Title	歴史的景観の理解を深める:物体検出と画像深度推定による江南伝統庭園の空間分析
Author(s)	GAO, CHAN
Citation	
Issue Date	2024-12
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/19682
Rights	
Description	Supervisor: KIM Eunyoung, 先端科学技術研究科, 博士

Doctoral Dissertation

Advancing the Understanding of Historic Landscapes: Spatial Analysis of Jiangnan Traditional Gardens through Object Detection and Image Depth Estimation

Gao Chan

Supervisor: KIM Eunyoung

Graduate School of Advanced Science and Technology

Japan Advanced Institute of Science and Technology

[Knowledge Science]

December 2024

Abstract

Jiangnan gardens exemplify traditional Chinese landscape design, blending natural aesthetics with architectural innovation. While extensive research has been conducted on these gardens, a comprehensive spatial analysis of their complex landscapes remains lacking. This research aims to enhance the understanding of historic landscapes through advanced spatial analysis of Jiangnan traditional gardens using object detection and image depth estimation. The primary objective is to develop and apply an improved object detection algorithm, tailored for the intricacies of Jiangnan gardens, to identify and catalog key visual elements like pavilions, rockeries, and plants. This research introduces enhancements to the YOLOv8 algorithm, including the Diverse Branch Block (DBB), which optimizes feature extraction across different scales; the Bidirectional Feature Pyramid Network (BiFPN), enhancing feature integration from multiple layers; and Dynamic Head Modules (DyHead), which dynamically adjust the detection heads for better object recognition performance. Concurrently, the research seeks to analyze the depth and complex spatial relationships within the gardens to understand their design and functional aesthetics better. Employing the enhanced YOLOv8 for object detection and the Marigold algorithm for depth estimation, the study has provided exceptional insights. YOLOv8 effectively cataloged various elements, while Marigold mapped their spatial interactions with high precision, revealing the interplay between architectural and natural features and enhancing understanding of the gardens' historical and cultural contexts. This integration of object detection with depth mapping offers a novel methodology for exploring complex cultural landscapes. The findings suggest substantial implications for enhancing virtual tours and educational programs, promoting broader access to these cultural heritage sites. Overall, this research not only enriches our understanding of Jiangnan traditional gardens but also advances methodologies for preserving and interpreting complex heritage sites, promising innovative solutions for challenges in historic landscape conservation.

Keywords:

Jiangnan Traditional Gardens, Space Analysis, Object Detection, YOLOv8, Image Depth, Marigold Algorithm

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor, Prof. KIM Eunyoung, whose guidance and support in my research and access to scientific equipment were indispensable for the training of my algorithms. Without this support, my research would not have been possible. I also extend my thanks to the supervisor of my minor research and to all the dissertation review committee supervisors who have contributed their expertise to my research.

Transitioning from professional guidance to personal support, I owe a special thank you to my parents, who provided immeasurable care to my little daughter during my most challenging times in my doctoral studies. Their unwavering support was crucial in allowing me to focus on my research.

Heartfelt thanks also go to my husband, who has been my steadfast companion. Whenever my research seemed overwhelming, he was there to lift my spirits, taking me into nature to rejuvenate. To my son, Jimmy, who bravely joined me in Japan and adapted to a Japanesespeaking kindergarten, your strength and independence in learning Japanese and communicating with peers fills me with pride. I am also immensely thankful to my daughter, who had to adapt to being weaned at only ten months old when I moved to Japan to pursue my doctoral studies. Under the loving care of her grandparents, she learned how to drink, how to walk, and has grown healthy and strong. I hope that my absence in my children's life is only temporary, and that my love will never be absent in the future.

Reflecting on the year 2020, amidst the pandemic, I found myself in China, attending online classes, studying four courses as a student, and teaching three as a lecturer. These experiences underscored the rapid passage of time and the extent of challenges I have managed to overcome. I am grateful for having dared to step out of my comfort zone during my doctoral studies, moving from design and painting to learning coding, Python, and the basics of network structures.

In conclusion, I am thankful for everything and everyone I have encountered along this journey. These experiences have profoundly shaped me into who I am today. I am committed to living passionately, pursuing what I truly love, and maintaining a lifelong attitude of learning and a perpetually youthful spirit.

Table of Contents

Chapter 1: Introduction 1
1.1 Background of Jiangnan Traditional Gardens 2
1.1.1 Historical Development and Societal Influence 3
1.1.2 Cultural Reflections 4
1.1.3 Architectural and Artistic Merit 5
1.1.4 Influence and Legacy 6
1.1.5 Ecological and Sustainable Practices 6
1.2 Analysis of Complex Spaces in Jiangnan Traditional Gardens 7
1.2.1 Spatial Philosophy and Design Principles 8
1.2.2 The Use of Borrowed Scenery 8
1.2.3 Architectural Elements and Their Spatial Roles 9
1.2.4 Water and Rockery: Essential Components 10
1.2.5 Pathways and Movement 10
1.2.6 Seasonal and Time Dynamics 11
1.2.7 Cultural Significance and Preservation 11
1.3 Research gaps 12
1.3.1 A review of spatial research methods on Jiangnan traditional gardens 12
1.3.2 Research gaps 13
1.4 Objectives of the Study 16
1.5 Structure of the Thesis 18
Chapter 2: Literature Review 20
2.1 The Space Layout of Jiangnan Traditional Garden 21
2.2 Literature review of object detection algorithms 27
2.2.1 The development history of object detection 27
2.2.2 Development and changes of YOLO architecture 29
2.2.3 Previous Studies on YOLO algorithm in object detect 33
2.3 Literature Review of Image Depth Analysis 36
2.3.1 Development of Depth Estimation Technology 36
2.3.2 Application of the Marigold Algorithm 37
2.4 Conclusion 37
Chapter 3: Methodology of Study 1 39
3.1 YOLOv8 Architecture and Network Structure 40
3.2 Overview of the Enhanced YOLOv8 Network Design 42
3.3 The Improvements of YOLOv8 43
3.3.1 Diverse Branch Block (DBB) 43
3.3.2 Bidirectional Feature Pyramid Network (BiFPN) 46
3.3.3 Dynamic Head Modules (DyHead) 49
3.4 Dataset Compilation and Characteristics 53
3.4.1 Image Collection 53
3.4.2 Classification of Garden Elements 56
3.4.3 Dataset Annotation 63

3.4.4 Data Preprocessing and Augmentation Methods 59
3.5 The Computer Configuration and Parameter Settings of the Model Training 65
3.5.1 Experimental Setup and Model Configuration 65
3.5.2 Training Details 66
3.6 Evaluation Metrics 67
3.6.1 Evaluation Metrics Employed: 67
3.6.2 Threshold Settings for Model Evaluation: 69
Chapter 4: Methodology of Study 2 71
4.1 Data Collection 72
4.2 Marigold Depth Estimation Model 72
4.2.1 Selection and Rationale for the Depth Estimation Model 72
4.2.2 Framework of the Marigold Depth Estimation Model 75
4.3 Key Considerations for Depth Map Analysis 78
4.3.1 Depth Continuity and Smoothness 78
4.3.2 Edge Preservation 79
4.3.3 Handling of Overlapping Structures 79
4.3.4 Response to Light and Shadows 79
4.3.5 Color and Texture Handling 80
4.3.6 Multiscale Feature Learning 80
Chapter 5: Results of Study 1 82
5.1 Performance of the YOLOv8n-modify Model 83
5.1.1 Precision Analysis 83
5.1.2 mAP Analysis 84
5.1.3 The recall rate 86
5.1.4 Losses 87
5.2 YOLOv8n-modify Model Comparative Analysis with YOLOv8n Models 89
5.2.1 Comparative of The F1 versus confidence curve 89
5.2.2 Comparative of The confusion matrix 90
5.3 Comparative Analysis with Alternative Methods 92
5.4 Ablation Study 99
5.5 Changes in Heatmaps of Attention Mechanisms During Model Training 104
Chapter 6: Results of Study 2 107
6.1 Performance of the Marigold Model 108
6.1.1 Depth Map Accuracy and Detail 108
6.1.2 Robustness to Environmental Variabilities 109
6.1.3 Handling of Complex and Overlapping Structures 109
6.2 Analysis of the depth map of specific scenes in the garden 110
6.2.1 The Zhuozheng Garden borrows the scenery of the North Temple Pagoda 110
6.2.2 Scene 2: The perspective of Wuzhu Youju Pavilion 115
6.2.3 Scene 3: The Perspective Obscured by a Foreground Window 119
6.2.4 Scene 4: Garden view at dusk 121
6.3 Comparison results with other models in different scenarios 123
6.3.1 Scene 1: The Zhuozheng Garden borrows the scenery of the North

Temple Pagoda 123
6.3.2 Scene 2: Pavilion by the water at dusk 125
6.3.3 Scene 3: Perspective Obscured by a Foreground Window 128
6.3.4 Scene 4: Garden view at dusk 130
Chapter 7: Discussion and conclusion 134
7.1 Discussion of Experimental Results 135
7.1.1 Discussion of study 1 135
7.1.2 Discussion of study 2 141
7.2 Conclusions 144
Chapter 8: Limitations, future research and contributions 148
8.1 Limitations of the Study 149
8.1.1. Limited Detection of Small Targets 149
8.1.2 Robustness to Occlusion and Complex Scenes 149
8.1.3 Limitations of the Marigold algorithm 150
8.1.4 The Recommendations for Model Improvement 150
8.2 Future Research 151
8.2.1Future Research on Cultural Heritage Protection 151
8.2.2. Advanced Areas of Algorithmic Research 152
8.2.3 Digital Twin Research Directions 153
8.3 Contributions to advance the understanding of historic landscapes research155
8.4 Contribution to Knowledge Science 157
List of publications and presentations 161
Chinese references 165

List of Figures

Figure 1-1: Zhuozheng Garden 8

Figure 1-2: Guanyun Stone in Suzhou Liu Garden 10

Figure 2-1: The Zhuozheng Garden 24

Figure 2-2: View from Yi Yu pavilion in Zhuozheng Garden 25

Figure 2-3: View from Wuzhu Youju pavilion in Zhuozheng Garden 26

Figure 2-4: View from "Xiao Feihong" and "Xiao Canglang" 26

pavilion in Zhuozheng Garden 26

Figure 2-5: ViolaJones detector from Viola, P., & Jones, M. (2001, December) 27

Figure 2-6: Detailed structure of Faster RCNN 29

Figure 2-7: The development of Object detection methods 30

Figure 3-1: Detailed structure of YOLOv8 42

Figure3-2: Six transformations to implement an inferencetime DBB by a regular convolutional layer 44

Figure 3-3: the framewoek of the BiFPN 47

Figure 3-4: Dynamic head (Dyhead) framework 49

Figure 3-5: The Dynamic head (Dyhead) framwork (Dai, X et al.,2021) 50

Figure 3-6: Attention heat map after applying the DyHead mechanism 52

Figure 3-7: Examples of traditional traditional gardens in Jiangnan 56

Figure 3-8: Annotation Process 64

Figure 3-9: Data Augmentation Techniques 62

Figure 4-1:Overview of the Marigold fine-tuning protocol (Ke, B., et al.,2024).CVPR 2024 (Oral, Best Paper Award Candidate) 76

Figure 4-2: Overview of the Marigold inference scheme.(Ke, B., et al.,2024).CVPR 2024 (Oral, Best Paper Award Candidate) 77

Figure 5-1: Performance parameters of model training for 500 rounds 83

Figure 5-2: F1 and confidence curve 89

Figure 5-3: Confusion matrix 90

Figure 5-4: Comparison of precision between YOLOv8n-modify and other models 95

Figure 5-5: Comparison of recall value between YOLOv8n-modify and other models 95

Figure 5-6: Comparison of mAP50 between YOLOv8n-modify and other models 98

Figure 5-7: Comparison of FPS between YOLOv8n-modify and other models. 99

Figure 5-8: Heatmap changes of the attention mechanism during model training 103

Figure 7-1: Comparison of YOLOv8n and YOLOv8n-modify in various scenes 139

List of Tables

Table 1-1 Types of Borrowed Scenes..9
Table 2-1：Examples of Chinese Jiangnan traditional Gardens 22
Table 3-1：Classification of Garden Elements 58
Table 5-1: YOLOv8n-modify Comparative Analysis with Alternative Methods 93
Table 5-2：Ablation Study.. 100

List of Abbreviations

Abbreviation	Description
BiFPN	Bidirectional Feature Pyramid Network
C2f	CSPLayer_2Conv
CAM	Channel Attention Mechanism
CBAM	Convolutional Block Attention Module
CNN	Convolutional neural networks
CSPLayer	Cross Stage Partical
CVPR	Computer Vision and Pattern Recognition
DBB	Diverse Branch Block
DCNNs	Deep convolutional neural networks.
DPM	Deformable Part Models
DyHead	Dynamic Head Modules
FPNs	feature pyramid networks
FPS	Frames Per Second
HOG	Histogram of Oriented Gradients
IDE	Integrated development environment
IoU	Intersection over Union
LDM	latent diffusion model
mAP	Mean Average Precision
mAP@50	Mean Average Precision at 50
mAP@50-95	Mean Average Precision from 50 to 95
MRO	Major Research Objective
NeRF	Neural Radiance Fields
NLP	Natural language processing
NMS	NonMaximum Suppression
RCNN	Regions with CNN features
RPN	Region Proposal Network
SAM	Spatial Transformer Networks
SGD	Stochastic Gradient Descent
SROs	Secondary Research Objectives
SSD	Single Shot MultiBox Detector
SVM	Support Vector Machine
VAE	Variational Autoencoder
YOLO	You Only Look Once

Chapter 1: Introduction

1.1 Background of Jiangnan Traditional Gardens

The traditional gardens of Jiangnan are quintessential elements of ancient Chinese heritage, encapsulating a profound historical significance and enriched with deep artistic and cultural nuances. These gardens, as holistic artistic endeavors, reflect the ancient dwellers' intense admiration and quest for harmony, balance, and the essence of natural beauty. They serve not only as mere aesthetic displays but also as embodiments of philosophical ideals and cultural virtues(Zhang Qingping, Li Xia, 2018).

Jiangnan gardens are exemplary of historic landscapes, embodying centuries of cultural, aesthetic, and ecological wisdom inherent to the region. These gardens are not merely ornamental but are deeply interwoven with the philosophical and practical aspects of traditional Chinese landscape architecture. They reflect the sophisticated art of space, harmony, and perspective, integrating natural elements with man-made structures to create serene environments that have been preserved and cherished through generations. This profound connection to history and nature makes Jiangnan gardens a vital study subject in the exploration of historic landscapes.

As visitors traverse these gardens, they are immersed in an intricately woven tapestry of art, architecture, horticulture, and poetry, among other artistic expressions. These elements are seamlessly integrated, crafting an ambiance that fluctuates between serene tranquility and vibrant vitality. Such experiences are designed to engage the senses while also fostering a contemplative understanding of nature and aesthetics .

The design and layout of Jiangnan traditional gardens demonstrate a meticulous attention to detail, where traditional craftsmanship is applied with precision. Designers have skillfully manipulated the limited space to arrange stones, water features, plant life, and architectural elements in a manner that creates an illusion of vastness and layered depth(Chen Fenfang, 2007). This careful arrangement allows for a dynamic interplay of light, shadow, and texture, enhancing the sensory experience of each visitor.

Each component of the garden is imbued with symbolic meaning. From the placement of rocks to the choice of plants and the design of water elements, every aspect is thoughtfully selected to convey specific cultural and philosophical messages. These gardens are not only a sanctuary for personal reflection but also a medium through which the ancient Chinese philosophies and cultural values are communicated and preserved(Gong Xinjun, 2023).

In essence, Jiangnan traditional gardens are more than just scenic spots; they are living museums of culture and artistry. They offer a profound insight into the philosophical underpinnings and cultural aspirations of their creators, providing a window into the soul of ancient Chinese civilization(Wang Zhigang, 2021). Here, every path and pavilion tells a story, every garden scene evokes emotion, and every natural element celebrates the timeless pursuit of beauty and understanding.

1.1.1 Historical Development and Societal Influence

Jiangnan traditional gardens emerged prominently during the Ming Dynasty, reaching their zenith in the Qing era. These gardens were created primarily by literati and affluent merchant families seeking tranquil sanctuaries away from urban life′s turbulence(Wang Jue, 2008). The prosperity of these times enabled these individuals to invest in largescale landscaping projects, seamlessly blending art, nature, and architecture. The design and spatial arrangements reflect aesthetic principles of simplicity, elegance, and modesty, with winding paths and strategically placed pavilions providing varied perspectives that emphasize the Chinese aesthetic concept of ″scenic spots,″ revealing new surprises at each turn.

These gardens were vital spaces for the scholarly elite to engage in intellectual pursuits such as calligraphy, painting, poetry, and playing the guqin(Wang Ming, 2007). Adorned with inscribed stones and plaques, they reflect the artistic and literary tastes of their creators.

Jiangnan traditional gardens, epitomized by the classical gardens of Suzhou, hold a revered place in the cultural and historical narrative of China. These gardens are not just scenic vistas but living canvases that illustrate the sophisticated integration of art, nature, and architecture. The historical significance of these gardens is deeply rooted in the social, intellectual, and aesthetic realms of Chinese history, particularly during the Ming and Qing dynasties.

The traditional gardens of Jiangnan were predominantly owned and cultivated by the literati, the scholarly elite of their times, who were deeply influenced by Confucian, Taoist, and Buddhist philosophies. They sought to design these spaces as extensions of their scholarly pursuits, imbuing them with philosophical and poetic significance. The gardens′ layouts emphasized high cultural values such as contemplation, aesthetic appreciation, and harmonious blending of human and natural elements. The spaces often served as settings for social gatherings where poetry, painting, and philosophical discussions flourished, enabling intellectual exchange and artistic collaboration.

1.1.2 Cultural Reflections

Culturally, Jiangnan traditional gardens manifest traditional Chinese values like harmony with nature, simplicity, and intellectual and artistic refinement. These principles are reflected in their design, which replicates natural landscapes in miniature, maintaining a balance between yin and yang and emphasizing the flow of energy. Through meticulous placement of rocks, water features, pavilions, and pathways, the designers achieved a dynamic flow that enhances the spiritual experience. Every element has symbolic meaning: rocks represent mountains, ponds mimic lakes, and plants are chosen not only for their aesthetic appeal but also for their deeper symbolic significance.

1.1.3 Architectural and Artistic Merit

Jiangnan gardens often employ the principle of ＂borrowed scenery＂, integrating distant landscapes into their design to make the garden appear larger and more interconnected with the surrounding environment.

The architectural elements like pavilions, corridors, and bridges are meticulously arranged to provide multiple vantage points. The interplay between built structures and natural elements embodies a unique relationship between humanity and the environment.

The use of rockeries, representing mountains, and water features like ponds or streams creates a miniature version of the natural world. These features emphasize the dynamic and unpredictable patterns of nature, offering a tranquil backdrop for contemplation.

The architecture within Jiangnan gardens blends seamlessly with the natural environment, adhering to principles that emphasize modesty, restraint, and elegance. The use of space and perspective in these gardens illustrates advanced understanding of landscape architecture that influences modern design principles today. The winding paths and strategically placed structures are designed to offer changing vistas and experiences, symbolizing the journey through life and the continuous discovery of new perspectives.

Historically, these gardens also played a role in the ecological management of their locales. They were engineered to support local flora and fauna while optimizing water management, which is evident in the sophisticated designs of their water systems and plant selections. This ecological awareness embedded in the traditional landscaping practices provides a historical precedent for sustainable design and conservation practices in modern landscape architecture.

1.1.4 Influence and Legacy

The influence of Jiangnan gardens extends beyond their physical boundaries. They have inspired numerous artists, poets, and philosophers over centuries and continue to be a key subject in cultural studies and traditional Chinese art. Their design principles have influenced not only other Chinese garden styles but also international landscape architecture, promoting a blend of functionality, beauty, and ecological sensitivity.

In essence, Jiangnan traditional gardens are a profound cultural heritage that encapsulates the philosophical, aesthetic, and ecological wisdom of ancient China. They serve as a bridge connecting the past to the present, offering insights into the Chinese way of life and aesthetics that continue to influence artistic and ecological thought globally.

Jiangnan traditional gardens represent a harmonious blend of art, philosophy, and nature that transcends time. Their historical and cultural significance is evident in their role as intellectual and artistic spaces, their embodiment of traditional Chinese aesthetics, and their innovative landscape architecture. The legacy of these gardens resonates globally, offering timeless insights into how humans can create spaces that nurture both body and spirit.

1.1.5 Ecological and Sustainable Practices

The ecological wisdom embedded in Jiangnan gardens is evident in their sustainability principles. The intricate network of ponds and streams is not only aesthetically pleasing but also serves as an effective water management system, mitigating flooding, providing irrigation, and supporting a diverse ecosystem of plants and animals. The gardens are designed to support various plant species, creating microhabitats that promote biodiversity based on both aesthetic and ecological considerations.

The ecological wisdom embedded in the construction of these gardens is reflected in their sustainability principles.

The intricate network of ponds and streams in Jiangnan gardens is not only aesthetically pleasing but also serves as an effective water management system. They mitigate flooding, provide irrigation, and support a diverse ecosystem of plants and animals.

The gardens are designed to support a variety of plant species, creating microhabitats that promote biodiversity. The selection of plants was based on both aesthetic and ecological considerations.

Many Jiangnan gardens were constructed on preexisting structures or altered landscapes, demonstrating an early form of adaptive reuse that aligns with modern sustainability ideals.

In summary, Jiangnan traditional gardens represent a harmonious blend of art, philosophy, and nature that transcends time. Their historical and cultural significance is evident in their role as spaces of intellectual and artistic activity, their embodiment of traditional Chinese aesthetics, and their innovative approach to landscape architecture. The legacy of these gardens continues to resonate globally, offering timeless insights into how humans can create spaces that nurture both body and spirit.

1.2 Analysis of Complex Spaces in Jiangnan Traditional Gardens

Jiangnan gardens, epitomizing the quintessence of traditional Chinese landscape design, embody a harmonious blend of natural beauty and architectural ingenuity. These gardens, developed over centuries, are not only a retreat from the everyday world but also a profound representation of cultural values and philosophical ideals. The spatial design of Jiangnan gardens is distinguished by its complexity and subtlety, aiming to reflect the natural world in a microcosmic form(Figure 1-1).

Garden space is an ideology that seeks harmony with nature under the idea of unity between heaven and man, going beyond physical space. It has an extending quality, using techniques to visually and spiritually expand the limited space, breaking through the constraints of the site, giving a sense of infinite space and imagination.

1.2.1 Spatial Philosophy and Design Principles

The design of Jiangnan gardens is deeply rooted in Taoist and Confucian principles, which emphasize harmony between humans and nature. This philosophy is expressed through the careful arrangement of space, where every element is placed to achieve balance and aesthetic unity. The garden spaces are designed to lead visitors on a journey through a series of meticulously crafted scenes, each framed by architectural elements such as moon gates and winding paths that guide the eye and control the perspective.

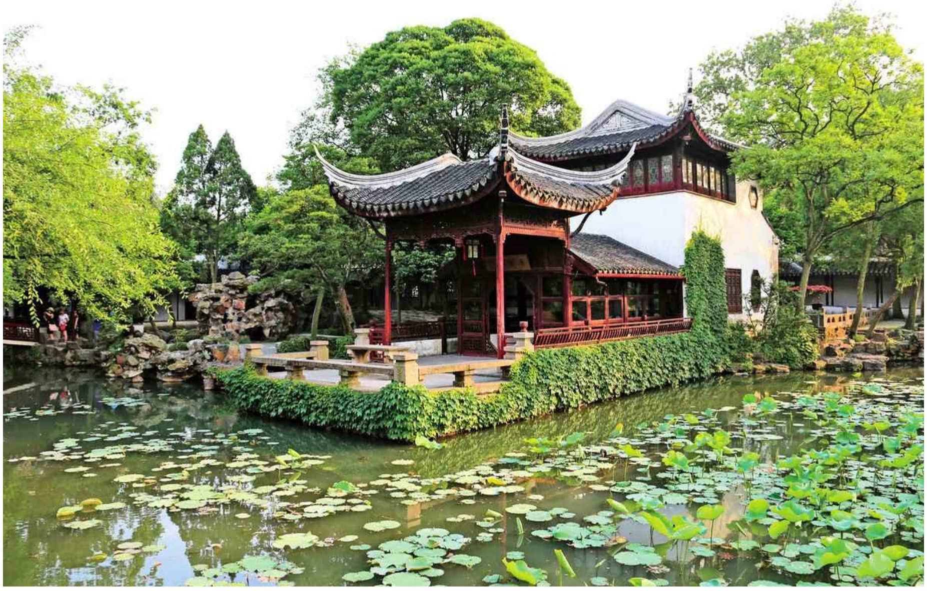


Figure 1-1: Zhuo Zheng Garden

1.2.2 The Use of Borrowed Scenery

One of the most distinctive features of Jiangnan gardens is the use of borrowed scenery ("Jie Jing"). This technique involves incorporating background landscapes into the garden's composition, thus extending the perceived depth and merging the garden with the surrounding nature seamlessly. Mountains, lakes, or other scenic views outside the garden boundaries are visually integrated, creating an illusion of an endless natural landscape. This method not only enhances the spatial depth but also enriches the garden's aesthetic and emotional impact.

Table 1-1 Types of Borrowed Scenes

Category	Definition	Examples
Borrowing scenery from Upward	Scenery borrowed from above, enlarging the vertical dimension of the garden.	1. Borrow from the sky and clouds 2. Borrow from high positioned architecture
Borrowing scenery from Downward	Scenery borrowed from below, exploiting lowlying areas to enrich the spatial layout.	1. Borrow from water reflections 2. Borrow from valleys and low lands
Borrowing scenery from a distance	Scenery borrowed from afar, extending the depth of the garden, adding layers.	1. View over the wall to 'borrow a scene' 2. borrow the beautiful scenery in the distance outside the garden
Borrowing scenery from adjacent areas	Scenery borrowed from adjacent areas, enhancing the garden's context by continuity.	1. Borrow from the side through corridors 2. Borrow from next door gardens
Borrowing scenery according to the season	Coordinating the natural changes of the four seasons, gardens primarily utilize the seasonal changes of plants	In spring, borrow the fluttering of winter jasmine; in summer, borrow the embellishments of lotus flowers; in autumn, borrow the straightness of bamboo; in winter, borrow the crystal clarity of icicles.

1.2.3 Architectural Elements and Their Spatial Roles

Architectural structures within Jiangnan gardens, such as pavilions, bridges, and towers, are strategically placed to create viewing points and to frame scenes artfully. These structures are often located at pivotal points to offer panoramic views of the garden, encouraging contemplation and interaction with the landscape. The architecture is not merely functional but also symbolic, with each element contributing to the narrative and thematic depth of the garden.

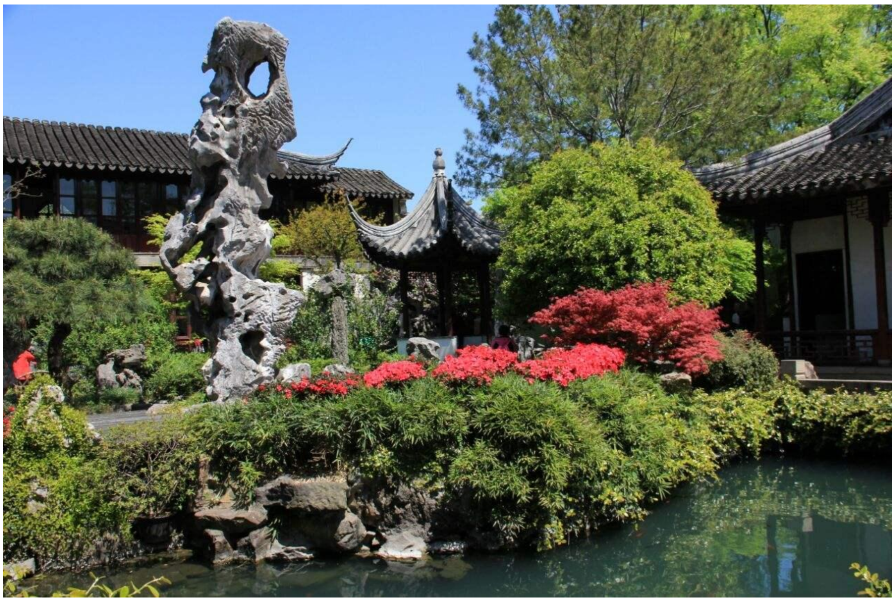


Figure 1-2: Guanyun Stone in Suzhou Liu Garden

1.2.4 Water and Rockery: Essential Components

Water features and rockeries are integral to Jiangnan gardens, serving both aesthetic and philosophical purposes. Water bodies, whether they are ponds, streams, or waterfalls, are central to the garden’s layout, reflecting sky and vegetation while providing a sensory contrast to the stone elements. Rockeries mimic natural mountains and hills, often constructed using scholar stones with unusual shapes and textures, representing the rugged beauty of nature(Figure 1-2).

1.2.5 Pathways and Movement

The pathways in Jiangnan gardens are deliberately meandering, designed to slow down the visitor’s pace and encourage a meditative interaction with the surroundings. The routes often lead to unexpected views and hidden enclaves, enhancing the garden’s mystery and depth. Each turn and bend is calculated to create anticipation and surprise, revealing the garden’s elements gradually and thoughtfully.

1.2.6 Seasonal and Time Dynamics

Jiangnan gardens are designed to offer varied experiences throughout different times of the day and across seasons. The placement and choice of plants, the orientation of structures, and the use of light and shadow are all planned to maximize the sensory experience in changing light conditions and seasonal transformations. This dynamic aspect ensures that the garden remains a source of perpetual discovery and delight.

1.2.7 Cultural Significance and Preservation

These gardens are not only spaces of aesthetic and ecological value but also of immense historical and cultural significance. They are living representations of China's artistic and philosophical traditions, requiring careful preservation and understanding. As urbanization and modern development pressures increase, the conservation of these gardens and their complex spatial arrangements becomes ever more critical.

