Monocular Visual Scene Analysis: Saliency Detection and 3D Face Reconstruction using GAN

Xiaoxu Cai

The thesis is submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy of the University of Portsmouth.

January 2021

Abstract

Visual scene analysis imitates the way humans perceive the outside world, which is essential for achieving computer intelligence. This thesis narrows down the scope of visual scene analysis to two fundamental tasks, namely detecting and reconstructing the object of interest in a scene. For a general scene consisting of multiple objects, it's a natural routine to screen out the most salient object first. For a human-centred scene, reconstructing the 3D geometry of the human face that occupies the central position in social communication and is highly deformable becomes one of the first priorities. Based on these two insights, the thesis studies the problems of saliency detection in a general scene and depth-to-3D face reconstruction in a human-centred scene. It deeply explores adapting the generative adversarial network – GAN that was initially proposed for image generation to solve the aforementioned problems.

For saliency detection, the thesis proposes a novel perceptual loss-guided GAN called as PerGAN. PerGAN applies a multi-scale discriminator and is trained with a perceptual loss that measures misdetection errors on the semantic feature level rather than the common pixel level of the generated saliency map. This enables an improved utilization of features across different image resolutions and those are semantically meaningful. The proposed method has been validated on benchmark datasets and outputs competitive saliency detection accuracy against the state-of-the-art.

For 3D face reconstruction from a depth image, the thesis first proposes to use the GAN to bridge the facial voxel grid and the depth data. The attention mechanism is incorporated into the GAN to regulate the learning process to weight higher on the intermediate features that are more relevant to predicting facial voxels. The resulting attention-guided GAN, or AGGAN in short, is trained and evaluated on synthesized depth images. Comparing with the previous methods that rely on a costly optimization-based 3D reconstruction process, the learning-based

AGGAN is more efficient and robust to depth images with noises and large facial poses. What's more, the use of synthetic data for training shows big potential on overcoming the shortage of depth images with 3D facial labels. Based on these results, the thesis continues to use the synthetic data for training the 3D face reconstruction network, meanwhile, incorporating unlabelled real depth images into the training procedure for obtaining a domain-adaptive reconstruction model. It employs a GAN to fill the domain gap between the synthetic and real depth images and learns a common feature embedding that is informative to both domains. The resulting reconstruction network shows a promising generalization ability to real-world depth images. Extensive experiments on mainstream real datasets demonstrate that the proposed domain-adaptive 3D face reconstruction method is competitive against the state-of-the-art.

Through developing the aforementioned algorithmic solutions to visual saliency detection and depth-to-3D face reconstruction, the thesis also gains first-hand experience on adapting GAN to different visual scene analysis tasks that are quite different from its familiar image generation task. The adaptation of GAN in this thesis ranges from binary saliency map generation, facial voxels prediction to domain alignment. This is supposed to be beneficial to propagate the GAN to a broader range of application scenarios that are not limited to visual scene analysis.

Contents

Declaration

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

List of Figures

List of Tables

List of Acronyms

3DMM	3 Dimensional Morphable Model
ACGAN	Auxiliary Classifier Generative Adversarial Network
AGGAN	Attention Guided Generative Adversarial Network
AR-GAN	Auxiliary Regressor Generative Adversarial Network
BEGAN	Boundary Equilibrium Generative Adversarial Network
BFM	Basel Face Model
CE	Cross Entropy
CGAN	Conditional GAN
CNN	Convolutional Neural Network
DCGAN	Deep Convolutional GANs
EMD	Earth Mover Distance
FCN	Fully Connected Network
GAN	Generative Adversarial Network
GT	Ground Truth
ICP	Iterative Closest Point
IoU	Intersection-over-Union
LPGAN	Laplacian Pyramid Generative Adversarial Network
MAE	Mean Square Error
MC	Marching Cubes
MLP	Multi-Layer Perceptron
NICP	Non-rigid Iterative Closest Point
PerGAN	Perceptual loss guided Generative Adversarial Network
PSR	Predicted Salient Region
ReLU	Rectified Linear Unit
SAGAN	Self-Attention Generative Adversarial Network
SLIC	Simple Liner Iterative Clustering
SPSR	Screened Poisson Surface Reconstruction

TOF	Time Of Fight
TPS	Thin Plate Splines
TSR	Target Salient Region
VGG-16	Visual Geometry Group’s Network with 16 layers

Acknowledgements

Foremost, I would like to greatly thank my supervisor Prof. Hui Yu for providing this wonderful PhD opportunity and giving me constant support over the past four years. I would also like to thank my master's instructor, Prof. Junyu Dong, for his encouragement, support, patience and everything I have learned from him.

To work and to party with my colleagues were marvellous. I would like to thank Miss Qiongdan Cao, Mr Yifan Xia, Mrs Xin Liu, Mr Muwei Jian, Mr Hao Fan, Miss Jianyuan Sun, Mr Yiming Wang, Mrs Weihong Gao, Mr Martin Kearl and Mr Shu Zhang. We have many good moments together in Portsmouth.

Last but not least, I dedicate this thesis to my husband Jianwen Lou, my parents and grandparents, with love. They are staying by my side all the way through this journey.

Dissemination

Cai, X., Yu, H., Lou, J., Zhang, X., Li, G., & Dong, J. (2021). 3D facial geometry recovery from a single depth view with attention-guided GAN. in submission.

Cai, X., Lou, J., Dong, J., & Yu, H. (2021). Domain adaptive single depth image 3D face reconstruction. in submission.

Cai, X., Lou, J., Dong, J., & Yu, H. (2020). Real-time 3D facial tracking via cascaded compositional learning. *IEEE Transactions on Image Processing*, under review.

Cai, X., Jian, M., Dong, J., Chen, R., Stevens, B.& Yu, H. (2020) Perceptual loss guided-GAN for saliency detection. *IEEE Transactions on Emerging Topics in Computing, under review*

Cai, X., and Yu, H. (2018) Saliency detection by conditional generative adversarial network. *In Ninth International Conference on Graphic and Image Processing (ICGIP 2017)*, vol. 10615, p. 1061541.

Lou, J., **Cai, X.**, Wang, Y., Yu, H., & Canavan, S. (2019). Multi-subspace supervised descent method for robust face alignment. *Multimedia Tools and Applications*, 78(24), 35455–35469.

Chapter 1

Introduction

1.1 Background

It's not what you look at that matters, it's what you see.

H.D. Thoreau (1817-62)

Glancing a new scene for no more than 250ms, we humans are able to describe the scene at the semantic level (M.C. Potter, 1976) such as name a few salient objects or recognize a human subject's facial expression. This capability is known as visual perception, an essential piece of human intelligence, that shapes the way we interact with the complicated outside world. From this perspective, without visual perception, purely human-like intelligence can never be envisaged for computers. This motivated an interesting and important research branch in computer vision, which is termed as visual scene analysis.

Visual scene analysis is a computer's version of visual perception, where the input stimulus becomes a digitized image and the perception is portrayed as image data processing. To analyse a visual scene, the computer must first identify objects and relationships between objects, labelling each correctly. This derives a more fundamental step which is about detecting the object of interest in an intricate scene or reconstructing the target object to recover its critical features as much as possible, thereby facilitating the aforementioned recognition task. At this stage, what act to do depends on the contents of the visual scene. This thesis divides the visual scene content into two categories, namely general and human-centred. For the general scene (see a. in Fig.1.1), it could consist of multiple objects in various categories without a unified focus, such as the traffic scene filled with vehicles, pedestrians, road signs, buildings and plants etc. In this case, the first priority of scene analysis would be finding out the most salient object. For the human-centred

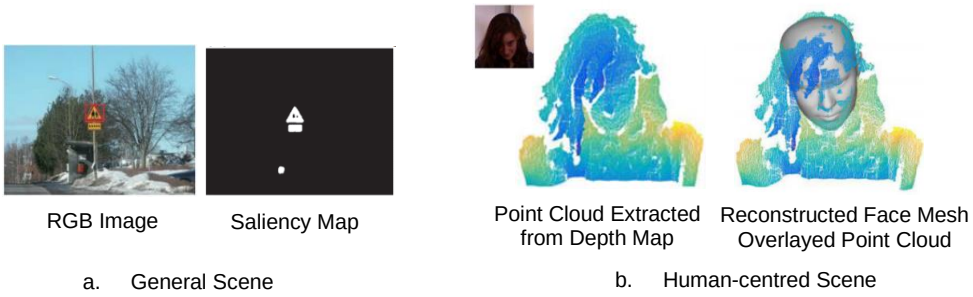

Figure 1.1: The examples of visual scene analysis.

scene, our attention will mostly be attracted by the human face as it plays a central role in social communication (Ekman & E. I. Rosenberg, 1997). Due to its anatomical diversity and high sensitivity, human face is rich with delicate features such as facial expressions and very informative, resulting that even a slight difference on its appearance could greatly affect the perception of the face. To attain those crucial facial features for further analysis, a promising way is to reconstruct the face as close as it was presented in the 3D physical space. This is particularly important when the input image is imperfect with noise, occlusion or observes a large head pose which makes facial features less identified. Based on these insights, the thesis addresses visual scene analysis by researching on two principal questions which are saliency detection for the general scene and 3D face reconstruction for the human-centred scene (see b. in Fig.1.1). More specifically, the research work concentrates on the most common setup – monocular visual scene capture, which in this thesis is a RGB image for saliency detection and a depth image for 3D face reconstruction.

In recent years, the field of visual scene analysis has been booming since the use of deep learning (Deng et al., 2019; N. Liu et al., 2018; J. Pan et al., 2017; L. Wang et al., 2015; Zhong et al., 2020). Some state-of-the-art deep learning models achieve near or even super-human performance. Such a big success is attributed to the rapid development of deep learning techniques, which can be split into two main streams: one adopts the convolutional neural network (CNN) (Simonyan &

Zisserman, 2014) to combine feature learning and output prediction into a hierarchical learning process, the other employs the generative adversarial network (GAN) (Goodfellow et al., 2014) to estimate generative models via an adversarial process. While the former method has been applied in a variety of visual detection, recognition and reconstruction tasks, the latter one is still centred around the image generation task. It is thus interesting to extend GAN to its unfamiliar tasks such as detection and reconstruction, a problem also remains largely unexplored. This motivates the thesis to adapt the GAN to solve the aforesaid saliency detection and 3D face reconstruction problems.

1.2 Problems and Challenges

The problems and challenges this thesis encountered with are twofold. First, adapting the GAN to different visual scene analysis tasks itself is nontrivial since the output modality varies across tasks. For example, in 3D face reconstruction from a single depth image, the typical output is a 3D face represented in voxels or a 3D point cloud which differs a lot from the 2D image that GAN normally generates. To handle the high dimensionality of the 3D data, a straightforward solution is to increase the number of convolutional kernels in GAN, but this will accordingly introduce much more difficulties into the network training. Second, each visual scene analysis task has its own issues, which is elaborated as follows:

1. Given a RGB image of a general scene, most existing studies (Hou et al., 2017; N. Liu et al., 2018; L. Wang et al., 2015; T. Zhao & Wu, 2019) in the field of saliency detection apply the CNN coupled with a pixel-wise loss function for labelling salient objects. Whereas this kind of approach has reported promising detection results on benchmark datasets, the use of the pixel-wise loss function is arguable. That loss function treats each pixel independently without considering the holistic semantic information the pixel and its neighbours or all the pixels as a whole that convey. The semantic information such as those regarding the shape is

however critical in determining an object's salient degree. As a result, it is essential to design a new loss function that effectively incorporates the holistic semantic information a salient region delivers to guide the saliency detection network training.

2. When moving into the task of 3D face reconstruction from a single depth image of a human-centred scene, it can be found that the learning-based approach has rarely been studied. Previous methods (Martin et al., 2014; Newcombe et al., 2011) are generally optimization-based, which reconstruct 3D face by fitting the input depth image with a parametric function where the 3D facial shape is either explicitly represented by a triangulated facial mesh(Amberg et al., 2007) or implicitly modelled by its level-set (Lorensen & Cline, 1987a). Those methods rely on a costly optimization procedure, let alone they are very sensitive to image noises, occlusions, big head poses that are common features of real-world depth images. As an alternative, learning-based approach builds the 3D face reconstruction function through an offline training process that learns the depth-3D mapping from a corpus comprising numerous depth images. During the online inference, the approach only needs to call the pre-built reconstruction function, which is quite efficient. The leveraging of a large-scale training corpus also makes the reconstruction function more robust to the aforementioned image imperfections. However, as its inherent deficiency, learning-based approach is data-intensive. This poses a big challenge on developing the method, especially for the fields like 3D face reconstruction in which collecting depth images and their 3D facial shape labels is very expensive and laborious. In a word, it is very timely and of great potential to develop a robust learning-based depth-to-3D face reconstruction method, but should cope with the training data problem.

1.3 Contributions

Targeting at tackling the aforementioned problems and challenges, the thesis develops three novel GAN-based networks for robust saliency detection and depth-to-3D face reconstruction. Its main contributions are as follows:

1. It proposes a perceptual loss-guided GAN - PerGAN for saliency detection from a single RGB image. Different from the conventional loss that penalizes pixel-wise misdetection, the perceptual loss evaluates the detection error on the semantic feature level of the generated saliency map. This guarantees a full utilization of high-level semantic information such as the object shape that is important in determining the saliency. PerGAN is further enhanced with a multi-scale discriminator and a two-level feature to saliency mapping that follows the coarse-to-fine strategy. Comparing against previous methods, the proposed PerGAN has shown improved detection accuracy on several widely-accepted benchmark datasets.

2. It develops an attention-guided GAN - AGGAN to predict facial voxels from a single depth image. To the best of our knowledge, this is the first work of its kind that utilizes GAN to construct the mapping between the depth data and the 3D facial geometry. The incorporation of the attention mechanism into GAN has demonstrated to be effective on handling facial voxels' spatial relationships both theoretically and experimentally. What's more, the thesis proposes to synthesize depth images and the corresponding 3D faces for training AGGAN, which provides an efficient solution to overcome the shortage of labelled 3D facial data. In contrast with the-state-of-the-art optimization-based methods, the proposed AGGAN exhibits much higher reconstruction precision on synthesized depth images with noises and large head poses. Whereas AGGAN is trained and evaluated on the synthetic data, it showcases the potential of generalizing to real-world depth images.

3. Inspired by AGGAN, the thesis further proposes a novel domain-adaptive 3D face reconstruction network. The proposed network accepts labelled synthetic

depth images and unlabelled real depth images that belong to different domains for training, and uses GAN to fill the synthetic-real domain gap while learning a common feature embedding. The resulting common embedding is domain-adaptive, thus enabling accurate 3D face reconstruction for both synthetic and real depth images. Extensive experiments on mainstream real depth image datasets demonstrate that the proposed method is competitive with the state-of-the-art 3D face reconstruction approaches.

4. The thesis also presents three successful adaptations of GAN beyond its familiar image generation task. It extends GAN to binary saliency map generation with a perceptual loss, 3D facial voxels prediction by incorporating the attention mechanism, and synthetic domain and real domain alignment, thereby strongly demonstrating the potential of GAN and providing first-hand experience on adjusting GAN to different visual scene analysis tasks.

1.4 Outline

The rest of the thesis is structured as follows:

Chapter 2 - Literature Review:** This chapter provides an extensive review on the two targeted visual scene analysis sub-fields, namely visual saliency detection and depth-based 3D face reconstruction, and recaps the development of GAN in recent years.

Chapter 3 - Perceptual Loss-Guided GAN for Saliency Detection: It develops a novel perceptual loss-guided GAN - PerGAN for detecting salient objects from a general-scene image. Instead of employing a pixel-wise difference loss to optimize the detection network, PerGAN applies a perceptual loss built upon the semantic feature level of the generated saliency map. This innovative design is supposed to utilize the key semantic information of a salient object such as its shape when performing detection. To further strengthen the predictive capability of PerGAN, the chapter equips the GAN with a multi-scale

discriminator and a two-level coarse-to-fine feature to saliency mapping. Experimental results on six benchmark datasets show that the proposed PerGAN outputs higher saliency detection accuracy than previous methods.

Chapter 4 - 3D Facial Geometry Recovery from a Depth View with Attention Guided GAN: This chapter proposes an attention-guided GAN - AGGAN to solve the problem of 3D face reconstruction from a single depth image. Specifically, AGGAN encodes the 3D facial geometry within a voxel space and enhances the GAN with the attention mechanism to model the ill-posed 2.5D depth-3D mapping. In contrast to previous works which are normally based on a costly optimization-based fitting procedure, AGGAN efficiently predicts a dense 3D voxel grid of the face from a single unconstrained depth view via a pre-trained inference network. To train and evaluate AGGAN, the chapter proposes to synthesize depth images and their ground-truth 3D face labels. Both qualitative and quantitative comparisons on synthetic depth images show that AGGAN recovers a more complete and smoother 3D facial shape, is able to handle a much wider range of view angles and resists more to image noise than those optimization-based methods. The experimental results also indicate that the AGGAN model trained with synthetic data has the potential of generalizing to noisy real depth images.

Chapter 5 - Domain Adaptive Single Depth Image 3D Face Reconstruction: It develops a domain-adaptive 3D face reconstruction network that works on both synthetic and real depth images. The proposed method requires only synthetic and unlabelled real depth images for training, while the trained network can generalize well to real images captured with commodity depth sensors. Its success can be mostly ascribed to a GAN that effectively regulates the learning of a feature embedding shared by synthetic and real depth images and makes the learned embedding domain adaptive. Both quantitative and visual comparisons on public datasets indicate that the proposed method produces competitive 3D face reconstruction results against the state-of-the-art methods.

Chapter 6 - Conclusions: This chapter summarises the thesis with an in-depth discussion on its contributions and the future work.

Chapter 2

Literature Review

Human's strong ability of describing the scene in a quick glance is known as visual perception, shaping the way human interacts with the complex world. Visual scene analysis is a computer's version of visual perception, which is a fundamental step for many recognition and classification tasks.

As mentioned in chapter 1, this thesis divides various images into general scene and human-centred scene. For general scene, the prior task of scene analysis is to find out the most salient object(s), a process named as saliency detection. In terms of another scene, 3D face reconstruction from the single depth image that is robust to illumination changes and occlusions, is a fantastic way for human-centred scene analysis since face is informative. The detailed review of saliency detection and 3D face reconstruction will be provided in this chapter. In addition, the development of basic framework-GAN is demonstrated as well.

The content of this chapter is organized as follows. Section 2.1 summarizes both the traditional and deep learning methods designed for detecting the salient objects from general scene. Section 2.2 concludes earlier and current models used for 3D face reconstruction from depth view for human-centred scene analysis. Section 2.3 introduces different variants of GAN which is the basic deep framework in this thesis. Section 2.4 concludes the literature review in both salient object detection and 3d face reconstruction.

2.1 Salient Object Detection

Salient object detection targets at locating pixels or regions which catch human attention most in a scene, is a fundamental and necessary step in visual scene

analysis. The outcomes provide primary visual cues to tasks such as object and activity recognition, thus facilitating a deep understanding of the target environment. Over the past two decades, the problem of saliency detection has been extensively studied in computer vision community with many approaches developed. Those approaches can generally be split into two categories, namely traditional methods and deep learning methods. Traditional methods focus on extracting handcrafted low-level features from images for saliency detection. In contrast, deep learning technologies instead utilize high-level semantic features and dramatically improve the saliency detection performance. This section will recap the representative studies regarding these two kinds of methods.

2.1.1 Traditional Methods

Most traditional methods first identify salient subsets from the given image, then integrate them to form complete salient objects. Salient subsets could be blocks which are image patches in a uniform and regular shape such as rectangle, or regions consisting of conceptually homogeneous image patches confined within sealed boundary. According to different salient subset types, traditional methods can be divided into two categories: block-based model and region-based model. The representative framework of each category can be seen in Fig.2.1

a) Block-based Model

Blocks-based models for salient object detection primarily were developed in the early stages. The seminal work which fosters salient object detection in a variety of communities was proposed by Itti et al. (Itti et al., 1998) in 1998. It computes a saliency map from an input image by incorporating the feature maps of color, intensity and orientation at multiple scales using centre-surround operation

Due to the lack of prior knowledge of salient objects, multi-scale feature contrast is often adopted for robust saliency detection. Ma et al. (Ma & Zhang, 2003) propose to extract the local contrast information of color, texture and shape at different scales, and then derive the saliency map from extracted maps by fuzzy

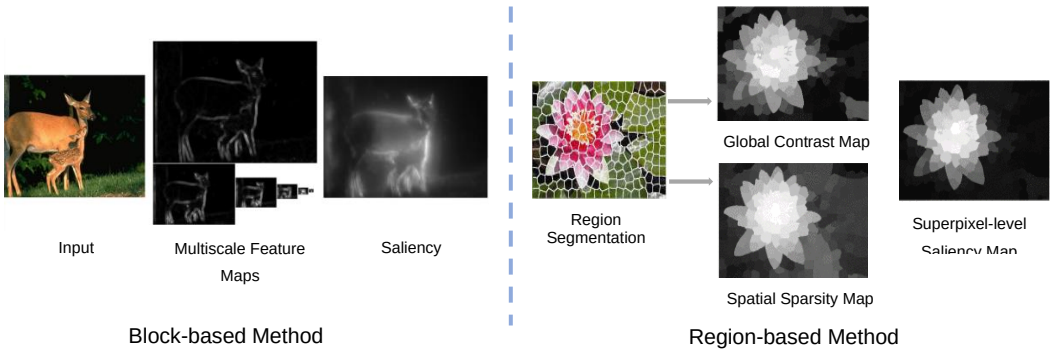

Figure 2.1: Examples of traditional salient object detection method.

growing approach. Observing more feature representations improve the performance of locating salient regions at the price of slowing down the calculation, Frintrop et al. (Frintrop et al., 2007) propose to compute the saliency maps by weighted combination of various feature maps extracted from different dimensions. The weight function is like a normalization operator presented in (Itti et al., 1998), promoting rare conspicuous maps while suppressing repetitive or similar maps. Unlike normalized weights, conditional random field is utilized to balance the multiple scale features of color and context to form the salient object (T. Liu et al., 2010). Valenti et al. (Valenti et al., 2009) propose to detect salient objects by exploring gradient slope information. The saliency maps generated using mentioned methods always are with high-contrast edges, to rise the response of salient object, Rosin et al. (Rosin, 2009) propose to produce multiple response maps with blobs from detected edge maps. It works for very simple salient objects.

Based on aforementioned methods, the contrast of edge is more intensive than that of salient objects in most cases, which makes the edge to be misled as salient objects. Additionally, it is hard to maintain the edge of salient object well, especially the high-resolution blocks are used. To resolve these issues, a series of region-based models are proposed to compute regional saliency map.

b) Region-based Model

Region-based models take the structure of region into consideration and regard the homogeneous region as the basic element instead of pixels, which normally provide the potential to design the efficient and fast saliency detection. Generally,

it segments an input into regions first, and then calculate visual cues of each region to form final saliency map. The homogeneous regions can be produced by popular mean-shift (Comaniciu & Meer, 2002), graph-based segmentation (Felzenszwalb & Huttenlocher, 2004), SLIC (Achanta et al., 2012) or Turbopixels (Levinshtein et al., 2009). In this section, these segmentation methods and the corresponding representative works will be introduced in detail.

Mean shift segmentation (Comaniciu & Meer, 2002) is an advanced and vertisale clustering technique. For each data point, mean shift defines a window around it and computes the mean value of point representations. Then it shifts the centre of window to the mean value and repeats the process until it converges. The initial trial of mean-shift segmentation-based saliency detection model is developed by Liu et al.(F. Liu & Gleicher, 2006). They propose to generate salient regions by segmenting the image based on the multi-scale color contrast of each pixels to enhance the saliency maps. Color (Z. Liu et al., 2010), texture and location (Q. Wang et al., 2012), spatial distribution (Ren et al., 2013) are also adopted to produce salient regions for saliency detection.

Graph-based segmentation (Felzenszwalb & Huttenlocher, 2004) generally represents the problem in term of a graph consisting of vertices and edges, and then cut the graph into pieces. The vertex corresponds to a pixel in the image and the edge refers the connection between neighbouring pixels. A weight is associated with each edge based on some property (eg. color, motion, location or difference in intensity) of the pixels that it connects. Jiang et al. (H. Jiang et al., 2011) propose to use the multiple scale context contrast to set edge weight and infer the regional saliency cues. Besides, the shape prior is extracted to be aligned with salient contour for fine saliency map estimation. Continuing in the same direction, color histogram (Cheng, Mitra, Huang, Torr, et al., 2014), spatial distribution (Z. Jiang & Davis, 2013), texture distinctiveness (Scharfenberger et al., 2013) and edge density (Jia & Han, 2013) are utilized to reconstruct saliency map.

SLIC(Achanta et al., 2012) is short for simple liner iterative clustering, which can be regarded as a special graph-based method. It uses a five-dimensional vector

which stores each pixel's positions in the CIELAB colour space and the image plane, and adopts the K-means to cluster pixels into super-pixels by evaluating two pixels' similarity in the aforementioned 5D space. SLIC has been applied in many works (B. Jiang et al., 2013; Singh et al., 2018; V.Singh et al., 2020; L. Wang, Jiang, et al., 2019; C. Yang et al., 2013) result from the fast calculation. After segmentation, Jiang et al.(C. Yang et al., 2013) propose a manifold ranking method to calculate both foreground and background visual cues of each super-pixel in CIE LAB color space. Similarly, both boundary and centred salient information are used to form regional cues via absorbing markov chain (B. Jiang et al., 2013). Recently, Gaussian mixture model (Singh et al., 2018; V.Singh et al., 2020) and progressive graph ranking method (L. Wang, Jiang, et al., 2019) are adopted for calculate the regional saliency.

Turbo-pixels(Levinshtein et al., 2009) segments an image into a lattice-like structure of compact regions (super-pixels) compactness constraint. Recent works (L. Xu et al., 2014; L. Zhang et al., 2018) adapted it to produce regular and compact super-pixels. After segmentation, the same features including color, texture, spatial distribution used in above methods are utilized to extract visual cues.

2.1.2 Deep Learning Methods

With the promising performance in image classification task(Krizhevsky et al., 2012) in 2012, deep learning-based techniques have been springing up for a variety of vision tasks. Currently, deep learning-based model is becoming the mainstream direction of salient object detection. According to the architecture of deep network, we categorize the deep salient object detection models into four classes, namely MLP-based, FCN-based, hybrid network-based and GAN-based. Fig.2.2 displays the simple frameworks of each model.

MLP-based model means to map the saliency cues of each super-pixel via MLP from their deep features extracted from CNN. As an initial attempt, He et al. (He et al., 2015) present to regress the salient value of each segmented region by feeding the deep features extracted from their color contrast to MLP-based model.

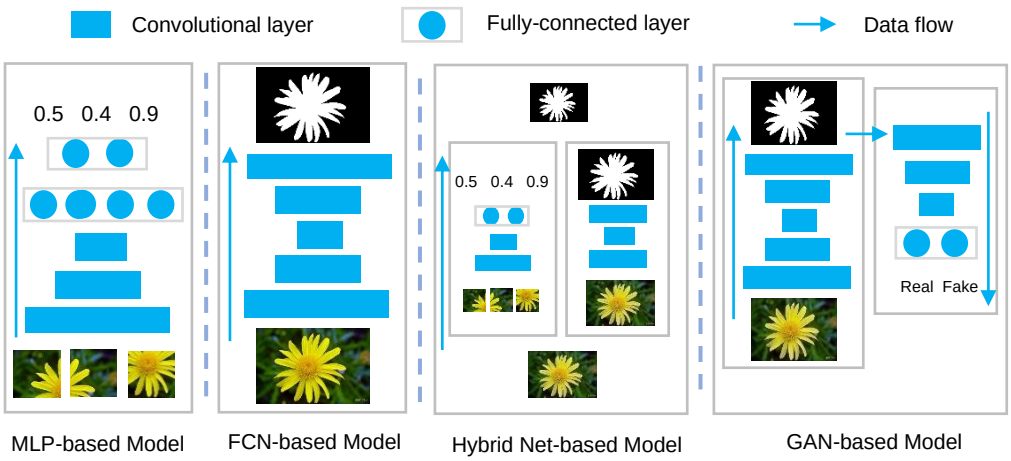

Figure 2.2: Simple examples of deep salient detection model.

Kim et al.(Kim & Pavlovic, 2016) propose a two-stage framework consisting of two MLP-based model to predict saliency map. During the first stage, unlike the saliency score prediction, the salient shape of each region is classified to produce a coarse saliency map. The shape class of each image batch is assigned through the selective research and clustering method. To guarantee the boundary information, the coarse map is refined in second stage by predicting saliency score of hierarchical segmented regions. Similarly, Zhao et al. (R. Zhao et al., 2015) propose a two-branch architecture which explores both local and global context via two MLP-based subnets. Compared with traditional salient object detection methods, MLP-based models improve the performance dramatically. But it ignores the global information of image since it relies on the segmented regions. Besides, One-by-one processing of super-pixels is quite time-consuming.

