Confidential thesis □

浙江大学

Zhejiang University-University of Illinois Urbana-Champaign Institute

Undergraduate Thesis (Design)

Thesis Title	Data-driven Sensing with Digital Twin: An Inception
Student Name	Bingjun Guo
Student ID	3210115445
Supervisor	Shurun Tan
Class of year	2025
Major	Electrical and Computer Engineering
Submission date	June 1st, 2025

Acknowledgements

I anticipated difficulty of the journey before I made up my mind to enter this international dual-degree program, but clearly not enough. These four year has exhibited more suffering than I've ever experienced in the past 18 years before college. I want to first give the gratitude to myself for the perfect preservation of enthusiasm in exploring and discovering into the unknown, as well as the courage and toughness to undertake all the challenges in past and future while rarely compromised.

Rather fortunately, I also underestimated how much I would grow as well as how much warm-hearted help I would receive. Specifically for this thesis, I want to thank Prof. Shurun Tan for his valuable guidance along with all the time and patience he put on me. I would like to thank Prof. Liangjing Yang and Prof. Gaoang Wang for their guidance on my another senior design during this tough transformation of my domain of interest. Meanwhile, I want to thank Prof. Wang for his suggestions regarding the last application cycle (although I didn't follow and went off the road) and all the helpful service he has provided as a responsible class mentor.

During my junior year in UIUC, I spent a wonderful year interning in the Data Mining Group, under the guidance of Prof. Jiawei Han and Dr. Bowen Jin, when my understanding of research and capabilities were dramatically improved. I cannot thank them too much for their help and trusts when I was just a beginner in research. Right before my junior year, I spent a month in the Center for Data Science, ZJU, where I received guidance from Prof. Xiaoye Miao, Dr. YangYang Wu, and Dr. Wei Ni. I would like to present my gratitude to them for all their patience in guiding me down the path of formal research. I would also like to thank Prof. Hongwei Wang for offering me a glimpse of what research was like when I didn't know a thing about it.

This journey would be incomplete without all the companionship and inspiration from my schoolmates. I would like to thank Qi Long, Honghui Chenyang, Yuxuan Li, Yuxi

Chen, Xiran Wu, Haotian Zhang, and so much more, for the unforgettable moments of fun and teamwork made possible by them. I also want to thank Dr. Yuxuan Jiang, Haotian Zhang, and all other members in VoICE A'Cappella for the excellent beginning of my college life that was filled up with colorful memories and melodies.

Lastly, I want to give my sincere gratitude to my family for their unconditional support in all aspects, including tolerating my impossible dreams that literally eliminate the time we could've spent together and sometimes hurt. My life in ZJUI began with an audition for the choir and VoICE A'Cappella, singing *Audition* in *La La Land*. Now it ends with me standing on the graduation stage as well as the motto,

"Here's to the fools who dream."

Regarding those who cares about me, I'm sometimes truly sorry about being such a fool, although in private I often feel relieved that this fact cannot be changed — and hopefully will never been.

Abstract

Remote sensing has long served as a scalable tool to observe forests, trees, and the broader ecological systems they exist in. While geometric analysis—such as classification, segmentation, and density estimation—has dominated the research landscape, much of the valuable information, such as internal moisture and other non-visual attributes, still remains underexplored. To go beyond geometry, microwave sensing has shown promise, especially when paired with simulation. However, simulations rely heavily on accurate 3D models, which are hard to come by—particularly for trees, whose structure is notoriously complex and messy.

This thesis proposes a method to infer 3D features of trees from single 2D images that leverages recent developments in generative models and realistic large-scale datasets, aiming for both realism and scalability. By combining the shape-awareness of volume generation with conventional growing-based modeling techniques, the approach leads to an affordable way to acquire 3D assets suitable for downstream tasks like physical simulation in remote sensing. By enabling efficient, large-scale 3D model generation, this work ultimately envisions a new paradigm for remote sensing that goes, quite literally, beneath the surface.

Additionally, this thesis reflects on a broader trend in role of data, especially for generative tasks, has started to overshadow traditional theory-heavy methods. Drawing on empirical experiments and observations, it argues that data—more than ever—is central to progress in both machine learning and scientific modeling under the current circumstance.

Contents

Chapter 1. Introduction

1.1 To Learn or Not to Learn

How much intelligence of intelligence is learnt?

The revolution of learning has swept across a wide range of domains including computer vision[1], robotics[2], biology[3], medical[4], finance[5], and way more. Taking one of the most representative field where the changes happened, computer vision, as an example. Conventionally, in order to extract meaningful pattern from images, filtering kernels were hard-implemented[6] and critical features were hand-selected by domain experts. However, nowadays the case has been a lot rarer for people to implement kernels or determine features themselves. Instead, convolutional kernels and feature weights are learnt from specific sets of data[7], or even an open set of data[8], while outperforming the expert models in numerous tasks. From the perspective of engineering, such approaches are often referred as “data-driven”, whose designs are supposed to be adaptive, flexible, and somehow unexpectedly powerful.

By way of comparison, data-driven approaches have yet to exert as profound an impact on remote sensing as they have on computer vision. In fact, the concept of remote sensing is closely related to computer vision, both of which attempt to “see” and “understand” things from highly unstructured signals that are either analog or digital. However, despite the remarkable advancement in computer vision brought by learning, model-based approaches are still dominating the mainstream in the field of remote sensing. What are the underlying causes of this phenomenon? Is there anything that we can do about it? Meanwhile, is this phenomenon present in other disciplines as well? How can we contribute in a more foundational manner through this lens of remote sensing?

1.2 Briefly on Remote Sensing

1.2.1 Sensing Modalities

We as human beings are known for now to rely on electromagnetic signals merely in the range of so called “visible light” frequency, which is, in fact, rather narrow among the whole spectrum known to us that exists in nature. However, modern technologies has employed the signals spanning a spectrum that is seriously wider, and each types of them demonstrates distinct characteristics while leading to distinct capabilities of sensing methods. The following are several most common and representative categories of them.

1. Visible light (380-780nm) is widely used in environmental monitoring via passive imaging, offering high-resolution visuals rich in shape, color, and texture. However, it lacks depth and non-visual data like moisture or density, limiting its scope.

2. LiDAR (Light Detection and Ranging) (905-1550nm) actively emits laser pulses to map environments in 3D with high precision[9]. While being powerful, its performance drops under poor weather conditions like fog or rain, and it often comes with high operational costs.

3. Microwave-based radar systems (mms to ms) operate with longer wavelengths, allowing them to penetrate clouds and weather. While ideal for all-weather, long-range sensing, they offer lower precision and typically require large antennas—an issue mitigated by Synthetic Aperture Radar (SAR)[10], which simulates larger antennas through motion.

1.2.2 Enhanced Sensing with Microwave

As discussed in the previous section, microwave benefits from it’s advantage in penetrability as a mean of remote sensing, which is a consequence of its distinct electromagnetic properties from visible light, thanks to its long wavelength. However, such properties also endow it with possibilities to carry back richer information than normal modalities such as visible light do. For example, visible light is majorly sensitive to colors and geometrical properties, while microwave is additionally sensitive to other surface parameters such as roughness or texture and carries back information associated with such parameters. Extracting this portion of information inherent in microwave sensing that surpasses the boundaries of geometry as well as classical forms of perception is highly valuable while challenging.