In conclusion, the Jiangnan traditional gardens are a profound cultural heritage that exemplifies the sophisticated use of space to mirror philosophical ideals. The complexity of their design and the subtlety of their execution make them not only a subject of aesthetic enjoyment but also of scholarly interest, particularly in the fields of architecture, landscaping, and cultural studies. The continuing study and preservation of these gardens are essential for maintaining their beauty and cultural relevance for future generations.

1.3 Research gaps

1.3.1 A review of spatial research methods on Jiangnan traditional gardens

Jiangnan gardens, an integral part of Chinese cultural landscapes, have garnered attention from scholars across disciplines like architecture, landscape architecture, and narratology, enriching our understanding of their spatial constructs and narrative capacities.

Traditional garden studies typically focus on the aesthetic and technical aspects, such as plant arrangement, water system design, and architectural layouts. These studies often employ case study methodologies, providing detailed accounts of specific gardens to elucidate the embedded traditional knowledge and craftsmanship. Through such meticulous documentation, scholars reveal how garden designs align with their cultural and environmental contexts, thus preserving and interpreting traditional gardening knowledge (Zhang, 2004; Zhou, 2010).

Architectural narratology offers a novel perspective for examining the spatial and narrative aspects of Jiangnan gardens. Gu (2022) investigates how garden spaces convey stories, analyzing the role of spatial structures in supporting narrative functions within the gardens (Gu, 2022). This approach underscores the potential of architectural theories in uncovering the deeper narrative mechanisms embedded in garden layouts.

Cultural geography and humanistic landscape approaches also contribute significantly to our understanding of gardens by emphasizing their cultural and symbolic meanings. These studies explore gardens' roles in cultural memory and identity, offering insights into how garden elements reflect and shape local characteristics and historical transformations (Wang, 2016).

The application of technological and digital tools, such as GIS and 3D modeling, has increasingly supported the research of garden spaces. These tools enable more precise analysis of spatial configurations and visual impacts, facilitating a deeper comprehension of design principles and visual narratives within the gardens (Liu, 2020).

Interdisciplinary research methodologies provide a broader perspective on the study of Jiangnan gardens. Integrating theories and methods from architecture, art history, literature, and ecology allows researchers to analyze the multifunctional roles of gardens comprehensively. Such approaches reveal how gardens function as ecosystems and their significance in cultural and social contexts, contributing to discussions on garden conservation, restoration, and innovation (Chen, 2016; Peng, 2018).

In summary, the study of the traditional spatial layouts of Jiangnan gardens is enriched by a diverse array of research methodologies that span various disciplines. This multifaceted approach not only deepens our historical and aesthetic appreciation of these gardens but also informs contemporary practices in garden conservation and landscape design. By synthesizing insights from architecture, narratology, and ecology, this field continues to evolve, offering new theoretical and practical frameworks for understanding and preserving the intricate beauty of Jiangnan gardens.

1.3.2 Research gaps

Despite the substantial body of research on Jiangnan traditional gardens, significant gaps remain in the comprehensive spatial analysis of these complex landscapes. Traditional studies have contributed immensely to our understanding of the cultural, aesthetic, and technical aspects of garden design, yet they often fall short in fully addressing the space and interactions between garden elements. The reliance on manual surveys, case studies, and two-dimensional representations limits the depth and breadth of the spatial insights that can be derived. While these methods provide detailed documentation, they typically focus on the visual and historical aspects rather than the intricate space and relationships that define the experience of Jiangnan gardens.

Gap 1: Limited Use of Advanced Technologies in space Analysis

Most existing research on Jiangnan gardens has relied heavily on descriptive analysis and traditional documentation methods, such as architectural drawings, textual records, and basic 3D modeling. While these approaches offer valuable historical and aesthetic insights, they do not fully exploit the capabilities of modern technologies in spatial analysis. The growing use of GIS, 3D modeling has helped bridge this gap, but these tools often focus on surface-level visual representations, lacking the capacity to fully capture and analyze the spatial relationships between objects within the garden landscape.

Currently, there is no integration of advanced neural network algorithms like YOLOv8 for object detection and Marigold for depth mapping to conduct detailed space analysis in Jiangnan gardens. This gap is particularly critical because the space of Jiangnan gardens are multi-layered and complex, involving a delicate balance of natural and man-made elements. Traditional methods struggle to capture this depth, leading to an incomplete understanding of how these elements interact across varying spatial dimensions and scales.

Gap 2: Lack of Automated Methods for Analyzing Garden Elements

Another major limitation of existing studies is their reliance on manual and often subjective methods to document and analyze the different elements of Jiangnan gardens. While interdisciplinary approaches—drawing from architecture, landscape design, narratology, and cultural geography—have been useful in deepening our understanding of the cultural and symbolic meanings of these gardens, they still require significant manual intervention. These methods are often case-specific and do not provide scalable solutions for analyzing large datasets or multiple gardens simultaneously.

The lack of automated tools that can detect, classify, and analyze garden elements like pavilions, rocks, trees, and water features at various scales highlights a major gap in the field. Presently, research falls short in automating the recognition of objects within garden spaces and efficiently mapping their spatial relationships. This could improve the consistency and depth of spatial analysis across different gardens. This shortfall limits the possibility for wider comparative studies and diminishes the effectiveness of ongoing monitoring and conservation efforts.

Gap 3: Insufficient Integration of Depth Information for Spatial Understanding

While some studies have applied GIS-based tools to study garden layouts, these methods often provide only superficial representations of space, focusing on horizontal and vertical dimensions without adequately addressing the depth and layering of elements within the gardens. Jiangnan gardens are renowned for their intricate use of spatial depth—through techniques like "borrowed scenery" and the careful manipulation of sightlines—but current methods fail to capture this complexity effectively.

Marigold’s depth estimation model fills this gap by offering a way to analyze the spatial depth relationships between garden objects, such as the distance between a pavilion and the surrounding trees or how a rock formation aligns with the water features. The lack of depth information in existing studies limits the ability to fully understand the space of these gardens, especially how different elements are designed to interact across varying distances and levels within the garden.

1.4 Objectives of the Study

Major Research Objective (MRO)

The main objective of this study is to advance the understanding of historic landscapes by spatial analysis of Jiangnan traditional gardens through object detection and image depth estimation.

This objective addresses the intricacies of spatial composition that define the aesthetic and functional essence of these gardens. By focusing on the space , this study seeks to elucidate the underlying principles that make Jiangnan gardens unique cultural and historical landscapes.

Importance of the MRO:

The spatial arrangement in Jiangnan gardens embodies a unique blend of natural beauty and architectural sophistication, representing a critical aspect of Chinese cultural heritage. These gardens are not only spaces of visual and ecological appeal but also of immense historical value and philosophical depth. Understanding these space allows for a more effective conservation strategy and enhances the scholarly appreciation of their design principles, contributing to both academic research and practical preservation efforts.

Secondary Research Objectives (SROs)

SRO1: To construct an enhanced object detection algorithm tailored for the traditional gardens of the Jiangnan region.

This objective focuses on identifying and cataloging the diverse visual elements within Jiangnan gardens, such as pavilions, rockeries, water features, and plants.

Recognizing these elements is essential for understanding the visual and functional complexity of the gardens. This involves improved the YOLOv8 algorithm for detecting and classifying these features accurately within the varied and intricate landscapes of the gardens.

SRO2: To analyze the depth of the complex spatial relationships in Jiangnan traditional gardens.

The second secondary objective involves analyzing the spatial relationships among the identified visual elements. This includes studying how these elements interact to create the layered and depthfilled perspectives that are characteristic of Jiangnan gardens. This analysis will utilize advanced spatial analysis techniques Marigold algorithm to quantify and model these relationships, providing a deeper understanding of the garden's layout and design principles.

Relationship between MRO and SROs

The Major and Secondary Research Objectives are intricately linked to form a comprehensive framework for studying Jiangnan gardens. SRO1 establishes the foundational knowledge of the garden’s components, which is crucial for conducting any further analysis on their spatial interactions as addressed in SRO2. Together, these objectives not only foster a thorough understanding of the physical and visual complexity of Jiangnan gardens but also contribute to the broader field of cultural heritage preservation by providing detailed insights into the design and conservation of these historical landscapes. This structured approach ensures that each step in the research contributes directly to achieving the overarching goal of understanding and preserving the unique space of Jiangnan gardens.

1.5 Structure of the Thesis

This thesis systematically discusses the methodologies, results, and in-depth analysis of Jiangnan's traditional gardens, focusing on novel techniques for detecting garden elements using the enhanced YOLOv8 network and monocular image depth estimation techniques. The document is organized into eight chapters as follows:

Chapter 1: Introduction

This chapter covers the background of Jiangnan traditional gardens, noting their unique aesthetics and complex spatial layouts. It discusses the importance of cultural heritage protection and digital preservation, identifies research gaps, and outlines the major and secondary research objectives, highlighting their interrelationships.

Chapter 2: Literature Review

This chapter provides a comprehensive review of spatial research methods on Jiangnan traditional gardens, the evolution of object detection technologies, studies on YOLO networks, and literature on image depth analysis.

Chapter 3: Methodology of Study 1

This chapter begins with an overview of the enhanced YOLOv8 network design, including components like the Diverse Branch Block (DBB), the Bidirectional Feature Pyramid Network (BiFPN), and the Dynamic Head Modules (DyHead). It details the dataset compilation, characteristics, image collection, annotation, data augmentation strategies, and describes the model training, validation processes, and evaluation metrics used.

Chapter 4: Methodology of Study 2

This chapter introduces the mechanisms of the Marigold Depth Estimation Model and key considerations for depth map analysis.

Chapter 5: Results of Study 1

This chapter demonstrates the performance of the enhanced YOLOv8 model, showing improvements in precision and mAP analysis. It offers a comparative analysis with previous models and provides insights into the detection of garden elements and its implications for digitizing cultural heritage.

Chapter 6: Results of Study 2

This chapter discusses the use of the Marigold algorithm, focusing on the technical aspects of depth map analysis in Jiangnan gardens. It includes discussions on depth continuity, smoothness, edge preservation, overlapping structure processing, responses to light and shadows, color and texture handling, and multi-scale feature learning.

Chapter 7: Discussion and Conclusion

This chapter summarizes the findings, discusses the contributions to cultural heritage research, and highlights the implications of the study.

Chapter 8: Limitations and Future Research

This chapter outlines the study's limitations, suggests directions for future research, and discusses the contributions of this study to the field of knowledge science.

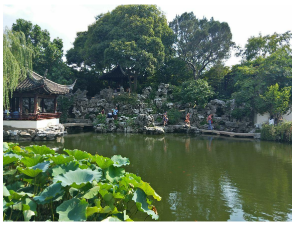
Chapter 2: Literature Review

1.1 The Space Layout of Jiangnan Traditional Garden

Traditional Jiangnan gardens possess irreplaceable classical artistic value and traditional historical and cultural significance unmatched by any other architectural form (Table 2-1). Their spatial layout serves as a vessel for carrying these artistic and cultural values. According to Peng Yigang's "Analysis of Chinese Classical Gardens" (Peng, Yi Gang, 1996), the spatial layout of Jiangnan traditional gardens can be categorized into three types: introverted, extroverted, and integrated. Introverted gardens feature buildings oriented towards inner courtyards, creating a centripetal layout with a strong sense of inward focus, as seen in Wuxi's Jichang Garden and Suzhou's Lingering Garden. Extroverted gardens, like Suzhou's Canglang Pavilion, center around buildings with surrounding courtyard landscapes, presenting a more centrifugal spatial arrangement. Integrated gardens, such as Suzhou's Zhuozheng Garden(Figure 2-1), are large enough to include both aforementioned layouts. These forms arise from the garden creators' philosophies, spatial needs, and traditional gardening techniques.

Furthermore, Jiangnan garden layouts exhibit four characteristics: sequentiality, permeability, subtlety, and contrast. The extension of garden spaces transcends physical boundaries, blending surrounding environments and landscapes into a "flowing space." Practically, this is achieved through the artistic segmentation of space using elements like flower walls, latticed windows, and covered bridges, allowing distant views through visual corridors. Spaces extend and permeate into each other, enriching the landscape's layers and creating the artistic effect of a "deep and profound courtyard." The contrasting element in garden spaces is omnipresent, including contrasts of solidity and void, light and shadow, size, and openness and enclosure. It can be summarized garden contrasts as seeing the large in the small, the real in the unreal, sometimes hidden, sometimes apparent, sometimes shallow, sometimes deep. Philosophically, these contrasts represent opposition and unity, cause and effect, crafting spatial variations that highlight characteristics and transform finite physical spaces into infinite spiritual realms.

Table 2-1：Examples of Chinese Jiangnan traditional Gardens

Garden Name	Location	Features	Famous Spots	Image
Zhuozheng Garden	Suzhou	Classic Suzhou garden with exquisite layout	Landscape Gallery, Covered Walkway, Small Bridges over Streams	
Liu Garden	Suzhou	Famous for its waterscapes and classical architecture	Five Peak Pavilion, Winding Corridors, Ancient Pavilions	
Wangshi Garden	Suzhou	Compact and exquisite with unique courtyard design	Secluded Paths, Flower Hall, Bamboo Grove	
Canglang Pavilion	Suzhou	Known for its waterscapes and ancient pavilions	Rippling Waves, Lake-view Pavilion, Tea House	
Lion Grove Garden	Suzhou	Famous for its rockeries and labyrinthine stone forests	Stone Caves among Trees, Pavilions, Water Ponds	
Yi Pu Garden	Suzhou	Known for its flowers and meandering waters	Babbling Streams, Flower Path, Fish Viewing Pond	
Ou Garden	Suzhou	A garden within a garden, with meticulous layout	Labyrinthine Corridors, Bridges, Ponds	

Tuisi Garden	Suzhou	Delicate and poetic garden	Distant Green Pavilion, Curved Bridge, Bamboo Path	
Shouxi Lake	Yangzhou	Scenic lake views, long corridors and bridges	White Pagoda, Little Gold Mountain, Fishing Platform	
Ge Garden	Yangzhou	Private garden with a unique style	Embracing Mountain Pavilion, Stone Boat, Garden Lake	
He Garden	Yangzhou	Famous for its garden architecture and water systems	Lotus Pond, Artificial Mountain, Covered Corridor	

The typical example: The Space Layout of Zhuozheng Garden

Zhuozheng Garden(Figure 2-1) holds a significant place among Jiangnan's private gardens, originating in the Ming dynasty and evolving through multiple renovations. Now, it is divided into three distinct sections: the east, middle, and west, with the central area being the most significant. This section constitutes a third of the garden's area, containing many of its highlights(Figure 2-2). It features a clear and layered landscape space, with a well-organized and coherent spatial layout.

Over different periods, the garden's layout has evolved, making the original and current garden layouts and architectural features distinct. This evolution showcases Zhuozheng Garden’s unique characteristics. Initially, the garden was designed to adapt to the local conditions, favoring abundant water landscapes(Figure 2-3). Historical records describe the garden as having many low-lying areas filled with water, which were then dredged and surrounded by trees and hills, creating a dynamic water garden landscape. This layout allowed for the integration of pavilions and bridges, adding to the garden’s architectural diversity.

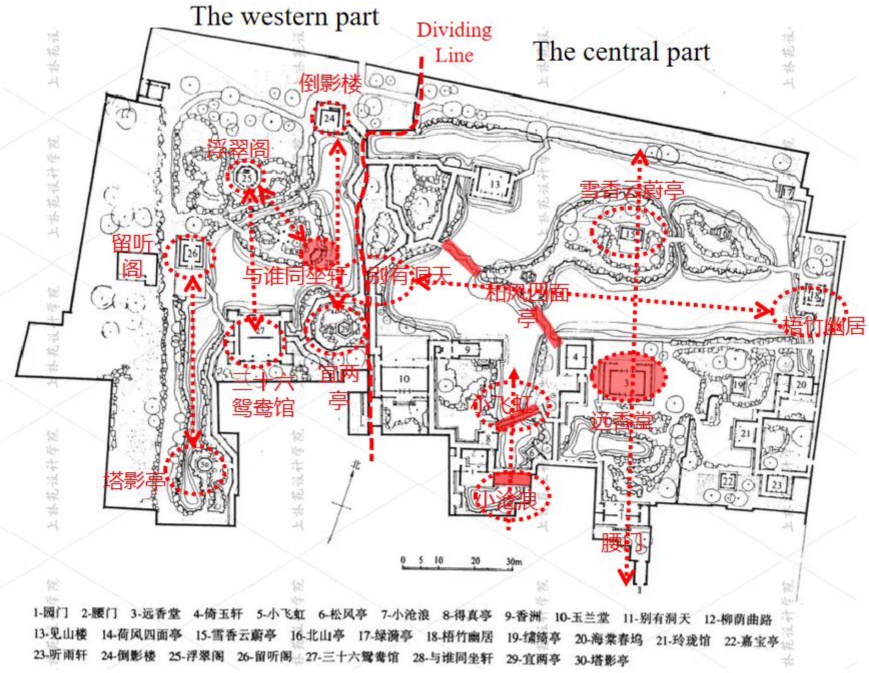

The western part
Dividing Line
The central part
倒影楼
浮翠阁
留听阁
与谁同坐轩
别有洞天
和风四面亭
雪香云蔚亭
梧竹幽居
三十六鸳鸯馆
宜两亭
塔影亭
小飞虹
远香堂
小沧浪
腰门
北
0 5 10 20 30m
1-园门 2-腰门 3-远香堂 4-倚玉轩 5-小飞虹 6-松风亭 7-小沧浪 8-得真亭 9-香洲 10-玉兰堂 11-别有洞天 12-柳荫曲路
13-见山楼 14-荷风四面亭 15-雪香云蔚亭 16-北山亭 17-绿漪亭 18-梧竹幽居 19-绣绮亭 20-海棠春坞 21-玲珑馆 22-嘉宝亭
23-听雨轩 24-倒影楼 25-浮翠阁 26-留听阁 27-三十六鸳鸯馆 28-与谁同坐轩 29-宜两亭 30-塔影亭

Figure 2-1: The Zhuozheng Garden

https://www.pinterest.com/pin/307230005825607479/

One of the standout features in Zhuozheng Garden is its extensive use of water, occupying a third of its space, earning it the nickname "water garden." Bridges like the Xiao Feihong, or 'Little Rainbow,'(Figure 2-4) enhance this watery landscape with their reflections shimmering in the water, reminiscent of rainbows. Compared to its original state, the current garden has fewer buildings, but those that remain are elegantly simple and complement the natural scenery without overwhelming it. The garden's architecture seems to float above the water, creating a distinct and picturesque setting that is enhanced by the surrounding flora.

Paths and corridors in Zhuozheng Garden are notably winding and deep, serving two purposes. They not only add aesthetic appeal and vitality to the space but also help in zoning and guiding visitors through the garden. The wave-shaped corridor in the western part of the garden was designed to be more creative than a simple wall, allowing views into different parts of the garden, linking and blending the landscapes seamlessly.

The central section of Zhuozheng Garden, being the core of the garden, contains multiple complex spatial forms. Although divided into different zones, the pathways interconnect these spaces, creating a cohesive and flowing visual journey through the garden. This structure, like poetry, has a rhythm to its spatial sequence, with open, semi-open, and enclosed spaces varying in size and composition, providing a dynamic and harmonious environment that enriches the visitor’s experience.

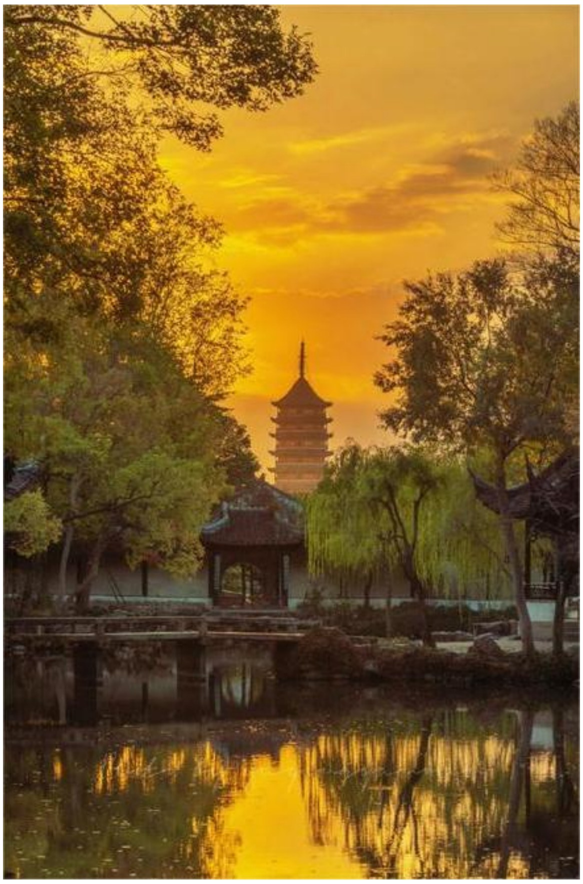


Figure 2-2: View from Yi Yu pavilion in Zhuozheng Garden

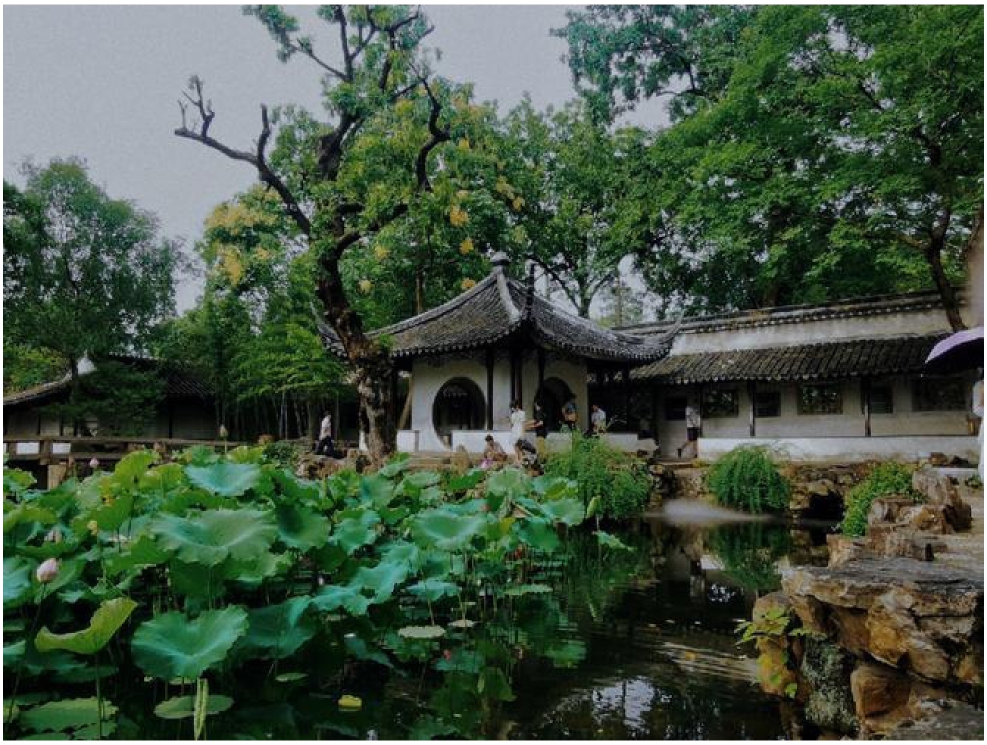


Figure 2-3: View from Wuzhu Youju pavilion in Zhuozheng Garden

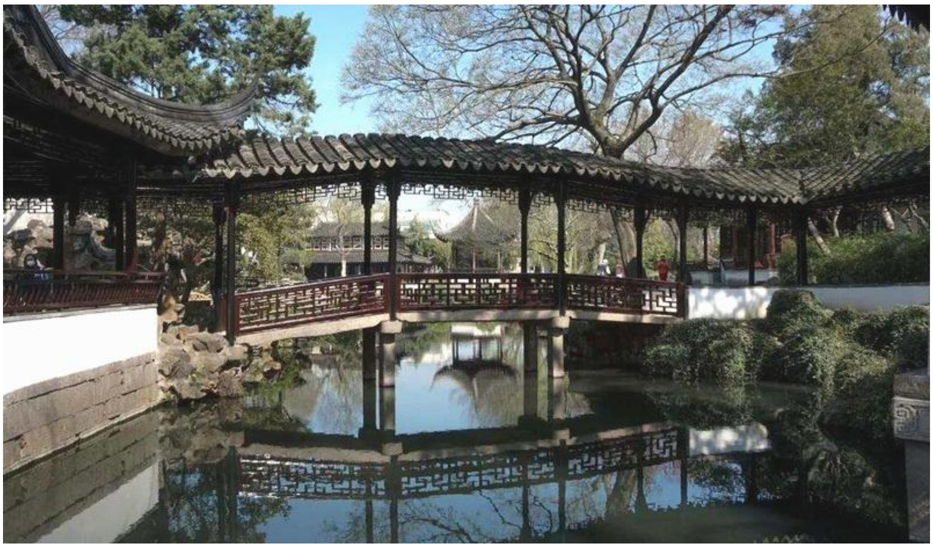


Figure 2-4: View from "Xiao Feihong" and "Xiao Canglang"

pavilion in Zhuozheng Garden

1.2 Literature review of object detection algorithms

2.2.1 The development history of object detection

The development of object detection techniques has evolved considerably over the years, transitioning from handcrafted, heuristic methods to the complex, deep learningbased systems widely used today. The earliest techniques largely relied on feature extraction methods designed specifically for the tasks at hand, leading to a surge in the use of specific feature descriptors and classifiers.

Early FeatureBased Methods

One of the most impactful early methods was the ViolaJones detector (Figure 2-5) , developed primarily for face detection(Viola & Jones, 2001). It utilized Haarlike features, which represented patterns of pixel intensity values, and a cascade classifier for rapid decisionmaking. This cascade structure enabled the classifier to reject large numbers of negative windows early in the process, making it remarkably fast. The success of this detector in facial recognition tasks demonstrated the power of using effective features and cascades for highspeed performance(Zhang & Viola, 2007).

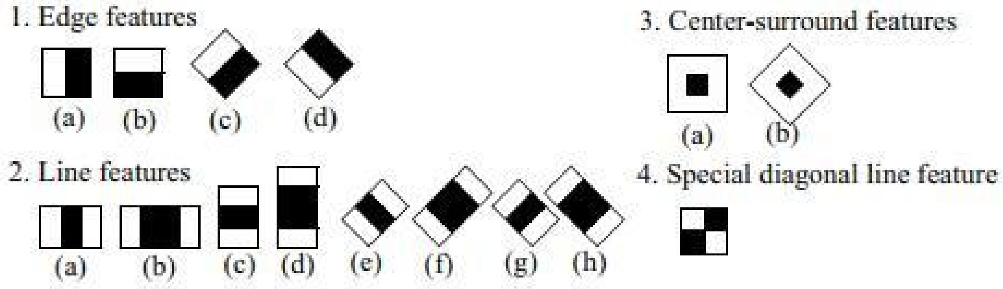


Figure 2-5: ViolaJones detector from Viola, P., & Jones, M. (2001, December)

As the field expanded beyond face detection to general object detection, researchers began exploring more sophisticated and flexible models. A notable leap forward was the Deformable Part Models (DPM). DPMs represented objects as compositions of various parts, allowing for deformation within a given range(Q. Zhu et al., 2006). The method used Histogram of Oriented Gradients (HOG) features and a latent SVM (Support Vector Machine) to identify objects at varying scales and orientations(Q. Zhu et al., 2006). This approach provided significantly improved results in detecting a wider range of object categories compared to rigid template methods(Viola & Jones, 2001).

The Deep Learning Paradigm Shift

The field underwent a revolutionary transformation with the emergence of deep learning and deep convolutional neural networks (DCNNs). One of the first significant contributions was the introduction of RCNN (Regions with CNN features)(Girshick et al., 2014). This method proposed a twostage process where regions were initially proposed using a selective search algorithm, and then a deep CNN was employed to classify each region proposal. Despite the impressive gains in detection precision, the process was computationally intensive due to the separate region proposal and feature extraction stages.

To tackle this inefficiency, Fast RCNN was introduced as a unified framework that merges region proposals and feature extraction within a single network. It utilizes ROI (region of interest) pooling to speed up the process (Liu et al., 2016). These enhancements significantly cut down on computation time and boosted detection accuracy. Nonetheless, the system continued to depend on external algorithms for proposing regions, which still posed a bottleneck.

Towards RealTime Detection

Building on previous developments, Faster RCNN (Figure 2-6) introduced the Region Proposal Network (RPN), which was integrated into the main network and designed to directly generate region proposals (Ren et al., 2015). By replacing traditional region proposal methods with the RPN, this model established an end-to-end deep learning framework that could generate high-quality proposals and classify them within a single seamless pipeline. This advancement represented a significant leap in both efficiency and precision compared to earlier models.

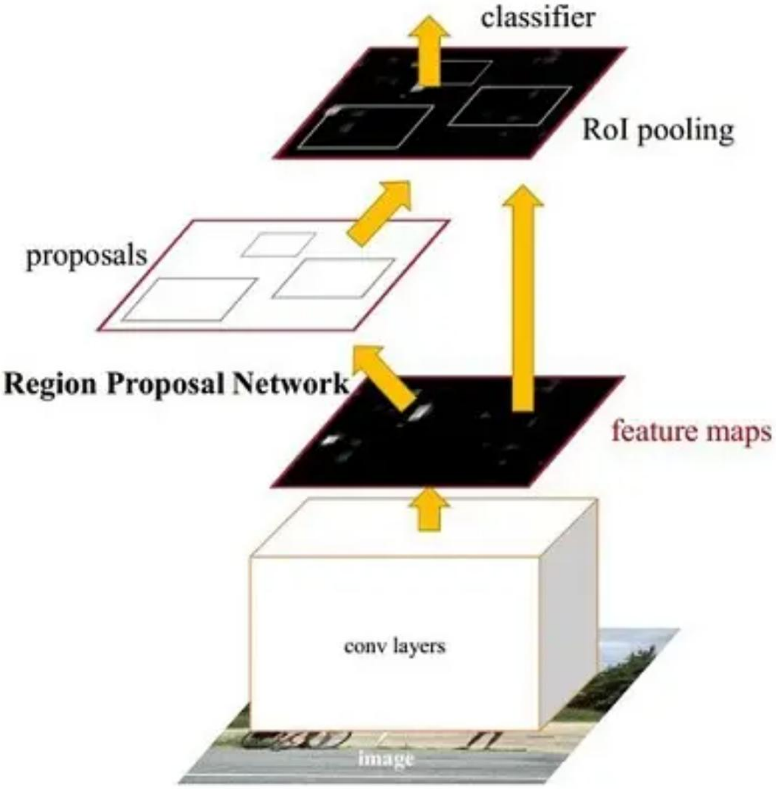


Figure 2-6: Detailed structure of Faster RCNN

While these multistage detectors offered strong precision, their relatively slower inference times limited their realworld applications, particularly in scenarios requiring realtime detection. To meet this challenge, singleshot detectors like YOLO (You Only Look Once)(Ren et al., 2015)and SSD (Single Shot MultiBox Detector)(Liu et al., 2016)were developed. They eliminated the need for separate region proposals, directly predicting bounding boxes and class probabilities across the entire image in one pass. YOLO used a gridbased approach, treating detection as a regression problem, while SSD introduced anchor boxes to handle varying aspect ratios more effectively.

