UNIVERSITY OF CALIFORNIA

Los Angeles

Image-to-image Translation by Deep Learning Model

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical and Computer Engineering

by

Weinan Song

2023

ABSTRACT OF THE DISSERTATION

Image-to-image Translation by Deep Learning Model

by

Weinan Song
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2023
Professor Lei He, Chair

Image-to-image translation is a fascinating and rapidly evolving field in computer vision and artificial intelligence. The problem involves transformation from an input image into an output image, while preserving certain semantic or structural information. This technology enables machines to convert data from one domain into another, offering a wide range of applications like artistic rendering and image restoration. In recent years, this field has garnered significant attention and research efforts, motivated by its potential to revolutionize various industries like entertainment and healthcare. In this dissertation, we address image-to-image translation challenges through a dual lens: cross-domain translation and cross-dimension translation. To be more precise, we present efficient and scalable approaches capable of accomplishing multi-domain translation within a unified framework. Additionally, we introduce an innovative 3D reconstruction method capable of generating three-dimensional representations from single 2D images. Through comprehensive experimentation on diverse datasets spanning multiple modalities, our findings not only validate the efficiency and effectiveness of our proposed methods but also signify a promising technological solution for facilitating efficient cross-domain and cross-dimension translation tasks.

The dissertation of Weinan Song is approved.

Fabien Scalzo

Xiang Chen

Lin Yang

Yingnian Wu

Lei He, Committee Chair

University of California, Los Angeles

2023

To my parents Xiaohui and Zongyan,

for their support and sacrifice.

To my little seal,

for her love and accompanying.

To my grandma,

I will always miss you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I am deeply appreciative of the unwavering support and invaluable guidance provided by my three advisors during my student age: Dr. Lei He, Dr. Zhiru Zhang, and Dr. Guangyu Sun. Their assistance was instrumental in both initiating and successfully completing my PhD journey. Additionally, I want to extend my heartfelt gratitude to my esteemed committee members: Dr. Yingnian Wu, Dr. Fabien Scalze, Dr. Xiang Chen, and Dr. Lin Yang, whose expertise and counsel proved indispensable throughout my research endeavor. Most importantly, I wish to convey my utmost respect and gratitude to three remarkable individuals who have profoundly shaped my values during my youth: Dr. Yingnian Wu, Dr. Meng Li, and Dr. Xingyu Mao. Their spirit and enthusiasm will continue to inspire and motivate me throughout my life.

Besides my committee members and advisors, I would like to extend my deepest gratitude to my collaborators Dr. Jianwen Xie, Dr. Kun Wang, whose insights and expertise have been invaluable to the success of my research projects. Their contributions have not only enhanced the quality of this work but have also enriched my own understanding and skills. I am also thankful for the opportunity to work with a talented and dedicated group of professionals: Yuan Liang, Jiawei Yang, Yaxuan Zhu, Ziheng Zhou, and Chengwen Liang. Their collective knowledge and experience have greatly contributed to my research work.

In addition, I want to say thanks for my new friends made in Los Angeles, Bay Area, Toronto, and Seattle for their support: Haoxin Zheng, Xuan Hu, Lin Du, Yu Zhao, Peng Wei, Xiusi Chen, Junheng Hao, Tong He, Ruiqi Gao, Yutan Gu, Siyuan Huang, Yuejiao Sun, Chen Wu, Yunxuan Yu, Dezhan Tu, Tiandong Zhao, Jiawei Zhang, Flora Fang, Weikun Han, Yifan Zhao, Hengjie Yang, Fang Lin, Hanzhi Xia, Siyou Pei, Siting Liu, Qi Wu, Fei Feng, Meixi Lin, Fangyao Liu, Junzi Tan, Yuxiao Xie, Guangwei Zhang, Can Tao, Chunji Wang, Yaxin Yang, Wanchun Wei, Yan Zhou, Tianhe Yu, Danmei Xu, Wenyan Zhao, Xitong Zhou, Yifei Xu, Yifeng Lan, Xiaofeng Gao, Minqi Liu, Xiaoyi Liu, Yu Shi, Xiaochen Liu, Ian Zhao,

Lily Zhang, Jiashang Liu, Haoming Chen, Erfan Xia, Rui Xu, Ying Zeng, Yichao Zhao, Min Fang, Kaida Zhang, Jiahui Xu, Xin Wei, Jianan Zheng, Weiwei Wu, Jiajia Wu, Mi Tang, Hao Mao. The memories of time together have been the most cherished moments of my life in North America.

In closing, I wish to convey my deepest appreciation to my parents, Xiaohui Song and Zongyan Ren, for their selfless support, and to my beloved girlfriend, Siyuan Liu, for her constant and heartwarming presence by my side. I would also like to express my gratitude and miss to my grandmother, Yulian Liu, who passed away in this year, never having the chance to witness my graduation. Her love and care during my childhood will forever remain in my heart.

VITA

2013–2017	B.S. in Electrical Engineering, Peking University
2021	Software Engineering Intern, Google Cloud
2022	Machine Learning Engineering Intern, Meta (Facebook)
2023	Software Engineering Intern, Google Cloud
2017–2023	Graduate Student Researcher & Teaching Fellow in Electrical and Computer Engineering, University of California, Los Angeles, US

PUBLICATIONS

Weinan Song, Gaurav Fotedar, Nima Tajbakhsh, Ziheng Zhou, Lei He, Xiaowei Ding. MDT-Net: Multi-domain Transfer by Perceptual Supervision for Unpaired Images in OCT Scan. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI-2023)

Weinan Song, Yuan Liang, Jiawei Yang, Kun Wang, Lei He. Oral-3d: Reconstructing the 3d structure of oral cavity from panoramic x-ray. In Proceedings of the AAAI conference on artificial intelligence (AAAI-2021)

Weinan Song, Yuan Liang, Jiawei Yang, Kun Wang, Lei He. T-Net: Learning Feature Representation With Task-Specific Supervision For Biomedical Image Analysis. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI-2021)

Yuan Liang, **Weinan Song**, Liang Qiu, Kun Wang, Lei He. X2teeth: 3d teeth reconstruction from a single panoramic radiograph. In Medical Image Computing and Computer

Assisted Intervention (MICCAI-2020)

Yuan Liang, **Weinan Song**, J.P. Dym, Kun Wang, Lei He. CompareNet: anatomical segmentation network with deep non-local label fusion. In Medical Image Computing and Computer Assisted Intervention (MICCAI-2019)

Ritchie Zhao, **Weinan Song**, Wentao Zhang, Tianwei Xing, Jeng-Hau Lin, Mani Srivastava, Rajesh Gupta, Zhiru Zhang Accelerating binarized convolutional neural networks with software-programmable FPGAs. In ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA-2017)

Weinan Song, Fanhui Zeng, Jingzhi Hu, Zhijun Wang, Xinyu Mao. An unsupervised-learning-based method for multi-hop wireless broadcast relay selection in urban vehicular networks. In IEEE 85th vehicular technology conference (VTC-2017)

CHAPTER 1

Introduction

1.1 Background

Image-to-image translation stands as a pivotal task in the realm of computer vision and image processing, striving to transform an input image from one domain to another while preserving its intrinsic content and structural characteristics. This versatile problem has garnered substantial attention across various industries, spanning medical imaging, artistic expression, autonomous driving, augmented reality, and more. The advent of deep learning models has ushered in a new era of possibilities for image-to-image translation, enabling the development of sophisticated algorithms that exhibit the capacity to convert images between different modalities, enhance their visual quality, or even generate entirely novel visual content.

The essence of image-to-image translation lies in the profound ability to bridge the semantic gap between different visual domains. This gap encompasses a multitude of scenarios, such as converting satellite images to maps, transforming sketches into realistic images, or changing day-time scenes into night-time renditions. The overarching goal is to facilitate meaningful and coherent transformations that are not only visually appealing but also contextually relevant. Deep learning models, with their ability to learn intricate patterns and representations from vast amounts of data, have proven to be formidable allies in achieving these objectives.

Within this landscape, image-to-image translation has witnessed a remarkable evolution

over the past decade, driven by innovations in deep neural networks, architectures, and training strategies. From the pioneering work of neural style transfer, generative adversarial networks, and energy-based model, this field has witnessed the development of an array of powerful tools and methodologies. These models have demonstrated exceptional prowess in tasks like photo-to-caricature translation, super-resolution, domain adaptation, single-image reconstruction, virtual try-on, photo inpainting, old picture restoration, and even the creation of artistic masterpieces by transferring the style of one image to another.

1.2 Research Objective

This dissertation intend to tackle image-to-image translation challenges from two distinct perspectives: cross-domain translation and cross-dimension translation: 1) Cross-domain translation involves translating an image from one domain to another within the same dimension. Our goal here is to develop an efficient and scalable method capable of handling multi-domain translation within a unified framework. 2) Cross-dimension translation, on the other hand, focuses on generating or reconstructing high-dimensional images from a lower-dimensional space, such as converting X-rays to CT images. In this context, we have innovatively proposed two methods for reconstructing teeth or dental structures in density from a single panoramic X-ray. These methods represent pioneering efforts in this specific application.

1.3 Dissertation Outline

The rest of this dissertation is arranged as follows:

- **Chapter 2** presents a multi-domain transfer model designed to overcome the challenges associated with domain shift in medical imaging, using perceptual supervision. This innovative approach simplifies the process of translating images across multiple

domains. Its effectiveness and efficiency are showcased in a specific application: fluid segmentation in Optical Coherence Tomography (OCT) datasets, in the task of data adaptation and augmentation.

- **Chapter 3** studies a cutting-edge energy-based framework for multi-domain image-to-image translation, characterized by four key components: a descriptor, a translator, a style encoder, and a style generator. The descriptor, an advanced multi-head energy-based model, is adept at representing multi-domain image distributions. This framework operates by taking an image from a specific source domain and using the translator to produce a corresponding image in the target domain, guided by a style code. This style code is either derived from a reference image via the style encoder or created by the style generator from random noise. To enhance the efficiency and scalability of our approach, we introduce a progressive cooperative learning strategy. The framework's effectiveness is underscored by robust empirical results, showcasing high-resolution image generation capabilities for both human and animal faces within our energy-based image translation framework.

- **Chapter 4** delves into a groundbreaking method for 3D teeth reconstruction using just a single 2D panoramic radiograph. This method stands apart from conventional single-object reconstruction techniques due to its unique challenge: the need to construct multiple objects at high resolution. This approach introduces an innovative framework that splits the reconstruction process into stages of teeth localization and individual tooth shape estimation. A key feature of this method is the implementation of a patch-based training strategy, enabling efficient end-to-end optimization. Extensive testing has demonstrated the method's ability to accurately reconstruct the 3D structure of dental cavities and capture intricate details for each tooth. These results indicate its potential as an effective solution for other complex multi-anatomy 3D reconstruction tasks. This work is a cooperation with Dr. Yuan Liang for the extension of [SLY21], where I am the original proponent of the innovative concept of transforming 2D images

into 3D space within the field of dental imaging in both works.

- **Chapter 5** introduces an innovative method for reconstructing 3D medical images from panoramic scans, marking a departure from previous techniques that depend on learning from paired 2D and 3D images or individual prior information. This novel approach utilizes an implicit representation model equipped with multi-head prediction, dynamic sampling, and adaptive rendering. It is capable of accomplishing detailed 3D dental reconstructions using only the projection data obtained during panoramic scans, which includes imaging direction and the projection image itself. This method, focused on reconstructing the 3D structure of the oral cavity, demonstrates state-of-the-art performance. Notably, it achieves this high level of accuracy and detail without the need for additional supervision or prior knowledge, setting it apart from existing generative models based on adversarial learning.

- **Chapter 6** concludes the dissertation, and discusses the current challenges and future work of image-to-image translation by deep learning models.

CHAPTER 2

Perceptual Learning for Multi-domain Translation

2.1 Domain Shift in Medical Image Analysis

Deep learning has proved effective in automating the diagnosis and quantification of various diseases and conditions. These models, however, tend to underperform in the presence of domain shifts, which commonly exist in medical imaging due to variance of scanner devices [AMD20], diversity of imaging protocols [GMK17], or deviation between real and synthesized data [SLY21]. In the absence of paired and labeled data, models based on cycle-consistent loss [ZPI17][RDF21] could achieve the image-to-image translation by simply learning from a cycle transfer process within two domains. However, such models are easy to lose the content consistency on diseased images (as shown in our experiment) during the transfer. Besides, the model can only learn a one-to-one domain transfer at one time, where the model complexity grows geometrically with the number of domains. In comparison, neural style transfer [GEB16] provides a promising solution to keep the content consistency by aligning the statistical distribution of medical images collected from different sources [MJG19][CMY20][MG19], while the model complexity problem remains as the optimization time increases linearly with the number of images in source domains.

To address the limitations above, we introduce *MDT-Net*, a multi-domain transfer model, to decouple the feature representation of anatomy structures and domain deviation of medical images with an encoder-decoder network and multiple domain transfer modules. Inspired by perceptual supervision [JAF16], *MDT-Net* preserves anatomical structures during trans-

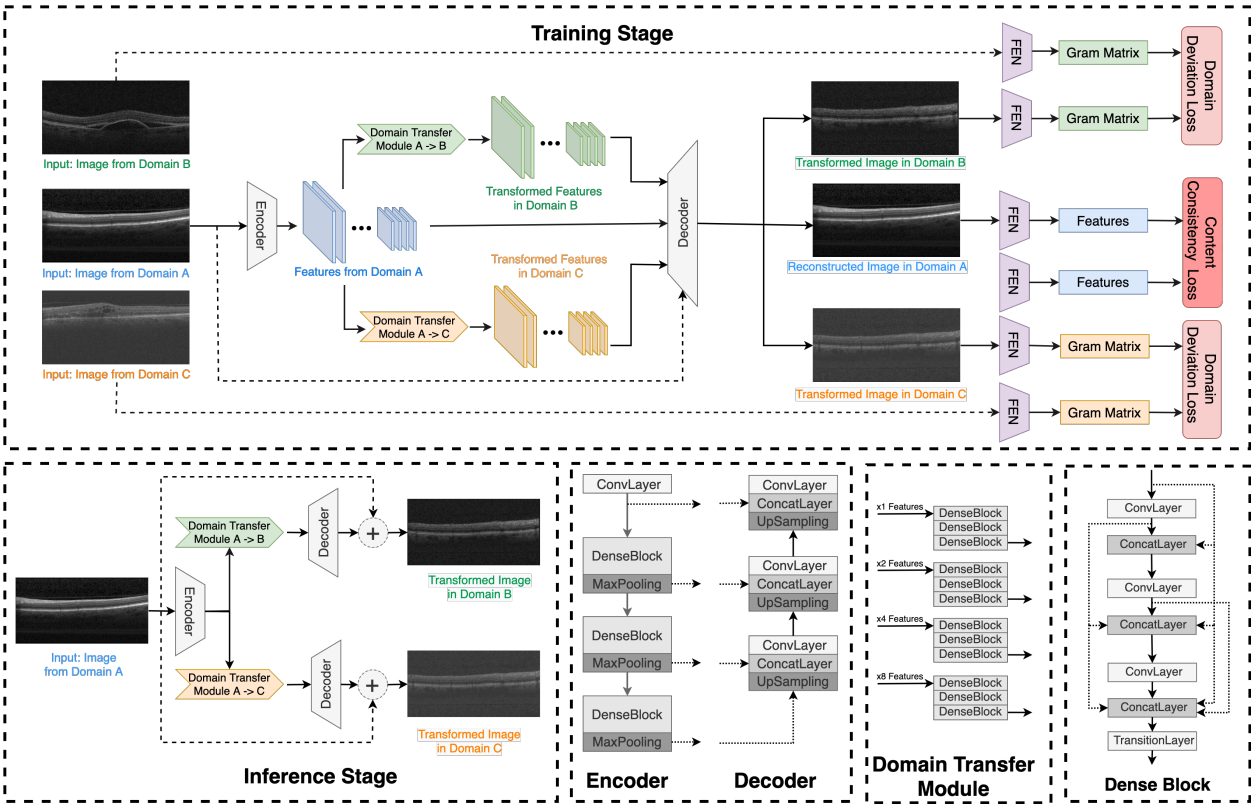


Figure 2.1: A figure illustration of our proposed model is shown in this picture. We use different colors to represent features and images in different domains, i.e., blue for the source domain, and orange and green for the two target domains.

lation by imposing content loss during the identical domain transfer and domain loss (some may call style loss) in the diverse domain transfer. Therefore, it can directly translate images into multiple target domains at one time without any reference images during the inference, where the translation time is independent of the number of source images. Moreover, the model complexity reduces from $\frac{n(n-1)}{2}$ to n when involving transfer among n domains compared with cycle-consistent-based models. To demonstrate the translation performance, we first compare the translated results in domain variance against the target domains and content similarity with the source content. Then we take these translated images as extra data to boost existing segmentation models. Extensive results show that our model can significantly

outperform other methods qualitatively and quantitatively.

2.2 Multi-domain Transfer by Perceptual Learning

2.2.1 Keeping Anatomy Consistency during Translation

As shown in Fig 2.1, *MDT-Net* consists of an encoder-decoder network ($f_e()$ and $f_d()$) to learn anatomy-consistent feature representation and multiple feature transfer modules ($t_i(), i = 1, 2 \cdots, X$) to learn domain transition, where each module learns feature translation to a target domain. The training process during the translation is composed of two circumstances: 1) identical domain transfer, where the model generates an image I' by $f_d(f_e(I))$ from an image I in the source domain, and 2) diverse domain transfer, where the model outputs a translated image I'_X into the target domain X via $f_d(t_X(f_e(I)))$. Since the domain transfer toward each target domain is learned explicitly by a feature transfer module, *MDT-Net* can directly translate images into multiple target domains without any reference images by forwarding deep features into different feature transfer modules respectively during the inference.