To address the shortcomings of MLP-based model, FCN-based model is proposed for pixel-wise saliency estimation is an end-to-end manner. Zhang et al.(P. Zhang, Wang, Lu, Wang, & Yin, 2017) designed a single-stream network based on encoder-decoder architecture for saliency detection. The developed network predicts the pixel-wise probabilistic map which highlights salient subject from the whole input image directly. It is the classical architecture of FCN-based model for saliency detection. To learn the multiple scale feature representations,

Li et al. (G. Li et al., 2017) proposed a multi-stream framework consisting of three single-stream networks, in which three scales of outputs are mapped from corresponding inputs and fused together to produce saliency map. Instead of reconstructing the saliency map through the last layers' output of single-frame network, Hou et al.(Hou et al., 2017) propose to fuse the outputs of each layer to generate the saliency map. Compared with the classical architecture, the deep side-outputs assist shallow layer to identify salient regions, and shallow side-outputs enrich details for deep layers as well. In (N. Liu et al., 2018), short connections are established between the output of deep layers and symmetric shallow layers based on a single-stream network. The combined features are refined recurrently in a top-down pathway to enhance the final output. In a follow-up work, attention model is integrated with classical model to select the most efficient channels and spatial information for saliency detection in (X. Zhang et al., 2018).

Considering the super-pixel level and pixel-wise saliency are predicted separately, many researchers propose to predict edge-preserving outcome with multi-scale context via hybrid network-based mode combing MLP- and FCN-based models together. Wang et al. (L. Wang et al., 2015) both local and global feature representation for more accurate salient cues prediction. A local pixel-wise estimation is established based on single-stream FCN-based subnet. Be combined with the geometric information of original input, the output of local branch is segmented and further fed to MLP-based global research branch to estimate the regional saliency values. Finally, the top K candidate regions are weighted and summarized to produce final saliency map. Similarly, (G. Li & Yu, 2016; Tang & X. Wu, 2016) also adopted hybrid network-based models for salient object detection.

GAN-based model can be regarded as a special hybrid network-based model, in which the MLP-based model deals with the whole input image instead of the segmented regions. As we all know that the training of GAN is a competing game between generator and discriminator. For GAN-based salient object detection model, the generator used for saliency map generation is based on FCN-based

model and the discriminator used for evaluate the generated results is based on binary MLP-classifier. The model has been adopted in many works (Cai & Yu, 2018; Ji et al., 2018; Mukherjee et al., 2019). Cai et al. (Cai & Yu, 2018) and Ji et al. (Ji et al., 2018) designed the generator by incorporating classical FCN-based model and short connections together for saliency detection. Based on this, (Mukherjee et al., 2019) propose to predict salient objects via minimizing cycle consistency loss. Additionally, a generator integrating capsule-module with FCN-based model is developed by (C. Zhang et al., 2019) for salient cues generation.

2.2 3D Face Reconstruction

Monocular 3D face reconstruction aims to recover 3D facial geometry from a single image. It is essential for face-centred visual scene understanding as face conveys message, emotion and intent, and occupies key place in human visual perception (M.Zollhöfer et al., 2018). RGB-based face reconstruction that relies on the texture and colour information provided by 2D images has been well studied (Feng et al., 2018; Gecer et al., 2019; Jackson et al., 2017; F. Liu et al., 2019). However, the performance cannot be guaranteed when the input image was captured under a poor lighting environment. Different from RGB image, depth image recording the distance between face surface and viewpoint is robust to illumination changes and occlusion. This thesis mainly focuses on reconstructing 3D face from the depth image.

This section is organised as follows. Section 2.2.1 describes the popular depth senor - Microsoft Kinect used for collecting depth data. Section 2.2.2 summarizes the widely used 3D face representations. Section 2.2.3 and 2.2.4 reviews the traditional and deep learning-based 3D face reconstruction method separately.

2.2.1 Kinect Sensor

Microsoft Kinect is an RGB-D sensor which equipped with both RGB and depth cameras. It is originally released as gaming peripheral allowing users to naturally interact with a computer. Its appearance has revolutionized the way people play games and how they experience entertainment. Recently, Kinect's impact has extended far beyond the gaming industry due to its low price and robust performance. It is a popular sensor nowadays, and many researchers utilize it to capture 3D face data (Baltrušaitis et al., 2012; Cao et al., 2013; Fanelli et al., 2011) for research.

Kinect sensor not only provides a size of 640×480 colour image from an RGB camera, but only offers the same size of depth map from depth sensor. The depth map records the distance between the surface of object and sensor. Currently, four versions of Kinect are available. But in this thesis, we just focus on Kinect V1 and V2 which has been used for depth image collection. Kinect V1 computes the distance based on the active structure light technology. To be specific, it projects a known pattern onto the scene and calculate the distortion of pattern to estimate the distance of points. Similarly, Kinect V2 is also based on the active ranging named time of flight (TOF) technology, but the emitted energy is different. With the known speed of light, TOF usually computes the duration time between the emission of a laser light and its return to the sensor after being reflected by a subject to infer the distance. To this end, the Kinect is with an energy projector, a depth camera and a RGB camera. An example structure is showed in Fig.2.3.

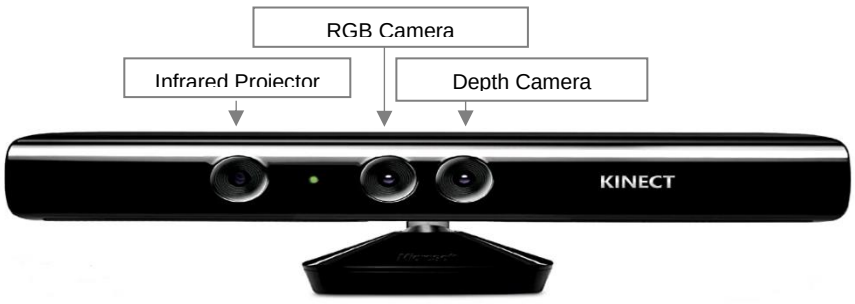

Figure 2.3: The structure of Microsoft Kinect V1.

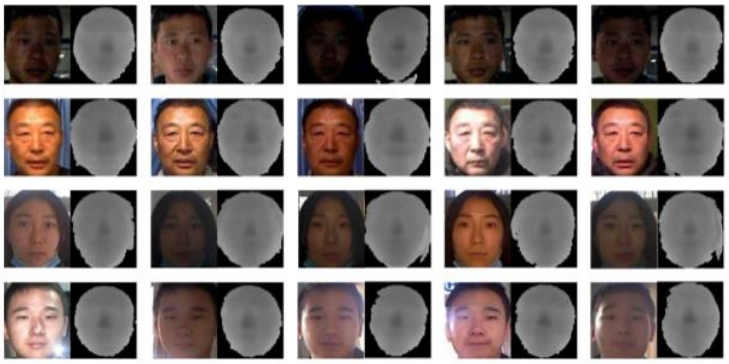

Figure 2.4: The samples captured by RGB and depth cameras (J. Luo et al., 2019).

Compared with traditional RGB cameras which are sensitive to light conditions, the depth camera (Kinect sensor) is robust to the illumination changes and occlusions. No matter how poor the lighting is, the depth camera can well capture the 3D Information. The examples of RGB image and corresponding depth map collected in different light condition can be seen in Fig.2.4.

2.2.2 3D Face Representation

In this section, we spotlight the 3D/2.5D face representations that widely used in deep learning nowadays. The example of each representation is showcased in Fig. 2.5.

a) Point Cloud

Point cloud is a group of vertices in world space. Each vertex is constituted with x, y and z coordinates, which record the physical location of the vertex in 3D space. It is a simple and direct data representation. But it is not in a uniform distribution. The region closed to view point usually is dense, while the far area is sparse. This feature results in irregular and unordered data format.

Deep learning methods are good at dealing with the regular data format like signal sequences, images, videos. Therefore, the irregular data format makes point cloud not a desired data format for deep learning until PointNet (Qi et al., 2017) was proposed. Instead of regarding the point cloud as a vector, PointNet ignores

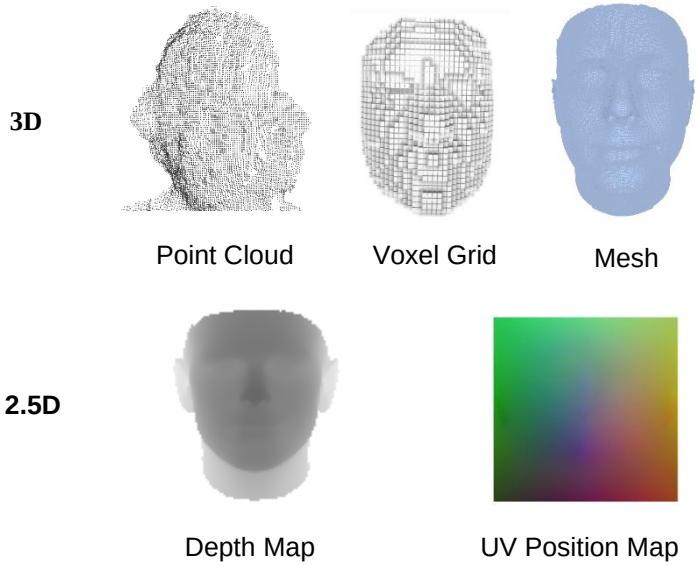

Figure 2.5: The examples of different face representations.

the data structure and processes each point separately. Later, Liu et al. (F. Liu et al., 2019) present to model the regular face point cloud from the unorganised data with the help of PointNet. Lombardi et al. (Lombardi et al., 2018) propose to predict the point cloud of face directly using variation autoencoder. What needs to be mentioned here is that the point cloud they used is extracted from synthetic data and is in regular data format.

b) Voxel Occupancy Grid

Voxel occupancy grids are the spatial representation of 3D face and environment. The occupancy grid is typically acquired through voxelization (Cohen-Or & Kaufman, 1995; Karabassi et al., 1999). It converts a continuous geometric shape into a set of voxels with each depicts the occupancy state of the corresponding point on the 3D face. Generally, the voxel stores a Boolean occupancy status 0 or 1, where 0 means the background cells while 1 reflects the face region cells, so it does not record the physical location. But it is easy to get the geometry of the object/face from voxel occupancy grid by inferring the location of a voxel based on its location relative to others.

Compared with irregular and unordered point cloud, occupancy grid is a regular data format, which contributes to a popular 3D data representation for deep neural networks (Jackson et al., 2017; B. Yang et al., 2017). Jackson et al. (Jackson et al., 2017) proposed to reconstruct 3D face in voxel space. Many existing works also applied occupancy grid to represent the object for 3d object reconstruction can refer to the comprehensive survey (B. Yang et al., 2017). In addition, the voxelized data (S.Shi et al., 2020; C. Wang, M.Cheng, et al., 2019) have also been used in many 3D object segmentation and recognition tasks. Fig.2.5 shows a sample of face in voxel space. It can be seen obviously that the surface of face is non-smooth, which can be resolved by increasing the resolution of grid. But the high resolution creates a huge unnecessary demand for computer storage and computational cost. That is why voxel-based representation is not suitable for representing high-resolution data.

c) Mesh

A mesh is a collection of vertices, edges and faces that defines the topology of a polyhedral object. Different from the unordered vertices in point cloud, each vertex in mesh is assigned with an index which is helpful for building edges and faces. The face consisting of the connection list (edges) and index (vertices) reports how object coordinates exist in 3D space. The face mesh sample can be seen in Fig.2.5.

It is also a popular data format for 3D face (object) reconstruction and has been widely used to deform from pre-defined mesh templates, limiting them to fixed mesh topologies. Specifically, (Tan et al., 2018; Yuan et al., 2020) propose to deform the facial meshes by using edge and normal information which are the distinctive attributes of the mesh. In contrast, Gkioxari et al. (Gkioxari et al., 2019) propose to predict the accurate mesh of object directly by fusing multiple 3D shape representations.

d) 2.5D Image

There are two kinds of 2.5D images which are able to store 3D face information. One kind is depth map recording the depth of visible points, and the other is UV

position map storing the location of all points. This section will separately introduce depth map and UV position map in detail at this section.

Depth map is an image or image channel that stores data with respect to the distance between a viewpoint and the surfaces of a face. It can be captured by 3D scanner or synchronized using Z-buffer from 3D mesh. Each pixel value in depth map only depends on the distance between sensor and face, which makes depth map robust to illumination changes and keep rich geometric information than RGB pictures. (Y. Guo et al., 2018; X. Zhu et al., 2017) propose to use a 3-channel position (depth) map storing x, y and z coordinates to represent 3D face and have got a better performance than single-channel depth image in reconstruction. The main deficiency of depth map is that only information of visible points is kept. Recovering the complete face geometry from a single depth map with large pose is still a difficult problem due to massive information missing.

UV map is the flat representation of the 3D model used for warping texture. It offers the accurate correspondence with the 3D face vertices and resolve the challenging occlusion problem resulted from head poses. Inspired by this phenomenon, Feng et al. (Feng et al., 2018) present a UV position map, which is a UV map recording the position information of 3D face and providing dense correspondence to the semantic meaning of each point. Recently, this new data format has attracted increasing interest in 3d face modelling (Bagautdinov et al., 2018). The only downside is that the UV map needs to be designed by artist manually.

2.2.3 Traditional Methods

In the last few years, many traditional methods had been developed for 3D face reconstruction from depth maps. Existing methods (Donne & Geiger, 2019; Fang et al., 2019; Newcombe et al., 2011) were able to obtain the promising 3D shape by fusing multiple views of depth maps. However, it is not applicable for the practical application because of the complexity of multiple depth maps acquisition. This thesis mainly focuses on monocular 3D face reconstruction. As mentioned in

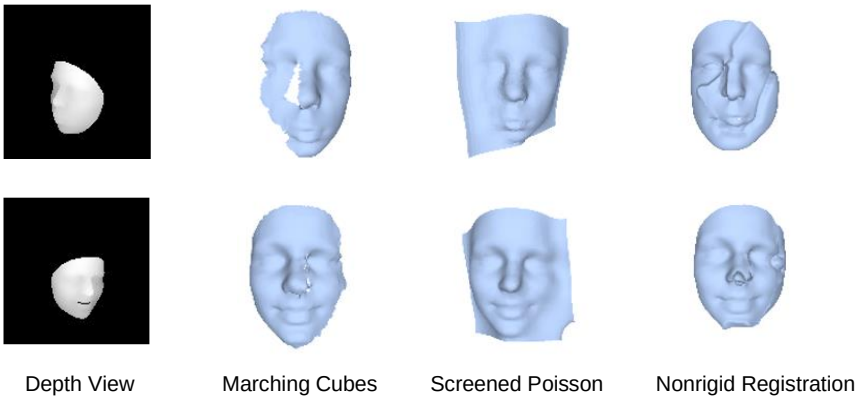

Figure 2.6: The visual results reconstructed by traditional methods.

section 2.2.2, depth image recording the depth information is a kind of 3D data representation. With the given intrinsic matrix of camera, it is easy to get the point cloud according to the process of projection. The problem is thus converted into recovering the object/face surface from the point cloud projected from a single depth view. The recovering methods can be roughly classified into two categories: zero set method and non-rigid registration. Both Marching Cubes and Screened Poisson are zero set method. Some recovered examples are reported in Fig.2.6.

Zero-set methods usually reconstruct the surface by designing a distance function which assigns to each point a signed distance to the surface. The definition of the appropriate function with a zero value for the sampled points and different to zero value for the rest can be regarded as polygonal representation of the object. In terms of the output, the methods can be broadly split into two parts, generating either a discrete surface, or an implicit function. For the former method (discrete surface), marching cubes (Lorensen & Cline, 1987b) is a typical and representative work. It extracts the surface of object/face by finding intersections between the input points and cubes of fitted grid, which voxelized the unordered input. Newman et al. (Newman & Yi, 2006) have extended it to improve the performance. The latter method (implicit function) typically utilizes the knowledge of the exterior and interior of the surface with an implicit function for reconstruction. Taking poisson reconstruction (Kazhdan et al., 2006)as an example,

the implicit function is defined using poisson equation to adjust the input point set. Functions such as moving least squares, basic functions with local support (C.Walder et al., 2006), poisson equation have been proposed to define the implicit function as well. Loss of the geometry precision in areas with extreme curvature, i.e., corners, edges is one of the main issues encountered. Moreover, pretreatment of information, by adopting some kind of filtering technique, also impacts the definition of the corners by softening them.

Unlike aforementioned methods, the reconstruction problem can be converted into the 3D non-rigid registration task, which aims to transfer the change in shape exhibited by raw input deformation onto the shape prior. It can be considered as warping the shape reference onto the raw point cloud. A typical solution is to register a facial template mesh to the given depth view using a deformation model based on smooth local affine transforms. Reliable correspondences which find the semantic mapping between the template and 3D points are required to indicate which parts of the two shapes should deform similarly. False correspondence can cause strong shape distortions that are inconsistent with the desired facial shape. Amberg et al. (Amberg et al., 2007) propose an optimal step Nonrigid Iterative Closest Point (NICP) framework for registration. To be specific, the nearest-point search was used to estimate the preliminary correspondence first. Then the deformation of the reference shape and the active stiffness is calculated according to fixed correspondence. The optimal deformation continues by searching new mappings from the displaced template vertices, which results in the dense correspondence. Alternatively, the best-known 3D Morphable Model (3DMM) proposed by Blanz and Vetter (Blanz & Vetter, 1999) is also used in 3D face shape recovery. It is a linear parametric model of 3D face shape and texture in low dimension space. To recover the face shape, the key is to fit the model to face images / face scans of previously unseen subjects. In initial 3DMM (Blanz & Vetter, 1999), Blanz and Vetter reduce the non-rigid registration problem into an image registration task by cylindrically unwarping the 3D facial mesh into 2D UV space. The dense one-to-one correspondence are established automatically through

finding the corresponding points in 2D images using gradient-based optical flow algorithm. They only align the facial meshes of 200 subjects with similar ethnicity and age. Beyond this constrained setting, the proposed method is fragile. To resolve this issue, Patel and Smith (Patel & Smith. W., 2009) suggest to manually annotate several key points of 3D face for alignment first, and then employ a Thin Plate Splines (TPS) (Bookstein & Green, 1993) warp to register the face scan into a reference shape in UV space. Some following-up works (Z. Fan et al., 2018; Gerig et al., 2018; Gilani et al., 2017; C. Zhang et al., 2016) have been proposed. Additionally, Rodala and Cosmo (Rodolà et al., 2017) propose to compute partial functional correspondence according to the perturbation analysis of Laplacian matrices. This method is resilient to missing parts or incomplete data in object reconstruction. But it has not been used in human face.

The non-rigid registration methods deeply rely on the correspondence which are prone to cause errors due to the unprecise estimation caused by iterative closet point or hand-selected facial features. Furthermore, such correspondences are inaccessible when the given depth view is noisy and non-frontal with partial facial regions occluded. In addition, the approximate alignment between input and facial template also needs to be done before registration. All these issues make 3D reconstruction from a single unconstrained depth view intractable with existing non-rigid registration methods.

2.2.4 Deep Learning Methods

Currently, there are two main kinds of deep models for reconstructing the 3D shape from a single depth map. One predicts the 3D shape from the depth map directly. The other instead predicts the deviation between the input and a shape prior, which is inspired by the non-rigid registration method mentioned above.

As an early attempt, data-driven methods are widely used in 3D object reconstruction. Many approaches (H. Fan et al., 2017; R.Girdhar et al., 2016; Wu et al., 2015) reconstruct objects in 3D voxel grid by combining the semantic labels of objects using CNNs. Without the category information, (Sharma et al., 2016)

and (B. Yang et al., 2018) proposed to learn the volumetric 3D object using autoencoder and conditional GAN separately. Instead of reconstructing 3D shape in voxel space, the deformation model (Yu et al., 2018) predicts the point cloud of 3D shape directly. Specifically, (Yu et al., 2018) propose to estimate new positions for all vertices of a reference shape according to the deform changes in input 3D data. In contrast, (Groueix et al., 2018; W. Wang, Ceylan, et al., 2019) infers the per-vertex displacement of the template, and then map them to each vertex for 3D shape reconstruction. Recently, a cage-based deep model is proposed to deform the input source object to target shape (Y. Wang et al., 2020). These mentioned methods are efficient in 3D object/human body reconstruction. However, they could not be used directly for 3D face reconstruction result from the complex and detailed structure of human face.

In terms of deep deformation models in 3D face reconstruction, most of the algorithms are developed for dense correspondence prediction via 3D-to-3D model fitting. Both (Tan et al., 2018) and Abrevaya et al. (Abrevaya et al., 2018) propose an autoencoder-based multilinear model which can accurately predict the 3D face mesh and decouple identity and expression variations. Specifically, a CNN-based encoder extracts the latent feature representations of depth data, from which the decoder performs a multilinear transformation to conduct 3D face fitting. However, both methods require 3D faces with an initial correspondence as input and the correspondence problem is considered in the restrictive space expressed by the model. To overcome the limitation, Liu et al.(F. Liu et al., 2019) propose an innovative framework to jointly learn a nonlinear face model from a diverse set of raw 3D scan databases and establish dense point-to-point correspondence among their scans. To be specific, as the most researcher usually do, they explore the use of PointNet (Qi et al., 2017) architectures for converting unordered point clouds to latent feature representations. These representations embed the facial identity and expression information separately, from which the decoder networks recover the new positions for all vertices of a face identity and the point-wise displacement for expressions accordingly. Since no correspondence label for real

scans are available, a self-supervised loss based on chamfer-distance between output and real scan is defined to optimize the final reconstruction in the training process. Although this method deals with the point cloud, with the given intrinsic matrix, it could recover the 3D facial geometry from a single depth map. However, this approach ignores the head pose information due to the data pre-processing which normalizes all the input point cloud into a unit sphere.

Compared with expensive and laborious ground truth generation for real collected data, synthetizing 3D facial mesh using existing statistical model and rendering them into depth image using classical Z-buffer is convenient and fast, and can supply unlimited mesh-depth image pairs. But the model solely trained on synthetic data usually perform poorly when tested on real collected depth map due to domain mismatch. Recently, domain adaptation which aims to align data that are sampled from different distributions has drawn a lot attention. Zhong et al. (Zhong et al., 2020) propose a cycleGAN-based domain adaptation framework to regress the 3D face geometry. The framework can be seen in Fig.2.7. Only paired synthetic samples and real depth image are fed to the network, no labels of real data are required. Instead of predicting the direct 3D face representation, it estimates the coefficients of the existing parametric face model. Point-wise

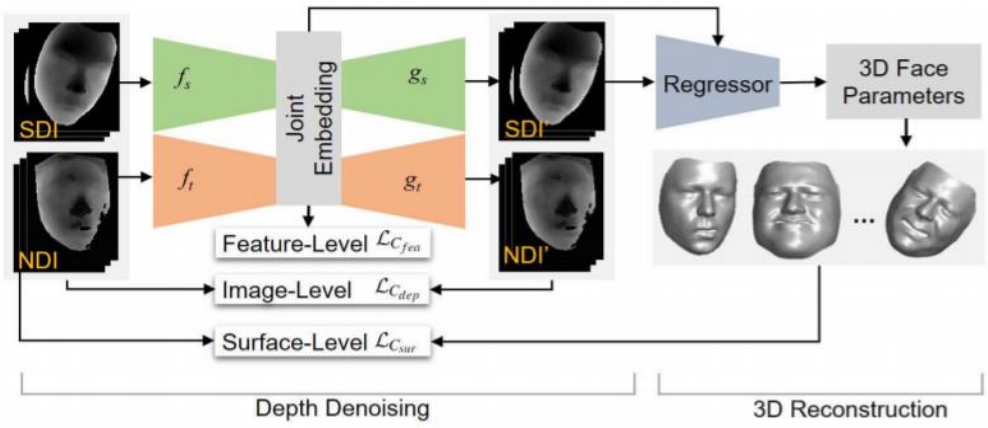

Figure 2.7: The framework of prior art method (Zhong et al., 2020) for face reconstruction from single depth image.

correspondences are extracted to guarantee the shape consistency between the output and input real depth. It is the first attempt to recover 3D facial geometry from real single depth image. Currently, very few explorations about leaning-based method have been made on monocular face reconstruction from depth image.

2.3 Generative Adversarial Network

This section recaps the generative adversarial network which is the deep learning framework used in this thesis. It is a group of artificial intelligence and is drawing growing interest due to its promising performance in image, speech, text generation tasks. GAN is built based on a distinctive framework, in which the generative models are built through an adversarial process. Specifically, the framework includes two parts: a generator capturing the data distribution, and a discriminator predicting the probability that a sample came from the input rather than output. During training, the connection between these two models is established via an adversarial process: generator tries to maximize the probability of discriminator making a mistake, while the discriminator aims to distinguish the real or fake. Both modules are optimized to improve their learning abilities over time. According to the research purposes, massive extensions have been made in the past few years. The representative architecture contributing to the major waves in the chronicle are shown in Fig. 2.8 and will be introduced in this section.

Original GANs (Goodfellow et al., 2014) is proposed by Goodfellow and his colleagues in 2014. For the architecture, fully-connected neural networks were adopted in both generator and discriminator, and the maxout function (Goodfellow et al., 2013) was employed in the discriminator while the ReLU activations (Jarrett et al., 2009) were used for all the layers except the final one with a sigmoid non-linear function in the generator. In order to avoid the overfitting of discriminator, the authors suggest to update discriminator K times and update generator once.

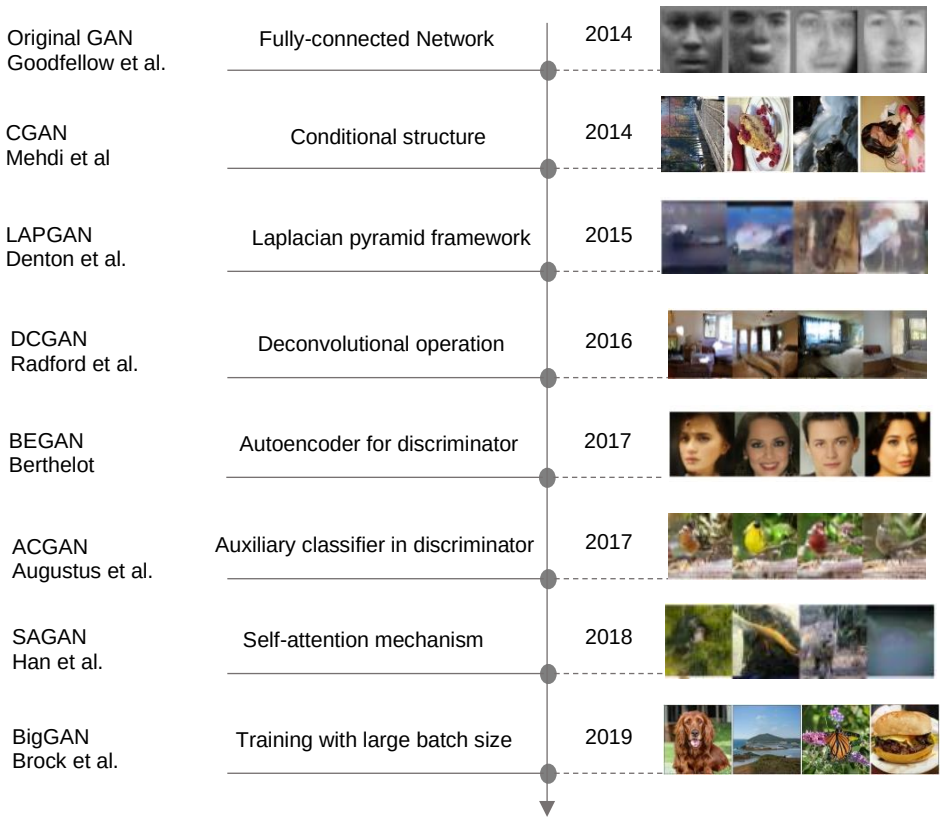

Figure 2.8: Timeline of GAN'S architecture-variants introduced in this thesis.

This architecture variant can generate simple images, and was evaluated on MNIST, CIFAR-10 and Toronto face dataset. It does not demonstrate good generalization performance for more complex image types.

Conditional GANs (Mirza & Osindero, 2014) extends the original GAN to a conditional model by feeding the extra auxiliary information including class labels or other semantic text to both discriminator and generator. The extra auxiliary information is combined with the prior input noise to feed to generator, which generates the conditional real-looking images. Meanwhile, the discriminator also takes the labels, which enhances the distinguishing ability. This type of GAN can generate the images conditioned on class labels, and was tested on MNIST, Yahoo Flickr Creative Common 100M datasets.