2.2.2 Development and changes of YOLO architecture

The YOLO (You Only Look Once) series has significantly advanced the field of object detection through its innovative approach of treating object detection as a regression problem(Figure 2-7). Introduced in 2016, the original YOLO algorithm divided the input image into an S x S grid, with each cell responsible for predicting bounding boxes and class probabilities. Despite its realtime detection capability at 45 fps and the ability to detect multiple objects per grid cell, it faced challenges with small and closely packed objects due to the grid system.

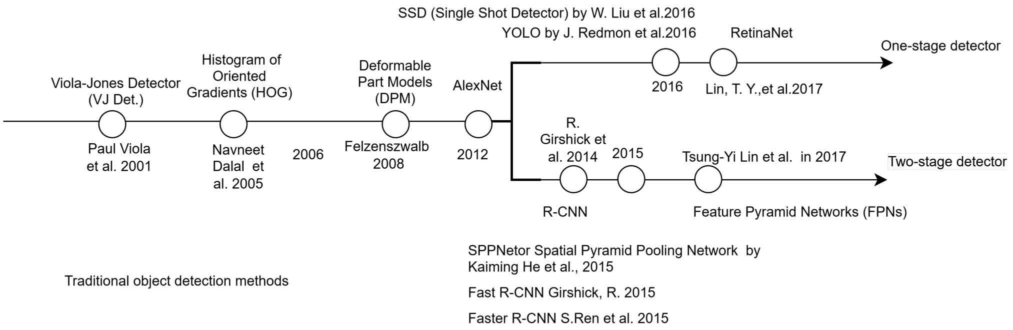

SSD (Single Shot Detector) by W. Liu et al.2016
YOLO by J. Redmon et al.2016
RetinaNet
One-stage detector
Viola-Jones Detector (VJ Det.)
Histogram of Oriented Gradients (HOG)
Deformable Part Models (DPM)
AlexNet
2016
Lin, T. Y.,et al.2017
Paul Viola et al. 2001
Navneet Dalal et al. 2005
2006
Felzenszwalb 2008
2012
R. Girshick et al. 2014
2015
Tsung-Yi Lin et al. in 2017
Two-stage detector
R-CNN
Feature Pyramid Networks (FPNs)
SPPNetor Spatial Pyramid Pooling Network by Kaiming He et al., 2015
Traditional object detection methods
Fast R-CNN Girshick, R. 2015
Faster R-CNN S.Ren et al. 2015

Figure 2-7: The development of Object detection methods

The subsequent iteration, YOLOv2, emerged in 2017 with architectural improvements such as batch normalization, anchor boxes, and multiscale training. These modifications enhanced detection precision, particularly through the introduction of anchor boxes which provided a predefined set of boxes to improve prediction stability and precision across different object sizes. However, the model still encountered difficulties with small objects and localization precision.

YOLOv3, released in 2018, further refined the model by incorporating a deeper network architecture, Darknet53, which utilized residual connections and multiscale feature maps. This version utilized three different scales of predictions to cater to small, medium, and large objects, significantly boosting the model's precision and robustness. Although it increased the computational load, the enhancements maintained a balance between speed and precision.

In 2020, YOLOv4 continued to advance the series by focusing on achieving a balance between detection speed and precision. It integrated the latest advancements such as CSPDarknet, PANet, and SAM, along with optimized training strategies and data augmentation techniques. The YOLOv4 architecture, while larger and more resourceintensive, managed to achieve realtime performance with enhanced precision.

Also in 2020, YOLOv5, developed by Ultralytics, marked a deviation from the original series, focusing on optimizing performance through a more lightweight architecture and ease of use. This iteration introduced novel augmentation techniques like Mosaic and AutoAugment, promoting faster inference speeds and versatility across different hardware configurations.

YOLOv7 in 2022 pushed the limits of computational efficiency with a new network structure designed to optimize both speed and precision. This latest version emphasizes model scaling and efficient architecture to set new benchmarks for realtime object detection performance.

Throughout its evolution, each version of YOLO has contributed to setting industry standards in the object detection domain, continuously influencing new research directions and applications. The series exemplifies a dynamic progression towards more sophisticated, efficient, and accessible object detection technologies, marking significant milestones in the landscape of computer vision.

The YOLO series has continued to evolve, reflecting ongoing improvements in machine learning technologies and their application in computer vision. Each iteration has systematically addressed the limitations of its predecessors while introducing new capabilities to enhance both precision and efficiency. This progression from YOLOv1 to YOLOv8 demonstrates a consistent commitment to optimizing the tradeoffs between speed and detection quality, which is critical for realtime applications.

The improvements in each version have been guided by both the need for better performance and the desire to simplify the operational demands of the algorithm in practical deployments. For instance, the enhancements in YOLOv3 through the introduction of multiscale feature maps and a more robust network backbone paved the way for the sophisticated architecture of YOLOv4 and YOLOv5, which integrated stateoftheart techniques such as CSPDarknet and advanced data augmentation to further push the boundaries of what could be achieved in terms of detection precision and speed.

YOLOv8, the most recent version in the YOLO series, incorporates a network structure built on the strengths of Feature Pyramid Networks and crosslayer connections to effectively integrate multiscale feature information. The architecture utilizes attention mechanisms and optimized strategies to boost detection precision and performance. The core structure of YOLOv8 includes a backbone network for feature extraction, typically involving deep convolutional neural network architectures like Darknet or ResNet, and a detection head with convolutional and fully connected layers that predict object bounding boxes and class probabilities.

The object detection task in YOLOv8 is approached as a regression problem. The network uses convolutional layers, pooling layers, and fully connected layers to predict object location and classification. Convolutional layers employ sliding kernels to capture local spatial structures from input data, while pooling layers minimize the dimensionality of feature maps by compressing and aggregating them through maxpooling. This strategy reduces computational demands and parameters while increasing translational invariance. The fully connected layer at the end of the network converts feature maps into object detection outputs.

Overall, YOLOv8 distinguishes itself through the effective integration of various architectural features, such as crosslayer connections, attention mechanisms, and a streamlined optimization strategy. This combination allows it to deliver highquality object detection by accurately locating objects and classifying them with minimal latency, marking a significant step forward in the YOLO series.

In academia and industry, the YOLO series has been pivotal in shifting the paradigm of how object detection can be approached and implemented. It has inspired numerous derivative works and adaptations, which have explored various aspects of the network architecture, training processes, and implementation strategies. The YOLO framework's influence extends beyond traditional applications, impacting sectors such as autonomous driving, surveillance, and interactive media, where the ability to quickly and accurately detect objects in realtime is paramount.

The continuous evolution of the YOLO series stands as a prime example of iterative innovation in artificial intelligence, with each new version improving upon the insights and limitations of its predecessors. As the field of computer vision expands, the legacy of the YOLO series in advancing object detection capabilities is set to inspire further breakthroughs, establishing it as a key benchmark in both the study and practical application of AI technologies.

2.2.3 Previous Studies on YOLO algorithm in object detect

The YOLO algorithm has proven instrumental across various research domains, demonstrating its adaptability and technical robustness.

（1）Construction and Urban Planning

Kumar et al. utilized YOLOv4 to enhance safety at construction sites by detecting fire and ensuring compliance with personal protective equipment (PPE), effectively illustrating the algorithm's capability in handling complex, dynamic environments in real time and improving safety protocols (Kumar et al., 2022). Zheng and Wu extended its application to ecological monitoring by employing YOLOv4Lite for accurate singletree detection in urban plantations, a crucial step for optimizing urban environmental resource management (Zheng, Y., et al., 2022).

（2）Civil Engineering and Autonomous Driving

Qiu et al. demonstrated the use of an improved YOLOv5 algorithm for detecting soil foreign objects using groundpenetrating radar in civil engineering projects, enhancing detection efficiency and precision, thus providing invaluable insights into subsurface anomalies (Qiu, Z., et al., 2022). Similarly, Li et al. showed how the RESYOLO model significantly improves vehicle detection in autonomous driving systems, ensuring high precision in object detection (Li, Y., et al., 2022).

（3）Heritage Conservation

In architectural heritage, Kaamin et al. highlighted the significance of UAVbased YOLOV systems for conducting visual inspections and longterm monitoring, which effectively disseminates structural health information and promotes the preservation of architectural integrity (Kaamin et al., 2017). Liu et al. achieved over 90% precision using YOLOv3 for detecting timbercrack damage in wooden architectural heritage in less than 0.1 seconds, showcasing the algorithm’s efficiency in preserving valuable structures (Liu, Y., et al., 2017).

Shi et al. and Barlindhaug leveraged YOLOV's AI capabilities to aid in metadata creation and ensure retention of contextual integrity vital for historical interpretation (Shi et al., 2023; Barlindhaug, G., 2022). Jadhav and Kurnthekar explored YOLOV’s role in restoration efforts, addressing challenges like pollution, agerelated degradation, and structural flaws (Jadhav, S., and Kurnthekar, B.). Petracek et al. utilized YOLOV’s collaborative aerial autonomy for documenting and monitoring historical structures, enabling precise identification of structural defects such as frame decay and vegetation overgrowth (Petracek, P., et al., 2023). Yang et al. used YOLOv4 to identify five types of damage in historical graybrick buildings, underscoring the model's indispensability in protecting and preserving ancient sites (Yang, X., et al., 2023).

（4）Garden Research

In garden applications, the YOLO algorithm is primarily used for monitoring plant changes. Soeb et al. proposed an AIbased tea disease detection and identification method using YOLOv7 based on 4,000 images collected from tea gardens, expected to significantly support the rapid identification and mitigation of tea diseases (Soeb et al., 2023). Lawrence demonstrated the effectiveness of YOLOV4 and YOLOV5n in detecting Redcockaded Woodpecker cavities in forested environments, achieving high mean Average Precision (mAP) and F1 score (Lawrence, B., 2023). Badgujar highlighted the potential of YOLOV in agriculture for realtime monitoring and object recognition (Badgujar et al., 2024), and Urmashev applied the YOLOV5 architecture in a weed detection system, obtaining good results in classifying lowresolution images of weeds (Urmashev, B., et al., 2024).

While the focus has been predominantly on specific biological species detection, there is a notable gap in studies on the overall design concept and structure of gardens. The design philosophy of Jiangnan traditional gardens requires a comprehensive, integrated approach for a deeper understanding, necessitating more advanced object detection methods for a thorough study and recognition of significant garden elements. These works collectively illustrate the adaptability and technological advancement of the YOLO algorithm, highlighting its efficacy in cultural and architectural heritage preservation, and solidifying its role as an indispensable tool for contemporary research and practical applications.

2.3 Literature Review of Image Depth Analysis

2.3.1 Development of Depth Estimation Technology

The development of depth estimation technology has undergone significant evolution, from early methods relying on traditional computer vision techniques like stereopsis and multiview geometry to contemporary deep learningbased approaches.

Initially, stereo vision used disparity between multiple camera views to estimate depth, but these methods faced challenges related to occlusion and lighting variations (Scharstein & Szeliski, 2002). With the rise of deep learning, convolutional neural networks (CNNs) revolutionized this field. Eigen et al. (2014) introduced a seminal model for depth prediction from a single image, using a multiscale deep network that greatly improved the prediction accuracy compared to traditional methods by using coarsetofine depth estimation. This method became a foundation for many subsequent models.

Recent advancements in deep learning have introduced new architectures like UNet and feature pyramid networks (FPNs), further improving accuracy. Kendall et al. (2017) integrated deep stereo regression in an endtoend learning framework, optimizing both photometric and geometric losses to achieve higher precision. More recently, neural radiance fields (NeRF) have emerged, enabling highquality 3D scene reconstruction from images, marking a significant leap in depth estimation technology (Mildenhall et al., 2020).

As deep learning models become more sophisticated, researchers are focusing on reducing their computational cost without sacrificing accuracy. Techniques like model distillation and lightweight network designs, such as MobileNet (Howard et al., 2017), are being actively explored to ensure that depth estimation models can be deployed in realtime systems like autonomous vehicles.

2.3.2 Application of the Marigold Algorithm

The Marigold algorithm represents an innovative breakthrough in depth estimation, combining both deep learning and traditional multiview geometry approaches. It excels in multimodal data fusion, integrating RGB and point cloud data (e.g., from LiDAR) to improve accuracy in depth estimation, particularly in challenging environments.

One of Marigold's key innovations is its ability to perform selfsupervised learning. Unlike many deep learning models that rely on vast amounts of labeled data, Marigold can predict depth by leveraging photometric consistency loss, significantly reducing the dependency on expensive labeled datasets (Kumar et al., 2024). This selfsupervision framework enables the model to adapt to various environments while maintaining high accuracy.

The algorithm has seen wide applications, particularly in autonomous driving, where it enhances environmental perception by accurately estimating road depth from multimodal inputs, even under difficult weather conditions (Chen et al., 2024). In the field of virtual reality, it has been used to generate more immersive environments through accurate 3D scene reconstruction (Johnson, 2023). Additionally, its multimodal fusion techniques are now being applied in medical imaging for precise 3D organ reconstruction from CT and ultrasound data, and in terrain modeling for environmental monitoring and disaster assessment (Kumar, 2023).

2.4 Conclusion

In this thesis, we explored the application of object detection algorithms within the domain of garden design, with a focus on Jiangnan traditional gardens. Notably, while the YOLO algorithm is extensively utilized in various fields, its use in garden design has primarily been restricted to species detection, often focusing on a singular species. This limitation highlights a significant gap in the current application of advanced detection technologies, which tend to overlook the broader potential for integration within complex garden environments.

The Marigold algorithm, a novel development presented at CVPR 2024, where it was recognized as a Best Paper Award candidate. Despite its promising capabilities in other applications, the Marigold algorithm has not yet been applied to the spatial study of traditional Jiangnan garden heritage. This oversight signals a crucial area for further research and development.

The Marigold algorithm represents a significant advancement in the analysis of the space within traditional Jiangnan gardens. When combined with YOLO algorithms, which are adept at identifying specific targets such as individual species of plants or architectural elements, Marigold extends this capability by deciphering the complex interrelations and layered spatial structures that characterize these gardens. This dual approach allows for a more nuanced understanding of how different elements within the gardens interact and coexist in a harmoniously designed environment.

In conclusion, the integration of Marigold and YOLO algorithms into the study and conservation of Jiangnan gardens promises a revolutionary shift in how we understand, interact with, and preserve these important cultural landscapes. This approach not only deepens our comprehension of their spatial and aesthetic complexities but also enhances our ability to protect and celebrate these gardens as vital components of our global heritage.

Chapter 3: Methodology of Study 1

3.1 YOLOv8 Architecture and Network Structure

YOLOv8 represents a paradigm shift in the progression of the YOLO series of object detection networks, incorporating a sophisticated architectural framework that substantially enhances the detection precision and efficiency. This latest iteration features a dynamic integration of a Feature Pyramid Network with crosslayer connections, which effectively facilitates the seamless blending of multiscale feature information across different layers of the network(Wang et al., 2023). This innovative approach allows for the utilization of attention mechanisms and sophisticated optimization strategies, dramatically improving both the precision and performance of the object detection tasks.

The architectural foundation of YOLOv8 is built around a robust backbone network that undertakes the crucial task of feature extraction from images. This backbone is typically composed of advanced deep convolutional neural network structures such as Darknet or ResNet, which are wellregarded for their deep architectural complexities and efficacies in processing intricate visual data. This is complemented by a detection head that includes a series of convolutional and fully connected layers, meticulously designed to predict bounding boxes and the probability classes of the objects detected with high precision.

In YOLOv8, object detection is ingeniously treated as a regression problem, where the model uses convolutional, pooling, and fully connected layers to predict both the location and class of objects within an image. The convolutional layers of the network employ sliding kernels that traverse the entire input data, extracting pivotal features while capturing essential local spatial structures of the images. The subsequent pooling layers serve a critical role in reducing the dimensionality of the feature maps produced by the convolutional layers. Through maxpooling operations, these layers effectively compress and aggregate the extracted features, significantly reducing the computational demand and the number of parameters required, which concurrently enhances the model’ s translational invariance.

A fully connected layer, strategically placed at the terminal end of the network, transforms the refined feature maps into the final outputs required for effective object detection. YOLOv8 meticulously calibrates the architecture and parameters of its convolutional, pooling, and fully connected layers(Bochkovskiy et al., 2020). It incorporates specialized components such as Anchor Boxes, which standardize the predictions of bounding boxes, Intersection over Union (IoU) thresholds that provide a metric for the precision of these predictions, and NonMaximum Suppression (NMS) techniques that effectively reduce redundancy in detection outputs(X. Zhu et al., 2021).

Furthermore, YOLOv8 incorporates a range of optimization technologies aimed at enhancing the model ’ s overall performance. These include sophisticated data augmentation strategies that increase the diversity and robustness of the training dataset, batch normalization techniques that help stabilize and accelerate the training process, and dropout methods designed to prevent the overfitting of the model to the training data. Each of these components is integrated within the network to synergistically boost its capability to accurately detect objects under various conditions.

The detailed structural components of YOLOv8 (Figure 3-1)are illustrated of the related technical documentation. While maintaining a backbone that resembles that of its predecessor, YOLOv5, YOLOv8 introduces several significant enhancements. Among these is the adjustment of the CSPLayer and the inclusion of the C2f module, which incorporates dual convolutions within the crossstage partial bottleneck. This innovative feature is tailored to merge highlevel features with rich contextual information effectively, thereby significantly enhancing the overall precision of the detection process.

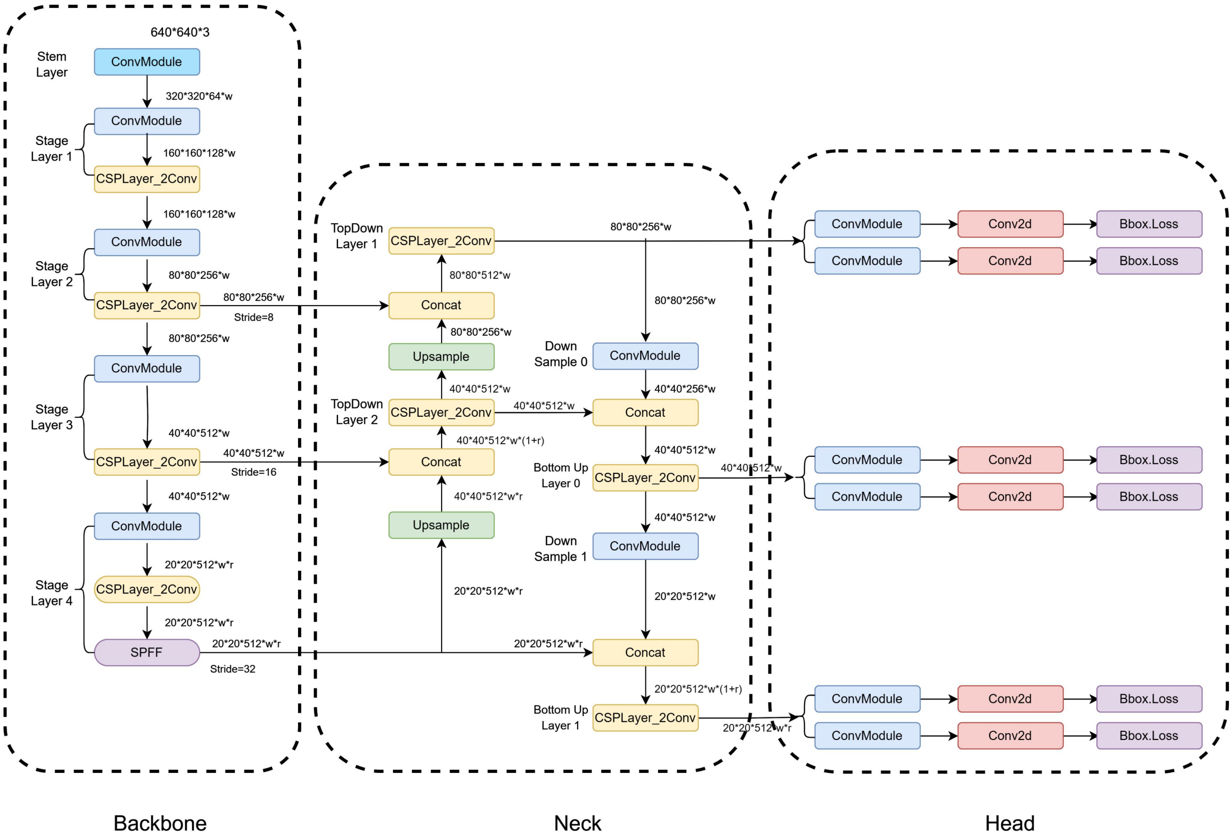


Figure 3-1: Detailed structure of YOLOv8

In summary, YOLOv8 stands as a cuttingedge advancement in the field of object detection, pushing the boundaries of technology with its refined algorithms and architectural innovations. This network not only improves the precision and speed of detection but also enhances the scalability and adaptability of the YOLO series, making it a formidable tool in the arsenal of modern automated visual recognition systems.

3.2 Overview of the Enhanced YOLOv8 Network Design

A modified YOLOv8 algorithm has been proposed in various research papers to enhance fault detection in smart additive manufacturing(Karna et al., 2023), automate tomato detection in agriculture, and enhance fire detection for early disaster prevention.

In the field of UAV multitarget detection, a model incorporating BiPANFPN for improved feature fusion, GhostblockV2 for reduced parameter count, and WiseIoU loss for bounding box regression has been proposed(Yiting, Li., et al., 2021), enhancing detection performance. For cardiac arrhythmia detection, a custom YOLOv8 model finetuned on ECG signals achieved high precision and realtime monitoring capabilities. In waste sorting, YOLOv8 demonstrated superior performance in automating waste classification, enhancing efficiency and safety in waste treatment processes. Additionally, an improved YOLOv8 algorithm for small target detection on flapping wing drones utilized a small object detection layer and Multihead selfattention, significantly improving mAP indicators on relevant datasets(Ma et al., 2023). These modifications showcase the versatility and effectiveness of adapting YOLOv8 for various specialized tasks.

These modifications include additional feature extraction layers, improved precision in complex environments, depthwise separable convolution techniques, attention gate modules, and feature enhancement modules. The proposed algorithms have shown significant improvements in detection performance, precision, recall rates, and mean average precision (mAP) while also reducing model size and maintaining realtime detection capabilities. The modifications aim to strike a balance between detection precision, model complexity, and realworld applicability across various domains, showcasing the versatility and effectiveness of the enhanced YOLOv8 models.

3.3 The Improvements of YOLOv8

3.3.1 Diverse Branch Block (DBB)

The Diverse Branch Block (DBB) represents a significant enhancement in convolutional network architecture, specifically tailored for improving the performance and precision of the model it integrates into. By innovating upon the conventional bottleneck structure found in the C2f layer, the DBB replaces the standard convolutional operations with a multibranch setup that incorporates varying receptive fields and complexities. This multibranch structure is heavily inspired by the inception architecture, known for its efficacy in handling different scales of input and extracting more nuanced features from the image data.

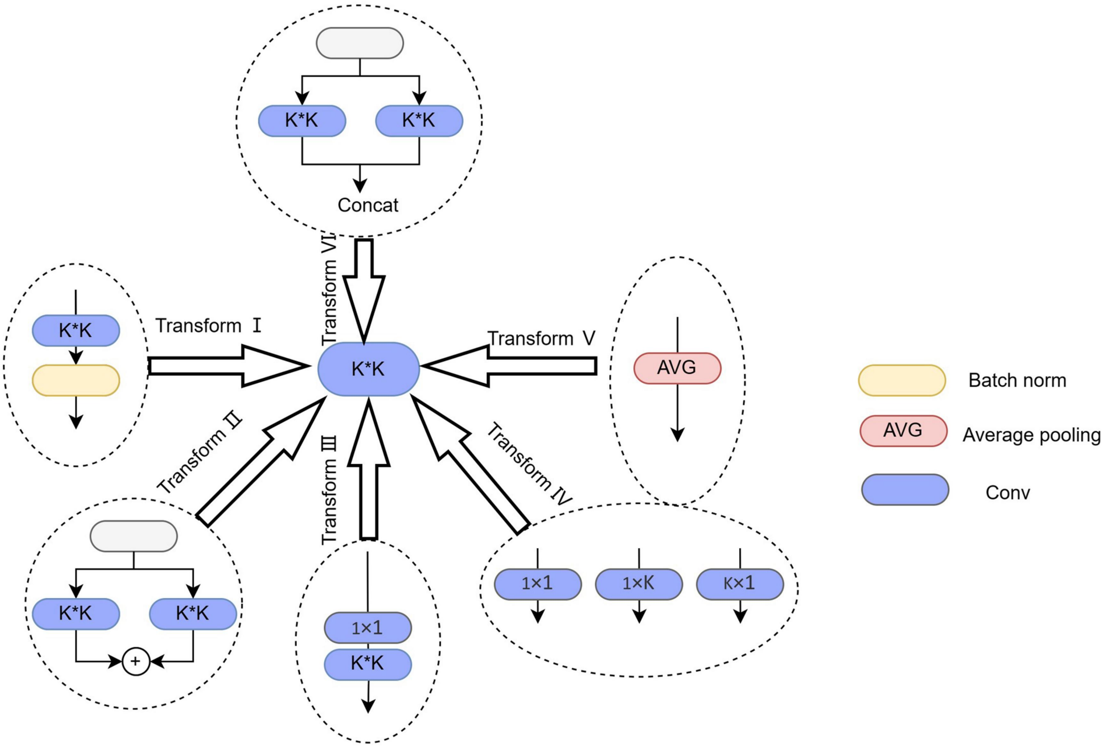


Figure 3-2: Six transformations to implement an inferencetime DBB by a regular convolutional layer

In detail, the DBB integrates a combination of multiscale convolutional processes, a sequence of 1 × 1 followed by K × K convolutions, average pooling, and a novel branch addition methodology. Each element of this configuration contributes uniquely to the model's capabilities. The multiscale convolutions allow the DBB to process inputs of varying sizes effectively, capturing details that might be missed by a single scale approach. Sequential convolutions, where a 1 × 1 convolution is followed by a K × K convolution, help in refining the feature map by first reducing the dimensionality before capturing more complex features. Average pooling reduces feature dimension while preserving essential information, and the branch addition strategy enhances the depth and robustness of the feature extraction process.

This advanced architectural design of the DBB enriches the feature space significantly. By providing multiple pathways and complexity levels through its diverse branches, it ensures a more comprehensive analysis of the input data, leading to a marked improvement in feature extraction capability. This complexity allows the model to discern and detect objects with higher precision, leveraging the enriched feature maps generated by the DBB.

Moreover, the adaptability of the DBB design plays a critical role in its practical application. While the DBB enhances the model's precision during the training phase by introducing a richer and more complex network structure, it also offers a unique advantage during the inference phase. Specifically, the DBB can be equivalently converted into a single convolution operation. This conversion capability is facilitated through six structural reparameterization methods, as depicted in Figure 3-2. These methods enable the transformation of the DBB back into a standard K×K convolution during realtime applications, thereby maintaining the enhanced detection precision achieved during training without incurring additional computational costs or complexity.

This feature ensures that the enhancements provided by the DBB during the training phase do not lead to increased inference times or resource demands during practical deployment. The ability to convert complex structures into simpler, computationally efficient formats during inference means that models equipped with the DBB can operate in realtime environments effectively, maintaining high precision without the burden of increased processing requirements. Thus, the DBB not only elevates the detection capabilities of models during development and training but also ensures these improvements are realistically sustainable and beneficial in everyday applications, making it an invaluable addition to modern convolutional network architectures.

3.3.2 Bidirectional Feature Pyramid Network (BiFPN)

Early detectors typically made predictions directly based on the pyramid feature hierarchy extracted from the backbone network. The Feature Pyramid Network (FPN) introduced a topdown approach to combine multiscale features. Based on FPN, the PANet adds an additional bottomup path aggregation network on top of FPN; NASFPN utilizes neural architecture search to automatically design the feature network topology. Although it achieves better performance, NASFPN requires thousands of GPU hours during the search process, and the generated feature network topology is irregular, making it difficult to interpret. However, BiFPN introduces learnable weights to gauge the importance of different input features while repeatedly applying both topdown and bottomup multiscale feature fusion.

The different detectors are illustrated as follows:(Figure 3-3)

(a) FPN introduces a topdown pathway to fuse multiscale features from levels 3 to 7 (P3 P7);

(b) PANet adds an additional bottomup pathway on top of FPN;

(c) NASFPN uses neural architecture search to find an irregular feature network topology, then repeatedly applies the same blocks;

(d) BiFPN features bidirectional crossscale connections and weighted feature fusion, offering a better tradeoff between precision and efficiency.