One of the promising approaches to extract super-geometrical information, i.e. surface parameters, is through radar simulation[11]. Given a 3D mesh of the detected object, SAR parameters, and set surface parameters of the object to specific values (initially randomized), propagation of microwave is rendered with differential ray tracing[12], inspired from computer graphics, and eventually leads to a synthetic image via simulated typical SAR imaging process. This image synthesized from simulation is then compared with actual SAR image of the target in the physical world, and the difference between the simulated SAR image and the actual SAR image is considered as the loss to optimize for the input variable, surface parameters, since the rendering process is differentiable. During the process, as the surface parameters are adjusted, the simulated SAR image will gradually converge to the actual SAR image. When the difference is under a appropriate tolerance, the optimized surface parameters of the target in simulation would be close enough to those of the actual target in the physical world.

1.3 Formation of the Problem and Objectives

1.3.1 Remote Sensing for Trees - Beyond Geometry

Remote sensing for trees has kept attracting attention for its ability to scalably monitor forest health, biodiversity, environmental change, and more. Most widely studied topics includes tree classification[13], tree coverage and density detection[14], and segmentation of tree structures[15], which mainly focus on geometric or spatial features of trees, leaving out the valuable information beyond geometry, such as soil moisture content and water content of vegetation.

This issue can be resolved with the property of microwave sensing mentioned in **1.2.2** as well as optimizing in simulation[11]. However, in acquisition of target surface parameters, the approach demands an accurate description of the target surface geometry, normally in the form of a 3D mesh, and acquisition of the 3D geometry can be just as challenging.

1.3.2 Background in Tree Modeling

3D modeling of trees has long been a prominent problem in the field of computer graphics[16], for its potential application in a wide range as well as the particular challenges it brought due to the especially complicated geometric structure of trees.

For computational approaches for tree modeling, a persistent conflict exists between flexibility and scalability. If in demand of tree models that are highly flexible and customized, e.g. close to certain realistic environments, tree models are normally built with fine-grained manual operation[17], which results in tremendously high labour costs to model a forest accurately.

An extensively studied solution to this issue in remote sensing is taking advantage of LiDAR point clouds[18][19]. Meanwhile, such approaches are restricted by the high expense of LiDAR sensing or deployment and potentially high operational costs. In addition, as most of the approaches are computational-based, they are usually less robust to measurement errors.

Benefiting from recent success in machine learning and computer vision, especially regarding generative models[20][21] (refer to **2.3** for detailed explanation), more and more studies focus on the tree geometry reconstruction from single images[22][23][24]

[25][26], which aim to both align with realistic features from the input tree images and preserve the completeness and details of tree structures well. In order to achieve the both goal, a number of them, which demonstrate ideal scalability[22][24][23], adopt the following procedure:

1. Generate a volume that aligns with geometry of the tree in image
2. Guide existing computational tree modeling methods (growing methods) with the volume boundaries

Empowered with capabilities of modern generative models to generate vague contents with remarkable alignment as well as the detail modeling process of conventional growing models, such approaches are able to generate realistic-aligned and detailed 3D models of trees at exceptionally low costs.

1.3.3 Objectives

This thesis is firstly dedicated to facilitate microwave sensing in going beyond geometric information, where the major challenge in the current scenario lies in the difficulty in acquiring the 3D assets of targets, that is, 3D models of trees, in cases of remote sensing for trees, which are specifically tricky to built. Hence, this thesis will propose a method to infer the 3D information from single images, which is expected to be both efficient and scalable.

Meanwhile, as the method is data-driven, this thesis would also demonstrate empirical studies regarding how data is significant to our generation as well as the broader set of tasks share certain similarities (see **2.3** for more details). Eventually, it will be concluded that how this approach that enables efficiently collection of realistic 3D assets could not only serve for current remote sensing methods with simulation[11], but also lead to a new paradigm for remote sensing with simulation.

1.4 Overview of the Thesis

Chapter 1, the introduction, is expected to demonstrate both problems at a broader scale and the specific problems whose solution is offered in this thesis, that is, data-driven challenges for remote sensing, and reconstruction of tree models, correspondingly.

Chapter 2 will defend for the effectiveness of novelty for this thesis – the role and significance of data to learn for especially the “generative” tasks. Ever since large language models started dominating the field of NLP, as well as a lot others, and was crowned with the name ”foundation models”, there has been voices (mainly around Silicon Valley) advocating that the center of research should be shifted in a more empirical manner, in contrast to the common and conventional views in computer science that the science is only called science with complicated math. Such voices were brought by the observation that the dramatic power of large language models turned out to have little to do with formal mathematical analysis but mostly empirical studies[27][28], and they believe that the next quantum leap for AI is likely to happen in exactly the same way. In fact, a portion of the voices, which are strong, and even wilder, that the view of *data-centric* has been increasingly supported over the years. While this thesis does not provide a conclusive evaluation of whether such approaches would lead to success, a number of empirical results on the basis of several experiments conducted during the writer’s undergraduate study as well as the corresponding analysis will be present in Chapter 2, listing evidences for the point of view that whether data should be centric to the future study or not, it’s indeed playing a pivotal role under the current circumstances.

Chapter 3 will demonstrate the method proposed to address the problem of 3D tree reconstruction – learning 3D priors with synthesized specialized data, at service of the final goal – sensing beyond geometry, as illustrated in Chapter 1. Chapter 4 will give a summary on both the philosophical and methodological discussions in this thesis. A conclusion of how far we still are from the final goal, as well as further ambitions that could reshape the landscape of remote sensing study, as how fields of computer vision and natural language processing have evolved over the years, will also be included, in response to the questions raised in **1.1**.

Chapter 2. Role of Data for Generation

As introduced in **1.4**, data has been considered as one of the key component (sometimes the only key component) to modern AI systems, since the emergence of large foundation models[27][29]. Although it remains debatable that whether the models relying on data so heavily is a good thing or not, existence of the phenomenon has been widely validated under the current circumstances. Several basic but novel formalizations and empirical results further supporting the phenomenon will be present in this Chapter.

2.1 Data-driven and Model-based

Before we go further, we would establish a comprehensive and novel formalization for the problem in a broad scope.

All of the problems concerned in this thesis, so are the most processes happening in the nature, can be described in the following form: a model $\mathscr{P}$, the input form I, and the output form O such that:

$$\mathscr{P}(I) = O$$

With an close-form and analyzable model built, e.g. a microwave propagation model, the process of mapping I to O through $\mathscr{P}$ is usually referred as *forward process*, and the process of somehow mapping O back to I is referred as the *inverse process*, respectively. We will retain this notation for input and output in the following section. Implementing such a model first, without regards to I or O, is normally referred as *model-based* approach.

