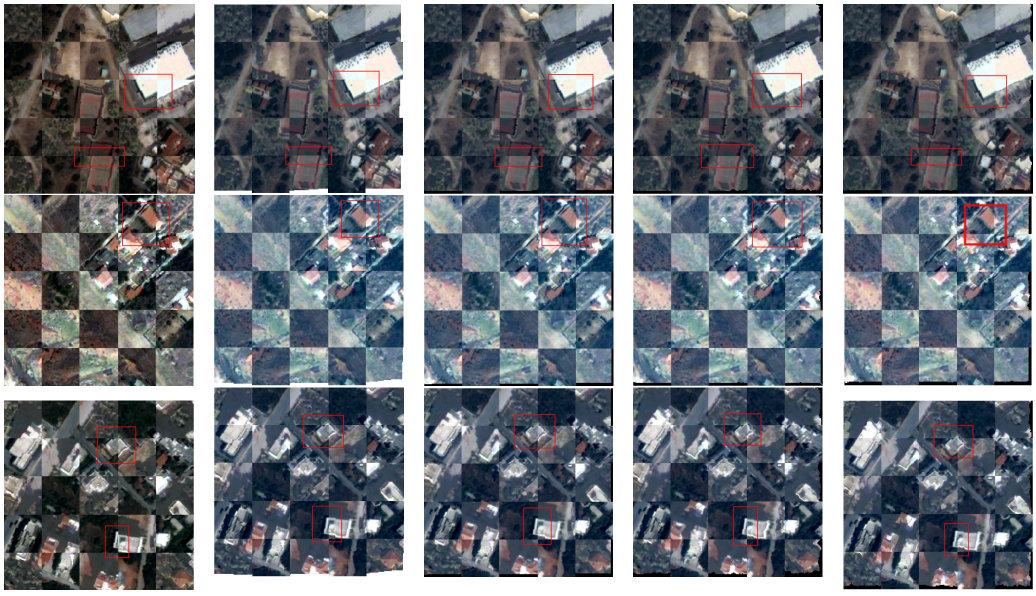



Figure 2.16: Qualitative evaluation for three different pairs of images. With red rectangles we indicate regions of interest.

| Method | $dx$ (pixel) | $dy$ (pixel) | $ds$ (pixel) | Time (s) |
|---|---|---|---|---|
| Unregistered | 7.3 | 6.3 | 9.6 | – |
| [Karantzalos 2014a] | 1.3 | 2.3 | 2.6 | ∼2 |
| Proposed, only affine | 2.5 | 2.8 | 3.7 | ∼0.02 |
| Proposed, only integral | 1.2 | 2.0 | 2.3 | ∼0.02 |
| Proposed | **0.9** | **1.8** | **1.9** | ∼0.02 |

Table 2.4: Errors measured as average euclidean distances between landmark locations in the moved and target images. $dx$ and $dy$ denote distances along $x$-, $y$-, respectively, while $ds$ denotes the average error along all axes.

well on large remote sensing imagery. [Karantzalos 2014a] used normalized cross correlation as the similarity metric.

In Figure 2.16 we present three different pairs of images using checkerboard visualizations between the target $R$ and warped image $D$ before and after the registration using the different tested approaches. Even if the initial displacements were quite important all the methods recover the geometry and register the pair of images. However, the proposed method trained only with the affine deformation fails to register accurately high buildings which have the largest deformations, due to the global nature of the transformation. Finally, the proposed method with only the deformable part, was slightly more difficult to be trained, proving that the additional linear component is a valuable part of the proposed framework.

For quantitative evaluation, a number of landmarks, mainly on the buildings corners were selected and their errors in each of the axes computed (Table 2.4). It

should be noted that for all the methods the same landmarks have been selected and around 10 image pairs were used to extract the landmarks. These landmarks contained mainly roofs of buildings as they were the ones presenting the higher registration errors. One can observe that the proposed method using only the affine component does not perform as well as the rest of the approaches as it fails to recover the geometry in places with local deformations. On the other hand the rest of the approaches report very low errors with the proposed method using both affine and integral parts performing slightly better. Finally, it should be noted that the proposed method is very fast, with inference time for an image pair of size $256 \times 256$ less than half a second, giving a big advantage for very large datasets such as the remote sensing ones, and allowing even real-time applications.

## 2.6 Summary

In this chapter we have developed deep autoencoders that can disentangle shape and appearance in latent representation space. We have shown that this method can be used for unsupervised groupwise image alignment. To achieve this, we have proposed a new module for dense deformation field regression that can be plugged into a neural network easily. Experiments show that this module helps generate more meaningful deformation grids that other methods, for example, predicting the residual grid. 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 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. Furthermore, applications of the deformation field modelling are shown to unsupervised registration problems in medical imaging and remote sensing, where it is shown to beat other contemporary methods. We will now extend the notions of 2D dense alignment to obtain 3D shape in the following chapter.

## 2.7 Contributions

This chapter presents a joint work to which several authors have contributed. My contributions to the work are as follows.

1. I contributed to building models and experiments.
2. I contributed to the evaluation of deformation field regresion and unsupervised landmark localisation.
3. I contributed to the application to medical image registration.
4. I contributed to the application to remote sensing.

Chapter 3

# Lifting Autoencoders: From 2D Dense Alignment to a 3D Morphable Model

In this chapter, we introduce lifting autoencoders (LAE), 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.

## 3.1 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.

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

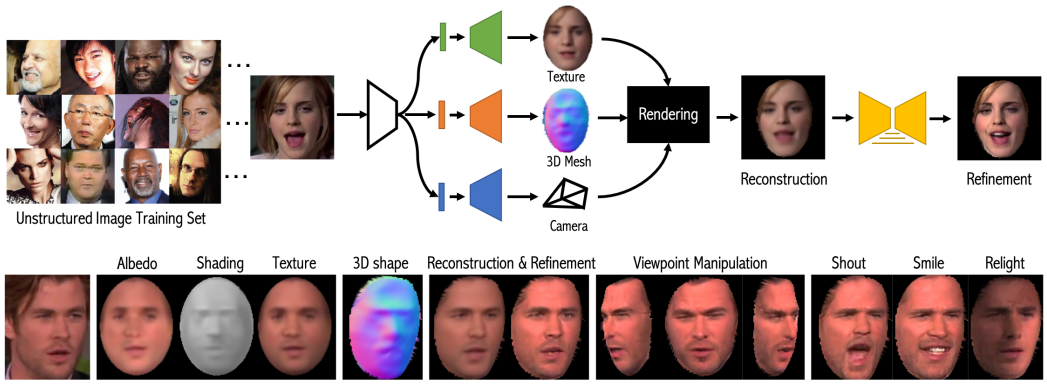



Figure 3.1: We introduce lifting autoencoders, a deep generative model of 3D shape variability that is learned from an unstructured photo collection without supervision. Having access to 3D allows us to disentangle the effects of viewpoint, non-rigid shape (due to identity/expression), illumination and albedo and perform entirely controllable image synthesis.

and expression, leading to even more controllable 3D generative models.

The resulting model allows us to generate photorealistic images of persons in a fully-controllable manner: we can manipulate 3D camera pose, expression, texture and illumination in terms of disentangled and interpretable low-dimensional variables.

Our starting point is the deforming autoencoder (DAE) model introduced in Chapter 2 to learn an unsupervised deformable template model for an object category. DAEs incorporate deformations in the generative process of a deep autoencoder by associating pixels with the UV coordinates of a learned deformable template. As such, they disentangle appearance and shape variability and learn dense template-image correspondences in an unsupervised manner.

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 DAE in terms of a 3D representation. For this we train a network task so as minimize a Non-Rigid SfM minimization 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 [Kato 2018b] 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 our spherical-harmonics-based [Zhang 2005, Wang 2007, Wang 2009] 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

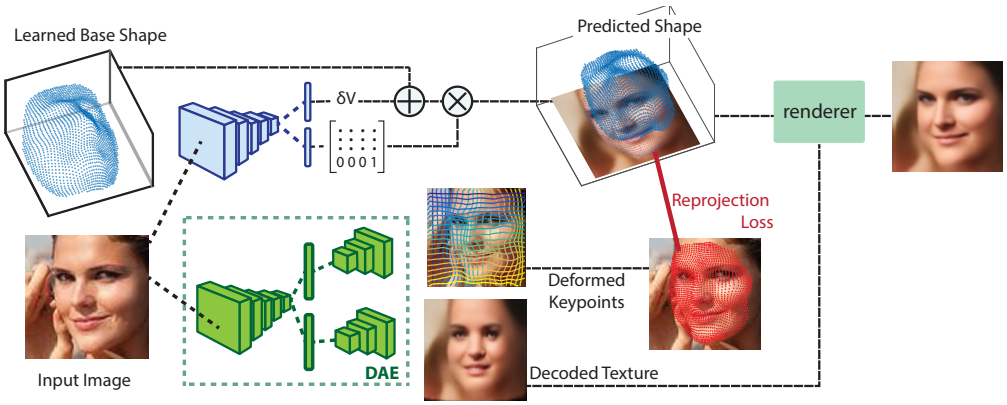



Figure 3.2: Lifting autoencoders bring Non-Rigid Structure from Motion (NRS*f*M) into the problem of learning disentangled generative models for object categories. We start from a deforming-autoencoder (DAE) that interprets images in terms of non-rigid, 2D warps between a template and an image. We train a lifting autoencoder network by minimizing a NRS*f*M-based reprojection error between the learned, 3D Morphable Model-based vertices and their respective DAE-based positions. Combined with a differentiable renderer providing 3D-to-2D information. and an adversarially trained refinement network this provides us with an end-to-end trainable architecture for photorealistic image synthesis.

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.

## 3.2 Related 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 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 [Chen 2016b, Worrall 2017, Memisevic 2010, Worrall 2016, Sundermeyer 2018], or even non-rigid, dense deformations from appearance [Zhou 2016b, Gaur 2017, Thewlis 2017c, Shu 2018, Wiles 2018b]. Several of these techniques have made it into some of the most compelling photorealistic, controllable generative models of object categories [Pumarola 2018, Karras 2019].

Moving closer to graphics, recent works have aimed at exploiting our knowledge about image formation in generative modeling by replicating the inner workings of graphics engines in deep networks. On the synthesis side, geometry-driven generative models using intrinsic images [Shu 2017, Alhaija 2018, Sengupta 2018b] or the 2.5D image sketch [Zhu 2018] as inputs to image synthesis networks have been shown to deliver sharper, more controllable image and video [Kim 2018] synthesis results. On the analysis side, several works have aimed at intrinsic image decomposition [Barrow 1978] using energy minimization, e.g [Gehler 2011, Kong 2014]. The disentanglement of image formation into all of its constituent sources (surface normals, illumination and albedo) was first pursued in [Barron 2013], where priors over the constituent variables were learned from generic scenes and then served as regularisers to complement the image reconstruction loss. More recently, deep learning-based works have aimed at learning the intrinsic image decomposition from synthetic supervision [Narihira 2015], self supervision [Janner 2017] or multi-view supervision [Yu 2018].

These works can be understood in D. Marr's terms as getting 2.5D proxies to 3D geometry, which could eventually lead to 3D reconstruction [Wu 2017]: texture is determined by shading, shading is obtained from normals and illumination, and normals are obtained from the 3D geometry. This leads to the task of 3D geometry estimation as being the key to a thorough disentanglement of image formation.

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 [Tulsiani 2017], videos [Novotny 2017] or some pre-existing 3D mesh representation that is the starting point for further disentanglement [Genova 2018, Sengupta 2018a, Yao 2018, Tewari 2018] or self-supervision [Zhou 2017]. 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. In [Kanazawa 2018b] the authors used segmentation masks and keypoints to learn a CNN-driven 3D morphable model of birds, trained in tandem with a differentiable renderer module [Kato 2018b]. Apart from the combination with an end-to-end learnable framework, this requires however the same level of manual annotation (keypoints and masks) that earlier works had used to lift object categories to 3D [Carreira 2016].

A similar approach has been proposed in [Tran 2018] to learn morphable models from keypoint annotations.

The LiftNet architecture proposed more recently by [Wiles 2018a] uses a 3D geometry-based reprojection loss to train a depth regression FCN by using correspondences of object instances during training. This however is missing the surface-based representation of a given category, and is using geometry only implicitly, in its loss function - the network itself is a standard FCN.

[Jimenez Rezende 2016] was the first work to propose unsupervised training of volumetric CNNs using toy examples and mostly binary masks. Most recently, a GAN-based volumetric model of object categories was introduced in [Henzler 2018], showing that one can recover 3D geometry from an unstructured photo collection using adversarial training. Still, this is far from a rendering pipeline, in the sense that the effects of illumination and texture are coupled together, and the volumetric representation implies limitations in resolution.

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 unorganised set of images, leading to controllable photorealistic image synthesis.

We do so by relying on on Non-Rigid Structure from Motion (NRS$f$M). Rigid SFM is a mature technology, with efficient algorithms existing for multiple decades years [Tomasi 1992, Hartley 2003], systems for large-scale, city-level 3D reconstruction were introduced a decade ago [Agarwal 2009], while high-performing systems are now publicly available [Schönberger 2016]. Rigid SFM has very recently been revisited from the deep learning viewpoint, leading to exciting new results [Ummenhofer 2017, Zhou 2017].

In contrast, NRS$f$M is still a largely unsolved problem. Developed originally to establish a 3D model of a deformable object by observing its motion [Bregler 2000] it was developed to solve increasingly accurately the underlying mathematical optimization problems [Torresani 2008, Paladini 2009, Akhter 2009, Dai 2014], extending to dense reconstruction [Garg 2013], lifting object categories from keypoints and masks [Carreira 2016, Kanazawa 2018b], incorporating spatio-temporal priors [Simon 2014] and illumination models [Liu-Yin 2017], while leading to impressively high-resolution 3D Reconstruction results [Gotardo 2015, Liu-Yin 2017, Hernandez 2017]. In [Kong 2019] 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. By contrast, our proposed method has a simple, linear model for the shape variability, as classical morphable models, but establishes the correspondences automatically.

Earlier NRS$f$M-based work has shown that 3D morphable model learning is possible in particular for human faces [Kemelmacher-Shlizerman 2013, Kemelmacher-Shlizerman 2012, Kemelmacher-Shlizerman 2011] by using a carefully designed, flow-based algorithm to uncover the organization of the image collection - effec-

tively weaving a network of connections between pixels of images, and feeding this into NRS$f$M. As we now show this is no longer necessary - we delegate the task of establishing correspondences across image pixels of multiple images to a Deforming Auto-Encoder [Shu 2018] and proceed to lifting images through an end-to-end trainable deep network as we now describe. . Several other works have shown that combining a prior template about the object category shape with video allows for an improved 3D reconstruction of the underlying geometry, both for faces [Thies 2016, Tewari 2018, Koujan 2018] and quadrupeds [Biggs 2018]. However, these methods still require multiple videos and a template, while our method does not. We intend to explore the use of video-based supervision in future work.

## 3.3 Lifting Autoencoders

We start by briefly describing how DAEs can help us learn structure from deformations, as this is the starting point of our work. We then turn to our contributions of 3D lifting in Section 3.3.1 and shape disentanglement in Section 3.4.2.

### 3.3.1 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 (NRS$f$M). Our central observation is that DAEs provide us with an image representation on which NRS$f$M 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 to then 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 optimise 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 NRS$f$M 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.

### 3.3.2 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, $S : \mathbf{c}_i \rightarrow \mathbb{R}^3$, where $\mathbf{c}_i = (u_i, v_i)$. Even though the underlying model is continuous, our implementation is discrete—we consider a set of 3D points sampled uniformly on a cartesian grid in intrinsic coordinates,

$$\mathcal{S}_i = S(\mathbf{c}_i),\ \mathbf{c}_i \in \mathcal{D} \times \mathcal{D}, \tag{3.1}$$

$$\text{with } \mathcal{D} = \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right\},\ i = 1, \ldots, N = (n+1)^2, \tag{3.2}$$

where $n$ determines the spatial resolution at which we discretise the surface. We begin by constructing a mesh of regular $n \times n$ 2D lattice using a standard triangulation technique as shown in Figure 3.3. Each vertex in this mesh is associated with a UV-coordinate.

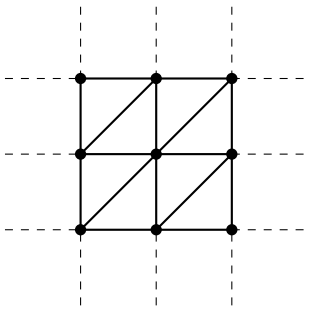


Figure 3.3: The triangulation of the regular 2D lattice of $(u_i, v_i)$ coordinates. This triangulation converts the set of points into a mesh that can be interpreted by a rendering module.

We parameterise 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 $\mathbb{B}^0$,

$$\mathcal{S}_i = \mathbb{B}_i^0 + \sum_{j=1}^{B} \mathbf{s}_j \mathbb{B}_i^j\,, \tag{3.3}$$

where $B$ is the number of elements in the number of components in the linear combination, excluding the mean shape. Here, $\mathbb{B}_i^j$ refers to the $i$-th element of the $j$-th component, which determines the contribution of this element towards the vertex cooresponding to the $i$-th UV-location, $\mathbf{c}_i$. One can think of the mean shape $\mathbb{B}^0$ as denoting *default* positions of the vertices $\mathcal{S}_i$. The residual $\sum_{j=1}^{B} \mathbf{s}_j \mathbb{B}_i^j$ can then be thought of as instance-specific deformation of the mean shape, because the *mixing* coefficients $\mathbf{s}_j$ are inferred from the image. In our formulation, the mean shape $\mathbb{B}^0$ as well as the components $\mathbb{B}^j$ are learnable parameters of the model, while the mixing coeffecients $\mathbf{s}_j$ are regressed from an image using a CNN.

In morphable models [Vetter 1997,Booth 2018] the mean shape and deformation basis elements are learnt by PCA on a set of aligned 3D shapes, but in our case we discover them from 2D by solving an NRS$f$M minimization problem that involves the projection to an unknown camera viewpoint.

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

$$\hat{\mathbf{x}}_i = \mathbf{P}\left[\mathbf{R}\mathcal{S}_i\right] + \mathbf{t}\,, \tag{3.4}$$

$$\mathbf{P} = \begin{bmatrix} \sigma & 0 & 0 \\ 0 & \sigma & 0 \end{bmatrix}\,, \tag{3.5}$$

where $\sigma$ 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 $(u_i, v_i)$, whereas its actual position is the one we obtain from the DAE. To find this 'ground truth' position, we find the image position $\mathbf{x}$ that the DAE's deformation decoder labels as $(u_i, v_i)$,

$$\mathbf{x}_i = \underset{\mathbf{x}\in\mathbb{R}\times\mathbb{R}}{\arg\min} \|W^{-1}(\mathbf{x}) - (u_i, v_i)\|_2\,. \tag{3.6}$$

Here, $W$ is the deformation predicted by the DAE, and $W^{-1}(\mathbf{x})$ denotes the point in the canonical space to which the inverse deformation of $W$ maps the point $\mathbf{x}$. In practice, we find $\mathbf{x}_i$ by warping $(u_i, v_i)$ under $W$ and locating the point in image coordinates that it warps to. This is effectively equivalent to placing a value of 1 at $(u_i, v_i)$ in the canonical space with 0s everywhere else, warping it under the deformation grid, and computing a weighted average of the responses this produces at all locations the image space, weighted by the intensity of response. For some $i$, $(u_i, v_i)$ might not produce a response anywhere, corresponding to cases where $(u_i, v_i)$ vanishes under the deformation. For such cases, we set a visibility variable $\nu_i$ to 0. All other UV-coordinates receive a visibility value of 1. Note that this is translates directly to occlusion of the associated mesh vertex $i$ after the corresponding rotation and projection are applied to $\mathcal{S}_i$. This process is equivalent to searching the image coordinates for a point $\mathbf{x}$ that would project to $(u_i, v_i)$ when warped using the inverse of the deformation inferred by the DAE. Since in practice, we work in a discrete setting, we use a warper that uses backward bilinear interpolation sampling.

We express this *reprojection objective* in terms of the remaining variables—

$$L(\mathbf{R}, \mathbf{t}, \sigma, \mathbf{s}, \mathbb{B}) = \sum_{i=1}^{N} \nu_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i(\mathbf{R}, \mathbf{t}, \sigma, \mathcal{S}, \mathbf{s})\|_2^2 \tag{3.7}$$

where we have expressed $\hat{\mathbf{x}}_i$ as a differentiable function of $\mathbf{R}, \mathbf{t}, \mathcal{S}, \mathbf{s}$ through Equation 3.4 and Equation 3.3.

For a set of $K$ images $(k = 1, \ldots, K)$ we denote by $(\mathbf{R}_k, \mathbf{t}_k), \sigma_k, \mathbf{s}_k$ the camera

and shape parameters for the $k$-th image, 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:

$$\mathcal{L}_{3D} = \sum_{k=1}^{K} L(\mathbf{R}_k, \mathbf{t}_k, \sigma_k, \mathbf{s}_k, \mathbb{B}) \tag{3.8}$$

### 3.3.3 LAE learning via Deep NRS$f$M

Minimizing the objective of Equation 3.8 amounts to the common Non-Rigid Structure-from-Motion objective [Bregler 2000, Torresani 2008, Paladini 2009, Akhter 2009, Dai 2014]. 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 $\mathbf{s}_k$, (b) the Euler angles/rotation matrix $\mathbf{R}_k$, and (c) the displacement vector $\mathbf{t}_k$ describing the camera position. In the limit of very large hidden vectors the related angle/displacement/coefficient heads could simply memorize the optimal values per image, as dictated by the optimization of Equation 3.8. With a smaller number of hidden units these heads learn to successfully regress camera and shape vectors and can generalize to unseen images. As such, they are components of a larger deep network that can learn to reconstruct an image in 3D— a task we refer to as Deep NRS$f$M.

If we only train a network to optimise 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 [Kato 2018b], 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 [Ronneberger 2015] 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

3.7 - we see that while keeping intact the image generation process, we achieve a substantially more realistic synthesis.

## 3.4 Geometry-Based Disentanglement

A lifting autoencoder provides us with a disentangled representation of images in terms of 3D rotation, non-rigid deformation, and texture, leading to controllable image synthesis.

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. These contributions rely on certain assumptions and data that are reasonable for human faces, but could also apply to several other categories.

We first describe in Section 3.4.1 how surface-based normal estimation allows us to disentangle appearance into albedo and shading using a physics-based model of illumination. In Section 3.4.2 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.

### 3.4.1 LAE-lux: Disentangling Shading and Albedo

Given the 3D reconstruction of a face we can use certain assumptions about image formation that lead to physically-plausible illumination modeling. For this we extend LAEwith albedo-shading disentangling, giving rise to LAE-lux where we explicitly model illumination.

As in several recent works [Shu 2017, Sengupta 2018b] we consider a Lambertian reflectance model for human faces and adopt the Spherical Harmonic model to model the effects of illumination on appearance [Zhang 2005, Wang 2007, Wang 2009]. We pursue the intrinsic decomposition [Barrow 1978] of the canonical texture $\mathsf{T}$ into albedo, $\mathsf{A}$ and shading, $\mathsf{S}$:

$$T = S \odot A \tag{3.9}$$

where $\odot$ denotes Hadamard product, by constraining the shading image to be connected to the normals delivered by the LAEsurface.

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:

$$S(x) = \langle L, H(N(x)) \rangle. \tag{3.10}$$

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. 3.9.

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 $S^{\text{render}}$ with spherical harmonics parameters $L$ and normal maps $N$and then use a U-Net to refine it, obtaining $S^{\text{adapted}}$.

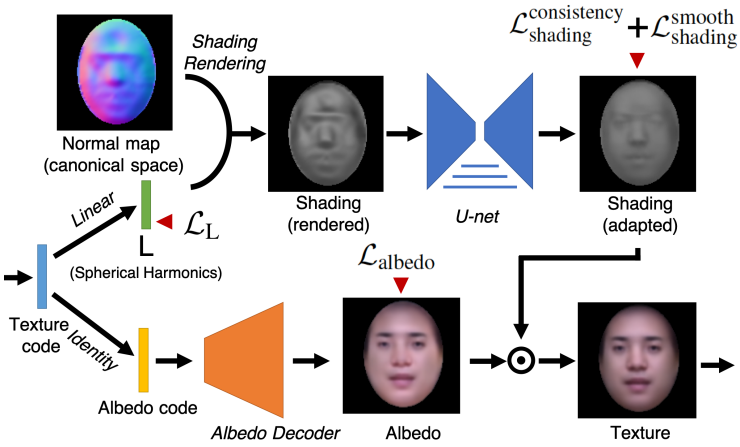



Figure 3.4: Texture decoder for LAE-lux: disentangling albedo and illumination with 3D shape and Spherical Harmonics representation for illumination.

In our experiments we have initialized LAE-lux with a converged LAE, discarded the last layer of the LAE's texture prediction and replaced it with the inrinsic predictor outlined above. The albedo image is obtained through an albedo decoder that has an identical architecture to the texture decoder in DAE. The latent code for albedo $Z_A$ and the spherical harmonics parameters $L$ are obtained as separate linear layers that process the penultimate layer of the texture encoder.

In training, only the texture decoders are updated while other encoding and decoding networks are fixed. When instead training everything jointly from scratch we observed implausible disentanglement results, presumably due to the ill-posed nature of the decomposition problem.

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 [Shu 2017] we employ intrinsic image-based smoothness losses on albedo and shading:

$$\mathcal{L}_{\text{shading}}^{\text{smooth}} = \lambda_{\text{shade}} \left\| \nabla S^{\text{adapted}} \right\|_2, \mathcal{L}_{\text{albedo}} = \lambda_{\text{albedo}} \left\| \nabla A \right\|_1, \tag{3.11}$$

where $\nabla$ represents the spatial gradient, which means that we allow the albedo to have sharp discontinuities, while the shading image should have mostly smooth variations [Samaras 2000]. 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 penalise its distance to the regressed illumination values:

$$\mathcal{L}_{\mathrm{L}} = \left\| L - \hat{L} \right\|_2 . \tag{3.12}$$

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 [Shu 2017], $\hat{L}$ serves as a reasonably rough approximation of the illumination coefficient and is used for weak supervision in Equation 3.12.

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

$$\mathcal{L}^{\text{consistency}}_{\text{shading}} = \text{Huber}(S^{\text{adapted}}, S^{\text{rendered}}), \tag{3.13}$$

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

### 3.4.2 Disentangling Expression, Identity and Pose

Having outlined our geometry-driven model for disentangling appearance variability into shading and albedo, we now turn to the task of disentangling the sources of shape variability.

In particular, we consider that 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.

We start by allowing our representation to separately model identity and expression, and then turn to forcing it to disentangle pose, identity and expression.

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:

$$\mathcal{S}_i(\mathbf{s}^I, \mathbf{s}^E) = \mathbb{B}^0_i + \sum_{s=1}^{I} \mathbf{s}^I_s \mathbb{B}^{I,s}_i + \sum_{s=1}^{E} \mathbf{s}^E_s \mathbb{B}^{E,s}_i \tag{3.14}$$

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 when watching a video of a single person, or a single person from multiple viewpoints one can enforce the identity expansion coefficients $\mathbf{s}^I$ to remain constant through a siamese loss [Koch 2015]. This would force the training

to model all of the person-specific variability through the remaining subspace, by changing the respective coefficients $\mathbf{s}^E$ per image.

Here we use the MultiPIE [Gross 2010] dataset to help disentangle the latent representation of person 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 this dataset to disentangle the latent representation for shape into distinct components.