The integration of Bidirectional Feature Pyramid Network (BiFPN) into the neck layer of YOLOv8 represents a substantial enhancement in feature fusion capabilities, primarily aimed at boosting the model's efficiency and precision in object detection and semantic segmentation. BiFPN optimizes feature fusion by simplifying the network structure through the elimination of singleinput nodes, enhancing connectivity among the same level nodes to improve feature integration without increasing computational load, and utilizing a repetitive, bidirectional path design to deepen feature fusion adaptively across multiple layers. These improvements ensure that YOLOv8 with BiFPN not only performs more efficiently but also achieves greater precision in processing complex visual data.

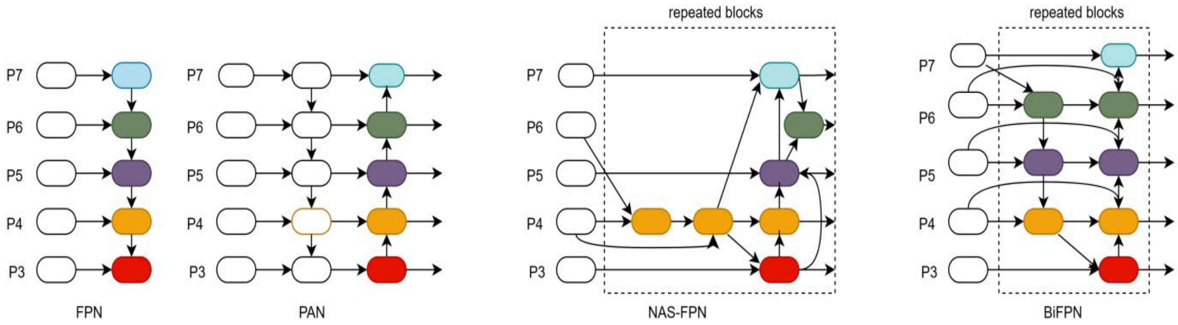


Figure 3-3: the framewoek of the BiFPN

The Bidirectional Feature Pyramid Network (BiFPN) represents a significant advancement in the architecture of feature fusion networks for computer vision tasks such as object detection and semantic segmentation. This optimized structure is specifically engineered to enhance both the efficiency and precision of feature fusion, thereby offering superior performance compared to traditional feature pyramid networks.

（1） Simplification of the Network Structure: Traditional feature pyramid networks often incorporate multiple nodes that process and transmit features through the network. However, some of these nodes may have only a single input, leading to redundant processing and inefficient use of computational resources. The BiFPN addresses this inefficiency by strategically removing such nodes, which reduces the complexity of the network architecture and minimizes unnecessary computational overhead. This streamlined structure not only speeds up the feature processing but also reduces the memory footprint, making the network more scalable and easier to deploy on devices with limited resources.

（2） Enhanced Feature Fusion: In traditional models, feature fusion typically occurs in a limited and sequential manner, which can restrict the thorough integration of contextual information across different scales. The BiFPN enhances this aspect by introducing additional connections when the input and output nodes reside at the same hierarchical level. These added connections promote a more comprehensive and effective fusion of feature maps, leveraging information from multiple scales without significantly increasing the computational cost. This results in a richer feature representation that is beneficial for capturing complex patterns and details in images, thereby improving the precision and robustness of the model.

（3） Reusing Feature Network Layers: Another key optimization in the BiFPN is the reuse of feature network layers across different paths. Each bidirectional path is treated as a separate layer within the network, and these layers can be repeated multiple times to deepen the feature fusion process. This approach not only increases the depth of the network without substantially increasing complexity but also offers greater flexibility in customizing the network architecture to better fit specific computational and resource constraints. By reusing layers, the BiFPN efficiently recycles computational resources to enhance feature extraction and integration, providing a scalable solution that adapts well to varying levels of computational availability.

Through these strategic optimizations, the BiFPN significantly enhances the performance of models on complex computer vision tasks. It does so by ensuring more efficient computation and better memory management while achieving deeper and more effective feature fusion. The result is a network capable of delivering higher precision and improved detection and segmentation results in realworld applications, making it a valuable tool for advancing the field of computer vision.

3.3.3 Dynamic Head Modules (DyHead)

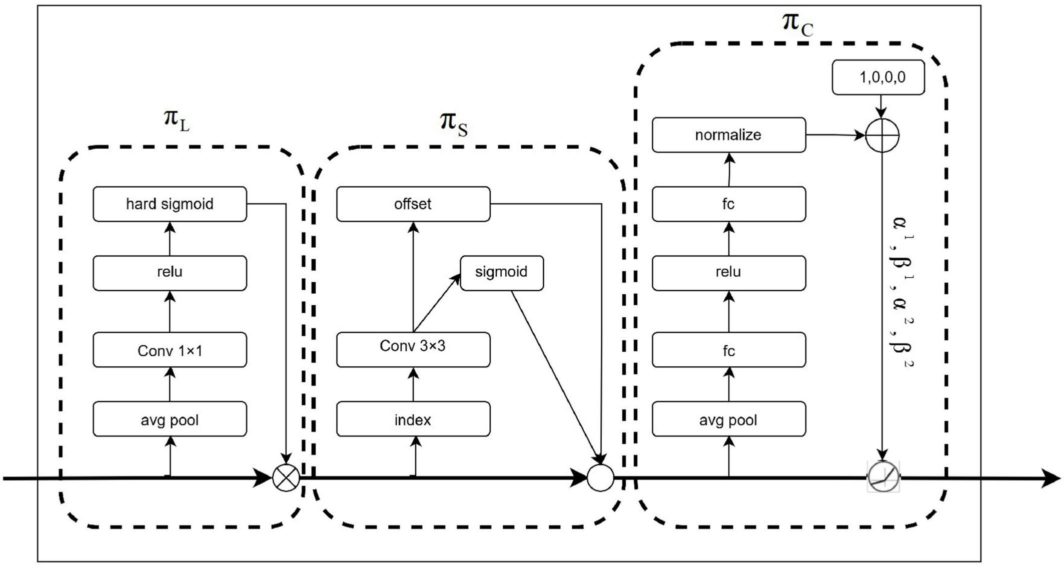


Figure 3-4: Dynamic head (Dyhead) framework

The new dynamic head framework (DyHead) (Figure 3-4) employed for the object detection head incorporates a comprehensive attention mechanism that skillfully amalgamates different forms of selfattention — namely, scale attention, spatial attention, and task attention. These mechanisms are applied coherently across varying feature levels, positions, and channel outputs, which are crucial for adapting to the dynamic nature of object detection tasks (Figure 3-5).

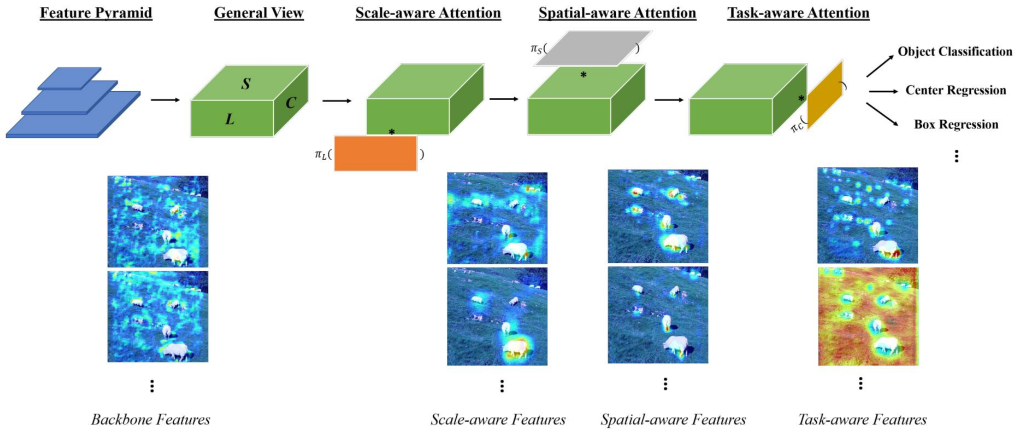


Figure 3-5: The Dynamic head (Dyhead) framwork (Dai, X et al.,2021)

（1） Scale Attention: This component of the DyHead framework focuses on scaleaware processing. In object detection, recognizing objects at different scales is crucial, especially in images with diverse object sizes. The scale attention mechanism adapts the neural network's focus across multiple scales, allowing it to pay more attention to relevant features at the appropriate scales. This adaptability is particularly beneficial in scenarios where the scale of objects varies significantly, helping to maintain high detection precision across all sizes.

（2） Spatial Attention with Deformable Convolution V2: The spatial attention component incorporates advanced features like Deformable Convolution V2, which includes capabilities for offset and feature amplitude modulation. This allows the attention mechanism to adapt to the geometric variations within the image. By focusing on regions of the image where the object contours and shapes are more complex or unusual, the network can better localize and recognize objects even in challenging spatial contexts. This adaptability is essential for accurate detection in images with complex backgrounds or overlapping objects.

（3）Task Attention and Channel Modeling: On the channel attention front, DyHead uses two fully connected neural networks to model the channelwise relationships within the feature maps. This type of attention assesses the importance of different channels in contributing to the taskspecific outputs, such as distinguishing between classes or recognizing subtle features of objects. By optimizing which features to enhance or suppress, the network becomes more efficient in focusing on relevant attributes, thus improving the precision of the detection.

π_L represents scaleaware attention. π_S represents spatial attention, employing Deformable Convolution V2 (including offset and feature amplitude modulation). π_C represents channel attention and channel modeling through two fully connected neural networks.

After implementing the DyHead mechanism, the system generates a pair of directionaware feature maps. These maps are designed to complement the input features effectively, thus enriching the object representations within the detection framework. The resulting visualizations, as illustrated in Figure 3-6, show clear distinctions between the input image and the processed heatmaps. In these heatmaps, the red areas highlight the features deemed crucial by the network—typically parts of the image containing objects of interest—while the blue areas denote background elements that are less relevant to the task. This distinction is evident in both simple and complex background scenarios, demonstrating the framework's ability to focus and differentiate essential details from noise.

The integration of these sophisticated attention mechanisms within the DyHead framework greatly enhances the representational capability of the object detection head, leading to more accurate and robust detections across varied and challenging visual environments. This innovative approach not only optimizes the performance in standard detection tasks but also provides a scalable solution that can adapt to complex, realworld scenarios where traditional methods might struggle.

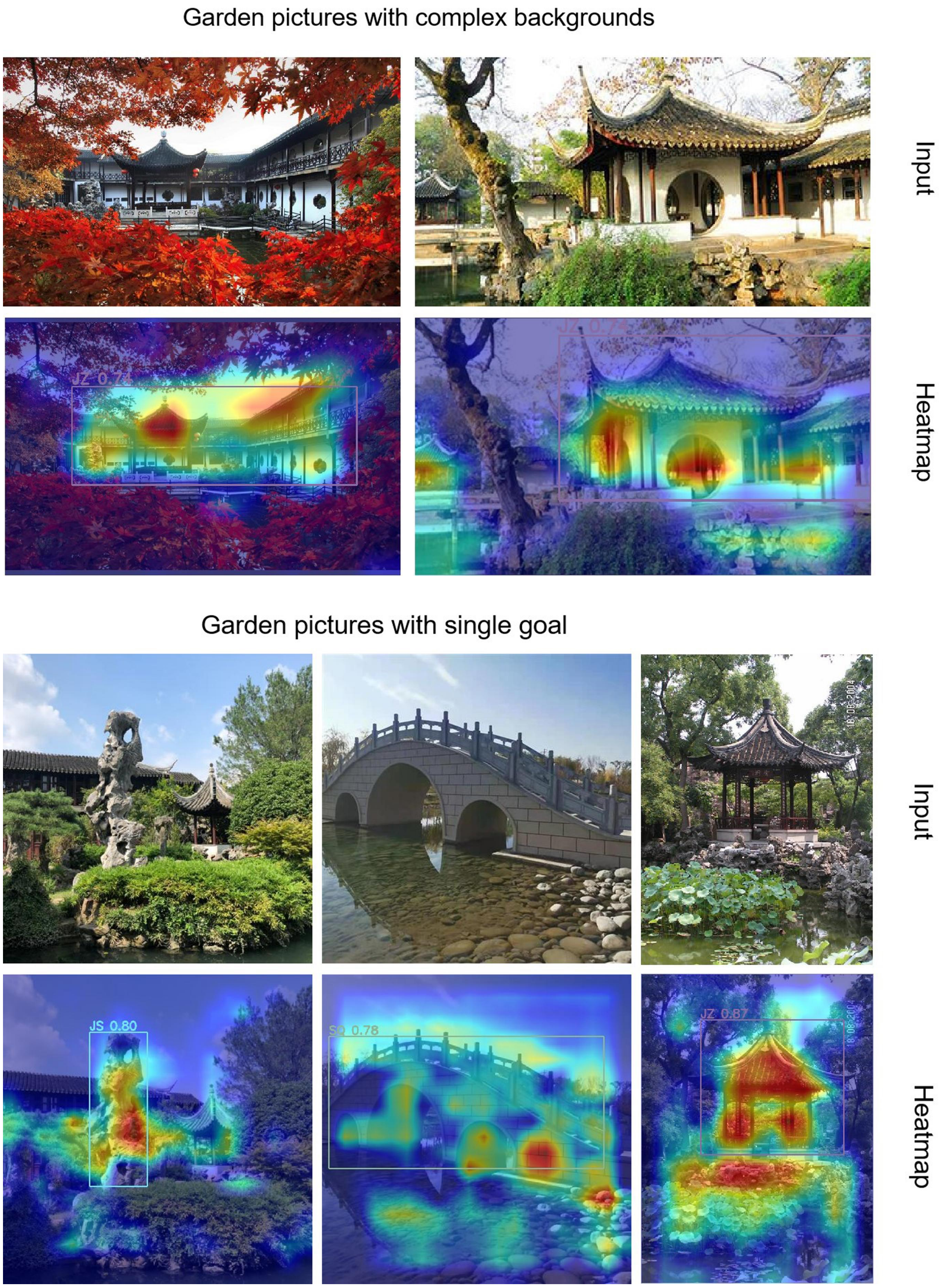

Garden pictures with complex backgrounds
Input
JZ 0.74
JZ 0.74
Heatmap
Garden pictures with single goal
Input
JS 0.80
SQ 0.78
JZ 0.87
Heatmap

Figure 3-6: Attention heat map after applying the DyHead mechanism

After applying the DyHead mechanism, we obtain a pair of directionaware feature maps. These maps can act complementarily on the input features, thereby enhancing the object representation. As shown in Figure 3-6, the top is the input image, while the bottom is the heatmap after attention mechanism processing. The left image shows a pure background, whereas the right image shows a complex background. The red part of the heatmap represents the information that the neural network considers to be more important, whereas the blue part represents background information unrelated to classification.

3.4 Dataset Compilation and Characteristics

3.4.1 Image Collection

The construction of a specialized largescale dataset dedicated to Jiangnan traditional gardens is a pivotal step toward advancing the research and understanding of these culturally significant landscapes. Recognizing the gap due to the absence of a public dataset, a meticulous approach was taken to compile a comprehensive collection of images that capture the essence and diversity of Jiangnan traditional gardens.

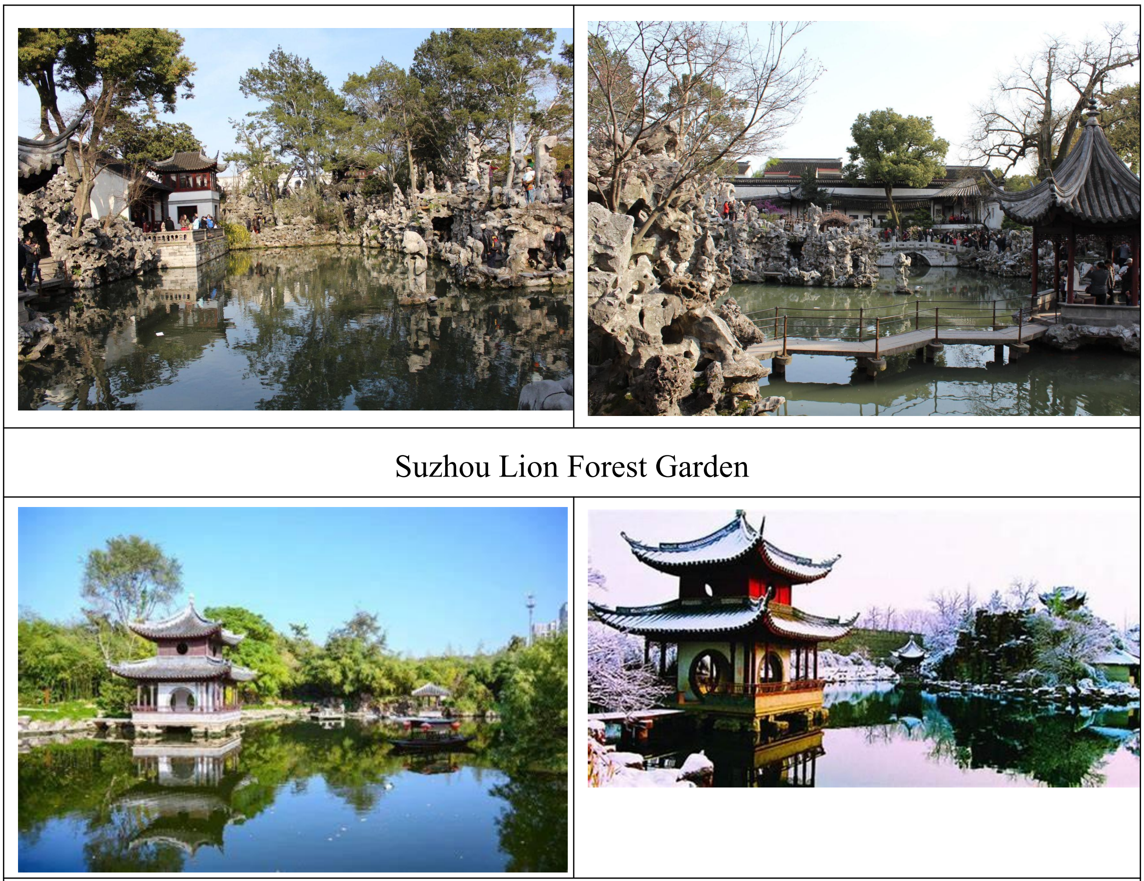
The dataset began with the collection of 3280 images sourced from various online platforms, selected for their representation of the quintessential elements of Jiangnan traditional gardens. This initial pool includes images from iconic gardens such as the "Four Famous Gardens": Nanjing Zhan Garden, Suzhou Liu Garden, Suzhou Zhuozheng Garden, and Wuxi Jichang Garden(Figure3-7).These gardens are renowned for their historical significance and unique aesthetic features, making them crucial subjects of study. The dataset is further enriched by including images from other notable gardens such as Shanghai Yu Garden, which is famous for its exquisite architectural details, and Yangzhou Slender West Lake, known for its scenic beauty and sweeping water vistas. Additional gardens like Ge Garden, He Garden, Suzhou Canglang Pavilion, Lion Forest Garden, and Nantong Shuihui Garden are also represented, ensuring a broad spectrum of Jiangnan garden styles and features.

Suzhou Liu Garden

Nanjing Zhan Garden

Suzhou Zhuozheng Garden

Wuxi Jichang Garden	
Yangzhou Ge Garden	
Yangzhou He Garden	
Suzhou Canglang Pavilion	



Suzhou Lion Forest Garden

Nantong Shuihui Garden

Figure 3-7: Examples of traditional traditional gardens in Jiangnan

To enhance the dataset's depth and authenticity, 1610 photos were personally captured in various Jiangnan traditional gardens. This effort was aimed at gathering unique perspectives and detailed visual information that may not be present in publicly available images. These personally captured photos include seasonal variations, different times of the day, and lessdocumented features or sections of the gardens, providing richer data for analysis.

3.4.2 Classification of Garden Elements

To effectively apply object detection algorithms in the recognition and classification of garden elements within Jiangnan traditional gardens, a welldefined classification system is crucial. The complex nature of traditional garden layouts and designs, particularly in garden architectures, presents a significant challenge due to the diversity in forms and styles these elements can exhibit. For instance, pavilions, which represent a specific type of garden architecture, exhibit a variety of structures and styles that could be classified according to multiple characteristics such as plans, roofs, and walls.

（1） Challenges in Detailed Classification:

Garden architectures, like pavilions, halls, and palaces, can vary greatly in their designs, encompassing a wide range of structural details and aesthetic nuances.

A highly granular classification approach, while academically sound, may lead to the practical issue of having insufficient sample sizes for each category, making it difficult for machine learning models to learn effectively. The diversity of styles within a single category can further complicate the detection process, as the algorithm may struggle to generalize across such varied examples.

（2）Proposed Classification Method:

To mitigate the issue of small sample sizes and to streamline the training process of object detection algorithms, a broader categorization strategy is employed.

Based on the book by Peng, Yi Gang, "Analysis of Chinese Classical Gardens" published in 1986 by China Architecture & Building, the elements of garden construction in Chinese classical gardens are categorized as follows (Table 3-1):

Table 3-1：Classification of Garden Elements

Major Category	Subcategory
Architectural Structures	Pavilions
	Towers
	Halls
	Corridors
Water Features	Ponds
	Streams
	Waterfalls
Rock Formations	Artificial mountains
	Rockeries
Plant Elements	Trees

	Flowers
	Bamboo
	Grass
Bridgestyle Architectural Features	Sculptures
	Bridges
	Lanterns
	Calligraphy and paintings
Functional Areas	Leisure areas
	Viewing spots
	Meditation spaces

This involves grouping garden elements into major classes that are sufficiently distinct yet encompass a variety of subtypes and styles within each category.

In our preliminary experiments, we found that water features are challenging to detect due to their reflective and transparent properties. Additionally, "Functional Areas" do not correspond to specific targets but are rather regions.Therefore, we have decided to exclude the two categories, Functional Areas and Water, from our classification. Currently, water is undetectable, which represents a limitation of our research.

（3）Detailed Categorization of Jiangnan traditional gardens:

Architecture (JZ): This category includes all major built structures such as halls, pavilions, and palaces. These structures are primarily used for dwellings or entertaining guests and can vary from simple singleroom structures to elaborate multilevel complexes. This broad category allows for the inclusion of a variety of architectural forms under a single label, facilitating easier data collection and model training.

Stone Bridges (SQ): Labeled as SQ, this class captures the variety of bridge forms found in these gardens, including curved, straight, crescent, moonviewing, and dragoncrossing bridges. Each type represents unique design elements and structural challenges, reflecting the artistry and functional considerations in traditional Chinese bridge construction.

Rockeries (JS): The JS category is subdivided into types like peak rock, waterstone, cave, and peculiar stone. Each subtype reflects different aspects of rockery design, from the simulation of natural landscapes with rocks and water to the creation of intricate cavelike structures that offer visitors a dynamic experience.

Plants (ZW): This classification encompasses all plant life within the gardens, including bonsais, flowers, vines, grasslands, and aquatic plants. It recognizes the critical role that flora plays in the aesthetics and ecological balance of Jiangnan gardens. This category is particularly diverse, covering a range of plant types and arrangements that are integral to the garden's design.

This classification framework not only respects the complexity and diversity of Jiangnan garden elements but also ensures that each category is robust enough to provide ample learning material for object detection algorithms. By adopting this approach, the potential of advanced machine learning techniques can be fully harnessed to enhance the understanding and preservation of these cultural landscapes, while also providing a structured way to analyze and interact with the rich data captured in the Jiangnan garden dataset.

3.4.3 Data Preprocessing and Augmentation Methods

The creation of a robust dataset for object detection in Jiangnan traditional gardens involves several crucial preprocessing and augmentation steps to enhance the model's precision and generalization capabilities. Given the intricate and diverse elements present in such gardens, a strategic approach to data preparation is necessary to ensure the effectiveness of the object detection algorithms.

（1）Data Augmentation Techniques:

Augmentation techniques were employed to artificially expand the diversity of the dataset, which is essential for training robust models capable of functioning accurately across realworld varying conditions(Figure 3-8):

Mirror Flipping: 400 images were mirrorflipped horizontally. This technique helps in reducing the model’s dependency on the orientation of objects, thereby enhancing its ability to recognize symmetrical structures irrespective of their directional orientation.

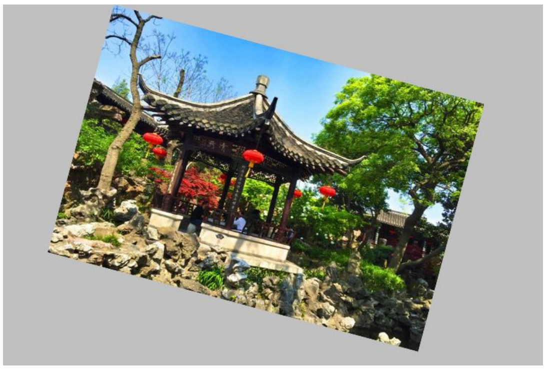
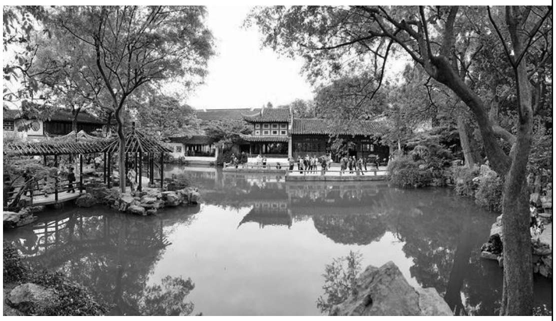
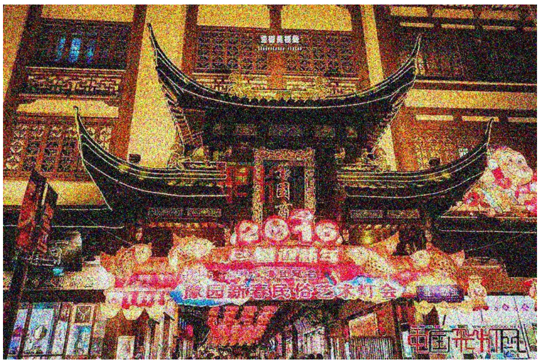
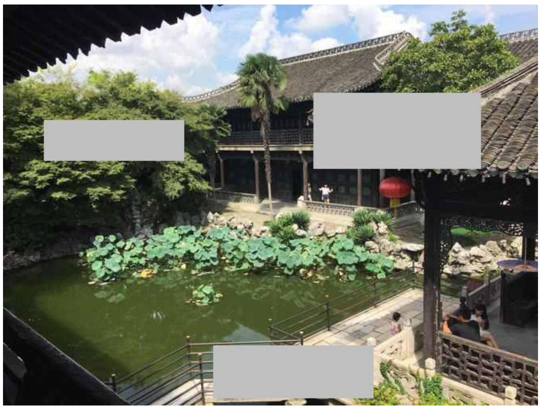
Random Flipping: 400 images underwent random flipping. This method increases the dataset's variance, enabling the model to learn to recognize objects from multiple angles and perspectives, thus improving its versatility.

Grayscale Conversion: Converting 400 images to grayscale simulates varying lighting conditions, from bright daylight to overcast or shaded environments, making the model more adaptive to changes in illumination.

Gaussian Noise Addition: Adding random Gaussian noise to 400 images helps the model train on data that might mimic the quality of older or lower resolution photographs, a common scenario when dealing with historical archives.

Random Occlusion: To further challenge the model, 400 images were randomly occluded. This step involves artificially blocking parts of objects to simulate scenarios where objects are partially obscured or overlap, training the model to recognize objects even under partially visible conditions.

The image before Mirror Flipping	The image after Mirror Flipping

	
The image before Random Flipping	The image after Random Flipping
	
The image before Grayscale Conversion	The image after Grayscale Conversion
	
The image before Gaussian Noise Addition	The image after Gaussian Noise Addition
	

The image before Random Occlusion	The image after Random Occlusion

Figure 3-8: Data Augmentation Techniques

（2） Image Preprocessing:

All images were resized to uniform dimensions of 640x640 pixels. This standardization is crucial as it ensures that the input to the neural network is consistent, which helps in optimizing the computational resources and enhances the network's ability to learn from different images effectively.

The pixel values of the images were normalized to a range between 0 and 1. This normalization not only helps in speeding up the convergence during training but also reduces the likelihood of model overfitting by smoothing out pixel intensity variations across the dataset.

（3） Dataset Division Using Holdout Method:

The dataset was meticulously divided using the holdout method to ensure that the model is tested and validated under unbiased conditions. The division ratios were set at 80% for training, 10% for testing, and 10% for validation. This resulted in 5512 training images, 689 test images, and 689 validation images. Such a distribution allows for comprehensive training while ensuring sufficient data for rigorous testing and validation of the model's performance.

（4） Consistency in Annotation:

Throughout the preprocessing and augmentation processes, maintaining the consistency and precision of the annotation data is paramount. This ensures that regardless of the transformations applied to the images, the position and class information of the bounding boxes are accurately retained relative to the modifications. Each annotation is carefully adjusted to reflect the transformations, whether it be resizing, cropping, or any form of augmentation, to maintain the integrity and relevance of the training data.

These comprehensive steps in dataset preparation not only enhance the training effectiveness but also ensure that the model developed is robust, accurate, and capable of performing well in diverse and challenging realworld conditions specific to Jiangnan traditional gardens.