In the domain of engineering, the concept of *data-driven* was initially dependent to *model-based*, distinguished in features of space for I. The space of I for a model that is not data-driven usually involves dimensions that are in discrete and limited distribution, i.e. requires a selection from a fixed set. For example, taking a smart phone as an agent model $\mathscr{A}$ that takes a touch on an application as input and returns a series of actions in form of display as output. If applications from external sources can be installed on the phone so that not only built-in apps are available, or it's even possible to design apps on the phone, then the phone is a data-driven model of agent. Formally, for such limited subspace of input $S = \{a, b, c, \dots\} \subset \mathbb{S}$, a data-driven design has the subspace defined by users and expands $I \in S$ into $I \in \mathbb{S}$.

However, nowadays, when we're talking about data-driven approaches in the sense of learning, we're actually being more ambitious. Instead of domain of I, what's expand comes into domain of $\mathscr{P}$, space for the model. Instead of relying on a fixed model, we wish to have $\mathscr{P}$ defined by users according to specific needs (while for most of the time the real reason would be that the classical modeling process is way too complicated). Models are usually on the basis of a meta-model and represented by parameters, and the space for $\mathscr{P}$ is expanded from a singleton space into the parameter space. For an AI system, such space for $\mathscr{P}$ can be considered as a number of models pretrained on different fields of data, i.e. a discrete set, and such approach is referred as Mix-of-Experts (MoE)[30].

2.2 A Canonical View of Generative Models

Based on the previous section, this section will establish a novel formalization of generative models, in emphasis of the importance of appropriate data (prior) to the so-called generative process. Generative models are terms corresponding to classification models and regression models. A classification model normally maps an input with a continuous distribution to a number of classes under a discrete distribution, mathematically:

$$\mathscr{C} : \mathbb{R}^n \mapsto \{\texttt{categories}\}$$

while a regression model maps such a input to a continuous distribution:

$$\mathscr{R} : \mathbb{R}^n \mapsto \mathbb{R}^m$$

Somehow $m = 1$ for a large portion of cases including house price predicting and sales revenue forecasting.

However, in the domain of machine learning, such functions are usually represented in terms of probability distributions, correspondingly:

$$\mathscr{C}(\cdot) = \arg\max_c P(c \mid \cdot) \in \{\texttt{categories}\}$$

$$\mathscr{R}(\cdot) = \arg\max_r P(r \mid \cdot) \in \mathbb{R}^m$$

in which $P(\cdot \mid \cdot)$ are conditioned discrete or continuous probability distributions.

It's worth noting that joint probability for multiple certain observations is **multiplication** of individual probabilities, but it's hard to optimize a multiplication for elements. This is why in actual optimization processes a *log* is usually seen before the probabilities, breaking the multiplication into **addition** that is friendly to optimization. For example, for independent variables $a, b, c, \cdots \in S$, we have:

$$\begin{aligned}
P(a,b,c,\dots) &= P(a)P(b)P(c)\cdots \\
&= \prod_{i \in S} P(i) \\
\log P(a,b,c,\dots) &= \log P(a) + \log P(b) + \log P(c) + \cdots \\
&= \sum_{i \in S} \log P(i)
\end{aligned}$$

Hence it can be seen that the mapping nature, or the predictive nature from a perspective of probability, of the classification or regression model, is possible and conventional to be described as a conditioned probability distribution, $P(O|I)$, with O as output and I

as input. What determines whether a model is a classification model or a regression model would be the discreteness of the targeted output form O.

This term leads to the conventional form of generative model, with which we are more interested in the distribution of the whole set of features, rather than inferring the rest given some of them. This view is under the consideration that both I and O are a subset of features of an entity, and eliminates the boundaries between what's given and what's desired. In other words, a generative model models the distribution of the whole set of features:

$$\begin{aligned} \text{discriminative model: } & P(O|I) \\ \text{generative model: } & P(I,O) \end{aligned}$$

(credits to ECE449, UIUC for this formalization)

This classification is both elegant and effective, especially for the models trained in a self-supervised manner[31], but probably a bit confusing under the more common circumstances nowadays. All those fancy AIGC products such as ChatGPT[28] and Stable Diffusion[29] claim themselves as generative models, but it doesn't make much sense with the definition of $P(I,O)$, whose results for generation should be an uncontrolled random set of (I,O) whose probability distribution is fixed.

Actually, it can be easier to understand how we are referring a model as generative with what we mentioned in the last chapter, forward and inverse process. Taking modeling the relationship between an image $I \in \mathbb{R}^{256,256}$ and a caption $C \in S\ s.t.\ |S| = 10$ as an example. It's clearly that the distribution of the former is much more complex to model than than the latter, since the former lies in a continuum with pretty high dimension while the latter is a selection from ten possible ones. Modeling the distribution of the latter variable can seem far easier than modeling the former one, and actually it was indeed the beginning of all. A long time ago, obtaining such a mapping attracts nearly all the attention: $\mathscr{C}(I) = C$, or in other words, $P(C|I)$, the discriminative model stated above. Recognizing hand-written numbers was a classical example. For this initial and more natural approach, let's refer it as the *forward process*.

Then what shall be the *inverse process*? Naturally, $\mathscr{C}'(C) = I$, i.e. $P(I|C)$. And this is actually what we usually refer as a generative model at these days, for example, generating an image from a caption. Although the caption used to generate images are usually not chosen from a fixed set nowadays, the distribution of the caption is actually still discrete, since a tokenizer would turn the input text into discrete tokens before guiding the generating process with it. Hence, freedom of the input form is still weaker than that of the output form, and this is why we need priors for such "generation" process.

In conclusion, we have demonstrated a straightforward formalization for modern generative models as a inverse process of the classical discriminative models whose definition has maintained consistence. The latter one maps an input from a complicated space into a less flexible output space, while the former one maps an input from the less flexible space to a more complicated space and thus has to take priors in consideration. Both processes are random.

By the way, a more appropriate way to give a definition could be giving up all the names and focusing on the space properties, which is not the focus of this thesis and seems requiring a portion of horrifying abstract math.

2.3 Data for More Effective Representations

As claimed in **2.1**, a data-driven model $\mathscr{P}$ involves a meta-model and a parameter space. Feedforward neural networks (FNNs)[32] has dominated the choice for meta-models nowadays.

2.3.1 Briefly on Feedforward Neural Network

A feedforward neural network consists of a series of layers through which information flows unidirectionally—from input to output—without cycles or feedback. Each layer performs a weighted transformation followed by a nonlinear activation. Formally, the transformation at layer l can be expressed as:

$$\mathbf{a}^{(l)} = f^{(l)}\left(\mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}\right)$$

where $\mathbf{a}^{(l-1)}$ is the activation from the previous layer, $\mathbf{W}^{(l)}$ is the weight matrix at layer l, $\mathbf{b}^{(l)}$ is the bias vector, and $f^{(l)}$ is a nonlinear function (e.g., ReLU, sigmoid, tanh). The final output of the network $\hat{\mathbf{y}}$ is given by:

$$\hat{\mathbf{y}} = \mathbf{a}^{(L)}$$

where L denotes the last layer. Training such a network involves minimizing a loss function $\mathscr{L}(\hat{\mathbf{y}}, \mathbf{y})$, typically using backpropagation and gradient descent:

$$\theta \leftarrow \theta - \eta \frac{\partial \mathscr{L}}{\partial \theta}$$

with $\theta \in \{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}$ and η denoting the learning rate.