2.2.2 Perceptual Supervision

Perceptual supervision is first proposed in [GEB16] and has been widely applied in style transfer between paintings and photograph by capturing implicit content features and texture statistics. Generally, the perceptual loss is calculated by a pre-trained feature extraction network (FEN) $\mathcal{F}$ to compute the reconstruction loss over content and style features. In our model, we define the loss function as a combination of $\mathcal{L}_{content}$ and $\mathcal{L}_{domain}$ following the perceptual loss as:

$$\mathcal{L} = \mathcal{L}_{content}(I, I') + \sum_X \alpha_X \cdot \mathcal{L}^X_{domain}(I_X, I'_X), \tag{2.1}$$

where I_X is the image randomly sampled in a target domain X. The content loss $\mathcal{L}_{content}$ and domain (style) loss $\mathcal{L}_{domain}$ are defined as:

$$\begin{aligned}\mathcal{L}_{content} &= \frac{1}{N_c} \sum_{l}^{l_1^c,\cdots,l_{N_c}^c} \left\|\mathcal{F}^l(I') - \mathcal{F}^l(I)\right\|^2 \\ \mathcal{L}_{domain} &= \frac{1}{N_d} \sum_{l}^{l_1^d,\cdots,l_{N_d}^d} \left\|\mathcal{G}(\mathcal{F}^l(I'_X)) - \mathcal{G}(\mathcal{F}^l(I_X))\right\|^2 .\end{aligned} \tag{2.2}$$

$\mathcal{F}^l(\cdot)$ denotes the features selected from the FEN and $\mathcal{G}(\cdot)$ is a function to compute the Gram matrix [GEB16], which has been widely used to compare the texture statistics in paintings.

2.2.3 Network Architecture

Our network architecture is developed based on StyleBank [CYL17]. We make several improvements to accommodate it to the domain transfer problem in medical images: 1) We apply transfer modules on multi-level features generated by the encoding network to learn feature translation. 2) The transfer modules of *MDT-Net* consist of multiple dense-connected convolution layers instead of a single convolution layer. 3) The model is trained to predict a residual image instead of the transfer result directly. Evaluation of these changes and generation comparisons between StyleBank and MDT-Net can be seen in the ablation study in section 2.4.

2.3 Experiment of Multi-domain Transfer on RETOUCH Dataset

2.3.1 Dataset

We use RETOUCH [BVK19] to validate the domain transfer capability of *MDT-Net*. The dataset contains 70 OCT scans taken by three different vendors (domains): 1) 24 from Cirrus, 2) 24 from Spectralis, and 3) 22 from Topcon. Each image is annotated with three kinds of pathological annotations i.e., intraretinal fluid, subretinal fluid and pigment epithelial

detachment for segmentation. We use C, S, T to represent each domain and randomly select 5 cases from each domain as test data for both domain transfer and segmentation tasks. To be noted, annotations are only used in segmentation models.

2.3.2 Evaluation

We use Fréchet Inception Distance (FID) [HRU17] and Learned Perceptual Image Patch Similarity (LPIPS) [ZIE18] to compare the domain similarity and content inconsistency of generated images. We also propose Domain Perceptual Distance (DPD), a combination of FID and LPIPS, as an overall evaluation metric to indicate the distance to the optimal results by:

$$DPD = FID + \lambda \cdot (1 - LPIPS) \times 100\%, \tag{2.3}$$

where we set $\lambda = 1$ in this work. For data adaptation and augmentation, we use averaged dice scores of the three segmentation targets as the evaluation metric. For comparison, we train four other unsupervised domain transfer models that are either based on cycle-consistent learning, i.e., CycleGAN [ZPI17] and StarGAN2 [CUY20], or perceptual learning, i.e., AdaIN [HB17] and StyleBank.

2.3.3 Training

Due to the various number of slices in three domains, we train the model for 32, 80, and 40 epochs for C, S, and T. The learning rate starts from 10^{-3} and decays by 0.1 in the middle. We use $\alpha_X = 10$ in $\mathcal{L}_{percep}$ and select the same layers as in [GEB16] from VGG-16 [SZ14] to obtain $\mathcal{L}_{content}$ and $\mathcal{L}_{domain}$.

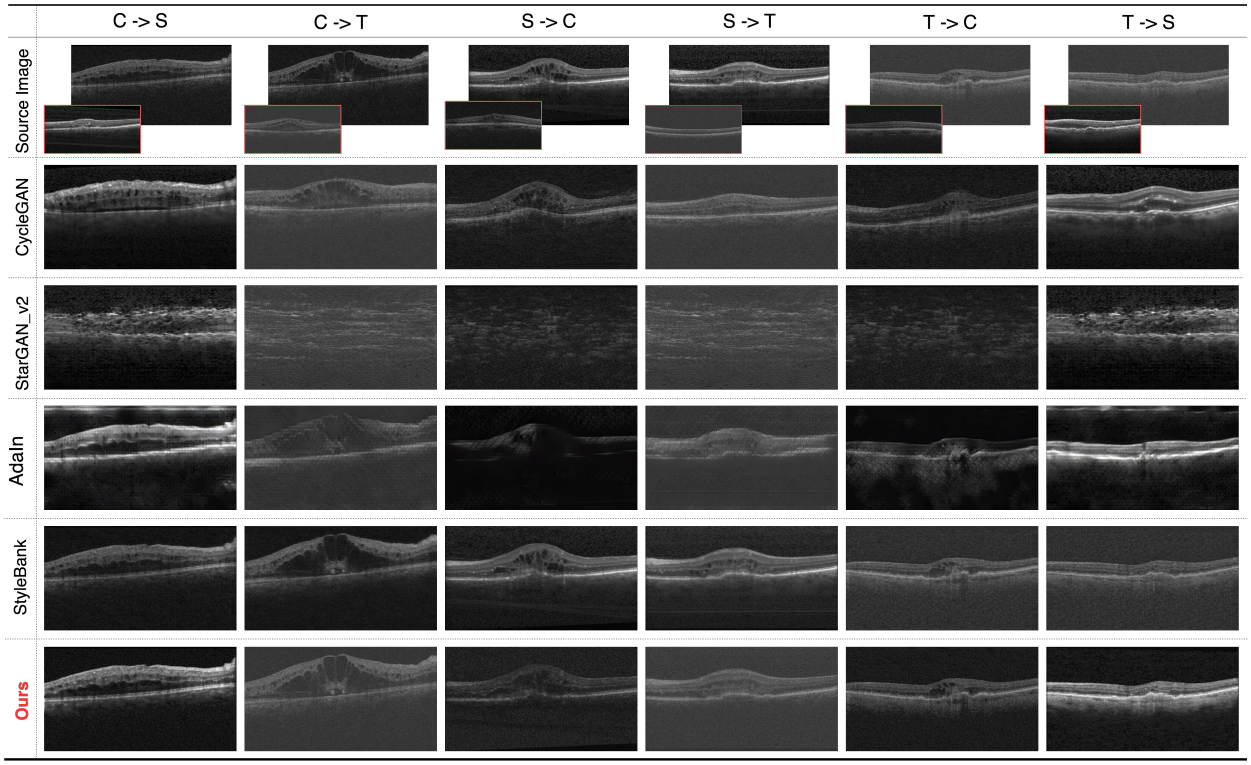


Figure 2.2: Example translated images from different models among domain C, S, and T are shown in this figure. We also place some real images in target domains (not used as reference) in the bottom-left of source images to facilitate comparison.

Method	CycleGAN	StarGAN2	AdaIN	StyleBank	**Ours**
FID↓	47.62	276.45	126.51	129.47	56.23
LPIPS↑	62.72	43.48	53.88	86.85	75.67
DPD↓	84.91	332.96	172.62	142.62	**80.56**

Table 2.1: Quantitative comparison of transfer results averaged in six types of domain transfer by FID, LPIPS and DPD.

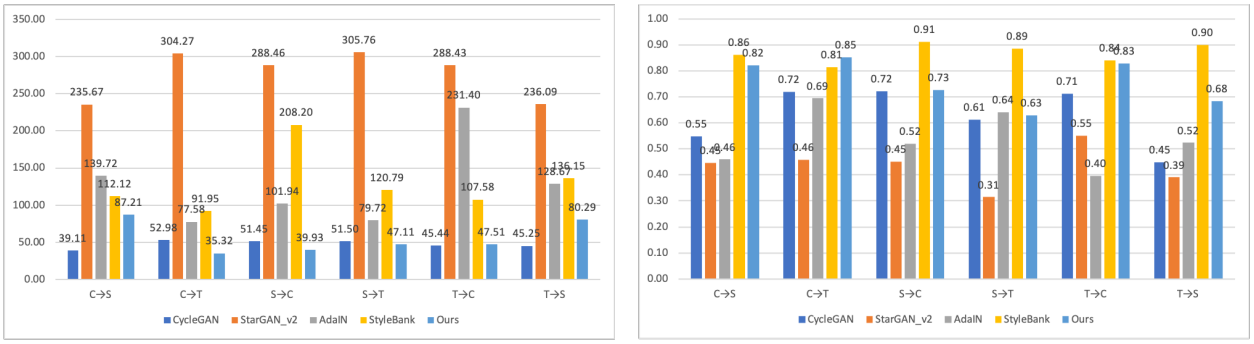


Figure 2.3: Details of the FID and LPIPS in 6 types of domain transfer.

2.4 Results of Domain Translation on OCT scans

2.4.1 Multi-domain Transfer

We first directly compare the translated images under the transfer among the three domains and show the results in Figure 2.4 and Table 2.4.1. Our proposed method can achieve the best balance between domain shift and content consistency. Compared with *MDT-Net*, CycleGAN can get excellent performance on domain shift but fails to keep the anatomical structure of the retina. StyleBank preserves the content during transformation but can not reasonably match the textures of target domains. As StarGAN2 fails to retain the content, we exclude its results in the latter experiments.

Method	C		S		T		Avg
	S	T	C	T	C	S	
-	59.5	70.0	59.1	64.8	54.8	72.1	63.4
CycleGAN	53.1	55.9	52.2	60.8	58.2	71.8	58.7
AdaIN	67.2	77.6	56.4	65.9	77.5	77.3	70.3
StyleBank	76.0	83.4	62.5	84.4	74.2	86.4	77.8
Ours	84.5	76.9	67.5	83.0	77.3	85.5	**79.1**

Table 2.2: Evaluating translated images by DeeplabV3+ in model adaptation.

2.4.2 Data Adaptation

In this experiment, we assume that only images in one domain are provided with annotations, while the segmentation model is expected to accommodate to the test data in the other two domains. This is a very common situation in clinical applications where the inference data could be collected from other sources (domains). We use DeeplabV3+ [CZP18] as the baseline segmentation model. We first train the model with images from one domain, then add translated images that share the same annotations with the source images. Therefore, the improved performance brought by these additional images can indicate the quality of domain transfer results. For example, for domain C, the gap between dice scores of the models trained with 1) C only and 2) C, C $\rightarrow$ S, and C $\rightarrow$ T can indicate the improved adaptation ability in unseen domains, i.e., S and T. As shown in Table 2.4.2, *MDT-Net* brings the biggest improvement and demonstrate the best transfer results.

2.4.3 Data Augmentation

Unlike the adaptation task where images from other domains are unseen during training, we take all images, including both original images from three domains and the transfer results generated by the six kinds of domain transfer within these three domains, as the training data

Segmentation Model	Method	C	S	T	Avg
U-Net	-	76.64	78.36	67.02	74.01
	CycleGAN	68.24	74.71	77.64	73.53
	AdaIn	69.55	73.78	78.18	73.84
	StyleBank	72.05	74.40	79.77	75.41
	Ours	83.97	71.74	73.35	**76.35**
Deeplab	-	83.37	78.91	87.85	83.38
	CycleGAN	79.73	88.90	85.53	84.72
	AdaIn	80.66	87.59	85.86	84.70
	StyleBank	83.44	88.83	80.52	84.13
	Ours	81.61	89.16	87.63	**86.13**
HR-Net	-	80.07	87.26	87.53	84.95
	CycleGAN	82.08	87.56	84.89	84.84
	AdaIn	82.20	80.58	87.97	83.58
	StyleBank	81.92	87.36	81.26	83.51
	Ours	81.46	89.87	87.42	**86.25**

Table 2.3: Evaluating translated images by multiple segmentation models in dice score for data augmentation.

for segmentation in this experiment. Similarly, we use improvements in segmentation accuracy to indicate the domain transfer performance. To avoid influence brought by variation in segmentation models, we introduce U-Net [RFB15] and HR-Net [WSC20] as additional baseline models. As shown in Table 2.4.3, our method can still best boost all three existing segmentation models, bringing about +2% in dice scores.

Method	*MDT-Net_D*	*MDT-Net_S*	*MDT-Net_19*	Baseline
FID↓	60.95	125.46	57.45	56.23
LPIPS↑	76.05	67.63	74.98	75.67
DPD↓	84.90	157.83	82.47	**80.56**

Table 2.4: Evaluation of translated images in the ablation study.

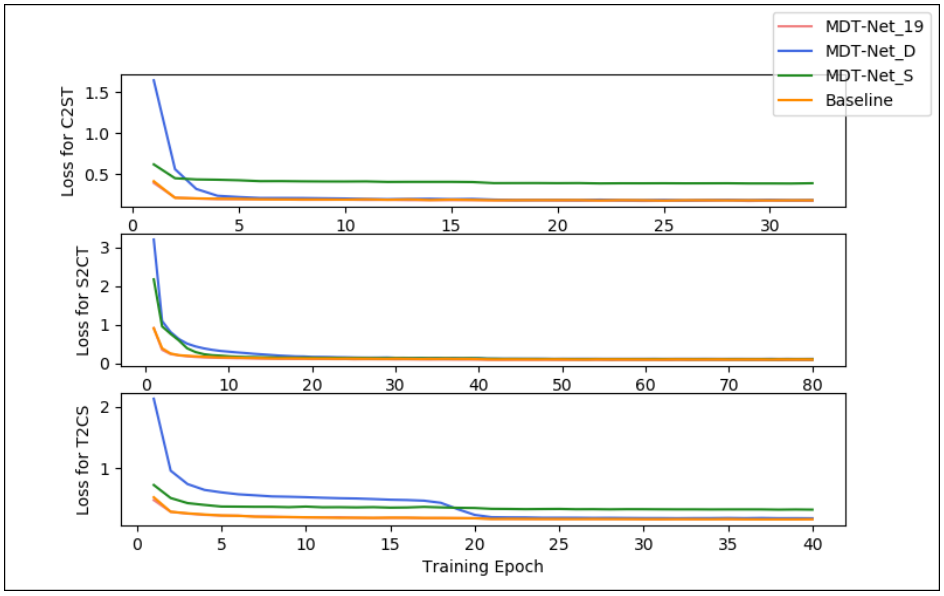


Figure 2.4: Change of the training loss in the ablation study.

2.4.4 Ablation Study

In this experiment, we change the architecture of *MDT-Net* for ablation study. *MDT-Net_D* removes the residual learning at the output, where the decoder directly generates the translated result. *MDT-Net_S* replaces the feature transfer module with the style bank structure as proposed in [CYL17]. *MDT-Net_19* replaces VGG-16 with VGG-19 as FEN. From Table. 2.4.3 we can find that our proposed feature transfer modules play the most important role during the domain transfer. Combined with Figure 2.4, we can see that changing FEN does not affect the result while removing the residual structure mainly increases the converging time.

CHAPTER 3

Progressive Energy-based Model for High-resolution Image Translation

3.1 Motivation

The task of image-to-image translation primarily involves the learning of mappings between different visual domains. This learning framework carries immense application value in the field of generative artificial intelligence, facilitating the development of various creative products for artificial intelligence-generated contents (AIGC). In this context, a "domain" refers to a collection of images belonging to a visually distinctive category such as the gender of a person and animal species. Within each domain, every image exhibits a unique appearance, encompassing image-specific elements such as hairstyle and makeup, commonly referred to as its "style". An ideal image-to-image translation framework should possess the ability to handle multiple domains, efficiently process high-resolution images, and provide diverse synthesis (i.e., one-to-many mapping) when translating to each target domain. By leveraging the representation power of an energy-based model and the sampling efficiency of a latent variable model, the Generative Cooperative Network [XLG18], also known as CoopNets, and its variants [XZL21, XZL22], have achieved impressive results in numerous computer vision tasks, such as image generation [XLG18, XZL21, XZL22], visual salient object detection[ZXZ22], supervised image-to-image translation [XZF22], and unsupervised image-to-image translation [XZF21]. However, while the cross-domain translation framework, CycleCoopNets [XZF21], has demonstrated success in unpaired image-to-image trans-

lation, it is only capable of learning the relation between two different domains at a time. Such an approach has a limited scalability to deal with multiple domains, as a separate model must be trained for each pair of domains. Besides, cooperative learning still faces challenges when it comes to translating high-resolution images. This is because the translation process involves sampling from the energy-based model via Langevin dynamics, which can be difficult to apply to high-resolution image spaces. To tackle the aforementioned challenges in the current cooperative learning (or more generally, energy-based learning) for multi-domain unsupervised image-to-image translation, this chapter proposes a novel cooperative learning framework, PMD-CoopNets, to ensure **scalability**, **flexibility**, **stability** and **efficiency** for applying energy-based framework to image-to-image translation.

To be specific, the PMD-CoopNets consists of four components: descriptor, translator, style generator and style encoder. (1) The descriptor is a multi-head energy-based model that represents a multi-domain image distribution, where each head of the energy function corresponds to one image domain. (2) The style generator is a multi-head latent variable model responsible for generating domain-specific style codes. It achieves this by transforming a Gaussian latent code into style codes. Each head in the style generator corresponds to one specific domain. (3) The style encoder extracts domain-specific style codes from an input image using a multi-head encoder. Each head of the encoder corresponds to a specific domain. (4) The translator is a style-controlled mapping, which takes an image and a style code as input, and then transforms the image into a translated image that reflects the desired style indicated by the style code. The style code can be obtained either from the style generator or the style extractor. The style generator, style encoder, and style-controlled translator can constitute a diversified image generator.

As to the learning, the multi-domain descriptor and the diversified image generator engage in a cooperative game, where the multi-domain descriptor guides the diversified image generator in aligning its mapping towards the target domains using MCMC teaching, while the image generator assists in expediting the descriptor's MCMC teaching process by provid-

ing a good initialization. Specifically, to enforce a meaningful latent space of style codes, we train the style-controlled translator and style encoder by reconstructing style codes that are randomly generated from the style generator. To enforce translated image to preserve the domain-invariant property of the input reference image, we train the translator with a cycle consistency loss. To enforce the one-to-many translation output, we regularize the translator via a diversity sensitive loss, such that, given an identical reference image, different style codes can lead to sufficiently diversified translated outputs.