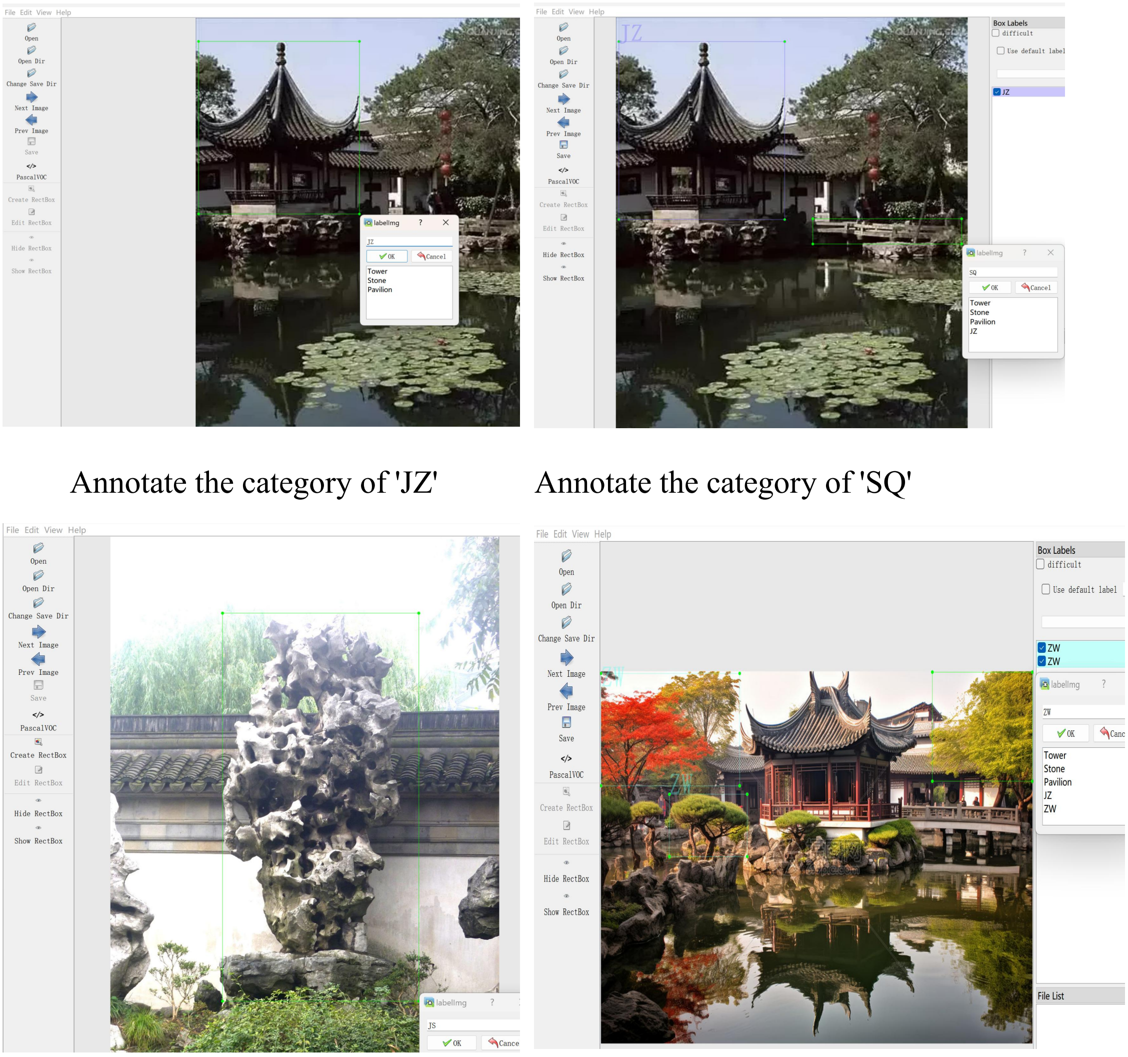
3.4.4 Dataset Annotation

Accurate dataset annotation is a cornerstone in the development of robust object detection systems, particularly when applied to specific architectural and natural features such as those found in Jiangnan traditional gardens. For the Jiangnan garden dataset, a meticulous annotation process was implemented using rectangular bounding boxes to delineate each object of interest within the images. Corresponding class labels were carefully assigned to each bounding box to facilitate precise object detection and classification.

（1） Annotation Process and Techniques:

Each image was annotated with rectangular bounding boxes, which are universally recognized for their simplicity and effectiveness in enclosing object extents. These boxes were drawn to ensure they tightly encompass the object, minimizing the inclusion of background or irrelevant features, which is crucial for highquality object detection training(Figure 3-9).

Objects within the images were classified into distinct categories based on their relevance and characteristics: 'JZ' for architectural elements, 'SQ' for stone bridges, 'JS' for rockeries, and 'ZW' for plants. These labels facilitate targeted analyses and applications, such as architectural studies or botanical research within the garden settings.



Annotate the category of 'JZ'

Annotate the category of 'SQ'

Annotate the category of 'JS'

Annotate the category of 'ZW'

Figure 3-9: Annotation Process

（2） Tools and Formats for Annotation:

The annotation process was carried out using the LabelImg tool, a popular choice for manually annotating object detection datasets. This tool provides a userfriendly interface for drawing bounding boxes and assigning labels, thereby enhancing the efficiency and precision of the manual annotation process.

For each annotated image, corresponding annotation files were created in YOLO format.In YOLO format, each line of the label file contains crucial information about a target, including its class identifier and the coordinates of its bounding box. This format is preferred for its simplicity and direct compatibility with YOLO (You Only

Look Once), a popular realtime object detection system.

（3） Attention to Detail in Annotation:

Special attention was paid to common issues such as vegetation occlusion, where plants might cover parts of architectural elements, and architectural segmentation, necessary when buildings of multiple categories appear interconnected or overlaid. These challenges require careful judgment to ensure annotations are as accurate as possible, aiding in the development of a dataset that truly reflects the complex realities of the gardens.

Buildings or structures that encompass elements from multiple categories were annotated separately for each distinct feature. This meticulous detail in annotation helps in more nuanced understanding and detection of hybrid structures commonly found in Jiangnan gardens, such as pavilions with integrated rockeries.

3.5 The Computer Configuration and Parameter Settings of the Model Training

In our comprehensive study of object detection within Jiangnan traditional gardens, we leveraged the advanced capabilities of the YOLOv8 model, specifically version 0.114 developed by Ultralytics. This choice was predicated on YOLOv8's reputation for high performance in realtime object detection scenarios, making it wellsuited for the detailed and dynamic environments presented by these gardens.

3.5.1 Experimental Setup and Model Configuration

We used Python 3.9.0 for its robust library ecosystem and compatibility with machine learning frameworks. VScode (1.76.0) served as the integrated development environment (IDE), providing a powerful and userfriendly interface for coding. The choice of CUDA 10.2 was crucial for leveraging the parallel processing capabilities of GPUs, essential for training deep learning models efficiently.

All computational tasks were performed on an NVIDIA TITAN V (12 GB), a choice driven by its excellent computational power and memory capacity, ideal for handling large datasets and intensive computing tasks required by deep learning models.

3.5.2 Training Details

（1）Optimizer Configuration:

The optimizer was set to 'auto', allowing the system to choose adaptively between SGD and Adam based on the specific gradient descent characteristics of the task at hand. This flexibility is beneficial for optimizing training under varying data conditions and model architectures.

（2）Learning Rates and Decay:

We started with an SGD initial learning rate of 1E2 and an Adam initial learning rate of 1E3. Weight decay was set at 5E4 to help prevent overfitting by penalizing larger weights. The momentum factor was fixed at 0.937, and three warmup stages with a momentum of 0.8 were utilized to stabilize the learning rates at the beginning of training.

（3）Learning Rate Scheduling:

The cosine annealing method was employed to decay the learning rate over the course of 500 epochs, facilitating finer adjustments in learning as the training progressed. This method helps in avoiding local minima and ensures more stable convergence.

（4）Batch Size and Training Duration:

The batch size was set at 16, balancing between computational efficiency and memory usage, allowing for a comprehensive learning process across epochs. The entire modeltraining process was completed in approximately 6.5 hours, demonstrating the efficiency of our setup.

（5） Model Size and Comparative Analysis:

The YOLOv8n model size was chosen for its optimal balance between speed and precision. This size is particularly suited for realtime applications where both performance metrics are critical. This choice ensures that the model is not overly demanding on computational resources, making it applicable for realtime processing.

To maintain fairness in the evaluation, all comparative models were also selected to be of the 'N' size. This standardization is critical to ensure that differences in performance metrics are attributable to model efficacy and not variations in computational complexity or capacity.

3.6 Evaluation Metrics

To rigorously assess the performance of the modified YOLOv8 model, dubbed YOLOv8modify, on our meticulously curated dataset of Jiangnan traditional gardens, we employed a comprehensive set of evaluation metrics. These metrics are critical in determining the model′s precision, efficiency, and overall capability in detecting and classifying diverse garden elements.

3.6.1 Evaluation Metrics Employed:

For model evaluation, images from the test set are input into the trained model to obtain prediction results. These predictions are then compared with the true labels to calculate various metrics.

In object detection tasks, a predicted bounding box is considered a true positive (TP) if the Intersection over Union (IOU) with a ground truth box exceeds a certain threshold; otherwise, it's considered a false positive (FP). False negatives (FN) are actual objects that the model failed to detect. The study employs several commonly used evaluation metrics in the field of object detection to assess the performance of the algorithm model. The primary metrics include Precision (P), Recall (R), Average Precision (AP), F1 Score, and Mean Average Precision (mAP).

（1） Precision: This metric indicates the proportion of correctly identified positive instances among all instances that the model classified as positive. High precision reflects the model's effectiveness in minimizing false positives, crucial for ensuring that only relevant objects are detected, avoiding erroneous identifications which can skew further analysis.

$$P = \frac{TP}{TP + FP}$$

（2）Recall: Recall measures the proportion of actual positive instances that the model correctly identified. This metric is particularly important in scenarios where missing an object can be more detrimental than mistakenly labeling a nonobject. High recall in our context ensures that the model is capable of detecting most, if not all, relevant objects in the garden scenes, a key factor for comprehensive analysis.

$$R = \frac{TP}{TP + FN}$$

（3）F1 Score: The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both the model′s precision and recall. An excellent F1 score indicates not only the model's precision but also its robustness, representing both aspects equally without favoring one over the other.

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

（4）Mean Average Precision (mAP): The mAP is a popular metric in object detection tasks that aggregates the average precisions across all classes at different recall levels, providing a holistic view of the model's performance across multiple object categories. It is especially useful in our dataset, which includes diverse elements such as architecture, plants, stone bridges, and rockeries.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} \int_{0}^{1} P(R)\mathrm{d}R$$

（5） Confusion Matrix: This visual tool helps in understanding the types of errors made by the model. By displaying the number of correct and incorrect predictions broken down by class, the confusion matrix provides insights into which categories are most problematic or most successfully detected, guiding further refinements in the model.

（6） FPS, or Frames Per Second, is a measure used to gauge how smoothly a video plays back or how quickly images are processed. In object detection, FPS shows the number of image frames that the detection network can handle each second. The higher the FPS, the quicker the system processes images, which is especially beneficial for real-time applications.

3.6.2 Threshold Settings for Model Evaluation:

The IOU threshold was set at 0.7, meaning the model needs to have a 70% overlap between the predicted bounding box and the ground truth to be considered a correct detection. This stringent threshold ensures high spatial precision in the model′s predictions, important for precise localization and sizing of detected objects.

A confidence threshold of 0.25 was established to filter out detections with lower prediction certainty. This threshold helps to balance between missing true positives (higher threshold) and increasing false positives (lower threshold), optimizing the model for practical application where some level of noise is tolerable.

These metrics and thresholds were chosen to provide a robust framework for impartially evaluating the experimental outcomes. By using these rigorous standards, we aim to accurately quantify the enhancements made in the YOLOv8n-modify model, ensuring that it meets the high demands of object detection within the rich and complex environments of Jiangnan traditional gardens. This thorough evaluation will help confirm the model's efficacy in realworld applications and guide future improvements.

Chapter 4: Methodology of Study 2

4.1 Data Collection

This study utilized a previously created dataset. The photos include seasonal changes, different times of the day, and less-documented features or areas within the garden, providing a richer dataset for analysis.

4.2 Marigold Depth Estimation Model

4.2.1 Selection and Rationale for the Depth Estimation Model

The choice of the Marigold depth estimation model for analyzing traditional Jiangnan gardens stems from its robust capability to handle complex and layered natural environments. The Marigold model, initially developed for diverse and intricate settings, is particularly suited to gardens due to its advanced algorithms that proficiently handle overlapping structures, varying textures, and the subtleties of natural light variations.

This model leverages a deep learning framework that is adept at discerning subtle variations in depth from highresolution images, making it an optimal choice for the detailed and nuanced landscapes found in Jiangnan gardens. Its ability to generate accurate depth maps from single images allows for detailed spatial analysis without the need for multiple sensor inputs, simplifying data acquisition and processing.

The selection of the Marigold depth estimation model for the analysis of traditional Jiangnan gardens is predicated on its formidable capacity to interpret and process the intricacies and multilayered characteristics of natural environments. This model, originally designed to tackle complex settings, has demonstrated remarkable effectiveness in garden landscapes, adapting seamlessly to their unique challenges. These include handling overlapping structures, managing diverse textures, and accommodating the subtle nuances introduced by natural lighting variations.

Depth Estimation in Complex Environments

Jiangnan gardens are characterized by their sophisticated artistry, involving meticulously arranged flora, serpentine water features, and delicately positioned architectural elements. These gardens are not merely areas of aesthetic and cultural value; they represent a complex interplay of history, art, and nature. The Marigold model is engineered to capture this complexity through its sophisticated algorithms that excel in detecting and delineating the depth of overlapping structures. For instance, where branches of an ancient pine might overlay a distant rockery, Marigold can accurately segment and assign correct depth values to each, despite their visual intertwinement.

Advanced Algorithmic Processing

The strength of the Marigold model lies in its advanced algorithmic framework that enables it to handle varied textures efficiently. In the context of Jiangnan gardens, where the texture varies significantly from the rough bark of ancient trees to the smooth, reflective surfaces of lakes, Marigold's texture handling capabilities ensure that each element is distinctly recognized and accurately processed. This texture differentiation is crucial for generating highfidelity 3D models of the gardens, which are invaluable for preservation efforts and virtual reality recreations.

Handling of Light and Shadow

One of the more challenging aspects of outdoor environments is the changing light conditions, which can dramatically affect the appearance and visibility of objects. Jiangnan gardens, with their open spaces and lush canopies, present a challenging array of light and shadow interplays. Marigold's deep learning algorithms are adept at differentiating between shadows and actual depth variations, a critical feature that prevents the misinterpretation of shadow as form. This capability is particularly important during the late afternoon when long shadows cast by the garden's architecture could obscure or distort underlying features.

Single Image Depth Mapping

Traditional depth mapping techniques often rely on inputs from multiple sensors or stereoscopic images to construct depth maps. However, Marigold breaks away from this norm by effectively generating accurate depth maps from single images. This feature is particularly advantageous for analyzing Jiangnan gardens, as it simplifies the data acquisition process—allowing for the capture of detailed spatial data with minimal equipment. Researchers and conservators can obtain comprehensive depth information using just a standard camera, significantly reducing the logistical complexity and costs associated with data collection.

Simplification of Data Acquisition and Processing

The ability of Marigold to process single images into detailed depth maps significantly simplifies the overall data acquisition and analysis process. This simplification is invaluable in scenarios where access to gardens is limited, or where minimal disturbance to the environment is paramount. By reducing the need for multiple sensor setups or complex riggings, Marigold facilitates a more efficient and less intrusive method of capturing the necessary data. This ease of use not only accelerates the research and documentation process but also makes it more feasible for ongoing monitoring and analysis of these culturally rich sites.

Implications for Cultural Preservation and Research

The application of the Marigold model in the study and conservation of Jiangnan gardens has profound implications. By providing detailed and accurate representations of garden layouts and features, the model aids in the documentation and preservation of these sites. These digital assets are crucial for historical research, providing scholars and conservators with the tools to analyze changes over time, assess the impact of environmental factors, and plan restoration projects with greater precision. Moreover, the visualizations generated can enhance public appreciation and understanding of Jiangnan gardens, promoting cultural heritage and encouraging support for its preservation.

Future Directions and Enhancements

Looking ahead, the potential expansions of the Marigold model include its adaptation for realtime processing and augmented reality applications. Such advancements could transform the way visitors interact with Jiangnan gardens, offering them immersive educational experiences through augmented tours that highlight historical and ecological insights. Additionally, ongoing improvements to the model's algorithms will enhance its accuracy and efficiency, ensuring that it remains at the forefront of technological innovations in cultural heritage preservation.

In conclusion, the Marigold depth estimation model represents a significant advancement in the field of digital heritage and environmental analysis. Its application to the Jiangnan gardens not only underscores its versatility and power but also highlights the growing importance of technology in the preservation and appreciation of our cultural legacies.

4.2.2 Framework of the Marigold Depth Estimation Model

The Marigold model operates on a convolutional neural network (CNN) architecture optimized for depth prediction. This framework is structured to extract multiscale features from input images, enabling the model to appreciate details at various scales—from the delicate patterns of individual leaves to the overarching layout of garden paths and structures. The model processes inputs through a series of convolutional layers that progressively encode higherlevel features, followed by upscaling mechanisms that reconstruct the depth map at the original resolution.

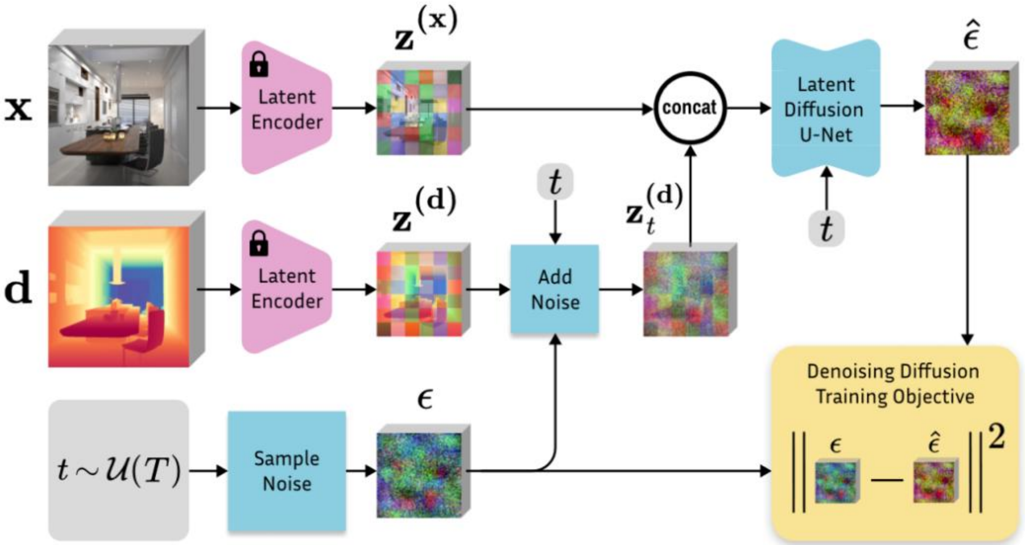


Figure 4-1: Overview of the Marigold fine-tuning protocol (Ke, B., et al.,2024).CVPR 2024 (Oral, Best Paper Award Candidate)

The Marigold model is a sophisticated framework based on latent diffusion models, designed to model the conditional distribution D(d|x) , where d represents depth and x denotes the given RGB image. This innovative approach employs a stepbystep addition and removal of noise to recover clear depth information from noisy data. During training, the model employs adversarial learning to finetune its parameters, aiming to minimize the denoising diffusion objective function. This effectively allows the model to predict corresponding depth maps from the input RGB images. The generative nature of this framework not only enhances the accuracy of depth estimation but also improves the model's ability to generalize across new domain data.

Network Architecture

At the core of Marigold lies the pretrained texttoimage latent diffusion model (LDM), which has been finetuned to suit the depth estimation tasks(Figure 4-1) . The architecture incorporates a frozen Variational Autoencoder (VAE) to encode images and depth information into a latent space. This is coupled with specifically tailored depth encoders and decoders. During the denoising process, the model adapts a modified UNet to perform conditional encoding of images, which enhances training efficiency while maintaining the capability to generate highresolution images. This carefully designed network architecture enables Marigold to process depth estimation tasks efficiently while leveraging robust image priors.

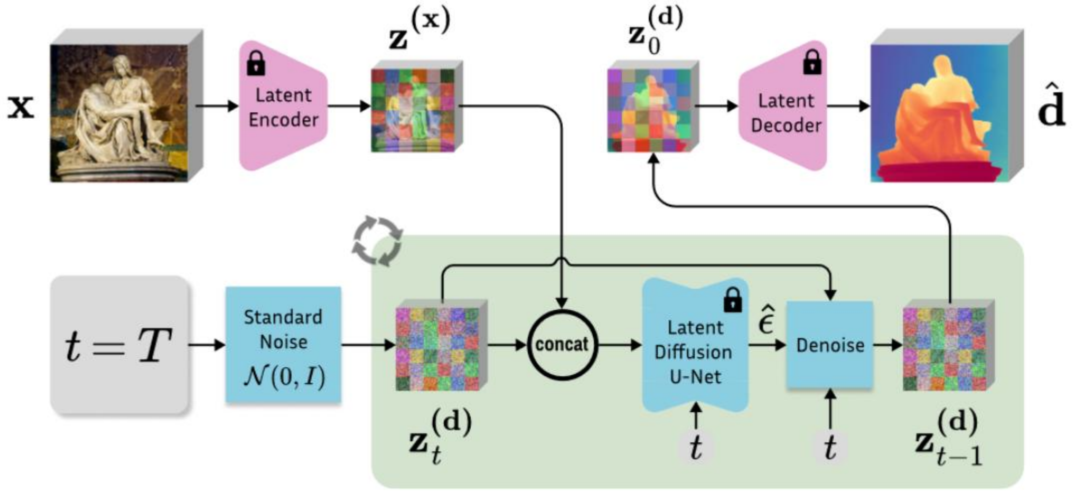


Figure 4-2: Overview of the Marigold inference scheme.(Ke, B., et al.,2024).CVPR 2024 (Oral, Best Paper Award Candidate)

The finetuning protocol of the Marigold model leverages the pretrained latent diffusion model (LDM) for depth estimation(Figure 4-2). By adopting affineinvariant depth normalization and training strategies on synthetic data, the model effectively adapts and learns from synthetic RGBD data obtained from shortterm training on a single GPU. This data does not require complex preprocessing, simplifying the training process. Additionally, the use of multiresolution noise and annealing schedules further enhances training efficiency and model performance. This protocol not only improves the accuracy of depth estimation but also ensures good generalization capabilities of the model across new scenes. This showcases the potential for achieving advanced performance by finetuning pretrained models with limited resources.

An overview of the Marigold finetuning protocol illustrates how pretrained latent diffusion models are repurposed for monocular depth estimation. The implementation of this protocol demonstrates effective depth map generation from single images using a combination of novel training strategies and a robust architectural framework.

4.3 Key Considerations for Depth Map Analysis

Depth map analysis within the framework of the Marigold model requires attention to several critical factors:

4.3.1 Depth Continuity and Smoothness

Ideal depth maps should show smooth gradations where the realworld geometry does not change abruptly. Discontinuities or abrupt changes should only appear where actual depth changes occur in the scene.

For Jiangnan gardens, where continuous elements like water bodies and rockeries prevail, the depth maps should reflect smooth transitions, preserving the natural flow of the landscape. The Marigold model is designed to ensure that depth continuity is maintained, and abrupt depth changes are only marked where actual physical barriers or elevation changes exist.

4.3.2 Edge Preservation

It's crucial that depth maps maintain sharp transitions at the edges of objects to avoid artifacts in applications like 3D reconstruction or augmented reality. Edges should be welldefined and not bleed into adjacent areas.

Accurate edge delineation is vital for maintaining the structural integrity of architectural features in gardens, such as pavilions and sculpted hedges. The Marigold model emphasizes edge preservation, ensuring that depth transitions at object boundaries are sharp and precise, facilitating highquality 3D reconstructions.

4.3.3 Handling of Overlapping Structures

Gardens often feature overlapping elements such as branches, leaves, pavilions, and distant structures. A robust depth map should correctly interpret these overlaps, distinguishing foreground from background appropriately.

The layered nature of Jiangnan gardens, with their overlapping foliage and nested spaces, poses a significant challenge for depth estimation. The Marigold model uses advanced segmentation algorithms alongside depth prediction to differentiate between foreground and background layers effectively, thus accurately modeling the depth of overlapping structures.

4.3.4 Response to Light and Shadows

Depth estimation models can sometimes misinterpret shadows as depth variations. Effective depth maps should differentiate between actual depth differences and shadowinduced apparent depth changes.

Shadows can significantly affect depth perception by creating false indications of depth discontinuities. The Marigold model incorporates lightingaware algorithms that distinguish between shadows and genuine depth variations, ensuring that shadows do not skew the depth analysis.

4.3.5 Color and Texture Handling

Algorithms may interpret areas with similar color or texture as being at the same depth, which isn't always the case. Accurate depth maps should transcend these visual similarities to derive actual depth based on structure.

In Jiangnan gardens, similar textures and colors across different depth planes can confuse standard depth estimation techniques. The Marigold model addresses this by integrating texture and color differentiation features within its depth estimation process, enabling it to overcome the challenges posed by visual homogeneity in garden environments.

4.3.6 Multiscale Feature Learning

Gardens contain elements varying from smallscale features like individual flowers to largescale features like towering trees or expansive lawns. Depth estimation models must therefore operate effectively across multiple scales to accurately capture the depth information of both minute and large objects. Multiscale feature learning involves analyzing the scene at various resolutions, allowing the model to understand and integrate the depth data from different perspectives and scales. This approach ensures that both the macro and micro aspects of a garden are represented accurately in the depth map, enhancing the overall model's utility for diverse applications.

By addressing these considerations, depth map analysis can significantly improve in accuracy and functionality, making it a powerful tool for applications that rely on precise spatial awareness and realistic digital interactions.

Chapter 5: Results of Study 1

5.1 Performance of the YOLOv8n-modify Model

5.1.1 Precision Analysis

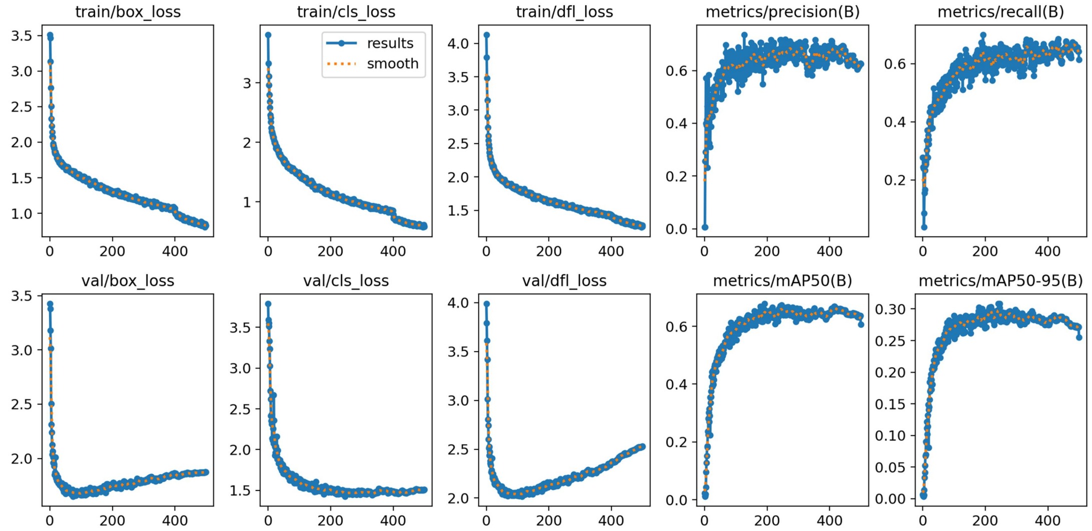


Figure 5-1: Performance parameters of model training for 500 rounds

The YOLOv8n-modify model exhibits a noteworthy learning trajectory throughout the 500 epochs, characterized by rapid gains in precision during the initial training phases(Figure 5-1). This rapid ascent in performance metrics highlights the model's capability to efficiently assimilate fundamental patterns and features from the training data. Specifically, the precision swiftly escalated from its minimum value of approximately 0.003, indicating an initial adjustment phase where the model began to learn discriminative features crucial for detecting objects.

Throughout the training process, the precision metric encountered periodic fluctuations, which can be quantified by observing a standard deviation of approximately 0.078 in the precision values. This variability is indicative of the model's interaction with complex and possibly noisy data segments within the training set. Such fluctuations might also reflect the adjustments in model parameters and the impact of different minibatch samples influencing the learning process at various stages.

Notably, the model achieved a semblance of midterm stability, as evidenced by the maintenance of precision within a higher range, predominantly between approximately 0.661 and 0.722. This stability phase suggests that the YOLOv8 model has effectively captured the principal characteristics of the target objects, likely due to a welltuned balance between the model's sensitivity and specificity, optimized through continuous learning and adaptation to the data's intricacies.

Furthermore, the maximum precision achieved during the training was about 0.722, which signifies the model's ability to reach a high level of precision under optimal conditions. This peak value can be seen as the model's performance ceiling under the current training configuration and data quality.

In summary, the YOLOv8n-modify model's performance over 500 epochs illustrates its robust learning capacity, punctuated by periods of significant precision gains and moderate variability, ultimately leading to a highprecision detection capability. This performance narrative underscores the model′s effectiveness in navigating through diverse challenges posed by the dataset, capturing essential data features, and achieving local optima that contribute to sustained high performance in object detection tasks.

5.1.2 mAP Analysis

The Mean Average Precision (mAP) metrics, specifically mAP@50 and mAP@5095, are pivotal indicators of object detection model performance. These metrics give us a nuanced view of the model's precision across different Intersection over Union (IoU) thresholds.

（1）For mAP@50:

Minimum Value:

The mAP@50 reaches its lowest at 0.00663, suggesting initial difficulty in achieving overlap with the ground truth.

Maximum Value:

It peaks at 0.64346, showing that at its best, the model has a robust detection capability at the 50% IoU threshold.

Mean Value:

With an average of 0.56361, the model consistently maintains high precision for this threshold over the epochs.

Standard Deviation:

A standard deviation of 0.09531 indicates moderate fluctuations in precision, which can be expected as the model learns and adapts to the data.

（2）For mAP@5095:

Minimum Value:

The lowest mAP@5095 is 0.00191, reflecting initial challenges the model faces when meeting more stringent IoU criteria.

Maximum Value:

The highest value recorded is 0.29632, which, while lower than the mAP@50, still represents a significant capability in detecting objects across a range of IoU thresholds.

Mean Value:

The mean of 0.24947 suggests that, on average, the model has a moderate level of precision across these more demanding IoU thresholds.