2.3.2 Representation Entanglement

A tricky experiment is conducted for this section in order to demonstrate the limitation of FNNs (credits to PSYC489, UIUC) and how involving appropriate data can resolve the issue.

With a FNN with exactly the same numbers of dimensions for input and output and a training set S, we conduct the following algorithm:

Basically, this procedure train the FNN to output the input for each element in the

Algorithm 1 Train FNN to Output Identity on Set S

```
Require: Training set $S = \{s_1, s_2, \ldots, s_n\}$, Feedforward Neural Network (FNN)
while testflag do
    for each $s \in S$ do
        $\hat{s}$ ←FNN(s)
        Backpropagate error
    end for

    testflag ← False

    for each $s \in S$ do
        $\hat{s}$ ←FNN(s)
        if $\hat{s} \neq s$ then
            Train FNN with input $s$ and target output $s$
            testflag ← True
        end if
    end for
end while
```

training set. With the training set designed with some visible patterns, we can then conduct tests to validate if the FNN successfully captures the pattern inherent in the training set.

Specifically in the experiment, a feedforward net with 8 nodes in the input layer, 8 nodes in the intermediate layer, and 8 nodes in the output layer was utilized, and two rounds of training together with corresponding tests were conducted.

The first round:

training patterns: $[1,0,0,0,0,0,0,0]$, $[0,1,0,0,0,0,0,0]$, $[0,0,0,0,0,0,1,0]$

test patterns: all the eight-dimensional one-hot vectors

results: $[0,0,0,0,0,0,1,0]$ for all test patterns

The result show that the FNN fails to learn the most effective representation for the training patterns, i.e. simply output the only activated node in input.

The second round (reserving the trained weights in the first round):

extra training patterns: $[1,1,1,0,0,0,0,0]$, $[0,0,0,1,1,1,0,0]$, $[0,1,1,1,0,0,0,0]$

test patterns: $[0,0,1,1,0,0,0,0]$, $[1,1,0,0,0,0,1,0]$, $[0,0,1,0,0,0,0,0]$, training patterns in the first round

results: $[0,1,1,1,0,0,0,0]$, $[1,1,1,0,0,0,0,0]$, $[0,1,1,1,0,0,0,0]$, $[1,1,1,0,0,0,0,0]$, $[0,1,1,1,0,0,0,0]$, $[0,0,0,1,0,1,0,0]$

It's demonstrated by the result that diversedata with denser features benefits the model's understanding of the training data. Not only the test patterns closer to the extra training patterns leads to more reasonable results, but also formerly trained data was better understood by the model, in terms of general positions of the activated nodes.

2.4 Transformer and Data-centric

Transformer[33] is a encoder-decoder model that was originally designed for translation tasks, but was found with "emergent abilities" if scaled in large size and fed with great volumes of data[27], and such models were later known as "pretrained language models" or "large language models", leading to revolutions in countless fields. In fact, quality of the data fed has been proved as one of the keys to the performance of large language models[28]. This section will list some of the observations found by the writer in experiments which aligns with such common sense.

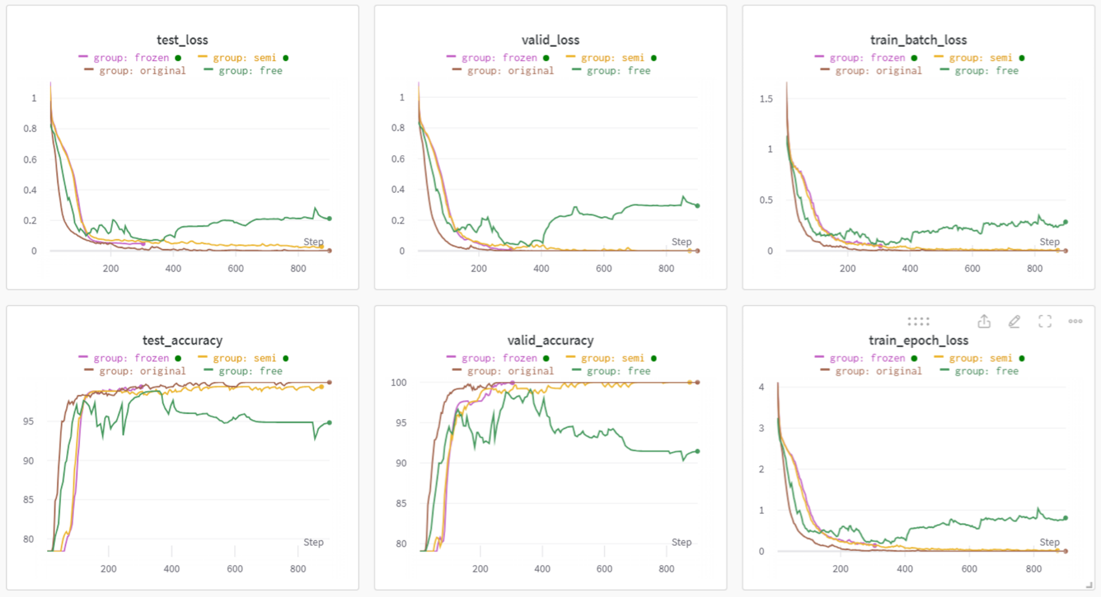

Figure 1: Training on a small amount of samples.

Figure 1 shows the training process of several models including a full scale BERT model[31] on a moderately small tabular dataset with 38 features and ≈ 200 samples. The pink, yellow, and brown lines represent training process of a specialized tabular model under three different frozen conditions, while the green lines represent training of the full scale pretrained BERT model. It's shown that for the small set of data, the pretrained language model converges slower and less effectively than the specialized model, and even appears to overfit after 400 steps of training, which is similar to the representation entanglement shown in **2.3**.

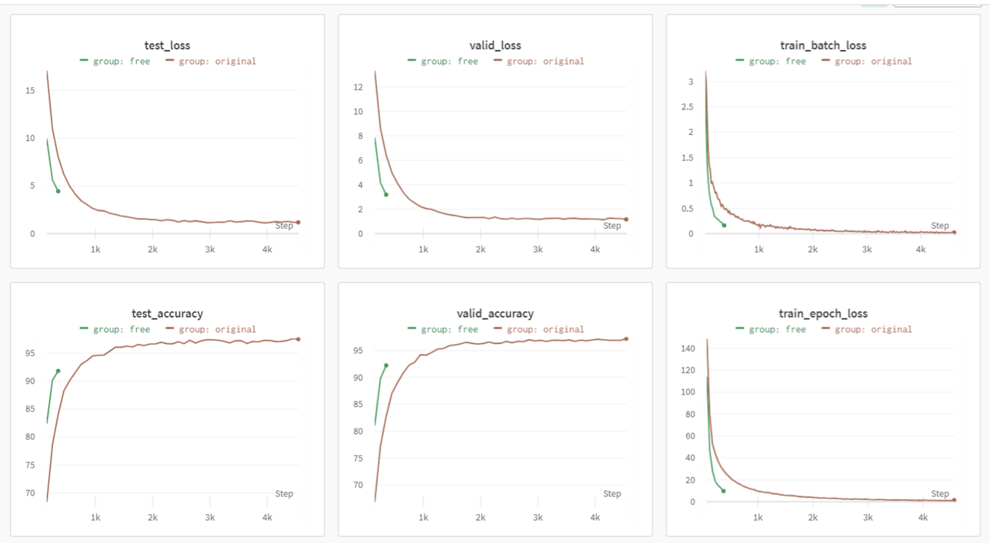

Figure 2: Training on a large amount of samples.