Laplacian Pyramid GANs (Denton et al., 2015) utilizes a cascade of convolutional networks within a Laplacian pyramid framework to generate images

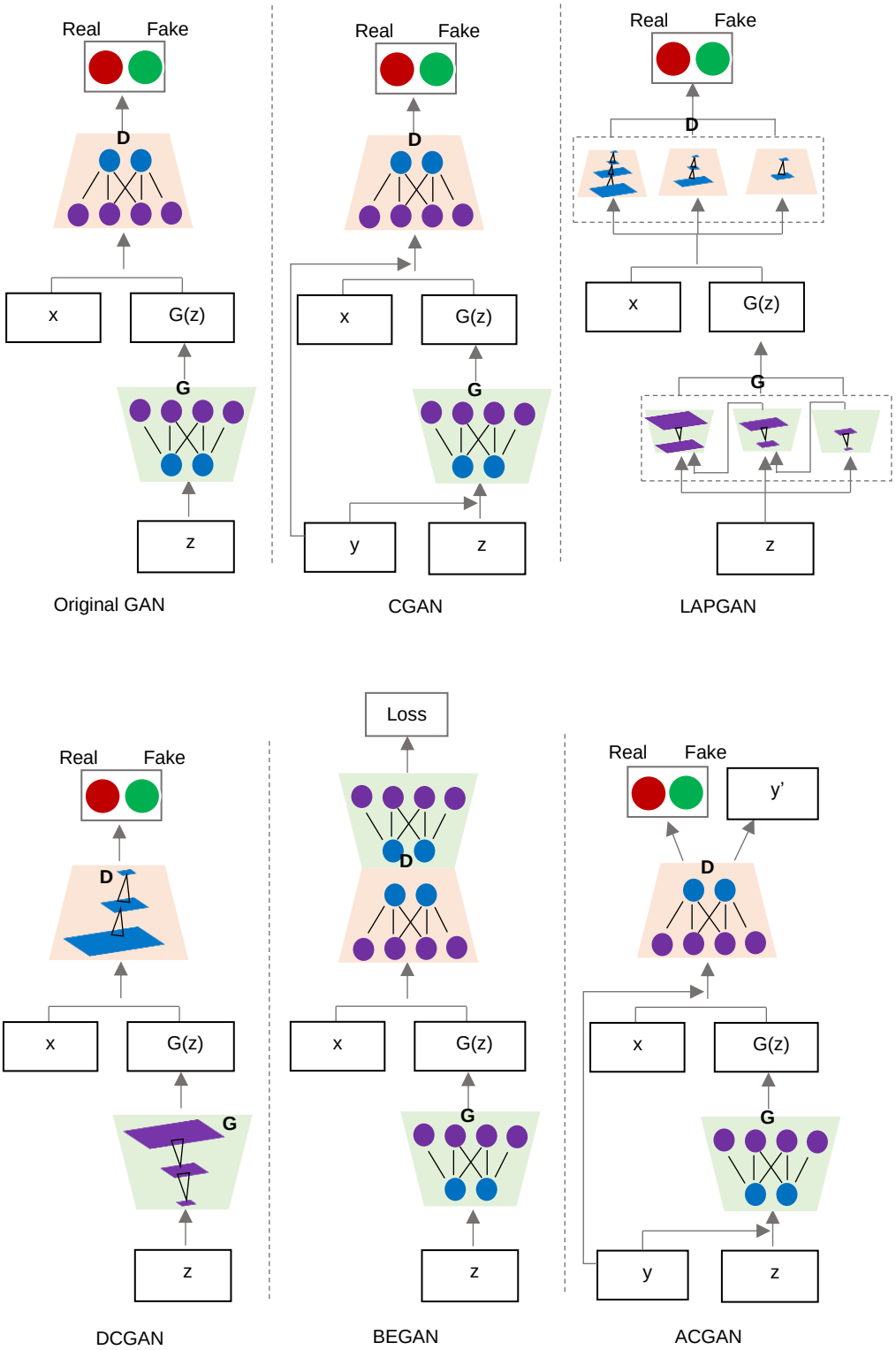

Figure 2.9: The structure of different GAN variants.

in a coarse-to-fine manner. It combines the conditional GAN model with a

Laplacian pyramid representation which is used to up-sample the image. There are a set of generators and discriminators to make the multiscale generation. The first generator usually produces a very small image, which can eliminate the unstable issue. And then the generated image is up-sampled through using Laplacian pyramid before feeding to the next generator. Each conditional generator produces the particular levels of details of an image in a Laplacian pyramid representation. Similarly, the discriminator distinguishes the real or fake samples from multiscale. What needs to be mentioned here is that the generators generally generate the image difference for the high-resolution images, which is much less complicated that the same size raw images. This structure benefits more stable training and high-resolution modelling. Extensive experiments have been done on three public datasets to prove its performance.

Deep Convolutional GANs (DCGANs) (Radford et al., 2015) adopts the convolutional networks in both generator and discriminator. Compared with original GAN, it mainly has three critical modifications, which results in stable training and high-resolution modelling. Firstly, all sampling operation are conducted using fractional-strided convolutions for generator and strided convolutions for discriminator rather than pooling layer. Secondly, to reduce the impact on poor initialization, batch normalization(Ioffe & C. Szegedy, 2015) is suggested to make input of each layer in a normal distribution. Thirdly, ReLU activation function (Jarrett et al., 2009)is utilized in generator for all layers except final one, which adopted Tanh as the non-linear function, while LeakyReLU activation(Maas et al., 2013) is utilized for all layers of discriminator. Compared with Relu ignoring the negative neuros, the LeakyRelu activation takes these neuros into consideration, which avoids the network to stuck a “dying state” circumstance (e.g., inputs smaller than 0 in ReLU). DCGANs are evaluated on Large-scale Scene Understanding (LSUN), ImageNet and the customized-assembled face dataset, and has got promising performance. It is a very important milestone in the GANs history and the deconvolution becomes the main architecture used in the generator. Due to the limit of the model capacity and the

optimization used in DCGAN, it is only successful on low-resolution and less diverse images.

Boundary Equilibrium GAN (BEGAN) (Berthelot et al., 2017) adopts an autoencoder architecture rather than general encoder framework in the discriminator. The structure of both generator and discriminator in BEGAN are very similar to the generator in DCGAN. Different from the general discriminator defined as a binary classifier, the discriminator in BEGAN is designed to reconstruct the input images. The distinguishing problem is converted to match the reconstruction loss distribution, which is an effective indirect method of matching data distributions and has been confirmed experimentally in the paper (Berthelot et al., 2017). In other words, the objective of the discriminator is to maximize the difference of the data distribution of the reconstruction losses of real samples and produced samples. Following the energy-based GAN (J. Zhao et al., 2016), Wasserstein distance is used to measure the difference. With the help of this equilibrium enforcing method, the BEGAN balances the generator and discriminator, especially at the early training stage. The model is trained on CelebA dataset and have got promising results.

ACGAN (Odena et al., 2017) is a new variant of the GAN architecture named auxiliary classifier GAN. It is an extension of CGAN, changing the discriminator to estimate the class label of a given image instead of receiving the label as input. In other words, the discriminator in ACGAN not only needs to predict whether the input image is real or fake as general discriminator usually do, but also must estimate the class label of the input. Compared with existing models, this variant is not absolutely novel, but it generates high-quality results and appears to stabilize training while learning a representation in the latent space that is independent of the class label. The model trained on CIFAR-10 and the ImageNet for all 1000 classes has improved the visual quality of the produced samples. However, the improvements highly rely on large-scale labelled database, which might cause challenges in real-world application.

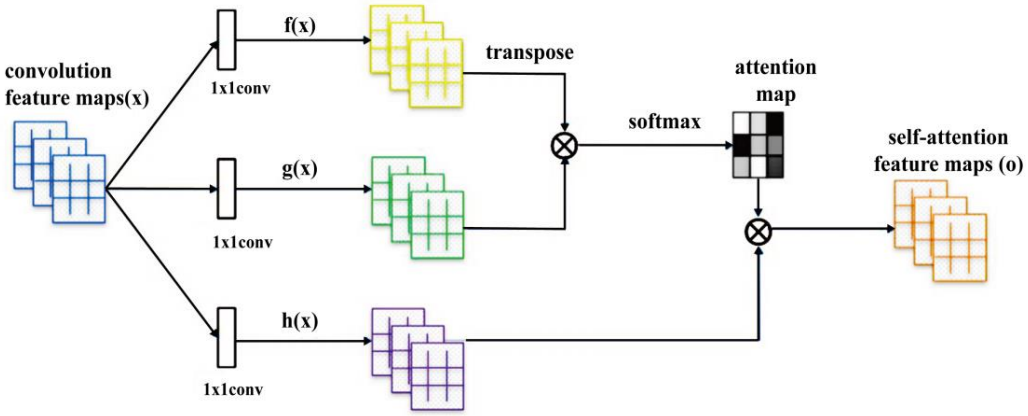

Figure 2.10: Self-attention module proposed by (H. Zhang et al., 2019).

Self-attention GAN (H. Zhang et al., 2019) is also a new variant of GAN. Compared with aforementioned GANs only capture the local spatial information and the receptive filed may not cover enough structure, it adopts the self-attention mechanism in both generator and discriminator to ensure large receptive field and without sacrificing computational efficiency. This mechanism can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. Benefiting from this mechanism, SAGAN (H. Zhang et al., 2019) is able to learn global, long-range dependencies for generating images. It has achieved great performance on multi-class image generation based on the ImageNet datasets.

Recently, a new GAN variant named BigGAN (Brock et al., 2018) has been proposed and has shown outstanding, large scale, indistinguishable, and high-quality image generation capacity. The training setting of BigGAN follows the SAGAN, in which the learning rate was halved and train two D steps per G step. This section summarizes following operations which make BigGAN scale-up the architecture: (1) Self-attention module contributing to the model diversity and

Hinge loss boosting more stable training. (2) the class information provided to the generator model via class-conditional batch normalization. (3) Update discriminator more than generator. (4) Moving average of model weights. More details can be referred to the original paper. This model can train bigger neural networks even with more parameters; create a more extremely detailed image with remarkable performance.

2.4 Conclusion

This chapter has briefly summarized the relative fields to the thesis. It comprises the literature on:

- The methods of salient object detection for general scene analysis,
- The approaches of 3D face reconstruction from single depth view for human-centred scene analysis,
- The variants of generative adversarial network provided the basic framework for this thesis.

To sum up, earlier saliency detection methods usually extract the salient object using the low-level features like pixel, color or texture. This kind of methods do not perform well on complex scenes such as complicated background and multi-objects. Currently, deep learning-based methods have been proposed and have got improvements. However, these methods generally take the pixel errors into consideration for saliency map generation, which ignores the shape and semantic information of the salient object. For 3D face reconstruction, traditional methods usually are optimized-based, which is not efficient for large poses and occlusions. Until to recent advance, very few efforts were made on learning-based technology to explore the 3D face reconstruction from single depth image.

Overall, compared with human performance, significant advances have been made on visual scene analysis, but it still remains challenging at current stage. The

aim of this thesis is to explore the potential of GAN for bridging the gap between human intelligence and algorithms of visual scene analysis.

Chapter 3
Perceptual Loss-Guided GAN for Saliency Detection

3.1 Introduction

Saliency detection attempting to highlight salient objects or regions by human visual and cognitive systems (Borji & Itti, 2012) is one of the key fundamental problems in psychology and computer vision. It has various applications including image segmentation (Zeng et al., 2019), object recognition (Zeng et al., 2019; Z. Q. Zhao et al., 2019) and visual scene understanding (Jeon & Kim, 2018) in computer vision. During the past few years, significant advances have been made on these tasks using both conventional classic methods (Cheng, Mitra, Huang, Torr, et al., 2014; Jian et al., 2018; Q. Liu et al., 2017; Ogasawara et al., 2017) and deep learning technologies (L. Wang et al., 2016; C. Zhang et al., 2019; P. Zhang, Wang, Lu, Wang, & Yin, 2017). Most of existing conventional saliency detection methods (Jian et al., 2018; Q. Liu et al., 2017; Ogasawara et al., 2017) mainly devotes to design the low-level saliency cues including color contrast, edge, pattern and texture, or extract the middle-level object information such as spatial context, shape and contour. However, these handcrafted features cannot make obvious contrast between background and the salient region if salient object located in complex scene scenarios. Deep learning methods including Convolutional Neural Network (CNN) (L. Wang et al., 2018, 2016; P. Zhang, Wang, Lu, Wang, & Yin, 2017) and Generative Network (GAN) (Cai & Yu, 2018; J. Pan et al., 2017; C. Zhang et al., 2019) were proposed to solve the existing problem due to its powerfulness of extracting high-level feature representations.

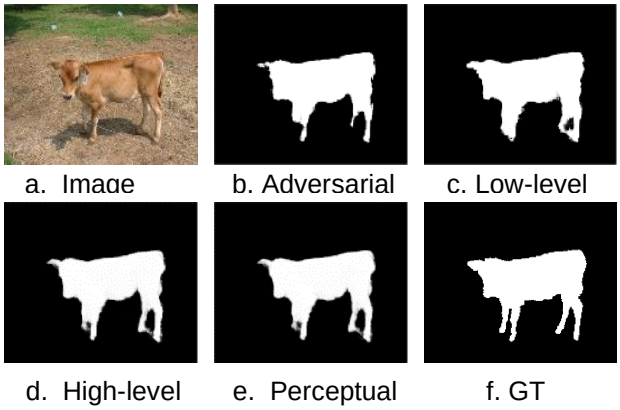

Figure 3.1: The salient object detection results using different loss functions. From b to e, the results are generated by the proposed structure guided by adversarial loss (b), adversarial loss and low-level feature measurements (c), adversarial loss and high-level feature measurements (d), adversarial loss and perceptual loss (e) respectively.

These kinds of network usually improve the accuracy and precision of prediction by minimizing the difference between the output and the target saliency map. Generally, the difference is represented by the loss function which is a significant part of improving the deep learning models performance. An appropriate loss function can not only help to speed up the convergence of gradient descent but also improve the naturalism and realism of the saliency map. For example, WGAN (Martin Arjovsky & Bottou, 2017) proposed to define the loss function by using Earth Mover Distance (EMD) instead of Jensen-Shannon divergence, which improved the performance of GAN deeply by making the network more stable and easier to train. However, how to choose or design an appropriate training loss for saliency detection is still an open problem.

A loss function usually consists of distance measurement methods and the feature representations of the sample. Currently, the deep learning methods (Cai & Yu, 2018; J. Pan et al., 2017) mainly focus on defining the loss function by calculating the mean square error or binary cross entropy between the generated and ground truth saliency map. Obviously, only the pixel difference is taken into consideration to optimize the deep model. But it is not robust enough to ensure the

shape and detailed information of the salient object. Recent work has demonstrated that the perceptual loss function based on feature loss extracted from several layers of the pretrained deep model can boost the high-quality image generation (Johnson et al., 2016).

To this end, this work pays attention to the profit of perceptual loss to generate high-quality saliency maps using GAN. In GANs, training is a competing game between two neural networks: the generator synthesizes samples to match with the training data; the discriminator distinguishes the real sample from the fake synthesized by the generator. In this chapter of the thesis, the real sample stands for the pair of input RGB image and the corresponding ground truth saliency map. Similarly, the fake means the pair of input image and the predicted saliency map. Specifically, GAN's generator consisting of deep convolutional neural network is trained for rough saliency map generation. Instead of only using L1 cost function relying on pixel values, a perceptual loss function depending on both high-level representations and low-level information extracted from the pretrained deep model is adopted, to measure the semantic and appearance difference between the salient object extracted by filtering input image with the output and the target saliency map. As shown in the Fig.3.2, the encode part of generator is initialized with the pretrained VGG-16 model (Simonyan & Zisserman, 2014) to get the high-level feature maps and the decoder produces a rough saliency map directly than convert the feature map to a latent feature vector. The rough saliency map was first refined by 2 fully convolution layers in generator and then further refined with a multiscale discriminator network which is trained to distinguish the real sample from the fake one.

In short, the contributions can be concluded as follows:

- A novel perceptual loss guided GAN (PerGAN) is developed for salient object prediction in still images. The perceptual loss is defined based on both low-level content cue and high-level semantic representation, which ensures more detailed information in edge localization and the completeness of the salient objects respectively.

- A multiscale discriminator is designed to distinguish the real and fake samples at different scales for further improving the learning ability of generator. Furthermore, the quality of output saliency map is guaranteed. Experimental results show that the multiscale discriminator performs between than general one.
- The simple coarse-to-fine strategy is adopted to convert the feature map to saliency map at the last two layers with 1×1 kernels in generator.

3.2 Related Work

3.2.1 Saliency Detection

Detecting salient objects has attracted much attention for past decades. In the early stages, the methods based on the hand-crafted feature representations like boundary (Q. Liu et al., 2017), color contrast (Ogasawara et al., 2017), texture pattern (Jian et al., 2018) were designed to estimate the salient objects. Generally, these low-level features are good at keeping the object structure and boundary information, but they are not able to represent the semantic information of the object.

More recently, deep learning models (W. Wang, Lai, et al., 2019) have shown superior performance in many image processing and analysis tasks including detecting salient objects over other state-of-the-art techniques. Previous deep models (Lee et al., 2017; L. Wang et al., 2015) typically chose the salient objects from the saliency score which was regressed using CNNs. Currently, fully convolutional networks were adopted to generate pixel-wise saliency maps in a direct end-to-end manner. Specifically, Hou et al.(Hou et al., 2017) present to embed the skip-layer with short connections within the holistically-nested edge detector architecture, which shared more spatial information with target maps. Wang et al. (L. Wang et al., 2015) proposed a recurrent fully convolutional networks based deep model for saliency detection. It resolved the gradient vanish

problem caused by very deep networks. Zhang et al. (P. Zhang, Wang, Lu, Wang, & Ruan, 2017) proposed to aggregate multi-level convolutional feature maps for salient object detection, which is efficient and flexible. To further explore the efficiency of the convolutional feature maps, Zhao et al. (T. Zhao & Wu, 2019) applied an attention strategy for choosing the features maps which were useful for saliency map generation since not all the feature maps include the salient cues. All aforementioned methods used low-level differences such as pixel errors to guide the network for training, which cannot ensure the high-level information for salient object prediction. Inspired by (Johnson et al., 2016), we propose a perceptual loss guided network, which reduces not only the appearance difference but also the semantic errors between the predicted and target saliency maps in optimization process, to generate necessary details for salient objects.

3.2.2 Generative Adversarial Network

Generative adversarial networks (GAN) (Goodfellow et al., 2014) has gained many promising performances in various visual analysis tasks, including high-resolution image generation (T. C. Wang et al., 2018) , font generation (Hayashi et al., 2019), 3D object reconstruction (B. Yang et al., 2018) and so on. It has been adopted for salient object detection as well. Cai et al. (Cai & Yu, 2018) proposed to adapt conditional GAN which integrated the u-net and skip-connections for saliency map generation. To reduce the blurred contour result from that the mean squared error filters the high spatial frequency, SalGAN (J. Pan et al., 2017) proposed to use the binary cross entropy to measure the pixel difference between the synthetic and marked saliency maps. Instead of using general discriminator to classify fake or real, Pan et al. (H. Pan et al., 2020) proposed to take the category label of the image into consideration and designed a supervised classifier for classification. But the downside is that the class labels were required. To improve the feature learning ability in a direct manner, Zhu et al. (Hu et al., 2018) developed a Laplacian pyramid-based generator to produce the smooth saliency maps. Recently, Zhang et al. (C. Zhang et al., 2019) was proposed to estimate the salient

object by adding capsule blocks in generator and discriminator. All these aforementioned networks were updated by minimizing the original adversarial losses which based on low-level feature representations. In terms of the optimization function, the perceptual loss has not been paid attention to by all the mentioned methods. Moreover, there has been less previous evidence for GAN with a multiscale discriminator in saliency detection.

3.2.3 Perceptual Loss Function

Loss function which depending on feature representations and measurement plays an important role in deep network. How to choose the feature representations and how to measure the feature difference are still open issues in saliency detection. Low-level feature representations presenting the color, boundary, texture information are widely used to ensure the edge information since salient object has the obvious contrast with the simple background. A number of recent works (Gatys et al., 2015; Jetley et al., 2016; Johnson et al., 2016; C. Wang et al., 2018) have optimized the loss function based on feature representations extracted from pre-trained network to generate high-quality images. Various information such as edge, colour, shape and semantic category was embedded in feature loss, which makes the output reasonable. To the best of our knowledge, the saliency detection issue has never been explored by GAN which is optimized by perceptual loss function. In this paper, the perceptual loss is considered to ensure the completeness of the object by understanding what it is. Specifically, the perceptual loss defined based on the high-level semantic information and low-level features is calculated to measure the perceptual difference between the colourful object maps, extracted via filtering the background information from the input RGB images. Of course, the perceptual difference is used for network optimization.

3.3 Perceptual Loss-Guided GAN

In this paper, to tackle the problem of salient objects detection, the GAN with perceptual loss is developed. Fig.3.2 illustrates the overall architecture of proposed PerGAN in detail. The designed PerGAN mainly has two components: the generator as in upper part and the multi-scale discriminators showed in lower half

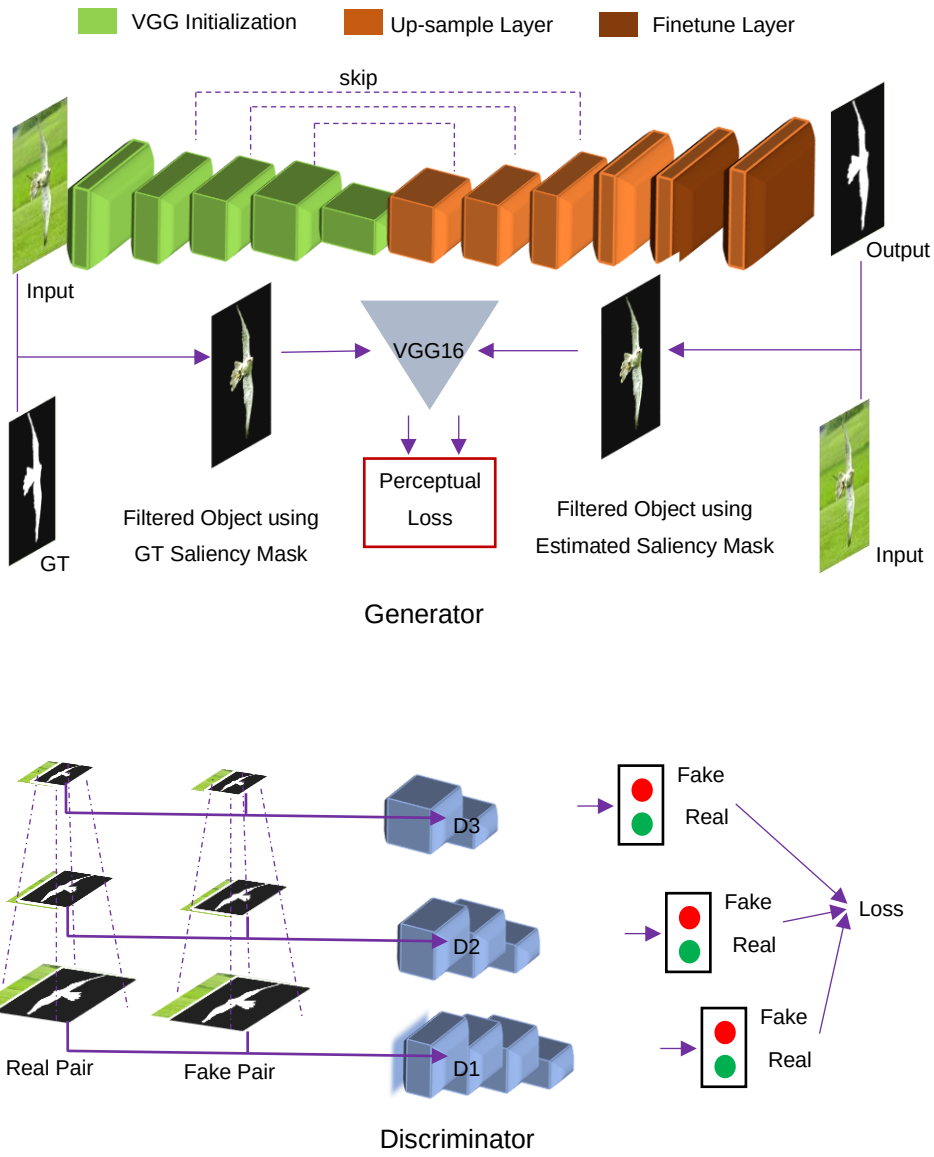

Figure 3.2: The framework of proposed PerGAN.

space. The training process is the competing game between two convolutional neural networks: a generator network which generates the salient map, and a discriminator network which aims at distinguishing the real sample between the ground truth saliency maps and generator's output. Furthermore, this section provides details on aforementioned module's structures and the objective loss functions used to guide the network.

3.3.1 Generator

The generator consists of an autoencoder structure and the skip connections between the encoder and decoder. Specifically, the encoding part comprises five convolutional layers whose filter size and stride step is 3×3 and 1×1 respectively. To down sample the output feature maps, the max pooling operation with 2×2 filter follows each convolutional layer, which make the size of feature scale with a factor ½ at each layer to reduce from 224 to 14. In contrast, the output channel of max pooling layer begins with 64, doubling at each subsequent layer to rise to 512. In order to extract the feature maps efficiently, the pre-trained VGG-16 model was imported to initialize the variables, which also saves lots of computing resources during training. Inspired by (T. Zhao & Wu, 2019) , the output of encoder is directly fed to decoder directly rather use linear transformation to convert them to hidden latent space. The decoding part is a symmetric structure with encoder. It has 5 transposed convolution layers with the strides of 2×2 replacing the up-pooling operation for generating the salient features maps and 2 fully convolutional layers with the size 1×1 of filters to fine-tune the coarse outputs. Instead of using the same filter of convolutional layer in encode, the filter size is increased to 5×5 for receiving more fields of salient objects. Both encoding layers and decoding layers are followed by a ReLU activation function (Jarrett et al., 2009) except for the last layer with a logistic regression converting the data distribution to [0, 1]. Skip-connection between encoder and decoder guarantee propagation of local structures. What needs to be emphasised here is

that the autoencoder is unable to learn desirable saliency map without the last two fine-tune layers. In training process, the generator was fed with RGB images and supervised by corresponding target saliency maps. In testing phase, only RGB images are required.

3.3.2 Discriminator

The discriminator devotes to distinguish estimated maps from the group of generator's outputs and target maps. In order to improve the distinguishing ability, a multiscale network structure is designed to extract the feature representations of input pairs at three different scales. Particularly, the discriminator has 3 sub-networks (named as D_1, D_2 and D_3) with 4, 3, 2 convolutional layers separately. The input pair are directly fed to D_1, and then is down sampled by stride 2 and 4 in width and height to feed to D_2 and D_3 separately. Similar to the encoder in generator, all sub-networks have an identical encoding architecture. Rather than max pooling operation, the down sampling of feature maps was conducted by convolution operation with strides 2×2. Every convolutional layer only was followed with the ReLU activation operations (Jarrett et al., 2009) except the last layer. The large filter with size 5×5 in D_1 and D_2 is also used to get big receptive field of salient objects for ensuring the overall cues of fake pairs is in line with the real pairs. In contrast, small kernels with shape 3×3 was adapted in D_3 to guarantee the local detailed information of fake samples.

3.3.3 Loss Functions

The total loss used to optimize the network includes three weighted components: an adversarial loss, a *L1* loss based on per-pixel for generator, and a perceptual loss.

a) Adversarial Loss

The adversarial loss, which is a reflection of how the generator could maximally deceive the distinguisher and how well the distinguisher could discriminate between real and fake samples, is developed as:

$$\mathcal{L}_{adv} = \sum_{k=1}^{3} E_{(x,y)}[\log(D_k(x,y))] + E_{(x,y')}[\log(1 - D_k(x, G(x)))] \quad (3\text{-}1)$$

As mentioned in the earlier section, D_k refers to the k^{th} discriminator and $G()$ stands for the generator. The feeding image is defined as x, and its corresponding saliency map is defined as y. In each iteration of our training process, two generator updates are performed followed by a discriminator update. The weight of the first element (adversarial loss) in total loss is ω_{adv}.

b) L1 Loss

Compared with *L2* loss, the per-pixel loss can reduce the artifacts of depth map and saliency map. The *L1* cost function between the target map T and the produced sample $G(x)$ is calculated as follows:

$$\mathcal{L}_{l1} = |G(x) - T| \quad (3\text{-}2)$$

To balance different losses, it is contributed to the total loss with weight ω_{L1}.

c) Perceptual Loss

The perceptual loss function used in this work includes two parts. One is the content loss function used for measuring the image-specific difference, the other is the style loss function aimed for calculating the difference between textures, shapes and colors. Both of them depend on the feature extracted from VGG-16 model. The feature examples can be seen in Fig.3.3.

Content loss: As mentioned in (Johnson et al., 2016), compared with matching the low-level pixel difference, the estimated salient objects and target RGB objects which were filtered out by saliency map S (or G(x)) and target map T is encouraged to have similar feature representations extracted from the deep model. To extract the feature representation of the generated and target salient objects, the

pre-trained VGG-16 is employed. The content loss of the two filtered images, after passing through a convolutional layer can be calculated as follows:

$$\mathcal{L}_{content} = \frac{\|V(T*x) - V(S*x)\|^2}{c*l*w} \tag{3-3}$$

Here, c, l and w means the output's channels, length and width separately. V() represents the non-linear transformation, which is performed by the VGG-16 network.