We denote by $S$ the concatenation of all shape parameters: $S = \left[\mathbf{s}^C, \mathbf{s}^I, \mathbf{s}^E\right]$ and turn to the task of forcing the different components of $S$ to behave as expected. We use facial expression distentangling as an example, and follow a similar procedure for pose and camera disentangling. Given an image $I_{\text{exp}}$ with known expression exp, we sample two more images. The first, $I^+_{\text{exp}}$ has the same facial expression but different identity, pose, and illumination conditions. The second, $I^-_{\text{exp}}$, has a different facial expression but the same identity, pose and illumination condition as $I_{\text{exp}}$. We use siamese training to encourage $I_{\text{exp}}$ and $I^+_{\text{exp}}$ to have similar latent representations for facial expression, and a triplet loss to ensure that $I_{\text{exp}}$ and $I^+_{\text{exp}}$ are closer in expression space than $I_{\text{exp}}$ and $I^-_{\text{exp}}$:

$$\mathcal{L}_{\text{expression}} = \mathcal{L}^{\text{similarity}}_{\text{expression}} + \mathcal{L}^{\text{triplet}}_{\text{expression}}, \text{where} \tag{3.15}$$

$$\mathcal{L}^{\text{similarity}}_{\text{expression}} = \left\| f_{\text{exp}}(I_{\text{exp}}) - f_{\text{exp}}(I^+_{\text{exp}}) \right\|_2, \tag{3.16}$$

$$\mathcal{L}^{\text{triplet}}_{\text{expression}} = \max(0, 1 + \left\| f_{\text{exp}}(I_{\text{exp}}) - f_{\text{exp}}(I^+_{\text{exp}}) \right\|_2 - \left\| f_{\text{exp}}(I_{\text{exp}}) - f_{\text{exp}}(I^-_{\text{exp}}) \right\|_2). \tag{3.17}$$

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

$$\mathcal{L}_{\text{disentangle}} = \mathcal{L}_{\text{expression}} + \mathcal{L}_{\text{identity}} + \mathcal{L}_{\text{pose}}, \tag{3.18}$$

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, $\lambda_{\text{disentangle}} = 1$.

### 3.4.3 Complete Objective

Having introduced the losses that we use for disentangling, we now turn to forming our complete training objective.

We control the model learning with a regularization loss defined as follows:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{scale}} \sum_{k=1}^{K} \|\sigma_k\|_2 + \lambda_{\text{shape}} \sum_{k=1}^{K} \left\| \sum_{s=1}^{S} \mathbf{s}^k_s \mathbb{B}^s \right\|_2, \tag{3.19}$$

where $\sigma$ is the scaling parameter in Equation 3.4 and $\sum_{s=1}^{S} \mathbf{s}_s \mathbb{B}^s_i$ 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, $\mathcal{L}_{\text{3D}}$, defined in Equation 3.8, we can write the complete objective function, which is trained end-to-end:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \; & \lambda_{\text{3D}} \cdot \mathcal{L}_{\text{3D}} + \\ & \lambda_{\text{disentangle}} \cdot \mathcal{L}_{\text{disentangle}} + \\ & \lambda_{\text{scale}} \cdot \mathcal{L}_{\text{scale}} + \\ & \lambda_{\text{shape}} \cdot \mathcal{L}_{\text{shape}}.\end{aligned} \tag{3.20}$$

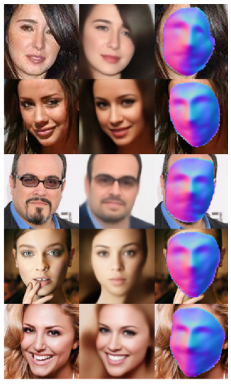
In our experiments, we used the scaling factor for the 3D reprojection loss, $\lambda_{\text{3D}} = 50$. This relatively high scaling factor was chosen so that the reprojection loss is not overpowered by other losses at later training iterations.

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

$$\mathcal{L}_{\text{lux}} = \mathcal{L}^{\text{smooth}}_{\text{shading}} + \mathcal{L}^{\text{consistensy}}_{\text{shading}} + \mathcal{L}_{\text{albedo}} + \mathcal{L}_{\text{L}}. \tag{3.21}$$

## 3.5 Experiments

(a) Visualising reconstructions from various yaw angles



(b) Inferred shape

Figure 3.5: VIsualisations of the learnt 3D shapes from various yaw angles. Our reconstructions respect prominent face features such as the nose, forehead, and cheeks, allowing us to rotate an object reconstruction in 3D.

### 3.5.1 Architectural Choices

Our encoder and decoder architectures are similar to the ones employed in [Shu 2018], but working on images of size $128 \times 128$ pixels instead of $64 \times 64$. We use convolutional neural networks with five stridedConv-batchNorm-leakyReLU layers in image encoders, which regress the expansion coefficients $\mathbf{s}$s. Image decoders consist similarly of five stridedDeconv-batchNorm-ReLU layers.

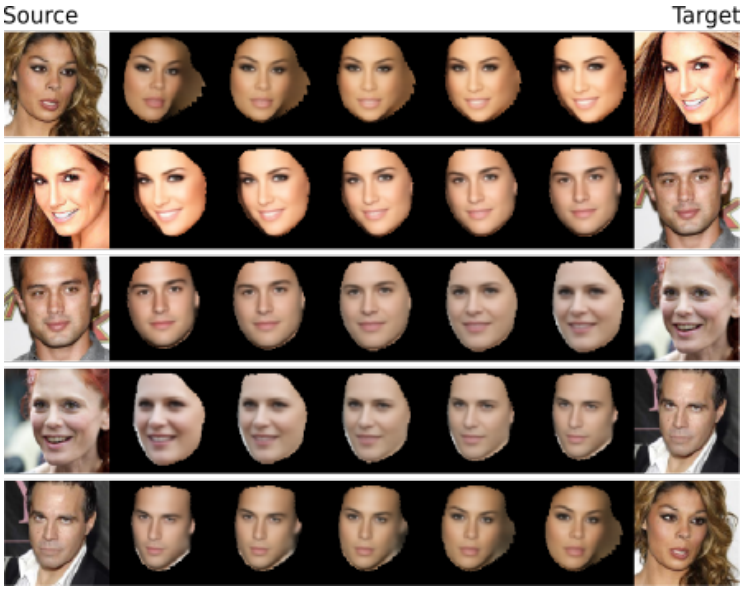



Figure 3.6: Interpolation on the shape, pose, and texture latent vectors. We show renderings of intermediate 3D shapes, with intermediate poses and textures, as we move around on all three latent spaces.

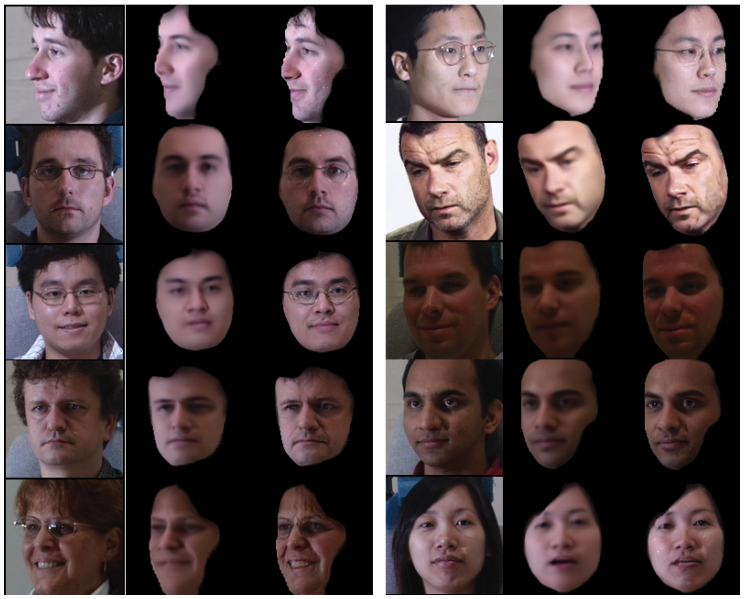


Figure 3.7: Photorealistic refinement: starting from an image reconstruction by an LAE(left), an adversarially-trained refinement network adds details to increase the photorealism of a face (right).

In all of these experiments the training process was started with a base learning rate of 0.0001, which was reduced by a factor of 0.5 every fifty epochs of training. We used the Adam optimiser [Kingma 2014a] and a batch size of 64. All training

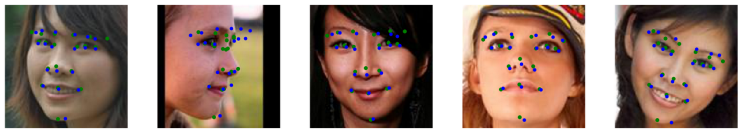


Figure 3.8: Landmark localization on a few AFLW2000 test images. We manually annotated landmarks in the UV space and visualized them after reprojection of the vertices. The LAEis able to localize landmarks for small pose variability. Ground-truth landmarks are shown in green, whereas the predicted ones are shown in blue.

images were cropped and resized to a shape of $128 \times 128$ pixels, while a mesh of size $64 \times 64$ was used in training. This allowed us to sample one keypoint for every four pixels in the UV space, making the mesh fairly high resolution. The mesh was initialized as a Gaussian surface, and was initially positioned so that it faces toward the camera.

### 3.5.2 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** [Liu 2015]: 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 [Zhang 2014a], 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** [Gross 2010]: 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** [Zhu 2017b]: 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.

### 3.5.3 Qualitative Results

In this section, we show examples of the learned 3D shapes. Figure 3.5 shows visualizations of 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. However, when using no pose information from Multi-PIE, and the completely unsupervised nature of our alignment, it is not able to properly decode side poses. This drawback is quickly overcome when we add weak pose supervision from the Multi-PIE dataset, as seen in Figure 3.7.

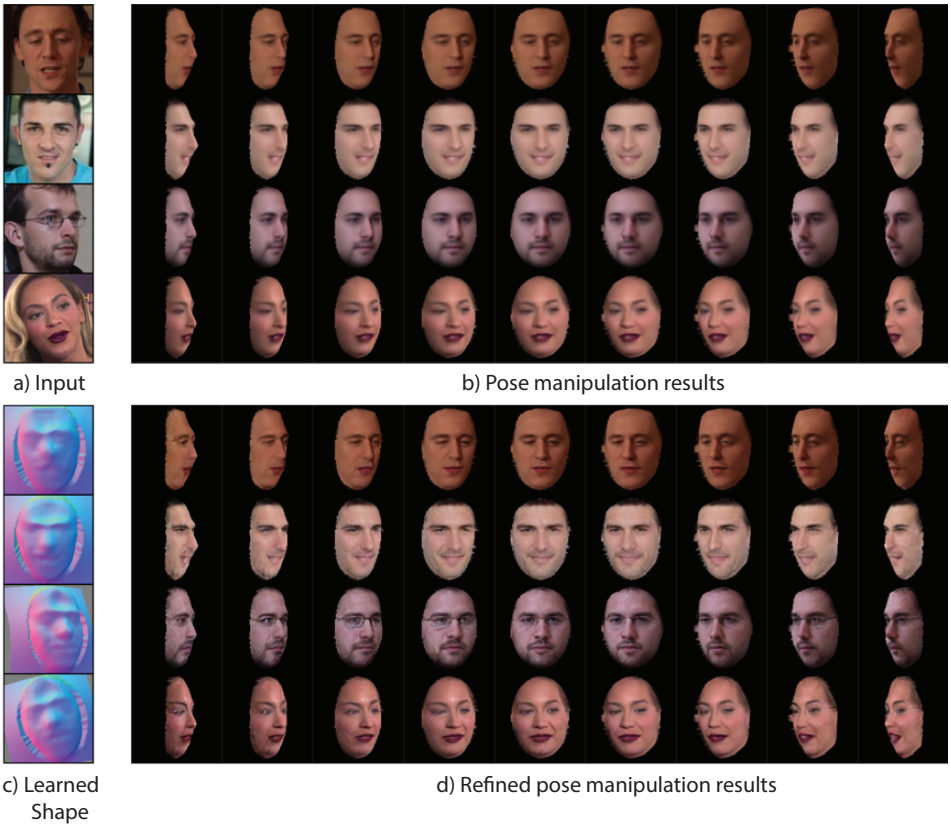



Figure 3.9: Changing Pose with LAE. Given input face image (a), LAElearns to recover the 3D shape (c), with which we can manipulate the pose of the faces (b). With the additional refinement network, we can enhance the manipulated face image by adding facial details (d) that better preserve the characteristic features of the input faces.

### 3.5.4 Face Manipulation Results

In this section, we show some results of manipulating the expression and pose latent spaces. In Figure 3.9 (b), we visualize the decoded 3D shape from input images in 3.9 (a) from various camera angles. Furthermore, in Figure 3.9 (d), we show results after passing the visualizations in Figure 3.9 (b) through the refinement network.

Similarly, in Figures 3.10 and 3.11 (a)-(e), we interpolate over the expression latent space from each of the images in (a) to the image in (b), and visualize the shape at each intermediate step in Figure (c), the output in (d), and the refined output in (e).

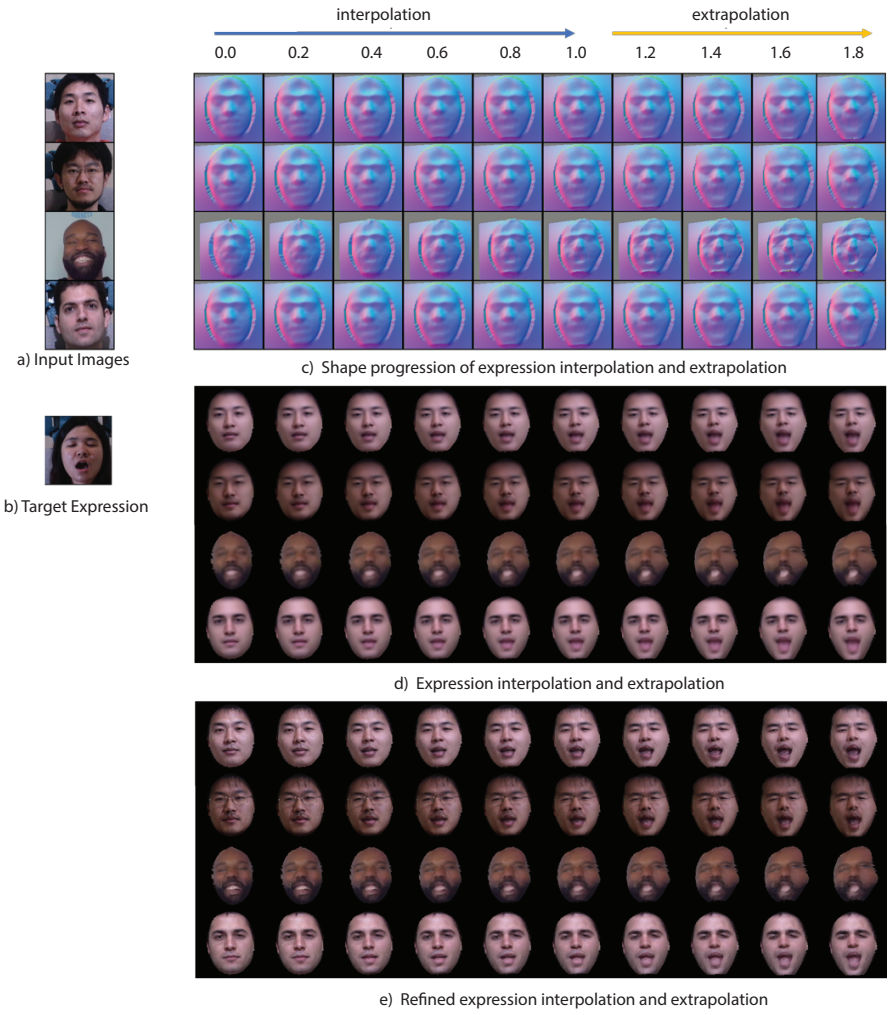



Figure 3.10: Changing Expression with LAE. With LAEwe can perform facial expression interpolation and extrapolation. Given the input faces (a), we can simply transfer the facial expression from another image (b) onto (a) with varying intensities by manipulating the learned expression representations. From (c,d,e) we observe continuous facial expression transformation from the input (a) to the target (b) (column 1 to 6), as well as realistic expression enhancements (column 7 to 10) via latent representation extrapolation (note the mouth and the eyes region).

Finally, in Figure 3.6, we interpolate over all three latent spaces—texture, pose, and shape.

### 3.5.5 Landmark Localization

Our system allows us to roughly estimate landmarks, by annotating them only once in the aligned, canonical space, as also shown by [Shu 2018]. Here we further visualize detected landmarks using the learned 3D shape in Figure 3.8 on some

| Method | NME |
|---|---|
| Thewlis *et al.* (2017) [Thewlis 2017a] | 6.67 |
| Thewlis *et al.* (2018) [Thewlis 2017b] | 5.83 |
| Jakub *et al.* (2018) [Jakab 2018b] | 2.54 |
| Shu *et al.* (2018), DAE, no regressor [Shu 2018] | 7.54 |
| Shu *et al.* (2018), DAE, with regressor [Shu 2018] | 5.45 |
| LAE, CelebA (no regressor) | 7.96 |
| LAE, CelebA (with regressor) | 6.01 |

Table 3.1: 2D landmark localization results for the proposed LAEcompared 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.

images from the AFLW2000-3D dataset.

### 3.5.6 Albedo-Shading Disentanglement

In Fig. 3.12 we show that with the disentangled physical representation for illumination, we can hallucinate illumination manipulation with LAE-lux.

### 3.5.7 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, 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 [Thewlis 2017a, Thewlis 2017b, Shu 2018], 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 3.1 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 [Pavlakos 2017].

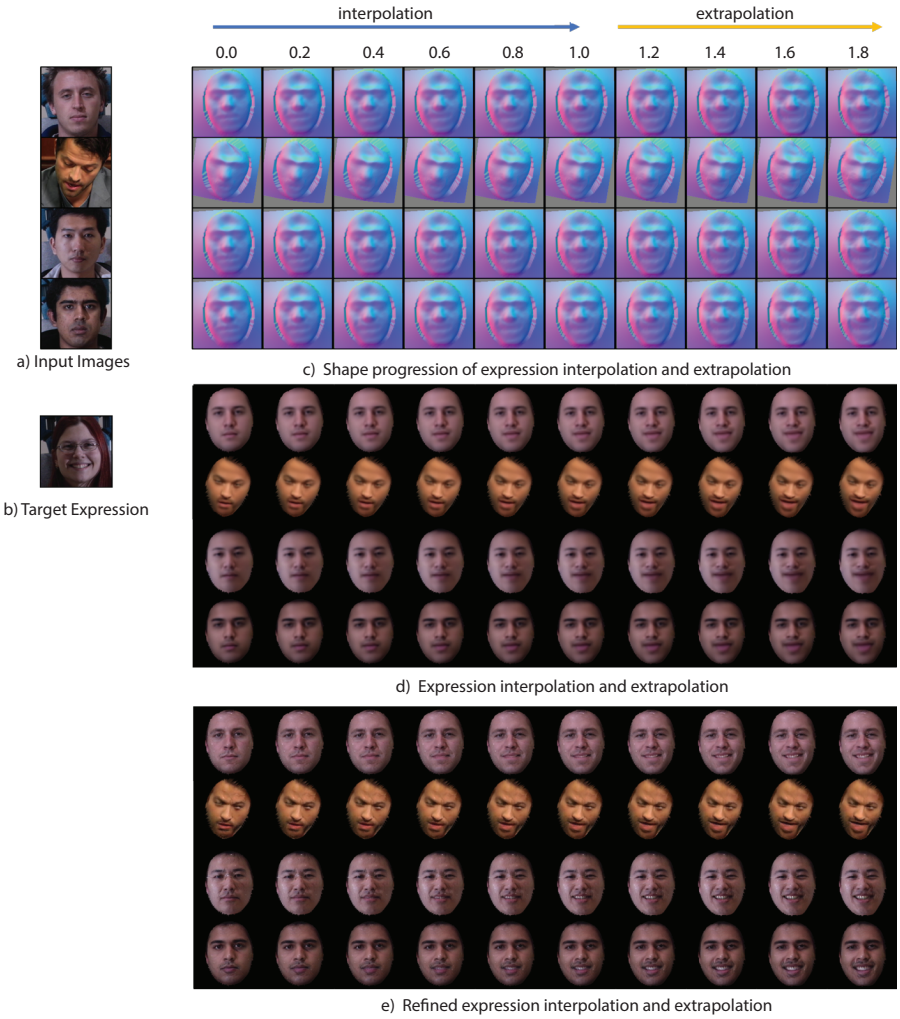



Figure 3.11: Changing Expression with LAE. With LAEwe can perform facial expression interpolation and extrapolation. Given the input faces (a), we can simply transfer the facial expression from another image (b) onto (a) with varying intensities by manipulating the learned expression representations. From (c,d,e) we observe continuous facial expression transformation from the input (a) to the target (b) (column 1 to 6), as well as realistic expression enhancements (column 7 to 10) via latent representation extrapolation (note the mouth and the eyes region).

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 3.2 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. 3.7, where we visualize the discovered shapes for both of these cases.

## 3.6 Summary

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 unorganised 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 synthetsis results with substantially more diverse categories than faces [Brock 2019, Karras 2018] - we anticipate that their combination with 3D representations like LAEs will further unleash their potential for controllable image synthesis.

## 3.7 Contributions

This chapter presents a joint work to which several authors have contributed. My contributions to the work are as follows.

1. I contributed to the models and experiments using deep NRS*f*M.
2. I contributed to the keypoint sampling to link DAE ground truth with LAE targets.
3. I contributed to the evaluation of unsupervised shape learning using landmark localisation and Procrustes analysis.

| Method | Rotation | Yaw angle | | | |
|---|---|---|---|---|---|
| | | $[0, 30]$ | $(30, 60]$ | $(60, 90]$ | All |
| 3DDFA [Zhu 2017b] | Y | 4.25±0.95 | 4.34±1.04 | 4.39±1.35 | 4.28±1.03 |
| (supervised) | N | 12.51±6.40 | 23.20±5.92 | 32.55±3.85 | 17.31±9.30 |
| PRNet [Feng 2018] | Y | 4.88±1.24 | 6.94±2.83 | 10.51±5.31 | 6.01±3.08 |
| (supervised) | N | 7.17±3.45 | 10.96±5.00 | 16.34±8.91 | 9.11±5.66 |
| 3D-FAN [Bulat 2017b] | Y | 2.73±1.38 | 2.48±2.24 | 3.74±2.95 | 2.84±1.92 |
| (supervised) | 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) | Y | 6.83±0.96 | 8.41±1.15 | 9.83±1.65 | 7.59±1.60 |
| CelebA+MultiPIE | N | 9.11±4.54 | 13.60±6.08 | 24.62±8.37 | 12.33±7.84 |

Table 3.2: 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.

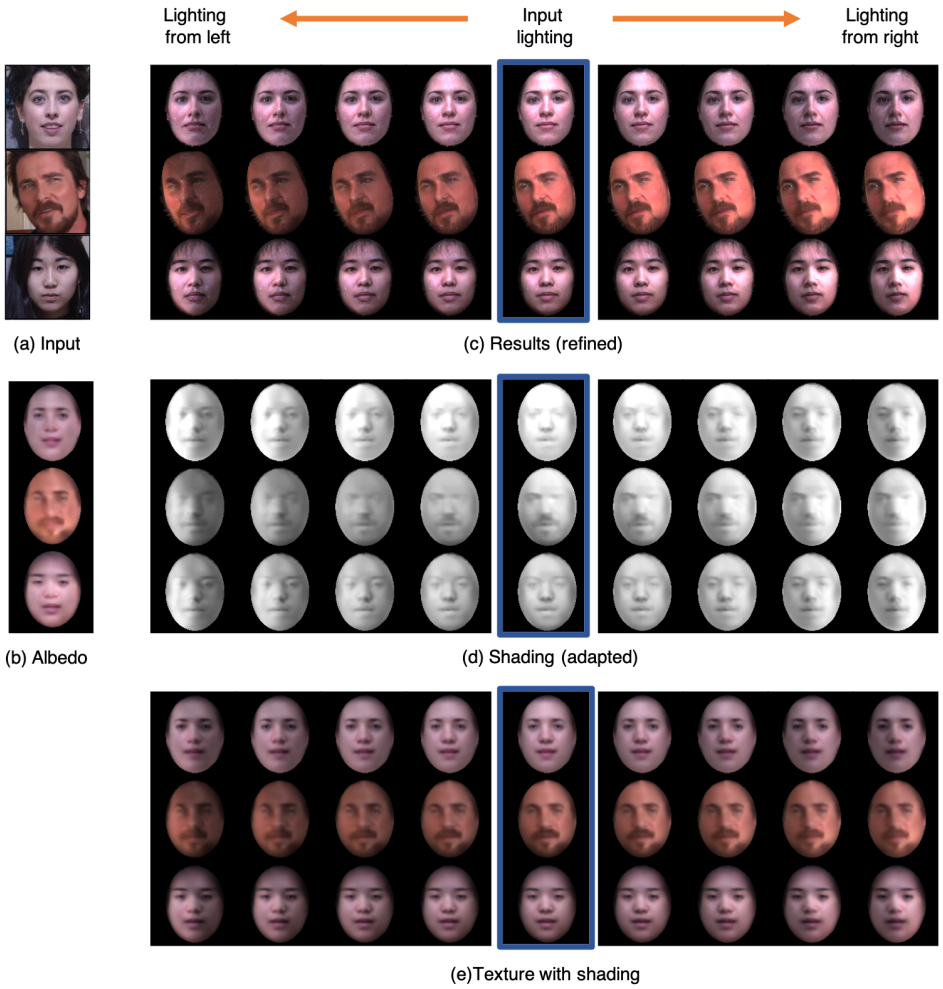



Figure 3.12: Lighting manipulation with LAE-lux. With disentangled albedo and shading and explicit shading representation using Spherical Harmonics, we can manipulate the illumination of faces. We show illumination editing of 3 examples from given input faces (a), to a hallucinated lighting from left ((c) - left side) and a hallucinated lighting from right ((c) - right side). Interpolation of Spherical Harmonics coefficients generates smooth transition of shading effect (d), combining with the learned albedo (b), we obtain the dense aligned texture with different illumination effect (e). Final results (c) are obtained by applying deformation learned in LAEto (e), and a refinement step.

Chapter 4

# Deep Multi-Instance Learning for Diagnosis of Lymphocytosis

The analysis of blood cell count is the most prescribed analysis in medical biology. It includes a quantitative count of different blood cell subsets, and the qualitative analysis of the blood smear, performed through a cytology exam. The latter is characterized by large intra- and inter-operator variability, as the differences between normal and abnormal cells are not evident in most of the cases. This being one of the first steps in the clinical diagnosis of lymphocytosis, it is important to investigate solutions to remove these intra- and inter-operator variabilities, as well as retain high performance. To this end, in this chapter, we explore a weakly-supervised deep learning architecture that combines information from the quantitative cell count, as well as the blood smears which form a part of the qualitative test. A repeatability test is also conducted on unseen examples in order to test the robustness of the proposed approach in a clinical diagnosis setting.

## 4.1 Introduction

Lymphocytosis (i.e., absolute lymphocyte count above $4 \times 10^9$/L) is a common finding, which can be either reactive (to infection, acute stress, and so on), or the manifestation of a lymphoproliferative disorder (a type of cancer of the lymphocytes). In existing clinical practice, diagnosis relies on visual microscopic examination of the blood cells (Figure 4.1) together with the integration of clinical attributes such as age and lymphocyte count. Taking into consideration the visual assessment of the entire set of blood cells and the clinical attributes, a diagnosis of the subtype of lymphoid malignancy is performed. On the positive side such practice is fast

and affordable, but it suffers from poor reproducibility. Additional clinical tests are required, with flow cytometry being the gold standard to definitively affirm the malignant nature of the lymphocytes. However, this analysis is relatively expensive and time consuming, and therefore cannot be performed for every patient in practice. Therefore, the development of automatic and accurate processes could lead to a better way to determine which patient should be referred for flow cytometry analysis in order to confirm or exclude a lymphoproliferative disorder, augmenting and assisting the assessment of the clinicians.