In the contrast, with brown lines still representing the specialized tabular model under the most powerful setting, and the green lines representing the pretrained language model, Figure 2 demonstrates training process of the two models on a dataset with a larger amount of samples (12881), during which the pretrained language model performed better than the last training and tend to out-perform the specialized one. However, due to computational constraints, training process of the two models ended up not completed.

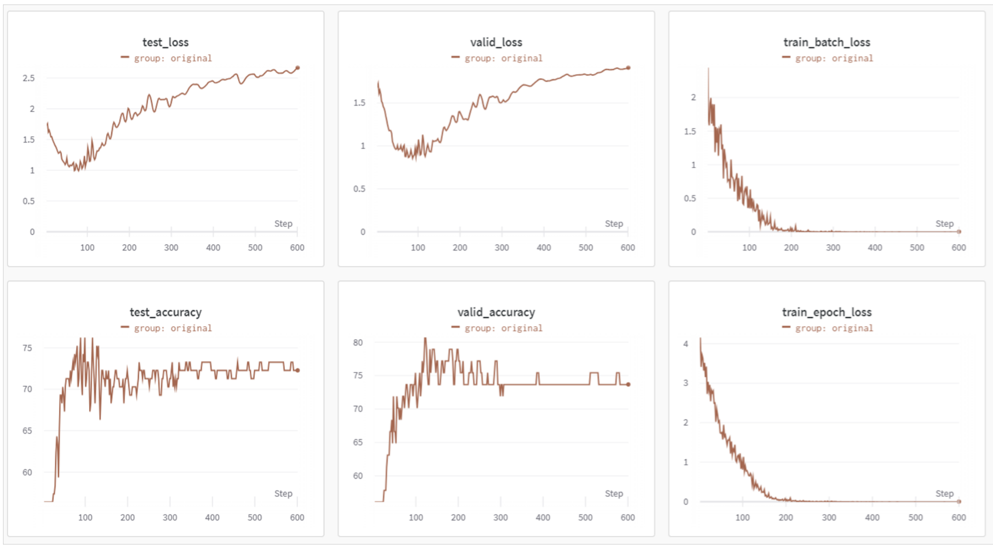

Figure 3: Training on samples with numerous features.

Previous experiments demonstrate how the amount of training data could affect the performance of powerful foundation models such as pretrained language model, while characteristics of training data can also affect the performance. Training of the specialized tabular model on another dataset with 279 features was conducted, shown in Figure 3. The model's accuracy on the test set converges at ≈ 72.5, which is critically lower. Due to computational constraints, again, training of the pretrained language model hasn't been conducted on the dataset, but is anticipated to excel according to the previous results.

Chapter 3. Learning 3D prior for Trees

3.1 Briefly on Neural Radiance Field

One of the most intuitive way to represent 3D structures is to directly encode the volume, which leads to issues including high storage costs and rendering costs. Instead of storing a 3D model explicitly, Neural Radiance Field (NeRF)[34] learns a continuous volumetric scene representation by mapping spatial coordinates and viewing directions to volume density and color using a neural network F_θ:

$$F_\theta : (\mathbf{x}, \mathbf{d}) \mapsto (\mathbf{c}, \sigma)$$

where the vector $\mathbf{x} \in \mathbb{R}^3$ denotes a 3D point in space, and $\mathbf{d} \in \mathbb{S}^2$ represents a viewing direction, constrained to lie on the unit sphere. The output of the function is a color vector $\mathbf{c} \in \mathbb{R}^3$ and a scalar volume density $\sigma \in \mathbb{R}_{\geq 0}$, which together describe how much light is emitted and absorbed at that point in the given direction. θ represents parameters of the neural network to be learnt.

To render an image, NeRF casts rays from the camera through the scene. Each ray is defined by:

$$\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}, \quad t \in [t_n, t_f]$$

where $\mathbf{o}$ is the ray origin, and $\mathbf{d}$ is the ray direction. The scalar variable t parameterizes the position along the ray between near and far bounds t_n and t_f.

The color observed along the ray, denoted as $\hat{C}(\mathbf{r})$, is computed by integrating the emitted color weighted by the volume density and accumulated transmittance along the ray:

$$\hat{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\, \sigma(\mathbf{r}(t))\, \mathbf{c}(\mathbf{r}(t), \mathbf{d})\, dt$$

The transmittance function $T(t)$ quantifies the probability that light travels from the camera origin to the point $\mathbf{r}(t)$ without any occlusion, which is given by:

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))\,ds\right)$$

Since this integral is intractable in closed form, it is approximated numerically by sampling N discrete points $\{t_i\}_{i=1}^{N}$ along the ray. The discrete approximation to the color is written as:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i\,(1-\exp(-\sigma_i\delta_i))\,\mathbf{c}_i$$

In this formulation, $\delta_i = t_{i+1} - t_i$ is the interval between adjacent samples, and the discrete transmittance term is:

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)$$

To train the NeRF model, the predicted color $\hat{C}(\mathbf{r})$ is compared against a ground-truth color $C(\mathbf{r})$ obtained from actual image data. The loss function is the mean squared error over all sampled rays:

$$\mathscr{L} = \sum_{\mathbf{r}\in\mathscr{R}} ||\hat{C}(\mathbf{r}) - C(\mathbf{r})||_2^2$$

where $\mathscr{R}$ is the set of all rays sampled from the training images.

3.2 Training Neural Radiance Field with Diffusion

Diffusion model[20][29] is a representative generative model evolving fast over the past few years, which models probability distribution of noises during gradual noising and denoising processes. In the training stage, noises are iteratively added to a natural image, which are expected to be learned by the model, typically a U-Net[35]. For inference, starting from pure noise, the model iteratively refines the sample through a learned reverse diffusion process, effectively generating realistic data step-by-step. As was claimed in **2.2**, generating 3D structures with optical methods would give rise to a strong demand for involving priors, as well as a powerful generative model. Diffusion model would be our first choice to learn 3D features of trees.

Let $C(\mathbf{r})$ denote the ground truth pixel color corresponding to the ray $\mathbf{r}$. The reconstruction loss would be:

$$\mathscr{L}_{\text{photometric}} = \sum_{\mathbf{r}\in\mathscr{R}} \left\|\hat{C}_\theta(\mathbf{r}) - C(\mathbf{r})\right\|_2^2$$

where $\mathscr{R}$ is the set of sampled camera rays from the training images.