Additionally, we propose to improve the cooperative learning algorithm by incorporating some loss terms to regularize the behaviors of the components in our framework. Firstly, we put an l_2 regularization on the output of the energy function of the descriptor to limit the magnitude of the energy values. To accelerate and stabilize the teaching process provided by the descriptor's MCMC, we propose to use the energy function to regularize the output of the translator. These regularization techniques significantly improve the performance of the cooperative learning.

To enhance efficiency, stability, and scalability, we present a progressive cooperative learning algorithm for our model. Our approach involves gradual expansion of all four components, initially operating on simpler low-resolution images. As the cooperative training proceeds, new layers are added to each component, enabling the model to handle more challenging high-resolution images. This progressive growth strategy significantly accelerates and stabilizes both training and sampling processes at higher resolutions. Moreover, it offers the flexibility and convenience to scale up the resolution of any pre-trained PMD-CoopNets.

We demonstrate the effectiveness of our proposed multi-domain translation model on the CelebA-HQ [KAL17] and AFHQ [CUY20] dataset. The translated examples exhibit high fidelity and are comparable to GAN-based multi-domain translation models. Furthermore, our progressive learning strategy improves the efficiency and stability of the original training process, particularly when it comes to translating high-resolution images. Our contributions are listed below:

- We propose a novel energy-based cooperative learning framework for multi-domain image-to-image translation. We build a single multi-head energy-based model to represent probability distributions of multiple domains, and train it with a translator, a style encoder, and a style generator using a cooperative manner.

- We present a novel progressive learning algorithm to optimize the training efficiency of our framework. Our approach adopts a progressive growth strategy, advancing all components from low resolution to high resolution. It yields a significant reduction in the total number of MCMC steps required for training and sampling from the high-resolution model.

- We propose regularization strategies to stabilize the cooperative learning, which include an energy-based regularization loss for the translator and a l_2 regularization loss for limiting the magnitude of the energy values of the descriptor. Significant performance gain are obtained from these regularization.

- We demonstrate strong empirical results on CelebA-HQ and AFHQ datasets to verify the proposed energy-based framework. Our method obtains state-of-the-art performance among existing energy-based image translation models.

3.2 Related Work

Energy-based Learning Training energy-based models (EBMs) [ZWM98, LCH06, Hin12] involves maximizing the likelihood of the observed data by adjusting the model's energy function parameters, which typically requires Markov chain Monte Carlo (MCMC) sampling to evaluate the intractable gradient [XLZ16, NHZ19, DM19]. Contrastive divergence (CD) [Hin02, DLT21] is an efficient approximation algorithm for training energy-based models by initializing the MCMC chains with observed data. [NHZ19] uses a noise-initialized non-convergent short-run MCMC to train an EBM, and obtains a valid flow-like generator trained

with moment matching estimation. [GSP21] defines a sequence of conditional EBMs and forms a denoising diffusion process. To avoid MCMC, [GNK20] brings in normalizing flow and trains an EBM by flow contrastive estimation. Learning an amortized sampler [KB16, XLG18, HNF19, KOG19, XKK21, GKH21] for EBMs is also an alternative strategy. Our method has a single multi-head EBM to represent multi-domain data distribution, and the image-to-image translator serves as a multi-domain amortized sampler for the EBMs.

Cooperative Learning Cooperative learning for energy-based models with MCMC teaching is first proposed in [XLG18], where the authors utilize an energy-based model as the descriptor and a latent variable model as the generator to speed up the learning of each other by maximum likelihood algorithms. During each training iteration, the descriptor generates samples by finite-step MCMC sampling with initialization by the generation from the generator for maximum likelihood estimation. Simultaneously, the sampling results from descriptor are used to directly supervise the generator, which is called MCMC teaching. Further research in [ZXL23] shows that this cooperative learning method could also provide a good start point for adversarial models with small computation overhead. Additionally, the model could also be extended for image-to-image translation [XZF21] with two pairs of descriptor and generator or used in saliency prediction [ZXZ22] by introducing a conditional latent variable model.

Progressive Learning The proposed idea of progressive cooperative learning is closely connected to the research conducted by [ZXL21], which involves the incremental growth of a single EBM. The multi-grid EBM framework [GLZ18], trains a series of EBMs simultaneously at various resolutions. The sampling process is conducted sequentially, starting from low-resolution and gradually progressing to higher resolutions, leveraging the lower resolution as a foundation for subsequent higher-resolution sampling. In contrast, our method, which combines the growth of an EBM with three mapping networks, introduces a more challenging and complex progressive learning strategy. It is important to note that while there are

several progressive learning frameworks based on Generative Adversarial Networks (GANs), our approach falls within the domain of energy-based learning. We need to carefully consider MCMC sampling when progressively expanding the energy function, as it plays a crucial role in both bottom-up energy mapping and top-down image generation.

3.3 Proposed Framework

Suppose we have unpaired images from multiple domains A, B, C, $\cdots$ with some shared high-level features, such as expressions in face images, our target is to learn a conditional generative model that maps an image into a target domain, which could be same as the source domain, with specific features. To achieve this, we propose a generative model that consists of four components, i.e., descriptor, style encoder, style generator, and translator. The latter three can form a diversified translator, which is trained with the descriptor in a cooperative learning manner. Let x be an observed image and y be its domain label. We also use y' to denote the label of target domain.

3.3.1 Multi-Domain Descriptor

The multi-domain descriptor is a multi-head energy-based model that specifies the probability distribution of each domain by

$$p_y(x;\theta) \propto \exp[D_y(x;\theta)], \tag{3.1}$$

where θ are parameters of the multi-head energy function D. For notation simplicity, we use $D_y(\cdot)$ to denote the negative energy for domain y. The descriptor are learned by multi-domain maximum likelihood estimation, which is equivalent to minimizing the Kullback-Leibler (KL) divergence between the data distribution $p_{\text{data}}(x,y)$ and the model $p_y(x;\theta)$. The gradient of the objective for learning the descriptor is given by

$$\nabla_\theta \mathcal{L}_{\text{ebm}}(\theta) = -\mathbb{E}_{p_{\text{data}}(x,y)}\{\nabla_\theta D_y(x;\theta) - \mathbb{E}_{p_y(x';\theta)}\left[\nabla_\theta D_y\left(x';\theta\right)\right]\}, \tag{3.2}$$

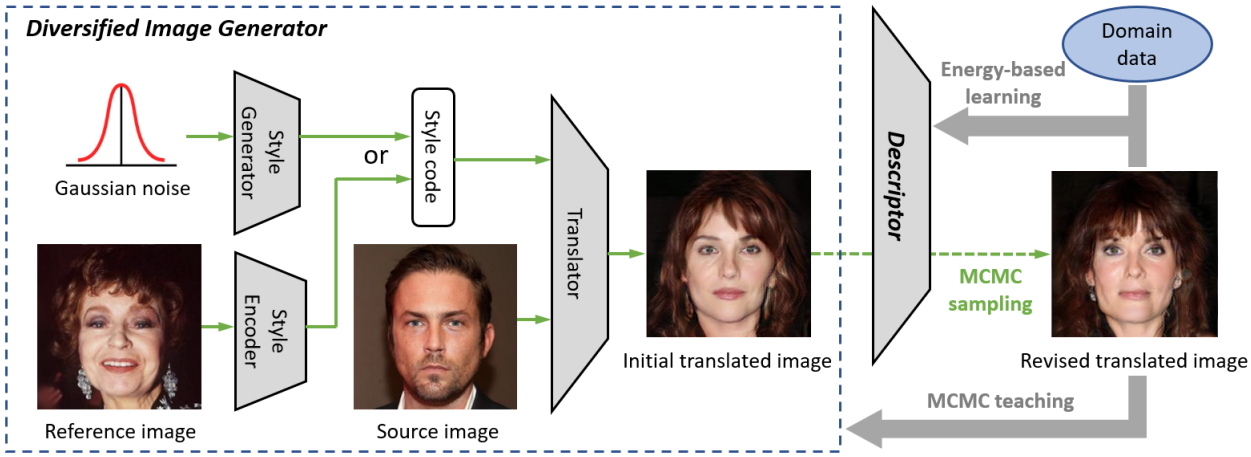


Figure 3.1: Diagram of energy-based cooperative learning for multi-domain image-to-image translation. The framework consists of a style generator, a style encoder, a translator and a descriptor. The first three components (i.e., style generator, style encoder, and translator) form a diversified image generator. Given a input source image, the translator can transform it into a target domain, which is specified by a style code. The style code can be obtained by sampling from the domain-specific style generator or extracted from a reference image by the style encoder. The descriptor is a multi-domain image distribution, which plays the role of guiding the translation such that the translated images can match the observed images in the target domain in terms of statistical property. All components are trained simultaneously in a cooperative learning scheme. The descriptor learns from the multi-domain training images by maximizing the data likelihood, while utilizing MCMC teaching to guide the training of the diversified image generator, which consists of a translator, a style encoder, and a style generator.

where $\mathbb{E}_{p_y(x';\theta)}$ denotes the expectation with respect to the EBM and we use x' in order to distinguish the random variable x in $\mathbb{E}_{p_{\text{data}}(x,y)}$ in the same equation. Suppose we observe a batch of training examples $\{(x_i, y_i)\}_i^n$, which is assumed to be from $p_{\text{data}}(x, y)$. The gradient in Eq.(3.3.1) can be approximated by

$$\nabla_\theta \mathcal{L}_{\text{ebm}}(\theta) \approx \nabla_\theta \left[\frac{1}{n} \sum_{i=1}^{n} D_{y_i}(x_i; \theta) - \frac{1}{n} \sum_{i=1}^{n} D_{y_i}(\tilde{x}_i; \theta) \right], \tag{3.3}$$

where for each observed domain y_i, we use Langevin dynamics to obtain the corresponding synthesized example $\tilde{x}_i$ as a sample from $p_{y_i}(x; \theta)$. With a specified step size δ, Langevin dynamics is performed by iterating the follow step:

$$\tilde{x}_{\tau+1} = \tilde{x}_\tau + \delta \nabla_x D_y(\tilde{x}_\tau; \theta) + \sqrt{2\delta} U_\tau, \;\; U_\tau \sim \mathcal{N}(0, I), \tag{3.4}$$

where τ indexes time step and $\tilde{x}_{\tau=0}$ is initialized by the output of a style-controlled image-to-image translator, which is presented in Section 3.3.2. A good initialization improves the efficiency of Langevin dynamics. To stabilize the EBM training, we also add an l_2 regularization on the energy outputs of both training examples and synthesized examples, which is

$$\mathcal{L}_{\text{energy}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \|D_{y_i}(x_i; \theta)\|^2 + \frac{1}{n} \sum_{i=1}^{n} \|D_{y_i}(\tilde{x}_i; \theta)\|^2. \tag{3.5}$$

3.3.2 Diversified Image Generator

Multi-Domain Style Generator Given a latent variables z and a domain label y, the multi-domain style generator can produce a domain-specific style code c by

$$c_y = G_y(z; \beta) + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, I), z \sim \mathcal{N}(0, I), \tag{3.6}$$

where ϵ is an observation residual and z follows a Gaussian prior distribution. G is a multilayer perceptron (MLP) with multiple output branches to produce style codes for multiple domains. The distribution of style code c conditioned on a domain y is given by $p_y(c; \beta) = \int p_y(c|z; \beta) p(z) dz$, which is more informative than the prior distribution $p(z)$ to

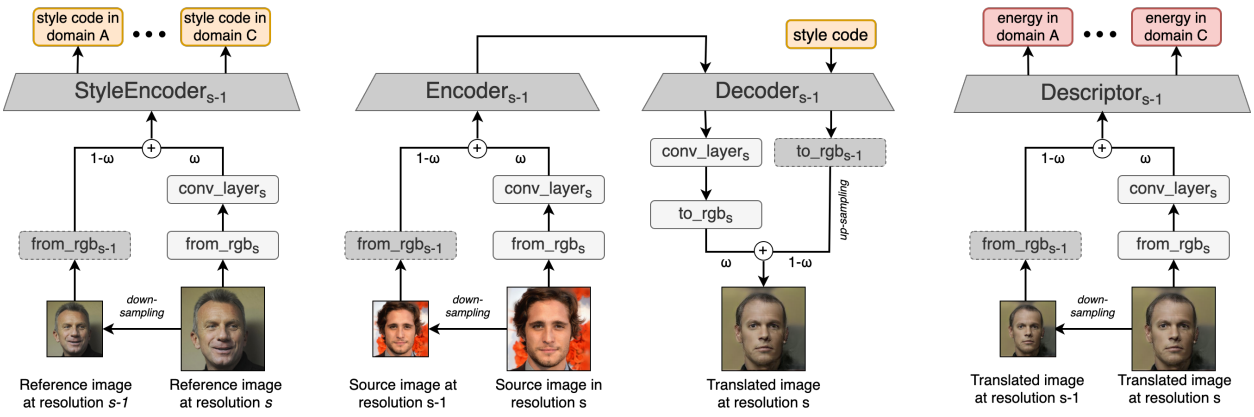


Figure 3.2: An illustration of the progressive strategy for the style encoder E, translator T, and descriptor D. Boxes in dark grey represent well-trained modules at resolution level $s-1$, while blocks in light gray represent the newly added parameters at the current resolution level s. The expansion of the model involves removing some incompatible parameters (depicted as dark grey boxes with dashed boundaries) and adding new parameters (depicted as light grey boxes). The output of the module that needs to be removed and the output of the module that needs to be added are fused using a transition factor ω, This factor starts from 0 and gradually increases to 1, controlling the percentage of contribution from the old and new modules. Left: style encoder. Middle: style-controlled image-to-image translator. Right: descriptor.

capture the underlying style space. The domain-specific style code c_y is directly used in the translator, which is presented in Section 3.3.2, for specifying the style and the target domain of the translated image.

Style Encoder The style encoder E is a multi-head bottom-up network that takes as input an image x and its corresponding domain label y and then outputs a domain-specific style code $c = E_y(x; \phi)$, where ϕ are parameters and $E_y(\cdot)$ denotes the output of E that corresponds to domain y.

Style-Controlled Image-to-Image Translator To achieve a one-to-many translation between domains, we build a style-controlled image-to-image translator. It is a conditioned encoder-decoder T that takes as input a source reference image x and a domain-specific style code c_y and outputs a translated image in target domain y, which is given by

$$x_y = T(x, c_y; \alpha) + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, I), c_y \sim p_y(c; \beta), \tag{3.7}$$

where α is the parameters of the neural network T. The randomness in the translated image, when given a reference image and the target domain, arises from the stochastic nature of the style codes, which follows a distribution defined by the style generator $p_y(c; \beta)$. The translator T and the style generator G forms a diversified translator. They are trained by the MCMC teaching loss [XLG18], which is

$$\mathcal{L}_{\text{teach}}(\alpha, \beta) = \mathbb{E}_{z,y,x}[\|\tilde{x}_{z,y,x} - T(x, G_y(z; \beta); \alpha)\|^2], \tag{3.8}$$

where $\tilde{x}_{z,y,x}$ denotes the Langevin synthesis from the descriptor, which is initialized by the output of $T(x, G_y(z; \beta); \alpha)$. That is, we set $\tilde{x}_{z,x,y,\tau=0} \leftarrow T(x, G_y(z; \beta)$ for Langevin dynamics in Eq.(3.4) to revolve $\tilde{x}_{z,y,x}$. Let $M_\theta q_{\alpha,\beta}(x)$ be the marginal distribution obtained by running Markov transition M_θ from $q(x; \alpha, \beta)$. At learning step $t + 1$, the gradient of the MCMC teaching loss in Eq.(3.8) is the gradient of $\text{KL}(M_{\theta^{(t)}} q_{\alpha^{(t)},\beta^{(t)}} || q_{\alpha,\beta})$, where $q_{\alpha,\beta}$ seeks to be the stationary distribution of M_θ, i.e., minimizing $\text{KL}(p_\theta || q_{\alpha,\beta})$. The effects of the MCMC

teaching loss include: (i) q can chase p toward p_{data} for MLE; (ii) q can serve as a good MCMC initializer for p for efficient MCMC sampling. To ensure diverse translator outputs, we regularize T by minimizing the negative diversity sensitive

$$\mathcal{L}_{\text{diverse}}(\alpha) = -\mathbb{E}_{z_1,z_2,y,x}[\|T(x, G_y(z_1;\beta);\alpha) - T(x, G_y(z_2;\beta);\alpha)\|_1]. \tag{3.9}$$

Since the translator is learned from unpaired data domains, to ensure the translated image $T(x, c; \alpha)$ to preserve the domain-invariant features of the source image x, we adopt the cycle consistency loss:

$$\mathcal{L}_{\text{cycle}}(\alpha) = \mathbb{E}_{z,y,x,y'}[\|x - x_{cycle}\|_1], \tag{3.10}$$

where $x_{cycle} = T(T(x, G_{y'}(z;\beta);\alpha), E_y(x;\phi);\alpha)$. To ensure any style code that is applied to the translated image can be retrieved back from the translated image by the style encoder, we also have a style code reconstruction loss

$$\mathcal{L}_{\text{style}}(\alpha,\phi) = \mathbb{E}_{z,y',x}[\|G_{y'}(z;\beta) - E_{y'}(T(x, G_{y'}(z;\beta);\alpha);\phi)\|_1]. \tag{3.11}$$

To further stabilize the cooperative training and accelerate the MCMC teaching effect, we propose to add the following energy-based regularization on the translator,

$$\mathcal{L}_{\text{mode}}(\alpha,\beta) = \mathbb{E}_{z,y',x}[D_{y'}(T(x, G_{y'}(z;\beta);\alpha);\theta)], \tag{3.12}$$

which can shift the translator mapping toward the low energy modes of the energy function.