Standard Deviation:

The standard deviation of 0.04488 denotes less variability compared to mAP@50, which may indicate that the model's performance across different IoU thresholds is relatively stable.

These statistics reveal that while the model excels in detecting objects at a lower IoU threshold (50%), as depicted by the higher mAP@50 scores, it faces more challenges as the IoU thresholds become more rigorous, as shown by the mAP@5095 scores. However, the overall upward trends in both metrics across epochs signal a solid capacity for learning and improving detection precision. The consistent performance in mAP@50 along with the significant improvements in mAP@5095 suggest that the YOLOv8n-modify model is effectively learning to generalize its detection capabilities, not just at a basic level of overlap, but across a spectrum of precision requirements. This analysis points to YOLOv8n-modify model that, while already performing well, has room for growth, especially in achieving precision at higher IoU thresholds, which could be an area of focus in future training iterations.

5.1.3 The recall rate

In the context of Figure 5-1, which plots various metrics over 500 epochs of YOLOv8 model training, the recall rate exhibits a multiphased evolutionary pattern. Initially, the recall rate starts at 0.27786, indicative of the model's nascent stage where it is only beginning to correctly identify a portion of relevant objects within the dataset. This is followed by a precipitous decline to 0.03762, reflecting a period of potential overfitting or an adjustment phase where the model's detection capacity temporarily wanes.

Following this decline, the recall metric enters a stage of recuperation and consistent improvement. This incremental ascent suggests iterative refinements in the model's ability to capture a greater fraction of the positive cases. The multistage upward trajectory, characterized by small peaks and valleys, mirrors the iterative learning process, where the model benefits from successive rounds of optimization, tuning, and possibly augmentation or correction within the training data.

Ultimately, the recall metric culminates at a value of 0.70001, a substantial enhancement from its nadir. This peak represents a level of maturity in the model's development where it has substantially enhanced its capacity to detect relevant objects. The presence of several stages in the recall rate's ascension likely indicates the model's navigation through various complexities within the data, such as class imbalances, ambiguous cases, and varying degrees of object occlusion or scale.

The fluctuating pattern before achieving stability could be attributed to the model encountering new data features, adapting to the nuanced variations within the dataset, and undergoing finetuning of the weights and biases within its architecture. Each incremental stage of increase in recall may also reflect the model overcoming specific detection challenges or benefiting from improved generalization as a result of optimized hyperparameters.

The overall trend of the recall rate, with its intricate pattern of dips and recoveries leading to a substantial final value, signifies a learning process punctuated by challenges and optimizations. This conveys the model's resilience and adaptive learning capacity to better predict and recall the relevant features over time, indicative of an increasingly robust object detection system.

5.1.4 Losses

(1)Training Set Losses

①Box Loss : The training box loss demonstrates a reduction from an initial value of 3.4978 to a minimum of 0.87896, with an average of 1.33207. The standard deviation of 0.32274 implies a relatively moderate spread in the box loss values, indicative of the model's increasing precision in predicting the bounding boxes over time.

②Classification Loss : The classification loss decreases markedly from 3.8463 to

0.57134. The mean classification loss stands at 1.15047, a testament to the model's improving ability to correctly classify objects. The higher standard deviation of 0.46063 in cls_loss suggests more variability in the model's classification performance during training.

③Distribution Fitting Loss : The distribution fitting loss, with values starting at 4.1473 and decreasing to 1.206, averaging at 1.59618, indicates the model's evolving capability to fit the probability distribution of the classes. A standard deviation of 0.32337 points to a stable learning curve with respect to this complex aspect of the loss.

(2) Validation Set Losses

①Box Loss : The validation box loss oscillates between 3.4224 and 1.7087, averaging at 1.86390. The relatively low standard deviation of 0.15586 signifies a consistent performance in the model's generalization to new data.

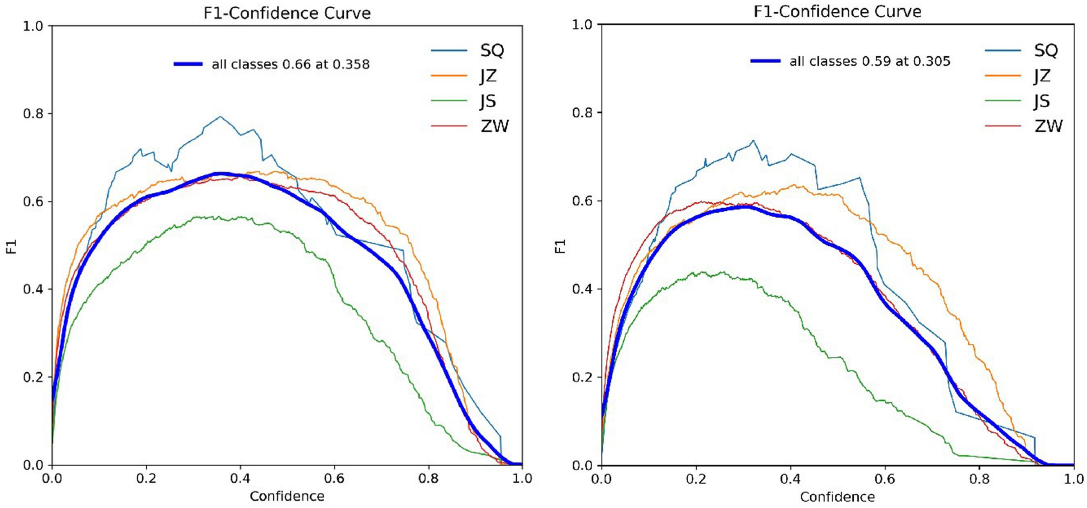
②Classification Loss : The cls_loss in the validation set shows the most significant variability, with a maximum of 11.718 and a minimum of 1.4783, averaging at 1.65115. This high variability, with a standard deviation of 0.57867, may highlight areas where the model's classification ability could be further improved, particularly under varying or unseen data conditions.

③Distribution Fitting Loss : For the validation dfl_loss, the range from 3.9463 to 1.9363 with an average of 2.10943 and a standard deviation of 0.19335 suggests the model's consistent but slightly less stable performance in distribution fitting when compared to the training set.

5.2 YOLOv8n-modify Model Comparative Analysis with YOLOv8n Models

5.2.1 Comparative of The F1 versus confidence curve

The F1 versus confidence curve in Figure 8 offers a comprehensive depiction of the model's performance across a spectrum of confidence thresholds. This curve provides valuable insights into the tradeoff between precision and recall, encapsulated by the F1 score—a harmonic mean of these two critical metrics(Figure 5-2). An indepth examination of the F1 score for YOLOv8n-modify reveals a consistent outperformance compared to its predecessor, YOLOv8, at a majority of the confidence thresholds. This is a clear indication of the modifications' efficacy, where the improved model not only maintains but elevates performance standards across varying levels of detection confidence.



YOLOv8n-modify

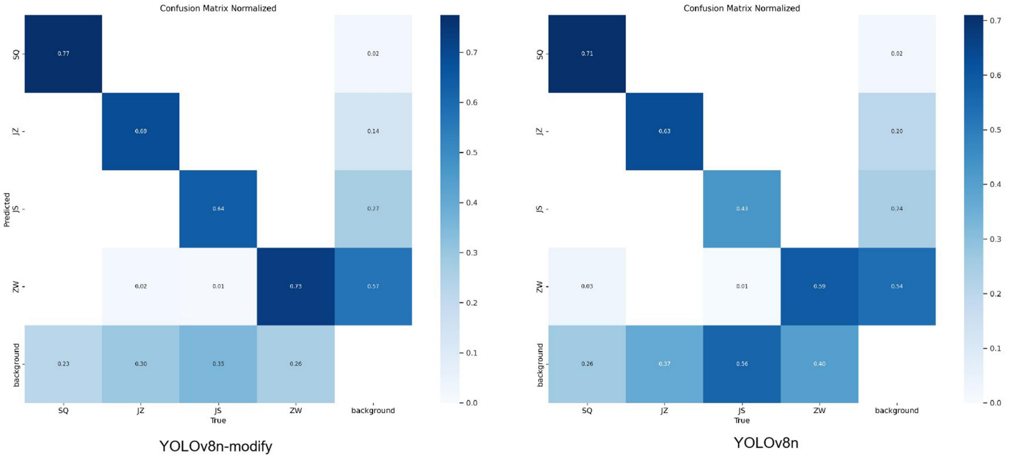
YOLOv8n

Figure 5-2: F1 and confidence curve

In the case of YOLOv8n-modify, the enhanced F1 values suggest that the modifications have yielded a model that is adept at balancing the recall and precision, ensuring that both the detection of relevant objects and the minimization of false positives are optimized. For instance, the model achieves an F1 score of 0.66 at a confidence threshold of 0.358. This score is 7% higher than that of the original YOLOv8 model, illustrating a notable advancement in the model ′ s predictive precision. This precise point of 0.358 for the confidence threshold emerges as an operational sweet spot, where the model's decisions about object presence are at their most balanced and reliable.

Additionally, the smoother nature of the curve for YOLOv8n-modify suggests a model that is not only more consistent in its performance across different thresholds but also possibly indicates a bettercalibrated probability output. In contrast to sharp peaks and troughs which may reflect erratic behavior in response to varying data inputs, a smoother F1 curve denotes a model that is unshaken by such variability, maintaining composure across the spectrum of decision thresholds.

5.2.2 Comparative of The confusion matrix



YOLOv8n-modify

YOLOv8n

Figure 5-3: Confusion matrix

The confusion matrix(Figure 5-3), a pivotal tool in evaluating the performance of classification models, reveals significant insights into the comparative effectiveness of YOLOv8modify versus its precursor, YOLOv8. By examining specific categories such as SQ, JZ, JS, and ZW, we observe marked improvements in recognition precision that speak volumes about the enhancements integrated into the modified version of the model.

In the modified model, YOLOv8modify, the recognition accuracies for the SQ, JZ, JS, and ZW categories are recorded at 0.77, 0.69, 0.64, and 0.73, respectively. These figures represent substantial improvements over the initial results of the original YOLOv8 model, where the corresponding accuracies were lower: 0.71 for SQ, 0.61 for JZ, 0.43 for JS, and 0.59 for ZW. This translates into improvements of 0.06, 0.08, 0.21, and 0.14 for each category, respectively. Such enhancements are not merely numerical gains; they are indicative of the model's refined ability to interpret and classify complex visual data more accurately. The significant leap in the precision of the JS category, in particular, from 0.43 to 0.64, underscores a remarkable advancement in distinguishing features that were previously challenging for the original model.

These improvements likely stem from a combination of optimized training parameters, enhanced feature extraction capabilities, and possibly more targeted data augmentation strategies that provide the model with a richer, more varied learning environment. The robustness of YOLOv8modify in identifying these categories, especially in a consistent manner across varied test cases, highlights its potential utility in realworld applications where such precision is crucial.

However, alongside these advancements, there are notable challenges as well. The model occasionally misclassifies background elements as belonging to the ZW category. This misclassification issue points towards a common challenge in object detection tasks—distinguishing between foreground objects and complex background patterns. The overlap in features or shared properties between the background and ZW categories, such as similar textures or color schemes, might be causing these errors. This suggests that while the model excels in recognizing distinct objects, its perceptual differentiation between objects and similarly featured backgrounds needs further refinement.

This phenomenon could potentially be addressed by integrating more nuanced background examples into the training set or by implementing more sophisticated segmentation algorithms that can better delineate objects from their surroundings. Additionally, enhancing the model's spatial reasoning capabilities might help reduce these kinds of misclassifications, thereby boosting its overall precision and reliability.

In conclusion, the analysis of the confusion matrix for YOLOv8modify and YOLOv8 demonstrates a clear path of progress with significant enhancements in categoryspecific recognition accuracies. Nonetheless, it also delineates areas needing attention, such as the differentiation between background and objects, which if addressed, could propel the model towards even higher levels of performance and reliability in diverse deployment scenarios.

5.3 Comparative Analysis with Alternative Methods

Table 5-1: YOLOv8n-modify Comparative Analysis with Alternative Methods

Modules	Adopted Modules	Precision (%)	Recall(%)	mAP0.5	mAP0.50.95	FPS(%)
YOLOv3n	none	55	48.6	53.1	25.8	27.2
YOLOv5n	none	57.4	51	51	22.4	38.3
YOLOv8n	none	57.4	53.5	55.3	26.1	38.9
YOLOv8n-modify	DBB+BiFPN +DyHead Attention	66.1	46.7	57.1	29.2	14.8

In a comprehensive evaluation of advanced object detection methodologies, the enhanced YOLOv8modify was rigorously compared against earlier iterations such as YOLOv3, YOLOv5, and the baseline YOLOv8. This comparison, as outlined in the figure and summarized in Table 5-1, was meticulously structured to ensure uniformity across various parameters. To achieve this, the operational environment and network configurations were standardized, and each model was trained to the point of convergence, ensuring that the performance metrics reflected the absolute potential of each methodology.

The comparative analysis, carried out using an identical test set, highlighted the superior capabilities of the enhanced YOLOv8modify method in object detection tasks. Specifically, this optimized version of YOLOv8 exhibited a precision rate of 66.1%, a recall of 46.7%, and an mAP50 of 57.1%. These metrics not only signify a robust enhancement over its predecessors but also underscore the methodological improvements incorporated into the model's design. The precision rate of over 66% is particularly noteworthy as it indicates a high likelihood that detected objects are correctly identified, which is critical in reducing false positives—a common challenge in complex detection environments.

The recall rate of 46.7%, while not as high as the precision, is significant in that nearly half of all actual objects in the images were correctly identified by the model. This balance between precision and recall is crucial for applications where missing an object can be just as detrimental as falsely identifying one. Furthermore, the mAP50 of 57.1% across varying IoU thresholds reflects a solid ability to maintain detection precision across different object sizes and overlaps, enhancing the model's utility in diverse realworld scenarios.

Moreover, the improvements in YOLOv8modify make it especially suitable for deployment in these Jiangnan garden scenarios, where the precision in detecting subtle differences between closely resembling objects can significantly enhance the user experience and operational efficiency. The application of this model in such settings demonstrates not only the versatility of YOLO architectures but also the potential for specialized adaptations to meet specific environmental demands.

The exemplary performance of YOLOv8modify, as highlighted in the comparative analysis against older YOLO versions, can be directly attributed to strategic enhancements integrated into its architecture. These enhancements include the introduction of the DBB (Diverse Branch Block) module into the Backbone Layer, the integration of BiFPN (Bidirectional Feature Pyramid Network) into the Neck Layer, and the incorporation of an attention mechanismbased object detection head, specifically the DyHead, into the Head Layer. Each of these modifications contributes uniquely to the model's improved precision and efficiency, making it particularly adept at handling the complexities of object detection in varied scenarios.

These targeted enhancements, each contributing to different stages of the detection process, synergize to produce a model that not only surpasses its predecessors in terms of raw performance metrics like precision, recall, and mAP50 but also in its ability to handle complex detection environments more effectively. By addressing specific challenges associated with feature extraction, integration, and contextual processing, YOLOv8n-modify showcases a profound improvement in detecting objects in the intricate settings of Jiangnan traditional gardens and similar environments, where traditional models might fail to deliver optimal performance. This demonstrates the power of thoughtful architectural modifications in pushing the boundaries of what's possible with object detection technologies.

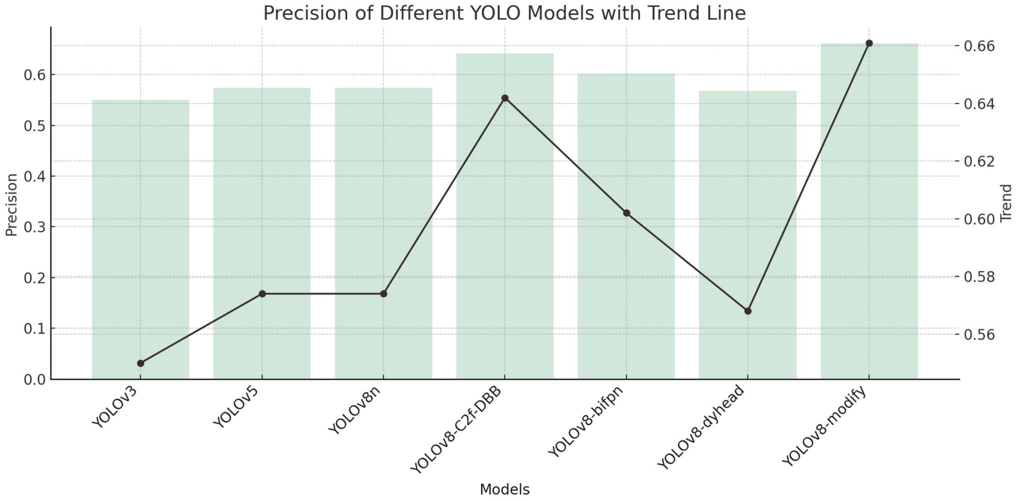


Figure 5-4: Comparison of precision between YOLOv8n-modify and other models

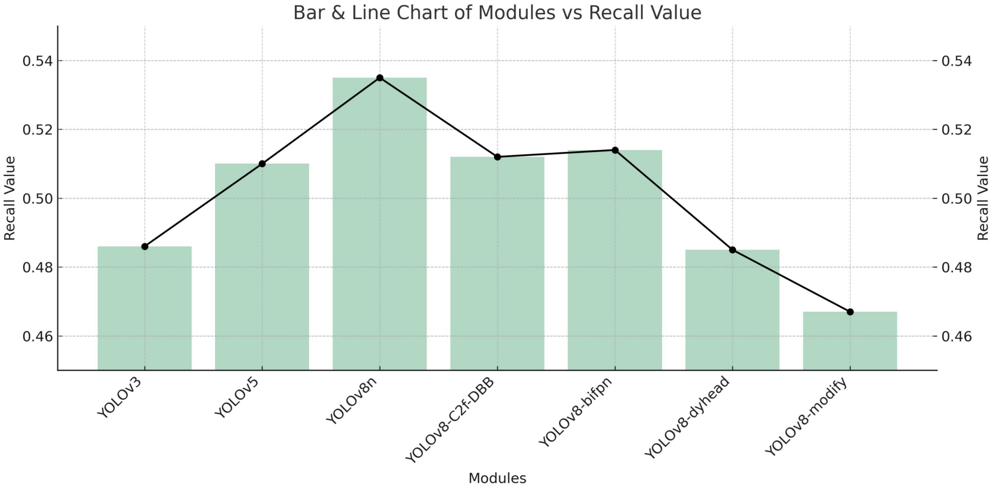


Figure 5-5: Comparison of recall value between YOLOv8n-modify and other models

Figure 5-4 provides a visual comparison of the precision metrics across different iterations of the YOLO model, emphasizing their performance in object detection. The chart utilizes light green bars to delineate the precision of each YOLO version, with an accompanying line chart that traces the trend of these accuracies over time or versions. Notably, the YOLOv8n-modify model stands out with a remarkable precision score of 0.661, surpassing its predecessors whose accuracies range between 0.55 and 0.65. This significant increment in precision indicates the efficacy of the architectural enhancements implemented in the YOLOv8n-modify model, particularly in its ability to correctly identify and classify objects with higher reliability. The graphical representation in Figure 5-4 effectively highlights the progressive improvement in precision, showcasing how each subsequent model version has built upon the previous iterations to achieve higher precision.

Figure 5-5 shifts focus to the recall rates of these models, plotting a range between 0.467 and 0.535. Intriguingly, YOLOv8n achieves the highest recall of 0.535 even without the integration of additional modules, suggesting that the base model was adept at capturing a majority of relevant objects within the scene. However, with the introduction of more sophisticated modules in YOLOv8n-modify, the recall slightly drops to 0.467. This decrease might seem counterintuitive given the increase in precision, but it underscores a strategic choice in model configuration: optimizing for high detection precision sometimes at the expense of capturing every possible object. This tradeoff can be particularly impactful in environments rich in complex visuals such as those found in the dataset, which includes dense greenery and architecturally complex structures that pose significant challenges in object detection.

The reduction in recall for YOLOv8n-modify could be attributed to several factors. Firstly, the complex image content within the dataset, including overlapping objects, varying textures, and subtle distinctions in the visual scene, makes it inherently difficult to maintain a high recall without sacrificing precision. The densely populated scenes in the images require the model to differentiate between highly similar objects and background clutter, a task that becomes increasingly challenging as the precision of the model increases. Secondly, the initial configuration of our algorithm was designed to prioritize precision over the breadth of detection. This configuration decision was likely made to ensure that the model minimizes false positives, a critical factor in applications where the cost of incorrect detection is high.

Overall, the balance between precision and recall observed in these models, as depicted in Figures 10 and 11, is a reflection of the underlying complexities and strategic decisions involved in model training and deployment. While YOLOv8n-modify shows an improved precision, indicating a superior capability in correctly identifying objects with fewer errors, the slight decrease in recall highlights the challenges and compromises that come with tuning the model to navigate and interpret highly complex visual data. This nuanced understanding of model performance, with a focus on optimizing specific metrics according to the needs of the application, is crucial for advancing the field of object detection and for deploying these models effectively in realworld scenarios.

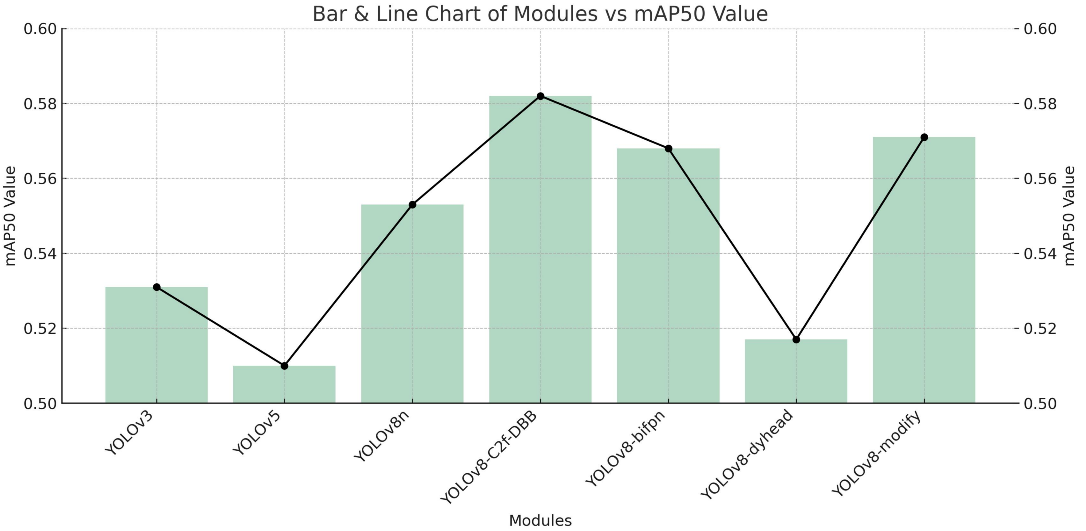


Figure 5-6: Comparison of mAP50 between YOLOv8n-modify and other models.

mAP50 is a critical metric in the evaluation of object detection models, particularly for understanding how well a model can detect objects with at least 50% overlap with the ground truth annotations. As highlighted in Figure 5-6, the mAP50 values across different YOLO models exhibit varied performance, with YOLOv8C2fDBB achieving the highest score of 0.582. This version, which incorporates a specific configuration of the DBB (Diverse Branch Block) module, showcases its efficacy in enhancing detection precision significantly. Close on its heels is YOLOv8n-modify, with an mAP50 of 0.571, indicating that despite not being the top performer, it still maintains a high level of precision in object detection tasks. This neartop performance in mAP50 reflects the successful integration of advanced modules and optimizations that focus on improving the precision and robustness of the detection capabilities.

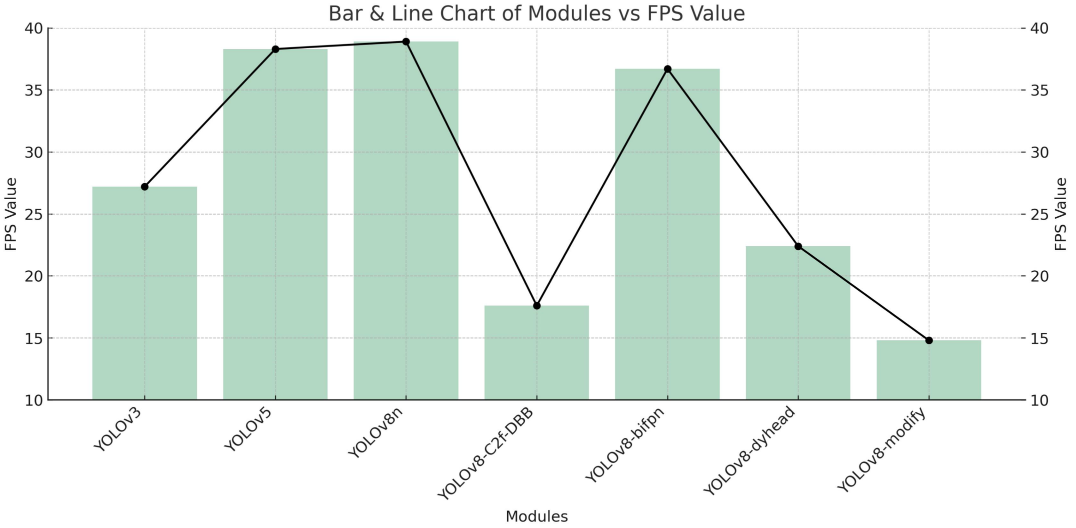


Figure 5-7: Comparison of FPS between YOLOv8n-modify and other models.

The FPS (Frames Per Second) metric, as discussed in Figure 5-7, serves as an indicator of the model's efficiency in processing video frames, which is crucial for applications requiring realtime analysis such as video surveillance and autonomous driving. In this regard, YOLOv8n stands out with the highest FPS at 38.9, suggesting that it is highly optimized for speed without necessarily incorporating the heavier computational modules found in other variants. This makes it exceptionally suitable for scenarios where speed is paramount. On the other hand, YOLOv8n-modify, which integrates several sophisticated features to boost detection precision, shows a considerably lower FPS of 14.8. This substantial decrease in frame rate underscores the tradeoff between computational complexity and realtime performance. The integration of complex modules such as BiFPN and DyHead, while beneficial for precision and feature integration, evidently imposes a heavier computational burden on the model, thus reducing its throughput.

This disparity in performance between the models highlights the inherent challenges in balancing detection precision with computational efficiency. YOLOv8n-modify, despite its commendable precision metrics, suffers in scenarios requiring swift frame processing, which could limit its application in realtime contexts. Conversely, the streamlined YOLOv8n, with its higher FPS, offers an alternative that, while slightly less accurate, provides much faster detections, making it more applicable for realtime applications.

In conclusion, the comparison of mAP50 and FPS across these YOLO variants illustrates the complex interplay between model architecture decisions and their practical implications. These metrics offer crucial insights into selecting the appropriate model variant based on specific requirements of precision and speed. Future improvements in YOLO models will likely continue to explore this balance, seeking ways to incorporate the benefits of high precision and robust feature detection without compromising significantly on the speed necessary for realtime applications.

5.4 Ablation Study

Table 5-2：Ablation Study

Modules	Add DBB	Add BiFPN Block	Add DyHead Attention	Precision (%)	Recall (%)	mAP0.5	mAP0.50.95
YOLOv8n	×	×	×	57.4	53.5	55.3	26.1
YOLOv8nC2fDBB	○	×	×	64.2	51.2	58.2	28.1

YOLOv8nbifpn	×	○	×	60.2	51.4	56.8	27.5
YOLOv8ndyhead	×	×	○	56.8	48.5	51.7	26.9
YOLOv8nbifpndyhead	×	○	○	61.1	52.7	60.6	28.9
YOLOv8nC2fDBBbifpn	○	○	×	64.8	53.2	62.5	30.1
YOLOv8nC2fDBBdyhead	○	×	○	63.2	52.0	58.8	29.8
YOLOv8n-modify	○	○	○	66.1	46.7	57.1	29.2

The comprehensive ablation studies conducted to evaluate the individual and combined impacts of the three strategic enhancements—DBB Module, BiFPN, and DyHead—on YOLOv8modify provide insightful data on how each contributes to the overall optimization for object detection in garden datasets (Table 5-2). These experiments, crucial for understanding the specific benefits of each module, were meticulously designed with uniform hyperparameters and pretraining weights across tests, ensuring that the observed improvements could be directly attributed to the architectural modifications rather than external variables.

1. DBB Module Enhancement:

The integration of the DBB (Diverse Branch Block) Module into the backbone of the network marks a significant stride in enhancing detection precision without adding undue complexity or computational overhead. This module employs a multibranch network architecture, where each branch processes the input with different convolutional strategies, thereby extracting a broader range of features. The results from the ablation study show that incorporating the DBB Module led to an enhanced detection precision, demonstrating its efficacy in enriching the model's feature extraction layer without significantly increasing the resource demands. This optimization allows the model to remain agile and efficient, crucial for realtime object detection applications.