Imaging offers great potential to analyze blood cells in a non-invasive and repeatable manner. In the last decade radiomics has emerged in oncology as a way to extract imaging features for diagnosis or prediction of treatment outcome or to be used as a surrogate of oncogenic processes that are difficult to explore by contextual biopsies [Ravi 2017, Limkin 2017, Sun 2018]. In a standard radiomic approach, tumors or regions of interest (ROI) are detected and outlined, and subsequently features are extracted describing, e.g., shape, texture, or morphology [Zacharaki 2009]. A detailed review of texture analysis methods focusing on microscopy images of cells or tissues can be found in [Cataldo 2017]. In such a setting, (i) the segmented ROIs and pre-defined features are choice-dependent, and (ii) feature extraction is performed independently from statistical modelling, thus diminishing the ability to find evidence-driven correlations, a thriving innovation in precision medicine. We argue that we have no evidence that those primitive features capture all the hidden information and that the independent processes of feature extraction and prediction modelling do not necessarily take full benefit of the richness of the information space. The problem becomes even more difficult when the disease category is given on the subject level and not the individual observation level. This is also the case for the diagnosis of lymphocytosis, where multiple images of blood cells are available for each subject (Figure 4.1) without all of them necessarily belonging to the same category (normal or abnormal). A patient is then associated with these images, with their number varying for each patient. Furthermore, it is difficult for biologists to annotate each individual image as either normal or abnormal, and even if they are able, this annotation suffers from inter-observer variability, so ground truth target variables are only available at the patient-level, the target being the nature of the symptoms, i.e., either reactive or tumoral. Finally, presence of abnormal lymphocytes does not guarantee tumoral nature of the symptoms.

In this chapter, we address these issues and present a novel approach for the challenging task of reliable diagnosis of lymphocytosis where our proposed approach is able to predict the nature of symptoms (reactive/tumoral) from an acquired set of blood smears. In particular, the contributions of this work are fourfold. First, we propose a multi-instance deep CNN for extracting visual representations from multiple microscopy images and associate them directly with patient's diagnosis. Second, we investigate how different aggregation methods for the multiple instance scores affect the model's predictions together with directly trained attention mechanisms. Thirdly, we introduce a mixture-of-experts model [Jacobs 1991] adapted to the problem in order to learn a classifier from both images and the patient's

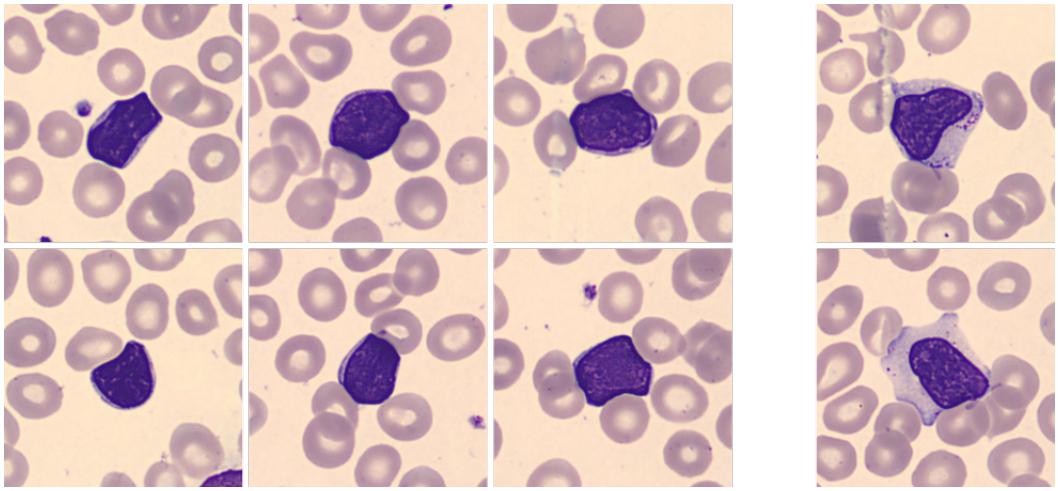


Figure 4.1: An example from lymphocyte images for a patient with a lymphocyte count of $6.967 \times 10^{10}$/L. The six images in the left group depict abnormal lymphocytes while the two in the right one are normal.

clinical attributes to render a diagnosis using information from both of them. Finally, we show comparisons with classical radiomic-based methods coupled with multi-instance classification, as well as recent deep learning-based attention methods reporting better performance for our proposed method.

This chapter is organized as follows. Section 4.2 discusses previous work on multiple-instance learning as well as deep learning applied to medical analysis. We then describe our method, and present its components and implementation details in Section 4.3, followed by descriptions of competing methods in Section 4.4. The dataset used for this study is introduced in Section 4.5, which is followed by a discussion of the evaluation setting and the results of our experiments (Section 4.6). An extensive comparison with other methods and ratings of clinical experts is presented along with a discussion of the results in Section 4.7.

## 4.2 Related Work

Multiple instance learning (MIL) [Keeler 1991, Dietterich 1997] applies to problems where objects (bags) are described by multiple observations (instances) with labels being provided only for the bags. It can be also considered as a form of weakly supervised learning. The challenge that arises for such representations is the lack of precise annotation for each individual instance, and the fact that some of the instances could lack information or encode even misleading information about the object's class (e.g. not all cells are malignant in a histopathology image with malignancy as shown in Figure 4.1). Several methods have been proposed exploiting local or global information and implementing different classifiers or mapping functions [Amores 2013, Foulds 2010b].

Specifically for histopathological image analysis, a variety of machine learning

techniques have been investigated and are exhaustively presented in various reviews [Gurcan 2009,Komura 2018]. Methods based on content-based image retrieval were very commonly used to address this problem [Zhang 2015,Sparks 2016]. Moreover, especially for the task of cell segmentation, a variety of methods have been proposed using bag-of-words [Caicedo 2009], support vector machines [Spanhol 2016], neural networks [Theera-Umpon 2007] or Gaussian mixture models [Dundar 2011]. However, all these methods use predefined hand-crafted radiomics features from the images and fail to take full benefit of the domain specificity. A recent work [Papastergiou 2018] exploits features from tensor decomposition for histopathological diagnosis in order to avoid dedicated feature extraction.

There are a lot of studies that adapt CNN models for MIL by using different pooling layers such as the maximum, mean, generalized mean, log-sum-exponentiation (LSE) [Ramon 2000] or the noisy-and function [Kraus 2016]. In particular, [Kraus 2016] presents a CNN architecture which classifies and segments microscopy images using an end-to-end multiple instance scheme. Further, [Ilse 2018] propose an attention-based multi-instance architecture to classify histophathological images.

Our approach extends the notions of these methods towards a generic multi-instance deep learning framework from weak annotations that are augmented by information relevant to the patient's clinical data.

## 4.3 Methodology

In this section, we present a deep learning-based framework for the task of predicting tumoral lymphocytosis. We will start with introducing the CNN-based models we used, and then describe a classical radiomics-based approach that we compare against.

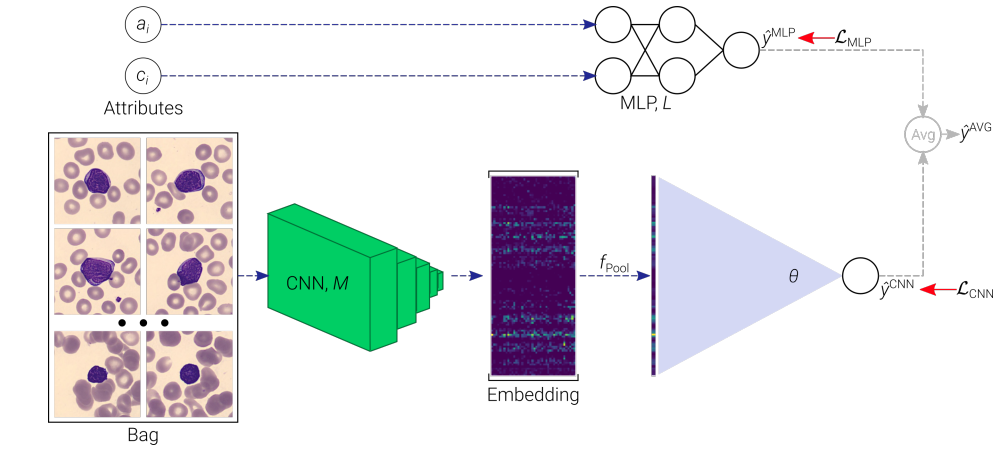
Let us first introduce some notation to describe the proposed approach. We are given a set of $N$ subjects, with a set of images being associated with each patient—the number of which can vary from one patient to the other— along with patient attributes, namely age and lymphocyte count. We represent the data of a patient as

$$S_i = \left( \left\{ X_i^j \right\}_{j=1}^{j=N_i}, a_i, c_i, y_i \right) , \tag{4.1}$$

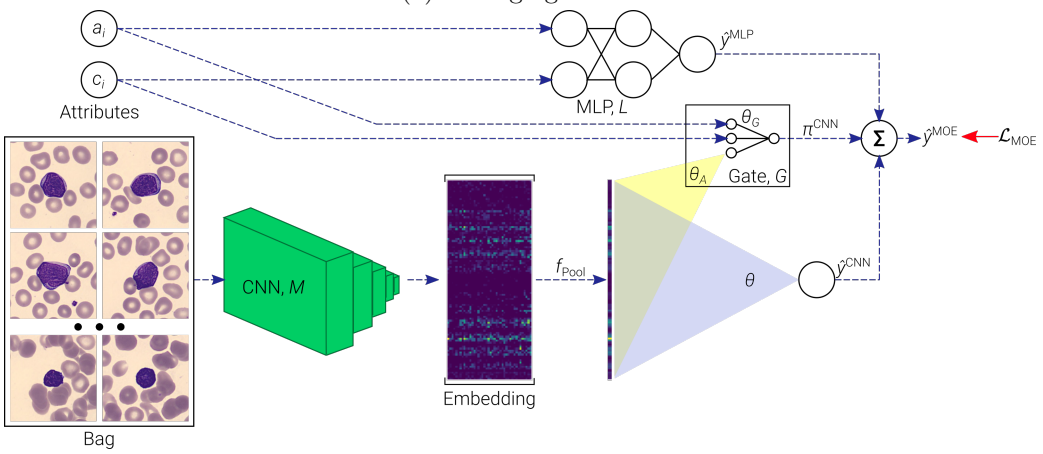
where $\{X_i^j\}$ represents the $N_i$ images obtained from the $i$-th patient, and $a_i$, and $c_i$ represent their age in years, and lymphocyte count in number of cells per litre of blood, respectively. The target class is represented by a binary variable $y_i \in \{0, 1\}$, with the values indicating a normal and malignant case, respectively. We will use this notation throughout the chapter, while further pertinent notation shall be introduced later.

### 4.3.1 Standard MIL assumption

In the standard MIL assumption each instance is considered to fall into one of the two categories—positive (1), or negative (0) [Dietterich 1997, Foulds 2010a].




(a) Averaging Model




(b) Mixture-of-Experts (MOE) model

Figure 4.2: Schematic representations of the two models—*top*: averaging model; and *bottom*: mixture-of-experts (MOE) model. Red arrows in the averaging model indicate that data flow through these arrows is not involved in the training phase, but only in the prediction phase. Further, $\Sigma$ in the MOE model refers to Equation 4.17. $\mathcal{L}_{\text{CNN}}$ and $\mathcal{L}_{\text{MLP}}$ in the figure indicate that these two sub-networks can be trained separately. We refer to the CNN model as the CNN trained only with $\mathcal{L}_{\text{CNN}}$, and the MLP model as the MLP trained only with $\mathcal{L}_{\text{MLP}}$.

Furthermore, the existence of one or more positive class instances in the bag renders the bag itself positive. Concretely, this can be written as

$$y_i = \begin{cases} 1 & \text{if } \sum_j y_i^j \geq 1\,, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

Since $y_i^j \in \{0, 1\}$, this equation can further be simplified as

$$y_i = \max_j y_i^j\,. \tag{4.3}$$

This, however, is quite a strong assumption for our problem.

Firstly it requires knowledge of the instance-level class which is not available. To address this challenge, many algorithms arbitrarily assume that each instance inherits the class from the bag it belongs to. Such an assumption is not suitable for our problem because the malignancy of an individual lymphocyte is uncertain looking only at the blood smear as cytologists can have differing opinions on the matter.

Secondly, the presence of only one *abnormal* lymphocyte does not justify a diagnosis of tumoral lymphocotycosis for the patient. To more closely adhere to the knowledge-based diagnostic approach followed by cytologists, we should draw inference form all instances in a bag.

This reasoning leads us to the choice of a more general aggregation approach. The aggregation function should be one that is invariant to permutation of the instances. Broadly, we can classify aggregation approaches into two classes instance-level and embedding-level MIL [Foulds 2010a]-

1. **Instance-level MIL**. This approach aggregates instance-level predictions to give bag-level predictions. Thus, a model predicts $y_i^j$, which is followed by an aggregation function to yield an estimate of $y_i$. Examples of aggregation functions that fall into this category are the max and mean functions [Hussain 2018], log-sum-exp [Ramon 2000], log-mean-exp, noisy-or [Maron 1998], and noisy-and [Kraus 2016].

2. **Embedding-level MIL**. In this approach, instead of aggregating predictions at the instance-level, a low-dimensional embedding of instances is learnt, and a bag-level classifier is trained on top of the aggregation of the embeddings of all instances in the bag. We shall refer to the vector resulting after the aggregation as the *pooled feature vector*, and the aggregation operation itself as *pooling*. This approach was employed in [Wang 2018], and also shown to perform well on document classification [Denil 2014, Kotzias 2014], as well as whole-slide histopathology images [Hou 2016, Ilse 2018].

It can easily be observed that the max approach discussed above is indeed invariant to permutation, and is an instance-level approach. In this chapter, we employ a deep-learning model with embedding-level pooling. The premise for using an embedding-level approach is based on the earlier discussion on why the standard multi-instance learning assumption is not a valid assumption for this problem. We will also introduce the pooling functions that were employed, and compare the performances of each of them.

### 4.3.2 Proposed Deep Learning Architecture

The proposed deep learning architecture consists of a convolutional neural network as a feature extractor. The CNN works on the entire (unmasked) lymphocyte image. We aim to predict the probability $P(y_i = 1 \mid S_i)$, where the variable $y_i = 1$ indicates

the presence of disease, and $y_i = 0$, otherwise. We investigate three neural network models to model this probability. We will first introduce the deep learning model which draws patient-level inference using only the images $\{X_i^j\}$. In this setting, we will refer to each patient as a *bag*, and to the $y_i$s as *bag-level labels*. Similarly, we call each image in $\{X_i^j\}$ an *instance*, and use $y_i^j$ to denote *instance-level* labels corresponding to $X_i^j$. In the supplied ground truth, we are provided only with $y_i$.

#### 4.3.2.1 CNN for Blood Smears

To generate the low-dimensional embeddings, we use a convolutional feature extractor. The feature extractor is followed by a pooling operation in the embedding space, and a classifier which predicts the probability of disease trained on top of the pooled representations. We design the model so that it is end-to-end trainable, in that it learns the low-dimensional embedding as well as the classifier jointly.

Let $M$ be the function that represents the feature extractor. $M$ operates on instances $(X_i^j)$ and generates an embedding in a low-dimensional space. Let $f_{\text{Pool}}$ represent a pooling function on these embeddings which is permutation-invariant. As mentioned before, we employ three pooling functions—the element-wise maximum function ($f_{\text{Max}}$), the element-wise average ($f_{\text{Mean}}$), and the log-sum-exp function ($f_{\text{LSE}}$). The pooling functions used in this chapter are defined as—

$$f_{\text{Max}}\left(\{\mathbf{h}_i^j\}\right) = \left(\max_j \mathbf{h}_i^j(k)\right)_{1\le k\le E}; \tag{4.4}$$

$$f_{\text{Mean}}\left(\{\mathbf{h}_i^j\}\right) = \frac{1}{N_i}\sum_j \mathbf{h}_i^j; \text{ and} \tag{4.5}$$

$$f_{\text{LSE}}\left(\{\mathbf{h}_i^j\}\right) = \frac{1}{r}\log\left(\frac{1}{N_i}\sum_j \exp\left(r\cdot\mathbf{h}_i^j\right)\right), \tag{4.6}$$

where $\mathbf{h}_i^j = M(X_i^j) \in \mathbb{R}^E$ represents the embedding of $X_i^j$, and $E$ is the dimension of this embedding. We will further represent the pooled embedding over all instances by the vector $\mathbf{p}_i$, i.e.,

$$\mathbf{p}_i = f_{\text{Pool}}\left(\{\mathbf{h}_i^j\}\right), \tag{4.7}$$

for $f_{\text{Pool}} \in \{f_{\text{Max}}, f_{\text{Mean}}, f_{\text{LSE}}\}$. For the convolutional neural network $M$, we use a ResNet [He 2016]. However, we set the the width of the ResNet as a hyperparameter of our model. Denoting by $K$, the the base "step size", shown in Table 4.1 is the architecture of the ResNet, with the number of channels doubling at each residual layer. In a standard ResNet [He 2016], $K$ is set to 64. However, the original ResNets were intended for large-scale computer vision applications, and as such use "wide" latent representations. Since our problem has limited data, we make a design choice to experiment with different values of $K$, where $K \in \{8, 16, 32, 64\}$.

Once we have an aggregate representation of a bag as an embedding in the low-dimensional space, we pass it through a linear classifier to predict the bag label.

The linear classifier assigns a score to the bag given by

$$\hat{y}_i^{\text{CNN}} = \boldsymbol{\theta}_C^\top \mathbf{p}_i + \beta \,, \text{and} \tag{4.8}$$

$$P(y_i = 1 \mid \{X_i^j\}) = \sigma\left(\hat{y}_i^{\text{CNN}}\right) . \tag{4.9}$$

Here, $\sigma(x) = \dfrac{1}{1+\exp(-x)}$ is the logistic function, and $\boldsymbol{\theta}_C$ and $\beta$ are, respectively, the weight vector and the bias of the classifier. Overall, this models the bag probability as a Bernoulli distribution. To train the model end-to-end, we use the negative log-likelihood. Concretely, the loss function is defined as

$$\mathcal{L}_{\text{CNN}} = -\ln \sigma(\hat{y}_i^{\text{CNN}}) \,\big|_{y_i=1} - \ln\left(1 - \sigma(\hat{y}_i^{\text{CNN}})\right) \big|_{y_i=0} \,. \tag{4.10}$$

We will henceforth refer to this model as CNN (in small caps). It should be noted that CNN uses only the images for diagnosis.

#### 4.3.2.2 Multi-layer Perceptron for Clinical Data

Since the clinical data $(a_i, c_i)$ are also helpful for cytologists during their diagnosis, we employ them as well in our model as an additional source of information. In order to integrate the clinical data, we use a multi-layer perceptron consisting of one hidden layer and one output layer to predict the probability of disease.

The multi-layer perceptron consists of an input layer with two units, connected to a hidden layer which also has two units. The sigmoid activation function is used in the hidden layer. The output layer has just one unit which represents the score of the classifier. We denote the score for a bag $i$ by $\hat{y}_i^{\text{MLP}}$. The multi-layer perceptron is also trained with the negative log-likelihood loss as described in Equation 4.10. Let $L$ represent the multi-layer perceptron. Then we can write

$$\hat{y}_i^{\text{MLP}} = L(a_i, c_i) \,; \tag{4.11}$$

$$P(y_i = 1 \mid a_i, c_i) = \sigma(\hat{y}_i^{\text{MLP}}) \,; \text{ and} \tag{4.12}$$

$$\mathcal{L}_{\text{MLP}} = -\ln \sigma(\hat{y}_i^{\text{MLP}}) \,\big|_{y_i=1} - \ln\left(1 - \sigma(\hat{y}_i^{\text{MLP}})\right) \big|_{y_i=0} \,. \tag{4.13}$$

We will henceforth refer to this model as MLP. It should be noted that MLP does not use images for diagnosis.

We have described two models that use different training data to predict the same variable. The two types of input data are not completely independent of each other, but the extent of the relationship is unknown at worst and difficult to model at best. Based on this, we combine the predictions of the two models in two possible ways.

#### 4.3.2.3 Averaging Model

The averaging model simply averages the two scores from these two models. The combined prediction is

$$\hat{y}_i^{\text{AVG}} = \frac{1}{2}\left(\hat{y}_i^{\text{CNN}} + \hat{y}_i^{\text{MLP}}\right) . \tag{4.14}$$

Since each predictor can be trained separately, there is no joint training in the averaging model. We refer to this model as AVG.

#### 4.3.2.4 Mixture-of-Experts Model

So far, we have represented a patient either as a set of images (CNN model), or as a set of clinical attributes (MLP model) and used different models for them. It is not unreasonable to assume that the two models might have disagreements over certain examples, as the biologists themselves are not always in agreement. It therefore makes sense to *choose* the better of the two models depending on the patient. To this end, we propose to employ a mixture-of-experts [Hampshire 1992, Jordan 1994, Moerland 1997] model to learn simultaneously from both, the images as well as the clinical attributes. However, instead of targeting cooperation between the models, in which minimization of the loss function targeting to minimise the prediction error over the average of all models' predictions, we wish to promote specialisation, such that each model specializes over a certain set of examples [Jacobs 1991]. More concretely, we have two "experts"—the CNN and the MLP—with a gating network weighting the contributions of the two experts. The gating network operates on the pooled features $\mathbf{p}_i$, as well the attributes $a_i$ and $c_i$, and outputs a set of mixing coeffecients. Such a model learns to output a mixture of probability distributions learnt by each of the experts. Examples of uses of a mixture-of-experts are applications to speech recognition [Jacobs 1991, Nowlan 1991, Waterhouse 1998] and disease classification [Ng 2007], among other tasks.

We formulate the gating network as an aggregation kernel learnt on the embedding space, followed by a linear layer to regress the contributions. The complete model used for the gating network is

$$\pi_i^{\text{CNN}} = G(\mathbf{p}_i, a_i, c_i) = \sigma\left(\boldsymbol{\theta}_G^\top \begin{bmatrix} \boldsymbol{\theta}_A^\top \mathbf{p}_i \\ a_i \\ c_i \end{bmatrix}\right) \text{ ; and} \tag{4.15}$$

$$\pi_i^{\text{MLP}} = 1 - \pi_i^{\text{CNN}} , \tag{4.16}$$

where $\pi_i^{\text{CNN}}$ and $\pi_i^{\text{MLP}}$ contributions of the CNN and the MLP, respectively. The final prediction of the mixture-of-experts model is given by

$$P(y_i = 1 \mid S_i) = \hat{y}_i^{\text{MOE}} = \pi_i^{\text{CNN}} \sigma\left(\hat{y}_i^{\text{CNN}}\right) + \pi_i^{\text{MLP}} \sigma\left(\hat{y}_i^{\text{MLP}}\right) , \tag{4.17}$$

Contrary to the parameter free AVG model, the mixture-of-experts uses the gating network is parameterized $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_G$ as trainable parameters, and hence can be

| Layer | Layer Name | Output Size | Residual Blocks |
|---|---|---|---|
| 1 | conv1 | $112 \times 112$ | - |
| 2 | conv2 | $56 \times 56$ | $\begin{bmatrix} 3 \times 3, K \\ 3 \times 3, K \end{bmatrix}$ |
| 3 | conv3 | $28 \times 28$ | $\begin{bmatrix} 3 \times 3, 2K \\ 3 \times 3, 2K \end{bmatrix}$ |
| 4 | conv4 | $14 \times 14$ | $\begin{bmatrix} 3 \times 3, 4K \\ 3 \times 3, 4K \end{bmatrix}$ |
| 5 | conv5 | $7 \times 7$ | $\begin{bmatrix} 3 \times 3, 8K \\ 3 \times 3, 8K \end{bmatrix}$ |
| 6 | flatten | $7 \cdot 7 \cdot 8K$ | - |

Table 4.1: Architecture of the convolutional neural network $M$ used to estimate $\mathbf{h}_i^j$. The input to the network is an image of size $224 \times 224$. Each row defines an operation, where each convolution is followed by batch normalisation and ReLU. The residual layers are layers 2-5. $K$ is a hyperparameter which determines the width of the residual network. We test with the values $\{8, 16, 32, 64\}$ for $K$.

trained end-to-end. The loss function to train this model encourages specialisation, in that it lets each expert concentrate on examples it can classify better. We use the following loss fuction—

$$\mathcal{L}_{\text{MOE}} = -\ln \hat{y}_i^{\text{MOE}} \big|_{y_i=1} - \ln \left(1 - \hat{y}_i^{\text{MOE}}\right) \big|_{y_i=0} \ . \tag{4.18}$$

We will henceforth refer to the mixture-of-experts model as MOE. We further explore three paradigms in the MOE framework—(P1) training the entire model end-to-end with no initialisation; (P2) initialising the cnn and mlp with models trained uniquely with $\mathcal{L}_{\text{CNN}}$ and $\mathcal{L}_{\text{MLP}}$, respectively, and then training only $G$; and (P3) initialising the cnn and mlp as before, and training end-to-end.

### 4.3.3 Training

We train our networks with the negative log likelihood loss. Depending on which model is used, we employ one of the losses out of $\mathcal{L}_{\text{CNN}}$, $\mathcal{L}_{\text{MLP}}$, and $\mathcal{L}_{\text{MOE}}$. Training is performed with standard backpropagation. The entire batch was used to get gradients at every epoch leads to a much more stable model than using a smaller batch size. However, for the CNN feature extractor, a smaller batch of size 32 was used. This means the batch normalisation layers of the CNN $M$ have as input a batch of size 32 and compute statistics accordingly. This number, however, can be increased without increase in GPU memory usage.

We train several models with different combinations of configurations, i.e., with varied combinations of $K$, $f_{\text{Pool}}$, training data (images, attributes), and averaging and mixture-of-experts models.

### 4.3.4 Overfitting

As we have very few training examples, we find that the model is susceptible to overfitting. With the available data and the randomly selected train, validation and test splits, the number of training examples is significantly smaller than the trainable parameters of our architectures.

To reduce overfitting, we introduce standard data augmentation during training of the CNN. We add random flips along the $x$- and $y$-axes, as well as random rotations from the set $\left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}$. We additionally use standard colour augmentation [Krizhevsky 2012]. We use a probability of 0.5 for each of the flips and rotations. We also use standard colour augmentation [Krizhevsky 2012]. More specifically, First, we perform PCA on RGB pixel values over the training dataset. Next, for a training image, three values, $\alpha_i$, are sampled from a normal distribution with mean 0 and standard deviation 0.1. The colour of the training image is then rescaled by adding

$$[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]\,[\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^\top \tag{4.19}$$

where $\mathbf{e}_i$ and $\lambda_i$ are, respectively, the eigenvectors and eigenvalues of the $3 \times 3$ covariance matrix of RGB pixel values over the entire training dataset.

Furthermore, we record the performance of the model on the validation set at each training epoch. We keep the model that performs best on the validation set at any point during training. Using the best model for evaluation is equivalent to employing early stopping, where we stop the training process before it has converged on the training set.