3.3.3 Cooperative Learning of Descriptor and Translator

Our full objective function of the descriptor is $\mathcal{L}_{\text{descriptor}} = \mathcal{L}_{\text{ebm}} + \lambda_{\text{energy}}\mathcal{L}_{\text{energy}}$ and the full objective function of the translator is $\mathcal{L}_{\text{translator}} = \mathcal{L}_{\text{teach}} + \lambda_{\text{diverse}}\mathcal{L}_{\text{diverse}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}} + \lambda_{\text{style}}\mathcal{L}_{\text{style}} + \lambda_{\text{mode}}\mathcal{L}_{\text{mode}}$, where λ_{energy}, λ_{diverse}, λ_{cycle}, λ_{style}, and λ_{mode} are hyperparameters. At each learning iteration, the cooperative learning algorithm alternates the following steps: (1) Generate an initial translated image via $\hat{x} = T(x, G_y(z))$; (2) Revise $\hat{x}$ by Langevin dynamics in Eq.3.4 to obtain $\tilde{x}$; (3) Update the parameters θ of descriptor by minimizing $\mathcal{L}_{\text{descriptor}}$; (4) Update the parameters α, ϕ, β of translator by minimizing $\mathcal{L}_{\text{translator}}$.

Algorithm 1 Progressive Cooperative Learning

Input: Multi-resolution data $\{(x_i^{(s)}, y_i^{(s)}),\ i = 1, ..., N; s = 1, ..., S\}$ **Output:** Model $E^{(S)}, T^{(S)}, D^{(S)}, G$

$E^{(0)} \leftarrow \varnothing, D^{(0)} \leftarrow \varnothing, T^{(0)} \leftarrow \varnothing$

for $s = 1, \cdots, S$ **do**

$m \leftarrow 0$

if $s = 1$ **then** $\omega \leftarrow 1$ **else** $\omega \leftarrow 0$

$E^{(s,\omega)} \leftarrow \text{expand}(E^{(s-1)})$

$D^{(s,\omega)} \leftarrow \text{expand}(D^{(s-1)})$

$T^{(s,\omega)} \leftarrow \text{expand}(T^{(s-1)})$

while $(m \leq N)$ **do**

Sample (x, y) and y'

Sample $z \sim \mathcal{N}(0, I)$

$c \leftarrow E_y^{(s,\omega)}(x)$ or $c \leftarrow G_{y'}(z)$

$\hat{x} \leftarrow T^{(s,\omega)}(x, c)$

Revise $\hat{x}$ to obtain $\tilde{x}$ by a K-step Langevin dynamics in Eq. (3.4).

Update descriptor $D^{(s,\omega)}$ with $\mathcal{L}_{\text{descriptor}}$

Update translator $\{E^{(s,\omega)}, T^{(s,\omega)}, G\}$ with $\mathcal{L}_{\text{translator}}$

$m \leftarrow m + n^{(s)}$

if $s \neq 1$ **then** $\omega \leftarrow \min(1, m/N)$ **else** 1

end while

end for

3.3.4 Progressive Cooperative Learning

The update of both descriptor and translator relies on the cooperative generation of MCMC synthesized examples, denoted as $\tilde{x}$. To significantly improve training efficiency, we propose

a progressive learning strategy for our cooperative learning framework. The algorithm gradually enhances the model resolution from low to high, while maintaining cooperative learning across all components at each resolution. The underlying motivation behind this strategy is that learning and sampling from a low resolution data domain is much more efficient. By leveraging a pre-trained low resolution model as a foundation, we can efficiently learn the next scale of the model, rather than starting from scratch. When expanding the current model to the next scale, each component's network structure undergoes modifications. New layers are added to handle higher resolution image inputs or outputs, while incompatible old layers are removed. The newly added layers are trained together with the remaining parameters. To ensure a smooth transition and prevent gradient exposure due to the addition of expanding blocks in each component, we propose to retain partial effects of the parameters that need to be removed while incorporating the effects of the newly added parameters. Throughout each resolution of learning, the impact of the removed parameters gradually diminishes until it becomes zero. Figure 3.2 illustrates the expanding strategy of each component at every level of resolution. Here, ω represents a transition factor that starts from 0 and increase to 1, controlling the percentage of effects from the parameters to be removed (depicted as dark grey boxes with dashed boundaries) and the parameters to be added (depicted as light grey boxes). For a complete description of the proposed progressive cooperative learning algorithm, please refer to Algorithm 1.

3.4 Experiment

3.4.1 Experiment Settings

Dataset and Evaluation Metrics To demonstrate the performance of our proposed multi-domain image-to-image translation framework, we test it on the CelebA-HQ [KAL17] and AFHQ [CUY20] datasets and compare it with several baselines. We use M and F to refer the domains of male and female in CelebA-HQ, and C, D and W to refer to the domains of

cat, dog, and wild animals in AFHQ. We only use the images and the corresponding domain labels from the datasets in our experiments. We evaluate the quality of translated images using the Fréchet Inception Distance (FID) [HRU17] and the Kernel Inception Distance (KID) [BSA18], which are widely used to measure the distance between the population of translated images and the population of original images in the target domain. A small FID or KID is desired to indicate that the translated distribution is very close to the target distribution.

Training and Network Architecture We use bottom-up convolution neural networks for the Descriptor and Style Encoder and a 8-layer MLP for the Style Generator. The Translator utilizes an encoder-decoder architecture with AdaIN [HB17] for style control in the decoding network. We start training our model with a resolution of 64×64, and then scale it up to 128×128 and 256×256. We step for 16 iterations for MCMC sampling in the beginning, decreasing by 4 steps after each progression. The hyper-parameters of λ_{energy}, λ_{diverse}, λ_{cycle}, λ_{style}, and λ_{mode} are set to be 1.

3.4.2 Diverse Image Generation

In this experiment, we use style codes that are randomly sampled from the style generator to generate diverse translated images. Examples of generation results for human face on the CelebA-HQ dataset and animal face on the AFHQ dataset can be seen in Figure 3.3. For each source image shown in the first row, we generate multiple outputs using random Gaussian noise. The qualitative results verify the diversity of the translated results from a source input image. We observe that, given a source image, our model can not only generate diverse translated images but also produce high-quality images that obtain the same attribute (e.g., expression) from the source image.

Figure 3.3: Qualitative results of diverse image generation for human face on CelebA-HQ dataset (left) and animal face on AFHQ dataset (right) are shown in this figure. Each column displays one example of one-to-many image generation. The first row displays source images. The rest four rows show different translated images, which are obtained by using four style codes randomly generated by the style generator. The style generator produces style codes by randomly sampling from Gaussian distribution.

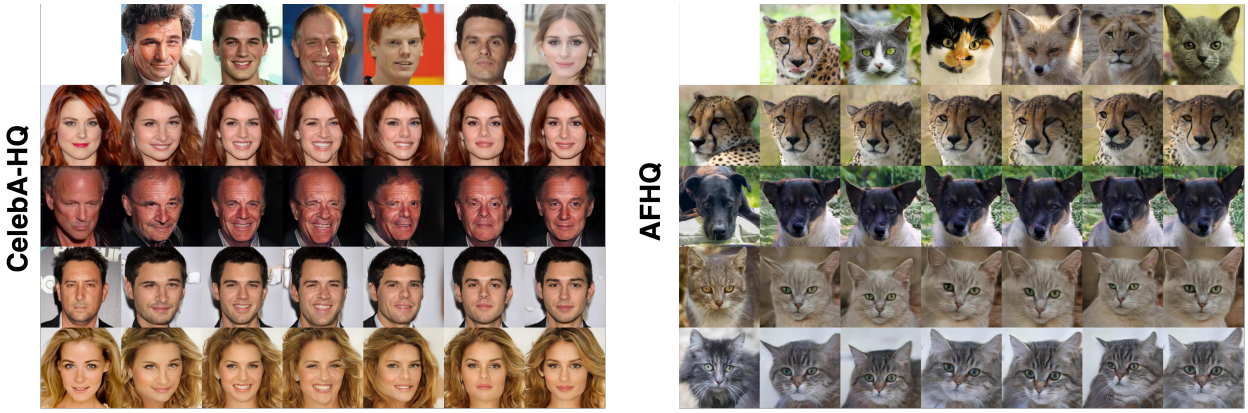


Figure 3.4: We show the translated images with style codes generated from the Style Encoder and reference images for human (left) and animal (right) face in this picture. The source images and reference images are put in the first row and first column. We could see that the face has successfully translated into target domains with consistency in expression.

Table 3.1: Evaluation on CelebA-HQ dataset for two-domain human face generation and AFHQ dataset dataset for three-domain animal face generation.

Method	Reference				Diverse			
	CeleA-HQ		AFHQ		CeleA-HQ		AFHQ	
	FID	KID	FID	KID	FID	KID	FID	KID
MUNIT[HLB18]	107.1	-	223.9	-	31.4	-	41.5	-
DRIT[LTH18]	53.3	-	114.8	-	52.1	-	95.6	-
MSGAN[MLT19]	39.6	-	69.8	-	33.1	-	61.4	-
StarGAN2[CUY20]	23.8	12.1	19.8	6.1	13.7	4.1	16.2	9.1
Liu[LSC21]	26.7	16.8	51.7	28.6	17.8	11.0	26.0	7.0
TUNIT[BCU21]	173.7	187.7	223.0	187.7	128.0	122.0	116.1	99.7
SwapAE[PZW20]	25.4	17.8	61.2	28.8	-	-	-	-
CLUIT[LSL21]	28.9	18.1	22.6	10.5	-	-	-	-
SMGAN[KCK21]	28.8	25.1	64.3	51.3	24.3	15.2	32.8	18.7
CycleCoop[XZF21]	-	-	-	-	131.0	124.7	-	-
EM-LAST[HMH22]	-	-	-	-	48.8	22.9	41.5	17.0
Ours	**21.0**	**7.7**	**19.0**	**6.1**	32.9	21.9	31.8	16.9

Table 3.2: Evaluation on specific domain translations by FID score.

Method	C→D	W→D	M→F
ILVR[CKJ21]	74.4	75.3	46.1
SDEdit[MHS21]	74.2	68.5	49.4
CUT[PEZ20]	76.2	92.9	31.9
C2F-EBM[ZXL21]	55.1	-	-
EM-LAST[HMH22]	69.4	72.5	47.8
EGSDE[ZBL22]	51.0	50.4	30.6
Ours (Diverse)	53.4	54.3	26.8
Ours(Reference)	**36.2**	**36.1**	**16.1**

3.4.3 Translation with Reference Image

We perform image-to-image translation by providing a reference image. We first adopt the style encoder to extract the style code from the provided reference image, and the apply the style code to the translator. Figure 3.4 show some qualitative results, where we take images in the first row as source images and images in the first column as reference images. The translation results are shown in the middle. Comparing results displayed in each row, we can observe that the human face in the source image can be clearly changed into the same gender and appearance of the face in the reference image, while keeping the facial expression consistent with that in the source domain.

3.4.4 Quantitative Comparison

We also compare the results of our translation results quantitatively by using style codes from Style Encoder by randomly selecting reference image in different domains or from Style Generator through sampling from Gaussian distribution with other baseline methods based

Table 3.3: Ablation Study on CelebA-HQ and AFHQ datasets in 64×64 resolution.

Removed Item	CelebA		AFHQ		Avg
	Reference	Diverse	Reference	Diverse	
baseline	15.1	14.3	12.4	19.6	15.4
Remove $\mathcal{L}_{diverse}$	16.3	17.1	36.4	35.2	26.3
Remove $\mathcal{L}_{cycle}$	111.0	127.3	NA	NA	119.2
Remove $\mathcal{L}_{energy}$	134.5	40.7	208.5	97.6	120.3
Remove $\mathcal{L}_{mode}$	NA	NA	277.2	217.6	247.4

on adversarial learning, score matching, or EBMs quantitatively. For each source image in the validation dataset, we obtain ten translated images for each target domain to compute the FID. Results for both human and animal face translation are shown in Table3.1. We also compare our results with some pair-wise translation models on specific domain transfer and summarize the results in Table 3.2. We could see that our model could significantly out perform existing cooperative learning methods with additional ability of guidance by reference images and reach comparable performance with GAN-based methods.

3.4.5 Ablation Study

We conduct an ablation study to evaluate the importance of each individual component proposed in our paper. In Table 3.3, we report the model performance in terms of FID by removing different key loss term (including $\mathcal{L}_{\text{diverse}}, \mathcal{L}_{\text{cycle}}, \mathcal{L}_{\text{energy}}, \mathcal{L}_{\text{mode}}$) from our objective function in our framework. We train our model in a 64×64 resolution setting on datasets CelebA-HQ and AFHQ without using the progressive learning strategy. We show results of image translation using style codes obtained from both style encoder and style generator and report average performance. NA means that the model fails in learning and can not generate meaningful results. We can see that the newly added regularization strategies for

the descriptor and the translator, i.e., $\mathcal{L}_{energy}$ and $\mathcal{L}_{mode}$, are essential for stabilizing the cooperative training. Especially, the energy-based regularization loss $\mathcal{L}_{mode}$ plays an important role to ensure that the translator can quickly catch up with the descriptor toward the data distribution during the cooperative training. The $\mathcal{L}_{energy}$ is useful to obtain performance gain by limiting the magnitude of the energy values. Also, we can find that the performance drops significantly when removing the cycle-consistency loss $\mathcal{L}_{cycle}$, which proves to be a key objective for unpaired cross-domain image translation task.

CHAPTER 4

3D Teeth Reconstruction from a Single Panoramic Radiograph

4.1 Motivation

X-ray imaging is a vital tool in dental diagnosis and surgical procedures due to its cost-effectiveness and lower radiation dose compared to Cone Beam Computed Tomography (CBCT). However, unlike CBCT, X-ray images cannot provide three-dimensional (3D) details about tooth volumes or spatial localization, limiting their use in several dental applications [BOP04, RKB05] such as micro-screw planning, root alignment assessment, and treatment simulations. Additionally, the interpretation of X-ray images, particularly involving volumetric radiation transport, typically requires experienced experts, as discussed in the work of [HRR18]. Therefore, enhancing X-ray images with 3D visualization could be immensely beneficial not only for clinical applications but also for patient education and physician training.

There have been several researches on the 3D reconstruction of a single tooth from its 2D scanning. For example, [MCR13] models the volume of a tooth from X-rays by deforming the corresponding tooth atlas according to landmark aligning. [AEF12, AFS14] reconstruct a tooth from its crown photo by utilizing the surface reflectance model with shape priors. Despite those work, no one has explored the 3D teeth reconstruction of a whole cavity from a single panoramic radiograph. This task is more challenging than the single tooth reconstruction, since not only tooth shapes but also spatial localization of teeth should be

estimated from their 2D representation. Moreover, all the existing methods of tooth reconstruction [AEF12, AFS14, MCR13] utilize ad-hoc image processing steps and handcrafted shape features. Currently, Convolutional Neural Networks (ConvNet) provide an accurate solution for single-view 3D reconstruction by discriminative learning, and have become the state-of-the-art for many photo-based benchmarks [CXG16, HRR18, TDB17]. However, the application of ConvNet on the teeth reconstruction has not yet been explored.

In this work, we pioneer the study of 3D teeth reconstruction of the whole cavity from a single panoramic radiograph with ConvNet. Different from most 3D reconstruction benchmarks [CFG15, SWZ18], which target at estimating a single volume per low-resolution photo, our task has the unique challenge to estimate the shapes and localization of multiple objects at high resolutions. As such, we propose *X2Teeth*, an end-to-end trainable ConvNet that is compact for multi-object 3D reconstruction. Specifically, *X2Teeth* decomposes the reconstruction of teeth for a whole cavity into two sub-tasks of teeth localization and patch-wise tooth reconstruction. Moreover, we employ the random sampling of tooth patches during training guided by teeth localization to reduce the computational cost, which enables the end-to-end optimization of the whole network. According to experiments, our method can successfully reconstruct the 3D structure of the cavity, as well as restore the teeth with details at high resolutions. Moreover, we show *X2Teeth* achieves the reconstruction Intersection over Union (IoU) of 0.6817, outperforming the state-of-the-art encoder-decoder method by 1.71× and retrieval-based method by 1.52×, which demonstrates the effectiveness of our method.

4.2 Methodologies

Figure 4.1 shows the overall architecture of our *X2Teeth*. We define the input of *X2Teeth* as a 2D panoramic radiograph (Figure 4.1(1)), and the output as a 3D occupancy grid (Figure 4.1(5)) of multiple categories for indicating different teeth. Different from the existing single-

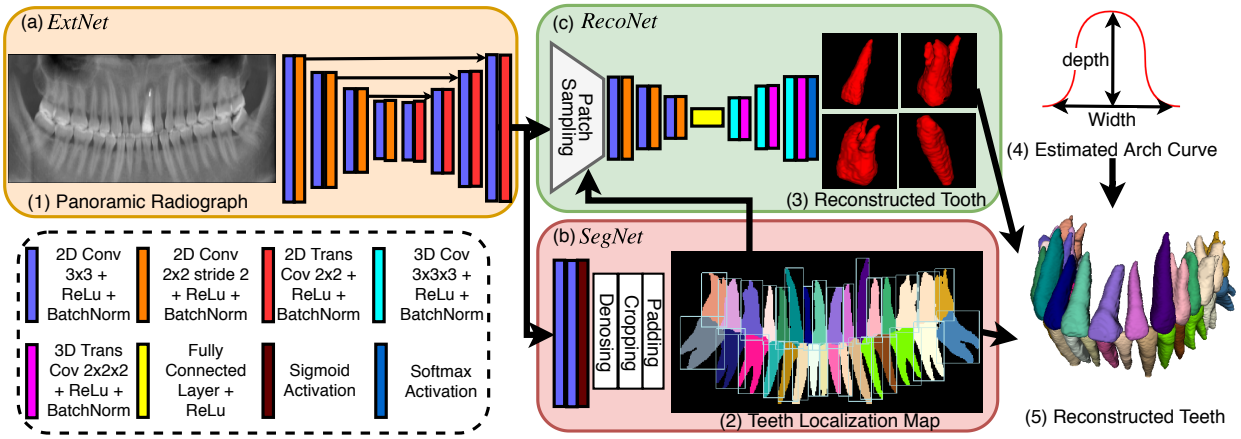


Figure 4.1: Overall architecture of *X2Teeth*. *X2Teeth* consists of three subnets: (a) *ExtNet*, (b) *SegNet* and (c) *ReconNet*. *ExtNet* captures deep representations of teeth from the input panoramic radiograph. Based on the representations, *SegNet* performs pixel-wise classification followed by segmentation map denoising for localizing teeth. *ReconNet* samples tooth patches from the derived feature map and performs single-shape reconstruction. The final reconstruction of the whole cavity is the assembling of the reconstructed teeth according to the teeth localization and arch curve that estimated via β function model. The whole model can be end-to-end trained.

shape estimations [CXG16, TDB17] that mostly employ a single encoder-decoder structure for mapping the input image to one reconstructed object, *X2Teeth* decomposes the task into object localization (Figure 4.1(b)) and patch-wise single tooth reconstruction (Figure 4.1(c)). As such, the reconstruction can be carried out at high resolutions for giving more 3D details under the computational constraint, since tensor dimensions within the network can be largely reduced compared to directly reconstructing the whole cavity volume. Moreover, both sub-tasks share a feature extraction subnet (Figure 4.1(a)), and the whole model can be end-to-end optimized by employing a sampling-based training strategy for the optimal performance. With the derived teeth localization and tooth volumes, the final reconstruction of the cavity is derived by assembling different objects along the dental arch that is estimated via a β function model.