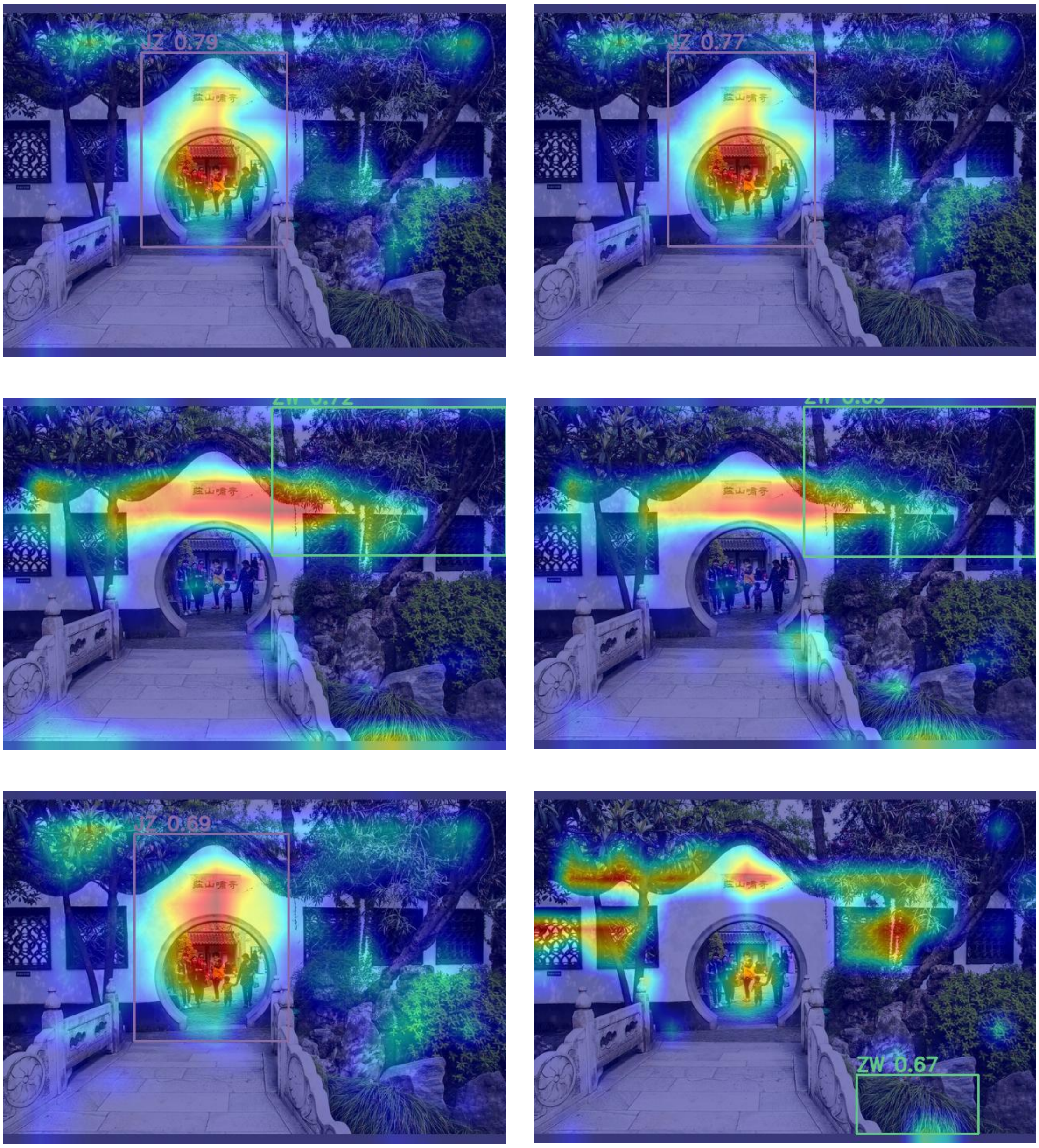
2. BiFPN Enhancement:

The BiFPN (Bidirectional Feature Pyramid Network) component, integrated into the neck layer of the model, significantly advances the feature fusion process. By facilitating a more effective integration of semantic information from across different scales, BiFPN ensures that both highlevel and lowlevel features are more accurately aligned and utilized for decisionmaking. This capability is especially beneficial in complex scenes typical of garden environments where the distinction between foreground objects and intricate backgrounds can be subtle. The ablation studies highlighted that BiFPN not only improves the precision of feature fusion but also enhances the model's precision in localizing and classifying objects, as reflected in the improved metrics.

3. DyHead Enhancement:

Finally, the implementation of DyHead, an attention mechanismbased detection head, introduces dynamic convolutional capabilities that adapt to the input features, overlaying scale, spatial, and taskspecific attentions. This development allows the model to concentrate its computational resources on the most informative features of the input, significantly refining its predictive precision. The DyHead's ability to adjust its focus based on the varying scales and complexities of the objects enhances all performance metrics, notably precision and recall. The ablation study results underscored DyHead's role in substantially boosting the model's overall efficacy.

The aggregated findings from these ablation studies, as shown in Table 5-2, indicate that the collective integration of DBB, BiFPN, and DyHead into the optimized YOLOv8modify leads to impressive performance metrics, with a precision of 66.1%, a recall of 46.7%, and an mAP50 of 57.1%. Each module's contribution is evident in these enhanced metrics, confirming that the strategic enhancements not only bolster the model's precision but also its efficiency and robustness across challenging detection environments. This synergistic integration exemplifies the potential of architectural innovations in advancing the field of object detection, particularly in specialized settings such as the complex landscapes of Jiangnan traditional gardens.



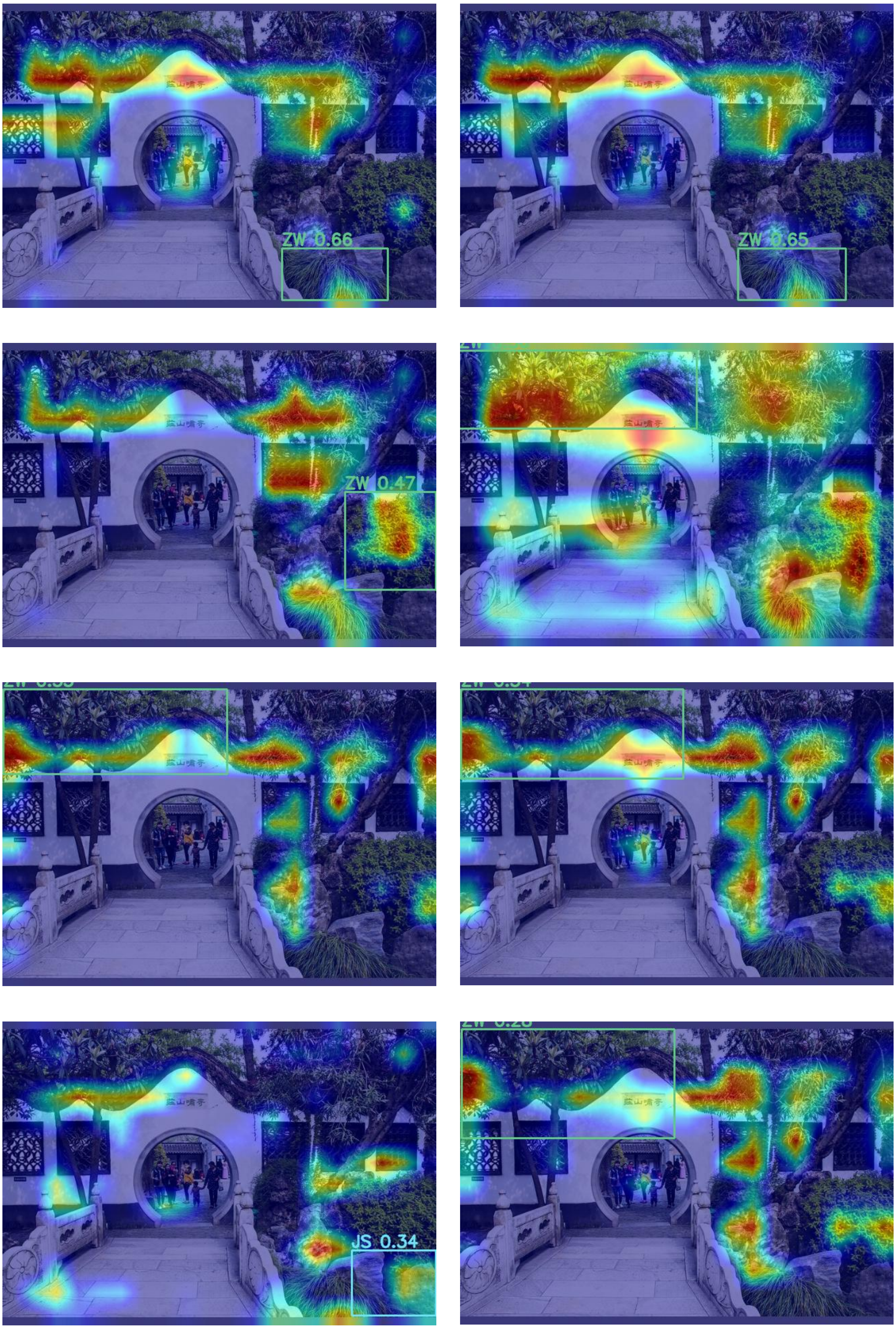


Figure 5-8: Heatmap changes of the attention mechanism during model training

5.5 Changes in Heatmaps of Attention Mechanisms During Model Training

The Figure 5-8 display changes in the heatmaps of attention mechanisms during the model training process. These heatmaps provide us with an intuitive way to observe how the model gradually learns and focuses on key features within the images. They show the model's attention to different areas of the image at various stages of training, indicated by changes in color gradients (typically from blue to red, representing a shift from lower to higher attention) and attention value changes.

The change in values, where higher values (close to 1.0) indicate a high degree of attention to a region, is because that area contains crucial information for classification or detection tasks. A decrease in values may indicate that the model is gradually learning to ignore less important features, or that the focus is shifting to more critical features after multiple iterations.

The change in color intensity from blue to red signifies a transition from low to high attention. Changes in the intensity and range of colors within the images can indicate optimizations in the model's feature extraction strategy. Initially, the entire image area may have a broad distribution of colors, but as training progresses, colors may become concentrated in certain key areas, indicating that the model is learning to focus resources on parsing these critical features.

These visual heatmaps show that in the early stages of model training, the distribution of attention points is quite broad, reflecting the model's attempt to capture all possible information to determine which features are most important. As training continues, the focus becomes more concentrated and precise, indicating that the model is becoming more efficient by focusing only on features that are most helpful for the final task. This optimization can reduce the waste of computational resources and improve processing speed and precision.

In the YOLOv8n-modify network model, the capabilities of the attention mechanism are directly related to the added DBB (Diverse Branch Block), BiFPN (Bidirectional Feature Pyramid Network), and DyHead modules. Each module enhances the model's attention distribution and processing capabilities for key information in images by improving feature extraction and integration methods. The following detailed explanations show how these modules interact with the attention mechanism:

（1）Diverse Branch Block (DBB)

DBB improves the convolutional layers in YOLOv8 by introducing a multibranch structure that allows the model to simultaneously process and recognize image information of different sizes and complexities. This structure, by expanding the model's receptive field and increasing the diversity of features, enables more refined and widespread attention, which is particularly important for enhancing the model's recognition capabilities in complex environments, improving precision and robustness in detection tasks.

（2）Bidirectional Feature Pyramid Network (BiFPN)

BiFPN significantly enhances feature integration efficiency and effectiveness by introducing bidirectional feature fusion paths and weighted feature fusion. This bidirectional connection not only enhances the flow of information between features extracted at different scales but also optimizes the model's attention distribution to key features by adjusting the influence of each feature through learnable weights. This helps the model more accurately identify objects in images, especially in scenarios involving size variations or different perspectives.

（3） DyHead

DyHead is a dynamic head network that introduces adaptive feature fusion and attention mechanisms during the model's prediction phase, enhancing the model's prediction precision and efficiency. DyHead dynamically adjusts the interaction and fusion strategy between feature layers, allowing the model to respond more flexibly to different inputs and scenarios. This dynamic adjustment directly affects how the model allocates its attention, focusing more on decisive features, thereby optimizing the overall detection performance.

By integrating the DBB, BiFPN, and DyHead modules, the YOLOv8n-modify model's attention mechanism is significantly enhanced. These modules work together to improve feature extraction, strengthen interactions between features, and optimize the model's dynamic response capabilities, collectively enhancing the model's ability to recognize and process key information in images. These improvements enable the model to exhibit higher precision in image processing and even better performance and adaptability in diverse environments.

Chapter 6: Results of Study 2

6.1 Performance of the Marigold Model

The Marigold depth estimation model's performance in generating depth maps of Jiangnan gardens represents a significant leap in digital imaging and analysis technologies. By adapting and finetuning a latent diffusion model for the task of depth estimation, the Marigold model offers an innovative approach to understanding and preserving the intricate landscapes of these historic sites. This section evaluates the performance of the Marigold model across various parameters critical for effective depth perception in complex natural and architectural environments.

6.1.1 Depth Map Accuracy and Detail

One of the most commendable aspects of the Marigold model is its exceptional accuracy in rendering depth maps. The precision with which the model captures the contours, elevations, and depressions of the garden terrain is striking. In traditional Jiangnan gardens, where subtle topographical variations play a significant role in overall aesthetics, the ability to accurately map these variations is crucial. Marigold's depth maps reveal not just the macro structures like pavilions and large rock formations, but also the finer landscape details such as small mounds, garden paths, and water edges which are essential for detailed architectural and archaeological analysis.

The high fidelity of these maps aids significantly in tasks requiring precise spatial reconstructions, such as virtual reality simulations and conservation planning. The level of detail ensures that digital replicas of the gardens are not only visually accurate but also spatially congruent with their realworld counterparts, providing a reliable foundation for virtual tourism and educational applications.

6.1.2 Robustness to Environmental Variabilities

Jiangnan gardens are characterized by their dynamic interaction with natural elements, particularly light and shadow, which can vary dramatically with the time of day and weather conditions. The robustness of the Marigold model in handling these variations is a testament to its advanced training and algorithmic design. The model's ability to differentiate between shadowinduced false depths and actual physical barriers is particularly noteworthy. This capability ensures that the depth maps remain consistent across different lighting conditions, providing reliability that is crucial for longitudinal studies and timelapse analyses where lighting conditions cannot be controlled.

Moreover, the model's performance in different seasons, capturing the seasonal dynamism of Jiangnan gardens—from the lush greenery of spring to the sparse, stark outlines of winter—demonstrates its adaptability and robustness. This seasonal resilience makes Marigold particularly suited for ecological and environmental studies, where understanding the impact of seasonal changes on garden topology is essential.

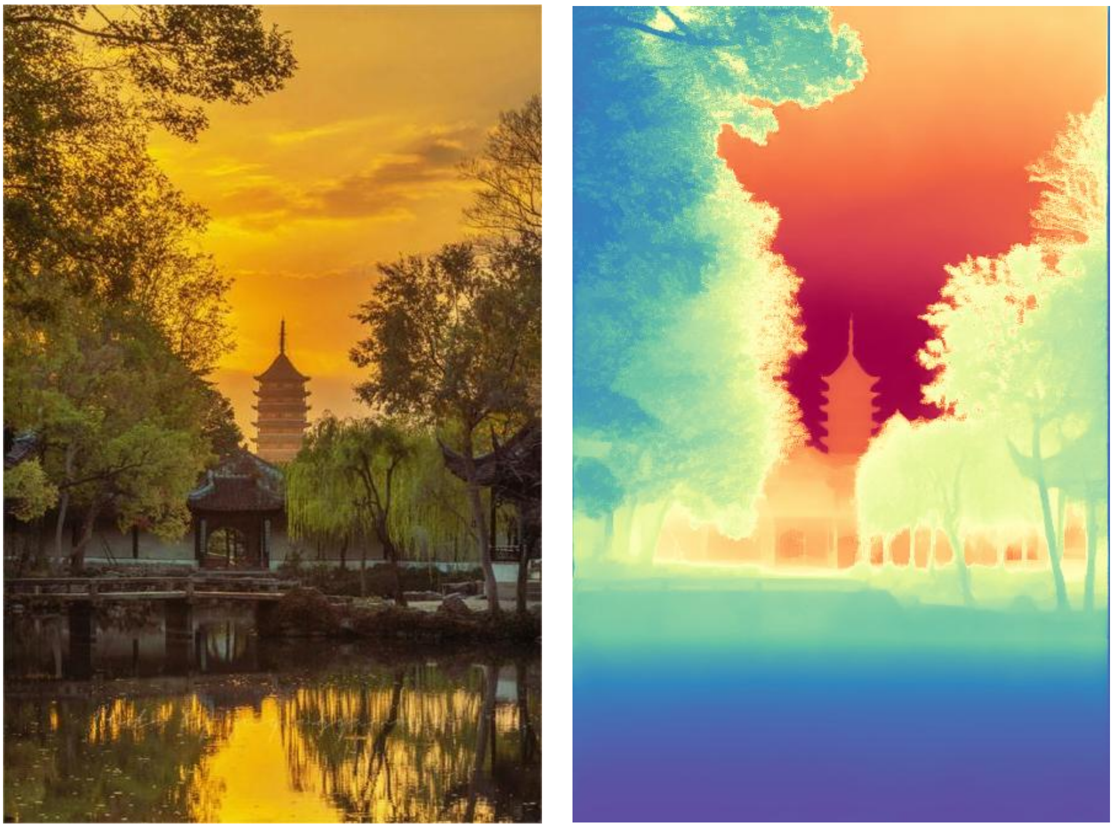
6.1.3 Handling of Complex and Overlapping Structures

The intricate layouts of Jiangnan gardens, with their overlapping foliage and multilayered design, pose significant challenges to depth estimation. Marigold's performance in managing these complexities is facilitated by its sophisticated multiresolution analysis capabilities, which allow the model to accurately parse and differentiate between closely situated or overlapping garden elements. This is crucial for maintaining the space and relational positioning that is characteristic of Jiangnan garden design, where the visual and spatial arrangement of elements is often intended to convey philosophical or poetic meanings.

Overall, the performance of the Marigold model in generating depth maps of Jiangnan gardens is exemplary. Its precision, robustness to environmental variabilities, adept handling of complex structures, and operational efficiency position it as a leading tool in the field of cultural heritage preservation and landscape analysis. The depth maps produced by the Marigold model not only serve as fundamental tools for conservationists and historians but also enrich the scientific community's understanding of the spatial dynamics at play in these historic sites. The ongoing improvements and adaptations of the model promise even greater contributions to the fields of archaeology, architecture, and landscape design, making it a cornerstone technology in the preservation of global cultural heritage.

6.2 Analysis of the depth map of specific scenes in the garden

6.2.1 The Zhuozheng Garden borrows the scenery of the North Temple Pagoda



The real image The depth map

Figure 6-1: The Zhuozheng Garden borrows the scenery of the North Temple Pagoda

In the heart of the Zhuozheng Garden, a masterstroke of classical Chinese landscape design is vividly demonstrated through the strategic incorporation of Beisi Pagoda's majestic silhouette, ingeniously borrowed from its distant location within the Bao'en Temple, situated 1.5 kilometers away(Figure 6-1). This artistic technique of "borrowing scenery" (借景), a quintessential element of traditional Chinese garden design, is exemplified here, not only enhancing the spatial depth of the garden but also enriching its narrative and aesthetic complexity.

Architectural Harmony and Visual Integration

As one meanders from the eastern part of the Zhuozheng Garden, passing through the Central Garden's arched gateways, they arrive at Yihong Pavilion, a vantage point that offers a breathtaking view of the garden's sophisticated design. Here, the viewer is greeted by a tranquil expanse of water that stretches outwards, intersecting elegantly with a curved bridge and a square pavilion. It is from this perspective that the grandeur of Beisi Pagoda looms into view, its reflection shimmering in the water, creating a layered visual feast of near and far, of reality and reflection.

This view cleverly integrates the towering pagoda, projecting its stature and spiritual significance into the garden's landscape. The reflection of the pagoda in the garden's pond adds a layer of depth—both literal and metaphorical—to the scene. The "distant" blur of the pagoda, the "height" of its spires, and the "depth" of its reflection converge to animate the entire garden, transforming it into a living painting that embodies dynamism and tranquility in equal measure.

Historical and Cultural Context

The Beisi Pagoda, originally erected during the period of the Three Kingdoms, boasts a venerable history of over 1700 years. It stands not only as a historical monument but also as a testament to the spiritual and cultural fabric of the era. Meanwhile, the Zhuozheng Garden itself, with its nearly 500year history, has been a canvas for dynastic changes, artistic evolution, and philosophical reflections through landscape art.

The decision to incorporate the view of Beisi Pagoda into the garden was a strategic one, addressing the inability to construct such grand structures within private estates due to restrictions of the era. By visually borrowing the pagoda, the garden designers effectively expanded the perceived boundaries of the garden, introducing a depth of field and a grandeur that the private space could not physically accommodate. This technique not only solved a spatial and regulatory limitation but also enhanced the garden’s aesthetic and spiritual dimensions, making the garden scene more profound and contemplative.

Artistic Techniques and Viewer Experience

The layout of the garden elements that lead to the pagoda view is intentionally designed to enhance the visitor's experience. As people walk through the garden, each feature is placed strategically to direct their attention and path, ultimately revealing the pagoda. This progression through the garden resembles a story unfolding, with each step introducing a new aspect, culminating in the dramatic sight of the pagoda surrounded by the garden's lush vegetation and architectural features.

In conclusion, the Zhuozheng Garden utilizes the technique of borrowing scenery to create a landscape that is deeply rooted in its cultural and historical context, while also offering a transcendent aesthetic experience. The integration of Beisi Pagoda into the visual narrative of the garden is a masterful use of natural and architectural harmony, enhancing the garden's depth and expanding its narrative scope. This strategic design choice not only showcases the ingenuity of classical Chinese garden design but also underscores the profound connection between landscape, architecture, and cultural memory in traditional Chinese aesthetics.

Depth Map Analysis by Marigold Model

Depth, Continuity, and Smoothness

The Marigoldgenerated depth map demonstrates a reasonable gradient from the foreground, where the garden elements are, to the background, featuring the pagoda and distant landscapes. This effective estimation indicates that the model accurately captures the spatial depth essential for realistic 3D representations, maintaining the visual flow that guides the viewer’s eye through the garden’s layout.

Edge Preservation

While the depth image retains the overall silhouette of the pagoda and surrounding trees reasonably well, it tends to lose some finer details such as branches and leaves. These elements appear overly smoothed or absent, likely due to the model's prioritization of larger, more distinct features. This prioritization is a common tradeoff in depth estimation processes, where the focus is on ensuring the clarity of major structural elements at the expense of fine textural details.

Handling of Overlapping Structures

The depth map exhibits proficiency in distinguishing between overlapping elements, such as branches in front of the pagoda or reflections in the water. However, there is some indistinct separation, which can occasionally lead to a somewhat flattened appearance where depth cues are crucial. Enhancing the model’s ability to differentiate these elements more clearly could further improve the depth perception and add to the realism of the 3D scene.

Light and Shadow in Depth Encoding

The depth image employs variations in brightness to encode depth, with brighter areas signifying closer proximity to the viewer and darker areas indicating greater distances. This method helps to articulate the spatial relationships within the scene effectively, contributing to the overall threedimensional feel of the image.

Macro and Micro Spatial Relationships

In the depth image, the algorithm captures both macro and micro spatial relationships well, distinguishing effectively between the foreground, midground, and background. This distinction enhances the depth accuracy and the overall threedimensional sensation of the scene, which is essential for digitally narrating the spatial borrowing techniques employed in the Zhuozheng Garden.

Conclusion

The analysis of the Marigold model's performance in generating depth maps of the Zhuozheng Garden, particularly focusing on the iconic Beisi Pagoda, offers a nuanced understanding of its capabilities and limitations in the context of heritage conservation and digital rendering. While the model demonstrates robust capabilities in capturing the overall spatial relationships and structural outlines, there are specific areas where refinement is necessary to fully realize the potential of digital depth mapping technologies for cultural and architectural applications.

The depth maps generated by the Marigold model effectively illustrate the space between the garden's foreground and the borrowed scenery of Beisi Pagoda in the background. The model successfully captures the grandeur and vertical dominance of the pagoda over the garden's landscape, which is crucial for maintaining the intended aesthetic and philosophical impact of the garden's design. The visual transition from the garden's intimate spaces to the imposing structure of the pagoda is smooth, reflecting the garden's design principle of guiding the viewer's gaze through a progressively unfolding natural tableau.

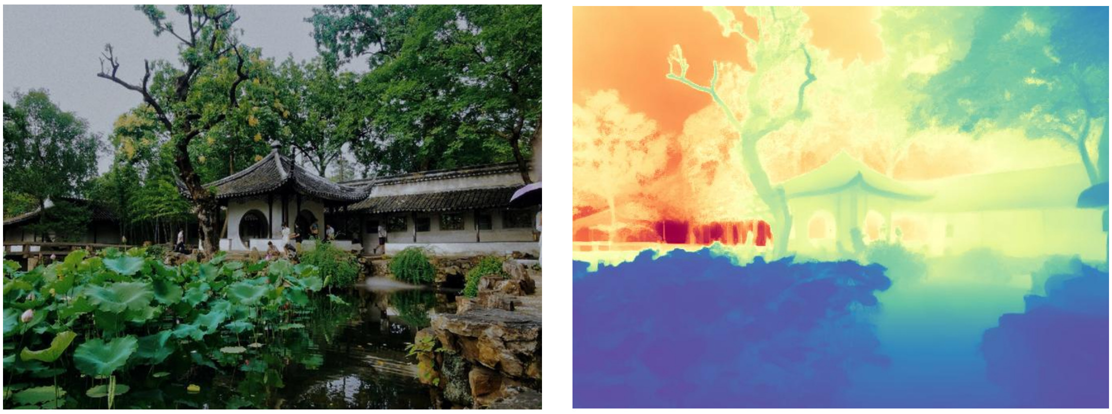
However, the depth maps show limitations in handling the finer details and textures, particularly in areas where complex overlapping structures, such as branches and leaves in front of the pagoda, are present. These elements are crucial for conveying the full texture and density of the garden's lush vegetation, yet they are often smoothed over or lost in the depth translation. This oversimplification can detract from the realistic portrayal of the garden's rich botanical details and may impact the accuracy of virtual recreations.

The Marigold model's depth maps successfully convey the 'borrowed scenery' technique, a fundamental aspect of Jiangnan garden aesthetics. The depth map maintains a clear distinction between the garden elements and the distant pagoda, effectively using gradations of color and light to denote spatial depth. This aspect of the depth map is particularly successful in illustrating how the garden space extends beyond its physical boundaries to include distant landmarks, enhancing the viewer's perception of depth and scale.

However, the effectiveness of these gradations in representing the true depth and layering of the scene could be enhanced by improving the model's ability to handle shadow and light interplay more subtly. The areas where light interacts with water and reflective surfaces need a more refined approach to capture the ephemeral qualities of light and reflection that are so characteristic of Jiangnan gardens.

In conclusion, while the Marigold model offers a promising approach to the digital preservation and analysis of cultural heritage sites, targeted improvements are necessary to fully capture the complex beauty and historical depth of Jiangnan gardens. These enhancements will not only improve the technical quality of the depth maps but also enrich the interpretive possibilities of these digital tools, bridging the gap between historical authenticity and modern technological capabilities.

6.2.2 Scene 2: The perspective of Wuzhu Youju Pavilion



The real image The depth map

Figure 6-2: The The perspective of Wuzhu Youju Pavilion

The Wuzhu Youju (Figure 6-2) is located at the eastern end of the central pool in the Zhuozheng Garden. It is a pavilion-style building with a square floor plan and a single-eaved, four-cornered spire roof, originally built during the Qing Dynasty. The pavilion backs onto a double-decker corridor that separates the eastern and central parts of the garden, faces a large pool to the west, and is north of a small curved bridge that connects to a mountain island in the pool and faces the Luyi Pavilion across the water. Unlike typical garden pavilions, this pavilion uses walls instead of columns and is surrounded by white walls on all sides. To facilitate the enjoyment of the scenery and communication between inside and outside, four circular openings are made on the square walls. Sitting on the stone bench at the center of the pavilion, one can view the landscape from different angles, much like admiring ancient round fan landscape paintings.The pavilion has four circular doorways, creating framed views through which one can enjoy the seasonal changes of spring mountains, summer lotuses, autumn leaves, and winter bamboo. In spring, it borrows the fluttering of primroses; in summer, it is adorned with lotuses; in autumn, it draws from the upright bamboo; and in winter, it captures the crystalline icicles.

The Marigold model, applied to the scenic depictions of the Zhuozheng Garden, specifically focusing on the Wuzhu Youju Pavilion, showcases the advanced capabilities and sophistication of modern depth estimation algorithms. This analysis evaluates the performance of the Marigold model in generating depth maps that effectively capture and enhance the scenic and cultural richness of Jiangnan gardens.

Depth Continuity and Smoothness

The depth maps generated by the Marigold model exhibit excellent depth continuity and smoothness, essential attributes for recreating the natural flow of garden landscapes. This seamless gradient is most evident as the viewer's eye transitions from the vibrant foreground, populated with water lilies, to the intricately structured Wuzhu Youju Pavilion in the background. This smooth gradation aids in reinforcing the spatial coherence of the garden, which is crucial for maintaining the narrative and aesthetic integrity of these culturally significant sites.

Edge Preservation

A notable strength of the Marigold model is its ability to preserve sharp edges within complex scenes, a critical factor in depth perception and threedimensional representation. The distinct outlines of trees, pavilion structures, and other architectural elements are maintained with high fidelity, ensuring that each component stands out clearly against the background. This sharp delineation is particularly important in Jiangnan gardens, where the clarity of architectural forms and natural elements plays a significant role in the overall visual impact of the scene.

Handling of Overlapping Structures

The Marigold model adeptly manages overlapping structures, effectively distinguishing between layers of foreground vegetation and architectural backdrops. This sophisticated depth layering capability allows the model to create a perceptible distinction between different spatial planes, enhancing the threedimensional realism of the garden. It can well distinguish the pavilion from the lotus in front, and from the branches in the back.The ability to differentiate these layers without confusion is testament to the model's advanced processing capabilities, which are crucial for detailed landscape analysis and cultural heritage conservation.

Light and Shadow Dynamics

The model's handling of light and shadow showcases its finesse in utilizing natural lighting to enhance depth perception. Shadows and light patches are not merely reproduced but are strategically used to accentuate depth cues, contributing significantly to the threedimensional feel of the scene. This approach is particularly effective in Jiangnan gardens, where the interplay of light and shadow is often used to create mood and highlight seasonal changes.

Color and Texture Processing

Color and texture processing in the Marigold model are intelligently adapted to maintain the scene's natural aesthetics while emphasizing depth through subtle shifts in hue and saturation. This careful adjustment ensures that the visual richness and detail of the original settings are preserved, allowing for a more authentic and engaging viewer experience. The model’s capability to handle these subtle nuances in color and texture is indicative of its potential to produce visually rich and accurate representations of complex environments.