***Style loss**:* The content loss is defined based on the shallow features which mainly contains edge and boundary information. Inspired by the recent work which built the style reconstruction loss (Gatys et al., 2015) through measuring the style feature representation between two different images, the style differences in patterns, shapes and colors between the predicted salient objects and the target salient objects is also computed, which will further ensure the salient region of the output maps. The style feature of a convolutional layer can be represented by the correlation between different feature maps of this layer. The correlation between a group of maps usually was stored in a Gram matrix. For the input f, the output $\varphi_j(f)$ of j^{th} layer is with size $C_j \times H_j \times W_j$. The components of Gram matrix are described by:

$$GM_j(x) = \frac{1}{C_j*H_j*W_j} \sum_h^{H_j} \sum_w^{W_j} \varphi_j(f)_{h,w,c}\, \varphi_j(f)_{h,w,c'} \tag{3-4}$$

And the size of Gram matrix is $C_j \times C_j$. The outputs of Conv1_2, Conv2_2, Conv3_3 and Conv4_3 are utilized to reconstruct 4 correlation matrices. Thus, the style reconstruction loss between the Gram matrices of the colourful predicted salient objects and target objects is defined by:

$$\mathcal{L}_{style} = \sum_{n=1}^{4} \|GM(V_n(S*f) - GM(V_n(T*f)\,)\|_F^2 \tag{3-5}$$

The perceptual loss is defined by weighting style loss and content loss as below:

$$\mathcal{L}_{percep} = \mathcal{L}_{content} + \lambda \mathcal{L}_{style} \tag{3-6}$$

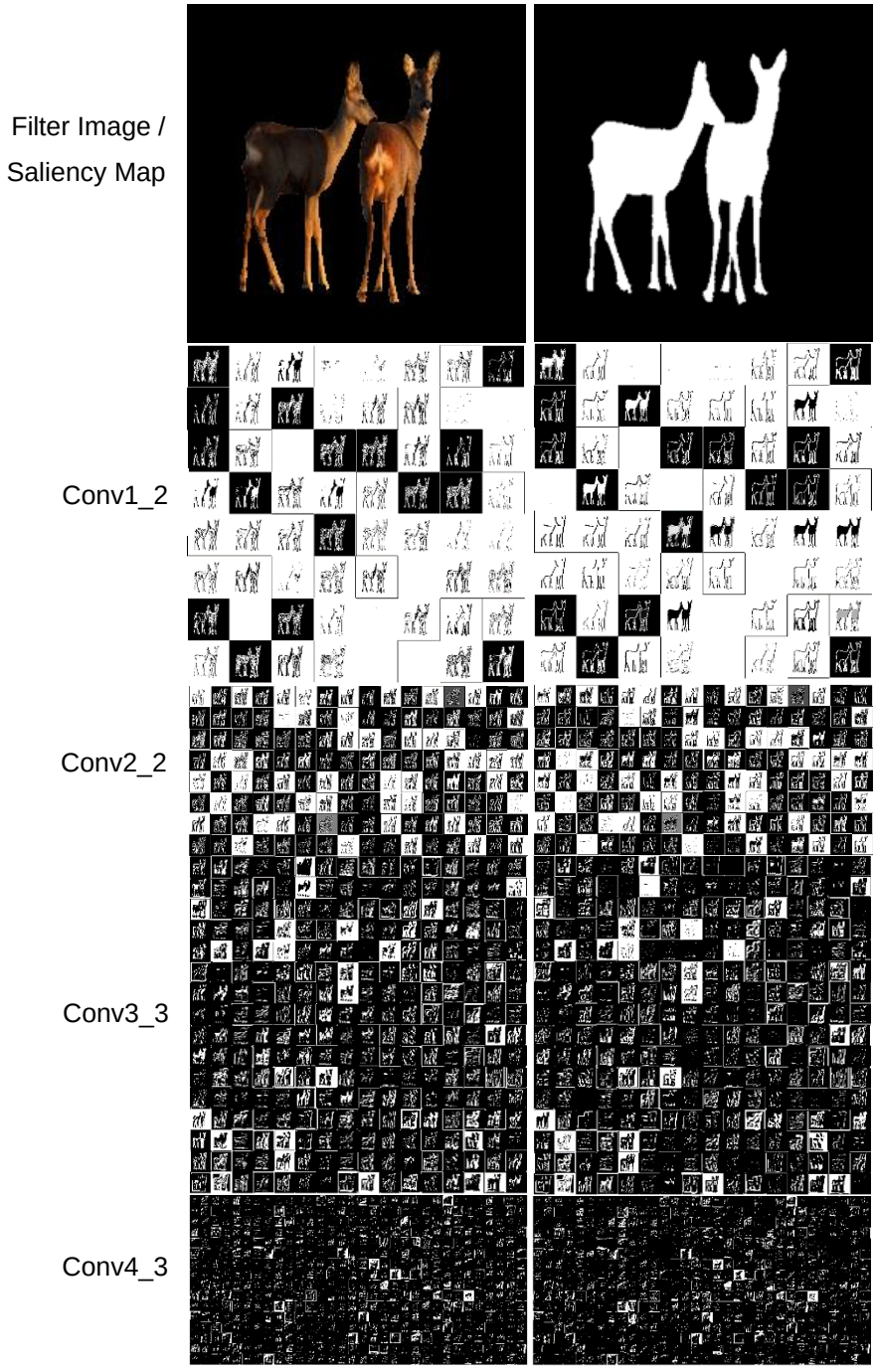

Figure 3.3: The feature map examples extracted from VGG-16. The output of Conv1_2 was adapted for building content loss. All feature maps were used to reconstruct the style loss. Obviously, the shallow layer usually extracts the obvious edge and shape information. While the deep layer records somewhat semantic information.

It is a necessary part of total loss as well. The weight for it in total loss is ω_{percep}.

Overall, the loss function used for guiding the whole network is as following:

$$\mathcal{L}_{total} = \omega_{adv} * \mathcal{L}_{adv} + \omega_{l1} * \mathcal{L}_{l1} + \omega_{percep} * \mathcal{L}_{percep} \quad (3\text{-}7)$$

3.4 Experimental Results

3.4.1 Datasets

Six public benchmark databases are used to demonstrate our experimental performance following the settings mentioned in (T. Zhao & Wu, 2019).

ECSSD (Yan et al., 2013) includes 1, 000 structurally complicated images obtained online. The corresponding binary ground truth saliency maps were annotated by five subjects. It is the extended dataset of CSSD (Yan et al., 2013).

PASCAL-S (Y. Li et al., 2014) covers 850 natural pictures with corresponding saliency segmentation masks which were annotated by twelve participants. We set the threshold as 0.5 to acquire binary ground truth saliency maps as mentioned in (G. Li & Yu, 2015). It also provides the eye fixation information of each image. In this paper, we just use binary saliency segmentation maps to measure our generated salient objects.

DUTS (L. Wang et al., 2017) is the largest dataset including 15572 complex images and the corresponding per-pixel annotations. It covers varied image contents, such as indoor, outdoor, human, animals and vehicles. The dataset is divided into two parts, 10552 for training and the remining 5019 images for testing.

DUT-OMRON (C. Yang et al., 2013) contains 5167 natural images. It not only provided the bounding boxes but also offered the corresponding pixelwise annotations. However, different observers have different opinions on the annotations, which contribute it to a challenging dataset. Most of the existing saliency models have not predicted very accurate saliency maps for this dataset.

HKU-IS (G. Li & Yu, 2015) covers 4447 pixelwise annotation ground truths which are in line with the natural RGB images. It is a challenging dataset since most images have more than two salient objects and inapparent colour contrast. The author grouped the database into three subsets: training set including 2450 samples, validation set containing 500 specimens and testing set comprising the resting 1447 pairs.

THUR15K (Cheng, Mitra, Huang, & Hu, 2014) includes 15k images and only about 6k images are with accurate pixelwise annotation. These annotated images were divided into 5 groups covering butterfly, coffee mug, dog, giraffe and plane. Most of the existing saliency prediction methods have not achieved a high accuracy on this dataset since it has many complex scene scenarios.

3.4.2 Training Process

The whole MSRA-10k dataset (Ogasawara et al., 2017) is used for the training process. To extend the training data, the data augment including flip and rotation is conducted. Approximately 80K images which are size of 224*224 were fed to the network to learn the mapping from RGB images to saliency maps. For the network setting, the optimization strategy mentioned in (Cai & Yu, 2018) is followed. The ω_{adv}, ω_{l1}, ω_{percep} mentioned in total loss are set as 1, 100 and 100 respectively, and the λ controlling the style loss is set as 10. The batch size is set to be 1. Adam optimizer is adopted with the momentum of 0.5 and the step size of 0.0002. The whole training process took about 50 hours (180k iterations) on a PC with a 4GHz Intel i7 processor and a Nvidia GTX 1080 GPU (8G RAM).

3.4.3 Evaluation Metrics

a) Measures

Four widely used performance measurements including precision (P), recall (R), F-measure score (F_β) and mean absolute error (MAE) were adopted to evaluate the proposed method. The threshold was changed from 0 to 255 to produce 256

values of every measurement per saliency map. All the evaluation performances were obtained via averaging the measures over saliency maps in the whole dataset.

Precision is to measure the quality of our predictions only based on what our network claims to be positive. In other worlds, it demonstrates the proportion of the predicted salient regions (PSR) within the target salient regions (TSR) over the predicted regions. In contrast, recall is to measure such with respect to the mistakes we did. It calculates the percentage of the number of detected salient object regions inside the target regions over those of the target regions. Both precision (P) and recall (R) are defined by:

$$precision = \frac{PSR \cap TSR}{PSR} \tag{3-8}$$

$$recall = \frac{PSR \cap TSR}{TSR} \tag{3-9}$$

Generally, precision and recall do not take the true negative salient scores into consideration. A balanced and complementary measurement calculating the average pixel-wise difference is needed. MAE is the classical and popular measurement to evaluate the difference. In our case, MAE between predicted saliency map S and its corresponding ground truth *T* is computed as:

$$MAE = \sum_{i=1}^{l*w} \frac{|S-T|}{l*w} \tag{3-10}$$

The parameter l and w stands for saliency map's length and width, respectively. According to measurement criterion both S and T are converted to [0, 1]. Besides, the overall performance F-measure taking precision and recall into consideration is also used for measurement. It is defined as:

$$F_\beta = \frac{(1+\beta^2)* P*R}{\beta^2*P+R} \tag{3-11}$$

β^2=0.3 is set to highlight more precision as mentioned in (G. Li & Yu, 2015). What needs to be mentioned is that the mean F-measure scores is calculated for comparison by averaging the value of precision and recall.

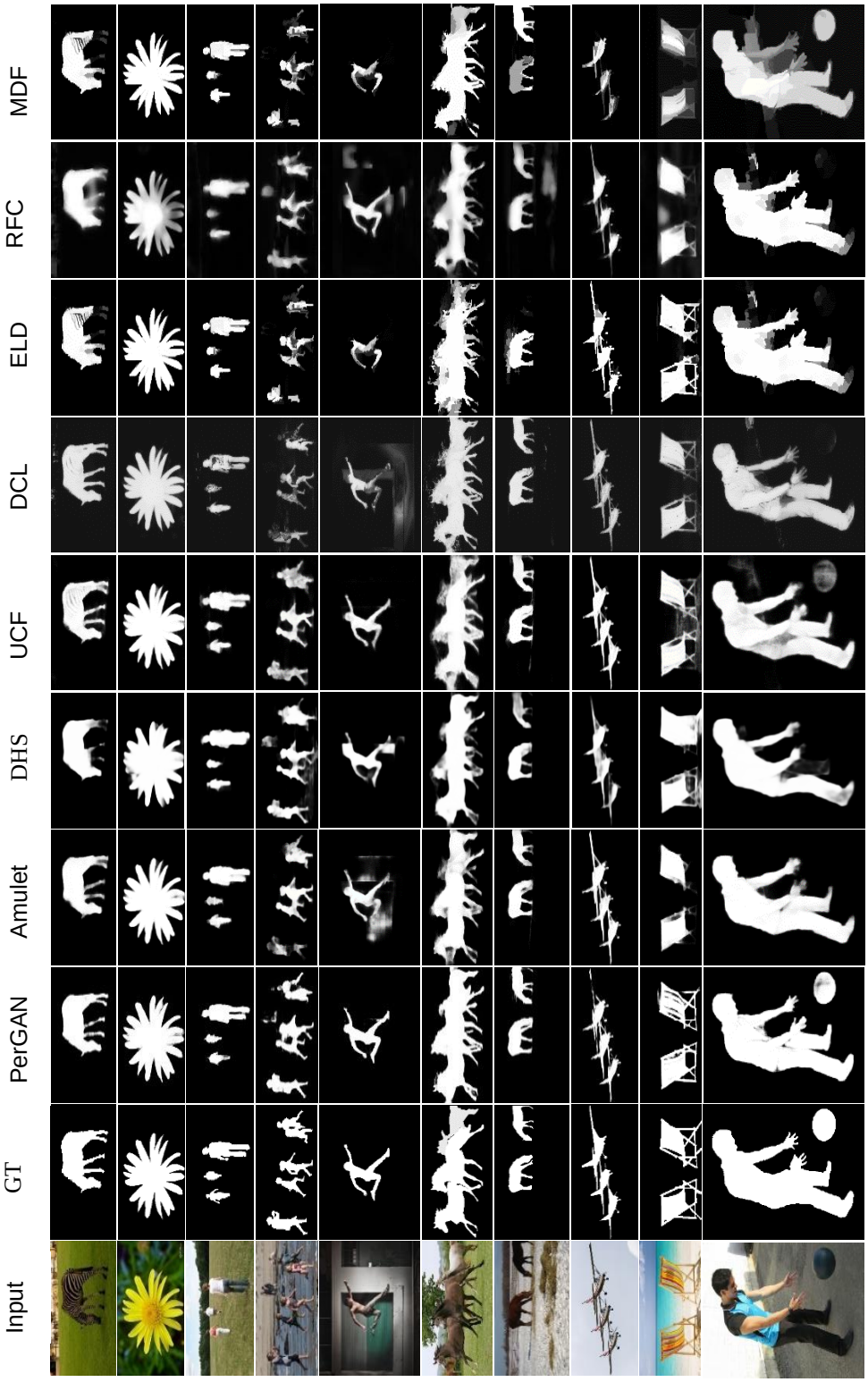

Figure 3.4: Saliency maps generated by state-of-the -art methods and proposed method.

The proposed method is compared with more than ten state-of-the-art salient object detection methods including Amulet (P. Zhang, Wang, Lu, Wang, & Ruan, 2017) , BL (Tong et al., 2015), CBGAN (C. Zhang et al., 2019), CF (Hassan et al., 2019), DCL (G. Li & Yu, 2016), DHS (N. Liu & Han, 2016), DMCN (Sun et al., 2019), DRFI (H. Jiang et al., 2013), DS (X. Li et al., 2016), ELD (Lee et al., 2017), KSR (T. Wang et al., 2016), LAWS (Qian et al., 2019), LEGS (L. Wang et al., 2015), MCDL (R. Zhao et al., 2015), MDF (G. Li & Yu, 2015), RFCN1 (L. Wang et al., 2016), RFCN2 (L. Wang et al., 2018), MSNSD (Liang et al., 2019), UCF (P. Zhang, Wang, Lu, Wang, & Yin, 2017) on six aforementioned datasets. For fair

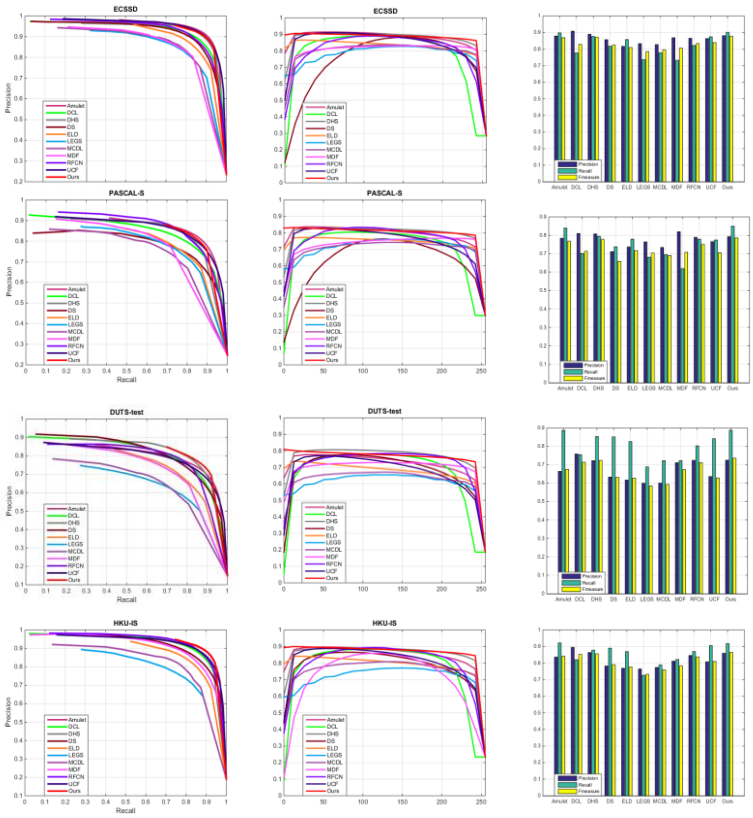

Figure 3.5: Comparisons among eleven deep learning-based salient object detection approaches on four challenging public datasets. Each row relates to a dataset. The left and middle columns are the precision-recall curves and the F-measure-threshold curves of different methods. The right column shows the average precision, recall and F-measure scores. The proposed method is comparable to the state-of-the art methods under all measurements.

Table 3.1: Comparison of Quantitative MAE (closer to zero is Better) and F-measure (Larger is Better). The top three results are marked with Red, Green and Blue.

Methods	ECSSD		PASCAL-S		DUTS-test		DUT-OMRON		HKU-IS		THUR15K	
	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β
Amulet	0.061	0.869	0.100	0.763	0.085	0.678	0.098	0.647	0.052	0.839	0.094	0.670
BL	0.217	0.684	0.249	0.574	0.238	0.490	0.239	0.499	0.207	0.660	0.219	0.530
CBGAN	0.089	0.858	0.176	0.686			0.113	0.689	0.092	0.859		
CF	0.115	0.692	0.192	0.037			0.112	0.6060				
DCL	0.151	0.827	0.181	0.714	0.149	0.714	0.097	0.657	0.136	0.853	0.161	0.676
DHS	0.059	0.871	0.095	0.773	0.067	0.724			0.054	0.852	0.082	0.673
DMCN	0.054	0.867	0.223	0.634			0.091	0.686	0.057	0.858		
DRFI	0.166	0.733	0.207	0.618	0.175	0.541	0.138	0.550	0.145	0.722	0.150	0.576
DS	0.124	0.826	0.176	0.659	0.091	0.632	0.120	0.603	0.078	0.785	0.116	0.626
ELD	0.082	0.810	0.123	0.718	0.093	0.628	0.092	0.611	0.074	0	0.098	0.634
KSR	0.135	0.782	0.157	0.704	0.121	0.602	0.131	0.591	0.120	0.747	0.123	0.604
LAWS	0.088	0.831	0.119	0.741	0.084	0.628	0.093	0.634	0.067	0.821	0.088	0.684
LEGS	0.119	0.785	0.155	0.697	0.138	0.585	0.133	0.592	0.119	0.723	0.125	0.607
MCDL	0.102	0.796	0.145	0.691	0.105	0.594	0.089	0.625	0.092	0.757	0.103	0.620
MDF	0.108	0.805	0.146	0.709								
MSNSD	0.171	0.777	0.151	0.792			0.109	0.688	0.071	0.837		
RFCN	0.109	0.834	0.133	0.751	0.090	0.712	0.111	0.627	0.089	0.835	0.100	0.695
RFCN2	0.067	0.871	0.105	0.778					0.055	0.856		
UCF	0.080	0.841	0.127	0.701	0.117	0.629	0.132	0.613	0.074	0.8080	0.112	0.645
Ours	0.052	0.878	0.091	0.782	0.064	0.732	0.086	0.677	0.041	0.864	0.081	0.687

comparisons, saliency maps used in this chapter is provided by the authors or generated by the recommended parameters and released code of existing methods. Fig.3.4 shows visual maps generated by mentioned state-of-the-art approaches and proposed technique. It can be seen that the proposed method completely highlights the salient objects even in very challenging scenes. In the fifth row of Fig.3.4, almost all methods could not disregard complex wall as the salient objects. Only UCF (P. Zhang, Wang, Lu, Wang, & Yin, 2017) and the proposed method detected the jumping man accurately and our method produced a smooth and detailed edge information. This is because the perceptual loss profits our method extracting both edge cues and semantic information to ensure the accurate saliency detection.

As shown in Fig.3.4, the results of proposed method have fewer missing pixels in predicted salient regions and have more detailed information in boundary localization. This is because the feature extracted by pre-trained VGG-16 not only includes the high-level semantic feature representations but also covers the low-level edge information. It is obvious that the feature maps in first two rows mainly capture the boundary localization and edge information and those in last two rows represent the abstract sematic information, which corresponds with the low-level feature representations and high-level feature representations, respectively.

For quantitative evaluation, P-R curves, F-measure curves and mean F-measure scores are illustrated in Fig.3.5. The Fig.3.5 obviously demonstrates that the proposed method is competitive over other methods on these four datasets under all evaluation metrics. Furthermore, F-measure score is the overall performance measurement as mentioned above, so a quantitative comparison MAE and mean F-measure is reported in Tab.3.1. It shows that the proposed approach locates in the first place on almost all datasets. In terms of F-measure scores, the PerGAN outperforms the rank 2 model by 0.7%, 0.8%, 0.5% over ECSSD, DUTS-test, HKU-IS respectively. For MAE, the proposed approach reduces the value by 0.2%, 0.4%, 0.3%, 0.3%, 1.1%, 0.1% on ECSSD, PASCAL-S, DUTS-test, DUT-OMRON, HKU-IS, THUR15K respectively.

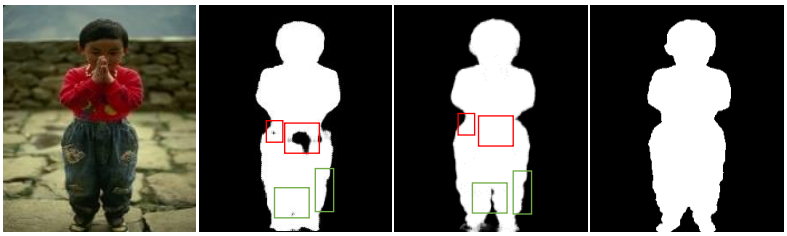

Figure 3.6: Visual examples generated by guiding with and without the perceptual loss (PL).

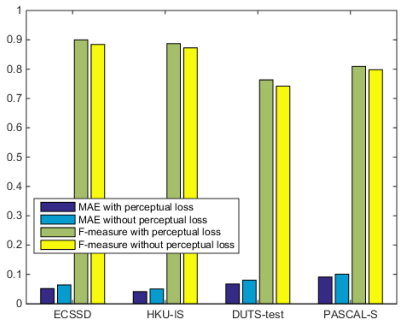

Figure 3.7: The bar chart of MAE (closer to zero is better) and F-measure (closer to one is better) with and without perception loss calculated from our method.

Table 3.2: The F-measure based on RGB images and binary saliency maps (SM) and MAE from the proposed method.

Dataset		ECSSD	PSACAL-S	DUTS-test	DUT-OMRON	HKU-IS
F_β	SM	0.8991	0.8094	0.7633	0.8871	0.7108
	RGB	0.9024	0.8111	0.7767	0.8904	07228
MAE	SM	0.0517	0.0909	0.0673	0.0410	0.0908
	RGB	0.0519	0.0909	0.0638	0.0408	0.0855

Table 3.3: Ablation study with different components combinations on ECSSD dataset.

CGAN	Multi-Discriminator	Fine-Tune	Perceptual Loss	MAE	F_β
×				0.108	0.778
×	×			0.0762	0.823
×	×	×		0.0611	0.851
×	×	×	×	**0.0519**	**0.878**

b) The Effectiveness of Perceptual Loss

The perceptual loss is employed to guide GAN for learning more detailed salient objects including both boundary localization and the shape completion.For the better visual perception, the saliency maps generated with and without the perceptual loss is showed in Fig.3.6. Note that perceptual loss not only ensures the completeness of the salient object but also obtains the detailed edge information. Fig. 3.7 shows the MAE and F-score measure calculated from PerGAN with and without perceptual loss. These results illustrate that the perceptual loss improves saliency prediction model deeply. Besides, salient RGB maps are nearly same as the binary saliency maps in providing feature representation for building the perceptual loss is concluded from Tab.3.2. The output of different layers for both filtered image and the saliency map are displayed in Fig.3.3. These feature maps are generally similar, which indicates that the saliency maps rely more on the boundary localization and shape information than colors, textures and other patterns.

c) Ablation Study

To investigate the importance of different modules in proposed approach, the ablation study is conducted. The corresponding measurements are displayed in Tab.3.3. The model containing all components (multi-scale discriminator, fine-tune operation and perceptual loss) achieves the best performance, which demonstrates that these three components are all necessary for the proposed approach to get the promising salient object detection result.

3.5 Conclusion

In this work, a novel method named PerGAN is designed for salient object detection. In particular, the perceptual loss is adopted to guide the GAN for learning the completeness and the detailed boundary localization of the salient objects. The loss function designed to optimize the network takes information of both the output saliency maps and the RGB salient objects into consideration. This design improves the network's performance effectively, especially in locating correct boundary and keeping the completeness of the saliency map. To further improve the learning ability of PerGAN, a multiscale discriminator is developed to distinguish the real and fake pairs at different scales. Experimental results on four challenging databases illustrate that the proposed approach is competitive over state-of-the-art methods.

Chapter 4

3D Facial Geometry Recovery from a Depth View with Attention-Guided GAN

4.1 Introduction

A number of artificial intelligent systems such as robots and agents are designed for interacting with humans via multiple facial sensing techniques and learning methods. In some of those systems, reconstructing 3D facial geometry from integrated depth sensors is a fundamental step to achieve accurate facial expression capture and recognition. With the continuously increasing sensing precision and portability, depth camera is becoming a critical tool in capturing 3D objects including the human face. For example, the Apple's TrueDepth camera has been successfully deployed in mobile devices to support 3D facial applications. This motivates an important research stream which aims to reconstruct 3D facial geometry from 2.5D depth views. Existing methods (Donne & Geiger, 2019; Fang et al., 2019; Newcombe et al., 2011) were able to obtain the promising 3D shape by fusing multiple views of depth maps. However, it is not applicable for the practical application because of the complexity of multiple depth maps acquisition. Compared with these approaches, recovering geometry from a single view is more feasible and convenient in real applications. Nevertheless, it is very challenging to recover 3D facial geometry precisely if there is only one depth view available. This is mainly because partial observation can be theoretically associated with an infinite number of possible 3D facial information, especially when the depth view is non-frontal with the depth information of the occluded facial parts missing (see Fig.4.1).

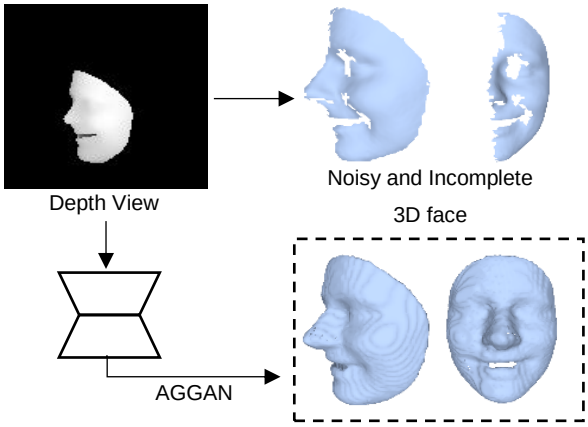

Figure 4.1: AGGAN can recover the complete 3D facial geometry from a noisy and non-frontal depth view.

The problem above can be interpreted as reconstructing a facial surface from 3D point cloud projected from the given depth view. This is a long-lasting research topic that has been extensively studied in computer graphics (Berger et al., 2014; Bernardini et al., 1999; Guennebaud & Gross, 2007; Hoppe et al., 1992; Kazhdan, 2005; Kazhdan & Hoppe, 2013; Lorensen & Cline, 1987a). Typical solutions reconstruct the surface by either fitting the points with a discrete grid (Guennebaud & Gross, 2007; Lorensen & Cline, 1987a) or using the zero set of an implicit function (Hoppe et al., 1992; Kazhdan, 2005; Kazhdan & Hoppe, 2013) such as the indicator function defining the interior and exterior of the object surface. However, these approaches degenerate sharply when dealing with noisy and non-frontal depth views, and normally can only recover partial 3D facial geometry. The problem can also be cast to 3D shape non-rigid registration (Amberg et al., 2007; H. Li et al., 2009; C. Luo et al., 2019; Sumner & Popović, 2004; Zollhöfer et al., 2014) which is also pervasive in computer graphics. Generally, non-rigid registration methods first build dense point correspondences between the projected 3D point cloud and a template 3D facial mesh, and then conforms the template mesh to the point cloud using the built correspondences. Whereas a complete 3D facial geometry can be acquired with such methods, facial parts occluded or missing in the depth view can rarely be warped correctly on the template because

false correspondences are prone to being found for them. Furthermore, these methods usually require certain hand-selected facial feature points to rigidly align the template with the point cloud for a promising registration initialization. In summary, existing methods can hardly handle imperfections in the given depth view such as the noise and missing data.

Existing methods merely utilize noisy 3D information embedded in the given imperfect depth view, while making no attempt to build and exploit a 3D facial point distribution which covers various facial geometries. With such a distribution, the reconstruction problem becomes generating or sampling 3D facial points from that distribution given a depth view as a conditional input, which could be solved efficiently by Generative Adversarial Networks (GAN) (Goodfellow et al., 2014). Accordingly, in this study a new variant of GAN named AGGAN is proposed to learn the highly-complicated conditional distribution of 3D facial geometry given its depth view from thousands of synthetic depth-3D pairs. First, the 3D facial geometry is encoded within a high-resolution voxel grid which has shown robustness in depicting 3D shapes (Jackson et al., 2017; J. W. Li et al., 2018; Woo et al., 2018; B. Yang et al., 2018). Then the GAN is guided to extract features that are more sensitive and discriminative in locating 3D facial points by incorporating the attention mechanism which has been validated in many other computer vision tasks (Jyoti et al., 2020; T. Zhao & Wu, 2019; Z. Zhu et al., 2019) . To build a generative model covering a variety of natural depth-3D mappings, large variations in head pose and facial expression together with random noise are introduced during synthesizing training depth views.