4.2.1 Model Architecture

Given the panoramic radiograph, our *X2Teeth* consists of three components: (1) a feature extracting subnet *ExtNet* for capturing teeth representations, (2) a segmentation subnet *SegNet* for estimating teeth localization, and (3) a patch-wise reconstruction subnet *ReconNet* for estimating the volume of a single tooth from the corresponding feature map patch. The detailed model configuration can be seen from the Figure 4.1.

4.2.1.1 *ExtNet*

As shown in Figure 4.1(a), *ExtNet* has an encoder-decoder structure consisting of 2D convolutions for capturing contexture features from the input panoramic radiograph (Figure 4.1(1)). The extracted features are at high resolutions as the input image, and are trained to be discriminative for both *SegNet* and *ReconNet* to increase the compactness of the network. *ExtNet* utilizes strided convolutions for down-sampling and transpose convolutions for up-sampling, as well as channel concatenations between different layers for feature fusion.

4.2.1.2 *SegNet*

Given the feature map of *ExtNet*, *SegNet* maps it into a categorical mask $Y_{seg} \in \mathbb{Z}^{H \times W \times C}$, where H and W are image height and width, while C denotes the number of categories of teeth. Especially, a categorical vector $y \in Y_{mask}$ is multi-hot encoded, since nearby teeth can overlap in a panoramic radiograph because of the 2D projecting. With the categorical mask, *SegNet* further performs denoising by keeping the largest island of segmentation per tooth type, and localizes teeth by deriving their bounding boxes as shown in Figure 4.1(2). As indicated in Figure 4.1(b), *SegNet* consists of 2D convolutional layers followed by a *Sigmoid* transfer in order to perform the multi-label prediction. In our experiments, we set $C = 32$ for modeling the full set of teeth of an adult, including the four wisdom teeth that possibly exist for some individuals.

4.2.1.3 *ReconNet*

ReconNet samples the feature patch of a tooth, and maps the 2D patch into the 3D occupancy probability map $Y_{recon} \in \mathbb{R}^{H_p \times W_p \times D_p \times 2}$ of that tooth, where H_p, W_p, D_p are patch height, width and depth, respectively. The 2D feature patch is cropped from the feature map derived from *ExtNet*, while the cropping is guided by the tooth localization derived from *SegNet*. Similar to [CXG16, HRR18], *ReconNet* has an encoder-decoder structure consisting of both 2D and 3D convolutions. The encoder employs 2D convolutions, and its output is flattened into a 1D feature vector for the fully connected operation; while the decoder employs 3D convolutions to map this feature vector into the target dimension. Since our *ReconNet* operates on small image patches rather than the whole x-ray, the reconstruction can be done at high resolutions for restoring the details of teeth. In this work, we set $H_p = 120$, $W_p = 60$, $D_p = 60$ since all teeth fit into this dimension.

4.2.1.4 *Teeth Assembling*

By assembling the predicted tooth volumes according to their estimated localization from x-ray segmentation, we can achieve the 3D reconstruction of the cavity as a flat plane without the depth information about the cavity. This reconstruction is an estimation for the real cavity that is projected along the dental arch. Many previous work has investigated the modeling and prediction of the dental arch curve [NHS01]. In this work, we employ the β function model introduced by [BHF98], which estimates the curve by fitting the measurements of cavity depth and width (Figure 4.1(4)). As the final step, our prediction of teeth for the whole cavity (Figure 4.1(5)) can be simply achieved by bending the assembled flat reconstruction along the estimated arch curve.

4.2.2 Training Strategy

The loss function of *X2Teeth* is composed of two parts: segmentation loss L_{seg} and patch-wise reconstruction loss L_{recon}. For L_{seg}, considering that a pixel can be of multiple categories because of teeth overlaps on X-rays, we define the segmentation loss as the average of dice loss across all categories. Denote the segmentation output Y_{seg} at a pixel (i, j) to be a vector $Y_{seg}(i, j)$ of length C, where C is the number of possible categories, then

$$L_{seg}(Y_{seg}, Y_{gt}) = 1 - \frac{1}{C} \sum_{C} \frac{\sum_{i,j} 2Y_{seg}(i,j)Y_{gt}(i,j)}{\sum_{i,j} (Y_{seg}(i,j) + Y_{gt}(i,j))}, \tag{4.1}$$

where Y_{gt} is the multi-hot encoded segmentation ground-truth. For L_{recon}, we employ the 3D dice loss for defining the difference between the target and the predicted volumes. Let the reconstruction output Y_{recon} at a pixel (i, j, k) be a Bernoulli distribution $Y_{recon}(i, j, k)$, then

$$L_{recon}(Y_{recon}, Y_{gt}) = 1 - 2\frac{\sum_{c=1}^{2} \sum_{i,j,k} Y_{recon}(i,j,k)Y_{gt}(i,j,k)}{\sum_{c=1}^{2} \sum_{i,j,k} (Y_{recon}(i,j,k) + Y_{gt}(i,j,k))}, \tag{4.2}$$

where Y_{gt} is the reconstruction ground-truth.

We employ a two-stage training paradigm. In the first stage, we train *ExtNet* and *SegNet*

for the teeth localization by optimizing L_{seg}, such that the model can achieve an acceptable tooth patch sampling accuracy. In the second stage, we train the whole *X2Teeth* including *ReconNet* by optimizing the loss sum $L = L_{seg} + L_{recon}$ for both localization and reconstruction. Note that Adam optimizer is used for optimization. For each GPU, we set the batch size of panoramic radiograph as 1, and the batch size of tooth patches as 10. Besides, standard augmentations are employed for images, including random shifting, scaling, rotating and adding Gaussian noise. Finally, we implement our framework in Pytorch, and trained for the experiments on three NVidia Titan Xp GPUs.

4.3 Experiments

In this section, we validate and demonstrate the capability of our method for the teeth reconstruction from the panoramic radiograph. First, we introduce our in-house dataset of X-ray and panoramic radiograph pairs with teeth annotations from experts. Second, we validate *X2Teeth* by comparing with two state-of-the-art single view 3D reconstruction methods. Finally, we look into the performance of *X2Teeth* on the two sub-tasks of teeth localization and single tooth reconstruction.

4.3.1 Dataset

Ideally, we need paired data of the panoramic radiographs and CBCT scans captured from the same subject to train and validate *X2Teeth*. However, in order to control the radiation absorbed by subjects, such data pairs can rarely be collected in clinical settings. Therefore, we take an alternative approach by collecting high resolution CBCT scans and synthesize their corresponding panoramic radiographs. Such synthesis is valid since CBCT scans contain full 3D information of cavity, while panoramic radiographs are the 2D projections of them. Several previous work has demonstrated promising results for high quality synthesis, and in our work, we employ the method of Yun *et al.* [YYH19] for building our dataset. Our in-

Method	IoU	DA	DF
3D-R2N2	0.398 ± 0.183	0.498 ± 0.101	0.592 ± 0.257
DeepRetrieval	0.448 ± 0.116	0.594 ± 0.088	0.503 ± 0.119
X2Teeth (ours)	**0.682 ± 0.030**	**0.702 ± 0.042**	**0.747 ± 0.038**

Table 4.1: Comparison of reconstruction accuracy between *X2Teeth* and general purpose reconstruction methods in terms of IoU, detection accuracy (DA) and identification accuracy (FA). We report each metric in the format of *mean* ± *std*.

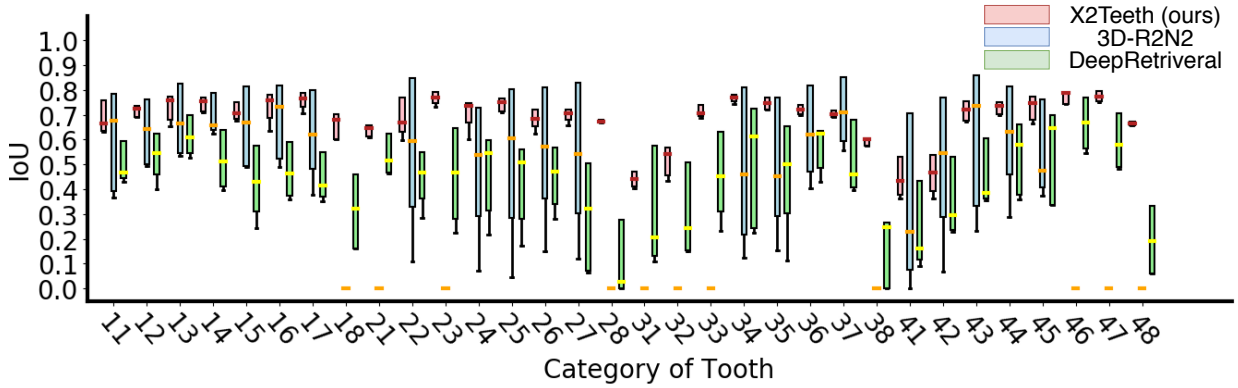


Figure 4.2: IoU comparison of different tooth types between *X2Teeth*, 3D-R2N2, and DeepRetrieval.

house dataset contains 23 pairs of 3D CBCT scans and panoramic radiographs, each with a resolution ranging form 0.250 mm to 0.434 mm. All CBCT scans and panoramic radiographs are first labeled with pixel-wise tooth masks by 3 annotators, and then reviewed by 2 board-certificated dentists. Finally, we randomly split the dataset into 15 pairs for training, 1 pair for validation, and 7 pairs for testing.

4.3.2 Overall Evaluation of Teeth Reconstruction

We compare our *X2Teeth* with two general purpose reconstruction methods that have achieved state-of-the-art performance as baselines: 3D-R2N2 [CXG16] and DeepRetrieval [TRR19].

3D-R2N2 employs an encoder-decoder network to map the input image to a latent representation, and reasons about the 3D structure upon it. To adapt 3D-R2N2 for high resolution X-rays in our tasks, we follow [TRR19] by designing the output of the model to be 128^3 voxel grids, and up-sampling the prediction to the original resolution for evaluation. DeepRetrieval is a retrieval-based method that reconstructs images by deep feature recognition. Specifically, 2D images are embedded into a discriminative descriptor by using a ConvNet [KSH12] as its representation. The corresponding 3D shape of a known image that shares the smallest Euclidean distance with the query image according to the representation is then retrieved as the prediction.

Quantitative Comparison. We evaluate the performance of models with intersection over union (IoU) between the predicted and the ground-truth voxels, as well as detection accuracy (DA) and identification accuracy (FA) [CLW19]. The formulations of the metrics are:

$$IoU = \frac{|D \cap G|}{|D \cup G|}, \quad DA = \frac{|D|}{|D \cap G|} \quad and \quad FA = \frac{|D \cap G|}{|D|}, \tag{4.3}$$

where G is the set of all teeth in ground-truth data, and D is the set of predicted teeth. As shown in Table 4.1, *X2Teeth* outperforms both baseline models significantly in terms of all three metrics. Specifically, *X2Teeth* achieves a mean IoU of 0.682, which outperforms 3D-R2N2 by 1.71×, and DeepRetrieval 1.52×. Similarly, Figure 4.2 reveals IoUs for all the 32 types of tooth among the three methods, where our method has the highest median and the smallest likely range of variation (IQR) for all tooth types, which shows the consistent accuracy of *X2Teeth*. Yet, we also find that all algorithms have a lower accuracy for wisdom teeth (numbering 18, 28, 38, and 48) than the other teeth, indicating that the wisdom teeth are more subject-dependent, and thus difficult to predict.

Qualitative Comparison. Figure 4.3 visualizes the 3D reconstructions of a panoramic radiograph (Figure 4.3(a)) from the testing set, which clearly shows our *X2Teeth* can achieve more appealing results than the other two methods. As for 3D-R2N2, its reconstruction (Figure 4.3(e)) misses several teeth in the prediction as circled with green boxes, possibly

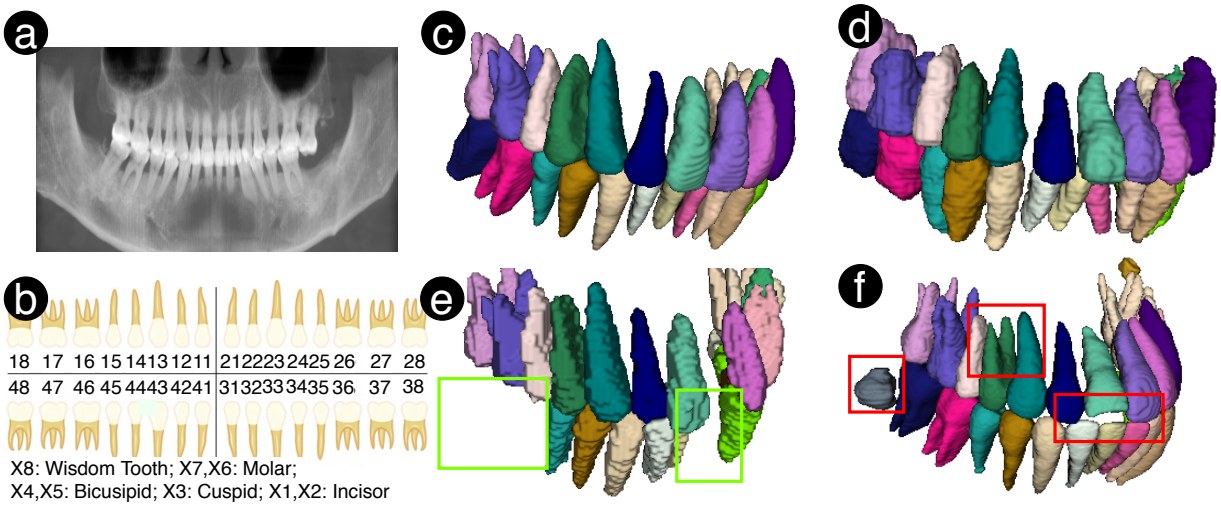


Figure 4.3: Comparison of the reconstruction between (d) *X2Teeth* (ours), (e) 3D-R2N2, and (f) DeepRetrieval. (a) shows the input panoramic radiograph from the testing set, (c) shows the ground-truth of reconstruction, and (b) is the teeth numbering rule.

because spatially small teeth can lose their representations within the deep feature map during the deep encoding process. The similar issue of missing tooth in predictions has also been previously reported in some teeth segmentation work [CLW19]. Moreover, the reconstruction of 3D-R2N2 has coarse object surfaces that lack details about each tooth. This is because 3D-R2N2 is not compact enough and can only operate at the compressed resolution. As for DeepRetrieval, although the construction (Figure 4.3(f)) has adequate details of teeth since its retrieved from high-resolution dataset, it fails to reflect the unique structure of individual cavity. The red boxes in Figure 4.3(f) point out the significant differences in wisdom teeth, tooth root shapes, and teeth occlusion between the retrieved teeth and the ground-truth. Comparing to these two methods, *X2Teeth* has achieved a reconstruction (Figure 4.3(d)) that can reflects both the unique structure of cavity and the details of each tooth, by formulating the task as the optimization of two sub-tasks for teeth localization and single tooth reconstruction.

4.3.3 Sub-task Evaluations

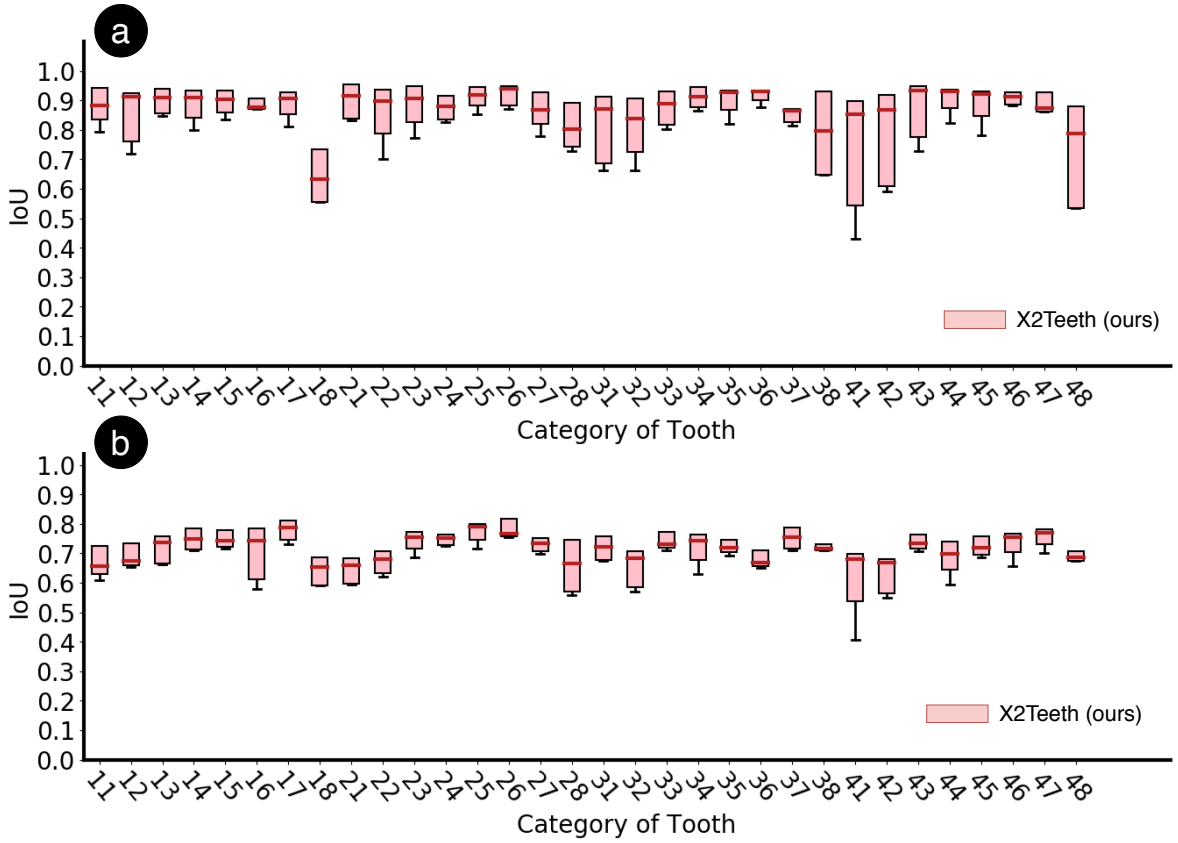


Figure 4.4: (a) Segmentation IoUs of various teeth for the teeth localization sub-task. (b) Reconstruction IoUs of various teeth for the single tooth reconstruction sub-task.