In addition to this reconstruction loss, we include a diffusion-based regularization term using Score Distillation Sampling (SDS). This term guides the rendered images to conform to a prior distribution learned by a pretrained denoising diffusion model ε_ϕ. We begin by rendering an image $\hat{C}_\theta$ from the NeRF model and perturbing it with Gaussian noise to obtain a noisy version:

$$x_t = \sqrt{\bar{\alpha}_t}\hat{C}_\theta + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathscr{N}(0, I)$$

The diffusion model ε_ϕ predicts the noise component from the noisy input, timestep t, and optional conditioning signal y, which is an image combined with an angle shift in our case. The SDS loss is defined by the discrepancy between the predicted and actual noise:

$$\mathscr{L}_{\text{SDS}} = \left\|\varepsilon_\phi(x_t, t, y) - \varepsilon\right\|_2^2$$

To incorporate this loss into NeRF training, we use the chain rule to compute gradients with respect to the network parameters:

$$\frac{\partial \mathscr{L}_{\text{SDS}}}{\partial \theta} \propto \frac{\partial \hat{C}_\theta}{\partial \theta} \cdot \left(\varepsilon_\phi(x_t, t, y) - \varepsilon\right)$$

The total training loss is then a weighted combination of the photometric loss and the SDS loss:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{photometric}} + \lambda_{\text{SDS}} \cdot \mathscr{L}_{\text{SDS}}$$

where λ_{SDS} is a weighting coefficient that balances the contribution of the diffusion prior. This combined objective allows NeRF to benefit from the precise supervision provided by real images while still leveraging the strong visual priors captured by the diffusion model, leading to improved generalization and visual quality.

Models such as Zero 1-to-3[36] has acquired such architecture for general object 3D reconstruction from single images. Our project will apply the model to represent tree geometrical features and train the model specifically on tree data, aiming to solve the specialize tasks of 3D reconstruction of tree shapes in an end-to-end and straight-forward manner.

3.3 Implementation

As claimed in Chapter 2, ideal data available to our generative model should be:

1. diverse
2. large in amount
3. high in quality & close to real world distribution

A common way to acquire specific data in large amounts is to synthesize data through simulation. However, such approach typically tend to overrepresent certain species or structural patterns and may fail to capture the complex variability found in real environments. To address this issue, we take advantage of Tree-D Fusion[22], a large-scale dataset ($\approx$ 600000 samples) constructed from realistic street images that provide more informative distributions. Such dataset reconstructed from realistic images at a large scale is expected to sufficiently benefit our proposed generative model.

Specifically regarding training process, the model is deployed on 2 Nvidia RTX 3090 GPUs. A portion of Tree-D Fusion, consisting of 15000 samples collected from Montreal is adopted as the training set. During the training process, each sample in the training set, that is, 3D mesh of a tree, is rendered from 12 different views. Then, rendered images along with the corresponding view angles are fed to the model for training.

3.4 Test Results

Results of two test samples are demonstrated, both of which are collected near the east gate of the campus of ZJU, Haining. For each sample, original segmented pictures of the tree are first given. Then, picture of the same tree taken from a view angle with a difference in approximately -90 degree, along with the synthesized image by the 3D model from such a view angle are provided. Meanwhile, images of sample trees synthesized from the bottom view, which are nearly impossible to taken in reality, are also appended.