Compared with existing methods on data generated from benchmark facial image datasets, the proposed data-driven AGGAN recovers a more complete and smoother 3D facial shape, while being able to handle a much wider range of view angles and more resistant to noise in the input depth view. Overall, the main contributions of this chapter are as follows:

- To the best of our knowledge, this is the first work of its kind that utilizes GAN to recover 3D facial geometry from a single unconstrained depth view.
- The incorporation of the attention mechanism into GAN can improve the precision of 3D facial geometry prediction is validated.
- This chapter showcases that using synthetic facial depth views for training is helpful in generalizing AGGAN to real depth views captured from depth cameras.

4.2 Related Work

4.2.1 3D Surface Reconstruction from Point Cloud

The area of 3D surface reconstruction has witnessed impressive progress in the last two decades (Berger et al., 2014). From the perspective of the reconstruction output, the proposed solutions can be broadly divided into two categories, producing either a discrete surface (Guennebaud & Gross, 2007; Lorensen & Cline, 1987a) or an implicit function (Hoppe et al., 1992; Kazhdan, 2005; Kazhdan & Hoppe, 2013). The first kind of solutions typically fits a regular grid to the given points such as the well-known Marching Cubes (Guennebaud & Gross, 2007; Lorensen & Cline, 1987a) which extracts the surface by finding intersections between the cubes of the grid and the points. The latter type utilizes the knowledge of the exterior and interior of the surface with an implicit function for reconstruction. The implicit function can have various forms such as a signed distance field (Hoppe et al., 1992) or an indicator function (Kazhdan, 2005), whereby the reconstructed surface is found by isocontouring for an appropriate isovalue. However, when the point density is low, there are outliers or missing data, these methods are prone to generating an incomplete surface that poorly approximates the desired object shape. As a result, they can hardly deal with a single unconstrained facial depth view which is often noisy and with a head pose.

4.2.2 3D Shape Non-rigid Registration

The concerned reconstruction problem can be projected into the 3D non-rigid registration framework (Amberg et al., 2007; H. Li et al., 2009; Sumner & Popović, 2004; Zollhöfer et al., 2014) if there is a facial geometry prior available. A typical solution is to register a facial template mesh to the given depth view using a deformation model based on smooth local affine transforms. Primarily, the registration process has to estimate reliable correspondences between the template and 3D points projected from the depth view for warping the template to match the underlying geometry of the captured depth data. False correspondence can cause strong shape distortions that are inconsistent with the desired facial shape. However, such correspondences are inaccessible when the given depth view is noisy and non-frontal with partial facial regions occluded. Moreover, a promising correspondence estimation often requires hand-selected facial feature correspondences (Amberg et al., 2007; Sumner & Popović, 2004) or a rigidly-aligned shape prior (H. Li et al., 2009; Zollhöfer et al., 2014) that offers a strong approximation of the target facial geometry. This is against with the most general setting where no facial geometry prior and feature point correspondences are available. All these issues make 3D reconstruction from a single unconstrained depth view intractable with existing non-rigid registration methods.

4.2.3 3D Reconstruction from a Single Depth View with Deep Learning

Whereas learning the 3D facial shape from a single depth view with data-driven deep neural networks remains almost unexplored, there are several studies (Dai et al., 2017; Song et al., 2017; Varley et al., 2017; Wu et al., 2015; B. Yang et al., 2018; Zou et al., 2017) working on single depth view 3D object reconstruction. However, the early approaches (Varley et al., 2017; Wu et al., 2015) apply a low resolution voxel grid ($\leqslant 40 \times 40 \times 40$) which can only preserve the coarse shape information of the object. To solve this problem, Dai et al. (Dai et al., 2017)

propose a two-stage pipeline: first using the neural network to predict a shape prior encoded with a 32×32×32 voxel grid from the given depth view, then synthesizing a higher resolution shape based on a pre-built shape database. Such a shape database is however very difficult to construct, especially for the human face which has extensive shape variations. The SSCNet (Song et al., 2017) extends the reconstruction to 3D indoor scene which contains multiple object categories and requires a much higher-resolution volumetric space for representation. The method leverages the synthetic scene data which provides both the depth view and the ground-truth voxel-level occupancy annotations, which significantly reduces the expense for collecting the high-resolution training data. Inspired by these studies, we propose to solve the ill-posed single depth view 3D face reconstruction with deep neural networks. To model the complex non-rigid facial shape motions and deformations within the network, we synthesize a large amount of training data by altering along the dimensions such as facial identity, expression and head pose.

4.3 Methodology

In contrast with existing methods focusing on modelling only the given imperfect depth data, the ill-posed single depth view 3D face reconstruction is proposed to be solved in a more data-driven manner. Specifically, an attention-guided GAN named as AGGAN (see Fig.4.2) is designed to model the complex 2.5D depth-3D relationship by learning from a large amount of synthesized training pairs. The generator of AGGAN approximates the real conditional distribution of 3D facial surface given its depth view. This data-driven prior is supposed to be more robust than manually specialized priors (e.g. distance field function (Hoppe et al., 1992), indicator function (Kazhdan, 2005) or template 3D facial mesh (Amberg et al., 2007; Sumner & Popović, 2004) used in previous methods on addressing challenging data imperfections such as noise, missing/occluded facial parts. In the following sessions, the proposed AGGAN and the training data synthesise are

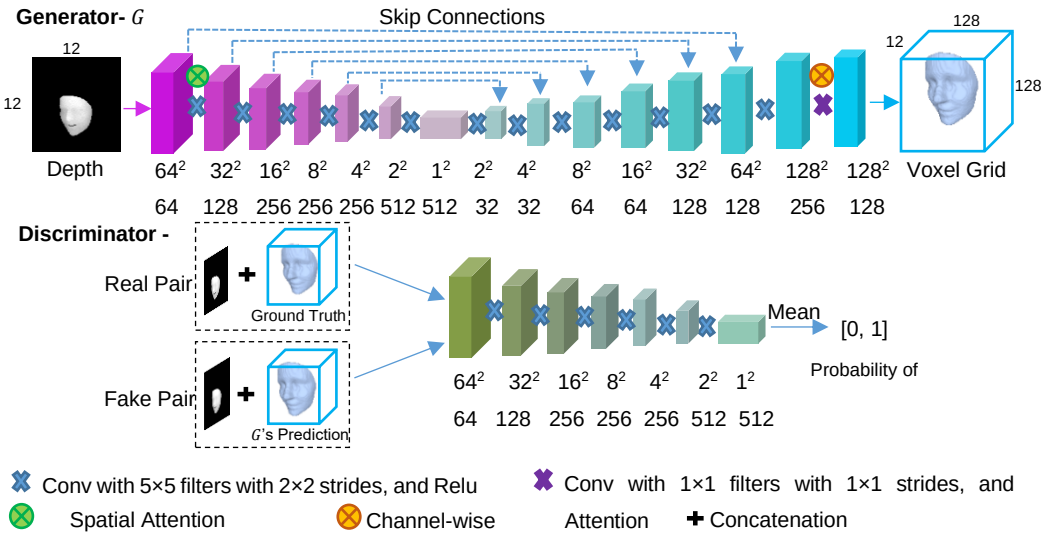

Figure 4.2: The architecture of AGGAN.

introduced in detail.

4.3.1 AGGAN

From previous work (Jackson et al., 2017; B. Yang et al., 2018) on 3D shape reconstruction, the voxel representation shows a promising ability in depicting 3D geometry and can be seamlessly processed by deep neural networks. Thus, 3D facial geometry is encoded within a 3D voxel grid whose voxel occupancy (1 for facial point and 0 for non-facial point) indicates if the current point belongs to the facial surface or not. The voxel grid resolution is set as 128×128×128 which was determined after balancing the grid's representation capability and the network's processing consumption.

Fig.4.2 illustrates the structure of AGGAN. During training, the generator G tries to learn the ground-truth 3D voxel grid which encodes the facial geometry from a 128×128 facial depth view. Coupling with the corresponding depth view, both G's prediction and its ground truth counterpart are then fed into the discriminator D for training a classifier to distinguish real reconstruction pairs (the pair of a depth view and its ground-truth voxel grid) from fake reconstruction pairs

(the pair of a depth view and its G prediction). G's outputs are forced to not only get close to the ground truth voxel grid but also maximize the probability of D making a mistake. This adversarial learning drives G to recover a faithful 3D facial geometry that matches the input depth view. Given a new facial depth view, G will be called to predict the 3D voxel grid that encodes the facial geometry.

a) Generator and Discriminator

The generator is a fully convolutional encoder-decoder network with skip-connections. The encoder consists of seven convolutional layers, each of which uses a bank of 5×5 filters with 2×2 strides and is followed with a Leaky ReLU activation (Maas et al., 2013). Without specification, the remaining network applies the same filter setup. From the first convolutional layer to the last one, the number of feature map channel is 64, 128, 256, 256, 256, 512 and 512 respectively. On the other side, the decoder comprises eight transpose-convolutional layers, the first seven of which are followed with Leaky ReLU activations, while the last one is followed with a sigmoid function to regulate the final output as the voxel occupancy probability. The number of each transpose-convolutional layer's output channel is 32, 32, 64, 64, 128, 128, 256 and 128. The last transpose-convolutional layer is for fine-tuning purpose and uses a bank of 1×1 filter with 1×1 stride. Skip-connections are built between encoder and decoder to guarantee the information sharing and prevent the gradient vanishing problem.

The discriminator accepts a 128×128×129 tensor concatenated by a facial depth view and a 3D voxel grid as input, and outputs a single scalar whose value is between 0 and 1 to specify the probability that the voxel grid fully matches the depth view. Excluding the input and the last layer, it has a same structure as the generator's encoder. The last layer calculates the mean of a 1×1×512 feature vector output from the previous layer. This mean feature is shown effective in stabilizing the adversarial training (B. Yang et al., 2018).

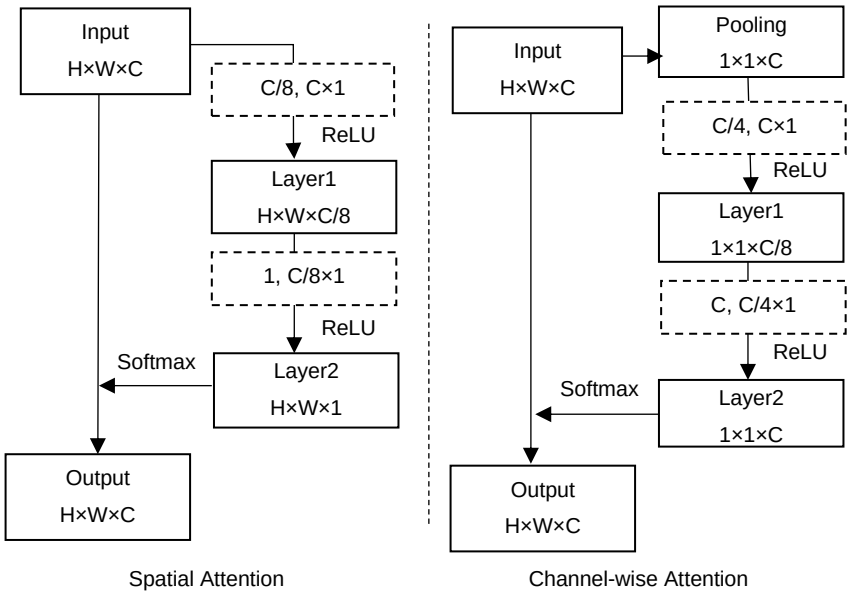

Figure 4.3: The attention modules in AGGAN.

b) Attention Mechanism

Spatial Attention. In general, the face occupies only a partial region in the depth view. The left background region is noisy and might mislead the neural network to learn less informative features for 3D facial geometry prediction. To force the network to focus more on the foreground facial region during feature learning, a spatial attention mechanism is incorporated into AGGAN' generator. After the first layer activation layer of the generator's encoder, two convolutional layers followed with a softmax function are applied on the low-level feature maps to generate a spatial weighting map (see Fig. 3):

$$\boldsymbol{SA} = F_{sa}(\boldsymbol{f}^l, \boldsymbol{W}_{sa}) \tag{4-1}$$

$$F_{sa}(\boldsymbol{f}^l, \boldsymbol{W}_{sa}) = softmax\Big(cv2_{sa}(cv1_{sa}(\boldsymbol{f}^l, \boldsymbol{W}_{sa}^1), \boldsymbol{W}_{sa}^2)\Big) \tag{4-2}$$

where $\boldsymbol{f}^l \in \mathbb{R}^{C \times HW}$ stacks C reshaped $1 \times HW$ low-level feature vectors output from the previous layer, F_{sa} is the mapping function whose parameters are denoted as $\boldsymbol{W}_{sa}$ and $\boldsymbol{SA}$. $\boldsymbol{SA}$ refers to the generated $H \times W \times 1$ spatial weighting map.

$cv1_{sa}(\cdot)$ and $cv2_{sa}(\cdot)$ represent two convolutional layers which use $\frac{C}{8}$ $C \times 1$ filters and a $\frac{C}{8} \times 1$ filter respectively, and whose parameters are $\boldsymbol{W}_{sa}^{1}$ and $\boldsymbol{W}_{sa}^{2}$. $softmax(\cdot)$ refers to the Softmax function. The final outputs of the spatial attention module can be obtained by weighting each previous feature map with $\boldsymbol{SA}$ (see Fig.4.3).

Channel-wise Attention. As reported in previous studies (Woo et al., 2018), different feature channels generated within convolutional neural networks correspond to different semantic information. Hence, the channel-wise attention mechanism is incorporated into AGGAN to weight heavier on feature channels that show higher relevance in predicting 3D facial voxel grid. The channel-wise attention module is adhered to the second-to-last transpose-convolutional layer of the generator's decoder, aiming to produce a weighting vector for feature channels (see Fig. 4.3):

$$\boldsymbol{CA} = F_{ca}(\boldsymbol{f}^{p}, \boldsymbol{W}_{ca}) \tag{4-3}$$

$$F_{ca}(\boldsymbol{f}^{p}, \boldsymbol{W}_{ca}) = softmax\big(cv2_{ca}(cv1_{ca}(\boldsymbol{f}^{p}, \boldsymbol{W}_{ca}^{1}), \boldsymbol{W}_{ca}^{2})\big) \tag{4-4}$$

where $\boldsymbol{f}^{p} \in \mathbb{R}^{C\times 1}$ is the feature vector obtained by max-pooling feature maps output from the previous layer, F_{ca} is the mapping function whose parameters are denoted as $\boldsymbol{W}_{ca}$ and $\boldsymbol{CA}$ is the generated $1 \times 1 \times C$ channel weighting vector. $cv1_{ca}(\cdot)$ and $cv2_{ca}(\cdot)$ represent two convolutional layers which use $\frac{C}{4}$ C $\times 1$ and C $\frac{C}{4} \times 1$ filters respectively, and whose parameters are $\boldsymbol{W}_{ca}^{1}$ and $\boldsymbol{W}_{ca}^{2}$. $softmax(\cdot)$ refers to the Softmax function. Then, each previous feature map is weighted by the specific channel weighting value in $\boldsymbol{CA}$ (see Fig.4.3).

c) Objective Functions

The overall objective function of AGGAN consists of two parts: an adversarial loss $\mathcal{L}_{adv}$ for the whole network and an additional 3D face reconstruction loss $\mathcal{L}_{recons3d}$ for the generator.

Adversarial Loss - $\mathcal{L}_{adv}$. To train a generator that is able to predict an accurate 3D voxel grid $\boldsymbol{y}$ from a depth view $\boldsymbol{x}$, the loss function of generator is shown in Eq.4-5. For the discriminator, the well-known WGAN-GP (Gulrajani et al., 2017) loss function is adopted (see Eq.4-6):

$$\mathcal{L}_{adv}^{g} = -\mathbf{E}[D(\boldsymbol{y}|\boldsymbol{x})] \tag{4-5}$$

$$\mathcal{L}_{adv}^{d} = \mathbf{E}[D(\boldsymbol{y}|\boldsymbol{x})] - \mathbf{E}[D(\hat{\boldsymbol{y}}|\boldsymbol{x})] + \lambda\mathbf{E}\left[\left(\left\|\nabla_{\boldsymbol{y}'} D(\boldsymbol{y}'|\boldsymbol{x})\right\|_2 - 1\right)^2\right] \tag{4-6}$$

where $\hat{\boldsymbol{y}}$ is the ground-truth 3D voxel grid corresponding with the input depth view $\boldsymbol{x}$ and $\boldsymbol{y}' = \epsilon\hat{\boldsymbol{y}} + (1-\epsilon)\boldsymbol{y}$, $\epsilon \sim U[0,1]$. λ balances between optimizing the gradient penalty and the original objective in WGAN.

3D Face Reconstruction Loss - $\mathcal{L}_{recons3d}$. Since the face only occupies a small part of the overall volume, most voxels in the grid tend to be empty and the estimated voxel occupancy is prone to false positive. Inspired by this observation, a modified binary cross-entropy loss function (Song et al., 2017; B. Yang et al., 2018) is utilized to weight the penalty on false positive estimations and the penalty on false negative estimations in terms of the ratio of occupied voxels in the ground truth grid:

$$\mathcal{L}_{recons3d}^{ce} = -\sum_{i=1}^{h\times w\times d}\begin{bmatrix}(1-\omega)\hat{y}_i \log y_i + \\ \omega(1-\hat{y}_i)\log(1-y_i)\end{bmatrix} \tag{4-7}$$

where h, w, d is the voxel grid's height, width and depth respectively. For voxel i, $\hat{y}_i$ is the ground truth occupancy state and y_i is the estimated occupancy state. ω denotes the ratio of occupied voxels in the ground truth grid. To further avoid false positive estimations, the $L1$ sparsity constraint is imposed on the predicted voxel grid $\boldsymbol{y}$:

$$\mathcal{L}_{recons3d}^{sparse} = |\boldsymbol{y}|_1 \tag{4-8}$$

Overall, the loss functions for generator and discriminator in AGGAN are as follows:

$$\mathcal{L}_G = \alpha\mathcal{L}_{adv}^{g} + \beta\mathcal{L}_{recons3d}^{ce} + \gamma\mathcal{L}_{recons3d}^{sparse} \tag{4-9}$$

$$\mathcal{L}_D = \mathcal{L}_{adv}^{d} \tag{4-10}$$

where α, β and γ are used to balance different loss terms, and their values are set empirically.

4.3.2 Data Synthesis

Collecting real facial depth views and their precise 3D data in a volume sufficient for training a deep network is laborious and expensive. However, it's easy to get a depth view given a 3D face, head pose and camera projection matrix. Considering there are many high-quality 3D face datasets (X. Zhang et al., 2013; X. Zhu et al., 2016) which cover a wide range of facial identities and expressions, synthesizing depth views from the known 3D facial data for training and validating AGGAN is proposed.

The dataset - 300W-LP proposed in (X. Zhu et al., 2016) is adopted for data synthesis. 300W-LP contains in-the-wild face images from four independent benchmark databases including HELEN (Zhou et al., 2013), LFPW (X. Zhu & Ramanan, 2012), IBUG (Sagonas et al., 2013), AFW (S. Zhu et al., 2015), and their 3D faces reconstructed by 3DMM fitting (X. Zhu et al., 2016). The reconstructed 3D faces capture the facial identity and expression exhibited in the images well and are represented with triangulated meshes that have a uniform topology. To introduce more variations in head pose, 300W-LP rotated the reconstructed 3D faces with multiple view angles and generated the corresponding RGB face images through image warping. This yields a dataset which contains more than 122K face images and their corresponding 3D face data. 300W-LP also provides the weak perspective projection to align each 3D facial mesh with the face in the image:

$$V_p = f \times P_r \times R \times V + T_{2d} \tag{4-11}$$

where V_p is the projected 3D face with its depth channel removed, V is the reconstructed 3D face, R is the rotation matrix, P_r is the orthographic matrix

$\begin{pmatrix}1&0&0\\0&1&0\end{pmatrix}$, f is the scale factor and T_{2d} is the translation vector defined on the 2D image plane.

With Eq.4-11, the 2D image pixel coordinates of each 3D facial vertex can be easily found. For a pixel in the depth view, this chapter finds out the 3D vertex that is projected onto it and visible to the camera using Z-Buffer, then fill in the pixel value with the found 3D vertex's depth value. A depth view aligned with the face image can be acquired after going through all pixels with the operation above. To reduce the size of AGGAN for more efficient training, all synthetic depth views are resized to 128×128 and the aligned 3D facial meshes are shrunk accordingly. Considering real-world depth views are noisy, random Gaussian noise is further added to the synthesized depth views. For the facial mesh, the neck and the ear part are removed to focus on the main face region. The resulting mesh contains about 35K vertices. Inspired by previous work on 3D shape reconstruction (Jackson et al., 2017; B. Yang et al., 2018), the voxel grid is used to preserve the 3D facial geometry. In particular, the facial mesh is voxelized to a $128 \times 128 \times 128$ grid aligned with the depth view. Comparing with the vector, the voxel grid models 3D geometry in a way much closer to the real-world representation.

4.4 Experiments

4.4.1 Experimental Setup

a) Datasets

Depth views synthesized from HELEN are used for training AGGAN, while the rest depth views synthesized from 300W-LP (X. Zhu et al., 2016) are used for testing. In total, there are 75,352 training samples and 47,098 testing samples. As mentioned in *Data Synthesis*, 300W-LP includes a variety of natural facial expressions and has been augmented to cover a wide range of head poses, e.g. with

yaw angles ranging from -90° to 90°. The synthetic depth views have also been perturbed with random Gaussian noise to further simulate imperfections in real depth views.

b) Implementation Details

The generator and discriminator of AGGAN are optimized in an alternate manner. The discriminator is updated with one gradient descent step, after which the generator is updated with two gradient descent steps. λ is set as 5 for gradient penalty in $\mathcal{L}_{adv}^{d}$. α, β and γ are set as 20, 100 and 20 respectively, which produces promising results in our experiment. The Adam solver is used for both the generator and discriminator with a batch size of 1.

c) Evaluation Metrics

Two metrics are used to quantify the difference between the predicted 3D facial voxel grid and the ground truth.

1) Mean Intersection-over-Union (IoU) (B. Yang et al., 2018):

$$IoU = \frac{\sum_{i=1}^{N}[C(y_i>T)\times C(\hat{y}_i)]}{\sum_{i=1}^{N}\left[C\left(C(y_i>T)+C(\hat{y}_i)\right)\right]} \tag{4-12}$$

where $C(\cdot)$ is an indicator function, y_i is the predicted occupancy state of the ith voxel, $\hat{y}_i$ is the corresponding ground truth, T is the threshold for voxelization, and N is the number of voxels in the grid. T is set as 0.5 in our experiments. The higher the IoU value, the better the 3D facial geometry recovery.

2) Mean value of standard Cross-Entropy loss (CE) (B. Yang et al., 2018):

$$CE = -\frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i \log(y_i) + (1-\hat{y}_i)\log(1-y_i)] \tag{4-13}$$

where N, y_i and $\hat{y}_i$ are the same as in (4-12). The lower the CE level is, the closer the 3D prediction to be either '0' or '1', which indicates a more robust and confident prediction.

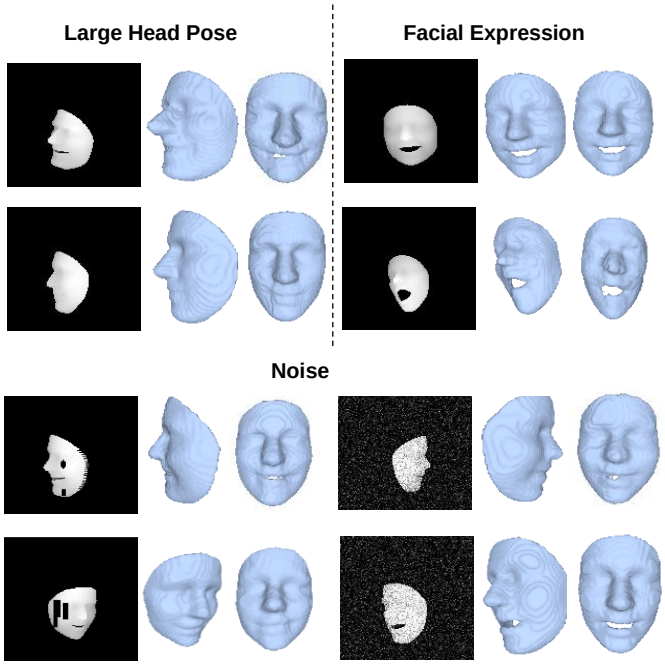

Figure 4.4: Example results of AGGAN for depth views with large head pose, facial expression and noise.

Table 4.1: IoU and CE values of testing results.

	IoU	CE
AFW (10414 samples)	0.9916	0.0517
IBUG (3572 samples)	0.9937	0.0490
LFPW (33112 samples)	0.9913	0.0523

4.4.2 Results

IoU and CE values calculated on the predictions of LFPW, IBUG and AFW in 300W-LP are reported in Tab. 4.1. Meanwhile, in Fig. 4.4 there are some visual results of the recovered 3D facial geometry for qualitative evaluation. As shown in Fig. 4.4, AGGAN can recover the 3D facial geometry well for different head poses, facial identities and expressions, and even when there are random noises or problematic holes in the given depth view. Hausdorff distance (two sets are close in the Hausdorff distance if every point of either set is close to some point of the

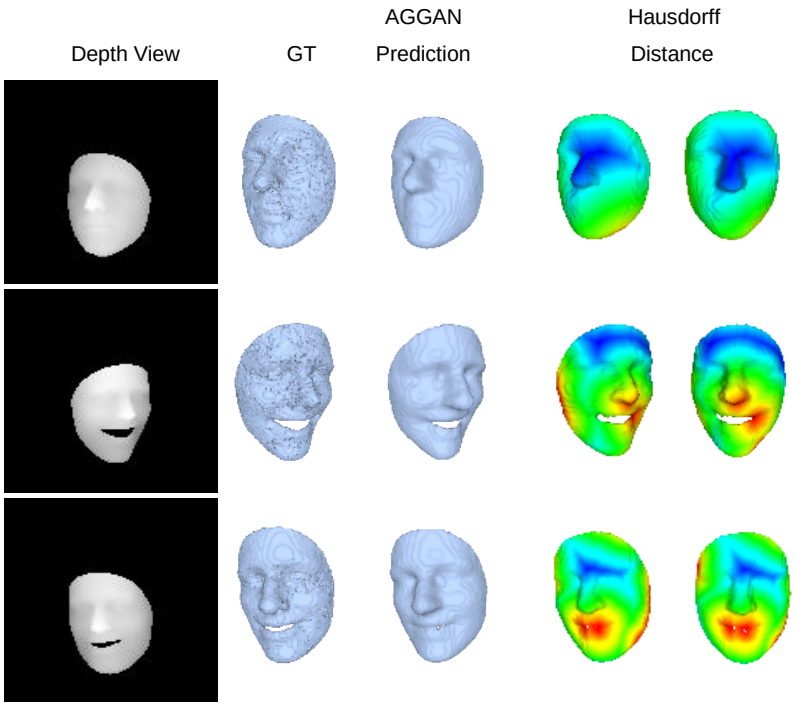

Figure 4.5: Comparison between the AGGAN prediction and the ground truth (GT). The Hausdorff distance between the GT and prediction is calculated and colorized on the predicted 3D face. Please note that the distance value increases from red to blue.

other set) between the predicted voxel grid and its ground truth is calculate and visualized (see Fig. 4.5, the distance value increases from red to blue). To further prove the accuracy of the facial identity prediction, the 3D results predicted from depth views with an identical facial identity but projected under different head poses are shown in Fig.4.6.

Comparison with Existing Methods. The proposed AGGAN is compared with some representative 3D surface reconstruction and non-rigid registration methods, including Marching Cubes (MC) (Lorensen & Cline, 1987a), Screened Poisson Surface Reconstruction (SPSR) (Kazhdan & Hoppe, 2013) and non-rigid ICP (NICP) (Sumner & Popović, 2004). For algorithms such as NICP that require connectivity, the Ball-Pivoting algorithm (Bernardini et al., 1999) is used to compute a triangle mesh interpolating the given facial point cloud. To get a

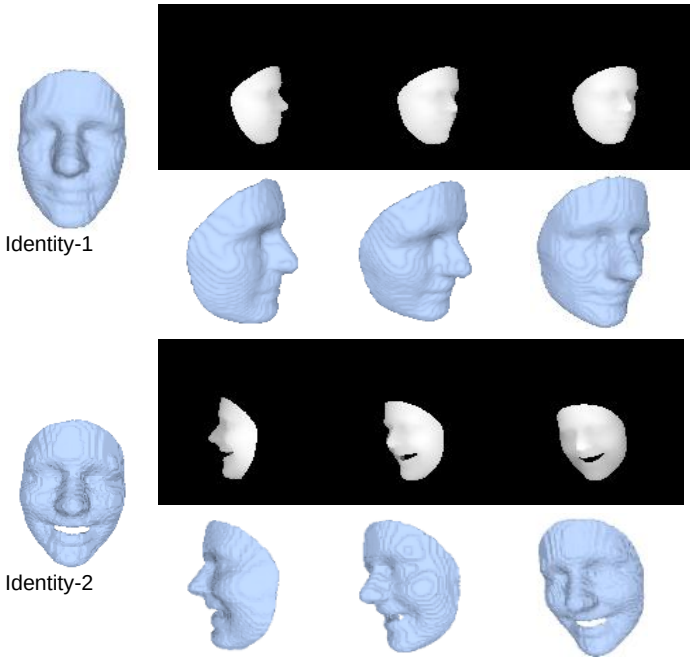

Figure 4.6: Results of AGGAN predicted from depth views with an identical facial identity however with different head poses.

promising result for NICP (Sumner & Popović, 2004), ICP is first applied to rigidly align the facial template with the given facial point cloud, then initialized the non-rigid registration with hand-selected facial landmarks. Since each aforementioned method reconstructs the 3D face in a distinct topology whose vertex amount and connectivity are different from each other, thus cannot be compared using IoU and CE. Alternatively, the visual comparison results are showcased in Fig.4.7. It can be seen 3D faces recovered by previous methods are severely distorted (Fig.4.7), when the input depth view is in a large head pose, incomplete and with prominent artefacts. In contrast, AGGAN is much more robust to data imperfections in the depth view and able to generate a complete and smooth 3D facial geometry with facial identity and expression well preserved. What's more, as shown in the third and fifth row in Fig.4.7, AGGAN can normally generate a 3D face smoother and denser than the ground truth since it predicts the probability of each voxel occupancy within the range of $[0, 1]$ continuously.