For better understanding the performance of *X2Teeth*, we evaluate its accuracy on the two sub-tasks of teeth localization and single tooth reconstruction. Figure 4.4(a) shows the IoUs of different teeth for the 2D segmentation, where our method achieves an average IoU of 0.847±0.071. The results validate that *X2Teeth* can accurately localize teeth, which enables the further sampling of tooth patches for the patch-based reconstruction. We also observe that the mean segmentation IoU for the 4 wisdom teeth (numbering X8) is 0.705±0.056, which is lower than the other teeth. This is possibly because they have lower contrasts with surrounded bone structures, such that are more challenging to segment. Figure 4.4(b)

demonstrates the IoUs of different types of teeth for the single tooth reconstruction, where our method achieves a mean IoU of 0.707±0.044. Still, wisdom teeth have the significantly lower mean IoU of 0.668±0.050, which can be contributed by the lower contrast with surroundings, less accurate localization, and the subject-dependent nature of their shapes. Moreover, incisor teeth (numbering X1 and X2) are observed to have less accurate reconstructions with the mean IoU of 0.661±0.031. We argue the reason can be their feature vanishing in the deep feature maps considering their small spatial size.

CHAPTER 5

3D Reconstruction from Single Image with Implicit Neural Representation

Radiological 3D reconstruction from limited 2D images has attracted increasing attention with the development of deep generative models in the past few years. Recent works like [SLY21, YGM19, HRR18, KLL19] have shown the feasibility of 3D reconstruction from only one or two X-ray images, which provides an alternative solution to 3D imaging where only 2D imaging equipment is available. Due to the low radiation generated by 2D imaging equipment, these methods also bring a new choice in radiological examination for patients who are sensitive to radiation. For example, research in [Bro09] shows that the X-ray imaging method could take as much as 200 less radiation than Cone Beam Computed Tomography (CBCT), a fast and low-radiation type of Computed Tomography (CT) and is widely used in dental radiology. Therefore, developing fast and accurate translation models could potentially bring great progress in medical imaging.

However, most of these cross-dimension translation models learn to explicitly generate a 3D image by auto-encoding and adversarial learning from paired X-ray images and CT scans. Consequently, the reconstruction quality is sensitive to the diversity and scale of training data. In dental imaging, restoring curved mandibular shapes brings additional challenges as only a single panoramic X-ray (PX) image is available. To solve this problem, recent studies like Oral-3D [SLY21] and X2Teeth [LSY20] utilize individual prior knowledge when training the model, i.e., dental arch shape extracted from buccal images or instance annotations of teeth at the pixel level. Yet these complicated operations could bring conspicuous

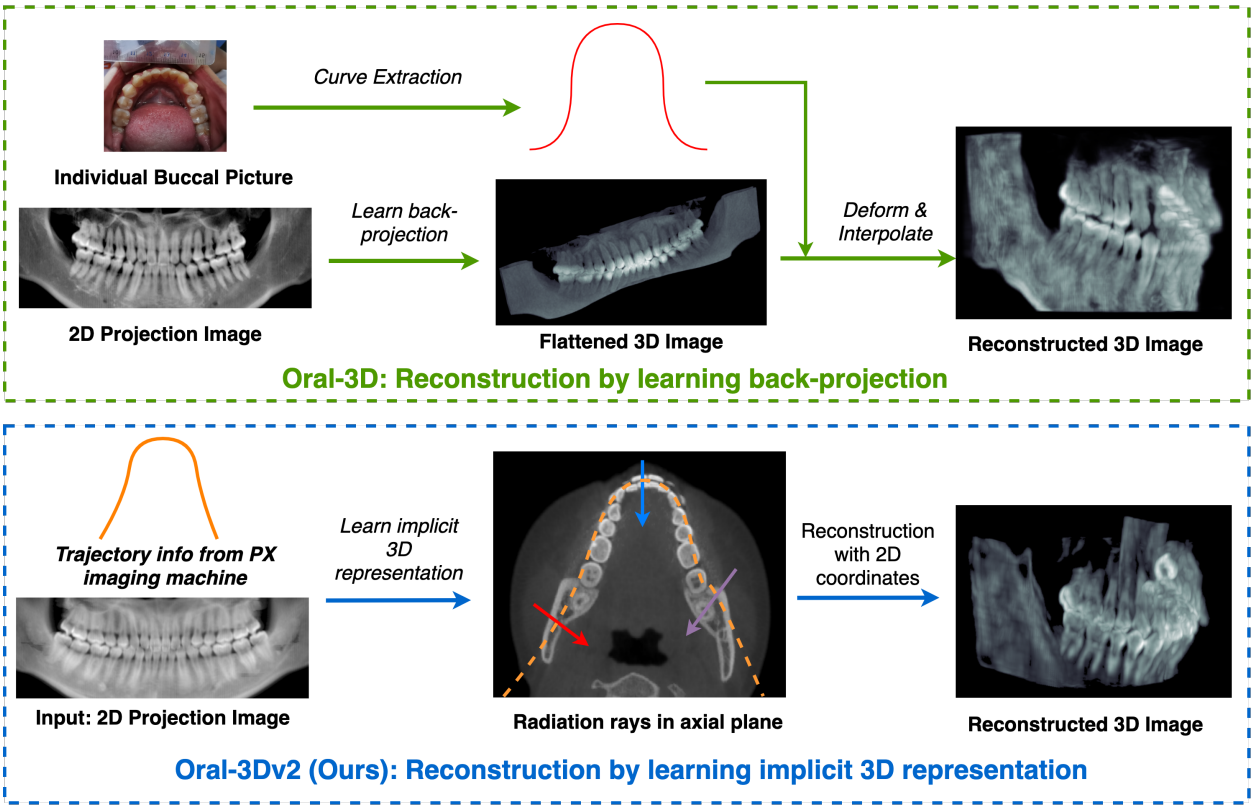


Figure 5.1: We compare our new model (blue) and Oral-3D (green) in this picture. Oral-3D first learns a back-projection model with paired images to generate a flattened 3D oral structure. Then it deforms the flattened image into a curved shape according to the individual dental arch shape acquired from the patient. In our model, we learn an implicit 3D representation of the oral structure only from the projection information, i.e., projection image and X-ray tube trajectory that is pre-defined by the equipment manufacturer and independent of individuality. After the model is well-trained, the 3D object is reconstructed by inferring the density distribution in 3D space from the implicit representation model and 2D coordinates.

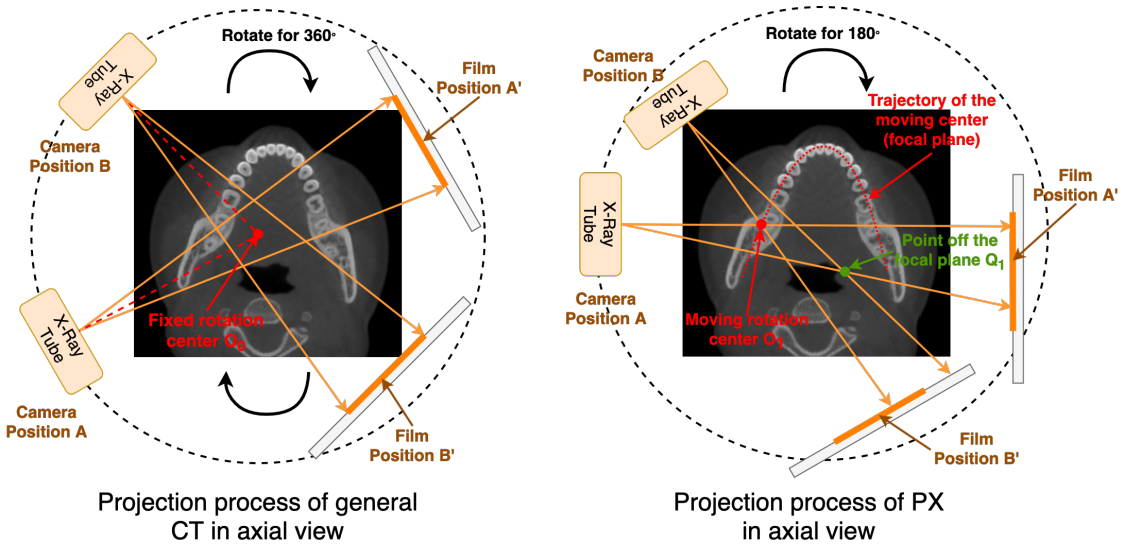


Figure 5.2: We show the comparison of imaging process of general CT (including CBCT) and PX in this picture. In CT, the X-ray tube and the film moves together around a fixed rotation center for 360 degrees, where the film receives all X-rays sent from the tube. In PX imaging, the X-ray tube and the film rotates around a moving center, whose trajectory fits the curve of the mandible. Therefore, points that are around and away from the trajectory receive different levels of radiation during the imaging. For example, when the tube and the film moves from A to B in the right picture, the red point is projected twice while the green point is only projected once. This could make the image show more information of the imaging target at the red point over the green point.

miss alignment during reconstruction, thus greatly hindering clinical applications in dental examinations. As a comparison, implicit representation models [MST21, MES22] provide a new solution in 3D reconstruction from 2D images. But these models rely on learning from abundant images viewed from various directions, which is hard to apply in radiology due to differences in imaging principles and inflexibility in imaging angles.

To address these limitations, we propose a new framework for 3D oral reconstruction from a single 2D panoramic X-ray (PX) image. Different from previous work like Oral-3D, which learns a back projection function to explicitly predict the reconstruction result by learning from paired images and prior knowledge of the individual dental arch shape, our model could learn 3D reconstruction simply from a single X-ray image with the projection settings from the imaging equipment. A comparison between Oral-3D and our method can be seen in Figure 5.1, where only projection data is required during the reconstruction in our method.

Unlike models in [CFB22, ZDW23] that utilize a single X-ray or two orthogonal X-ray images, our method could utilize the rich projection information during a panoramic scan with our advanced architecture. Specifically, we use a deep learning network to learn a mapping function between coordinates and density values of voxels in the 3D space, i.e., Hounsfield Unit (HU). To take advantage of the imaging process in panoramic imaging, we propose a multi-head model that outputs a bunch of voxel values at the same time given a 2D coordinate, which proves to be both efficient and effective over existing implicit representation models. Furthermore, to accommodate the imaging object in radiology, we utilize a dynamic sampling strategy to improve the reconstruction quality by acquiring points along radiation rays in random resolutions. Extensive experiments show that our model could significantly outperform state-of-the-art methods in 3D oral reconstruction both qualitatively and quantitatively. In conclusion, we summarize our contribution as follows:

- Different from previous approaches in 3D oral reconstruction, such as Oral-3D[SLY21] and X2Teeth[LSY20], our model could achieve superior performance without training

from any paired data, individual prior knowledge, or annotations.

- We propose an efficient implicit 3D representation model that maps a 2D coordinate into a bunch of 3D density values. This could reduce the computation complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ when reconstructing a $N \times N \times N$ object during both training and inference.

- We also propose a dynamic sampling strategy when sampling points from radiation rays with an adaptive projection method. This could encourage the model for higher reconstruction quality by learning a smooth density distribution in the 3D space .

5.1 Background and Related Works

5.1.1 Radiology in dental imaging

There are mainly two radiological imaging methods in dental health, i.e., CBCT and PX. CBCT generates a 3D image of the oral cavity with rich spatial information of teeth, thus widely used in orthodontics and tumor surgery. As a comparison, PX is a faster and lightweight method used in the examination before pulling or planting teeth, where a 2D panoramic picture is taken of all the teeth along the mandibular curve. We show illustrations of these two imaging methods viewed in the axial plane in Figure 5.2. In CBCT, as shown in the left image, the X-ray tube and the film moves around a fixed center for 360°. The 3D image is then reconstructed from sinogram signals in 2D space [Her09], which is feasible as each point is projected from different directions during the imaging. In PX, the X-ray tube and the film move around a moving center from one side to the other. The trajectory, also named the focal plane, generally fits the curved shape of the mandible, leading to different projection levels for tissues at various locations. For example, as shown in the right picture, the red point at O_1 located on the moving trajectory is projected twice while the green point at Q_1 off the moving trajectory is projected only once when the X-ray tube and the film

move from A to B. Therefore, the image shows stronger signals for tissues at O_1 than Q_1, thus generating a clear picture of objects around the focal plane. Like CBCT, points around the focal plane also receive multiple projections in PX but are not used to recover any 3D information during imaging. This feature is taken advantage of our proposed model for 3D oral reconstruction.

5.1.2 Implicit representation in 3D reconstruction

Implicit representation has been demonstrated to be a promising method in the task of 3D reconstruction since the work of neural radiance field (NeRF) [MST21], where the researchers use the deep neural network to map 5D coordinates of spatial location and viewing direction into the density and emitted radiance of a voxel. During the inference, the model could generate images from any position by rendering along the rays sent from the observation point. Based on this framework, D-NeRF [PCP21] takes time as additional input to the system for the reconstruction of dynamic scenes. Nerfies [PSB21] use an additional continuous volumetric deformation field to generate deformable photo-like scenes. Although our method also utilizes implicit 3D representation, there are still big differences due to the characteristic of the imaging process in radiology: 1) The movement of an X-ray tube has less degree of freedom (DoF) than a camera, thus leading to limited projection rays in both directions and origins. 2) The predicted density distribution represents the values of HU instead of the differential probability. 3) The reconstruction object should be view-independent.

5.1.3 Cross-dimension translation in radiology

Cross-dimension translation in radiology images between 2D and 3D by deep neural networks starts from the work of [HRR18], where the authors use an encoding-decoding network to learn a back projection function that maps a 2D projection image into 3D density volumes for the skull of mammals. Following this work, [YGM19, KDK20] improve the reconstruc-

tion quality for the abdomen and knees by utilizing bi-planar X-ray images and adversarial networks. In dental healthcare, Oral-3D [SLY21] first uses a single panoramic X-ray image to reconstruct the 3D oral structure. X2Teeth [LSY20] trains three networks to reconstruct and segment the teeth in 3D space with annotated X-ray images. Our model can be seen as an extension of these works that focus on the same problem but with a different technical solution: 1) In contrast to learning explicitly by auto encoding or adversarial learning, our method learns the representation of the 3D object in an implicit way. 2) Our model relies no more on paired 2D and 3D images or individual prior knowledge to restore the mandibular curve.

5.2 Methodologies

5.2.1 Problem Definition

Given a pair of projection image I and the trajectory of the rotation center O during the PX imaging, the object is to find an implicit 3D representation $V : \mathbf{p} \rightarrow h$ that maps 3D coordinates $\mathbf{p}$ into HU values h and minimizes the mean square error against the projection image given the imaging function $F(\cdot)$. The problem can be defined as:

$$\arg\min_{V(\cdot)} ||F(V, O) - I||_2. \tag{5.1}$$

With the sampled rotation center point at O_i and the corresponding projection image I_i, the reconstruction problem in Eq (5.1) could be solved by optimizing the below objective function:

$$L_{obj} = \sum_{i=1}^{N} ||f(V(\mathbf{p_1}), V(\mathbf{p_2}), \cdots, V(\mathbf{p_m})) - I_i||_2, \tag{5.2}$$

where $\mathbf{p_1}, \cdots, \mathbf{p_m}$ are the coordinates of points sampled along the radiation ray sent from the X-ray tube, and f is the projection function that maps multiple voxel values into a single one. To distinguish with existing NeRF-like models V_{NeRF}, we refer to our implicit representation model as V_{NeXF} (short for neural X-ray field) to represent the field function

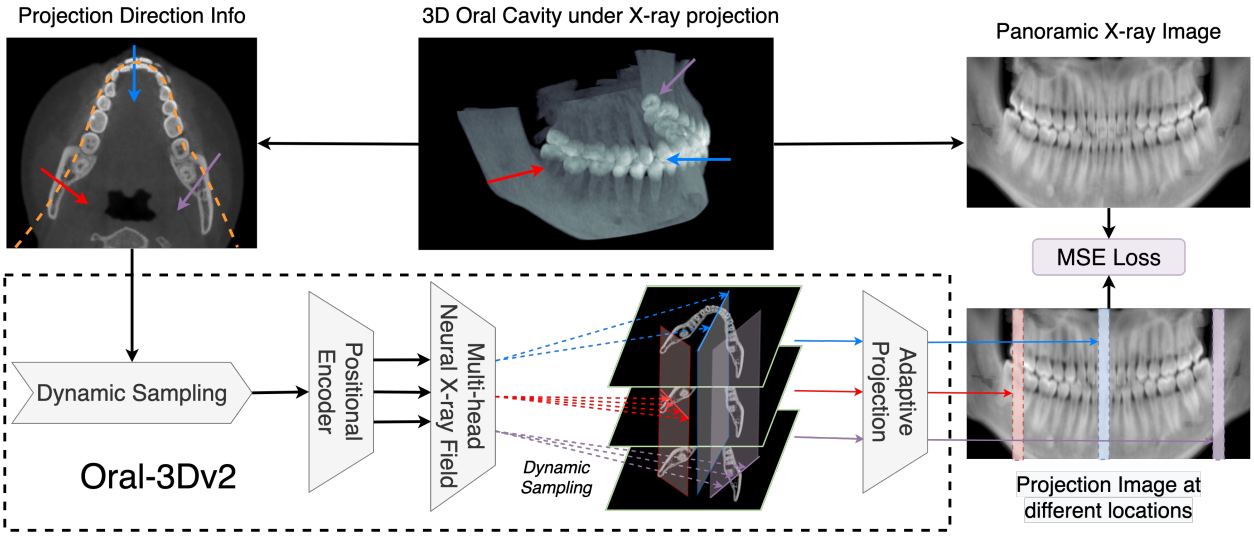


Figure 5.3: This image provides an overview of our model, i.e., Oral-3Dv2. Starting with radiation rays, we use a dynamic sampler to acquire sample points on each ray at random sampling rates. Then, we employ our proposed multi-head neural X-ray field (NeXF) with a positional encoder to predict densities in the 3D space. The NeXF outputs a bunch of HU values from a single 2D coordinate. Next, we generate a projection image adapting to the dynamic resolution during sampling. Finally, we calculate the MSE loss between the projection slice and the ground-truth image to update parameters of our implicit representation model.

in X-ray imaging.