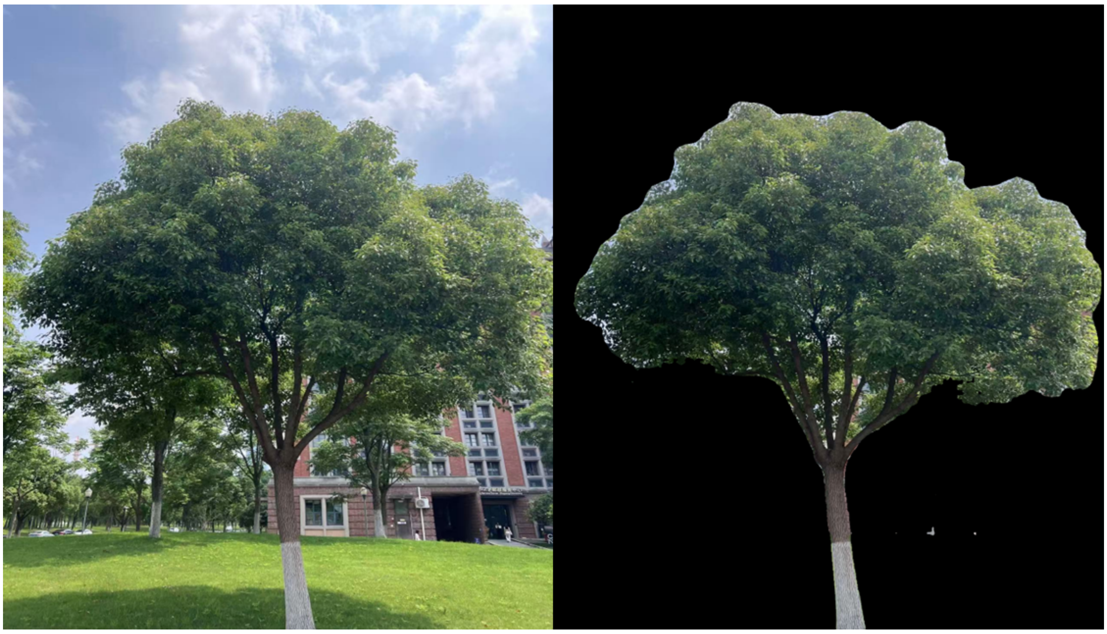

Figure 4: Sample 1

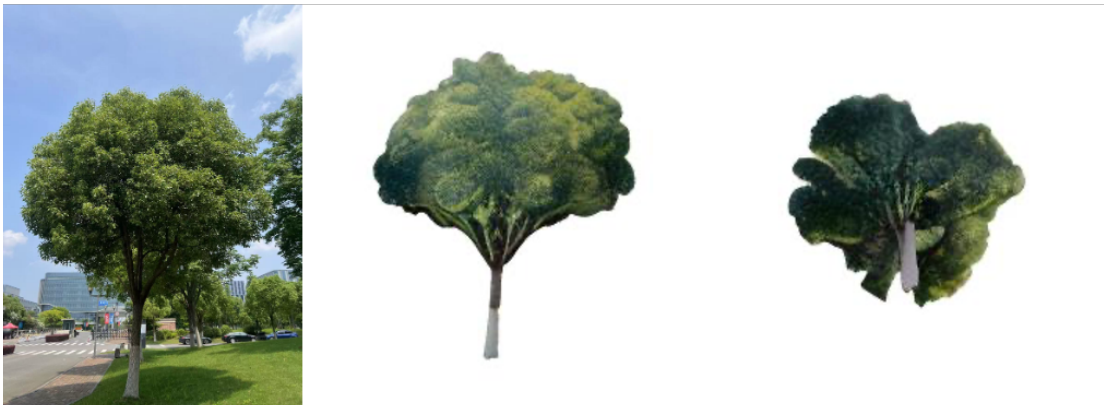

Figure 5: Rotated sample 1 and synthesized results

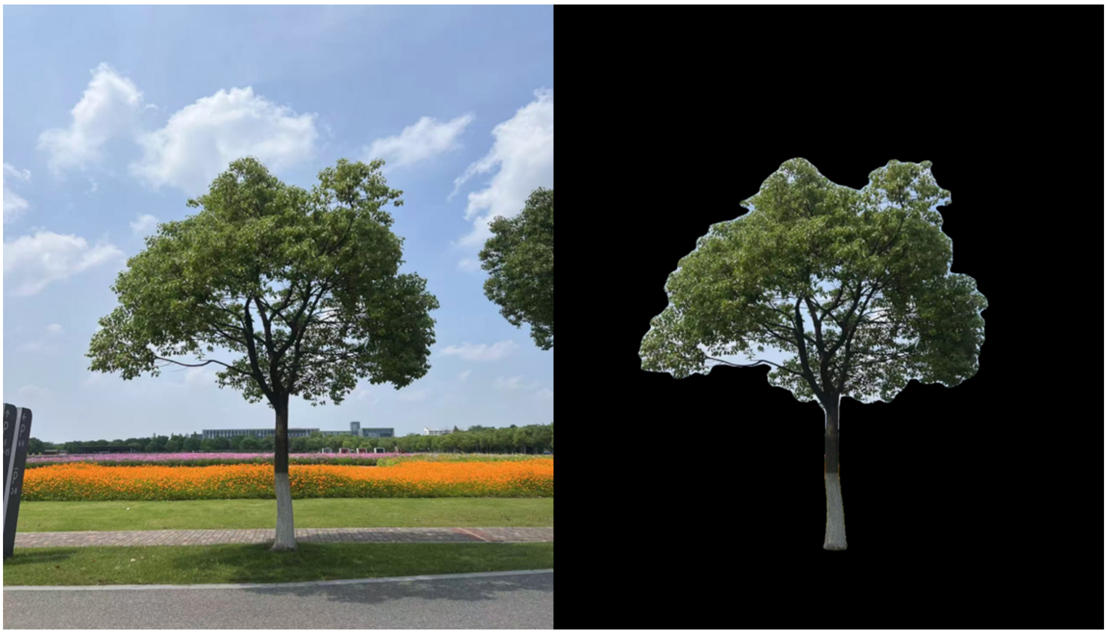

Figure 6: Sample 2

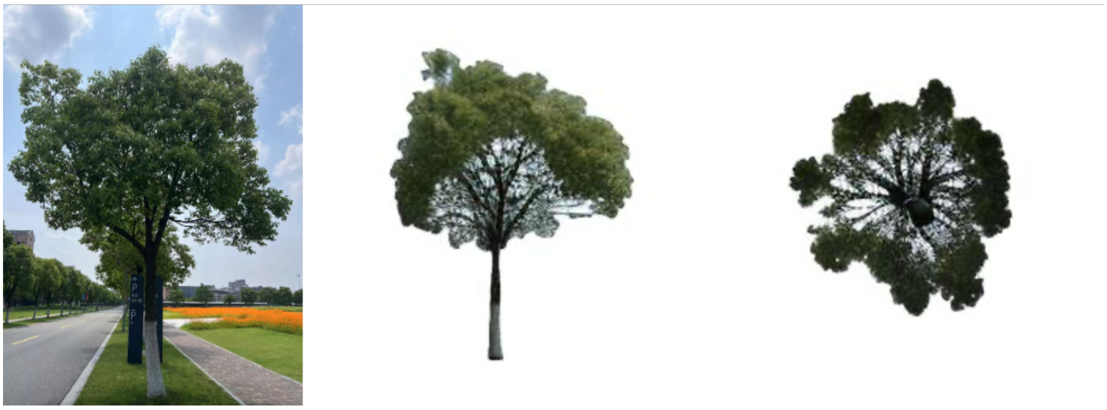

Figure 7: Rotated sample 2 and synthesized results

Chapter 4. Conclusion

4.1 Summary

This thesis aims to address a fundamental challenge in remote sensing of trees: the need to move beyond purely geometric representations towards a richer, more comprehensive understanding of tree properties. Traditional approaches to 3D tree modeling—whether through manual, labor-intensive techniques or costly LiDAR acquisition—face significant limitations in scalability and expense. Moreover, existing geometric models alone are insufficient to capture critical information such as vegetation moisture and soil parameters, which are essential for advanced sensing modalities like microwave remote sensing.

In response, a data-driven methodology is proposed, which learns 3D features of trees from single 2D images by leveraging the recent advances in generative models. This approach leads to a feasible pipeline that potentially integrates volumetric generation aligned with image data and conventional tree growth modeling and thus strikes a careful balance between realism, detail, and scalability. This method not only facilitates the generation of accurate 3D models at significantly reduced cost but also serves as a crucial enabler for simulation-based remote sensing methods that seek to extract electromagnetic parameters beyond geometry.

Beyond the immediate technical contribution, the thesis situates this work within the broader context of AI and data-centric scientific research. It underscores the critical role of data—both in quantity and quality—and the importance of model architectures such as Transformers and Diffusion models in achieving state-of-the-art generative capabilities. This shift from purely theoretical modeling to data-driven empirical approaches is, well, argued to be pivotal for future breakthroughs in remote sensing and beyond.

4.2 Future Work

Looking forward, several promising avenues for future research emerge from this work.

A primary direction is to extend the current framework to directly generate full tree structures conditioned on input images, thereby streamlining the pipeline and enhancing the fidelity and completeness of the 3D reconstructions. Insights from computational approaches for tree modeling such as graph structuring can be referred. Also, the volume extraction from NeRF representations as will as volume-guided tree modeling methods are yet to be validated specifically for the proposed framework.

One of the most compelling long-term visions inspired by this work is the development of comprehensive digital twins of forest ecosystems. These digital twins would act as highly detailed, dynamic simulations of real-world environments, integrating geometry, physical properties, and environmental processes at multiple scales. Such simulations could generate vast amounts of synthetic data to train and validate remote sensing models, facilitate “what-if” scenario analyses, and support decision-making in environmental monitoring and management. By effectively bridging the gap between data-driven AI and physics-based simulation, digital twins hold the promise to revolutionize remote sensing, enabling more accurate, scalable, and actionable environmental insights.

Moreover, while the current approach focuses primarily on inferring geometric properties, there is significant potential to broaden this scope by incorporating electromagnetic parameters directly into the modeling process. Advances in AIGC such as diffusion model and NeRF, coupled with improvements in computer graphics and radar simulation, could enable the direct inference of physical properties such as moisture content and soil characteristics, provided that sufficient synthesized training data can be produced.