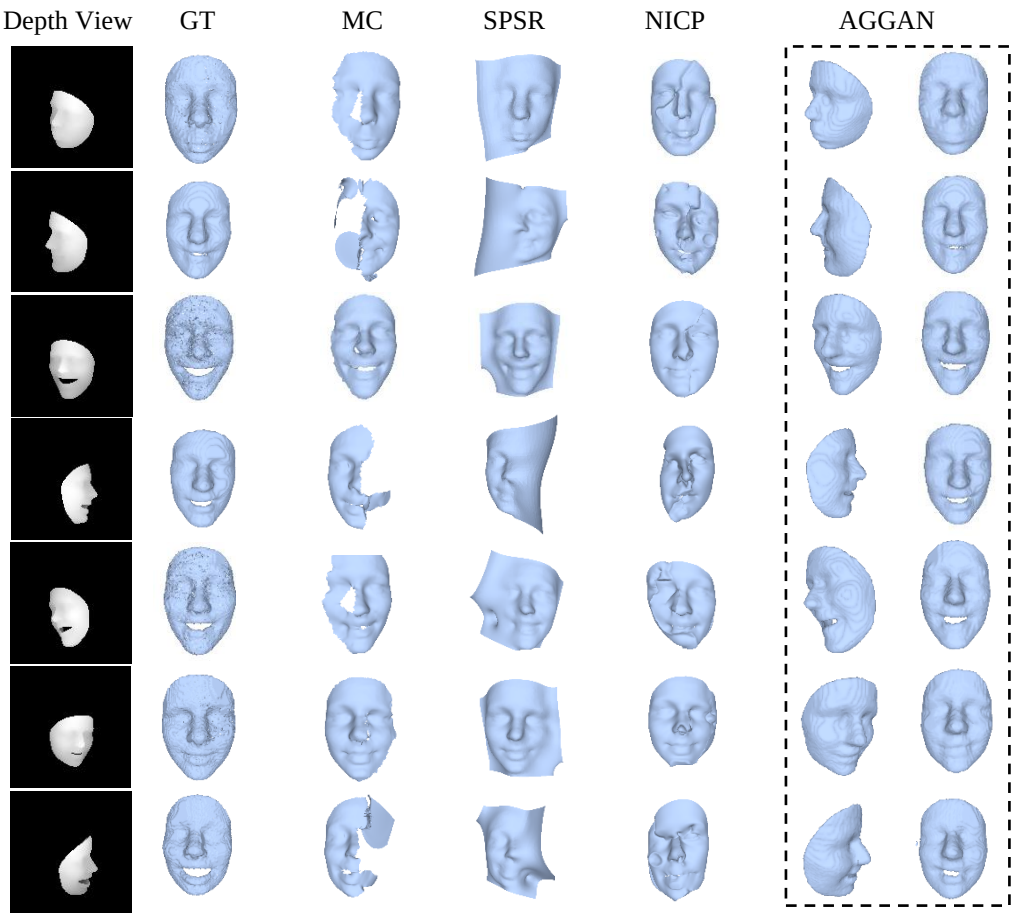

Figure 4.7: Comparison between AGGAN and existing methods on challenging depth views.

Table 4.2: Results of ablation study on a subset of IBUG.

Attention	Sparsity	IoU	CE
		0.9917	0.1151
×		**0.9932**	**0.0940**
	×	0.9928	0.0995
×	×	0.9927	0.1004

Ablation Analysis. The importance of the sparsity constraint and the attention module is investigated. Specifically, four different AGGAN models which cover all possible combinations (with/without sparsity and with/without attention) of the

two modules were trained on HELEN (Zhou et al., 2013) and tested on a subset of IBUG (Sagonas et al., 2013)- a challenging dataset which contains facial images of very large head pose and facial expression. IoU and CE levels of these four models on the testing set are listed in Tab.4.2. It can be found that both sparsity and attention help improve AGGAN's prediction accuracy when they work independently. Moreover, the model with the attention module outputs the best result, which verifies the significance of attention in AGGAN and implies that there might conflict between sparsity and attention during the network learning process.

4.4.3 Limitations and Prospect

A few artefacts can be observed around the facial boundary in the predictions of AGGAN. For example, as shown in Fig.4.4, the inner mouth region cannot be fully recovered when the facial expression is a big open mouth. This is mainly due to that the voxel grid used is not dense enough. When voxel occupancy states were predicted mistakenly, the resulted due to that the voxel grid used is not dense enough. When voxel occupancy states were predicted mistakenly, the resulted artefacts would be obvious. This problem can be alleviated by using a denser voxel grid or applying a better prior to restrict the voxel occupancy state for forming a reasonable face, e.g. using a mean face with neutral facial expression as a template grid and driving AGGAN to predict the difference between the template and the target face. Although AGGAN has been validated on the synthetic data, it shows its potential for the application of real depth views captured from depth cameras. For example, as shown in the 3rd and 4th column of Fig.4.4, AGGAN can recover 3D facial geometry accurately when there are random noises or even problematic holes (please note that these holes were not simulated in the training data) in the depth view. To fill in the gap between the synthetic data and real data, a promising direction is to train a network learning the common feature representation of the synthetic and real depth views. In this way, the synthetic data can be sufficiently utilized while much less real depth views will need to be collected.

4.5 Conclusion

This chapter proposes to model the ill-posed 2.5D facial depth-3D mapping with a novel attention-guided GAN structure - AGGAN in a data-driven manner. AGGAN is validated on synthetic depth views which cover a wide range of facial identities, expressions and head poses. When dealing with noisy and non-frontal facial depth views, AGGAN is still capable of recovering the 3D structure of the missing/occluded facial parts with facial identity and expression being accurately preserved, and thus significantly outperforms previous methods. Moreover, AGGAN is resilient to data imperfections in the depth view such as random noise and problematic holes, and hence has a potential of being applied to real depth views captured by depth cameras.

Chapter 5

Domain Adaptive Single Depth Image 3D Face Reconstruction

5.1 Introduction

Reconstructing dense 3D facial geometry from visual input is crucial for many face applications such as face manipulation (Thies et al., 2016)and facial animation (Cao et al., 2015). The vast majority of existing approaches are developed for reconstruction from ubiquitous RGB face images(Lin et al., 2020; W. Zhu et al., 2020). However, those approaches cannot cope well with image degradations induced by poor lighting, let alone the corresponding 2D-to-3D reconstruction process is ill-posed by nature and prone to generating implausible results when the head pose is large. In contrast, reconstruction from depth images is more robust to adverse lighting and pose conditions, since the depth image directly captures 3D geometric information. It tends to be one of the key technologies required in consumer face applications, especially after the advent of more advanced and lightweight depth cameras, e.g. Apple TrueDepth. However, depth-based 3D face reconstruction is challenging due to the imperfect depth data caused by device noise and natural modes of variations such as head pose and expression. It becomes more intractable if there is only a single depth image available, a setting that is common in real-world applications.

The problem above is mostly about adapting a source 3D face to fit with the target depth image. The 3D face can be represented as a mesh (Paysan et al., 2009), a group of discrete voxels (Jackson et al., 2017) or the level set of an implicit function (e.g. signed distance function) (Q. Xu et al., 2019). Traditional approaches (Amberg et al., 2007) address the fitting problem with an optimization

process constrained by hand-crafted priors, e.g. point-to-point correspondences between the source and the target. In practical settings, valid priors are difficult to acquire in a purely automated manner, as the real depth image is normally noisy, sparse and contains occlusions. More recent methods (Zhong et al., 2020) instead use deep neural networks to learn such prior knowledge from a training corpus, and exploit it for accurate inference (or reconstruction) from an unconstrained depth image during testing. However, the data-driven learning framework they applied usually consumes a great amount of depth-3D training pairs, whose collection procedure is expensive and laborious. This problem can be alleviated by utilizing synthetically generated depth images for training, but closing the synthetic-real domain gap is nontrivial.

To tackle the aforementioned issues, a novel domain-adaptive 3D face reconstruction method is proposed in this chapter. The proposed method requires only synthetic and unlabelled real depth images for training, while the trained model can generalize well to images captured with commodity depth sensors such as Kinect. Its core is a disentangled domain-adaptive neural network which consists of two sub-networks – PoseNet and ShapeNet for predicting head pose and facial shape under a canonical pose respectively. In 3D facial geometry, head pose represents the rigid component, while facial shape represents the non-rigid component. It implies that these two geometry attributes lie in different fields. Based on this insight, this chapter proposes to tailor the domain adaptation approach for training the two sub-networks. Specifically, PoseNet is trained with synthetic data first, then fine-tuned on real depth images. ShapeNet instead employs a more complicated adversarial domain adaptation framework during training (Sankaranarayanan et al., 2018). This disentangled learning process differs from the previous methods(Zhong et al., 2020) which applied a unified domain-adaptive network for pose and shape estimation. It effectively simplifies the source (synthetic) and target (real) distributions, thus reducing the complexity of the domain shift problem (S. J. Pan & Q. Yang, 2009).

Inspired by (Y. Guo et al., 2018), the depth image was converted into 3D point

coordinates in camera space with a known camera intrinsic matrix before being fed to the reconstruction network. In this way, the network can directly infer head pose from the corresponding position image. On the network's output side, the 3D vertex offsets from a neutral mean face are adopted for a more concentrated data distribution to reduce the learning difficulty. These two operations naturally adapt the reconstruction network to 3D data.

The proposed method is extensively evaluated on challenging benchmark datasets – FaceWarehouse (Cao et al., 2013), Biwi (Fanelli et al., 2011), ICT-3DHP (Baltrušaitis et al., 2012), which contain noisy real depth images covering a wide range of head poses and facial expressions. The experimental results show proposed method outperforms the state-of-the-art (Deng et al., 2019; Lin et al., 2020; Martin et al., 2014; Zhong et al., 2020) in both pose and expression estimation. In summary, the main contributions in this chapter are:

- A novel disentangled domain-adaptive network is proposed for single depth image 3D face reconstruction. The proposed network decouples the prediction of head pose and facial shape into two subnetworks which are trained with different unsupervised domain adaptation methods.
- A robust reconstruction pipeline is designed, explicitly handling 3D data by using the position image as input and 3D vertex offsets as output.
- The successful reconstructions of 3D facial meshes from very sparse and noisy real depth maps with the proposed method is evaluated.

5.2 Related Work

5.2.1 3D Face Reconstruction from a Single Depth Image

a) Traditional Methods

In traditional computer graphics approaches, the reconstruction problem is solved with an optimization process that adapts a voxel grid, an implicit function or a 3D facial mesh to fit with the point cloud embedded in the depth image. Accordingly,

the reconstructed 3D face is the intersection between the voxel grid and the point cloud (Lorensen & Cline, 1987b) , an iso-surface represented by the level set of the implicit function (Kazhdan & Hoppe, 2013), or a deformed 3D mesh (Amberg et al., 2007). For the first two cases, it will output an implausible reconstruction result (e.g. incomplete or unsmooth 3D facial surface), when the given depth image is noisy, sparse or contains a big head pose. For the last case, it requires valid point-to-point correspondences to constrain the highly nonlinear fitting process, a problem also known as non-rigid shape registration. Reliable correspondences usually have to be obtained by hand annotation, especially when the depth image is imperfect with a lot of noise and a large facial pose. However, this prerequisite cannot be satisfied in a fully automated reconstruction pipeline, where no manual intervention is allowed.

b) Deep Learning Methods

In recent years, using deep neural networks to reconstruct a watertight surface/mesh from an unordered set of sparse, noisy 3D points (e.g. those captured with the commodity depth sensor) is attracting increased interest (Chibane et al., 2020; Groueix et al., 2018; W. Wang, Ceylan, et al., 2019; Y. Wang et al., 2020). The relevant methods first encode the target point cloud using a latent feature vector, from which they then predict the 3D shape deformation represented by either per-vertex displacements from the source shape (Groueix et al., 2018; W. Wang, Ceylan, et al., 2019) or an implicit field (Chibane et al., 2020; Y. Wang et al., 2020). Typical feature encoders are PointNet (Qi et al., 2017) for unstructured point cloud or a common convolutional neural network for discrete voxels and images (Chibane et al., 2020). It can be found that existing methods rarely did pose estimation during reconstruction. Instead, they required the source and target shapes to be rigidly aligned with each other beforehand. This conflicts with most real-world scenarios in which the pose is unknown and should be estimated as well. What's more, most of the methods was designed to reconstruct 3D shape for objects and human/animal bodies, recovering only coarse-grained shape features. It is unclear if they can also handle the human face, whose fine-scale geometry

details are essential to maintain its expressive power and should be well preserved during reconstruction.

To the best of our knowledge, there is only one method (Zhong et al., 2020) has been developed till now for learning 3D face from a single depth image. The method employed a CycleGAN-based domain adaptation framework to learn domain-invariant features from synthetic and real depth images for estimating 3D facial geometry parameters. The learned network showed good 3D reconstruction performance when testing on noisy real depth images. However, the method adopted a unified domain-adaptive network for inferring head pose and facial shape, which actually have different distributions as they encode rigid and non-rigid geometry components respectively. This chapter proposes to decouple the pose and shape estimation with two different domain-adaptive networks. Specifically, the fine-tuning strategy is utilized to fill the synthetic-real domain gap when training the pose prediction network, while applying the adversarial domain adaptation approach to train the more intractable facial shape prediction network.

5.2.2 Unsupervised Domain Adaptation

Unsupervised domain adaptation (Wilson & D. J. Cook, 2020) aims to learn a predictive model that can perform well on target domain using only labelled source and unlabelled target samples for training. It is helpful to many practical learning tasks where the target data labels are difficult to acquire, e.g. in our case, obtaining the ground-truth 3D facial geometries of depth images is costly and tedious. A popular approach to this problem is constructing a common representation space in which the two domains are close to each other. In the context of deep learning, this can be achieved by either fine-tuning a pre-trained network on unlabelled target data (K. Wang et al., 2020) or applying an adversarial loss in the representation space (Sankaranarayanan et al., 2018), depending on the severity of domain shift. Experimental results show that directly applying the head pose prediction network trained on synthetic data to real depth images already can

produce promising results. After tuning the network on real depth images, the pose prediction accuracy can be further improved to a much higher level. It can be ascribed to the rigid nature of head pose, which makes the conditional probability distributions of pose given a depth image are naturally close to each other between the synthetic and real domains. However, this intuitive strategy doesn't work on training the facial shape prediction network which models non-rigid geometry variations. To solve this problem, following (Sankaranarayanan et al., 2018), an auxiliary GAN is adopted to further push the shared feature embedding to be domain invariant. Specifically, the generator is trained to produce source-like images from the embedding, while the discriminator is trained to not only distinguish input source images from generated images, but also force the generated image to preserve the same shape label as that of the input source image. With this disentangled domain adaptation learning framework, the 3D face reconstruction network is able to generalizes well to real depth images.

5.3 Method

This section elaborates the proposed method from the following four aspects: 3D face representation, depth image pre-processing and domain-adaptive 3D reconstruction network.

5.3.1 3D Face Representation

In this chapter, a dense triangle mesh is applied to represent 3D facial shape. The mesh is the combination of a point cloud $\mathbf{V} = [{\mathbf{v}_1}^T, {\mathbf{v}_2}^T, \cdots, {\mathbf{v}_n}^T]^T$ of $n = 29{,}678$ vertices $\mathbf{v}_i = [x_i, y_i, z_i]^T$ and a predefined connectivity (Paysan et al., 2009). For head pose, the quaternion $\mathbf{q} \in \mathbb{R}^4$ is utilized to represent rotation and $\mathbf{t} \in \mathbb{R}^3$ is used to represent translation respectively.

5.3.2 Depth Image Pre-processing

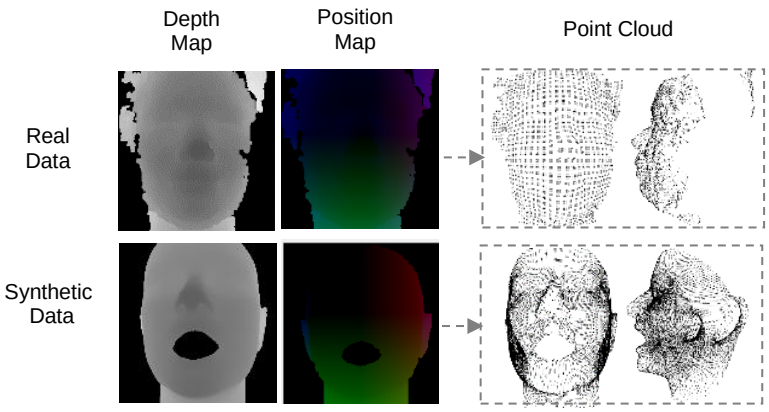

Figure 5.1: The pre-processing and presentation of input data.

Directly feeding depth image and camera intrinsic parameters into the reconstruction network is intuitive and feasible. However, the semantic meanings of camera parameters and their relationship with the depth data are somewhat obscure to the network. To tackle this issue, this chapter follows (Y. Guo et al., 2018) and adopts an additional pre-processing step on the original depth image with the camera intrinsic matrix, which converts all pixel values into the corresponding 3D points' coordinates in camera space. This results in a position image (see Fig.5.1) that explicitly stores 3D geometric information, while can be easily processed with 2D convolutions by the reconstruction network.

5.3.3 Domain-adaptive 3D Reconstruction Network

Harvesting real depth images with accurate 3D facial geometry label is time-consuming and expensive. This makes it impractical to develop a robust 3D face reconstruction network with fully supervised training. To address this problem, this chapter develops a novel domain adaptation framework that utilizes synthetically generated data and unlabelled real depth images to train the reconstruction network (see Fig.5.2). Considering head pose (rigid) and facial shape (non-rigid) model different components of 3D facial geometry, the developed framework decouples the prediction of these two attributes into two

subnetworks – ShapeNet and PoseNet, with each is trained by a custom domain adaptation method. The resulting 3D reconstruction network is domain-adaptive and can generalize well to noisy real depth images captured with commodity depth cameras. It achieves competitive accuracy in predicting both the head pose and facial shape when comparing with the state-of-the-art approaches(Deng et al., 2019; Feng et al., 2018; Jackson et al., 2017; Lin et al., 2020; Zhong et al., 2020).

a) ShapeNet

The geometric variation is diverse and spans widely in 3D space. To narrow down its distribution, we factor out the rigid component – head pose and concentrate on predicting 3D shape in this subnetwork - ShapeNet. To further reduce the learning difficulty, ShapeNet is designed to predict vertex displacements $\Delta\mathbf{V} = \left[\Delta\mathbf{v}_1{}^T, \Delta\mathbf{v}_2{}^T, \cdots, \Delta\mathbf{v}_n{}^T\right]^T$ from a reference 3D face $\overline{\mathbf{V}}$ under a canonical pose. The reference face is the mean face calculated from the Basel Face Model (BFM)(Paysan et al., 2009), and being rotated along the X axis by π to make it head up and face towards the negative Z axis.

Intuitively, the overall goal of ShapeNet is to learn from the position image space $\mathbf{X} = \{x^i\}_{i=1}^N$ a feature embedding map $F: \mathbf{X} \to \mathbb{R}^{d_F}$ and a regression function $R: \mathbb{R}^{d_F} \to \mathbf{Y}$ ($\mathbf{Y} = \{\Delta\mathbf{V}^i\}_{i=1}^N$). F and R can be easily modelled as deep neural networks and trained in a fully supervised manner given synthetic position images (position images of synthetic depth images) and their ground-truth shape labels. However, the model trained purely on the synthetic domain can hardly be applied to real position images (position images of real depth images) due to the domain gap between the synthetic and real data. To make the model adaptive to different domains particularly the real domain, this chapter employs a generate-to-adapt method (Sankaranarayanan et al., 2018) which augments F and R with an auxiliary GAN during training (see Fig.5.2). The auxiliary GAN is mainly tasked with generating synthetic-like images from the joint feature embedding output from F for both synthetic and real position images. It doesn't require ground-truth shape labels for real position images. Within such a training process, the joint

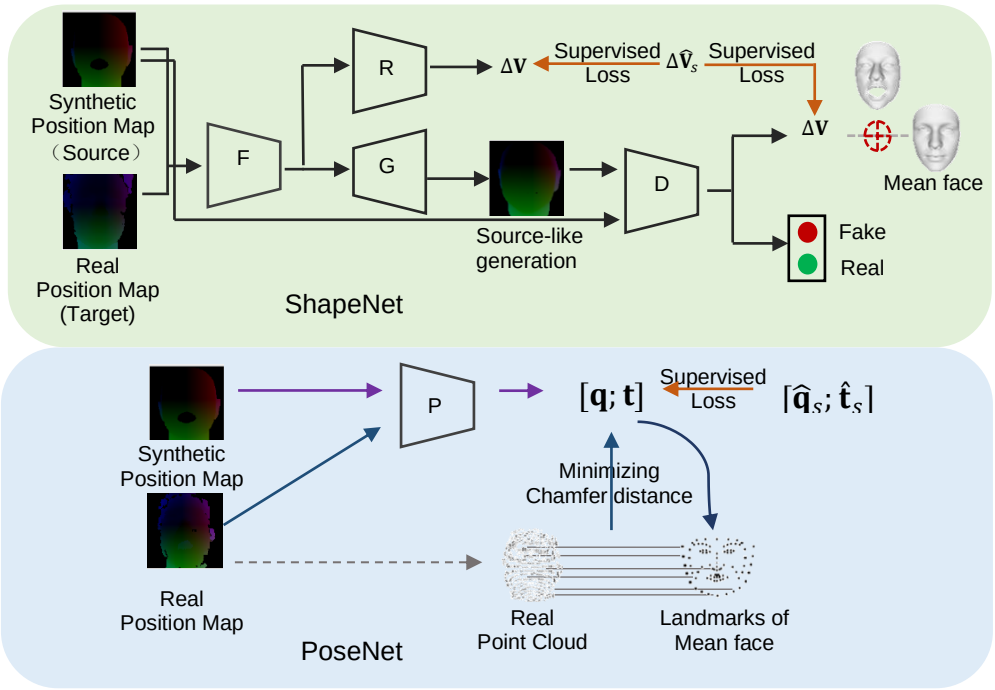

Figure 5.2: The Framework of proposed approach.

embedding continuously incorporates information from the real domain, while keeping aligning with the synthetic domain. In the test phase, the auxiliary GAN model is discarded and only the trained F - R pair is used to do 3D face reconstruction (see Fig.5.2). To better align with the context of domain adaptation, in the following, the terms source and target domains (images) are used to represent the aforementioned synthetic and real domains (images) respectively. For consistency, the source images, their ground-truth shape labels and the target images are then denoted as $\mathbf{X}_s = \{x_s^i\}_{i=1}^{N_s}$, $\hat{\mathbf{Y}}_s = \{\Delta\hat{\mathbf{V}}_s^i\}_{i=1}^{N_s}$ and $\mathbf{X}_t = \{x_t^i\}_{i=1}^{N_t}$.

The original generate-to-adapt method proposed in (Sankaranarayanan et al., 2018) adopts an auxiliary classifier GAN (AC-GAN) (Odena et al., 2017), which is designed for classification purpose. To solve the regression problem in this study, the multi-class classifier in AC-GAN is replaced with a regressor customed for predicting $\Delta\mathbf{V}$ and propose the auxiliary regressor GAN (AR-GAN). The key features of AR-GAN are as follows:

(a) The generator G accepts the concatenation of the feature embedding $F(x)$, a random noise $z \in \mathbb{R}^{d_z}$ sampled from $\mathcal{N}(0,1)$ and a one hot vector $l \in \mathbb{R}^2$ encoding whether the position image x is from the source domain or not as input. Note that explicitly conditioning G with the one hot vector is to exclude the information of image type (source or target) from the feature embedding. It is supposed to facilitate the learning of a common, semantically-consistent representation of source and target position images, thereby enhancing the generalization ability of the subsequent shape regression R.

(b) The discriminator D takes the source position image x_s or the generated image $G([F(x), z, l]), x \in \{\mathbf{X}_s, \mathbf{X}_t\}$ as input (for clarity, we use $x_{sg} = G([F(x_s), z, l])$ and $x_{tg} = G([F(x_t), z, l])$ to represent the image generated from x_s and x_t respectively), and makes two predictions: i) $D_{adv}(x)$ - the probability of x being a real source image, which is the core outcome of the GAN's adversarial training procedure. ii) $D_{reg}(x_{sg})$ - the $\Delta\mathbf{V}$ regressed from the generated image. It is designed for source generated image x_{sg} only whose ground-truth shape label $\Delta\widehat{\mathbf{V}}_s$ is available. As shown in previous studies(Odena et al., 2017), forcing D to predict side information ($\Delta\mathbf{V}$ in our case) can improve the model's performance on the original image generation task. The corresponding gradients can further be propagated to update the parameters of F.

Training ShapeNet is nontrivial as it involves four components (F, R, D, G) and two different kinds of training data (labelled synthetic position images and unlabelled real position images). In this study, the AR-GAN which lies at the core position of aligning the two domains is optimized in the first place, then move on to optimizing F and R.

Based on the two predictions of D, the loss function of D - L_D comprises two parts - $L_{D_{adv}}$ and $L_{D_{reg}}$:

$$L_{D_{adv}} = \begin{Bmatrix} \mathbb{E}_{x_s \sim \mathbf{X}_s} - \left[\log\left(D_{adv}(x_s)\right) + \log\left(1 - D_{adv}\left(x_{sg}\right)\right)\right] \\ + \ \mathbb{E}_{x_t \sim \mathbf{X}_t} - \left[\log\left(1 - D_{adv}\left(x_{tg}\right)\right)\right] \end{Bmatrix} \tag{5-1}$$

$$L_{D_{reg}} = \mathbb{E}_{x_s \sim \mathbf{X}_s}\left[\text{smooth}_{L_1}\left(D_{reg}\left(x_{sg}\right) - \Delta\widehat{\mathbf{V}}_s\right)\right] \tag{5-2}$$

D is trained to minimize $L_D = L_{D_{adv}} + L_{D_{reg}}$. G is then updated to generate realistic source images while preserving their original shape labels by minimizing the following loss function:

$$L_G = \mathbb{E}_{x_s \sim \mathbf{X}_s}\begin{bmatrix} \log\left(1 - D_{adv}\left(x_{sg}\right)\right) \\ + \text{smooth}_{L_1}\left(D_{reg}\left(x_{sg}\right) - \Delta\widehat{\mathbf{V}}_s\right) \end{bmatrix} \tag{5-3}$$

After optimizing D and G, we update R's parameters using source images and their shape labels in a fully supervised manner. The loss function for minimization is as follows:

$$L_R = \mathbb{E}_{x_s \sim \mathbf{X}_s}\left[\text{smooth}_{L_1}\left(R\left(F(x_s)\right) - \Delta\widehat{\mathbf{V}}_s\right)\right] \tag{5-4}$$

Finally, F is updated using the gradients back-propagated from the optimization procedures of D, G and R. The overall objective of F is to learn a feature embedding that not only keeps the dominant information of shape labels, but also is shared by the source domain and the target domain. Such an objective can be concretized as learning an embedding that maximizes the prediction accuracy of R and makes the images generated by G indistinguishable from each other (in this study, this is achieved by making all the generated images source-like). The resulting objective function is defined as follows:

$$L_F = L_R + \lambda_1 L_{D_{adv,t}} + \lambda_2 L_{D_{reg}} \tag{5-5}$$

$$L_{D_{adv,t}} = \mathbb{E}_{x_t \sim \mathbf{X}_t} - \left[\log\left(1 - D_{adv}\left(x_{tg}\right)\right)\right] \tag{5-6}$$

Minimizing the first item of L_F encourages the learned embedding to incorporate the essential shape information, while minimizing the last two items leads the embedding to be domain-invariant. Specifically, minimizing $L_{D_{adv,t}}$

makes the image generated from the target domain as source-like as possible. Similarly, minimizing the discriminative loss $\mathbb{E}_{x_s \sim \mathbf{X}_s}\left[\log\left(1 - D_{adv}\left(x_{sg}\right)\right)\right]$ will force the image generated from the source domain to be source-like, which can also be used to update F. However, such a loss cannot guarantee that the learned source embedding keeps the original shape information. To solve this problem, we use the regression loss $L_{D_{reg}}$ instead to update F. In equation (5-5), λ_1 and λ_2 are weighting coefficients for balancing the three energy items. They are empirically set in our experiments. According to the updating order above, D, G, R, F are updated iteratively until the corresponding loss converges to an acceptable level.

b) PoseNet

From the perspective of rigid head pose, the domain gap between the source and target position image distributions is assumed to be small. Based on this hypothesis, instead of using the complex adversarial domain adaptation method, we propose to fine-tune a neural network pre-trained on labelled source data with unlabelled target data to achieve domain-adaptive head pose estimation from a single position image. The resulting network P (see Fig.5.2) is named as PoseNet. It is a convolutional neural network which accepts a position image as input and outputs a rotation quaternion $\mathbf{q}$ and a translation vector $\mathbf{t}$. The training of PoseNet involves two stages. At the first stage, we use source images $\mathbf{X}_s = \left\{x_s^i\right\}_{i=1}^{N_s}$ and their pose labels $\left\{\hat{\mathbf{q}}_s^i, \hat{\mathbf{t}}_s^i\right\}_{i=1}^{N_s}$ to train PoseNet by minimizing the following loss function:

$$L_P^s = \mathbb{E}_{x_s \sim \mathbf{X}_s}\left[\text{smooth}_{L_1}\left(\left(P(x_s) - [\hat{\mathbf{q}}_s; \hat{\mathbf{t}}_s]\right) \circ [\alpha; 1]\right)\right] \tag{5-7}$$

α is a weighting coefficient for balancing between the rotation quaternion and the translation vector. The primed PoseNet is then tuned with unlabelled target images $\mathbf{X}_t = \left\{x_t^i\right\}_{i=1}^{N_t}$ via minimizing the landmark-based Chamfer distance,

$$L_P^t = \sum_{i=1}^{m} \min_{\mathbf{v}_t \in \mathbf{V}_t} \|\mathbf{v}_t - \Phi(\bar{\mathbf{v}}_i)\|_2^2 \tag{5-8}$$

where $\mathbf{V}_t$ is the noisy point cloud extracted from the target position image x_t, $\bar{\mathbf{v}}_i$ is the ith landmark (see Fig.5.2 (PoseNet) for the applied landmark markup) of the transformed reference 3D face $\bar{\mathbf{V}}$ and $\Phi(\cdot)$ represents the rigid transformation driven by the predicted pose vector $P(x_t)$. When $\min_{\mathbf{v}_t \in \mathbf{V}_t} \|\mathbf{v}_t - \Phi(\bar{\mathbf{v}}_i)\|_2^2 > \epsilon$, $\mathbf{v}_t$ is treated as a noisy point. The corresponding Chamfer distance is thus invalid and will not be counted. Through registering the two groups of landmarks as shown in the equation (5-7), PoseNet can be updated towards predicting more accurate head pose parameters, which in turn will facilitate the seeking of more matchable landmarks on the target point cloud $\mathbf{V}_t$. As this self-supervised training procedure iterates, PoseNet continuously generalizes to target position images.