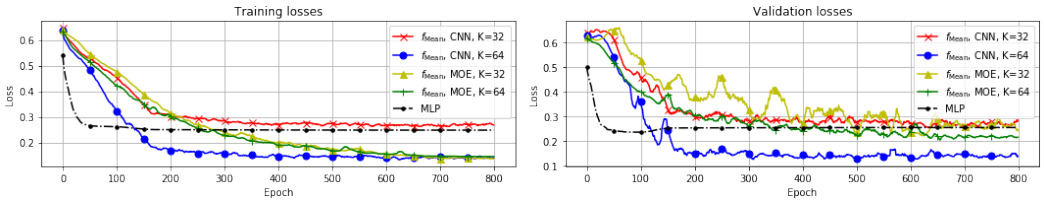



Figure 4.3: Training and validation losses for different models using the $f_{\text{Mean}}$ pooling function with varying ResNet widths.

### 4.3.5 Implementation Details

The code was written in Python with the PyTorch [Paszke 2017] library, and executed on a machine equipped with a NVIDIA GTX 1080 GPU, a 12-core 3.5 GHz processor, and 32 gigabytes of memory. The models were trained using the Adam optimiser [Kingma 2014b], starting with a learning rate of 0.0001, and decay it by a factor of 0.1 every 96,000 iterations. We use $\beta_1 = 0.9$ and a weight decay of 0.0005. We train for 220,000 iterations for training. This number is not fixed, because we choose the model that generalises best on the validation set during training. Training took about one and a half days per model.

The original images in our dataset are of size $360 \times 360$, but we resize them to $224 \times 224$, as we observed that we do not lose any significant information under the resizing operation, and it allows us to curb overfitting as well as use less memory overall. The images are then centered using the per-pixel mean over the entire dataset.

## 4.4 Compared Methods

In this section the methods that are used for comparison are presented together with their implementation details.

### 4.4.1 Classical Radiomics

To evaluate the performance of our proposed method we compare it with an MIL framework using classical radiomics.

#### 4.4.1.1 Feature Extraction

Before we employ an MIL scheme, we need to extract radiomics features. These features must be extracted from the area of interest, i.e., the lymphocytes in the images. Under this framework, a segmentation of lymphocytes in needed to compute several imaging and shape characteristics. The lymphocytes were automatically segmented in each image instance $X_i^j$ by first smoothing every image using the Simple Linear Iterative Clustering (SLIC) superpixels algorithm [Achanta 2012]. SLIC is a gradient ascent method implementing a local $K$-means clustering to generate a $K$-superpixel segmentation. Since the best value for $K$ is not known in advance, we perform multiple segmentations for different values of $K$ and then created an average segmentation for each instance and each of the RGB image channels. This multi-scale fuzzy segmentation step did not require any parameter tuning and aimed to smooth the boundaries of ambiguous regions while at the same time retain the crisp boundaries of regions that were present in more scales. The smoothed RGB image was then segmented using the $K$-means clustering algorithm in HSV (hue, saturation, value) colour representation scale using $K = 3$. The three obtained clusters represented *i)* the lymphocytes (with the cytoplasm), *ii)* all other cells, and *iii)* the background. If several lymphocytes were found in an image, only the largest of them was retained and used for feature extraction. The radiomic analysis was based on 94 features extracted from each of the segmented blood smear images per subject. These features are described in detail below.

**Shape** The shape of the largest lymphocyte in each image was described by 12 features: area, major axis length, minor axis length, eccentricity, convex area, filled area, Euler number, equivalent diameter, solidity, extent, and perimeter calculated in two ways using different weights for diagonal pixels and corners.

**Image Statistics** 3 intensity statistics (minimum, maximum, average) were extracted for each of the 3 RGB channels inside the region of interest.

**Texture** 24 texture variables [Haralick 1973] including the average fractal dimension and statistical measures (autocorrelation, contrast, correlation$_1$, correlation$_2$, cluster prominence, cluster shade, dissimilarity, energy, entropy, homogeneity$_1$, homogeneity$_2$, maximum probability, sum of squares, sum average, sum variance, sum entropy, difference variance, difference entropy, information measure of correlation$_1$, information measure of correlation$_2$, inverse difference, normalised inverse difference, moment normalised inverse difference) from the gray-level co-occurrence matrix were calculated for pairs of pixel in 0°, 45°, 90°, 135°, for each of the 3 channels.

**Density** The number of lymphocytes in the image was used as a measure of cell density.

#### 4.4.1.2 MIL Training

The feature vectors from all blood smear images of each subject comprised a multiple instance dataset which was introduced into a MIL classifier for prediction of lymphocytosis. We investigated several standard MIL classifiers from the multiple instance learning literature, such as the expectation maximization maximum diverse density (EMDD) [Zhang 2002], multi-instance support vector machine (MI-SVM) [Andrews 2003], multi-instance learning in embedded subspaces (MILES) [Chen 2006], but the best performing, which was finally selected, was the specialising MIL (SPEC_MIL) which is a generalisation of MI-SVM. The only hyperparameter in this algorithm is the fraction of positive instances, which was tuned by 3-fold cross validation on the training set and then fixed to the value attaining most often the highest classification accuracy. Two experiments were performed. The one relied only on the radiomics features whereas the other included also $a_i$ and $c_i$. Integration of the clinical variables with the radiomics features was performed in an early phase and led to a joint dataset that was introduced into the multi-instance classifiers.

### 4.4.2 Attention-based Methods

We also experimented with the attention-based model recently proposed in [Ilse 2018]. In this approach, a CNN is trained along with an attention mechanism which learns to focus on discriminative images in data. This approach employs a pooling function $f_{\text{Pool}}$ in the latent space which is effectively a weighted average, with the weights being determined by the attention mechanism. This can formally be written as

$$\mathbf{m}_i^j = \text{ReLU}(\mathbf{V}_0 \mathbf{h}_i^j)\,, \tag{4.20}$$

$$\mathbf{p}_i = f_{\text{Attention}}\left(\{\mathbf{h}_i^j\}\right) = \sum_{j=1}^{N_i} w_i^j \mathbf{m}_i^j\,, \tag{4.21}$$

where $\mathbf{m}_i^j$ are lower-dimensional embeddings of $\mathbf{h}_i^j$, and the weights $w_i^j$ are given by

$$w_i^j = \frac{\exp\left(\mathbf{W}^\top \tanh\left(\mathbf{V}_1 \mathbf{m}_i^j\right)\right)}{\sum_{k=1}^{N_i} \exp\left(\mathbf{W}^\top \tanh\left(\mathbf{V}_1 \mathbf{m}_i^j\right)\right)}. \tag{4.22}$$

$\mathbf{W} \in \mathbb{R}^{L\times 1}$, $\mathbf{V}_0 \in \mathbb{R}^{D\times E}$, and $\mathbf{V}_1 \in \mathbb{R}^{L\times D}$ are learnable parameters of the attention mechanism. We used $L = 16$ and $D = 64$ in our experiments, We observed that higher values for $L$ and $D$ caused the models to overfit.

## 4.5 Dataset

Blood smears and patient attributes were collected from 204 patients. The inclusion criteria were (a) a lymphocyte count above $4 \times 10^9$/L, and (b) absence of opposition to the research. The blood smears were automatically produced by a Sysmex automat, and the nucleated cells were automatically photographed with a DM-96 device (Cellavision). All the cells labelled as lymphocytes by the DM-96 device were used for analysis. To determine the presence of the lymphocytosis, flow cytometry was used incorporating a panel of antibodies for the diagnosis of lymphoproliferative disorders (CD3, CD4, CD5, CD8, CD10, CD56, CD20, CD19, kappa, lambda). In our dataset, the minimum and maximum number of images per patient were 16 and 198, with a mean and standard deviation of 82 and 45, respectively.

#### 4.5.0.1 Data split

The training cohort used of all our models consists of 142 subjects with 44 reactive and 98 malignant cases. The validation cohort consists of 21 subjects with 6 reactive and 15 malignant cases, while the test cohort includes 42 subjects with 13 reactive and 29 malignant examples.

## 4.6 Experimental Results

| Model | $K = 8$ | | | $K = 16$ | | | $K = 32$ | | | $K = 64$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $f_{\text{Max}}$ | $f_{\text{Mean}}$ | $f_{\text{LSE}}$ | $f_{\text{Max}}$ | $f_{\text{Mean}}$ | $f_{\text{LSE}}$ | $f_{\text{Max}}$ | $f_{\text{Mean}}$ | $f_{\text{LSE}}$ | $f_{\text{Max}}$ | $f_{\text{Mean}}$ | $f_{\text{LSE}}$ |
| CNN | 0.60 | 0.82 | 0.82 | 0.47 | 0.88 | 0.72 | 0.47 | 0.95 | 0.77 | 0.89 | **0.96** | 0.82 |
| AVG | 0.87 | 0.90 | 0.91 | 0.88 | 0.91 | 0.89 | 0.88 | 0.94 | 0.89 | 0.90 | 0.95 | 0.90 |
| MOE-P1 | - | 0.90 | - | - | 0.92 | - | - | 0.94 | - | - | 0.94 | - |

Table 4.2: Area under the receiver operating characteristics curve for various models and configurations. Similarly, CNN is trained only using the images. MOE-P1, 2, and 3 refer to the three paradigms explored when training the MOE model (see Section 4.3.2.4).

In this section, results from the proposed and compared algorithms are presented. The results obtained from the algorithms are also compared with the visual

| **Method** | Data | Sensitivity | Specificity | Accuracy | Balanced Accuracy |
|---|---|---|---|---|---|
| Biologists | imgs, attrs | 0.7529 ± 0.0953 | 0.7885 ± 0.0126 | 0.7639 ± 0.0690 | 0.7707 ± 0.0705 |
| MLP | attrs | 0.8621 | 0.6923 | 0.8095 | 0.7772 |
| Radiomics | imgs | 1.0000 | 0.3846 | 0.8095 | 0.6923 |
| | imgs, attrs | 0.8966 | 0.6923 | 0.8333 | 0.7944 |
| Deep MIL [Ilse 2018], $K = 32$ | imgs | 0.9655 | 0.6923 | 0.8810 | 0.8289 |
| Deep MIL [Ilse 2018], $K = 64$ | imgs | 1.0000 | 0.2308 | 0.7619 | 0.6154 |
| $f_{\text{Mean}}, K = 32$, CNN | imgs | 0.9310 | 0.6923 | 0.8571 | 0.8117 |
| $f_{\text{Mean}}, K = 32$, AVG | imgs, attrs | 0.9310 | 0.6923 | 0.8571 | 0.8117 |
| $f_{\text{Mean}}, K = 32$, MOE-P1 | imgs, attrs | 0.8621 | 0.8462 | 0.8571 | **0.8541** |
| $f_{\text{Mean}}, K = 32$, MOE-P2 | imgs, attrs | 0.8621 | 0.6923 | 0.8095 | 0.7772 |
| $f_{\text{Mean}}, K = 32$, MOE-P3 | imgs, attrs | 0.8621 | 0.6154 | 0.7857 | 0.7387 |
| $f_{\text{Mean}}, K = 64$, CNN | imgs | 0.8621 | 0.8462 | 0.8571 | **0.8541** |
| $f_{\text{Mean}}, K = 64$, AVG | imgs, attrs | 0.9310 | 0.6154 | 0.8571 | 0.8117 |
| $f_{\text{Mean}}, K = 64$, MOE-P1 | imgs, attrs | 0.8621 | 0.8462 | 0.8571 | **0.8541** |
| $f_{\text{Mean}}, K = 64$, MOE-P2 | imgs, attrs | 0.8621 | 0.8462 | 0.8571 | **0.8541** |
| $f_{\text{Mean}}, K = 64$, MOE-P3 | imgs, attrs | 0.8966 | 0.6923 | 0.8333 | 0.7944 |

Table 4.3: Different evaluation metrics for models discussed in this chapter, evaluated on the testing cohort for the diagnosis of lymphocytosis. The second column signifies the type of incorporated training data (imgs: images, attrs: clinical attributes, i.e. age and lymphocyte count), as explained in section A.2.1.7.

assessment annotations from 12 different biologists from the Lyon University Hospital. The biologists provided their diagnoses for each of the patients of the test cohort, based on the images and the supporting clinical data. The obtained results are evaluated and compared based on the following metrics: sensitivity, specificity, accuracy, balanced accuracy and in terms of area under receiver operating characteristic curve (ROC-AUC).

In Table 4.2 different components of the proposed method are evaluated in terms of ROC-AUC. The tested aggregation functions and the width of ResNet ($K$) are evaluated for the CNN, AVG, and MOE-P1 models. The best performance is obtained by $K = 64$ and using the $f_{\text{Mean}}$ pooling operation for both models. Based on these observations we performed the evaluation of the MOE model only for the $f_{\text{Mean}}$ pooling operation.

In Table 4.3 a comparison between the different methods is presented. To calculate the evaluation metrics we did not perform any optimisation for the threshold value and a value of 0.5 is used to separate healthy and diseased patients for all the methods. In general, the predictions of the biologists have a very wide variation that can reach even 9% for the sensitivity metric. Moreover, we can observe that a lot of information is captured by patient attributes as a relatively simple MLP performs better than the experts in almost all of the evaluation metrics and achieves similar balanced accuracy with them. This good influence of the patient attributes is also indicated in classical radiomics as our experiments indicate a boost in the

overall and balanced accuracy when the attributes are combined with the predefined features extracted from the images. However, the performance of the classical radiomics is inferior to the one reported by the attention based method [Ilse 2018]. The latter obtains quite high sensitivity but relatively low specificity indicating that this method detects much more false positives for the diseased category.

Table 4.3 summarises also the performance of our proposed method using different configurations and parameters. The most of them outperform all baselines with all the metrics, while all of them are higher than 70%. In particular, different configurations of the proposed method namely the $f_{\text{Mean}}, K = 32$, MOE-P1, $f_{\text{Mean}}, K = 64$, CNN and $f_{\text{Mean}}, K = 32$, MOE-P1 report the highest balanced accuracy while they also report very high values for the rest of the metrics. This proves the robustness of the method on different configurations. Moreover, one very interesting point is that the proposed model based solely on imaging information can perform similarly with models that use additional source of information about the patients.

For a better and more complete evaluation of the reported methods we compare the areas under their ROC (Figure 4.4) curves (ROC-AUC) in Table 4.4. In general all the methods report ROC-AUC greater than 0.83 with the proposed method using the $f_{\text{Mean}}$ reporting values higher than 0.91 proving its robustness and stability. Finally, a simple MLP using only the clinical information of the patients reports an ROC-AUC of 0.88 which is at least 3% lower than the models that are using information produced by the images.

| Model | ROC-AUC |
|---|---|
| Biologists' average | 0.9204 |
| MLP | 0.8912 |
| Radiomics, img | 0.8727 |
| Radiomics, img+attr | 0.8329 |
| Deep MIL [Ilse 2018], $K = 32$ | 0.9151 |
| Deep MIL [Ilse 2018], $K = 64$ | 0.9390 |
| $f_{\text{Mean}}, K = 32$, CNN | 0.9416 |
| $f_{\text{Mean}}, K = 64$, CNN | **0.9629** |
| $f_{\text{Mean}}, K = 32$, MOE-P1 | 0.9416 |
| $f_{\text{Mean}}, K = 32$, MOE-P2 | 0.9178 |
| $f_{\text{Mean}}, K = 32$, MOE-P3 | 0.9151 |
| $f_{\text{Mean}}, K = 64$, MOE-P1 | 0.9443 |
| $f_{\text{Mean}}, K = 64$, MOE-P2 | 0.9178 |
| $f_{\text{Mean}}, K = 64$, MOE-P3 | 0.9390 |

Table 4.4: A comparison of the CNN, MOE using the $f_{\text{Mean}}$ pooling function with attention models by ROC-AUC together with the comparisons with the attention module, classical radiomics and an MLP-only model trained only with the patient attributes. We also show a comparison against the average prediction of the twelve expert biologists.

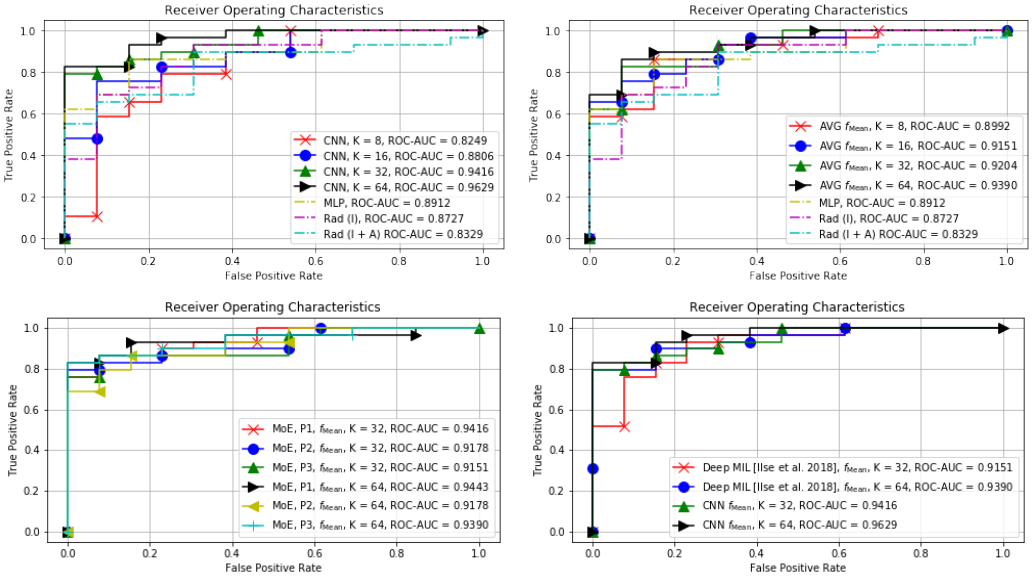



Figure 4.4: Receiver operating characteristics for methods discussed in this chapter. *Top left*: curves for the CNN-only models compared with the MLP as well as radiomics; *top-right*: curves for the averaging model; *bottom-left*: the mixture-of-experts model; and *bottom-right*: comparison of the CNN models with attention models from [Ilse 2018]. All CNN models shown here use the $f_{\text{Mean}}$ aggregation function.

### 4.6.1 Repeatability

In order for our system to be used in clinical practice, the proposed models need to be robust in terms of repeatability. That is to say, the proposed models should arrive at the same conclusion as long as a clinically relevant set of images is sampled from a patient for testing. To this end, we design a test to assess the performance of our model over several observations on the same patient. On five additional patients, five different sets of blood smears were extracted, where the patients had representatives of both reactive and tumoral cases. This results in five different sets of images per patient. Ideally, each set of images for a patient should result in the same conclusion after analysis. The goal of this test of repeatability is then to evaluate the performance of the proposed models per smear for each patient and examine the variance that is introduced in them.

In Table 4.5, we list the result of the best CNN, MLP, and MOE models. The true behaviour for each patient is listed in row 2, while the prediction of each of the two models is listed in columns 4, 5, and 6. For the *prediction* row of CNN and MOE-P1, each column indicates which image set (row 3) was used for diagnosis, whereas the *maj. vote* row is the diagnosis obtained by a majority vote over the predictions on the image sets. We note that the MOE model is more stable in terms of its conclusion with much fewer intra-patient disagreements, whereas the images-only CNN model is more sensitive to the set of sampled images as there are more

intra-patient disagreements. By majority vote, the MOE model is also able to give the correct prediction for each patient while the CNN and MLP models fail.

## 4.7 Discussion

To the best of our knowledge this is the first study that provides a deep learning-based method for accurate diagnosis of lymphocytosis. Our proposed method is evaluated with standard multi instance learning schemes that are used in literature and other recently proposed deep learning based methods while it is also compared with the visual assessment of 12 different biologists. Our experiments indicate the superiority of our method, showing the potential of such a tool for automated diagnosis of lymptocytosis in clinical practice.

Different pooling operators, network parameters (number of channels $K$), configurations (using images and clinical attributes) and training strategies are reported in this study in order to show the behavior of our proposed method. Starting with the pooling operators, $f_{\text{Max}}$ and $f_{\text{LSE}}$ show comparatively poor performance over the tested values of $K$. We postulate that the $f_{\text{Max}}$ operator is too strong for the problem at hand, while the $f_{\text{LSE}}$ operator—being a smooth approximation to the maximum—performs better than $f_{\text{Max}}$. However, both $f_{\text{Max}}$ and $f_{\text{LSE}}$ fail to capture the relationship between the instance and the bags unlike $f_{\text{Mean}}$, which reports the best performance for all the configurations.

Our experiments further indicate that the models work better when a wider ResNet is used, i.e., a higher value of $K$. In particular, narrower networks seem not powerful enough to capture all the available information and learn enough features from the data. This is in accordance with other studies in literature [**?**]. However, both $K = 32, 64$ perform similarly with both values reporting very close performance (Tables 4.2 and 4.3). This result also alleviates overfitting concerns with wider models.

Concerning the training strategies used for the MOE model, our experiments indicate that the MOE model works better with the P1 training paradigm, where both models are trained end-to-end without initialisation. This is expected behaviour under the training loss used. The training is done to encourage specialisation, but under the P2 and P3 paradigms, the participating experts (CNN and MLP) have been pre-trained to fit the entire data instead of specialising over a portion of it, whereas under the P1 paradigm, they are uninitialised.

We argue that the attention-based models [Ilse 2018], which are one of the competing methods, have lower performance than the proposed approach because of the high variance in the number of images per patient. As the attention-based models use a softmax function to compute image weights (Equation 4.22), these weights tend to become skewed when there are several *important* examples in the set. This renders the learning the classifier a more difficult task.

Here it is worth mentioning that almost all the methods reach similar and higher performances compared to the experts indicating the high potentials of such a

tool in clinical practice. Our experiments also show that a deep learning-based method is able to extract more discriminative features than a classical radiomics-based approach. The performance of the images-only model (CNN) shows that it is possible to extract and exploit information from blood smears using an automated tool. However, it is still sensitive to the set of images extracted as seen in Section **??**. The MOE model, on the other hand, is able to correct errors that the CNN and MLP models were individually making. This indicates that while neither of clinical attributes and images alone is enough to make a reliable diagnosis, the MOE model is able to combine information from them for the correct diagnosis. This demonstrates the robustness of the MOE model to data acquisition.

Finally, another aspect that can be taken into account is the time-efficiency of the proposed approaches. The proposed methods are fast when drawing inference, making the entire process rapid and efficient. For our test cohort which contains 42 patients, testing required 30s in all, which corresponds to about 0.72s per test example. This time is better than the one needed by a biologist who may need considerably longer for the examination of one case.

## 4.8 Summary

This chapter presents a deep MIL scheme for reliable diagnosis of lymphocytosis. Imaging features from lymphocytes which are extracted automatically are coupled with patient attributes in a dynamic way taking advantage of all available information for each patient. Our method has been validated under different training schemes and different pooling operators proving its robustness and accuracy. Moreover, it has been also evaluated against human experts, classical radiomics MIL frameworks and recently proposed attention-based methods. We also propose a mixture-of-experts model which combines information from acquired blood smears and clinical attributes of a patient for a more robust assessment. Overall, we found that deep learning based approaches outperform conventional methodologies while models that are based only on the images report better performance demonstrating their diagnostic capacity for lymphocytosis prediction. A repeatability test also evaluates the robustness of the CNN and the MOE models and demonstrates that the MOE model is indeed able to combine information from the two sources efficiently (attributes and images) for a more reliable diagnosis.

As we can also see from Table 4.3, our method outperforms the biologists' average prediction. With all our experiments, we demonstrated that our method can give a reliable tool to biologists in order to assist them in their everyday practice, being deployed in real-life scenarios. However, further tests, especially using datasets from different hospitals, must be undertaken in order to extensively validate the accuracy of the method. This is the principal immediate future direction of our work.

| **Patient** | | 1 | | | | | 2 | | | | | 3 | | | | | 4 | | | | | 5 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Ground truth** | | Reactive | | | | | Reactive | | | | | Reactive | | | | | Tumoral | | | | | Tumoral | | | | |
| **Image set** | | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| **cnn** | predictions | R | T | R | T | T | T | T | T | T | T | T | R | R | R | T | T | T | T | T | T | T | T | T | T | T |
| | majority vote | Tumoral | | | | | Tumoral | | | | | Reactive | | | | | Tumoral | | | | | Tumoral | | | | |
| **mlp** | predictions | Reactive | | | | | Tumoral | | | | | Tumoral | | | | | Tumoral | | | | | Reactive | | | | |
| **moe-P1** | predictions | R | R | R | R | R | R | T | R | T | R | R | R | R | R | R | T | T | T | T | T | T | R | T | T | T |
| | majority vote | **Reactive** | | | | | **Reactive** | | | | | **Reactive** | | | | | **Tumoral** | | | | | **Tumoral** | | | | |

Table 4.5: Test of repeatability. Please see text for more details.

Chapter 5

# Conclusion and Future Work

Within this thesis, we have discussed two challenging problems in computer vision and medical imaging. Below, the contributions of this thesis are summarised in more detail with discussions of potential future work to advance the ideas presented in this thesis.

- We propose a simple autoencoder model, deforming autoencoder (DAE), to discover a canonical space for an object category and simultaneously infer a dense mapping between images and the canonical space. It is demonstrated quantitatively that objects are well aligned in this canonical space.
- Modelling appearance in this canonical space allows us to disentangle object appearance and shape, with the shape being modelled by the dense deformation grid. This is achieved via. a simple, novel method of regressing the deformation grid which allows the shape decoder to predict semantically meaningful deformations. A differentiable warping operation based on bilinear interpolation allows use of reconstruction loss on the warped appearance.
- Applications of this method of modelling of the deformation grid to medical imaging and remote sensing demonstrate efficient unsupervised registration of two images are demonstrated. For the medical imaging application, we register 3D MRI scans of the lung taken during inspiration phase with those taken during the expiration phase. Similarly, we register satellite images of the same regions taken one year apart.
- We show that the canonical space facilitates further disentanglement of appearance (texture) into albedo and shading. This is achieved by regressing shading and albedo separately—with the shading being a single channel

image—and weighting the albedo with the shading map. Since shading and albedo both exist in the canonical coordinate system, warping only one of them using the deformation grid gives us shading and albedo maps in the image space.

- We also introduce a weakly-supervised variant of the DAE, which introduces a classification component to the latent space. This component is shared by both, the appearance and shape decoders, and is learnt using the negative log-likelihood loss. We demonstrate that this allows us to learn multiple appearance spaces for datasets containing several clusters of examples.

- We extend the notion of unsupervised dense alignment in 2D to recover 3D shape using NRS$f$M. Using an autoencoder, image-specific mesh deformation and pose parameters are predicted, while a base mesh is learnt simultaneously. The system is trained using reprojection loss, with the ground truth for keypoints supplied by the DAE.

- Estimating shading using cues from the learnt morphable model is demonstrated. The learnt shape is used to predict a shading map which is then refined using a U-Net, resulting in albedo-shading disentanglement in the template space of the LAE.

- The utility of weak supervision for the LAE is further demonstrated by utilising it for enforcing disentanglement of pose, expression, and identity in the LAE. We demonstrate qualitatively and quantitatively that adding the weak supervision signal also helps us recover better shapes.

- An application of weakly-supervised learning to medical imaging is also demonstrated. A multi-instance framework is evaluated in a mixture-of-experts model for the diagnosis of tumoral lymphocytosis, where the system is able to beat the average predictions of 12 biologists.

- The proposed method is compared against classical radiomics and attention-based methods and is shown to perform better than them, thus making a strong case for deployment in real-world.

With the contributions of the thesis summarised, a few possible future research directions based on this thesis are stated below.