5.2.2 Overview

We show an overview of our proposed model in Figure 5.3, where paired rays and rendering results are taken as input to train the implicit representation model V_{NeXF}. Given the direction and origin of the projection ray inferred from the moving trajectory $T(O)$ of the X-ray tube, we first generate points along the radiation ray at a random sampling rate. The sampled coordinates are then taken as the input of a positional encoding module, followed by our proposed NeXF model, to generate the projection results. The model is updated according to Eq. 5.2 until converge. Although our framework looks similar to NeRF-like models, we have three major differences due to the feature of PX imaging, where the radiation rays are almost parallel to the axial plane. First, our NeXF has a multi-head structure, whose input is a 2D coordinate and output is a bunch of voxel values in the same axial location. Second, we use a dynamic sampling strategy instead of a pair of coarse and fine networks to improve the reconstruction quality. Third, our model is view-independent as it is unreasonable for various density values for the same voxel in radiology.

5.2.3 Dynamic Sampling

NeRF-based models generally utilize a pair of coarse and fine networks to determine the sampling rate along the rays due to multiple free spaces and occluded regions in their 3D objects viewed from the outside. However, this is not applicable to radiology as the aim of imaging is to observe the inside structure of the object. Therefore, points along the radiation rays should be evenly sampled to evenly indicate the density variance in 3D space. To accommodate this, we propose a dynamic sampling strategy that acquires points from radiation rays in a random resolution to improve spatial smoothness without introducing additional new networks. As shown in Figure 5.3, radiation rays sent from different directions

(represented by the red, blue, and purple arrows) acquire different numbers of sampling points when generating projection images. We show that the variance in sampling rate during projection in training could significantly improve the reconstruction quality in the ablation experiments.

5.2.4 Positional Encoding

Positional encoding has been widely used in implicit 3D representation models due to the tendency of learning low-frequency details as revealed in recent research like [RBA19, TSM20]. To solve this spectral bias problem, frequency encoding is introduced in [MST21, BMT21] to encourage the model to exploit high-dimension spatial information during reconstruction. We follow the same way as in [VSP17] that utilizes multi-resolution sequence to encode the coordinate value p from $\mathbf{p}$ into L levels of embedding as:

$$\begin{aligned} Enc(p) =&(\sin\left(2^0 p\right), \sin\left(2^1 p\right), \cdots, \sin\left(2^{L-1} p \right. \\ &\left. \cos\left(2^0 p\right), \cos\left(2^1 p\right), \cdots, \cos\left(2^{L-1} p\right)\right) \end{aligned} \tag{5.3}$$

5.2.5 Multi-head Neural X-ray Field

Different from NeRF-based models, where the camera has more freedom in position and angle, the X-ray tube in radiological scans generally moves in a fixed trajectory, leading to a limited direction and origin of radiation rays during the imaging. For example, radiation rays that pass through the oral cavity are approximately parallel to the axial plane. Taking advantage of this feature, we propose a different radiance field model that predicts a bunch of voxel values in the 3D space from a 2D coordinate. Given that the 3D object in radiology should be view-independent, our implicit representation model V_{NeXF} can be defined as:

$$V_{NeXF} : (x, y) \rightarrow (v_{x,y,1}, v_{x,y,2}, \cdots, v_{x,y,z_n}), \tag{5.4}$$

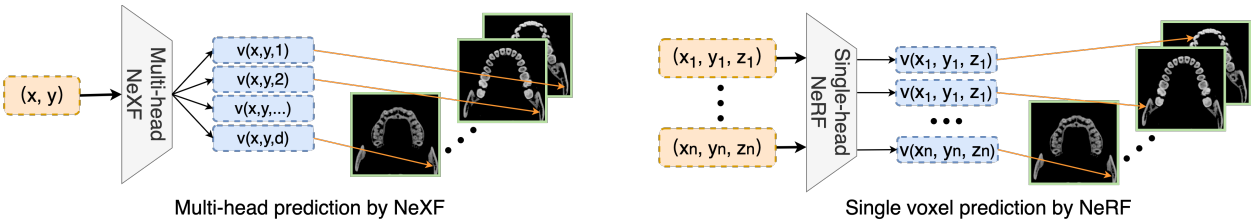


Figure 5.4: We show the comparison of implicit representation model between NeXF and NeRF models in this picture. The NeRF-like models have a single-head structure that outputs the specific voxel value of the given input. However, in NeXF the model only takes in a 2D coordinate by predicts a bunch of voxel values with its multi-head architecture. This architecture could best fit the imaging process of PX and significantly decrease the computation complexity during both training and inference.

in comparison to V_{NeRF} defined as:

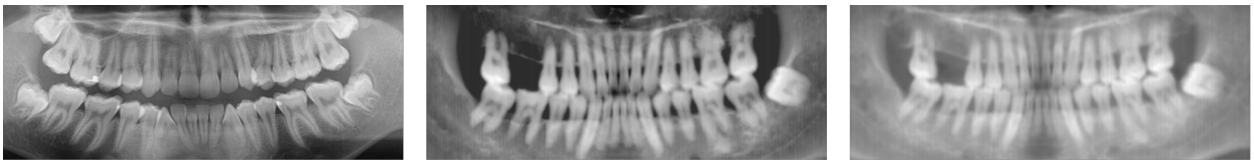
$$V_{NeRF} : (x, y, z, \theta, \phi) \rightarrow v_{x,y,z}. \tag{5.5}$$

We compare the difference between V_{NeXF} and V_{NeRF} in Figure 5.4. V_{NeXF} uses a multi-head architecture that takes in a 2D coordinate as input and outputs z_n number of voxel values, where z_n is the same as the resolution of reconstruction object in z axis. In contrast, V_{NeRF} only predicts a single value per 5D coordinate. Therefore, V_{NeXF} can reduce the computational complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(N^2)$ compared with V_{NeRF} during the reconstruction of $N \times N \times N$ object.

5.2.6 Adaptive Projection

Following Beer-Lambert absorption-only model [Dri03], the fraction α of radiation arriving at the film after traveling volumes with spatially-varying density $\mu(t)$ along a ray parameterized with variable t within $[t_n, t_f]$ could be expressed as:

$$\alpha = \exp \int_{t_n}^{t_f} \mu(t) dt, \tag{5.6}$$



Real PX Image | Projection with soft rendering | Projection without soft rendering

Figure 5.5: Comparison of different rendering methods in PX imaging. We can see that with soft rendering the generated PX image has a closer contrast with the real PX image (obtained from Internet). The real PX image looks more clear due to the high resolution of the PX machine.

Method	PSNR	Dice	SSIM	Overall
NAF [ZZL22]	18.35±0.86	57.20±3.94	60.69±2.69	65.93
GAN [GPM20]	16.71±0.89	75.10±1.46	63.96±7.03	76.93
ResEncoder [HRR18]	18.26±0.50	72.67±1.56	62.52±5.56	75.49
Oral-3D [SLY21]	18.59±0.70	76.88±1.26	65.94±4.24	78.60
Ours	**20.34**±0.62	75.34±3.95	**81.06**±1.61	**86.04**

Table 5.1: Evaluation of 3D oral reconstruction by PSNR, SSIM(%), and Dice.

where $\mu(t)$ is the attenuation coefficient and could be converted into a HU value by:

$$H(\mu) = 1000 \times \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}} - \mu_{\text{air}}}, \tag{5.7}$$

where μ_{water} and water are constant values and μ is the accumulated attenuation coefficient along the ray path. By sampling along the radiation ray, Eq. (5.6) can be converted into:

$$\alpha = \exp(\sum_{i}^{\lfloor t_f - t_n \rfloor} \mu_i). \tag{5.8}$$

Therefore, the projection function $f(\cdot)$ in Eq. (5.2) can be adaptively expressed with our proposed implicit representation model V and the dynamic sampling rate N_s as:

$$f(\cdot) = H(\frac{\sum_i^{\lfloor N_s(t_f - t_n) \rfloor} \mu_i}{N_s}) = H(\frac{\sum_i^{\lfloor N_s(t_f - t_n) \rfloor} H^{-1}(V(\mathbf{p_i})}{N_s})), \tag{5.9}$$

where p_i is the i-th sample point within $[t_n, t_f]$.

5.3 Experiments

5.3.1 Dataset

We collect a dataset consisting of 80 CBCT dental scans as groundtruth of the 3D oral structure and source images to simulate PX imaging. We divide the model into two groups: 1) 60 cases used for training models based on auto-encoding and adversarial learning, and 2) 20 cases used for inference and validation for all models. The CBCT scan is resized into a size of $288 \times 256 \times 160$ using trilinear interpolation to minimize influence brought by imaging machines.

5.3.2 PX Imaging Simulation from CBCT

The moving trajectory of rotation center in PX imaging is fitted by the beta function as:

$$y = 256 - beta(x/288, 3.6, 3.6) * 100 - 25. \tag{5.10}$$

We split the trajectory curve equally into 576 pieces and assume the radiation rays evenly cross each small curve in angles between $-\pi/4$ and $\pi/4$. Research in [SS06][Arm06] show that HU is unreliable in CBCT scans due to variations in gray-scale values for different areas in the scan, especially when the imaging areas have the same density but different relative positions. Therefore, we follow the same method proposed in [YYH19, SLY21] during projection to get realistic PX images from CBCT. Then the projection function $f(\cdot)$ in Eq. (5.9) can be rewritten into $\hat{f}(\cdot)$:

$$\hat{f}(\cdot) = S \cdot \log(\sum_{i}^{\lfloor N_s(t_f - t_n) \rfloor} e^{\frac{V(\mathbf{p_i})+C}{S}}) - \log N_s - C, \tag{5.11}$$

where $C = H(\mu_{air})$. Comparisons among real PX image and simulated images generated by $f(\cdot)$ and $\hat{f}(\cdot)$ can be seen in Fig 5.5, where PX images simulated by $\hat{f}(\cdot)$ has a more closer contrast as real images.

5.3.3 Hyper-parameters and Network Architecture

We select $S = 1200$ in Equation (5.11) to distinguish air and soft tissues in HU. The sampling rate N_s for each radiation ray during training follows a uniform distribution in $[0.25, 1, 25]$. The level L used in positional encoding is selected to be 16 with the normalizaton of coordinates into $[-1, 1]$. For the multi-head NeXF, we use a 12-layer fully-connected neural network with residual connections and set the number of heads as 160, which is consistent with the CBCT data.

5.3.4 Training and Evaluation

The model is trained for 20k iterations with a batch size of 64. The model is optimized by Adam with a learning rate of 0.0001. We use structural similarity index measure (SSIM) [WBS04], dice coefficient (DC), and peak signal-to-noise ratio (PSNR) to evaluate the reconstruction results. We also use the averaged score proposed in [SLY21] as the overall metric.

5.3.5 Baseline Models

We compare our method with baseline models that can be grouped into two categories. The first group including GAN [GPM20], Oral-3D [SLY21], and Res-Encoder [HRR18]. These models are trained with the 60 paired simulated X-ray images and CBCT images and learn the prediction of explicit 3D representation with either adversarial learning or auto-encoding. We put our model in the second group with NAF [ZZL22], another implicit representation with the same framework attenuation coefficients in 3D space.

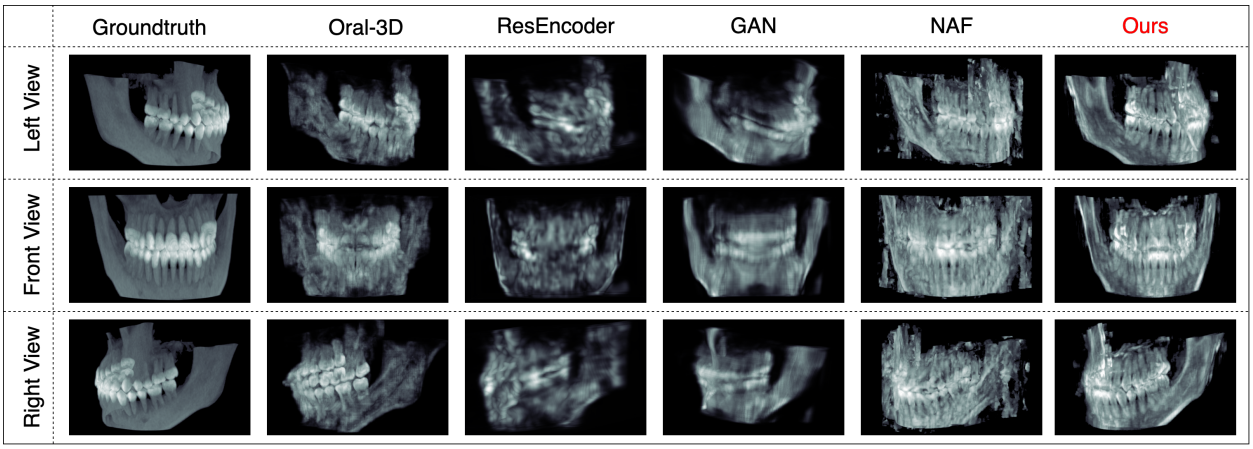


Figure 5.6: Comparison of 3D oral reconstruction by different methods from PX imaging. The reconstruction results are shown by maximum projection to compare density details. We could easily find that our method show the best performance with clear density density distributions and teeth boundaries.

M	D	P	PSNR	SSIM(%)	Dice(%)	Overall	Drop
✗	✓	✓	16.68±0.74	73.62±5.49	61.25±4.57	72.76	-13.28
✓	✗	✓	16.80±0.71	61.44±5.87	73.29±3.24	72.91	-13.13
✓	✓	✗	16.57±1.08	63.63±3.07	70.28±3.28	72.25	-13.79

Table 5.2: Ablation study by removing each component in our proposed method. M: Multi-head Prediction, D: Dynamic Sampling, P: Change $\hat{f}(\cdot)$ to $f(\cdot)$ in training

5.4 Results

5.4.1 Qualitative Comparison

We first show qualitative comparison in Figure 5.6 to compare the reconstruction results of baseline models. More results for Oral-3D and our model can be seen in supplemental materials. We can see that although ResEncoder and GAN could restore the curved shape of mandible without any prior knowledge, these models fail to recover the detail density distribution in the reconstruction results. For NAF, the model could recover the curved shape and density variance. But the results contain too much noise and is hard to identify the teeth shape. For Oral-3D, the model could restore both shape and teeth details with the help of individual dental arch shape. However, its reconstruction quality is obviously lower than our method, especially for the details of density change between teeth root and the mandible.

5.4.2 Quantitative Comparison

We then show the quantitative comparison by the proposed metrics in Table 5.2.5. The dice score is computed by setting a threshold at 500 HU to extract the bone from soft tissues. We could see our model could significantly outperform other models, with improvement of +5 in SSIM and +7.5 in the overall score against the state-of-the-art method without training on paired images or deformation by individual prior knowledge. To be noted, Oral-3D has a better Dice score but lower performance in PSNR and SSIM. This is consistent with the visualized results shown in Figure 5.6, where Oral-3D restores less density details.

5.4.3 Ablation Study

We conduct an ablation study to evaluate the contribution of each component in our model: 1) replace the multi-head field function with a single-head predictor and taking in 3D co-

ordinates as input for the positional encoder; 2) use a fixed sampling rate of $N_s = 1$ to generate sample points on projection rays; 3) change the rendering function in Eq. (5.11) to Eq. (5.9) when training the model. We use the letters M, D, and P to represent these changes. Results are shown in Table 5.4, where the performance drops significantly (about -13 in Overall) when changing any module. We could see the dynamic sampling strategy can greatly improve the reconstruction quality without introducing additional models. And the multi-head architecture has stronger ability in implicit representation in radiation imaging.

CHAPTER 6

Conclusion and Future Work

6.1 Research Summary

In this dissertation, we have delved into the realm of image-to-image translation from two distinct angles: cross-domain translation and cross-dimension translation. We succinctly conclude the four innovative methods as following:

6.1.1 Perceptual Learning for Multi-domain Translation

We introduce *MDT-Net* to achieve multi-domain transfer within one single model trained by unpaired and unlabeled images with perceptual supervision. We disentangle the anatomy content and domain variance by an encoder-decoder network and multiple domain-specific transfer modules. Furthermore, extensive experiments on the task of transfer among three domains of OCT images have validated the advantage of *MDT-Net* qualitatively and quantitatively.

6.1.2 Progressive Energy-based Model for High-resolution Image Translation

We present a novel approach that combines energy-based learning, MCMC sampling, cooperative learning, and progressive learning for unpaired multi-domain image-to-image translation in this chapter. Our method includes a multi-head energy-based model as a descriptor, capturing the multi-domain image distribution, and a diversified image-to-image translator for cross-domain one-to-many mapping. To train both the descriptor and translator, we

introduce a multi-domain MCMC teaching algorithm. Additionally, we propose progressive learning to enhance the scalability and efficiency. Experimental results demonstrate that our approach achieves comparable performance to adversarial learning frameworks and sets a new benchmark in energy-based image-to-image translation methods.

6.1.3 3D Teeth Reconstruction from a Single Panoramic Radiograph

We initialize the study of 3D teeth reconstruction of the whole cavity from a single panoramic radiograph. In order to solve the challenges posed by the high resolution of images and multi-object reconstruction, we propose *X2Teeth* to decompose the task into teeth localization and single tooth reconstruction. Our *X2Teeth* is compact and employs sampling-based training strategy, which enables the end-to-end optimization of the whole model. Our experiments qualitatively and quantitatively demonstrate that *X2Teeth* achieves accurate reconstruction with tooth details. Moreover, our method can also be promising for other multi-anatomy 3D reconstruction tasks.