It’s worth noting that this thesis has claimed a great significance on role of data under the current development of AI. However, it proposes here to maintain a critical perspective regarding the reliance on data. While data-driven approaches have demonstrated extraordinary potential, they should not be regarded as an end in themselves. At the end of the day, we want models to be “data-driven”, but probably not “data-dependant”. Robust learning-based remote sensing systems, and so are all the other intelligent systems, will likely require methods that balance data-driven inference with principled modeling, enabling models to extrapolate beyond existing data and explore novel or poorly understood scenarios. From the writer’s perspective, developing hierarchical or abstract representations could be key to achieving this, allowing models to generalize and reason more effectively, as the writer has been deeply affected by such methodology during the undergraduate study in computer architectures.

In conclusion, the progress made in this thesis exemplifies how combining advanced learning architectures with rich, synthesized data can address critical bottlenecks in remote sensing of trees. The foundational role of effective learning methods such as Transformers and Diffusion models is clear, yet the future success of this field will depend on continued innovation in how models learn and what they learn—toward the ultimate goal of creating comprehensive, scalable, and insightful remote sensing paradigms.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *Journal of Machine Learning Research*, vol. 22, no. 30, pp. 1–82, 2021.

[3] J. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021. DOI: 10.1038/s41586-021-03819-2.

[4] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[5] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, 2020.

[6] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th. Pearson, 2018.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, 2012, pp. 1097–1105. [Online]. Available: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[8] M. Caron, H. Touvron, I. Misra, *et al.*, "Emerging properties in self-supervised vision transformers," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

[9] J. Shan and C. K. Toth, *Topographic Laser Ranging and Scanning: Principles and Processing*. CRC Press, 2018.

[10] J. C. Curlander and R. N. McDonough, *Synthetic Aperture Radar: Systems and Signal Processing*. Wiley, 1991.

[11] J. Wei, Y. Luomei, X. Zhang, and F. Xu, "Learning surface scattering parameters from sar images using differentiable ray tracing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024. DOI: 10.1109/TGRS.2024.3459620.

[12] T.-M. Li, M. Aittala, F. Durand, and J. Lehtinen, "Differentiable monte carlo ray tracing through edge sampling," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–11, 2018.

[13] F. E. Fassnacht, H. Latifi, A. Ghosh, *et al.*, "Review of studies on tree species classification from remotely sensed data," *Remote Sensing of Environment*, vol. 186, pp. 64–87, 2016.

[14] M. Brandt, C. J. Tucker, A. Kariryaa, *et al.*, "An unexpectedly large count of trees in the west african sahara and sahel," *Nature*, vol. 587, no. 7832, pp. 78–82, 2020. DOI: 10.1038/s41586-020-2824-5. [Online]. Available: https://doi.org/10.1038/s41586-020-2824-5.

[15] M. Aubry-Kientz, A. Laybros, B. Weinstein, *et al.*, "Multisensor data fusion for improved segmentation of individual tree crowns in dense tropical forests," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 3927–3936, 2021. DOI: 10.1109/JSTARS.2021.3069159.

[16] J. L. Cárdenas, C. J. Ogayar, F. R. Feito, and J. M. Jurado, "Modeling of the 3d tree skeleton using real-world data: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 12, pp. 4920–4935, 2023. DOI: 10.1109/TVCG.2022.3193018.

[17] Interactive Data Visualization, Inc. "Speedtree." Accessed: 2025-06-09. (2024), [Online]. Available: https://store.speedtree.com/.

[18] H. Xu, N. Gossett, and B. Chen, "Knowledge and heuristic-based modeling of laser-scanned trees," *ACM Trans. Graph.*, vol. 26, no. 4, 19–es, Oct. 2007, ISSN: 0730-0301. DOI: 10.1145/1289603.1289610. [Online]. Available: https://doi.org/10.1145/1289603.1289610.

[19] S. Du, R. Lindenbergh, H. Ledoux, J. Stoter, and L. Nan, “Adtree: Accurate, detailed, and automatic modelling of laser-scanned trees,” *Remote Sensing*, vol. 11, no. 18, p. 2074, 2019.

[20] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *arXiv preprint arXiv:2006.11239*, 2020.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[22] J. J. Lee, B. Li, S. Beery, *et al.*, “Tree-d fusion: Simulation-ready tree dataset from single images with diffusion priors,” in *European Conference on Computer Vision*, Springer, 2024, pp. 439–460.

[23] Z. Liu, K. Wu, J. Guo, Y. Wang, O. Deussen, and Z. Cheng, “Single image tree reconstruction via adversarial network,” *Graphical Models*, vol. 117, p. 101 115, 2021, ISSN: 1524-0703. DOI: https://doi.org/10.1016/j.gmod.2021.101115. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1524070321000205.

[24] B. Li, J. Kałużny, J. Klein, *et al.*, “Learning to reconstruct botanical trees from single images,” *ACM Trans. Graph.*, vol. 40, no. 6, Dec. 2021, ISSN: 0730-0301. DOI: 10.1145/3478513.3480525. [Online]. Available: https://doi.org/10.1145/3478513.3480525.

[25] H. Huang, G. Tian, and C. Chen, “Evaluating the point cloud of individual trees generated from images based on neural radiance fields (nerf) method,” *Remote Sensing*, vol. 16, no. 6, 2024, ISSN: 2072-4292. DOI: 10.3390/rs16060967. [Online]. Available: https://www.mdpi.com/2072-4292/16/6/967.

[26] Y. Li, Z. Liu, B. Benes, X. Zhang, and J. Guo, “Svdtree: Semantic voxel diffusion for single image tree reconstruction,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 4692–4702. DOI: 10.1109/CVPR52733.2024.00449.

[27] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165.

[28] OpenAI, *GPT-4 technical report*, https://openai.com/research/gpt-4, Accessed: 2025-05-26, 2023.

[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 684–10 695, 2022.

[30] N. Shazeer, A. Mirhoseini, K. Maziarz, *et al.*, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations (ICLR)*, 2017.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019.

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org.

[33] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008. [Online]. Available: https://arxiv.org/abs/1706.03762.

[34] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 405–421.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[36] R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, *Zero-1-to-3: Zero-shot one image to 3d object*, 2023. arXiv: 2303.11328 [cs.CV].

Author's Biography

1. Personal Information

   Name: Bingjun Guo

   Gender: Male

   Date of Birth: 2003-04-16

2. Education

   2021.09-2025.06 Zhejiang University, Bachelor's Degree

   2018.09-2021.06 Shenzhen Middle School

3. Awards

   Spring 2024, Fall 2025 Dean's List, UIUC

   Summer 2023 Outstanding Summer Research Project

   Spring 2023 Finalist in Mathematical Contest in Modeling

4. Research Experience

   2024.01-2024.10 Research Intern, Data Mining Group, UIUC

   2023.06-2023.08 Research Intern, Center for Data Science, Zhejiang University

   2022.06-2022.07 Research Intern, Data Science and Knowledge Engineering Group, ZJU-UIUC Institute

5. Publication

   Bowen Jin, Ziqi Pang, Bingjun Guo, Yu-Xiong Wang, Jiaxuan You, Jiawei Han, "InstructG2I: Synthesizing Images from Multimodal Attributed Graphs", NeurIPS 2024.