## 5.1 Future Work

### 5.1.1 Morphable Model Learning

Morphable model generation, as summarised in Chapter 1 is a cumbersome process. It involves gathering subjects and acquiring images under controlled conditions of pose, illumination, expression, et cetera. However, morphable models also give us

extremely high resolution textures and realistic object shapes. In this thesis, we attempted to bridge the gap between the two by proposing a framework for unsupervised morphable model learning coupled with photo-realistic image formation using only weak supervision and no ground truth sparse or dense landmarks.

We learn from existing works and trends in 3D reconstruction [Blanz 2003b, Matthews 2004, Torresani 2008, Garg 2013, Güler 2017, Kanazawa 2018a, Booth 2018, Feng 2018] that correspondence plays a central role in capturing object shape. In Chapter 2, we have followed a similar approach by addressing the task of unsupervised dense alignment of object categories. As Chapter 3 builds on top of this, improving this unsupervised dense alignment can lead to much better model recovery. Indeed, as is shown in Figure 5.1, using dense correspondences obtained using DenseReg [Güler 2017] to train the LAE results in highly accurate 3D shape recovery.

Figure 5.1: Shape recovery with LAE using dense correspondences obtained through DenseReg. For each case, column 1 shows the input image, column 2 shows the mean shape, column 3 shows the recovered shape (i.e., mean shape + image-specific deformations), and column 4 shows the reconstructed image (using the correct pose parameters). For columns 2 and 3, the three rows show visualisations from different camera angles.

#### 5.1.1.1 Part-based models

Extending the unsupervised dense alignment approach to other, more complicated object categories is the first future direction for achieving fully-unsupervised model learning. Following the recent work DensePose [Güler 2018], which attempts dense regression of UV coordinates for the human body, and their follow-up HoloPose [Guler 2019], a part-based model can help facilitate the dense regression problem. This is also shown in [Kokkinos 2007] where a part-based model improves the learning of an active appearance model, while [Lorenz 2019] have shown unsupervised part-based disentanglement of shape and appearance. A important future direction in this area is investing the integration of object parts into DAEs.

#### 5.1.1.2 UV spaces

Recent unsupervised approaches have also demonstrated the power of learning efficient UV representations. Canonical surface mapping [Kulkarni 2019], a recent work, aims at learning canonical UV spaces for unseen objects using a template and ground truth segmentation masks. Their approach uses a cyclic consistency loss, much like the equivariance constraints demonstrated in Section1.1.1.3, in order to learn UV mappings between the image space and the template shape. Another recent approach [Wu 2020] exploits the symmetry of objects to learn canonical shapes of objects. Their approach does not need template shapes or ground truth masks and can discover 3D shapes from a bag of 2D images. While symmetry assumption at first thought seems rather strong, it is able to capture a surprisingly large category of objects, for example, cars, faces, cats.

### 5.1.2 Medical Imaging

With the rising interest in machine learning applications to medical imaging, there is lot of potential for active use of deep learning in medical analysis. We will note some possible future research directions related to this thesis here.

Cancer research is one of the largest domains with interest in machine learning [Litjens 2017]. With whole-slide image analysis becoming more and more studied [Wang 2017, Komura 2018], there is significant room for novel research in this domain. An important component of cancer research is disease response prediction. This is important as it can regular treatment costs substantially.

Specifically in the case of breast cancer, tumor-infiltrating lymphocytes (TILs) are shown to exhibit high correlation with achieving pathological complete response and survival without medical events [Salgado 2015]. Pathologists hence spend a considerable amount of time "scoring" the percentage of TILs from whole-slide images during diagnosis. Recent deep learning approaches that try to score TILs automatically are largely supervised approaches [Saltz 2018, Le 2019] and semi-supervised [Abousamra 2019]. [Hou 2019a] is a recently proposed generative model for histopathology images which can be used in place of pathologists' annotations for training supervised architecture for scoring TILs. An alternative to these approaches is using pathologists' scores of TILs as weak labels to learn significant regions in whole-slide images. This is indeed a very active area of research and a principal future research direction of the work presented in this thesis.

While TIL-count prediction is significant given diagnosis and medicine, unsupervised and weakly-supervised segmentation of lymphocytes can improve scoring of TILs. To this end, [Hou 2019b] were one of the first to propose a fully-unsupervised autoencoder for segmentation of nuclei in histopathology images. Their method is demontrated to be efficient via. experiments showing pre-trained unsupervised autoencoder indeed outperformed other nuclei segmentation models. They follow up with synthetic histopathology patch generation [Hou 2019a]. While these methods are shown to perform well on related classification tasks, robust, fully-unsupervised

segmentation of lymphocyte cells remains to be explored.

# Appendices

Chapter A

# Appendix: Additional Details and Results

## A.1 Deforming Autoencoders

In this section, we note additional results and architecture and implementation details of deforming autoencoders.

### A.1.1 Additional Results

#### A.1.1.1 Ablation Study: Dimension of $\mathbf{Z}_T$

In this section, we show experimental results on single deformed MNIST images of the digit 3 (Figure A.1) as well as in-the-wild faces (without masking) from the MAFL dataset (Figure A.2) to demonstrate the effect of varying the dimension of $\mathbf{Z}_T$.

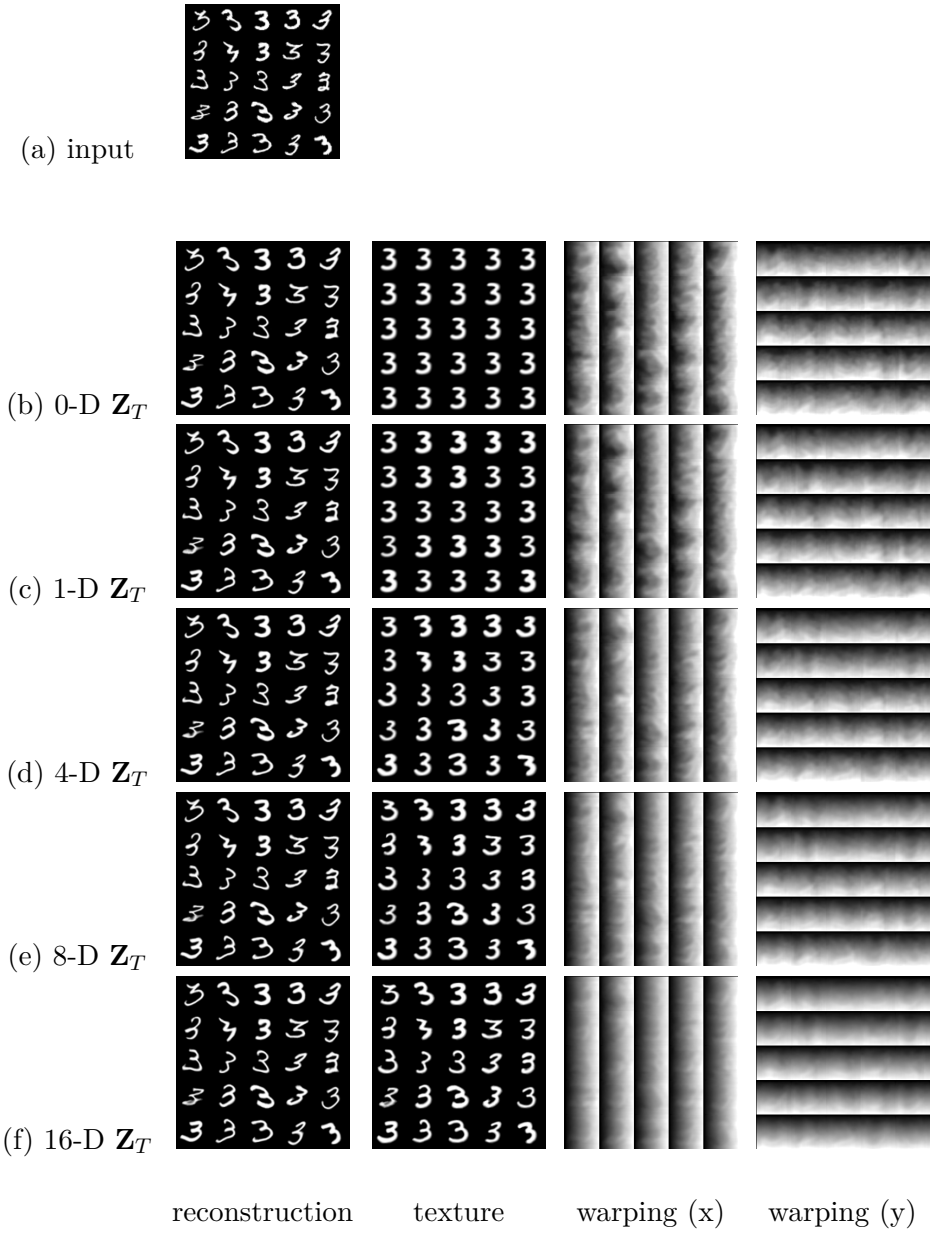



Figure A.1: Effect of varying the dimensionality of the latent vector for the texture encoding, $\mathbf{Z}_T$: The dimension of $\mathbf{Z}_T$ is 0 for (b), 1 for (c), 4 for (d), 8 for (e), 16 for (f). $Z_W$ is fixed to 128. When $\mathbf{Z}_T$ is 0-Dimensional, the texture decoder is forced to generate an identical texture for every image (b). When we increase the dimension of $\mathbf{Z}_T$ to 1, the texture decoder learns to align the pose (c) with varying stroke width. When further increasing the dimension of $\mathbf{Z}_T$, the network learns a more diverse texture map for each image (d, e, f).


Figure A.2: Effect of varying the dimensionality of the latent vector for the texture envoding, $\mathbf{Z}_T$, on the MAFL face dataset; $Z_W$ is fixed to 128. The problem is ill-posed and affords many solutions; if $\mathbf{Z}_T$ is set to be 0D dimension, the texture becomes a "bag of colored pixels" which, when deformed (at will) can reconstruct an image. Increasing the dimension of $\mathbf{Z}_T$ (4-32D) lets the network generates aligned texture maps and more exact appearance; further increasing $\mathbf{Z}_T$ (128-D) reduces the alignment effect.

### A.1.1.2 Methods for deformation modelling

In this section, we demonstrate the effect of using different warping modules.

We first show additional comparisons between using our proposed *affine + integral* warping and a non-rigid warping field directly output from a convolutional decoder for non-rigid deformation modelling (Figure A.3).

We visualize the utility of *affine* and *integral* warping modules in our network with face images (Figure A.4). We can see that the affine transformation handles global pose variance (Figure. A.4-(b)) but not local non-rigid deformation. Our proposed integral warping module aligns the faces in a non-rigid manner (Figure A.4-(c)). Incorporating both deformation modules improves the non-rigid alignment (Figure A.4-(d)).

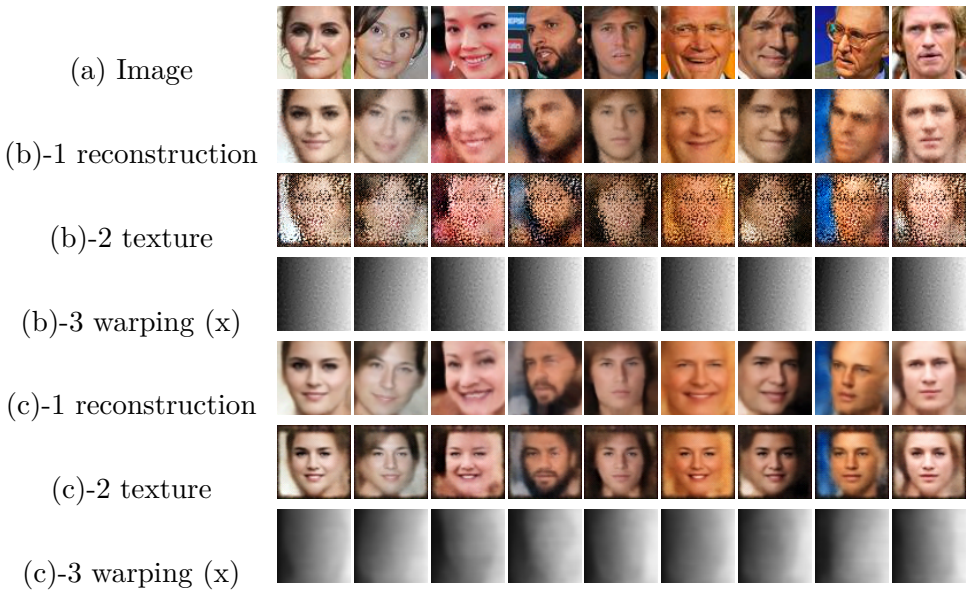



Figure A.3: Comparison between using our proposed *affine + integral* warping modules (c) and using a warping field directly predicted from a convolutional decoder (b) for non-rigid deformation modelling. Our non-rigid deformation modelling generates better reconstructions and visually plausible texture maps.

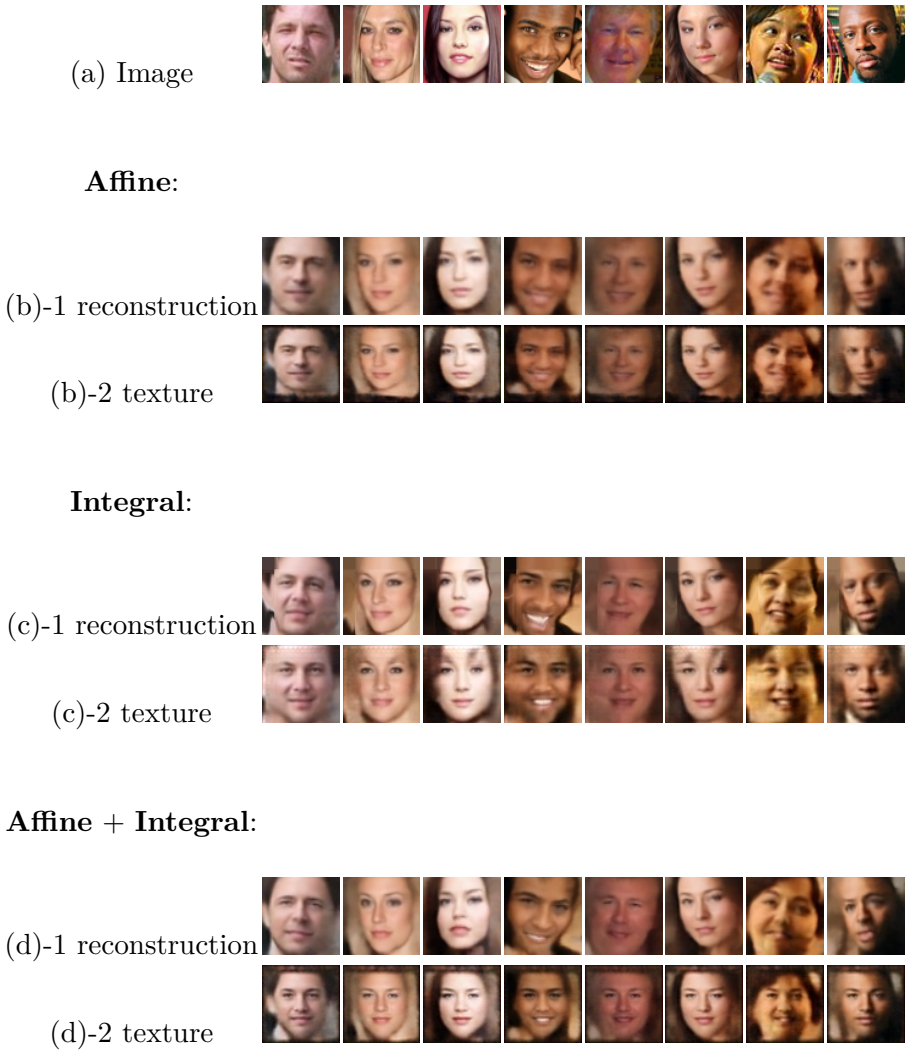



Figure A.4: Effect of affine and integral warping modules using in our network, using faces in-the-wild. The affine transformation can handle global pose variation, as shown in (b) but not local non-rigid deformation- eyes, noses, or other landmarks are not aligned in the decoded texture images. The proposed integral warping module aligns the faces in a non-rigid manner (c), but in an exaggerated manner, causing smears in the texture image, e.g. around eyebrows. Incorporating both deformation modules improves the non-rigid alignment (d). In this experiment, we set $Z_A = 32$, $\mathbf{Z}_T = 32$ and $Z_W = 32$.

### A.1.1.3 Latent Manifold Traversal

We provide additional results and comparisons with a plain autoencoder on traversing the learnt manifolds. In addition to Figure 13 in our manuscript, we provide two more sets of results in Figure A.5 and Figure A.6. Compared to a plain autoencoder, our deforming autoencoder not only generates better reconstructions, but also learns a better face manifold - interpolating between learnt latent representations generates sharper and more realistic face images. For this experiment, we use the convolutional encoder and decoder architecture as described in Sec. A.1.2.1.

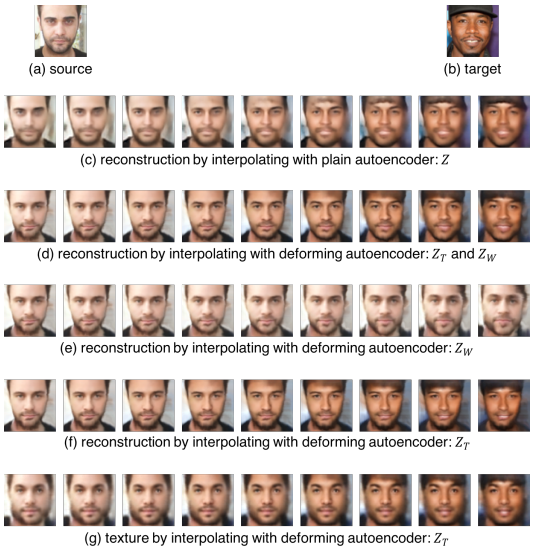



Figure A.5: Interpolating learnt representations using networks learnt on MAFL dataset. Deforming autoencoder learns better latent representations for face compared to a plain autoencoder. By interpolating the latent representations $\mathbf{Z}_T$ and/or $Z_W$, we observe smooth transition of pose, shape and skin texture. Interpolated results also stays on the face manifold and, generates more realistic image compared to a plain autoencoder.

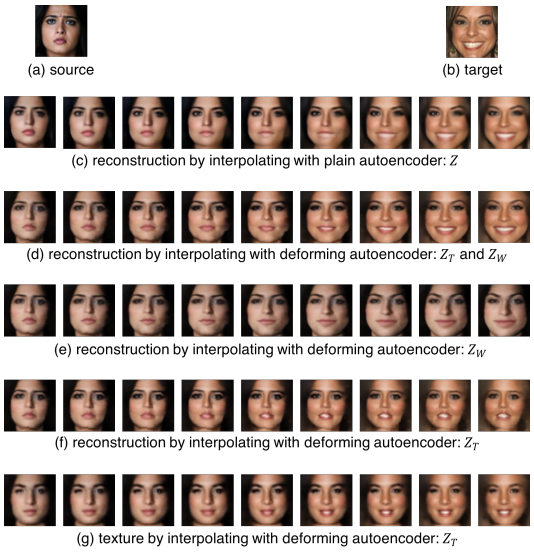



Figure A.6: Interpolating learnt representations using networks learnt on MAFL dataset. Deforming autoencoder learns better latent representations for face compared to a plain autoencoder. By interpolating the latent representations $\mathbf{Z}_T$ and/or $Z_W$, we observe smooth transition of pose, shape and skin texture. Interpolated results also stays on the face manifold and, generates more realistic image compared to a plain autoencoder.

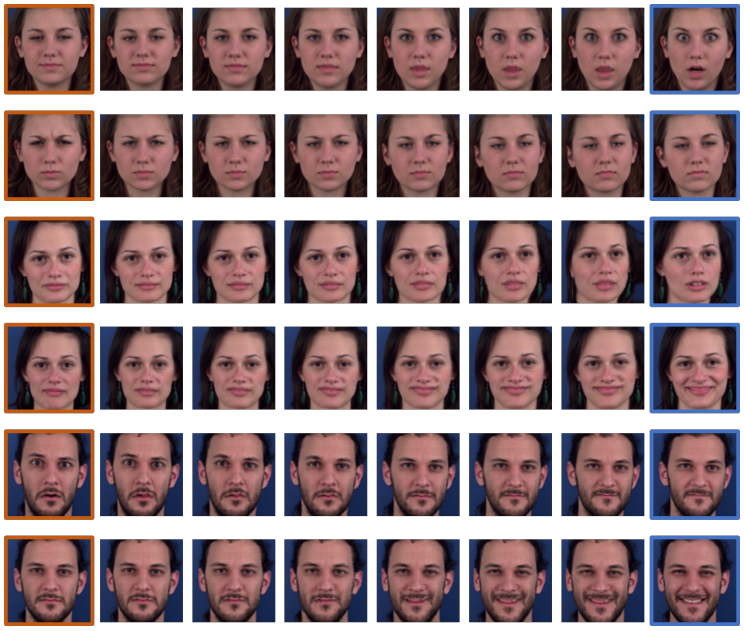


Figure A.7: Expression interpolation: Trained on the MUG facial expression dataset, our network is able to disentangle the facial expression deformation and encode this information in a meaningful latent representation. By interpolating the latent deformation representation from the source (in orange) to the target (in blue), our network generates sharp images and smooth deformation interpolation between expressions as shown in each row. In this experiment, ther model for each subject is independently trained, where we set dimension of $\mathbf{Z}_T$ to 0 (assuming single texture for each subject) and dimension of $Z_W$ to 128.

#### A.1.1.4 Intrinsic Decomposition with DAE

In Fig. A.8 we provide additional results of unsupervised intrinsic disentangling for faces-in-the-wild using Intrinsic-DAE. Using the architecture and objective functions described in Sec. 2.3 of the main paper the network learns to bring faces under different poses and illumination conditions, shown in Fig. A.8-(a), to a canonical view, as shown in Fig. A.8-(d), while separating the shading, shown in Fig. A.8-(b) and albedo, shown in Fig. A.8-(c) components in the canonical view using two independent decoders. With the learnt deformation from the deformation decoder, we can warp the aligned shading and aligned albedo to its original view as in the input image, as shown in Fig. A.8-(e,f).

In Fig. A.9, we provide additional results for "changing lighting direction" of a face image using Intinsic-DAE. We show that even without explicitly modelling of geometry, we can simulate smooth and reasonable lighting direction changes in the image by interpolating the learnt latent representation for shading, as shown in Fig. A.9-a-(4),b-(4).

For Intrinsic-DAE, we use the DenseNet architecture as the encoders and decoders (A.1.2.2). The network is trained with a subset of $200,000$ images in the CelebA dataset. The dimensions of latent representations are: 16 for albedo, 16 for shading, and 128 for deformation field.

### A.1.2 Architectural Details

#### A.1.2.1 Convolutional Encoders and Decoders

In our experiments, where input images are of size $64 \times 64 \times \text{Nc}$ (Nc is 1 for MNIST and 3 for faces), we use identical architectures for convolutional encoders and decoders.

The encoder architecture is

```
Conv(32)-LeakyReLU-Conv(64)-BN-LeakyReLU-Conv(128)->
 ->BN-LeakyReLU-Conv(256)-BN-LeakyReLU-Conv(Nz)->
 ->Sigmoid;
```

while the decoder architecture is

```
ConvT(256)-BN-ReLU-ConvT(128)-BN-ReLU-ConvT(64)->
 ->BN-ReLU-ConvT(32)-BN-ReLU-ConvT(32)-BN-ReLU-ConvT(Nc)->
 ->Threshold(0,1),
```

where

- `Conv(n)`: convolution layer with $n$ output feature map;
- `ConvT(n)`: transposed convolution (deconvolution) layer with $n$ output feature map;

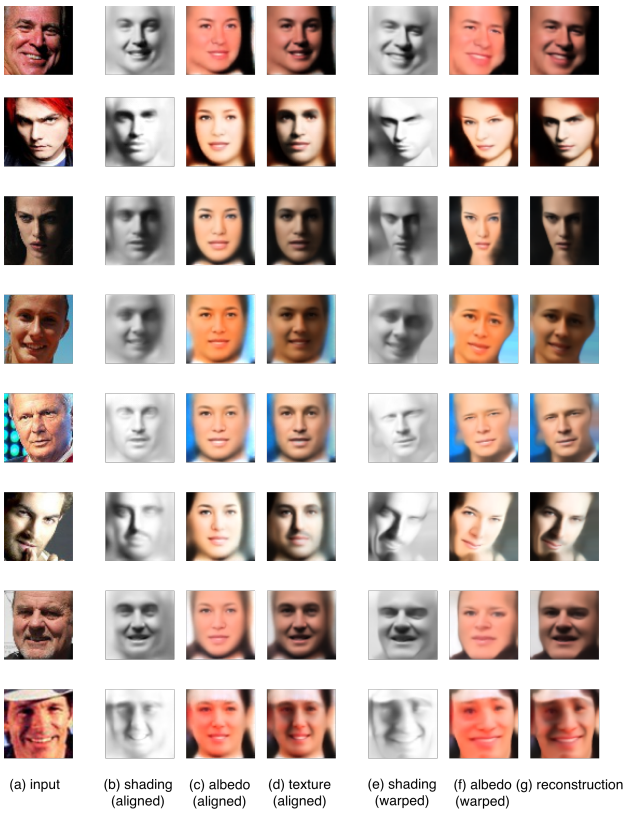



Figure A.8: Unsupervised intrinsic decomposition of faces-in-the-wild using Intrinsic-DAE: The network learns to bring faces under different poses and illumination conditions (a) to a canonical view (d), and further separate the shading (b) and albedo (c) component in the canonical view using two independent decoders. With the learnt deformation from the deformation decoder we can warp the aligned shading and aligned albedo to its original view as in the input image (e,f).

- `BN`: batch normalization layer
- `Nz`: latent representation dimension
- `Nc`: number of output image channel

#### A.1.2.2 DenseNet-style Encoders and Decoders

For DenseNet-style architectures, we employ dense convolutional connections. The architecture for the encoder is

```
BN-ReLU-Conv(32)-DBE(32,6)-TBE(32,64,2)->
 ->DBE(64,12)-TBE(64,128,2)-DBE(128,24)-TBE(128,256,2)->
 ->DBE(256,16)-TBE(256,Nz,4)-Sigmoid;
```

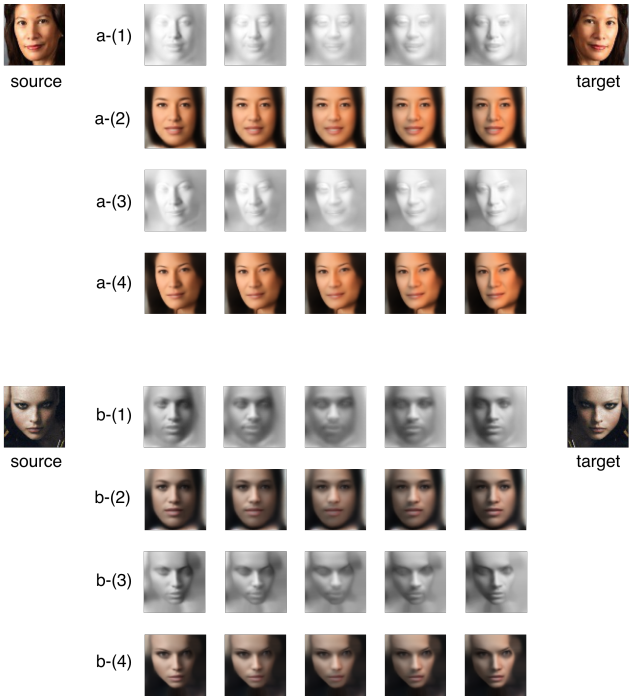



Figure A.9: Lighting manipulation by interpolating latent representation of shading: Intrinsic-DAE allows us to disentangle a latent representation for shading for a given face image in an unsupervised manner. Therefore, manipulating the shading component will result in lighting effects in the output images. In this experiment, we interpolate the latent representation of shading from source to target, which is the mirror of the source with reversed lighting direction. In the result, we can observe that, even without explicitly modelling geometry in our network, we can simulate smooth lighting direction change in both the shading (a-(3), b-(3)) and the final reconstruction (a-(4), b-(4)).

whereas the architecture for the decoder is

```
BN-Tanh-ConvT(256)-DBD(256,16)-TBD(256,128)->
 ->DBD(128,24)-TBD(128,64)-DBD(64,12)-TBD(64,32)->
 ->DBD(32,6)-TBD(32,32)-BN-Tanh-ConvT(Nc)-Threshold(0,1),
```

where

- `DBE(n,k)`: A dense encoder block with $k$ $3 \times 3$ convolutions with $n$ channels.
- `TBE(m,n,p)`: An encoder transition block of $1 \times 1$ convolutions with $m$ input channels and $n$ output channels. Also includes a max-pooling operation of size $p$.
- `DBD(n,k)`: A dense decoder block with $k$ $3 \times 3$ transposed convolution operations with $n$ channels.