6.1.4 3D Reconstruction from Single Image with Implicit Neural Representation

we propose a new method for reconstructing the 3D oral structure from projection information in panoramic X-ray imaging. We utilize an implicit representation model with multi-head architecture to accommodate the imaging process of PX and a dynamic sampling strategy to refine the reconstruction results. Unlike existing deep learning models like Oral-3D, our method does not require extensive patient data or dense annotations to reconstruct the complicated structure of oral cavity. Extensive experiments show that our model significantly outperforms state-of-the-art models both qualitatively and quantitatively with clear density details of teeth and the mandible in the reconstructed oral structure. Furthermore, the complexity analysis show that our method has great potential in clinical applications

with the low radiation and comparable reconstruction speed.

6.2 Conclusion and Future Work

Within this thesis, we present a comprehensive array of methodologies and strategies designed to advance the field of image-to-image translation, with a focus on both cross-domain and cross-dimension perspectives. Our extensive experimentation across the four chapters demonstrates the remarkable efficiency and effectiveness of our proposed approaches in successfully addressing the inherent challenges of image-to-image translation tasks.

Nonetheless, significant challenges persist. For instance, within the domain of multi-domain translation, the task of domain transfer, especially in the context of super-resolution, such as with 4K images, continues to pose formidable challenges in terms of achieving stability and efficiency with energy-based models. Furthermore, in the arena of cross-domain translation, bridging the performance gap arising from disparities between simulated and real images remains an ongoing area of concern. Moreover, the enticing prospect of applying reconstruction techniques to decode magnetic signals in MRI imaging represents uncharted territory in this field. These captivating opportunities for further exploration will be deferred to future research endeavors.

REFERENCES

[AEF12] Aly S Abdelrahim, Moumen T El-Melegy, and Aly A Farag. "Realistic 3d reconstruction of the human teeth using shape from shading with shape priors." In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 64–69. IEEE, 2012.

[AFS14] Aly S Abdelrehim, Aly A Farag, Ahmed M Shalaby, and Moumen T El-Melegy. "2D-PCA shape models: Application to 3D reconstruction of the human teeth from a single image." In *Medical Computer Vision. Large Data in Medical Imaging: Third International MICCAI Workshop, MCV 2013, Nagoya, Japan, September 26, 2013, Revised Selected Papers 3*, pp. 44–52. Springer, 2014.

[AMD20] Shahab Aslani, Vittorio Murino, Michael Dayan, Roger Tam, Diego Sona, and Ghassan Hamarneh. "Scanner invariant multiple sclerosis lesion segmentation from MRI." In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 781–785. IEEE, 2020.

[Arm06] Robert T Armstrong. "Acceptability of cone beam CT vs. multi-detector CT for 3D anatomic model construction." *Journal of Oral and Maxillofacial Surgery*, **64**(9):37, 2006.

[BCU21] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. "Rethinking the truly unsupervised image-to-image translation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14154–14163, 2021.

[BHF98] Stanley Braun, William P Hnat, Dana E Fender, and Harry L Legan. "The form of the human dental arch." *The Angle Orthodontist*, **68**(1):29–36, 1998.

[BMT21] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5855–5864, 2021.

[BOP04] Stéphanie Buchaillard, Sim Heng Ong, Yohan Payan, and Kelvin WC Foong. "Reconstruction of 3D tooth images." In *2004 International Conference on Image Processing, 2004. ICIP'04.*, volume 2, pp. 1077–1080. IEEE, 2004.

[Bro09] Sharon L Brooks. "CBCT dosimetry: orthodontic considerations." In *Seminars in Orthodontics*, volume 15, pp. 14–18. Elsevier, 2009.

[BSA18] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. "Demystifying mmd gans." *arXiv preprint arXiv:1801.01401*, 2018.

[BVK19] Hrvoje Bogunović, Freerk Venhuizen, Sophie Klimscha, Stefanos Apostolopoulos, Alireza Bab-Hadiashar, Ulas Bagci, Mirza Faisal Beg, Loza Bekalo, Qiang Chen, Carlos Ciller, et al. "RETOUCH: the retinal OCT fluid detection and segmentation benchmark and challenge." *IEEE transactions on medical imaging*, **38**(8):1858–1874, 2019.

[CFB22] Abril Corona-Figueroa, Jonathan Frawley, Sam Bond-Taylor, Sarath Bethapudi, Hubert PH Shum, and Chris G Willcocks. "Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray." In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3843–3848. IEEE, 2022.

[CFG15] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. "Shapenet: An information-rich 3d model repository." *arXiv preprint arXiv:1512.03012*, 2015.

[CKJ21] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. "Ilvr: Conditioning method for denoising diffusion probabilistic models." *arXiv preprint arXiv:2108.02938*, 2021.

[CLW19] Zhiming Cui, Changjian Li, and Wenping Wang. "ToothNet: automatic tooth instance segmentation and identification from cone beam CT images." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6368–6377, 2019.

[CMY20] Pietro Antonio Cicalese, Aryan Mobiny, Pengyu Yuan, Jan Becker, Chandra Mohan, and Hien Van Nguyen. "StyPath: Style-Transfer Data Augmentation for Robust Histology Image Classification." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 351–361. Springer, 2020.

[CUY20] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. "Stargan v2: Diverse image synthesis for multiple domains." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8188–8197, 2020.

[CXG16] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction." In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pp. 628–644. Springer, 2016.

[CYL17] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. "Stylebank: An explicit representation for neural image style transfer." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1897–1906, 2017.

[CZP18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[DLT21] Yilun Du, Shuang Li, Joshua B. Tenenbaum, and Igor Mordatch. "Improved Contrastive Divergence Training of Energy-Based Models." In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021.

[DM19] Yilun Du and Igor Mordatch. "Implicit generation and generalization in energy-based models." In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, (NeurIPS)*, 2019.

[Dri03] Ronald G Driggers. *Encyclopedia of Optical Engineering: Las-Pho, pages 1025-2048*, volume 2. CRC press, 2003.

[GEB16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

[GKH21] Will Sussman Grathwohl, Jacob Jin Kelly, Milad Hashemi, Mohammad Norouzi, Kevin Swersky, and David Duvenaud. "No MCMC for me: Amortized sampling for fast and stable training of energy-based models." In *The ninth International Conference on Learning Representations, ICLR*, 2021.

[GLZ18] Ruiqi Gao, Yang Lu, Junpei Zhou, Song-Chun Zhu, and Ying Nian Wu. "Learning Generative ConvNets via Multi-Grid Modeling and Sampling." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9155–9164, 2018.

[GMK17] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram Van Ginneken, et al. "Transfer learning for domain adaptation in mri: Application in brain lesion segmentation." In *International conference on medical image computing and computer-assisted intervention*, pp. 516–524. Springer, 2017.

[GNK20] Ruiqi Gao, Erik Nijkamp, Diederik P Kingma, Zhen Xu, Andrew M Dai, and Ying Nian Wu. "Flow contrastive estimation of energy-based models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2020.

[GPM20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial networks." *Communications of the ACM*, **63**(11):139–144, 2020.

[GSP21] Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P. Kingma. "Learning Energy-Based Models by Diffusion Recovery Likelihood." In *The ninth International Conference on Learning Representations, ICLR*, 2021.

[HB17] Xun Huang and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

[Her09] Gabor T Herman. *Fundamentals of computerized tomography: image reconstruction from projections*. Springer Science & Business Media, 2009.

[Hin02] Geoffrey E Hinton. "Training products of experts by minimizing contrastive divergence." *Neural computation*, **14**(8):1771–1800, 2002.

[Hin12] Geoffrey E. Hinton. "A Practical Guide to Training Restricted Boltzmann Machines." In *Neural Networks: Tricks of the Trade - Second Edition*, pp. 599–619. Springer, 2012.

[HLB18] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. "Multimodal unsupervised image-to-image translation." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

[HMH22] Giwoong Han, Jinhong Min, and Sung Won Han. "EM-LAST: Effective Multidimensional Latent Space Transport for an Unpaired Image-to-Image Translation With an Energy-Based Model." *IEEE Access*, **10**:72839–72849, 2022.

[HNF19] Tian Han, Erik Nijkamp, Xiaolin Fang, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. "Divergence Triangle for Joint Training of Generator Model, Energy-Based Model, and Inferential Model." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8670–8679, 2019.

[HRR18] Phlipp Henzler, Volker Rasche, Timo Ropinski, and Tobias Ritschel. "Single-image tomography: 3D volumes from 2D cranial X-rays." In *Computer Graphics Forum*, volume 37, pp. 377–388. Wiley Online Library, 2018.

[HRU17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." *Advances in neural information processing systems*, **30**, 2017.

[JAF16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." In *European conference on computer vision*, pp. 694–711. Springer, 2016.

[KAL17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. "Progressive growing of gans for improved quality, stability, and variation." *arXiv preprint arXiv:1710.10196*, 2017.

[KB16] Taesup Kim and Yoshua Bengio. "Deep directed generative models with energy-based probability estimation." *arXiv preprint arXiv:1606.03439*, 2016.

[KCK21] Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. "Exploiting spatial dimensions of latent in gan for real-time image editing." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 852–861, 2021.

[KDK20] Yoni Kasten, Daniel Doktofsky, and Ilya Kovler. "End-to-end convolutional neural network for 3D reconstruction of knee bones from bi-planar X-ray images." In *Machine Learning for Medical Image Reconstruction: Third International Workshop, MLMIR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 3*, pp. 123–133. Springer, 2020.

[KLL19] Hangkee Kim, Kisuk Lee, Dongchun Lee, and Nakhoon Baek. "3D reconstruction of leg bones from X-ray images using CNN-based feature analysis." In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 669–672. IEEE, 2019.

[KOG19] Rithesh Kumar, Sherjil Ozair, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. "Maximum entropy generators for energy-based models." *arXiv preprint arXiv:1901.08508*, 2019.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*, **25**, 2012.

[LCH06] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. "A tutorial on energy-based learning." *Predicting structured data*, **1**(0), 2006.

[LSC21] Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, and Marco De Nadai. "Smoothing the disentangled latent style space for unsupervised image-to-image translation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10785–10794, 2021.

[LSL21] Hanbit Lee, Jinseok Seol, and Sang-goo Lee. "Contrastive learning for unsupervised image-to-image translation." *arXiv preprint arXiv:2105.03117*, 2021.

[LSY20] Yuan Liang, Weinan Song, Jiawei Yang, Liang Qiu, Kun Wang, and Lei He. "X2teeth: 3d teeth reconstruction from a single panoramic radiograph." In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd*

International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23, pp. 400–409. Springer, 2020.

[LTH18] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. "Diverse image-to-image translation via disentangled representations." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 35–51, 2018.

[MCR13] Laura Mazzotta, Mauro Cozzani, Armando Razionale, Sabrina Mutinelli, Attilio Castaldo, Armando Silvestrini-Biavati, et al. "From 2d to 3d: Construction of a 3d parametric model for detection of dental roots shape and position from a panoramic radiograph—a preliminary report." *International journal of dentistry*, **2013**, 2013.

[MES22] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. "Instant neural graphics primitives with a multiresolution hash encoding." *ACM Transactions on Graphics (ToG)*, **41**(4):1–15, 2022.

[MG19] Agnieszka Mikołajczyk and Michał Grochowski. "Style transfer-based image synthesis as an efficient regularization technique in deep learning." In *2019 24th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pp. 42–47. IEEE, 2019.

[MHS21] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. "Sdedit: Guided image synthesis and editing with stochastic differential equations." *arXiv preprint arXiv:2108.01073*, 2021.

[MJG19] Chunwei Ma, Zhanghexuan Ji, and Mingchen Gao. "Neural style transfer improves 3D cardiovascular MR image segmentation on inconsistent data." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 128–136. Springer, 2019.

[MLT19] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. "Mode seeking generative adversarial networks for diverse image synthesis." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1429–1437, 2019.

[MST21] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM*, **65**(1):99–106, 2021.

[NHS01] Hassan Noroozi, Tahereh Hosseinzadeh Nik, and Reza Saeeda. "The dental arch form revisited." *The Angle Orthodontist*, **71**(5):386–389, 2001.

[NHZ19] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. "Learning non-convergent non-persistent short-run MCMC toward energy-based model." *Advances in Neural Information Processing Systems*, **32**, 2019.

[PCP21] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. "D-nerf: Neural radiance fields for dynamic scenes." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10318–10327, 2021.

[PEZ20] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. "Contrastive learning for unpaired image-to-image translation." In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 319–345. Springer, 2020.

[PSB21] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. "Nerfies: Deformable neural radiance fields." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5865–5874, 2021.

[PZW20] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. "Swapping autoencoder for deep image manipulation." *Advances in Neural Information Processing Systems*, **33**:7198–7211, 2020.

[RBA19] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. "On the spectral bias of neural networks." In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

[RDF21] Mengwei Ren, Neel Dey, James Fishbaugh, and Guido Gerig. "Segmentation-Renormalized Deep Feature Modulation for Unpaired Image Harmonization." *IEEE Transactions on Medical Imaging*, **40**(6):1519–1530, 2021.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

[RKB05] Alireza Rahimi, Ludger Keilig, G Bendels, Reinhard Klein, Thorsten M Buzug, Iman Abdelgader, Marcus Abboud, and Christoph Bourauel. "3D Reconstruction of dental specimens from 2D histological images and μCT-Scans." *Computer Methods in Biomechanics and Biomedical Engineering*, **8**:167–176, 2005.

[SLY21] Weinan Song, Yuan Liang, Jiawei Yang, Kun Wang, and Lei He. "Oral-3d: Reconstructing the 3d structure of oral cavity from panoramic x-ray." In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 566–573, 2021.

[SS06] Gwen RJ Swennen and Filip Schutyser. "Three-dimensional cephalometry: spiral multi-slice vs cone-beam computed tomography." *American Journal of Orthodontics and Dentofacial Orthopedics*, **130**(3):410–416, 2006.

[SWZ18] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. "Pix3d: Dataset and methods for single-image 3d shape modeling." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2974–2983, 2018.

[SZ14] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*, 2014.

[TDB17] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. "Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs." In *Proceedings of the IEEE international conference on computer vision*, pp. 2088–2096, 2017.

[TRR19] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. "What do single-view 3d reconstruction networks learn?" In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3405–3414, 2019.

[TSM20] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. "Fourier features let networks learn high frequency functions in low dimensional domains." *Advances in Neural Information Processing Systems*, **33**:7537–7547, 2020.

[VSP17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems*, **30**, 2017.

[WBS04] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. "Image quality assessment: from error visibility to structural similarity." *IEEE transactions on image processing*, **13**(4):600–612, 2004.

[WSC20] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. "Deep high-resolution representation learning for visual recognition." *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[XKK21] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. "VAEBM: A symbiosis between variational autoencoders and energy-based models." In *The ninth International Conference on Learning Representations, ICLR*, 2021.

[XLG18] Jianwen Xie, Yang Lu, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. "Cooperative training of descriptor and generator networks." *IEEE transactions on pattern analysis and machine intelligence*, **42**:27–45, 2018.

[XLZ16] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. "A Theory of Generative ConvNet." In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *International Conference on Machine Learning (ICML)*, 2016.

[XZF21] Jianwen Xie, Zilong Zheng, Xiaolin Fang, Song-Chun Zhu, and Ying Nian Wu. "Learning cycle-consistent cooperative networks via alternating MCMC teaching for unsupervised cross-domain translation." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10430–10440, 2021.

[XZF22] Jianwen Xie, Zilong Zheng, Xiaolin Fang, Song-Chun Zhu, and Ying Nian Wu. "Cooperative Training of Fast Thinking Initializer and Slow Thinking Solver for Conditional Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, **44**(8):3957–3973, 2022.

[XZL21] Jianwen Xie, Zilong Zheng, and Ping Li. "Learning Energy-Based Model with Variational Auto-Encoder as Amortized Sampler." In *Thirty-Fifth AAAI Conference on Artificial Intelligence, (AAAI))*, pp. 10441–10451, 2021.

[XZL22] Jianwen Xie, Yaxuan Zhu, Jun Li, and Ping Li. "A Tale of Two Flows: Cooperative Learning of Langevin Flow and Normalizing Flow Toward Energy-Based Model." In *International Conference on Learning Representations (ICLR)*, 2022.

[YGM19] Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, and Yefeng Zheng. "X2CT-GAN: reconstructing CT from biplanar X-rays with generative adversarial networks." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10619–10628, 2019.

[YYH19] Zhaoqiang Yun, Shuo Yang, Erliang Huang, Lei Zhao, Wei Yang, and Qianjin Feng. "Automatic reconstruction method for high-contrast panoramic image from dental cone-beam CT data." *Computer methods and programs in biomedicine*, **175**:205–214, 2019.

[ZBL22] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. "Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations." *Advances in Neural Information Processing Systems*, **35**:3609–3623, 2022.

[ZDW23] Chulong Zhang, Jingjing Dai, Tangsheng Wang, Xuan Liu, Yinping Chan, Lin Liu, Wenfeng He, Yaoqin Xie, and Xiaokun Liang. "XTransCT: Ultra-Fast Volumetric CT Reconstruction using Two Orthogonal X-Ray Projections via a Transformer Network." *arXiv preprint arXiv:2305.19621*, 2023.

[ZIE18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. "The unreasonable effectiveness of deep features as a perceptual metric." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[ZPI17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[ZWM98] Song Chun Zhu, Yingnian Wu, and David Mumford. "Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling." *International Journal of Computer Vision*, **27**:107–126, 1998.

[ZXL21] Yang Zhao, Jianwen Xie, and Ping Li. "Learning energy-based generative models via coarse-to-fine expanding and sampling." In *International Conference on Learning Representations*, 2021.

[ZXL23] Yang Zhao, Jianwen Xie, and Ping Li. "CoopInit: Initializing Generative Adversarial Networks via Cooperative Learning." *arXiv preprint arXiv:2303.11649*, 2023.

[ZXZ22] Jing Zhang, Jianwen Xie, Zilong Zheng, and Nick Barnes. "Energy-based generative cooperative saliency prediction." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3280–3290, 2022.

[ZZL22] Ruyi Zha, Yanhao Zhang, and Hongdong Li. "Naf: Neural attenuation fields for sparse-view cbct reconstruction." In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pp. 442–452. Springer, 2022.