5.4 Experiments

The proposed method is evaluated in this section. It is validated on three mainstream public datasets and compared with the state-of-the-art in both 3D facial shape reconstruction and head pose estimation.

5.4.1 Implementation Details

a) Network Structure

As illustrated in Fig.5.2, the proposed ShapeNet comprises four parts - F, R, D and G:

(a) F is an encoder that accepts a 160x160x3 position image of a human face as input and outputs a 512-dim feature vector. It consists of 11 convolutional layers, with each except for the last layer is followed by batch normalization(Ioffe & C. Szegedy, 2015) and leaky ReLU(Maas et al., 2013). For the first 10 convolutional layers, there are two types of convolutions. One uses a kernel of 4x4 and a stride of 2, the other uses a kernel of 3x3 and a stride of 1. The two different convolution operations are called in an alternating manner. F's last convolutional layer applies a kernel of 5x5 and a stride of 1, and is followed by leaky ReLU only.

(b) R is a three-layer MLP. It maps the concatenation of the 512-dim feature embedding output from F, a random noise vector and a one hot vector, first to a 1024-dim vector, and then to an 89,034-dim vector which saves the 3D vertex displacements. R's last layer and its previous layer is connected by leaky ReLU.

(c) G is implemented with a convolutional neural network whose structure is mostly symmetric to the F's. The main difference is that G's last convolutional layer is followed by a Tanh function instead of the leaky ReLU for the activation unit.

(d) D first applies a F-like structure to learn from the generated position image a 512-dim feature embedding. Then it maps the embedding to not only the probability of being a real position image through a fully connected layer and a sigmoid function, but also an 89,034-dim displacement vector through a MLP that has the same structure as that of R.

PoseNet is a typical convolutional neural network, which can be viewed as a combination of ShapeNet's F and R. PoseNet's first part applies the same structure as that of F, while its second part is a three-layer MLP similar as R. The MLP maps the 512-dim feature embedding sequentially to a 256-dim vector and a 7-dim vector which can directly be divided into a rotation quaternion and a translation vector.

b) Training

The proposed networks require both labelled synthetic (source) data and unlabelled real (target) data for training. For ShapeNet training, a batch of synthetic position images along with their labels are first fed into it to update its D, G, R, F sequentially by minimizing the losses defined in Eq.5-1 to Eq.5-5. Then, keeping G and R fixed, D and F are further updated using a batch of real position images. When updating F, the weighting coefficients λ_1 and λ_2 in Eq.5-5 are empirically set as 0.03 and 0.1. The batch size is set to 16 through this and all the following experiments. ShapeNet is optimized via Adam with an initial learning rate of 0.0005 which is decreased at least by half every 10 epochs. In our

experiments, ShapeNet can produce satisfactory results after 30 training epochs applying the following combination of learning rates – {0.0005, 0.0001, 0.00005}. For PoseNet, we first use labelled synthetic data to optimize the network by minimizing the loss defined in Eq.5-6, where the weighting coefficient α compensating for the numerical scale difference between the rotation quaternion and the translation vector is set to 350. The network is then updated using unlabelled real data by minimizing the self-supervised loss defined in Eq.5-7 where the threshold ϵ for finding valid matching landmarks is set as 30(mm). Similar to ShapeNet, PoseNet is also optimized via Adam, but with a fixed learning rate of 0.0001. It is able to produce accurate pose estimation in about 150 training epochs with the first 100 epochs trained with synthetic data and the last 50 epochs fine-tuned with real data.

During the experiment, reusing the real data in each training epoch can facilitate the learning of a domain-adaptive model. It is very important especially when the number of real data is limited such as in our case. To achieve this goal, several copies of real position images are created in the training set and synthesize the same amount of position images for training. This means that each real position image will be used more than one time in an epoch and make more contributions to optimizing the network. Experimental results show that using a real position image three times in an epoch makes a good trade-off between the training cost and the resulting model's prediction ability.

c) Datasets

To train and evaluate the proposed domain-adaptative 3D face reconstruction method, three public datasets (see Tab.5.1) – FaceWarehouse (Cao et al., 2013), Biwi (Fanelli et al., 2011) and ICT-3DHP (Baltrušaitis et al., 2012) are used. The three datasets provide numerous real-world depth images which cover a wide range of facial expressions, head poses and identities. FaceWarehouse consists of 3,000 640x480 depth images captured from 150 human subjects. For each subject, 20 different facial expressions ranging from neutral expression, mouth stretch to eye closed, while posed near-frontally to the camera were captured with the Kinect

Table 5.1: The details of real datasets and synthetic datasets.

Dataset	Subjects	Expressions	Pose	Total Numbers
FaceWarehouse	150	20	Near Frontal	3,000
BiWi	20	X	Various	15,678
ICT-3DHP	10	X	Various	14,202
SynData1	450	20	Near Frontal	9,000
SynData2	165	3	Various	47,034
SynData3	143	3	Various	42,606

sensor. Biwi contains 15,678 640x480 depth images of 20 people turning their head with almost the same facial expression while captured by a range scanner. Its head pose range covers about $\pm 90^{\circ}$ yaw and $\pm 45^{\circ}$ pitch rotations. Similarly, ICT-3DHP contains 10 depth sequences (about 1,400 640x480 frames in each sequence and 14,202 frames in total) of 10 people turning their head while captured by the depth sensor. In this study, due to its diversity in facial expression, FaceWarehouse (Cao et al., 2013) is mainly used to assess the 3D facial expression reconstruction performance of the proposed method. Biwi (Fanelli et al., 2011) and ICT-3DHP (Baltrušaitis et al., 2012) instead are mainly used to assess the method's head pose estimation performance. In addition, these two datasets provide an accurate head pose label for each depth image, hence enabling precise quantitative evaluation on pose estimation.

This chapter also synthesizes depth images which are available with correct 3D facial geometry labels for training. BFM (Paysan et al., 2009) is applied to model facial identity and blendshapes generated from FaceWarehouse (Cao et al., 2013) to model facial expression. To ease the training difficulty of unsupervised domain adaptation, a specific dataset is synthesized to simulate and pair with each of the three real datasets. Such a synthetic dataset applies blendshape coefficients, head pose parameters and the camera intrinsic matrix which are estimated from (or provided by) the corresponding real dataset for 3D face generation and depth

image rendering. The synthetic datasets are named as SynData1, SynData2 and SynData3 which correlate with FaceWarehouse, Biwi and ICT-3DHP respectively. Tab.5.1 lists their details. It can be found from the table that the synthesized depth images are three times more than the real depth images. As discussed in the "*Training*" part, this is for reusing the real data during network optimization.

5.4.2 Results on Face Reconstruction

a) Comparison on FaceWarehouse

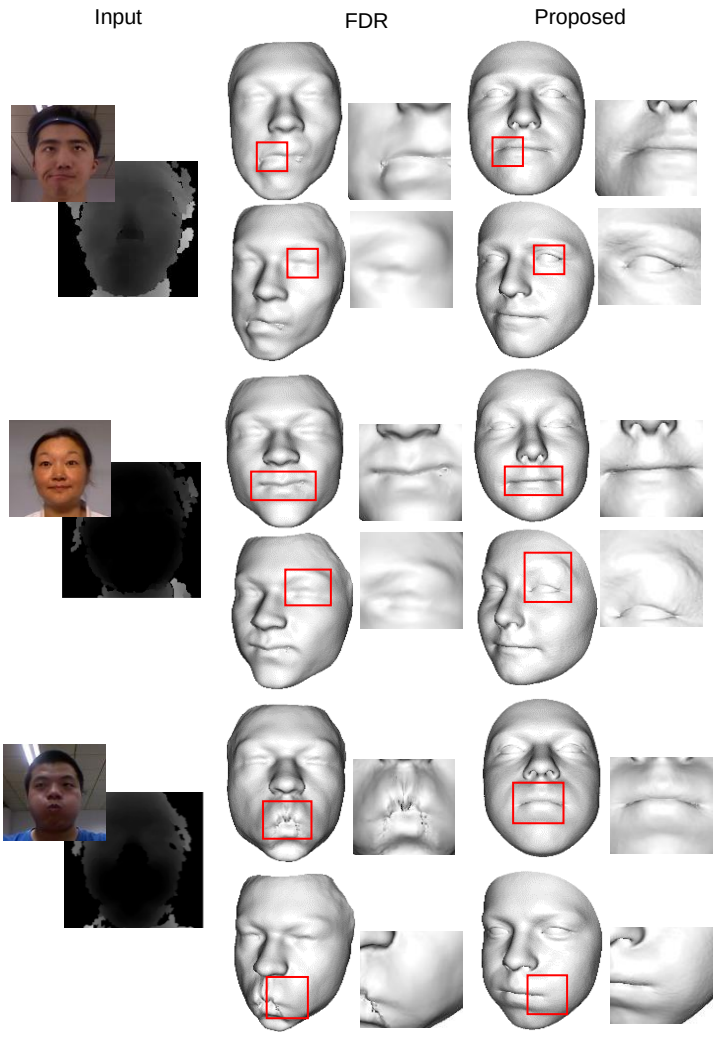

Figure 5.3: Comparison with depth-based method proposed by Zhong et al. (Zhong et al., 2020). The RGB image is shown only for better comparison.

As forementioned, FaceWareHouse (Cao et al., 2013) is mainly used to assess the 3D facial expression reconstruction performance due to its diversity in facial expressions. The proposed approach is firstly compared with FDR (Zhong et al., 2020), a recent and unique depth-based approach with self-supervised learning that estimates unrestricted face shapes. Based on the results provided by FDR (Zhong et al., 2020), three subjects from FaceWarehouse are displayed in Fig.5.3. From the Fig.5.3, the proposed approach produces detailed faces with the better identity and more accurate expressions. For example, the third subject in is cheek-bulging which the result of proposed approach reveals, however, the FDR (Zhong et al., 2020) is just pouting. Moreover, some artifacts from FDR (Zhong et al., 2020) can be observed around mouth region while results of proposed approach are much more pleasing.

The proposed method is then compared, with five deep RGB-based models of VRN (Jackson et al., 2017), PRN (Feng et al., 2018), Deng et al. (Deng et al., 2019), 3DDFA2 (J. Guo et al., 2020), and GCN (Lin et al., 2020). Both GCN (Lin et al., 2020) and Deng et al. (Deng et al., 2019) generate the same face shape since GCN adopts the Deng et al. for shape regression, so just results of Deng et al.(Deng et al., 2019) are showcased in Fig.5.4. The displayed samples were selected randomly. All visual results of compared methods are generated though the released pre-trained models. What needs to be mentioned here is that five key points (two eye centres, nose centre and two mouth corners) are required to be marked manually before the image are fed to Deng et al. (Deng et al., 2019). As exhibited in Fig.5.4, all the prior art RGB-based methods cannot perform well or even fail to reconstruct the 3D face with the unsymmetric facial expressions, while our method is still able to reconstruct the promising 3D face expressions thanks to the depth information.

Comparison on Biwi and ICT-3DHP. To further evaluate the proposed approach under large pose and occlusions, the recovered shapes are compared on two public datasets covering various poses information: Biwi (Fanelli et al., 2011)

Figure 5.4: Visual results of state-of-the-art RGB-based methods and the proposed approach. The RGB image is shown only for better comparison.

and ICT-3DHP (Baltrušaitis et al., 2012), with fore-mentioned RGB-based approaches. However, the demo of VRN (Jackson et al., 2017) cannot detect the face from the profile images. Here, the visual results of PRN (Feng et al., 2018), 3DDFA2 (J. Guo et al., 2020) and Deng et al. (Deng et al., 2019)are showcased in

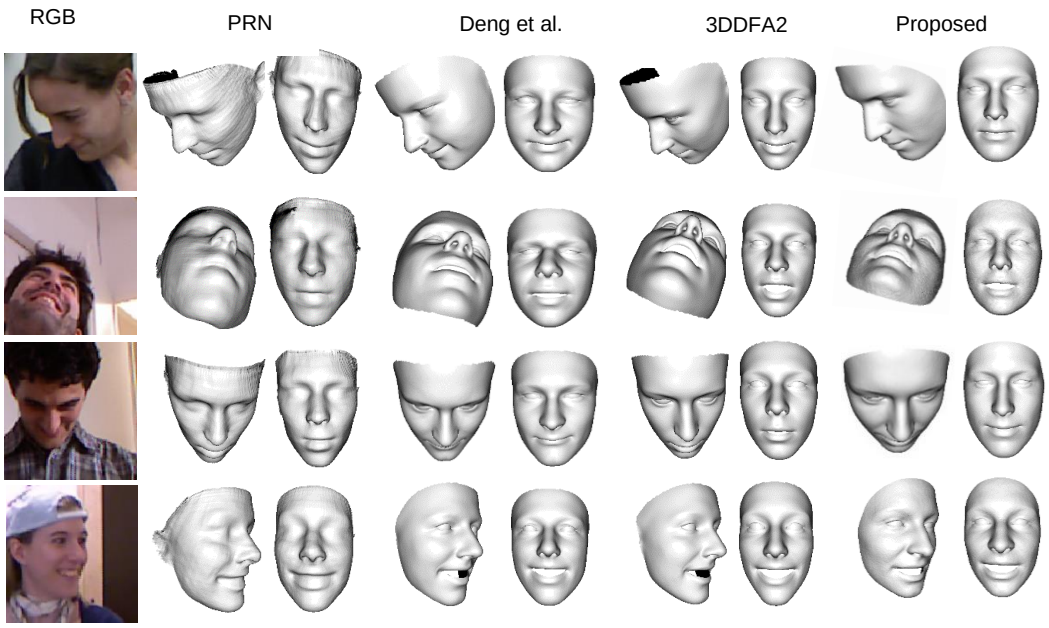

Figure 5.5: Comparison with prior art under large pose and occlusions. Not that the input to Deng et al. (Deng et al., 2019) needs to be marked 5 key points manually. The RGB image is shown only for better comparison.

Table 5.2: Evaluations on Biwi Dataset. 'A' means All sequences and 'P' means Partial sequences. The top three results are marked with Red, Green and Blue.

Methods	Testing Set	Data		Errors		
		Depth	RGB	Pitch (°)	Yaw (°)	Roll (°)
RF	A	×		8.5	8.9	7.9
Martin	A	×		2.5	3.6	2.6
CLM-Z	A	×		12	14.8	23.3
TSP	A	×		3.0	3.9	2.5
PSO	A	×		6.6	11.1	6.7
Li*	A	×	×	1.7	2.2	3.2
Ours	A	×		2.6	2.5	2.6
Poseidon*	P	×	×	1.6	1.7	1.8
FDR	P	×		2.7	2.9	3.0
Ours	P	×		2.1	2.1	2.3

Fig.5.6. By visual inspection, the proposed method is comparable with the state-of-the-art RGB-based methods even no rough landmark information needs to be provided.

5.4.3 Results on Head Pose Estimation

The proposed approach is also compared on both Biwi and ICT-3DHP datasets, with the state-of-the-art depth-based head pose estimation methods (Baltrušaitis et al., 2012; Fanelli et al., 2011; S. Li et al., 2015; Martin et al., 2014; P.Padeleris et al., 2012; Papazov et al., 2015; Zhong et al., 2020). Tab.5.2 Shows the mean average errors about the rotation (Euler) angles on Biwi Dataset. Similarly, the rotation errors about ICT-3DHP dataset are displayed in Tab.5.3. Note that the results of the reference approaches are taken directly from the corresponding papers.

Among all the depth-based head pose estimation algorithms, the proposed approach delivers almost the lowest errors on Biwi dataset. Partially, this can be due to that the position map fed to the network embeds the pose information in a more explicit way than the original depth map that is commonly used by the previous methods. The results demonstrate the superiority of the proposed method

Table 5.3: Evaluations on ICT-3DHP Dataset. 'A' means All sequences and 'P' means Partial sequences. * means deal with both RGB and depth images. The top three results are marked with Red, Green and Blue.

Method	Testing set	Errors		
		pitch	yaw	roll
RF	A	9.4	7.2	7.5
CLM-Z	A	7.1	6.9	10.5
Li*	A	3.1	3.3	2.9
Ours	A	6.5	7.1	5.7
Poseidon*	P	4.9	4.4	5.1
Ours	P	4.6	4.7	5.1

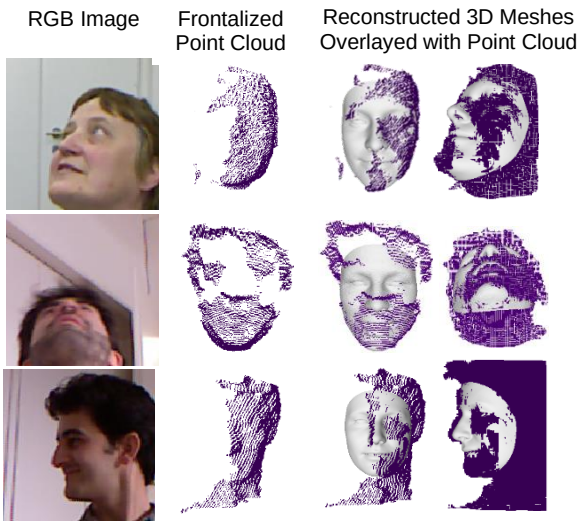

Figure 5.6: The examples of head pose estimation by the reconstructed face model.

which neither requires labelled data for training (Fanelli et al., 2011; Martin et al., 2014) or a compute-intensive optimization step during model adaptation (Baltrušaitis et al., 2012; S. Li et al., 2015; Martin et al., 2014; P.Padeleris et al., 2012). Although no appearance information is utilized, the developed approach is comparable with the prior art Li (S. Li et al., 2015) and Poseidon(Borghi et al., 2017) that employed both depth and RGB data. Similarly, the conclusion can be made on the ICT-3DHP dataset. The proposed approach achieves very pleasing performance for the translation on Biwi dataset as well. Moreover, the model of PoseNet only needs 135s to test all the samples from Biwi, which means it can be used in real-time head pose estimation benefit from the fast speed (more than 100 frames per second).

To further demonstrate the robustness of our PoseNet, the reconstructed 3D meshes overlayed with input point cloud (extracted from position maps) on BIWI and ICT-3DHP datasets are provided for visual inspection in Fig.5.6. The developed approach is robust to the large poses and occlusions, not only can handle rotation information under large pose, but also estimate the good translation data. What's more, it is also effective to the facial expression variations.

5.4.4 Ablation Study

The ablation experiments are conducted to explore the effectiveness of different elements in head pose estimation. Two main elements including training strategy and training data have been taken into consideration for head pose estimation. The effect of each element is evaluated on Biwi dataset in Tab.5.4. The results from the self-supervised model only trained on real data has not been produced since the model was non-convergence. This is due to that the noisy information from depth (position) image confused the model during the training. Obviously, the self-supervised finetuning manner performs best among all the strategies. The self-supervised fine-tune model improves the performance largely than the supervised model trained on synthetic data. The reason can be explained by following: the synthetic data and real data are in the similar feature domain, but the noise of real data is easy to mislead the feature learning. Fortunately, self-supervised fine-tuning based on chamfer distance is effective to extract the useful feature in real depth since it is good at finding the correspondence between different point clouds and matching them. Compare with domain adaptation, the self-supervised fine-tuning not only save the training time but also reduce the computation cost.

Table 5.4: The training details of head pose estimation models on Biwi dataset.

Learning Framework (Manner)	Supervised	×			×	
	Self-supervised		×			×
	Domain Adaptation (ShapeNet)			×		
Data (Position Map)	Synthetic	×		×	×	
	Real		×	×		×
Training Cost	Per Epoch(s)	120	135	650	120	135
	Epochs	150	150	35	100	50
Model Size	(M)	19.5	19.5	106.3	19.5	
Errors	Pitch	8.9	-	**2.5**	2.6	
	Yaw	5.1	-	2.6	**2.5**	
	Roll	5.3	-	3.1	**2.6**	

5.5 Conclusion

This chapter proposes a learning-based approach for 3D face reconstruction from a single depth map collected by the economic depth sensor Kinect. Specifically, the proposed approach mainly includes two parts: ShapeNet designed to predict the displacement of each vertex for reconstructing the 3D face and PoseNet developed to estimate head pose. The ShapeNet is designed inspired by generate-to-adapt domain adaptation framework and trained on source domain (labelled synthetic data) and target domain (un-labelled real data). Similarly, the PoseNet is also trained on these two kinds of data, but the structure is based on deep neural network. Both quantitive and qualitative experiments are conducted on three public datasets to evaluate proposed method which outperforms the compared prior art.

Chapter 6

Conclusion and Future Work

This thesis focused on two fundamental tasks in visual scene analysis which are saliency detection for the general scene and depth-based 3D face reconstruction for the human-centred scene. After extensively reviewing the principal studies in related areas, it identified problems and challenges of each task and resorted to the generative adversarial network - GAN for robust solutions. Its main contributions are summarised as follows:

In Chapter 3, the thesis proposed a perceptual loss-guided GAN – PerGAN for saliency detection from a general-scene RGB image. PerGAN is trained with a perceptual loss that measures misdetection on the semantic feature level rather than the pixel level of the estimated saliency map. It explicitly incorporates high-level semantic information like the object shape into salient object inference. PerGAN is further strengthened with a multi-scale discriminator for extracting useful information from the input image and the generated saliency map in different resolutions. Experimental results on four challenging databases demonstrated that PerGAN is competitive against the state-of-the-art methods. More specifically, PerGAN delivers improved performance on locating the salient object's boundary and preserving its completeness.

From Chapter 4, the thesis turned to address another important visual scene analysis task, namely 3D face reconstruction from a depth image of a human-centred scene. It proposed to use GAN to learn directly from the depth image a facial voxel grid that explicitly depicts the 3D facial shape with a number of discrete voxels. It further integrated the attention mechanism into the GAN for weighting heavier on those intermediate features that show higher relevance in predicting the facial voxel grid. The resulting network is named as Attention-

guided GAN - AGGAN. After being trained on a large-scale dataset of synthesized depth images that cover a wide range of facial expressions and head poses, AGGAN is able to accurately predict the facial voxel grid given a new depth image with big facial poses and noises. This provides an efficient learning-based alternative to existing solutions that normally rely on a costly optimization-based 3D surface recovery process and are sensitive to image noises and facial poses. Experimental results also indicated that the AGGAN model trained on the synthetic data has the potential of generalizing to noisy real depth images.

Encouraged by the outcomes of Chapter 4, the thesis continued to synthesizing depth images to train the 3D face reconstruction network in Chapter 5. For adapting the reconstruction model to real depth images, it proposed to apply both labelled synthetic data and unlabelled real data for model training, while employing the domain adaption technique to learn a common feature embedding that is informative to both real and synthetic domains. The core of the proposed method is two independent domain-adaptive convolutional neural networks for predicting head pose and a normalized 3D facial shape respectively. To achieve domain-adaptive, the former network is trained with a fine-tuning method, while the latter one is trained with a GAN-based generate-to-adapt method. The GAN here is used to fill the domain gap between the synthetic and real data. This is fundamentally different from the previous two methods where GAN directly forms the predictive model. The proposed domain-adaptive 3D face reconstruction network has been validated on three public real datasets, and showed state-of-the-art performance on both 3D facial shape recovery and head pose estimation.

In conclusion, the thesis developed novel and robust algorithmic solutions to both visual saliency detection and depth-to-3D face reconstruction. It also validated GAN's performance in three different scenarios (generating binary saliency map, predicting facial voxel grid and align different domains) that deviate a lot from its familiar area (image generation). This provides valuable experience on adapting GAN to various visual scene analysis tasks, which we believe is beneficial to the further development of GAN.

As future works, there are at least two promising directions to extend this thesis:

1) For saliency detection, it can be found that the current method is object-agnostic. In other words, it does not split the detected salient region into objects when there are two or more objects included. However, humans are capable of detecting saliency at instance level. It is therefore important to achieve instance-level saliency detection. To this end, a future research could be attaching an auxiliary network to the proposed PerGAN for further segmenting the detected salient regions to objects. A correct object segmentation can in turn improve the saliency detection accuracy.

2) In Chapter 5, the current method applies two independent networks (ShapeNet and PoseNet) for predicting head pose and normalized 3D facial shape respectively. This reduces the overall learning difficulty to some extent, but also cuts off some useful communications between these two networks. For example, an accurate 3D facial shape output from the ShapeNet instead of a mean shape can improve the effectiveness of the Chamfer distance-based loss applied for training the PoseNet. From this point, it is interesting and promising to liaise these two networks at the training phase. Specifically, the outputs of one network can be used to construct the loss for updating the other network. Theoretically, the two networks can be updated in an alternating manner until both of them reach to a stable point.

References

Abrevaya, V., Wuhrer, S., & Boyer, E. (2018). Multilinear autoencoder for 3d face model learning. *Proc.WACV*, 1–9.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2274–2282.

Amberg, B., Romdhani, S., & Vetter, T. (2007). Optimal step nonrigid ICP algorithms for surface registration. *Proc.CVPR*, 1–8.

Bagautdinov, T., Wu, C., Saragih, J., Fua, P., & Sheikh, Y. (2018). Modeling facial geometry using compositional vaes. *Proc.CVPR*, 3877–3886.

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2012). 3D constrained local model for rigid and non-rigid facial tracking. *Proc.CVPR*, 2610–2617.

Berger, M., Tagliasacchi, A., Seversky, L., Alliez, P., Levine, J., Sharf, A., & Silva, C. (2014). *State of the art in surface reconstruction from point clouds*.

Bernardini, F., Mittleman, J., Rushmeier, H., Silva, C., & Taubin, G. (1999). The ball-pivoting algorithm for surface reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, *5*(4), 349–359.

Berthelot, D., Schumm, T., & L. Metz. (2017). Began: Boundary equilibrium generative adversarial networks. *ArXiv Preprint ArXiv:1703.10717*.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proc.CGIT*, 187–194.

Bookstein, F. L., & Green, W. D. K. (1993). A thin-plate spline and the decomposition of deformations. *Mathematical Methods in Medical Imaging*, *2*, 14–28.

Borghi, G., Venturelli, M., Vezzani, R., & R.Cucchiara. (2017). Poseidon: Face-from-depth for driver pose estimation. *Proc.CVPR*, 4661–4670.

Borji, A., & Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207.

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *ArXiv Preprint ArXiv:1809.11096*.

C.Walder, B.Schölkopf, & O.Chapelle. (2006). Implicit Surface Modelling with a Globally Regularised Basis of Compact Support. *Computer Graphics Forum*, *25*(3), 635–644.

Cai, X., & Yu, H. (2018). Saliency detection by conditional generative adversarial network. *Proc.ICGIP*.

Cao, C., Bradley, D., Zhou, K., & Beeler, T. (2015). Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics*, *34*(4), 1–9.

Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2013). Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, *20*(3), 413–425.

Cheng, M. M., Mitra, N. J., Huang, X., & Hu, S. M. (2014). Salientshape: group saliency in image collections. *The Visual Computer*, *30*(4), 443–453.

Cheng, M. M., Mitra, N. J., Huang, X., Torr, P. H., & Hu, S. M. (2014). Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(3), 569–582.