- `TBD(m,n)`: A decoder transition block of $4 \times 4$ convolutions, stride of 2 and padding of 1. It has $m$ input channels, and $n$ output channels.

We describe the tensor sizes for intermediate convolution operations in Tables A.1 and A.2.

| Conv Encoder | | Conv Decoder | |
|---|---|---|---|
| Output Size | Operation | Output size | Operation |
| $32 \times 32 \times 32$ | $4 \times 4$ Conv(32) | $4 \times 4 \times 256$ | $4 \times 4$ ConvT(256) |
| $16 \times 16 \times 64$ | $4 \times 4$ Conv(64) | $4 \times 4 \times 128$ | $4 \times 4$ ConvT(128) |
| $8 \times 8 \times 128$ | $4 \times 4$ Conv(128) | $4 \times 4 \times 64$ | $4 \times 4$ ConvT(64) |
| $4 \times 4 \times 256$ | $4 \times 4$ Conv(256) | $4 \times 4 \times 32$ | $4 \times 4$ ConvT(32) |
| `Nz` | $4 \times 4$ Conv(`Nz`) | $4 \times 4 \times 32$ | $4 \times 4$ ConvT(32) |
| | | $4 \times 4 \times$`Nc` | $4 \times 4$ ConvT(`Nc`) |

Table A.1: Tensor sizes for intermediate convolutional operations in the convolutional encoder and decoder architectures. The output shape denoted $h \times w \times C$, where $h$ and $w$ are height and width of the feature maps, respectively, and $C$ is the number of channels.

| Dense Conv Encoder | | Dense Conv Decoder | |
|---|---|---|---|
| Output Size | Operation | Output size | Operation |
| $32 \times 32 \times 32$ | $4 \times 4$ Conv(32) | $4 \times 4 \times 256$ | $4 \times 4$ ConvT(256) |
| $32 \times 32 \times 32$ | `DBE(32,6)` | $4 \times 4 \times 256$ | `DBD(256,16)` |
| $16 \times 16 \times 64$ | `TBE(32,64,2)` | $8 \times 8 \times 128$ | `TBD(256,128)` |
| $16 \times 16 \times 64$ | `DBE(64,12)` | $8 \times 8 \times 128$ | `DBD(128,24)` |
| $8 \times 8 \times 128$ | `TBE(64,128,2)` | $16 \times 16 \times 64$ | `TBD(128,64)` |
| $8 \times 8 \times 128$ | `DBE(128,24)` | $16 \times 16 \times 64$ | `DBD(64,12)` |
| $4 \times 4 \times 256$ | `TBE(128,256,2)` | $32 \times 32 \times 32$ | `TBD(64,32)` |
| $4 \times 4 \times 256$ | `DBE(256,16)` | $32 \times 32 \times 32$ | `DBD(32,6)` |
| `Nz` | `TBE(256,Nz,4)` | $64 \times 64 \times 32$ | `TBD(32,32)` |
| | | $64 \times 64 \times$`Nc` | $3 \times 3$ ConvT(`Nc`) |

Table A.2: Tensor sizes for intermediate convolutional operations in the dense encoder and decoder architectures. The output shape denoted $h \times w \times C$, where $h$ and $w$ are height and width of the feature maps, respectively, and $C$ is the number of channels.

## A.2 Lifting Autoencoders

In this section, we note some additional implementation details for lifting autoencoders.

## A.2.1 Additional Details

### A.2.1.1 Data Processing

In our experiments, we used images of size $128 \times 128 \times 3$ pixels, which were cropped from the CelebA and MultiPIE datasets using ground-truth bounding boxes.

For CelebA images, the cropping was performed by extracting a square patch around the face with side-length equal to the length of the longer side of the bounding box. It was then adjusted so that it lies entirely inside the image (by translating it horizontally or vertically, or even scaling it down if necessary). Finally, we tightened the resulting box by 12 pixels from each side as the bounding boxes are quite loose crops, and resized the resulting square image to $128 \times 128$. We use all images from CelebA for training (about $200,000$ images) except the MAFL test set which is contained entirely in CelebA (1000 images).

For MultiPIE dataset, we crop the face images according to landmarks positions on the eyes, the corner of mouth, and the width of the frontal face. Specifically, we use the mean coordinates of the 4 landmarks as the center of the crop, and use $1.4\times$ the width of the face as the width of the images. We use the method proposed in [Bulat 2017b] to detect the landmarks. For each person, the crop is identical across all illumination condition for the same camera.

### A.2.1.2 Architecture Details

We used convolutional encoders and decoders similar to the ones described in [Shu 2018]. We detail the architectures here again for completeness. The convolutional encoder architecture is—

```
C onv(32)-LeakyReLU-Conv(64)->
   ->BN-LeakyReLU-Conv(128)->
   ->BN-LeakyReLU-Conv(256)->
   ->BN-LeakyReLU-Conv(256)->
   ->BN-LeakyReLU-Conv(Nz)->
   ->Sigmoid;
```

while the convolutional decoder architecture is— beginverbatim C onvT(512)-BN-ReLU-ConvT(256)-> ->BN-ReLU-ConvT(128)-> ->BN-ReLU-ConvT(64)-> ->BN-ReLU-ConvT(32)-> ->BN-ReLU-ConvT(32)-> ->BN-ReLU-ConvT(Nc)-> ->Threshold(0,1). endverbatim

### A.2.1.3 Refinement Networks

The refinement set-up consists of a *generator* network, and a *discriminator* network. The generator is a standard UNet [Ronneberger 2015] for $128 \times 128$ images that are downsampled to $1 \times 1$ in the innermost latent layer.

The discrminiator is a PatchGAN discrminator [Isola 2016] with the following architecture—

```
C onv(64)-LeakyReLU-Conv(128)-BN->
   ->LeakyReLU-Conv(256)-BN->
   ->LeakyReLU-Conv(512)-BN->
   ->LeakyReLU->Conv(1)
```

In all these descriptions, `Conv(x)` signifies a 2D convolution layer with `x` channels, a kernel size of $4 \times 4$, a stride of 2, and a padding of 1. Similarly for `ConvT(x)`, except that it signifies a deconv layer.

#### A.2.1.4 Implementation Details

We implemented our system in Python 3.6 using the PyTorch library. We use convolutional, activation, and batch norm layers predefined in the `torch.nn` module, and take advantage of the Autograd [Paszke 2017] framework to take care of the gradients required by backpropagation.

#### A.2.1.5 Rotation Modeling

Modelling rotations using quaternions has several advantages over modelling them using Euler angles, including computational ease, less ambiguity, and compact representation [Dam 1998]. Quaternions were also employed by [Kanazawa 2018b] to model mesh rotations. Following these works, we also use quaternions in our framework to model rotations, by regressing them from the camera latent space, and normalizing them to unit length.

#### A.2.1.6 The Neural Mesh Renderer

The Neural Mesh Renderer [Kato 2018b] is a recently proposed module that can be inserted into a neural network to enable end-to-end training with a rendering operation. The renderer proposes approximate gradients to learn texture and shape given the output rendering. The original module was released in Chainer [Kato 2018a], but we use a PyTorch port of this module, which is a publicly-available re-implementation [nr2 2018]. The renderer in our framework accepts a texture image, the mean shape, the deviation from the mean shape, and the camera parameters to output a 2D reconstruction of the original image.

#### A.2.1.7 Training Procedure

To train the LAE, we first train a DAE on the training data. We then fix the DAE and use it to extract dense correspondences between the image space and the canonical space. These correspondences are used in the objective of the 3D reprojection loss (Equation 3.7 and Equation 3.8).

To obtain image-specific camera, translation, and shape estimates, we train another convolutional encoder. This encoder learns a disentangled latent space where the shape estimates and camera and translation estimates are encoded by

different vectors. For the MultiPIE experiments, the shape latent vector is further divided into identity and expression vectors. We use linear layers to regress camera, translation, and shape estimates from their latent encodings.

We train our system using the Adam [Kingma 2014a] optimiser for all learnable parameters. We start with a learning rate of 0.0001, which is decayed every 50 training epochs by a factor of 0.5. We train for a total of 400 epochs.

Chapter B

# Appendix: Synthèse de la Thèse

Les données sont indispensables à tout système d'apprentissage automatique. Certains domaines spécialisés requièrent des données labellisées qui peuvent parfois être difficiles à obtenir et onéreuses. Par exemple, recueillir des données médicales demande du temps et des efforts importants aux docteurs et aux biologistes. Dans cette thèse, nous proposons des méthodes d'apprentissage sur données non-labellisées et faiblement labellisées pour deux problématiques en vision par ordinateur et imagerie médicale. Notre premier sujet d'étude, en vision par ordinateur, concerne l'apprentissage faiblement supervisé d'un modèle déformable. Dans un second temps, nous traitons du diagnostic automatique d'hyperlymphocytosis en utilisant une base de donnée faiblement labellisées.

La première contribution de cette thèse (Chapitre 2) est un modèle appelé Auto-Encodeur Déformable (DAE), utilisé pour l'apprentissage non-supervisé de l'alignement 2-D dense d'images d'une classe donnée. Ce modèle est capable d'identifier un espace canonique pour cette classe d'objets à partir des images non-alignées et non-labellisées. Nous proposons d'utiliser un auto-encodeur équipé d'un moyen de séparation de deux caractéristiques importantes dans son espace latent, à savoir l'apparence des objets dans l'espace canonique et la déformation dense associée permettant de retrouver l'image réelle à partir de cette apparence. Une nouvelle technique pour prédire la déformation à partir d'un vector latent est également proposée, qui nous permet de trouver la déformation la plus significative. Nous constatons que cette façon de prédire la déformation mène à de meilleurs résultats que celle qui prédit le changement en position de chaque pixel. L'évaluation de l'alignement dans l'espace canonique est faite par une méthode de localisation de repères trouvés sur l'objet, et correspond à l'erreur moyenne normalisée suivant cette métrique. Nous montrons que les résultats obtenus surpassent l'état de l'art. En outre, le système proposé permet une meilleurs séparation, en albédo et ombre,

une tâche rendue plus aisée dans l'espace canonique grâce à l'alignement. Pour conclure, nous illustrons l'application de cette méthode à d'autres domaines, à savoir, l'alignement d'IRM de poumons et d'images satellites.

Dans le Chapitre 3, nous étendons le modèle et la méthode d'alignement 2-D dense non-supervisé proposés dans le Chapitre 2 au cas de la 3-D afin d'apprendre un modèle tridimensionnel pour les visages humains. Dans le cas où la sortie du DAE est considérée comme la vérité terrain, nous montrons qu'il est possible d'obtenir un modèle 3-D à partir d'une méthode basée sur *la structure non-rigide du mouvement*. Nous utilisons un auto-encodeur basé sur le même principe que le DAE pour apprendre un espace latent qui est utilisé pour prédire des paramètres spécifiques pour la modélisation d'une image réelle. Ces paramètres incluent les coordonnées et le point de vue de la caméra, ainsi que la déformation d'un modèle moyen associé à cette image. Ces paramètres et le modèle moyen déterminent le maillage d'une surface dans un environmment 3-D représentant la forme du visage. Le modèle moyen est appris en même temps que ces paramètres et n'est pas initialisé par un modèle de déformation 3-D. Nous proposons également un modèle plus approfondi utilisant de la supervision faible pour séparer plusieurs types de déformation, à savoir l'angle de vue de la caméra, le modèle 3-D propre au visage de la personne considérée, et son expression faciale. Le modèle final est évalué de la même manière que le DAE, c'est-à-dire par l'utilisation de repères faciaux, cependant, dans le cas de ce modèle les repères sont localisés en 3-D. L'évaluation de la forme est aussi faite par une analyse procrustéenne sur la base de données AFLW2000-3D, en alignant la forme estimée par notre modèle et la vérité terrain.

Dans le Chapitre 4, nous étudions l'apprentissage profond faiblement supervisé appliqué au domaine d'imagerie médicale. Plus précisément, nous nous concentrons sur une problématique liée au cancer du sang. Une partie importante du diagnostique d'un patient présentant un symptôme d'hyperlymphocytosis—un nombre absolu de lymphocyte très élevé, au-dessus de $4 \times 10^9$ par litre—est de déterminer la cause de ce dérèglement. Ce symptôme se manifeste soit en réaction à une infection (le syndrôme est alors qualifié de réactif), soit en raison d'un cancer (syndrôme tumoral). On note une certaine variabilit'e du résultat concernant le diagnostique d'un même patient réalisé par plusieurs biologistes, ainsi qu'au sein d'un ensemble de diagnostiques réalisés par un même praticien. Il nous faut donc un moyen de surmonter cette variabilité en gardant un bon niveau de la performance en prédiction. La vérité terrain, c'est-à-dire le vrai diagnostique, n'étant obtenue qu'après des tests complémentaires, cette problématique appartient à la catégorie de l'apprentissage faiblement supervisé, le but étant de prédire la nature de la maladie. Pour cela nous proposons un modèle convolutif pour encoder les images appartenant à un patient, suivi par la concaténation de l'information contenue dans toutes les images, le résultat de cette opération est classifié soit en tumoral, soit en réactif. Nous proposons également un modèle *mixture-of-experts* pour accumuler l'information venant de différents attributs des patients, à savoir l'âge et le nombre absolu de lymphocytes. Nous démontrons également par un teste de reproductibilité que le modèle *mixture-of-experts* est plus reliable que l'autre qui n'utilise que les images pour la prédiction.

Nos résultats montrent que les modèles proposés sont de performances comparables à celles des biologistes, et peuvent donc les aider dans l'élaboration de leur diagnostique.

# Bibliography

[Abousamra 2019] Shahira Abousamra, Le Hou, Rajarsi Gupta, Chao Chen, Dimitris Samaras, Tahsin Kurc, Rebecca Batiste, Tianhao Zhao, Shroyer Kenneth and Joel Saltz. *Learning from Thresholds: Fully Automated Classification of Tumor Infiltrating Lymphocytes for Multiple Cancer Types*, 2019. (Cited on page 98.)

[Achanta 2012] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk. *SLIC Superpixels Compared to State-of-the-Art Superpixel Methods*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, 2012. (Cited on page 86.)

[Afifi 2017] Mahmoud Afifi. *Gender recognition and biometric identification using a large dataset of hand images*. CoRR, vol. abs/1711.04322, 2017. (Cited on page 38.)

[Agarwal 2009] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz and Richard Szeliski. *Building Rome in a day*. In IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009, pages 72–79, 2009. (Cited on page 55.)

[Aifanti 2010] Niki Aifanti, Christos Papachristou and Anastasios Delopoulos. *The MUG Facial Expression Database*. In 11th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2010, Desenzano del Garda, Italy, April 12-14, 2010. IEEE, 2010. (Cited on page 36.)

[Akhter 2009] Ijaz Akhter, Yaser Sheikh, Sohaib Khan and Takeo Kanade. *Nonrigid structure from motion in trajectory space*. In Advances in neural information processing systems, pages 41–48, 2009. (Cited on pages 19, 55 and 59.)

[Alansary 2016] Amir Alansary, Konstantinos Kamnitsas, Alice Davidson, Rostislav Khlebnikov, Martin Rajchl, Christina Malamateniou, Mary Rutherford, Joseph V Hajnal, Ben Glocker, Daniel Rueckert*et al.* *Fast fully automatic segmentation of the human placenta from motion corrupted MRI.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 589–597. Springer, 2016. (Cited on page 22.)

[Alhaija 2018] Hassan Abu Alhaija, Siva Karthik Mustikovela, Andreas Geiger and Carsten Rother. *Geometric Image Synthesis.* CoRR, vol. abs/1809.04696, 2018. (Cited on page 54.)

[Allen 2003] Brett Allen, Brian Curless, Brian Curless and Zoran Popović. *The space of human body shapes: reconstruction and parameterization from range scans.* In ACM transactions on graphics (TOG), volume 22, pages 587–594. ACM, 2003. (Cited on page 17.)

[Amberg 2007] Brian Amberg, Sami Romdhani and Thomas Vetter. *Optimal step nonrigid ICP algorithms for surface registration.* In 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2007. (Cited on page 17.)

[Amit 1991] Yali Amit, Ulf Grenander and Mauro Piccioni. *Structural image restoration through deformable templates.* Journal of the American Statistical Association, vol. 86, no. 414, 1991. (Cited on page 29.)

[Amores 2013] Jaume Amores. *Multiple instance classification: Review, taxonomy and comparative study.* Artificial Intelligence, vol. 201, pages 81 – 105, 2013. (Cited on page 77.)

[Andrews 2003] Stuart Andrews, Ioannis Tsochantaridis and Thomas Hofmann. *Support vector machines for multiple-instance learning.* In Advances in neural information processing systems, pages 577–584, 2003. (Cited on page 87.)

[Antipov 2017] Grigory Antipov, Moez Baccouche and Jean-Luc Dugelay. *Face aging with conditional generative adversarial networks.* In 2017 IEEE International Conference on Image Processing (ICIP), pages 2089–2093. IEEE, 2017. (Cited on page 8.)

[Avants 2008] B.B. Avants, C.L. Epstein, M. Grossman and J.C. Gee. *Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain.* Medical Image Analysis, 2008. (Cited on page 47.)

[Badrinarayanan 2017] Vijay Badrinarayanan, Alex Kendall and Roberto Cipolla. *Segnet: A deep convolutional encoder-decoder architecture for image segmentation.* IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pages 2481–2495, 2017. (Cited on page 21.)

[Baker 2004] Simon Baker, Iain Matthews and Jeff Schneider. *Automatic construction of active appearance models as an image coding problem.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 10, pages 1380–1384, 2004. (Cited on page 15.)

[Barber 2003] David Barber and Felix V. Agakov. *The IM Algorithm: A Variational Approach to Information Maximization.* In NIPS, 2003. (Cited on page 8.)

[Barron 2013] Jonathan Barron and Jitendra Malik. *Shape, Illumination, and Reflectance from Shading.* Rapport technique UCB/EECS-2013-117, EECS Department, University of California, Berkeley, May 2013. (Cited on page 54.)

[Barrow 1978] Harry Barrow, J Tenenbaum, A Hanson and E Riseman. *Recovering intrinsic scene characteristics.* Comput. Vis. Syst, vol. 2, pages 3–26, 1978. (Cited on pages 54 and 60.)

[Bengio 2007] Yoshua Bengio, Pascal Lamblin, Dan Popovici and Hugo Larochelle. *Greedy layer-wise training of deep networks.* In Advances in neural information processing systems, pages 153–160, 2007. (Cited on page 5.)

[Biggs 2018] Benjamin Biggs, Thomas Roddick, Andrew W. Fitzgibbon and Roberto Cipolla. *Creatures great and SMAL: Recovering the shape and motion of animals from video.* CoRR, vol. abs/1811.05804, 2018. (Cited on page 56.)

[Bilen 2016] Hakan Bilen and Andrea Vedaldi. *Weakly supervised deep detection networks.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2846–2854, 2016. (Cited on page 10.)

[Blanz 1999] Volker Blanz, Thomas Vetter*et al. A morphable model for the synthesis of 3D faces.* In Siggraph, volume 99, pages 187–194, 1999. (Cited on pages 4, 16, 17 and 24.)

[Blanz 2003a] Volker Blanz, Curzio Basso, Tomaso Poggio and Thomas Vetter. *Reanimating faces in images and video.* In Computer graphics forum, volume 22, pages 641–650. Wiley Online Library, 2003. (Cited on page 17.)

[Blanz 2003b] Volker Blanz and Thomas Vetter. *Face Recognition Based on Fitting a 3D Morphable Model.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 9, pages 1063–1074, 2003. (Cited on pages 15, 17, 29 and 97.)

[Booth 2018] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway and Stefanos Zafeiriou. *Large Scale 3D Morphable Models.* International Journal of Computer Vision, 2018. (Cited on pages 4, 58 and 97.)

[Bregler 2000] Christoph Bregler, Aaron Hertzmann and Henning Biermann. *Recovering non-rigid 3D shape from image streams.* In cvpr, volume 2, page 2690. Citeseer, 2000. (Cited on pages 18, 55 and 59.)

[Bristow 2015] Hilton Bristow, Jack Valmadre and Simon Lucey. *Dense Semantic Correspondence Where Every Pixel is a Classifier*. In ICCV, 2015. (Cited on page 31.)

[Brock 2019] Andrew Brock, Jeff Donahue and Karen Simonyan. *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. In International Conference on Learning Representations, 2019. (Cited on page 71.)

[Bulat 2017a] Adrian Bulat. *3D-FAN and LS3D-W Github repository*. https://github.com/1adrianb/face-alignment, 2017. (Cited on page 18.)

[Bulat 2017b] Adrian Bulat and Georgios 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. (Cited on pages 18, 72 and 115.)

[Caicedo 2009] Juan C. Caicedo, Angel Cruz and Fabio A. Gonzalez. *Histopathology Image Classification Using Bag of Features and Kernel Functions*. In Carlo Combi, Yuval Shahar and Ameen Abu-Hanna, editeurs, Artificial Intelligence in Medicine, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. (Cited on page 78.)

[Cao 2013] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong and Kun Zhou. *Facewarehouse: A 3d facial expression database for visual computing*. IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 3, pages 413–425, 2013. (Cited on page 17.)

[Carreira-Perpinan 2005] Miguel A Carreira-Perpinan and Geoffrey E Hinton. *On contrastive divergence learning*. In Aistats, volume 10, pages 33–40. Citeseer, 2005. (Cited on page 5.)

[Carreira 2016] Joao Carreira, Sara Vicente, Lourdes Agapito and Jorge Batista. *Lifting object detection datasets into 3d*. IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 7, pages 1342–1355, 2016. (Cited on pages 20, 54 and 55.)

[Cataldo 2017] Santa Di Cataldo and Elisa Ficarra. *Mining textural knowledge in biological images: Applications, methods and trends*. Computational and Structural Biotechnology Journal, vol. 15, 2017. (Cited on page 76.)

[Chandra 2013] Siddhartha Chandra, Shailesh Kumar and CV Jawahar. *Learning multiple non-linear sub-spaces using k-rbms*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2778–2785, 2013. (Cited on page 5.)

[Chaudhry 2017] Arslan Chaudhry, Puneet K Dokania and Philip HS Torr. *Discovering class-specific pixels for weakly-supervised semantic segmentation*. arXiv preprint arXiv:1707.05821, 2017. (Cited on page 10.)

[Chen 2006] Yixin Chen, Jinbo Bi and James Ze Wang. *MILES: Multiple-instance learning via embedded instance selection.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 12, pages 1931–1947, 2006. (Cited on page 87.)

[Chen 2014] Ting Chen and Christophe Chef d Hotel. *Deep learning based automatic immune cell detection for immunohistochemistry images.* In International workshop on machine learning in medical imaging, pages 17–24. Springer, 2014. (Cited on page 21.)

[Chen 2016a] Jianxu Chen and Chukka Srinivas. *Automatic lymphocyte detection in H&E images with deep neural networks.* arXiv preprint arXiv:1612.03217, 2016. (Cited on page 21.)

[Chen 2016b] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever and Pieter Abbeel. *InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets.* In Neural Information Processing Systems, 2016. (Cited on pages 8, 17, 30 and 54.)

[Christ 2017] Patrick Ferdinand Christ, Florian Ettlinger, Felix Grün, Mohamed Ezzeldin A Elshaera, Jana Lipkova, Sebastian Schlecht, Freba Ahmaddy, Sunil Tatavarty, Marc Bickel, Patrick Bilic*et al. Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks.* arXiv preprint arXiv:1702.05970, 2017. (Cited on page 21.)

[Çiçek 2016] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox and Olaf Ronneberger. *3D U-Net: learning dense volumetric segmentation from sparse annotation.* In International conference on medical image computing and computer-assisted intervention, pages 424–432. Springer, 2016. (Cited on page 21.)

[Cinbis 2016] Ramazan Gokberk Cinbis, Jakob Verbeek and Cordelia Schmid. *Weakly supervised object localization with multi-fold multiple instance learning.* IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 1, pages 189–203, 2016. (Cited on page 10.)

[Coates 2011] Adam Coates, Andrew Ng and Honglak Lee. *An analysis of single-layer networks in unsupervised feature learning.* In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 215–223, 2011. (Cited on page 5.)

[Community 2019] 3DMM Community. A curated list of 3d morphable models. https://github.com/3d-morphable-models/curated-list-of-awesome-3D-Morphable-Model-software-and-data, 2019. (Cited on page 18.)

[Cootes 1992] Timothy F Cootes, Christopher J Taylor, David H Cooper and Jim Graham. *Training models of shape from sets of examples.* In BMVC92, pages 9–18. Springer, 1992. (Cited on page 13.)

[Cootes 1995] Timothy F Cootes, Christopher J Taylor, David H Cooper and Jim Graham. *Active shape models-their training and application.* Computer vision and image understanding, vol. 61, no. 1, pages 38–59, 1995. (Cited on page 14.)

[Cootes 1998] Timothy F Cootes, Gareth J Edwards and Christopher J Taylor. *Active appearance models.* In European conference on computer vision. Springer, 1998. (Cited on pages 14, 15, 29, 31 and 32.)

[Cootes 2004] Timothy F Cootes, Cristopher J Taylor*et al.* *Statistical models of appearance for computer vision*, 2004. (Cited on page 16.)

[Dai 2014] Yuchao Dai, Hongdong Li and Mingyi He. *A simple prior-free method for non-rigid structure-from-motion factorization.* International Journal of Computer Vision, vol. 107, no. 2, pages 101–122, 2014. (Cited on pages 19, 55 and 59.)

[Dai 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu and Yichen Wei. *Deformable Convolutional Networks.* In ICCV, 2017. (Cited on pages 20 and 31.)

[Dam 1998] Erik B. Dam, Martin Koch and Martin Lillholm. *Quaternions, interpolation and animation.* Rapport technique, University of Copenhagen, 1998. (Cited on page 116.)

[Dawn 2010] Suma Dawn, Vikas Saxena and Bhudev Sharma. *Remote sensing image registration techniques: A survey.* In International Conference on Image and Signal Processing, pages 103–112. Springer, 2010. (Cited on page 47.)

[Denil 2014] Misha Denil, Alban Demiraj, Nal Kalchbrenner, Phil Blunsom and Nando de Freitas. *Modelling, Visualising and Summarising Documents with a Single Convolutional Neural Network.* Rapport technique arXiv:1406.3830, University of Oxford, 2014. (Cited on page 80.)

[Desjardins 2008] Guillaume Desjardins and Yoshua Bengio. *Empirical evaluation of convolutional RBMs for vision.* DIRO, Université de Montréal, pages 1–13, 2008. (Cited on page 6.)

[Dietterich 1997] T. G. Dietterich, R. H. Lathrop and T. Lozano-PÃ©rez. *Solving the multiple instance problem with axis-parallel rectangles.* Artificial Intelligence, vol. 89, no. 1, 1997. (Cited on pages 10, 77 and 78.)

[Donner 2006] Rene Donner, Michael Reiter, Georg Langs, Philipp Peloschek and Horst Bischof. *Fast active appearance model search using canonical correlation analysis.* IEEE transactions on pattern analysis and machine intelligence, vol. 28, no. 10, pages 1690–1694, 2006. (Cited on page 16.)

[Dundar 2011] Murat Dundar, Sunil S Badve, Gökhan Bilgin, Vikas C. Raykar, Rohit K. Jain, Olcay Sertel and Metin Nafi Gürcan. *Computerized classification of intraductal breast lesions using histopathological images.* IEEE Transactions on Biomedical Engineering, vol. 58, pages 1977–1984, 2011. (Cited on page 78.)

[Dürer 1534] Albrecht Dürer. Four books on human proportion. Hieronymus Andreae, called Formschneyder, 1534. (Cited on page 12.)

[Edwards 1998] Gareth J Edwards, Christopher J Taylor and Timothy F Cootes. *Interpreting face images using active appearance models.* In Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pages 300–305. IEEE, 1998. (Cited on page 16.)

[Egger 2019] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz and Thomas Vetter. *3D Morphable Face Models – Past, Present and Future*, 2019. (Cited on page 18.)

[Feng 2018] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang and Xi Zhou. *Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network.* In ECCV, 2018. (Cited on pages 72 and 97.)

[Ferrante 2017] Enzo Ferrante, Puneet K. Dokania, Rafael Marini and Nikos Paragios. Deformable registration through learning of context-specific metric aggregation, pages 256–265. Springer International Publishing, 2017. (Cited on pages 46 and 47.)

[Foulds 2010a] James Foulds and Eibe Frank. *A review of multi-instance learning assumptions.* The Knowledge Engineering Review, vol. 25, no. 1, pages 1–25, 2010. (Cited on pages 10, 78 and 80.)

[Foulds 2010b] James Foulds and Eibe Frank. *A Review of Multi-Instance Learning Assumptions.* The Knowledge Engineering Review, vol. 25, 03 2010. (Cited on page 77.)

[Frey 2003] Brendan J. Frey and Nebojsa Jojic. *Transformation-Invariant Clustering Using the EM Algorithm.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 1, pages 1–17, 2003. (Cited on page 31.)

[Funkhouser 2003] Thomas Funkhouser, Patrick Min, Michael Kazhdan, Joyce Chen, Alex Halderman, David Dobkin and David Jacobs. *A search engine for 3D models.* ACM Transactions on Graphics (TOG), vol. 22, no. 1, pages 83–105, 2003. (Cited on page 12.)

[Garg 2013] Ravi Garg, Anastasios Roussos and Lourdes 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. (Cited on pages 19, 20, 55 and 97.)

[Garrido 2016] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez and Christian Theobalt. *Reconstruction of personalized 3D face rigs from monocular video.* ACM Transactions on Graphics (TOG), vol. 35, no. 3, page 28, 2016. (Cited on page 20.)

[Gaur 2017] Utkarsh Gaur and B. S. Manjunath. *Weakly Supervised Manifold Learning for Dense Semantic Object Correspondence.* In ICCV, 2017. (Cited on pages 31 and 54.)

[Gehler 2011] Peter V. Gehler, Carsten Rother, Martin Kiefel, Lumin Zhang and Bernhard Schölkopf. *Recovering Intrinsic Images with a Global Sparsity Prior on Reflectance.* In NIPS, 2011. (Cited on page 54.)

[Genova 2018] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic and William T. Freeman. *Unsupervised Training for 3D Morphable Model Regression.* In CVPR, 2018. (Cited on pages 17, 18 and 54.)

[Gerig 2018] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn and Thomas Vetter. *Morphable face models-an open framework.* In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 75–82. IEEE, 2018. (Cited on page 17.)

[Gidaris 2018] Spyros Gidaris and Nikos Komodakis. *Dynamic few-shot visual learning without forgetting.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4367–4375, 2018. (Cited on page 23.)

[Gkioxari 2019] Georgia Gkioxari, Jitendra Malik and Justin Johnson. *Mesh R-CNN.* arXiv preprint arXiv:1906.02739, 2019. (Cited on page 3.)

[Goodfellow 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. *Generative adversarial nets.* In Advances in Neural Information Processing Systems, pages 2672–2680, 2014. (Cited on pages 7 and 41.)

[Gotardo 2015] Paulo FU Gotardo, Tomas Simon, Yaser Sheikh and Iain Matthews. *Photogeometric scene flow for high-detail dynamic 3d reconstruction.* In Proceedings of the IEEE International Conference on Computer Vision, pages 846–854, 2015. (Cited on page 55.)

[Gower 1975] John C Gower. *Generalized procrustes analysis.* Psychometrika, vol. 40, no. 1, pages 33–51, 1975. (Cited on page 14.)

[Gross 2005] Ralph Gross, Iain Matthews and Simon Baker. *Generic vs. person specific active appearance models.* Image and Vision Computing, vol. 23, no. 12, pages 1080–1093, 2005. (Cited on page 16.)

[Gross 2010] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade and Simon Baker. *Multi-PIE.* Image Vision Comput., vol. 28, no. 5, pages 807–813, May 2010. (Cited on pages 63 and 66.)

[Güler 2017] Rıza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou and Iasonas Kokkinos. *DenseReg: Fully Convolutional Dense Shape Regression In-the-Wild.* In Computer Vision and Pattern Recognition, 2017. (Cited on pages 18, 31 and 97.)

[Güler 2018] Riza Alp Güler, Natalia Neverova and Iasonas Kokkinos. *DensePose: Dense Human Pose Estimation In The Wild.* In CVPR, 2018. (Cited on pages 23 and 97.)

[Guler 2019] Riza Alp Guler and Iasonas Kokkinos. *Holopose: Holistic 3d human reconstruction in-the-wild.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10884–10894, 2019. (Cited on page 97.)

[Gurcan 2009] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot and B. Yener. *Histopathological Image Analysis: A Review.* IEEE Reviews in Biomedical Engineering, vol. 2, 2009. (Cited on page 78.)

[Hampshire 1992] J. B. Hampshire and A. Waibel. *The Meta-Pi network: building distributed knowledge representations for robust multisource pattern recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14, no. 7, pages 751–769, July 1992. (Cited on page 83.)

[Haralick 1973] R. M. Haralick, K. Shanmugam and I. Dinstein. *Textural Features for Image Classification.* IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pages 610–621, 1973. (Cited on page 87.)

[Hartley 2003] Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision. Cambridge University Press, 2003. (Cited on page 55.)

[He 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. *Deep residual learning for image recognition.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. (Cited on page 81.)

[He 2017] Kaiming He, Georgia Gkioxari, Piotr Dollar and Ross Girshick. *Mask R-CNN.* CVPR, 2017. (Cited on page 3.)

[Heisele 2007] Bernd Heisele, Thomas Serre and Tomaso Poggio. *A component-based framework for face detection and identification.* International Journal of Computer Vision, vol. 74, no. 2, pages 167–181, 2007. (Cited on page 17.)

[Henzler 2018] Philipp Henzler, Niloy Mitra and Tobias Ritschel. *Escaping Plato's Cave using Adversarial Training: 3D Shape From Unstructured 2D Image Collections*. arXiv preprint arXiv:1811.11606, 2018. (Cited on page 55.)

[Hernandez 2017] Matthias Hernandez, Tal Hassner, Jongmoo Choi and Gérard G. Medioni. *Accurate 3D face reconstruction via prior constrained structure from motion*. Computers & Graphics, vol. 66, pages 14–22, 2017. (Cited on page 55.)

[Higgins 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed and Alexander Lerchner. *$\beta$-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*. ICLR, vol. 2, no. 5, page 6, 2017. (Cited on page 17.)

[Hinton 1981] Geoffrey E. Hinton. *A Parallel Computation that Assigns Canonical Object-Based Frames of Reference*. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, IJCAI '81, Vancouver, BC, Canada, August 24-28, 1981, pages 683–685, 1981. (Cited on page 31.)

[Hinton 2002] Geoffrey E Hinton. *Training products of experts by minimizing contrastive divergence*. Neural computation, vol. 14, no. 8, pages 1771–1800, 2002. (Cited on page 5.)

[Hinton 2006] Geoffrey E Hinton and Ruslan R Salakhutdinov. *Reducing the dimensionality of data with neural networks*. science, vol. 313, no. 5786, pages 504–507, 2006. (Cited on pages 5 and 7.)

[Hinton 2011] Geoffrey E. Hinton, Alex Krizhevsky and Sida D. Wang. *Transforming Auto-Encoders*. In Artificial Neural Networks and Machine Learning - ICANN 2011 - 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I, pages 44–51, 2011. (Cited on page 31.)

[Hinton 2016] Geoffrey Hinton. *Neural Networks for Machine Learning Lecture 15a*. 2016. (Cited on page 6.)

[Hou 2016] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis and Joel H Saltz. *Patch-based convolutional neural network for whole slide tissue image classification*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2424–2433, 2016. (Cited on pages 11, 21, 23 and 80.)

[Hou 2019a] Le Hou, Ayush Agarwal, Dimitris Samaras, Tahsin M Kurc, Rajarsi R Gupta and Joel H Saltz. *Robust Histopathology Image Analysis: To Label or to Synthesize?* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8533–8542, 2019. (Cited on page 98.)

[Hou 2019b] Le Hou, Vu Nguyen, Ariel B Kanevsky, Dimitris Samaras, Tahsin M Kurc, Tianhao Zhao, Rajarsi R Gupta, Yi Gao, Wenjin Chen, David Foran*et al. Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images.* Pattern recognition, vol. 86, pages 188–200, 2019. (Cited on page 98.)

[Hu 2018] Jie Hu, Li Shen and Gang Sun. *Squeeze-and-excitation networks.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7132–7141, 2018. (Cited on page 45.)

[Huang 2007] G. B. Huang, V. Jain and E. Learned-Miller. *Unsupervised Joint Alignment of Complex Images.* In 2007 IEEE 11th International Conference on Computer Vision, 2007. (Cited on page 13.)

[Huang 2017] Gao Huang, Zhuang Liu, Laurens van der Maaten and Kilian Q Weinberger. *Densely connected convolutional networks.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. (Cited on pages 21 and 36.)

[Huang 2018] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu and Jingdong Wang. *Weakly-supervised semantic segmentation network with deep seeded region growing.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7014–7023, 2018. (Cited on page 10.)

[Huber 2016] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch and Josef Kittler. *A multiresolution 3D morphable face model and fitting framework.* In Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2016. (Cited on page 17.)

[Hussain 2017a] Mohammad Arafat Hussain, Alborz Amir-Khalili, Ghassan Hamarneh and Rafeef Abugharbieh. *Collage CNN for renal cell carcinoma detection from CT.* In International Workshop on Machine Learning in Medical Imaging, pages 229–237. Springer, 2017. (Cited on page 22.)

[Hussain 2017b] Mohammad Arafat Hussain, Alborz Amir-Khalili, Ghassan Hamarneh and Rafeef Abugharbieh. *Segmentation-free kidney localization and volume estimation using aggregated orthogonal decision CNNS.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 612–620. Springer, 2017. (Cited on page 22.)

[Hussain 2018] Mohammad Arafat Hussain, Ghassan Hamarneh and Rafeef Garbi. *Noninvasive Determination of Gene Mutations in Clear Cell Renal Cell Carcinoma Using Multiple Instance Decisions Aggregated CNN.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 657–665. Springer, 2018. (Cited on page 80.)

[Hwang 2016] Sangheum Hwang and Hyo-Eun Kim. *Self-transfer learning for weakly supervised lesion localization.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 239–246. Springer, 2016. (Cited on page 23.)

[Ilse 2018] Maximilian Ilse, Jakub M. Tomczak and Max Welling. *Attention-based Deep Multiple Instance Learning.* In ICML, 2018. (Cited on pages 11, 23, 78, 80, 87, 89, 90, 91 and 92.)

[Isola 2016] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou and Alexei A Efros. *Image-to-Image Translation with Conditional Adversarial Networks.* arxiv, 2016. (Cited on pages 8, 42 and 115.)

[Jacobs 1991] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan and Geoffrey E. Hinton. *Adaptive Mixtures of Local Experts.* Neural Comput., vol. 3, no. 1, March 1991. (Cited on pages 76 and 83.)

[Jaderberg 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman and Koray Kavukcuoglu. *Spatial Transformer Networks.* CoRR, vol. abs/1506.02025, 2015. (Cited on pages 20, 21, 31 and 32.)

[Jain 1996] Anil K. Jain, Yu Zhong and Sridhar Lakshmanan. *Object matching using deformable templates.* IEEE Transactions on pattern analysis and machine intelligence, vol. 18, no. 3, pages 267–278, 1996. (Cited on page 12.)

[Jakab 2018a] Tomas Jakab, Ankush Gupta, Hakan Bilen and Andrea Vedaldi. *Unsupervised learning of object landmarks through conditional image generation.* In Advances in Neural Information Processing Systems, pages 4016–4027, 2018. (Cited on page 9.)

[Jakab 2018b] Tomas Jakab, Ankush Gupta, Hakan Bilen and Andrea 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, editeurs, Advances in Neural Information Processing Systems 31, pages 4016–4027. Curran Associates, Inc., 2018. (Cited on page 69.)

[Janner 2017] Michael Janner, Jiajun Wu, Tejas D. Kulkarni, Ilker Yildirim and Josh Tenenbaum. *Self-Supervised Intrinsic Image Decomposition.* In NIPS, 2017. (Cited on page 54.)

[Jeon 2017] Yunho Jeon and Junmo Kim. *Active convolution: Learning the shape of convolution for image classification.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4201–4209, 2017. (Cited on page 20.)

[Jia 2017] Zhipeng Jia, Xingyi Huang, I Eric, Chao Chang and Yan Xu. *Constrained deep weak supervision for histopathology image segmentation.* IEEE

transactions on medical imaging, vol. 36, no. 11, pages 2376–2388, 2017. (Cited on page 22.)

[Jimenez Rezende 2016] Danilo Jimenez Rezende, S. M. Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg and Nicolas Heess. *Unsupervised Learning of 3D Structure from Images*. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, editeurs, Advances in Neural Information Processing Systems 29, pages 4996–5004. Curran Associates, Inc., 2016. (Cited on page 55.)

[Jojic 2003] Nebojsa Jojic, Brendan J. Frey and Anitha Kannan. *Epitomic analysis of appearance and shape*. In 9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France, pages 34–43, 2003. (Cited on page 31.)

[Jones 1998] Michael J Jones and Tomaso Poggio. *Multidimensional morphable models*. In Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), pages 683–688. IEEE, 1998. (Cited on page 15.)

[Jordan 1994] Michael I Jordan and Robert A Jacobs. *Hierarchical mixtures of experts and the EM algorithm*. Neural computation, vol. 6, no. 2, pages 181–214, 1994. (Cited on page 83.)

[Kallenberg 2016] Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Y Ng, Pengfei Diao, Christian Igel, Celine M Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer*et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring*. IEEE transactions on medical imaging, vol. 35, no. 5, pages 1322–1331, 2016. (Cited on page 23.)

[Kamnitsas 2017] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert and Ben Glocker. *Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation*. Medical image analysis, vol. 36, pages 61–78, 2017. (Cited on page 21.)

[Kanazawa 2018a] Angjoo Kanazawa, Michael J. Black, David W. Jacobs and Jitendra Malik. *End-to-end Recovery of Human Shape and Pose*. In Computer Vision and Pattern Regognition (CVPR), 2018. (Cited on page 97.)

[Kanazawa 2018b] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros and Jitendra Malik. *Learning Category-Specific Mesh Reconstruction from Image Collections*. In ECCV, 2018. (Cited on pages 20, 54, 55 and 116.)

[Kantorov 2016] Vadim Kantorov, Maxime Oquab, Minsu Cho and Ivan Laptev. *Contextlocnet: Context-aware deep network models for weakly supervised localization*. In European Conference on Computer Vision, pages 350–365. Springer, 2016. (Cited on page 10.)

[Karantzalos 2014a] K. Karantzalos, A. Sotiras and N. Paragios. *Efficient and Automated Multimodal Satellite Data Registration through MRFs and Linear Programming*. In 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014. (Cited on pages 47 and 48.)

[Karantzalos 2014b] Konstantinos Karantzalos, Aristeidis Sotiras and Nikos Paragios. *Efficient and automated multimodal satellite data registration through MRFs and linear programming*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 329–336, 2014. (Cited on page 47.)

[Karras 2018] Tero Karras, Timo Aila, Samuli Laine and Jaakko Lehtinen. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. In International Conference on Learning Representations, 2018. (Cited on page 71.)

[Karras 2019] Tero Karras, Samuli Laine and Timo Aila. *A style-based generator architecture for generative adversarial networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4401–4410, 2019. (Cited on pages 8 and 54.)

[Katharopoulos 2019] Angelos Katharopoulos and Francois Fleuret. *Processing Megapixel Images with Deep Attention-Sampling Models*. In International Conference on Machine Learning, pages 3282–3291, 2019. (Cited on page 23.)

[Kato 2018a] Hiroharu Kato. *The Neural Mesh Renderer on Github*. `https://github.com/hiroharu-kato/neural_renderer`, 2018. (Cited on page 116.)

[Kato 2018b] Hiroharu Kato, Yoshitaka Ushiku and Tatsuya Harada. *Neural 3D Mesh Renderer*. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. (Cited on pages 52, 54, 59 and 116.)

[Keeler 1991] James D Keeler, David E Rumelhart and Wee Kheng Leow. *Integrated segmentation and recognition of hand-printed numerals*. In Advances in neural information processing systems, pages 557–563, 1991. (Cited on pages 10 and 77.)

[Kemelmacher-Shlizerman 2011] Ira Kemelmacher-Shlizerman and Steven M Seitz. *Face reconstruction in the wild*. In 2011 International Conference on Computer Vision, pages 1746–1753. IEEE, 2011. (Cited on page 55.)

[Kemelmacher-Shlizerman 2012] Ira Kemelmacher-Shlizerman and Steven M Seitz. *Collection flow*. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 1792–1799. IEEE, 2012. (Cited on pages 13 and 55.)

[Kemelmacher-Shlizerman 2013] Ira Kemelmacher-Shlizerman. *Internet based morphable model*. In Proceedings of the IEEE International Conference on Computer Vision, pages 3256–3263, 2013. (Cited on pages 24 and 55.)

[Kim 2018] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer and Christian Theobalt. *Deep Video Portraits.* ACM Trans. Graph., vol. 37, no. 4, pages 163:1–163:14, July 2018. (Cited on page 54.)

[Kingma 2013] Diederik P Kingma and Max Welling. *Auto-encoding variational bayes.* arXiv preprint arXiv:1312.6114, 2013. (Cited on page 6.)

[Kingma 2014a] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization.* CoRR, vol. abs/1412.6980, 2014. (Cited on pages 46, 65 and 117.)

[Kingma 2014b] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization.* CoRR, vol. abs/1412.6980, 2014. (Cited on page 85.)

[Koch 2015] Gregory Koch, Richard Zemel and Ruslan Salakhutdinov. *Siamese neural networks for one-shot image recognition.* In ICML Deep Learning Workshop, volume 2, 2015. (Cited on page 62.)

[Kokkinos 2007] Iasonas Kokkinos and Alan L. Yuille. *Unsupervised Learning of Object Deformation Models.* In ICCV, 2007. (Cited on pages 15, 31 and 97.)

[Komura 2018] Daisuke Komura and Shumpei Ishikawa. *Machine Learning Methods for Histopathological Image Analysis.* Computational and Structural Biotechnology Journal, vol. 16, 2018. (Cited on pages 78 and 98.)

[Kong 2014] Naejin Kong, Peter V. Gehler and Michael J. Black. *Intrinsic Video.* In Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II, pages 360–375, 2014. (Cited on page 54.)

[Kong 2019] Chen Kong and Simon Lucey. *Deep Interpretable Non-Rigid Structure from Motion.* CoRR, vol. abs/1902.10840, 2019. (Cited on page 55.)

[Kotzias 2014] Dimitrios Kotzias, Misha Denil, Phil Blunsom and Nando de Freitas. *Deep multi-instance transfer learning.* arXiv preprint arXiv:1411.3128, 2014. (Cited on page 80.)

[Koujan 2018] Mohammad Rami Koujan and Anastasios Roussos. *Combining Dense Nonrigid Structure from Motion and 3D Morphable Models for Monocular 4D Face Reconstruction.* In CVMP, 2018. (Cited on page 56.)

[Kraus 2016] Oren Z. Kraus, Jimmy Lei Ba and Brendan J. Frey. *Classifying and segmenting microscopy images with deep multiple instance learning.* Bioinformatics, vol. 32, no. 12, pages i52–i59, 2016. (Cited on pages 11, 78 and 80.)

[Krizhevsky 2012] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. *Imagenet classification with deep convolutional neural networks.* In Advances in

Neural Information Processing Systems, pages 1097–1105, 2012. (Cited on pages 21 and 85.)

[Kulkarni 2019] Nilesh Kulkarni, Abhinav Gupta and Shubham Tulsiani. *Canonical Surface Mapping via Geometric Cycle Consistency.* arXiv preprint arXiv:1907.10043, 2019. (Cited on page 98.)

[Lample 2017] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer and Marc'Aurelio Ranzato. *Fader Networks: Manipulating Images by Sliding Attributes.* CoRR, vol. abs/1706.00409, 2017. (Cited on page 30.)

[Larochelle 2008] Hugo Larochelle, Dumitru Erhan and Yoshua Bengio. *Zero-data learning of new tasks.* In AAAI, volume 1, page 3, 2008. (Cited on page 23.)

[Le 2019] Han Le, Rajarsi Gupta, Le Hou, Shahira Abousamra, Danielle Fassler, Tahsin Kurc, Dimitris Samaras, Rebecca Batiste, Tianhao Zhao, Alison L Van Dyke*et al.* *Utilizing Automated Breast Cancer Detection to Identify Spatial Distributions of Tumor Infiltrating Lymphocytes in Invasive Breast Cancer.* arXiv preprint arXiv:1905.10841, 2019. (Cited on page 98.)

[Learned-Miller 2005] Erik G Learned-Miller and Vidit Jain. *Many heads are better than one: Jointly removing bias from multiple MRIs using nonparametric maximum likelihood.* In Biennial International Conference on Information Processing in Medical Imaging, pages 615–626. Springer, 2005. (Cited on page 12.)

[Learned-Miller 2006] Erik G. Learned-Miller. *Data Driven Image Models through Continuous Joint Alignment.* PAMI, 2006. (Cited on pages 12, 13 and 31.)

[LeCun 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner*et al.* *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, vol. 86, no. 11, pages 2278–2324, 1998. (Cited on page 5.)

[LeCun 2015] Yann LeCun, Yoshua Bengio and Geoffrey Hinton. *Deep learning.* nature, vol. 521, no. 7553, pages 436–444, 2015. (Cited on page 3.)

[Ledig 2017] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang*et al.* *Photo-realistic single image super-resolution using a generative adversarial network.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4681–4690, 2017. (Cited on page 8.)

[Lee 2009] Honglak Lee, Roger Grosse, Rajesh Ranganath and Andrew Y Ng. *Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations.* In Proceedings of the 26th annual international conference on machine learning, pages 609–616. ACM, 2009. (Cited on pages 3 and 6.)

[Lee 2011] Honglak Lee, Roger Grosse, Rajesh Ranganath and Andrew Y Ng. *Unsupervised learning of hierarchical representations with convolutional deep belief networks.* Communications of the ACM, vol. 54, no. 10, pages 95–103, 2011. (Cited on page 6.)

[Lee 2019] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee and Sungroh Yoon. *FickleNet: Weakly and Semi-Supervised Semantic Image Segmentation Using Stochastic Inference.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5267–5276, 2019. (Cited on pages 4 and 11.)

[Leopold 2001] David A Leopold, Alice J O'Toole, Thomas Vetter and Volker Blanz. *Prototype-referenced shape encoding revealed by high-level aftereffects.* Nature neuroscience, vol. 4, no. 1, page 89, 2001. (Cited on page 17.)

[Li 2006] Fei-Fei Li, Rob Fergus and Pietro Perona. *One-shot learning of object categories.* IEEE transactions on pattern analysis and machine intelligence, vol. 28, no. 4, pages 594–611, 2006. (Cited on page 23.)

[Li 2016] Chuan Li and Michael Wand. *Precomputed real-time texture synthesis with markovian generative adversarial networks.* In European Conference on Computer Vision, pages 702–716. Springer, 2016. (Cited on page 42.)

[Li 2017] Tianye Li, Timo Bolkart, Michael J Black, Hao Li and Javier Romero. *Learning a model of facial shape and expression from 4D scans.* ACM Transactions on Graphics (TOG), vol. 36, no. 6, page 194, 2017. (Cited on page 17.)

[Li 2019a] Jiayun Li, Wenyuan Li, Arkadiusz Gertych, Beatrice S Knudsen, William Speier and Corey W Arnold. *An attention-based multi-resolution model for prostate whole slide imageclassification and localization.* arXiv preprint arXiv:1905.13208, 2019. (Cited on page 11.)

[Li 2019b] Shaohua Li, Yong Liu, Xiuchao Sui, Cheng Chen, Gabriel Tjio, Daniel Shu Wei Ting and Rick Siow Mong Goh. *Multi-Instance Multi-Scale CNN for Medical Image Classification.* In MICCAI, 2019. (Cited on page 11.)

[Liang 2018] Luming Liang, Mingqiang Wei, Andrzej Szymczak, Anthony Petrella, Haoran Xie, Jing Qin, Jun Wang and Fu Lee Wang. *Nonrigid iterative closest points for registration of 3D biomedical surfaces.* Optics and Lasers in Engineering, vol. 100, pages 141–154, 2018. (Cited on page 17.)

[Limkin 2017] Elaine Johanna Limkin, Roger Sun, Laurent Dercle, Evangelia I Zacharaki, Charlotte Robert, Sylvain Reuzé, Antoine Schernberg, Nikos Paragios, Eric Deutsch and Charles Ferté. *Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology.* Annals of Oncology, vol. 28, no. 6, 2017. (Cited on page 76.)

[Litjens 2017] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken and Clara I. SÃ¡nchez. *A survey on deep learning in medical image analysis.* Medical Image Analysis, vol. 42, pages 60 – 88, 2017. (Cited on page 98.)

[Liu-Yin 2017] Qi Liu-Yin, Rui Yu, Lourdes Agapito, Andrew Fitzgibbon and Chris Russell. *Better together: Joint reasoning for non-rigid 3d reconstruction with specularities and shading.* arXiv preprint arXiv:1708.01654, 2017. (Cited on pages 20 and 55.)

[Liu 2015] Ziwei Liu, Ping Luo, Xiaogang Wang and Xiaoou Tang. *Deep Learning Face Attributes in the Wild.* In Proceedings of International Conference on Computer Vision (ICCV), December 2015. (Cited on pages 37 and 66.)

[Lorenz 2019] Dominik Lorenz, Leonard Bereska, Timo Milbich and Bjorn Ommer. *Unsupervised Part-Based Disentangling of Object Shape and Appearance.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10955–10964, 2019. (Cited on page 97.)