Chibane, J., T.Alldieck, & Pons-Moll, G. (2020). implicit functions in feature space for 3d shape reconstruction and completion. *Proc.CVPR*, 6970–6981.

Cohen-Or, D., & Kaufman, A. (1995). Fundamentals of surface voxelization. *Graphical Models and Image Processing*, *57*(6), 453–461.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603–619.

Dai, A., Ruizhongtai Qi, C., & Nießner, M. (2017). Shape completion using 3d-encoder-predictor cnns and shape synthesis. *Proc.CVPR*, 5868–5877.

Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., & Tong, X. (2019). Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. *Proc.CVPR Workshop*.

Denton, E. L., Chintala, S., & Fergus, R. (2015). Deep generative image models using a laplacian pyramid of adversarial networks. *Proc. NIPS*, 1486–1494.

Donne, S., & Geiger, A. (2019). Learning non-volumetric depth fusion using successive reprojections. *Proc.CVPR*, 7634–7643.

Ekman, P., & E. I. Rosenberg. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

Fan, H., Su, H., & L.J. Guibas. (2017). A point set generation network for 3d object reconstruction from a single image. *Proc.CVPR*, 605–613.

Fan, Z., Hu, X., Chen, C., & Peng, S. (2018). Dense semantic and topological correspondence of 3d faces without landmarks. *Proc.ECCV*, 523–539.

Fanelli, G., Gall, J., & Gool, L. V. (2011). Real time head pose estimation with random regression forests. *Proc.CVPR*, 617–624.

Fang, L., Wang, Z., Chen, Z., Jian, F., Li, S., & He, H. (2019). 3D shape reconstruction of lumbar vertebra from two X-ray images and a CT model. *IEEE/CAA Journal of Automatica Sinica*.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, *59*(2), 167–181.

Feng, Y., Wu, F., Shao, X., Wang, Y., & Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. *Proc.ECCV*, 534–551.

Frintrop, S., Klodt, M., & Rome, E. (2007). A real-time visual attention system using integral images. *Proc. ICCV*.

Gatys, L., Ecker, A. S., & Bethge, M. (2015). Texture synthesis using convolutional neural networks. *Proc. NIPS*, 262–270.

Gecer, B., Ploumpis, S., Kotsia, I., & Zafeiriou, S. (2019). Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. *Proc.CVPR*, 1155–1164.

Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., & Vetter, T. (2018). Morphable face models-an open framework. *Proc.FG*, 75–82.

Gilani, S., Mian, A., & Eastwood, P. (2017). Deep, dense and accurate 3D face correspondence for generating population specific deformable models. *Pattern Recognition*, *69*, 238–250.

Gkioxari, G., Malik, J., & Johnson, J. (2019). Mesh r-cnn. *Proc.ICCV*, 9785–9795.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Proc. NIPS*, 2672–2680.

Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., & Y. Bengio. (2013). Maxout networks. *Proc.ICML*, 1319–1327.

Groueix, T., Fisher, M., Kim, V. G., Russell, B. C., & Aubry, A. (2018). 3d-coded: 3d correspondences by deep deformation. *Proc. ECCV*, 230–246.

Guennebaud, G., & Gross, M. (2007). Algebraic point set surfaces. *Proc. ACM SIGGRAPH*, 23-es.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Proc. NIPS*, 5767–5777.

Guo, J., X.Zhu, Yang, Y., Yang, F., Lei, Z., & Li, S. Z. (2020). Towards fast, accurate and stable 3d dense face alignment. *Proc.ECCV*.

Guo, Y., Zhang, J., Cai, L., Cai, J., & Zheng, J. (2018). Self-supervised CNN for Unconstrained 3D Facial Performance Capture from an RGB-D Camera. *ArXiv Preprint ArXiv:1808.05323*.

Hassan, M. U., Niu, D., Zhao, X., Shohag, M. S. A., Ma, Y., & Zhang, M. (2019). Salient object detection based on CNN fusion of two types of saliency models. *Proc.IVCNZ*, 1–6.

Hayashi, H., Abe, K., & Uchida, S. (2019). GlyphGAN: style-consistent font generation based on generative adversarial networks. *Knowledge-Based Systems*, *186*.

He, S., Lau, R., Liu, W., Huang, Z., & Yan, Q. (2015). Supercnn: A superpixelwise convolutional neural network for salient object detection. *International Journal of Computer Vision*, *115*(3), 330–344.

Hoppe, H., DeRose, T., Duchamp, T., McDonald, J., & Stuetzle, W. (1992). Surface reconstruction from unorganized points. *Proc.CGIT*, 71–78.

Hou, Q., Cheng, M. M., Hu, X., Borji, A., Tu, Z., & Torr, P. H. (2017). Deeply supervised salient object detection with short connections. *Proc.CVPR*, 3203–3212.

Hu, D., Dai, L., Luo, Y., Zhang, G., Shao, X., Itti, L., & Lu, J. (2018). Multi-scale adversarial feature learning for saliency detection. *Symmetry*, *10*(10), 457.

Ioffe, S., & C. Szegedy. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *ArXiv Preprint ArXiv:1502.03167*.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Jackson, A. S., Bulat, A., Argyriou, V., & Tzimiropoulos, G. (2017). Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. *Proc. ICCV*, 1031–1039.

Jarrett, K., Kavukcuoglu, K., M.A.Ranzato, & Y.LeCun. (2009). What is the best multi-stage architecture for object recognition? *Proc.ICCV*, 2146–2153.

Jeon, J., & Kim, M. (2018). Discovering latent topics with saliency-weighted LDA for image scene understanding. *IEEE MultiMedia*, *26*(3), 56–68.

Jetley, S., Murray, N., & Vig, E. (2016). End-to-end saliency mapping via probability distribution prediction. *Proc.CVPR*, 5753–5761.

Ji, Y., Zhang, H., & Wu, Q. (2018). Saliency detection via conditional adversarial image-to-image network. *Neurocomputing*, *316*, 357–368.

Jia, Y., & Han, M. (2013). Category-independent object-level saliency detection. *Proc.ICCV*, 1761–1768.

Jian, M., Qi, Q., Dong, J., Yin, Y., & Lam, K. M. (2018). Integrating QDWD with pattern distinctness and local contrast for underwater saliency detection. *Journal of Visual Communication and Image Representation*, *55*, 31–41.

Jiang, B., Zhang, L., Lu, H., Yang, C., & Yang, M. (2013). Saliency detection via absorbing markov chain. *Proc.ICCV*, 1665–1672.

Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., & S., L. (2011). Automatic salient object segmentation based on context and shape prior. *Proc.BMVC*, 9.

Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., & Li, S. (2013). Salient object detection: a discriminative regional feature integration approach. *Proc.CVPR*, 2083–2090.

Jiang, Z., & Davis, L. S. (2013). Submodular salient region detection. *Proc.CVPR*, 2043–2050.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *Proc.ECCV*, 694–711.

Jyoti, V., Gupta, S., & Lahiri, U. (2020). Understanding the role of objects in joint attention task framework for children with autism. *IEEE Transactions on Cognitive and Developmental Systems*.

Karabassi, E., Papaioannou, G., & Theoharis, T. (1999). A fast depth-buffer-based voxelization algorithm. *Journal of Graphics Tools*, *4*(4), 5–10.

Kazhdan, M. (2005). Reconstruction of solid models from oriented point sets. *Proc.ESGP*, 73-es.

Kazhdan, M., Bolitho, M., & Hoppe, H. (2006). Poisson surface reconstruction. *Proc.ESGP*, 7.

Kazhdan, M., & Hoppe, H. (2013). Screened poisson surface reconstruction. *ACM Transactions on Graphics*, *32*(3), 1–13.

Kim, J., & Pavlovic, V. (2016). A shape preserving approach for salient object detection using convolutional neural networks. *Proc.ICPR*, 609–614.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Proc. NIPS*, 1097–1105.

Lee, G., Tai, Y. W., & Kim, J. (2017). ELD-net: an efficient deep learning architecture for accurate saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(7), 1599–1610.

Levinshtein, A., Stere, A., Kutulakos, K. N., Fleet, D. J., Dickinson, S. J., & Siddiqi, K. (2009). Turbopixels: Fast superpixels using geometric flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(12), 2290–2297.

Li, G., Xie, Y., Lin, L., & Yu, Y. (2017). Instance-level salient object segmentation. *Proc.CVPR*, 2386–2395.

Li, G., & Yu, Y. (2016). Deep contrast learning for salient object detection. *Proc.CVPR*, 478–487.

Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. *Proc.CVPR*, 5455–5463.

Li, H., Adams, B., Guibas, L. J., & Pauly, M. (2009). Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics*, *28*(5), 1–10.

Li, J. W., Gao, W., & Wu, Y. H. (2018). Elaborate scene reconstruction with a consumer depth camera. *International Journal of Automation and Computing*, *15*(4), 443–453.

Li, S., Ngan, K. N., Paramesran, R., & L. Sheng. (2015). Real-time head pose tracking with online face template reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(9), 1922–1928.

Li, X., Zhao, L., Wei, L., Yang, M. H., Wu, F., Zhuang, Y., Ling, H., & Wang, J. (2016). Deepsaliency: multi-task deep neural network model for salient object detection. *IEEE Transactions on Image Processing*, *25*(8), 3919–3930.

Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. *Proc.CVPR*, 280–287.

Liang, Y., Liu, H., & Ma, N. (2019). A novel deep network and aggregation model for saliency detection. *The Visual Computer*, 1–13.

Lin, J., Yuan, Y., Shao, T., & Zhou, K. (2020). Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks. *Proc.CVPR*, 5891–5900.

Liu, F., & Gleicher, M. (2006). Region enhanced scale-invariant saliency detection. *Proc.ICME*, 1477–1480.

Liu, F., Tran, L., & Liu, X. (2019). 3d face modeling from diverse raw scan data. *Proc.ICCV*, 9408–9418.

Liu, N., & Han, J. (2016). Dhsnet: deep hierarchical saliency network for salient object detection. *Proc.CVPR*, 678–686.

Liu, N., Han, J., & Yang, M. (2018). Picanet: Learning pixel-wise contextual attention for saliency detection. *Proc.CVPR*, 3089–3098.

Liu, Q., Hong, X., Zou, B., Chen, J., Chen, Z., & Zhao, G. (2017). Hierarchical contour closure-based holistic salient object detection. *IEEE Transactions on Image Processing*, *26*(9), 4537–4552.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H. (2010). Learning to Detect A Salient Object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 353–367.

Liu, Z., Xue, Y., Shen, L., & Zhang, Z. (2010). Nonparametric saliency detection using kernel density estimation. *Proc.ICIP*, 253–256.

Lombardi, S., Saragih, J., Simon, T., & Sheikh, Y. (2018). Deep appearance models for face rendering. *ACM Transactions on Graphics*, *37*(4), 1–13.

Lorensen, W. E., & Cline, H. E. (1987a). Marching cubes: a high resolution 3D surface construction algorithm. *ACM Siggraph Computer Graphics*, *21*(4), 163–169.

Lorensen, W. E., & Cline, H. E. (1987b). Marching cubes: A high resolution 3D surface construction algorithm. *ACM Siggraph Computer Graphics*, *21*(4), 163–169.

Luo, C., Zhang, J., Yu, J., Chen, C. W., & Wang, S. (2019). Real-time head pose estimation and face modeling from a depth image. *IEEE Transactions on Multimedia*, *21*(10), 2473–2481.

Luo, J., Zhang, J., & B. Deng. (2019). Robust RGB-D face recognition using attribute-aware loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(10), 2552–2566.

M.C. Potter. (1976). Short-term Conceptual Memory for Pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5).

M.Zollhöfer, Thies, J., Garrido, P., D.Bradley, T.Beeler, P.Pérez, Stamminger, M., Nießner, M., & C.Theobalt. (2018). State of the art on monocular 3D face reconstruction, tracking, and applications. *In ACM Computer Graphics Forum*, *37*(2), 523–550.

Ma, Y., & Zhang, H. (2003). Contrast-based image attention analysis by using fuzzy growing. *Proc.ACM Multimedia*, 374–381.

Maas, A. L., Hannun, A. Y., & A. Y. Ng. (2013). Rectifier nonlinearities improve neural network acoustic models. *Proc.ICML*.

Martin Arjovsky, S. C., & Bottou, L. (2017). Wasserstein generative adversarial networks. *Proc.ICML*.

Martin, M., Camp, F. V. D., & R. Stiefelhagen. (2014). Real time head model creation and head pose estimation on consumer depth cameras. *Proc.3D Vision*, 641–648.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *ArXiv Preprint ArXiv:1411.1784*.

Mukherjee, P., Sharma, M., Makwana, M., Singh, A., Upadhyay, A., Trivedi, A., Lall, B., & Chaudhury, S. (2019). DSAL-GAN: Denoising based Saliency Prediction with Generative Adversarial Networks. *ArXiv Preprint ArXiv:1904.01215*.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., & Fitzgibbon, A. (2011). KinectFusion: Real-time dense surface mapping and tracking. *Proc. ISMAR*, 127–136.

Newman, T. S., & Yi, H. (2006). A survey of the marching cubes algorithm. *Computers & Graphics*, *30*(5), 854–879.

Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. *Proc.ICML*, 2642–2651.

Ogasawara, K., Miyazaki, T., Sugaya, Y., & Omachi, S. (2017). Object-based video coding by visual saliency and temporal correlation. *IEEE Transactions on Emerging Topics in Computing*.

P.Padeleris, Zabulis, X., & A. A. Argyros. (2012). Head pose estimation on depth data based on particle swarm optimization. *Proc.CVPR Workshop, 42–49*.

Pan, H., Niu, X., Li, R., Shen, S., & Dou, Y. (2020). Supervised adversarial networks for image saliency detection. *Proc.ICGIP*.

Pan, J., Sayrol, E., Nieto, X. G. I., Ferrer, C. C., Torres, J., McGuinness, K., & OConnor, N. E. (2017). Salgan: visual saliency prediction with adversarial networks. *Proc.CVPR Workshop*.

Pan, S. J., & Q. Yang. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Enginering*, *22*(10), 1345–1359.

Papazov, C., Marks, T. K., & M. Jones. (2015). Real-time 3D head pose and facial landmark estimation from depth images using triangular surface patch features. *Proc.CVPR*.

Patel, A., & Smith. W. (2009). 3d morphable face models revisited. *Proc.CVPR*, 1327–1334.

Paysan, O., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. *Proc.ICAVSBS*, 296–301.

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc.CVPR*, 652–660.

Qian, M., Qi, J., Zhang, L., Feng, M., & Lu, H. (2019). Language-aware weak supervision for salient object detection. *Pattern Recognition*, *96*, 106955.

R.Girdhar, Fouhey, D. F., Rodriguez, M., & A. Gupta. (2016). Learning a predictable and generative vector representation for objects. *Proc.ECCV*, 484–499.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *ArXiv Preprint ArXiv:1511.06434*.

Ren, Z., Gao, S., Chia, L., & Tsang, I. W. (2013). Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(5), 769–779.

Rodolà, E., Cosmo, L., Bronstein, M. M., Torsello, A., & Cremers, D. (2017). Partial functional correspondence. *Computer Graphics Forum*, *36*(1), 222–236.

Rosin, P. L. (2009). A simple method for detecting salient regions. *Pattern Recognition*, *42*(11), 2363–2371.

S.Shi, C.Guo, L.Jiang, Wang, Z., J.Shi, Wang, X., & S. Li. (2020). Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. *Proc.CVPR*, 10529–10538.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. *Proc.ICCV Worshop*, 397–403.

Sankaranarayanan, S., Balaji, Y., Castillo, C. D., & R. Chellappa. (2018). Generate to adapt: Aligning domains using generative adversarial networks. *Proc.CVPR*, 8503–8512.

Scharfenberger, C., Wong, A., Fergani, K., Zelek, J. S., & Clausi, D. A. (2013). Statistical textural distinctiveness for salient region detection in natural images. *Proc.CVPR*, 979–986.

Sharma, A., Grau, O., & Fritz, M. (2016). Vconv-dae: Deep volumetric shape learning without object labels. *Proc.ECCV*, 236–250.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Singh, N., Arya, R., & Agrawal, R. K. (2018). Performance enhancement of salient object detection using superpixel based Gaussian mixture model. *Multimedia Tools and Applications*, *77*(7), 8511–8529.

Song, S., Yu, F., Zeng, A., Chang, A. X., Savva, M., & Funkhouser, T. (2017). Semantic scene completion from a single depth image. *Proc.CVPR*, 1746–1754.

Sumner, R. W., & Popović, J. (2004). Deformation transfer for triangle meshes. *ACM Transactions on Graphics*, *23*(3), 399–405.

Sun, D., Wu, H., Ding, Z., Li, S., & Luo, B. (2019). Salient object detection based on deep multi-level cascade network. *Proc.ICBICS*, 86–95.

Tan, Q., Gao, L., Lai, Y. K., & Xia, S. (2018). Variational autoencoders for deforming 3d mesh models. *Proc.CVPR*, 5841–5850.

Tang, Y., & X. Wu. (2016). Saliency detection via combining region-level and pixel-level predictions with CNNs. *Proc. ECCV*, 809–825.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & M. Nießner. (2016). Face2face: Real-time face capture and reenactment of rgb videos. *Proc.CVPR*, 2387–2395.

Tong, N., Lu, H., Ruan, X., & Yang, M. H. (2015). Salient object detection via bootstrap learning. *Proc.CVPR*, 1884–1892.

V.Singh, Kumar, N., & N. Singh. (2020). A hybrid approach using color spatial variance and novel object position prior for salient object detection. *Multimedia Tools and Applications*, 1–23.

Valenti, R., Sebe, N., & Gevers, T. (2009). Image saliency by isocentric curvedness and color. *Proc.ICCV*, 2185–2192.

Varley, J., DeChant, C., Richardson, A., Ruales, J., & Allen, P. (2017). Shape completion enabled robotic grasping. *Proc.IROS*, 2442–2447.

Wang, C., M.Cheng, Sohel, F., Bennamoun, M., & J.Li. (2019). NormalNet: A voxel-based CNN for 3D object classification and retrieval. *Neurocomputing*, *323*, 139–147.

Wang, C., Xu, C., Wang, C., & Tao, D. (2018). Perceptual adversarial networks for image-to-image transformation. *IEEE Transactions on Image Processing*, *27*(8), 4066–4079.

Wang, K., Xie, J., Zhang, G., Liu, L., & J.Yang. (2020). Sequential 3d human pose and shape estimation from point clouds. *Proc.CVPR*, 7275–7284.

Wang, L., Jiang, B., Tu, Z., Hussain, A., & Tang, J. (2019). Robust pixelwise saliency detection via progressive graph rankings. *Neurocomputing*, *329*, 433–446.

Wang, L., Lu, H., Ruan, X., & Yang, M. H. (2015). Deep networks for saliency detection via local estimation and global search. *Proc.CVPR*, 3183–3192.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., & Ruan, X. (2017). Learning to detect salient objects with image-level supervision. *Proc.CVPR*, 136–145.

Wang, L., Wang, L., Lu, H., Zhang, P., & Ruan, X. (2018). Salient object detection with recurrent fully convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(7), 1734–1746.

Wang, L., Wang, L., Lu, H., Zhang, P., & Ruan, X. (2016). Saliency detection with recurrent fully convolutional networks. *Proc. ECCV*, 825–841.

Wang, Q., Yuan, Y., & Yan, P. (2012). Visual saliency by selective contrast. *IEEE Transactions on Circuits and Systems for Video Technology*, *23*(7), 1150–1155.

Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. *Proc.CVPR*, 8798–8807.

Wang, T., Zhang, L., Lu, H., Sun, C., & Qi, J. (2016). Kernelized subspace ranking for saliency detection. *Proc. ECCV*, 450–466.

Wang, W., Ceylan, D., Mech, R., & Neumann, U. (2019). 3dn: 3d deformation network. *Proc.CVPR*, 1038–1046.

Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., & Yang, R. (2019). Salient object detection in the deep learning era: An in-depth survey. *ArXiv Preprint ArXiv:1904.09146.*

Wang, Y., Wang, N., Kim, V. G., Chaudhuri, S., & Sorkine-Hornung, O. (2020). Neural Cages for Detail-Preserving 3D Deformations. *Proc.CVPR*, 75–83.

Wilson, G., & D. J. Cook. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology*, *11*(5), 1–46.

Woo, S., Park, J., Lee, J. Y., & So Kweon, I. (2018). Cbam: Convolutional block attention module. *Proc. ECCV*, 3–19.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., & Xiao, J. (2015). 3d shapenets: A deep representation for volumetric shapes. *Proc.CVPR*, 1912–1920.

Xu, L., Zeng, L., & Wang, Z. (2014). Saliency-based superpixels. *Signal, Image and Video Processing*, *8*(1), 181–190.

Xu, Q., Wang, W., Ceylan, D., Mech, R., & U. Neumann. (2019). Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Proc. NIPS*, 492–502.

Yan, Q., Xu, L., Shi, J., & Jia, J. (2013). Hierarchical saliency detection. *Proc.CVPR*, 1155–1162.

Yang, B., Rosa, S., Markham, A., Trigoni, N., & Wen, H. (2018). Dense 3D object reconstruction from a single depth view. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(12), 2820–2834.

Yang, B., Wen, K., Wang, S., Clark, R., Markham, A., & Trigoni, N. (2017). 3d object reconstruction from a single depth view with adversarial learning. *Proc.ICCV Worshop*, 679–688.

Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M. H. (2013). Saliency detection via graph-based manifold ranking. *Proc.CVPR*, 3166–3173.

Yu, L., Li, X., Fu, C., Cohen-Or, D., & Heng, P. A. (2018). Pu-net: Point cloud upsampling network. *Proc.CVPR*, 2790–2799.

Yuan, Y., Lai, Y., Yang, J., Duan, Q., Fu, H., & Gao, L. (2020). Mesh variational autoencoders with edge contraction pooling. *Proc.CVPR Workshop*, 274–275.

Zeng, Y., Zhuge, Y., Lu, H., & Zhang, L. (2019). Joint learning of saliency detection and weakly supervised semantic segmentation. *Proc. ICCV*, 7223–7233.

Zhang, C., Smith, W., Dessein, A., Pears, N., & Dai, H. (2016). Functional faces: Groupwise dense correspondence using functional maps. *Proc.CVPR*, 5033–5041.

Zhang, C., Yang, F., Qiu, G., & Zhang, Q. (2019). Salient object detection with capsule-based conditional generative adversarial network. *Proc.ICIP*, 81–85.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. *Proc.ICML*, 7354–7363.

Zhang, L., Wang, S., & Wang, X. (2018). Saliency-based dark channel prior model for single image haze removal. *IET Image Processing*, *12*(6), 1049–1055.

Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2017). Amulet: aggregating multi-level convolutional features for salient object detection. *Proc. ICCV*, 202–211.

Zhang, P., Wang, D., Lu, H., Wang, H., & Yin, B. (2017). Learning uncertain convolutional features for accurate saliency detection. *Proc.CVPR*, 212–221.

Zhang, X., Wang, T., Qi, J., Lu, H., & Wang, G. (2018). Progressive attention guided recurrent network for salient object detection. *Proc.CVPR*, 714–722.

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., & Liu, P. (2013). A high-resolution spontaneous 3d dynamic facial expression database. *Proc.FG Workshop*, 1–6.

Zhao, J., Mathieu, M., & Lecun, Y. (2016). Energy-based generative adversarial network. *ArXiv Preprint ArXiv:1609.03126*.

Zhao, R., Ouyang, W., Li, H., & Wang, X. (2015). Saliency detection by multi-context deep learning. *Proc.CVPR*, 1265–1274.

Zhao, T., & Wu, X. (2019). Pyramid feature attention network for saliency detection. *Proc.CVPR*, 3085–3094.

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: a review. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(11), 3212–3232.

Zhong, Y., Pei, Y., Li, P., Guo, Y., Ma, G., Liu, M., Bai, W., & Zha, H. (2020). Face Denoising and 3D Reconstruction from A Single Depth Image. *Proc.FG*, 32–39.

Zhou, E., Fan, H., Cao, Z., Jiang, Y., & Yin, Q. (2013). Extensive facial landmark localization with coarse-to-fine convolutional network cascade. *Proc.ICCV Worshop*, 386–391.

Zhu, S., Li, C., Loy, C. C., & Tang, X. (2015). Face alignment by coarse-to-fine shape searching. *Proc.CVPR*, 4998–5006.

Zhu, W., Wu, H., Chen, Z., Vesdapunt, N., & B. Wang. (2020). ReDA: Reinforced differentiable attribute for 3D face reconstruction. *Proc.CVPR*, 4958–4967.

Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2016). Faces alignment across large poses: A 3d solution. *Proc.CVPR*, 146–155.

Zhu, X., Liu, X., Lei, Z., & Li, S. Z. (2017). Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(1), 78–92.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. *Proc.CVPR*, 2879–2886.

Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., & Bai, X. (2019). Progressive pose attention transfer for person image generation. *Proc.CVPR*, 2347–2356.

Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., & Stamminger, M. (2014). Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics*, *33*(4), 1–12.

Zou, C., Yumer, E., Yang, J., Ceylan, D., & Hoiem, D. (2017). 3d-prnn: generating shape primitives with recurrent neural networks. *Proc.CVPR*, 900–909.

Appendix

FORM UPR16

Research Ethics Review Checklist

Please include this completed form as an appendix to your thesis (see the Research Degrees Operational Handbook for more information

UNIVERSITY OF PORTSMOUTH

Postgraduate Research Student (PGRS) Information			**Student ID:**	836242
PGRS Name:	Xiaoxu Cai			
Department:	CTS	**First Supervisor:**	Prof. Hui Yu	
Start Date: (or progression date for Prof Doc students)	**08/04/2017**			
Study Mode and Route:	Part-time ☐ Full-time ☒	MPhil ☐ PhD ☒	MD ☐ Professional Doctorate ☐	

Title of Thesis:	Monocular Visual Scene Analysis: Saliency Detection and 3D Face Reconstruction using GAN
Thesis Word Count: (excluding ancillary data)	26600

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES ☒ NO ☐
b) Have all contributions to knowledge been acknowledged?	YES ☒ NO ☐
c) Have you complied with all agreements relating to intellectual property, publication and authorship?	YES ☒ NO ☐
d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES ☒ NO ☐
e) Does your research comply with all legal, ethical, and contractual requirements?	YES ☒ NO ☐

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	ETHIC-2019-431

If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:

Signed *(PGRS)*:	蔡晓旭	**Date:** 31/01/2021

UPR16 – April 2018

Certificate of Ethics Review

Project Title: Unrestricted 3D Facial Reconstruction by using Deep Learning

Name: Xiaoxu Cai **User ID:** 836242 **Application Date:** 12-Apr-2019 12:41 **ER Number:** ETHIC-2019-431

You must download your certificate, print a copy and keep it as a record of this review.

It is your responsibility to adhere to the University Ethics Policy and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers and University Health and Safety Policy.

It is also your responsibility to follow University guidance on Data Protection Policy:

- General guidance for all data protection issues
- University Data Protection Policy

You are reminded that as a University of Portsmouth Researcher you are bound by the UKRIO Code of Practice for Research; any breach of this code could lead to action being taken following the University's Procedure for the Investigation of Allegations of Misconduct in Research.

Any changes in the answers to the questions reflecting the design, management or conduct of the research over the course of the project must be notified to the Faculty Ethics Committee. **Any changes that affect the answers given in the questionnaire, not reported to the Faculty Ethics Committee, will invalidate this certificate.**

This ethical review should not be used to infer any comment on the academic merits or methodology of the project. If you have not already done so, you are advised to develop a clear protocol/proposal and ensure that it is independently reviewed by peers or others of appropriate standing. A favourable ethical opinion should not be perceived as permission to proceed with the research; there might be other matters of governance which require further consideration including the agreement of any organisation hosting the research.

(A1) Please briefly describe your project:: **The aim of my project is to reconstruct 3d face from a single image by using deep learning methods. I will use some publicly available face datasets to train my network and have a test.**
(A2) What faculty do you belong to?: **CCI**
(A3) I am sure that my project requires ethical review by my Faculty Ethics Committee because it includes at least one material ethical issue.: **No**
(A5) Has your project already been externally reviewed?: **No**
(B1) Is the study likely to involve human participants?: **No**
(B2) Are you certain that your project will not involve human subjects or participants?: **Yes**
(C6) Is there any risk to the health & safety of the researcher or members of the research team beyond those that have already been risk assessed?: **No**
(D2) Are there risks of damage to physical and/or ecological environmental features?: **No**
(D4) Are there risks of damage to features of historical or cultural heritage (e.g. impacts of study techniques, taking of samples)?: **No**
(E1) Will the study involve the investigator and/or any participants in activities that could be considered contentious, unacceptable, or illegal, or in any other way harmful to the reputation of the University of Portsmouth?: **No**
(E2) Are there any potentially socially or culturally sensitive issues involved? (e.g. sexual, political, legal/criminal or financial): **No**
(F1) Does the project involve animals in any way?: **No**
(F2) Could the research outputs potentially be harmful to third parties?: **No**
(G1) Please confirm that you have read the University Ethics Policy and have considered the implications for your project.: **Confirmed**
(G2) Please confirm that you have read the UK RIO Code of Practice for Research and will conduct your project in accordance with it.: **Confirmed**
(G3) The University is committed to The Concordat to Support Research Integrity.: **Confirmed**
(G4) Submitting false or incorrect information is a breach of the University Ethics Policy and may be considered as misconduct and be subject to disciplinary action. Please confirm you understand this and agree that the information you have entered is correct.: **Confirmed**