[Lowe 2004] David G Lowe. *Distinctive image features from scale-invariant keypoints.* International journal of computer vision, vol. 60, no. 2, pages 91–110, 2004. (Cited on page 13.)

[Lu 2016] Xiaoguang Lu, Daguang Xu and David Liu. *Robust 3D organ localization with dual learning architectures and fusion.* In Deep Learning and Data Labeling for Medical Applications, pages 12–20. Springer, 2016. (Cited on page 22.)

[Lucas 1981] Bruce D Lucas, Takeo Kanade*et al. An iterative image registration technique with an application to stereo vision.* 1981. (Cited on page 16.)

[Makhzani 2013] Alireza Makhzani and Brendan Frey. *K-sparse autoencoders.* arXiv preprint arXiv:1312.5663, 2013. (Cited on page 6.)

[Makhzani 2015] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow and Brendan Frey. *Adversarial autoencoders.* arXiv preprint arXiv:1511.05644, 2015. (Cited on page 5.)

[Malsburg 1981] Christoph Malsburg. *The correlation theory of brain function.* In Internal Report 81-2. Gottingen Max-Planck-Institute for Biophysical Chemistry., 1981. (Cited on page 31.)

[Maron 1998] Oded Maron and Tomás Lozano-Pérez. *A framework for multiple-instance learning.* In Advances in neural information processing systems, pages 570–576, 1998. (Cited on page 80.)

[Masci 2011] Jonathan Masci, Ueli Meier, Dan Cireşan and Jürgen Schmidhuber. *Stacked convolutional auto-encoders for hierarchical feature extraction.* In International Conference on Artificial Neural Networks, pages 52–59. Springer, 2011. (Cited on page 6.)

[Mathieu 2015] Michael Mathieu, Camille Couprie and Yann LeCun. *Deep multi-scale video prediction beyond mean square error.* arXiv preprint arXiv:1511.05440, 2015. (Cited on page 6.)

[Matthews 2004] I. Matthews and S. Baker. *Active Appearance Models Revisited.* IJCV, 2004. (Cited on pages 14, 15, 16, 24, 29, 31, 32 and 97.)

[Memisevic 2010] Roland Memisevic and Geoffrey E. Hinton. *Learning to Represent Spatial Transformations with Factored Higher-order Boltzmann Machines.* Neural Computation, 2010. (Cited on pages 30, 31 and 54.)

[Milletari 2016] Fausto Milletari, Nassir Navab and Seyed-Ahmad Ahmadi. *V-net: Fully convolutional neural networks for volumetric medical image segmentation.* In 2016 Fourth International Conference on 3D Vision (3DV), pages 565–571. IEEE, 2016. (Cited on page 21.)

[Moerland 1997] Perry Moerland. *Mixtures of experts estimate a posteriori probabilities.* 1997. (Cited on page 83.)

[Moeskops 2016] Pim Moeskops, Jelmer M Wolterink, Bas HM van der Velden, Kenneth GA Gilhuijs, Tim Leiner, Max A Viergever and Ivana Išgum. *Deep learning for multi-task medical image segmentation in multiple modalities.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 478–486. Springer, 2016. (Cited on page 22.)

[Narihira 2015] Takuya Narihira, Michael Maire and Stella X. Yu. *Direct Intrinsics: Learning Albedo-Shading Decomposition by Convolutional Regression.* In ICCV, 2015. (Cited on page 54.)

[Neverova 2018] Natalia Neverova and Iasonas Kokkinos. *Mass Displacement Networks.* In BMVC, 2018. (Cited on page 31.)

[Ng 2007] Shu-Kay Ng and Geoffrey J McLachlan. *Extension of mixture-of-experts networks for binary classification of hierarchical data.* Artificial Intelligence in Medicine, vol. 41, no. 1, pages 57–67, 2007. (Cited on page 83.)

[Ng 2011] Andrew Ng*et al.* *Sparse autoencoder.* CS294A Lecture notes, vol. 72, no. 2011, pages 1–19, 2011. (Cited on pages 6 and 7.)

[Novotny 2017] David Novotny, Diane Larlus and Andrea Vedaldi. *Learning 3d object categories by looking around them.* In Proceedings of the IEEE International Conference on Computer Vision, pages 5218–5227, 2017. (Cited on page 54.)

[Nowlan 1991] Steven J Nowlan and Geoffrey E Hinton. *Evaluation of adaptive mixtures of competing experts.* In Advances in neural information processing systems, pages 774–780, 1991. (Cited on page 83.)

[nr2 2018] *A PyTorch port of the Neural Mesh Renderer on Github.* https://github.com/daniilidis-group/neural_renderer, 2018. (Cited on page 116.)

[Olshausen 1995] Bruno A. Olshausen, Charles H. Anderson and David C. Van Essen. *A multiscale dynamic routing circuit for forming size- and position-invariant object representations.* Journal of Computational Neuroscience, vol. 2, no. 1, pages 45–62, 1995. (Cited on page 31.)

[Oquab 2014] Maxime Oquab, Leon Bottou, Ivan Laptev and Josef Sivic. *Learning and transferring mid-level image representations using convolutional neural networks.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1717–1724, 2014. (Cited on page 10.)

[Oquab 2015] Maxime Oquab, Léon Bottou, Ivan Laptev and Josef Sivic. *Is object localization for free?-weakly-supervised learning with convolutional neural networks.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 685–694, 2015. (Cited on page 10.)

[Paladini 2009] Marco Paladini, Alessio Del Bue, Marko Stosic, Marija Dodig, Joao Xavier and Lourdes 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. (Cited on pages 19, 55 and 59.)

[Papadopoulos 2019] A. Papadopoulos, K. Kyritsis, S. Bostanjopoulou, L. Klingelhoefer, R. K. Chaudhuri and A. Delopoulos. *Multiple-Instance Learning for In-The-Wild Parkinsonian Tremor Detection.* In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 6188–6191, July 2019. (Cited on page 11.)

[Papandreou 2008] George Papandreou and Petros Maragos. *Adaptive and constrained algorithms for inverse compositional active appearance model fitting.* In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008. (Cited on page 16.)

[Papandreou 2015] George Papandreou, Iasonas Kokkinos and Pierre-André Savalle. *Modeling local and global deformations in Deep Learning: Epitomic convolution, Multiple Instance Learning, and sliding window detection.* In CVPR, 2015. (Cited on page 31.)

[Papastergiou 2018] Thomas Papastergiou, Evangelia Zacharaki and Vasileios Megalooikonomou. *Tensor Decomposition for Multiple-Instance Classification of High-Order Medical Data.* Complexity, 12 2018. (Cited on page 78.)

[Park 2017] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan and Alexander C Berg. *Transformation-grounded image generation network for novel 3d view synthesis.* In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 702–711. IEEE, 2017. (Cited on page 30.)

[Paszke 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga and Adam Lerer. *Automatic differentiation in PyTorch.* In NIPS 2017 Autodiff Workshop, 2017. (Cited on pages 85 and 116.)

[Patel 2009] Ankur Patel and William AP Smith. *3d morphable face models revisited.* In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 1327–1334. IEEE, 2009. (Cited on page 17.)

[Pavlakos 2017] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G Derpanis and Kostas Daniilidis. *6-DoF Object Pose from Semantic Keypoints.* In ICRA, 2017. (Cited on page 69.)

[Paysan 2009] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani and Thomas Vetter. *A 3D face model for pose and illumination invariant face recognition.* In 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pages 296–301. Ieee, 2009. (Cited on pages 17 and 24.)

[Pierrard 2019] Régis Pierrard, Jean-Philippe Poli and Céline Hudelot. *A New Approach for Explainable Multiple Organ Annotation with Few Data.* In IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI), Macao, Macau SAR China, August 2019. (Cited on page 23.)

[Pinheiro 2015] Pedro O Pinheiro and Ronan Collobert. *From image-level to pixel-level labeling with convolutional networks.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1713–1721, 2015. (Cited on page 10.)

[Pumarola 2018] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu and Francesc 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. (Cited on pages 8 and 54.)

[Radford 2015] Alec Radford, Luke Metz and Soumith Chintala. *Unsupervised representation learning with deep convolutional generative adversarial networks.* arXiv preprint arXiv:1511.06434, 2015. (Cited on page 7.)

[Ramon 2000] Jan Ramon and Luc De Raedt. *Multi instance neural networks.* In Proceedings of the ICML-2000 workshop on attribute-value and relational learning, pages 53–60, 2000. (Cited on pages 78 and 80.)

[Ravi 2017] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo and G. Yang. *Deep Learning for Health Informatics*. IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, Jan 2017. (Cited on page 76.)

[Richardson 2016] Elad Richardson, Matan Sela and Ron Kimmel. *3D face reconstruction by learning from synthetic data*. In 2016 Fourth International Conference on 3D Vision (3DV), pages 460–469. IEEE, 2016. (Cited on pages 4 and 18.)

[Romdhani 2005] Sami Romdhani and Thomas Vetter. *Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior*. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pages 986–993. IEEE, 2005. (Cited on pages 17 and 18.)

[Ronneberger 2015] Olaf Ronneberger, Philipp Fischer and Thomas 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. (Cited on pages 21, 59 and 115.)

[Ruifrok 2001] Arnout C Ruifrok, Dennis A Johnston*et al. Quantification of histochemical staining by color deconvolution*. Analytical and quantitative cytology and histology, vol. 23, no. 4, pages 291–299, 2001. (Cited on page 21.)

[Russell 2011] Chris Russell, Joao Fayad and Lourdes Agapito. *Energy based multiple model fitting for non-rigid structure from motion*. In CVPR 2011, pages 3009–3016. IEEE, 2011. (Cited on page 19.)

[Sabour 2017] Sara Sabour, Nicholas Frosst and Geoffrey E. Hinton. *Dynamic Routing Between Capsules*. CoRR, vol. abs/1710.09829, 2017. (Cited on page 31.)

[Salgado 2015] Roberto Salgado, Carsten Denkert, Christine Campbell, Peter Savas, Paolo Nuciforo, Claudia Aura, Evandro De Azambuja, Holger Eidtmann, Catherine E Ellis, Jose Baselga*et al. Tumor-infiltrating lymphocytes and associations with pathological complete response and event-free survival in HER2-positive early-stage breast cancer treated with lapatinib and trastuzumab: a secondary analysis of the NeoALTTO trial*. JAMA oncology, vol. 1, no. 4, pages 448–455, 2015. (Cited on page 98.)

[Salimans 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford and Xi Chen. *Improved techniques for training gans*. In Advances in neural information processing systems, pages 2234–2242, 2016. (Cited on page 7.)

[Saltz 2018] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Re-

becca Batiste*et al.* *Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images.* Cell reports, vol. 23, no. 1, pages 181–193, 2018. (Cited on page 98.)

[Samaras 2000] Dimitris Samaras, Dimitris Metaxas, Pascal Fua and Yvan 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. (Cited on page 61.)

[Sanyal 2019] Soubhik Sanyal, Timo Bolkart, Haiwen Feng and Michael J Black. *Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7763–7772, 2019. (Cited on page 18.)

[Schönberger 2016] Johannes Lutz Schönberger and Jan-Michael Frahm. *Structure-from-Motion Revisited.* In Conference on Computer Vision and Pattern Recognition (CVPR), 2016. (Cited on page 55.)

[Sclaroff 1998] Stan Sclaroff and John Isidoro. *Active blobs.* In Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), pages 1146–1153. IEEE, 1998. (Cited on page 15.)

[Sela 2017] Matan Sela, Elad Richardson and Ron Kimmel. *Unrestricted facial geometry reconstruction using image-to-image translation.* In Proceedings of the IEEE International Conference on Computer Vision, pages 1576–1585, 2017. (Cited on page 8.)

[Sengupta 2017] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo and David Jacobs. *SfSNet: Learning Shape, Reflectance and Illuminance of Faces in the Wild.* arXiv preprint arXiv:1712.01261, 2017. (Cited on page 30.)

[Sengupta 2018a] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo and David W. Jacobs. *SfSNet : Learning Shape, Reflectance and Illuminance of Faces in the Wild.* In CVPR, 2018. (Cited on page 54.)

[Sengupta 2018b] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo and David W. Jacobs. *SfSNet: Learning Shape, Refectance and Illuminance of Faces in the Wild.* In Computer Vision and Pattern Regognition (CVPR), 2018. (Cited on pages 8, 54 and 60.)

[Seyedhosseini 2013] Mojtaba Seyedhosseini, Mehdi Sajjadi and Tolga Tasdizen. *Image segmentation with cascaded hierarchical models and logistic disjunctive normal networks.* In Proceedings of the IEEE International Conference on Computer Vision, pages 2168–2175, 2013. (Cited on page 21.)

[Shu 2017] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman and Dimitris Samaras. *Neural Face Editing with Intrinsic Image Disentangling.* In CVPR, 2017. (Cited on pages 8, 30, 35, 41, 54, 60, 61 and 62.)

[Shu 2018] Zhixin Shu, Mihir Sahasrabudhe, Rıza Alp Güler, Dimitris Samaras, Nikos Paragios and Iasonas Kokkinos. *Deforming Autoencoders: Unsupervised Disentangling of Shape and Appearance.* In European Conference on Computer Vision, 2018. (Cited on pages 54, 56, 64, 68, 69 and 115.)

[Simon 2014] Tomas Simon, Jack Valmadre, Iain Matthews and Yaser Sheikh. *Separable spatiotemporal priors for convex reconstruction of time-varying 3D point clouds.* In European Conference on Computer Vision, pages 204–219. Springer, 2014. (Cited on page 55.)

[Skourt 2018] Brahim Ait Skourt, Abdelhamid El Hassani and Aicha Majda. *Lung CT Image Segmentation using deep neural networks.* Procedia Computer Science, vol. 127, pages 109–113, 2018. (Cited on page 21.)

[Smolensky 1986] Paul Smolensky. *Information processing in dynamical systems: Foundations of harmony theory.* Rapport technique, Colorado Univ at Boulder Dept of Computer Science, 1986. (Cited on page 5.)

[Sotiras 2013] Aristeidis Sotiras, Christos Davatzikos and Nikos Paragios. *Deformable medical image registration: A survey.* IEEE transactions on medical imaging, vol. 32, no. 7, page 1153, 2013. (Cited on page 45.)

[Spanhol 2016] F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte. *A Dataset for Breast Cancer Histopathological Image Classification.* IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, 2016. (Cited on page 78.)

[Sparks 2016] Rachel Sparks and Anant Madabhushi. *Out-of-Sample Extrapolation utilizing Semi-Supervised Manifold Learning (OSE-SSL): Content Based Image Retrieval for Histopathology Images.* Scientific Reports, vol. 6, page 27306, 06 2016. (Cited on page 78.)

[Sun 2018] Roger Sun, Elaine Johanna Limkin, Maria Vakalopoulou, Laurent Dercle, Stephane Champiat, Shan Rong Han, Loic Verlingue, David Brandao, Andrea Lancia, Samy Ammari, Antoine Hollebecque, Jean-Yves Scoazec, AurÃ©lien Marabelle, Christophe Massard, Jean-Charles Soria, Charlotte Robert, Nikos Paragios, Eric Deutsch and Charles Ferté. *A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study.* The Lancet Oncology, vol. 19, no. 9, 2018. (Cited on page 76.)

[Sundermeyer 2018] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker and Rudolph 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. (Cited on pages 8 and 54.)

[Suwajanakorn 2018] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson and Mohammad Norouzi. *Discovery of latent 3d keypoints via end-to-end*

*geometric reasoning*. In Advances in Neural Information Processing Systems, pages 2059–2070, 2018. (Cited on page 9.)

[Tang 2011] Yichuan Tang and Ilya Sutskever. *Data normalization in the learning of restricted Boltzmann machines*. Department of Computer Science, University of Toronto, Technical Report UTML-TR-11-2, 2011. (Cited on page 4.)

[Tewari 2017] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez and Christian Theobalt. *Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction*. In Proceedings of the IEEE International Conference on Computer Vision, pages 1274–1283, 2017. (Cited on pages 4, 8 and 18.)

[Tewari 2018] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer and Christian Theobalt. *FML: Face Model Learning from Videos*. CoRR, vol. abs/1812.07603, 2018. (Cited on pages 4, 18, 54 and 56.)

[Theera-Umpon 2007] N. Theera-Umpon and S. Dhompongsa. *Morphological Granulometric Features of Nucleus in Automatic Bone Marrow White Blood Cell Classification*. IEEE Transactions on Information Technology in Biomedicine, vol. 11, no. 3, 2007. (Cited on page 78.)

[Thewlis 2017a] 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. (Cited on page 69.)

[Thewlis 2017b] James Thewlis, Hakan Bilen and Andrea Vedaldi. *Unsupervised learning of object frames by dense equivariant image labelling*. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, editeurs, Advances in Neural Information Processing Systems 30, pages 844–855. Curran Associates, Inc., 2017. (Cited on pages 9 and 69.)

[Thewlis 2017c] James Thewlis, Hakan Bilen and Andrea Vedaldi. *Unsupervised learning of object frames by dense equivariant image labelling*. In NIPS, 2017. (Cited on pages 30, 31, 43, 44 and 54.)

[Thies 2016] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt and Matthias Nießner. *Face2Face: Real-Time Face Capture and Reenactment of RGB Videos*. In CVPR, 2016. (Cited on page 56.)

[Thompson 1917] D'Arcy Wentworth Thompson. On growth and form. Cambridge University Press, 1917. (Cited on page 12.)

[Thong 2018] William Thong, Samuel Kadoury, Nicolas Piché and Christopher J Pal. *Convolutional networks for kidney segmentation in contrast-enhanced*

*CT scans*. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, vol. 6, no. 3, pages 277–282, 2018. (Cited on page 22.)

[Tipping 1999] Michael E Tipping and Christopher M Bishop. *Probabilistic principal component analysis*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 61, no. 3, pages 611–622, 1999. (Cited on page 19.)

[Tomasi 1992] Carlo Tomasi and Takeo Kanade. *Shape and motion from image streams under orthography: a factorization method*. International Journal of Computer Vision, vol. 9, no. 2, pages 137–154, 1992. (Cited on page 55.)

[Torresani 2008] Lorenzo Torresani, Aaron Hertzmann and Chris Bregler. *Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors*. IEEE transactions on pattern analysis and machine intelligence, vol. 30, no. 5, pages 878–892, 2008. (Cited on pages 19, 24, 55, 59 and 97.)

[Tran 2018] Luan Tran and Xiaoming Liu. *Nonlinear 3D Face Morphable Model*. In In Proceeding of IEEE Computer Vision and Pattern Recognition, Salt Lake City, UT, June 2018. (Cited on page 55.)

[Trigeorgis 2016] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos and Stefanos Zafeiriou. *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. (Cited on page 31.)

[Tulsiani 2017] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros and Jitendra 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. (Cited on page 54.)

[Tzimiropoulos 2017] Georgios Tzimiropoulos and Maja Pantic. *Fast algorithms for fitting active appearance models to unconstrained images*. International journal of computer vision, vol. 122, no. 1, pages 17–33, 2017. (Cited on page 16.)

[Ummenhofer 2017] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy and Thomas Brox. *Demon: Depth and motion network for learning monocular stereo*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5038–5047, 2017. (Cited on page 55.)

[Vakalopoulou 2016] Maria Vakalopoulou, Konstantinos Karantzalos, Nikos Komodakis and Nikos Paragios. *Graph-based registration, change detection, and classification in very high resolution multitemporal remote sensing data*.

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 9, no. 7, pages 2940–2951, 2016. (Cited on page 47.)

[Vetter 1997] Thomas Vetter, Michael J. Jones and Tomaso 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. (Cited on page 58.)

[Walker 2002] Kevin N Walker, Timothy F Cootes and Christopher J Taylor. *Automatically building appearance models from image sequences using salient features.* Image and Vision Computing, vol. 20, no. 5-6, pages 435–440, 2002. (Cited on page 15.)

[Wang 2007] Yang Wang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou 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. (Cited on pages 52 and 60.)

[Wang 2009] Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang and D. Samaras. *Face Relighting from a Single Image under Arbitrary Unknown Lighting Conditions.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, no. 11, pages 1968 –1984, nov. 2009. (Cited on pages 52 and 60.)

[Wang 2017] Chaofeng Wang, Jun Shi, Qi Zhang and Shihui Ying. *Histopathological image classification with bilinear convolutional neural networks.* In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 4050–4053. IEEE, 2017. (Cited on page 98.)

[Wang 2018] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai and Wenyu Liu. *Revisiting multiple instance neural networks.* Pattern Recognition, vol. 74, pages 15–24, 2018. (Cited on pages 11 and 80.)

[Waterhouse 1998] Steven Richard Waterhouse. *Classification and regression using mixtures of experts.* PhD thesis, Citeseer, 1998. (Cited on page 83.)

[Wiles 2018a] O. Wiles and A. Zisserman. *3D Surface Reconstruction by Pointillism.* In ECCV Workshop on Geometry Meets Deep Learning, 2018. (Cited on page 55.)

[Wiles 2018b] Olivia Wiles, A Sophia Koepke and Andrew 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. (Cited on pages 9 and 54.)

[Worrall 2016] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov and Gabriel J. Brostow. *Harmonic Networks: Deep Translation and Rotation Equivariance.* In CVPR, 2016. (Cited on pages 30, 31 and 54.)

[Worrall 2017] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov and Gabriel J. Brostow. *Interpretable Transformations with Encoder-Decoder Networks*. In CVPR, 2017. (Cited on pages 8, 30 and 54.)

[Wu 2017] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T. Freeman and Joshua B. Tenenbaum. *MarrNet: 3D Shape Reconstruction via 2.5D Sketches*. In NIPS, 2017. (Cited on page 54.)

[Wu 2020] Shangzhe Wu, Christian Rupprecht and Andrea Vedaldi. *Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild*. In CVPR, 2020. (Cited on page 98.)

[Xiancheng 2018] W Xiancheng, L Wei, M Bingyi, J He, Z Jiang, W Xu, Z Ji, G Hong and S Zhaomeng. *Retina blood vessel segmentation using a U-net based Convolutional neural network*. In Procedia Computer Science: International Conference on Data Science (ICDS 2018), Beijing, China, pages 8–9, 2018. (Cited on page 21.)

[Xiao 2004] Jing Xiao, Jin-xiang Chai and Takeo Kanade. *A closed-form solution to non-rigid shape and motion recovery*. In European conference on computer vision, pages 573–587. Springer, 2004. (Cited on page 19.)

[Xie 2016] Junyuan Xie, Ross Girshick and Ali Farhadi. *Unsupervised deep embedding for clustering analysis*. In International conference on machine learning, pages 478–487, 2016. (Cited on page 5.)

[Xu 2017] Yan Xu, Yeshu Li, Zhengyang Shen, Ziwei Wu, Teng Gao, Yubo Fan, Maode Lai, I Eric and Chao Chang. *Parallel multiple instance learning for extremely large histopathology image analysis*. BMC bioinformatics, vol. 18, no. 1, page 360, 2017. (Cited on pages 22 and 23.)

[Yao 2018] Shunyu Yao, Tzu Ming Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, Bill Freeman and Josh Tenenbaum. *3D-aware scene manipulation via inverse graphics*. In Advances in Neural Information Processing Systems, pages 1891–1902, 2018. (Cited on page 54.)

[Yu 2015] Rui Yu, Chris Russell, Neill DF Campbell and Lourdes Agapito. *Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video*. In Proceedings of the IEEE International Conference on Computer Vision, pages 918–926, 2015. (Cited on page 20.)

[Yu 2017] Lequan Yu, Xin Yang, Hao Chen, Jing Qin and Pheng Ann Heng. *Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images*. In Thirty-first AAAI conference on artificial intelligence, 2017. (Cited on page 21.)

[Yu 2018] Ye Yu and William A. P. Smith. *InverseRenderNet: Learning single image inverse rendering*. CoRR, vol. abs/1811.12328, 2018. (Cited on page 54.)

[Yuille 1991] Alan L Yuille. *Deformable templates for face recognition.* Journal of Cognitive Neuroscience, vol. 3, no. 1, 1991. (Cited on page 29.)

[Zacharaki 2009] Evangelia, I. Zacharaki, Sumei Wang, Sanjeev Chawla, Dong Soo Yoo, Ronald Wolf, Elias R. Melhem and Christos Davatzikos. *Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme.* Magnetic Resonance in Medicine, vol. 62, no. 6, 12 2009. (Cited on page 76.)

[Zhang 2002] Qi Zhang and Sally A Goldman. *EM-DD: An improved multiple-instance learning technique.* In Advances in neural information processing systems, pages 1073–1080, 2002. (Cited on page 87.)

[Zhang 2005] Lei Zhang, Sen Wang and Dimitris 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. (Cited on pages 52 and 60.)

[Zhang 2014a] Zhanpeng Zhang, Ping Luo, Chen Change Loy and Xiaoou Tang. *Facial Landmark Detection by Deep Multi-task Learning.* In European Conference on Computer Vision 2014, 2014. (Cited on pages 9 and 66.)

[Zhang 2014b] Zhanpeng Zhang, Ping Luo, Chen Change Loy and Xiaoou Tang. *Facial Landmark Detection by Deep Multi-task Learning.* In Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI, pages 94–108, 2014. (Cited on page 37.)

[Zhang 2015] X. Zhang, W. Liu, M. Dundar, S. Badve and S. Zhang. *Towards Large-Scale Histopathological Image Analysis: Hashing-Based Image Retrieval.* IEEE Transactions on Medical Imaging, vol. 34, no. 2, 2015. (Cited on page 78.)

[Zhang 2016] Zhanpeng Zhang, Ping Luo, Chen Change Loy and Xiaoou Tang. *Learning deep representation for face alignment with auxiliary attributes.* IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 5, pages 918–930, 2016. (Cited on page 44.)

[Zhang 2018] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He and Honglak Lee. *Unsupervised Discovery of Object Landmarks as Structural Representations.* In CVPR, 2018. (Cited on page 9.)

[Zhou 2016a] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva and Antonio Torralba. *Learning deep features for discriminative localization.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2921–2929, 2016. (Cited on page 10.)

[Zhou 2016b] Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qi-Xing Huang and Alexei A. Efros. *Learning Dense Correspondence via 3D-Guided Cycle Consistency.* In CVPR, 2016. (Cited on pages 31 and 54.)

[Zhou 2017] Tinghui Zhou, Matthew Brown, Noah Snavely and David 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. (Cited on pages 54 and 55.)

[Zhou 2018] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh and Jianming Liang. *Unet++: A nested u-net architecture for medical image segmentation.* In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pages 3–11. Springer, 2018. (Cited on page 21.)

[Zhu 2016] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi and Stan Z Li. *Face alignment across large poses: A 3d solution.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 146–155, 2016. (Cited on pages 9 and 17.)

[Zhu 2017a] Jun-Yan Zhu, Taesung Park, Phillip Isola and Alexei A Efros. *Unpaired image-to-image translation using cycle-consistent adversarial networks.* In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017. (Cited on page 8.)

[Zhu 2017b] Xiangyu Zhu, Zhen Lei, Stan Z Li*et al.* *Face Alignment in Full Pose Range: A 3D Total Solution.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. (Cited on pages 17, 66 and 72.)

[Zhu 2018] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum and William T. Freeman. *Visual Object Networks: Image Generation with Disentangled 3D Representation.* CoRR, vol. abs/1812.02725, 2018. (Cited on page 54.)

[Zöllei 2005] Lilla Zöllei, Erik Learned-Miller, Eric Grimson and William Wells. *Efficient population registration of 3D data.* In International Workshop on Computer Vision for Biomedical Image Applications, pages 291–301. Springer, 2005. (Cited on page 13.)