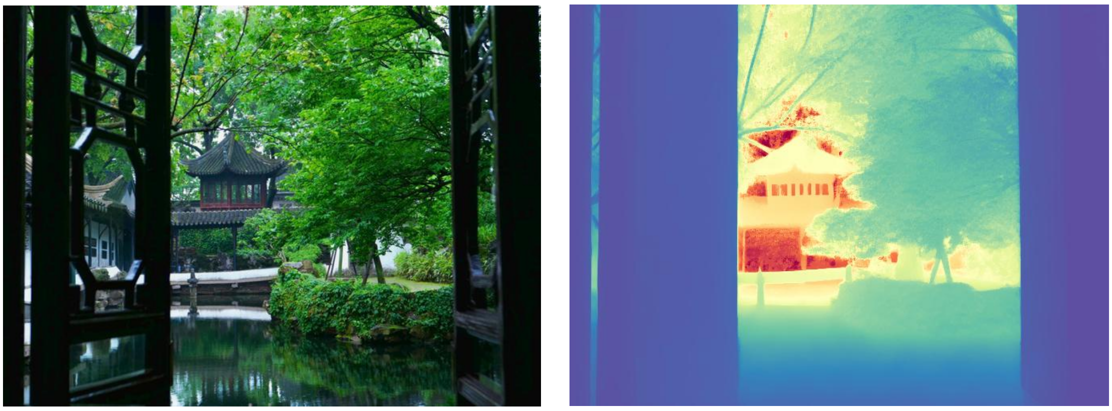
MultiScale Feature Learning

The incorporation of multiscale feature learning in the Marigold model allows it to effectively capture both fine details and broader structural elements. This feature is crucial in environments like Jiangnan gardens, where the diversity of visual elements ranges from the delicate textures of plants and water surfaces to the robust forms of pavilions and sculptures. The model's ability to represent all aspects of the scene with clarity and depth accuracy is a significant advantage for applications in cultural heritage visualization and sophisticated landscape analysis.

Conclusion

The performance of the Marigold model in the context of Jiangnan gardens demonstrates its excellence in depth mapping and its applicability to cultural heritage visualization. The model's ability to accurately capture and enhance the spatial and aesthetic qualities of the garden scenes, together with its sophisticated handling of complex visual data, positions it as an invaluable tool for both conservationists and researchers interested in the digital preservation and analysis of historic landscapes.

6.2.3 Scene 3: The Perspective Obscured by a Foreground Window



The real image The depth map

Figure 6-3: The Perspective Obscured by a Foreground Window

Foreground to Background Transition

The depth map effectively layers the garden scene, starting from the water in the foreground to the pavilion in the background(Figure 6-3). The transition is handled in such a way that there is a perceptible gradation in depth, which helps the observer's eye move naturally through the space from the water's edge to the distant structures. This gradation is crucial in a garden setting, where the sense of depth contributes to the overall serenity and aesthetic pleasure of the scene.

Edge Clarity

In the depth map, the edges of the pavilion and the trees are distinctly marked, preventing the blending of these key architectural and natural elements with the background. This clarity ensures that the pavilion, an important cultural symbol, stands out against the softer edges of the garden's foliage. Clear edge definition is vital in maintaining not just the visual integrity but also the focus on significant elements within the garden.

Overlapping Structures

The treatment of overlapping structures like trees in front of the pavilion is done with care, ensuring that each maintains its visual independence. This separation is subtle but effective, allowing for each component to be appreciated individually while contributing to the whole. The algorithm’s ability to maintain this separation helps in preserving the layered look typical of Jiangnan gardens, where depth and perspective are used to create a sense of expansive space.

Lighting and Shadowing Effects

The depth map uses variations in color intensity to simulate shadows and highlights, which adds a layer of realism to the image. This simulation helps in enhancing the threedimensional feel of the scene, making the digital image more lifelike. The subtle use of shadowing under the pavilion and around the trees adds depth, suggesting the time of day and the direction of sunlight, which are important for setting the scene's mood.

In this Jiangnan garden scene, the depth map generated by the Marigold algorithm not only enhances the visual appeal of the garden but also adds a layer of interpretative depth that could be useful for various applications, including virtual tours and educational content. The ability to distinctly navigate through layers of the scene—water, foliage, and architecture—via a digital interface showcases the potential of depth mapping technology in bringing traditional garden landscapes to life in a virtual or augmented reality setting.

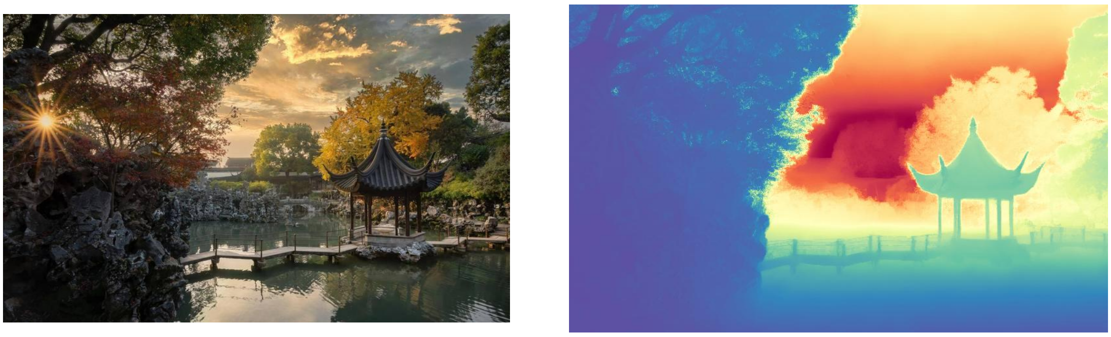
This precise and nuanced handling of traditional elements in a depth map underscores the potential of such technology in preserving and interpreting cultural heritage landscapes, providing a bridge between traditional aesthetics and modern visualization techniques.

6.2.4 Scene 4: Garden view at dusk

This depth map of a traditional Jiangnan garden scene, centered on a beautifully structured pavilion surrounded by lush foliage and rock formations, is a masterful representation of depth and perspective, achieved through the Marigold algorithm. The analysis will focus on how the depth map enhances the structural integrity, key features, and overall aesthetic of the garden scene.

Structural Integrity and Key Features

The depth map preserves the architectural essence and intricate details of the pavilion, which serves as the focal point of this garden scene(Figure 6-4). The contours and edges of the pavilion are sharply defined against the backdrop, which ensures that this cultural symbol stands out prominently. This clarity is crucial not only for visual appeal but also for maintaining the architectural integrity within the context of the garden's design.



The real image The depth map

Figure 6-4: The garden view at dusk

The surrounding rock formations and the intricate details of the tree branches are also well captured, each layer distinctly separated from the others. This separation is vital for understanding the layout and space inherent in Jiangnan garden design, where every element has both an aesthetic and symbolic function.

Light and Shadow Play

The depth map effectively uses the contrast between sunlit areas and shadows to simulate the natural interplay of light within the garden. This contrast not only enriches the threedimensional perception but also keeps the scene visually balanced, aligning it with realworld perspectives. The way light filters through the foliage and casts shadows on the pavilion and rocks adds depth and realism, making the digital representation more lifelike and immersive.

Depth Continuity and Smooth Transitions

The depth map showcases excellent continuity, with a smooth gradient that transitions seamlessly from the foreground elements to the background. This smooth transition across various spatial layers helps guide the viewer's eye through the garden scene, enhancing the natural flow and spatial narrative of the Jiangnan garden. Such depth management is critical for preserving the viewer's sense of immersion and for facilitating a deeper understanding of the garden's layout.

Handling of Overlapping Structures

The clarity and distinction with which overlapping structures like trees in front of the pavilion are handled are commendable. The depth map ensures that these elements do not merge into a single plane, which is a common challenge in twodimensional representations. Instead, each element is given its own space and depth, which contributes to the overall depth accuracy and enhances the viewer's ability to distinguish between the various components of the garden.

MultiScale Feature Learning

The application of multiscale feature learning by the Marigold algorithm allows for an intricate capture of both fine and extensive details, offering a comprehensive and immersive depth analysis. This technique ensures that both the macro elements of the garden, such as the pavilion and large rocks, and the micro elements, such as individual leaves and smaller plant details, are equally emphasized. This balance is critical for a holistic representation that respects the garden's natural and architectural elements.

Conclusion

The Marigold algorithm’s depth map of this Jiangnan garden scene excels in enhancing the structural integrity, handling complex overlapping structures, and simulating natural light play, which are all crucial for a realistic and engaging representation. By maintaining the visual richness and providing a detailed depth perspective, this technological application not only serves as a tool for digital conservation and virtual tourism but also as a means to deepen appreciation for the cultural and aesthetic values embodied in traditional Chinese gardens.

6.3 Comparison results with other models in different scenarios

6.3.1 Scene 1: The Zhuozheng Garden borrows the scenery of the North Temple Pagoda

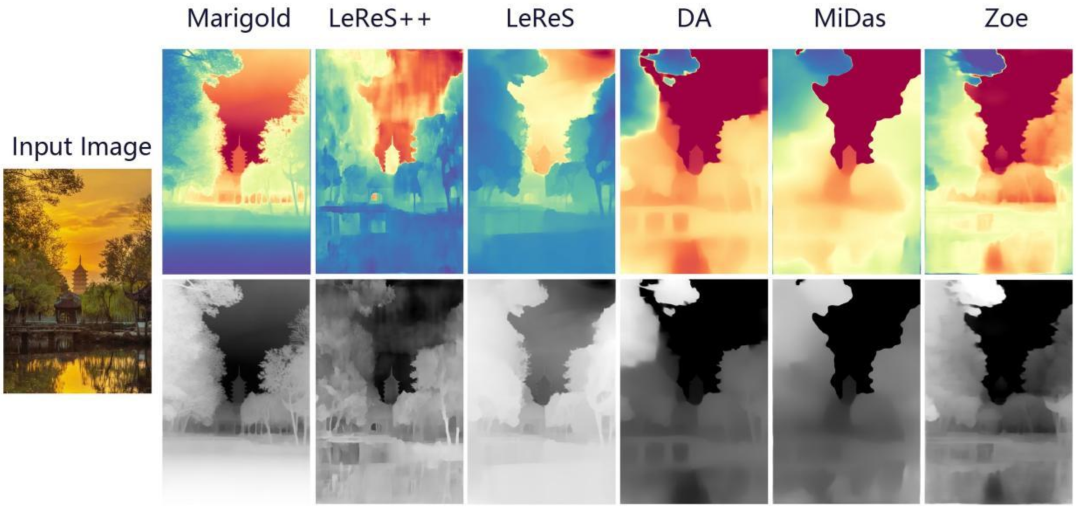


Figure 6-5: The Comparisons of different models of Zhuozheng Garden borrows the scenery of the North Temple Pagoda

In the comparative analysis of depth map algorithms applied to a classic scene from The Zhuozheng Garden, which artfully incorporates the distant North Temple Pagoda into its visual narrative, the performance of Marigold stands out against other models such as LeRes, MiDas, DA, LeRes++, and Zoe(Figure 6-5).

Depth Rendering and Color Gradation

Marigold excels in rendering depth with a refined approach to color gradation, enhancing the visual layering from the foreground to the background elements of the scene. This capability allows Marigold to provide superior edge clarity, making it easier to distinguish between the intricate details of the garden's architecture and its natural elements. In contrast, models like LeRes and MiDas tend to produce muddier transitions in areas where distinct separation between elements is crucial. Their less precise handling of color gradation blurs the lines between different planes, reducing the depth accuracy and visual clarity of the scene.

Edge Clarity and Transition Smoothness

Unlike Zoe, which produces vivid contrasts yet struggles with smooth depth transitions, Marigold maintains a balance between vibrant color contrasts and seamless depth progression. This balance is crucial for scenes like The Zhuozheng Garden, where the gradual transition from the water's edge to the lush foliage and ultimately to the architectural marvel of the North Temple Pagoda must be depicted with both clarity and aesthetic finesse.

Overlapping Structures and Detail Preservation

While DA and LeRes++ show adeptness in handling overlapping structures, they fall short of Marigold in maintaining the natural look and detailed textures that are signature to Jiangnan garden aesthetics. Marigold's sophisticated shadow response and preservation of multiscale details ensure that every aspect of the scene, from the smallest leaf to the sweeping arches of the pavilion, is rendered with high fidelity. These features are critical for achieving realistic depth perception in complex scenes, where the interaction between light, shadow, and texture plays a pivotal role in the overall realism.

Conclusion

In this comparative evaluation, Marigold's ability to render depth with enhanced color gradation and superior edge clarity positions it as the leading choice for scenarios that require detailed and realistic depth mapping. Its proficient handling of complex garden scenes, with a particular strength in maintaining natural aesthetics and detailed textures, sets it apart from other models. The results underscore Marigold’s suitability for cultural heritage visualization, where accuracy in depth perception and the preservation of visual integrity are paramount.

6.3.2 Scene 2: Pavilion by the water at dusk

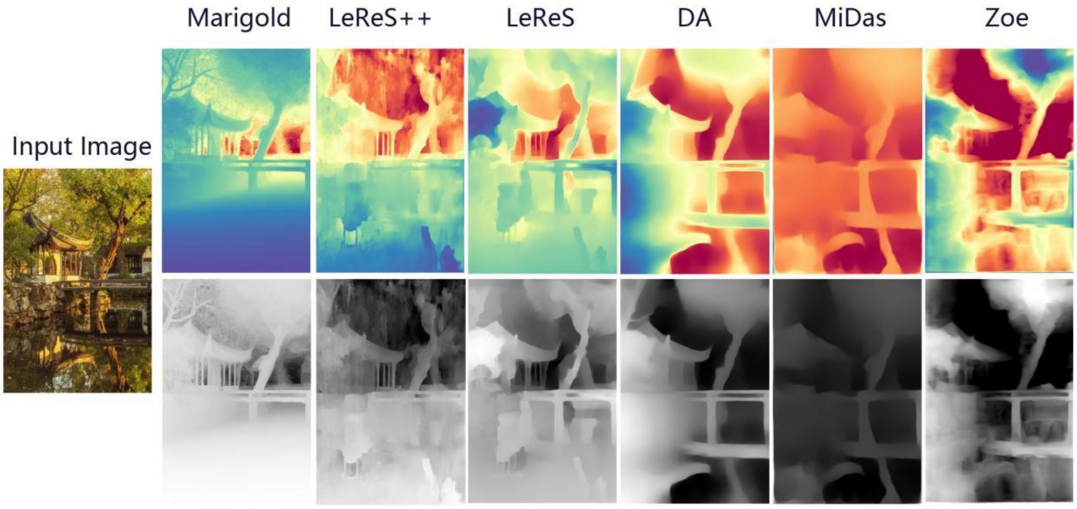


Figure 6-6: The Comparisons of different models of Pavilion by the water at dusk.

In the scene of pavilion by the water at dusk(Figure 6-6), Marigold stands out for its effective rendering of depth with enhanced color gradation, which is evident in the vivid and distinct transitions from the foreground elements to the architectural features in the background. This model excels in maintaining edge clarity, which is critical in scenes with complex structures and overlapping elements. The edges of pavilions and foliage are sharply defined, allowing these elements to stand out against a smoothly graduated background, enhancing the spatial narrative of the scene.

LeRes and its advanced version LeRes++ show different strengths and weaknesses. While LeRes++ manages overlapping structures better than its predecessor, both struggle to achieve the natural look that Marigold offers. Their depth maps tend to lack the fine detail and texture fidelity found in Marigold's outputs, particularly in how they handle light and shadow interplay, which is essential for realistic depth perception.

DA exhibits strong capabilities in handling overlapping structures, possibly surpassing Marigold in this regard. However, it does not maintain the natural appearance of the scene as effectively. DA’s depth maps often appear slightly artificial or enhanced, lacking the subtle gradations of light and texture that contribute to a lifelike portrayal of the garden environment.

MiDas tends to produce depth maps with muddier transitions in complex scenes, which can obscure the finer details and textures of the garden’s elements. While it provides strong contrast in some areas, it lacks the finesse required to accurately represent the delicate balance of light and shadow that characterizes Jiangnan garden scenes.

Zoe produces depth maps with vivid contrasts and is particularly strong in highlighting the most dramatic elements of a scene. However, it often fails to provide the smooth depth transitions that are necessary for a realistic and cohesive depth perception. Zoe's outputs can sometimes appear overly stylized, which might detract from the naturalistic portrayal required for historical or cultural heritage visualizations.

In comparing these models, Marigold’s depth maps are superior in terms of color gradation and edge clarity, crucial for detailed and realistic depictions of traditional gardens. Its ability to balance vivid color contrasts with smooth depth transitions allows for a more authentic and engaging visualization of the scene. Although DA and LeRes++ handle overlapping structures effectively, they do not achieve the same level of detail and natural look that Marigold does, especially in complex lighting conditions.

Marigold’s proficiency in multiscale detail preservation and its sophisticated shadow response make it an excellent tool for cultural heritage visualization and landscape analysis. Its outputs not only preserve the visual integrity of the scene but also enhance the viewer's understanding of spatial dynamics, making it particularly valuable for educational and preservation purposes related to Jiangnan gardens.

6.3.3 Scene 3: Perspective Obscured by a Foreground Window

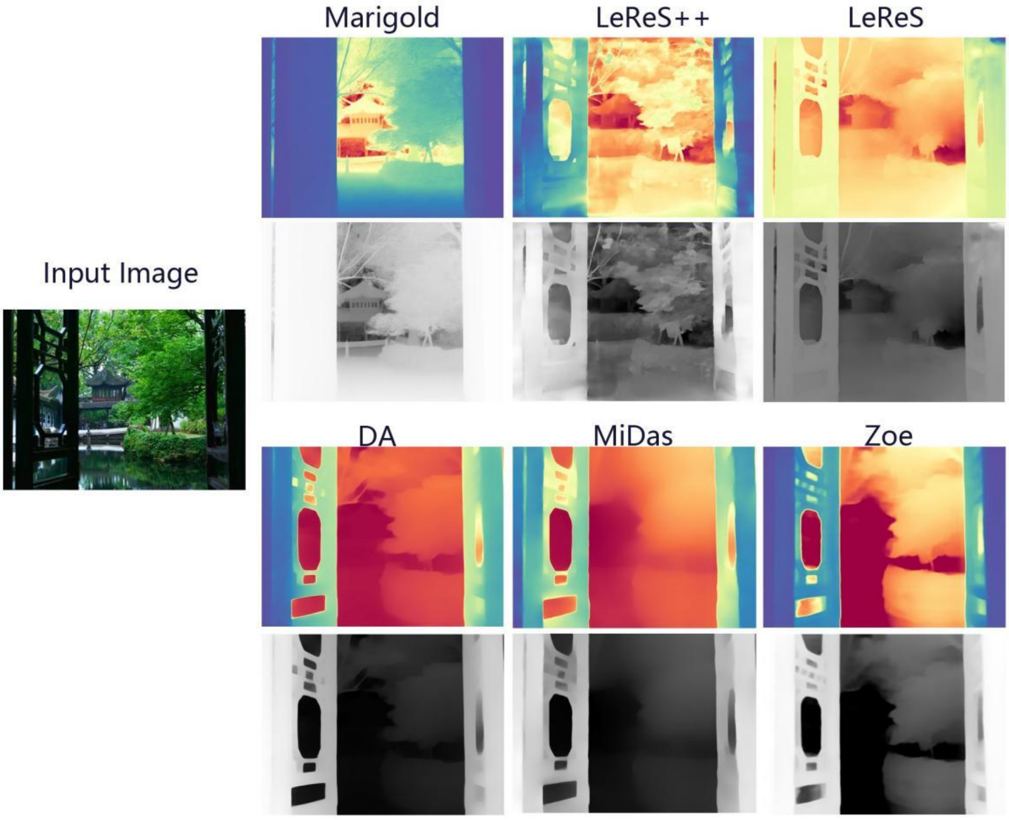


Figure 6-7: The Comparisons of different models of Perspective Obscured by a Foreground Window

In the scene of perspective obscured by a foreground window(Figure 6-7), Marigold uses vivid and varied color gradients to emphasize depth layers, enhancing the visual differentiation between foreground, midground, and background elements. The use of bright, saturated colors for closer objects and cooler tones for distant objects helps in creating a strong depth perception.Marigold maintains a good balance of detail across different parts of the image. The structure of the pavilion and the surrounding foliage are both distinct and wellseparated, aiding in the clarity of the scene’s spatial layout.

There are also some disadvantages, such as over-smoothing.Some areas, especially where fine details like branches and leaves are present, appear slightly smoothed out, potentially losing some textural nuances of the garden.

LeRes and LeRes++

Both versions, especially LeRes++, handle overlapping structures better, effectively distinguishing between the pavilion’s architecture and the surrounding trees.There are also some disadvantages. The depth maps tend to lack the natural look and feel that Marigold offers, with some areas appearing overly processed or artificial, particularly in how they handle shadow and light transitions.

DA (Depth Anything)

DA exhibits a robust dynamic range, displaying sharp contrasts between light and dark areas, which can help in accentuating depth cues in certain aspects of the scene.But The depth map can appear slightly artificial due to the aggressive contrast, which might not always convey the subtlety of natural light as effectively.

MiDas

Effective at discriminating different depth levels, which is visible in the clear demarcation between the pavilion and the background landscape.The depth map often lacks the color vibrancy seen in Marigold, making the scene feel less lively and dynamic.

Zoe

Zoe produces vivid contrasts and is particularly strong in highlighting dramatic elements of the scene through high saturation and contrasting colors.The approach can sometimes result in a loss of subtlety in depth transitions, making the scene appear less realistic and more stylized than might be desired for accurate depth perception.

Summary of Comparative Analysis

In this comparative evaluation, Marigold’s depth maps are noted for their superior color gradation and edge clarity, which are crucial for depicting complex and detailed garden scenes with a high degree of realism. Marigold's ability to maintain a natural appearance while providing detailed textural information sets it apart from other models, which sometimes struggle with either artificiality or lack of detail differentiation.

LeRes++ and DA show promise in handling complex structures and dynamic ranges but do not achieve the same level of natural look and fine detail preservation as Marigold. MiDas and Zoe, while effective in certain aspects, tend to either underperform in realistic color representation or overshoot in stylization, respectively.

The results highlight Marigold’s suitability for applications requiring accurate and visually appealing depth perception, making it a preferable choice for cultural heritage visualization and sophisticated landscape analysis.

6.3.4 Scene 4: Garden view at dusk

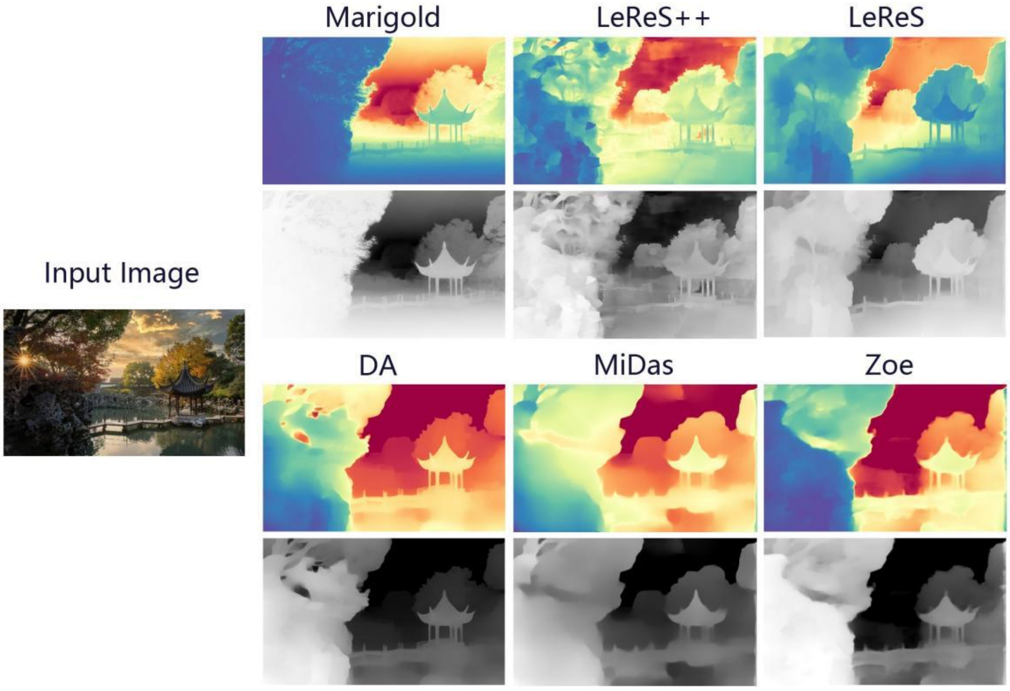


Figure 6-8: The Comparisons of different models of garden view at dusk

In this depth map comparison(Figure 6-8), we observe how different algorithms interpret the serene scene of a traditional Jiangnan garden at dusk, with the focus on a small pavilion by a pond surrounded by lush trees. This scenario is excellent for examining how each algorithm handles transitions between varying depths, color gradation, and the clarity of structural elements against a complex natural background.

Marigold

Marigold produces a vibrant and color-rich depth map that emphasizes the foreground elements with warm hues while cooling the background to enhance the sense of depth. This results in a visually appealing contrast that captures the eye and clearly distinguishes between different spatial planes.

It maintains excellent detail in the pavilion’s architecture, where the edges are crisply defined against the softer background, helping to preserve the structural integrity of the garden's focal point.

While it excels in color differentiation, Marigold might slightly over-emphasize color saturation, which could potentially distract from finer texture details, especially in areas of dense foliage.

LeRes and LeRes++

Both algorithms show a competent handling of depth transitions, with LeRes++ providing a more nuanced gradient that better reflects the natural flow of the scene from the water in the foreground to the trees in the background.They manage to keep the pavilion discernible, though with less color vibrancy compared to Marigold.

The depth maps can appear somewhat washed out, lacking the depth of color that might be necessary to fully convey the mood of a dusk scene.

DA (Depth Anything)

DA uses sharp contrast to define edges, which could be effective for scenes requiring clear delineation between elements such as water, pavilions, and sky.It shows a good balance in shadow regions, helping to add depth without losing detail in darker areas.

The high contrast approach may result in an unnatural feel, particularly in a natural setting like a garden, where softer transitions might better represent the ambient lighting.

MiDas

MiDas offers a decent overall depth perception, with the vertical structures like trees and the pavilion rendered with reasonable accuracy.It provides a solid middle ground in terms of both color balance and edge clarity, making it a versatile choice if less artistic stylization is desired.

The algorithm may struggle with color depth, particularly in rendering the lushness of vegetation with the same vibrancy that Marigold achieves.

Zoe

Zoe produces the most dramatic and artistic interpretation, with bold color shifts that could be appealing for more stylized visual applications.It emphasizes the silhouette of the pavilion against a dramatically colored sky, which could be useful for highlighting architectural features in artistic renditions.

The artistic approach may not always be suitable for applications needing realistic depth mapping, as it sacrifices some detail for the sake of visual impact.

Conclusion

In this comparative analysis, Marigold demonstrates its strength in providing a visually rich and detailed depth map that enhances the scene's natural beauty while maintaining the clarity and integrity of critical architectural elements. LeRes and LeRes++ offer a more subdued, but accurate depth portrayal, suitable for more technical applications. DA and MiDas provide balanced alternatives with distinct advantages in edge definition and versatility, respectively. Zoe stands out for artistic interpretations but may not always align with realistic depth mapping needs. Each algorithm has its niche, making the choice dependent on the specific requirements of the visualization or analysis project.

Chapter 7: Discussion and conclusion

7.1 Discussion of Experimental Results

7.1.1 Discussion of study 1

(1) Integration of DBB into the Backbone Layer

The integration of the Dense Block Backbone (DBB) into the backbone layer of object detection models has significantly improved the precision of detecting various objects. This enhancement is primarily due to the DBB's innovative multibranch structure that enriches the model's feature space. Unlike traditional approaches that increase computational demands, the DBB maintains a balance, boosting detection capabilities without adding computational overhead during the inference phase.

DBB utilizes advanced multiscale convolution techniques, which allow the model to capture features at various scales effectively. This capability is critical in handling objects of different sizes within the same scene. Sequential convolution technologies further refine these features by optimizing the receptive field, which is crucial for recognizing finer details and contextual relationships in images.

Furthermore, the DBB architecture simplifies the computational process during the inference phase by consolidating multiple convolutions into a single convolution operation. This optimization reduces the time and resources required for processing, thus maintaining high efficiency without compromising on the model's performance. The strategic integration of these technologies ensures that the model not only performs well in standard benchmark tests but also operates efficiently in realworld applications where quick and accurate detection is essential.

Overall, the DBB's integration into the backbone layer represents a significant advancement in object detection technology. By enhancing feature extraction and optimization while controlling computational complexity, the DBB provides a robust framework that improves the precision and efficiency of object detection models, making them more practical for widespread use in various demanding applications.

(2) Implementation of BiFPN in the Neck Layer

The implementation of BiFPN (Bidirectional Feature Pyramid Network) in the neck layer of object detection models significantly enhances the network's capability to perform feature fusion efficiently and accurately. BiFPN improves upon traditional feature pyramid networks by optimizing crossscale connections, which are vital for integrating semantic information from different layers of the network. This optimized integration facilitates a more robust and nuanced understanding of the images processed, crucial for accurate object detection across various scales.

BiFPN simplifies the overall network structure by streamlining feature transmission paths. It minimizes redundancy by selectively enhancing relevant feature pathways and pruning less useful connections. This selective strengthening and elimination process not only reduces computational load but also ensures that the most informative features are efficiently utilized. The network achieves this by iteratively repeating the feature fusion steps, each time deepening the fusion process to incorporate more complex and comprehensive feature interactions.

Moreover, BiFPN introduces additional connections within the network without substantially increasing the computational demands. This is achieved through an innovative design that leverages both topdown and bottomup pathways, allowing for continuous refinement of features at every level of the network. These pathways enhance the flow of information, making the feature fusion process more dynamic and responsive to the varying demands of realtime object detection.

The iterative deepening of the feature fusion in BiFPN also means that each layer or level of the network can refine its understanding based on both the immediate and more globally processed features from other layers. This bidirectional exchange is crucial for developing a more accurate and scalable model that can adapt to different detection scenarios efficiently.

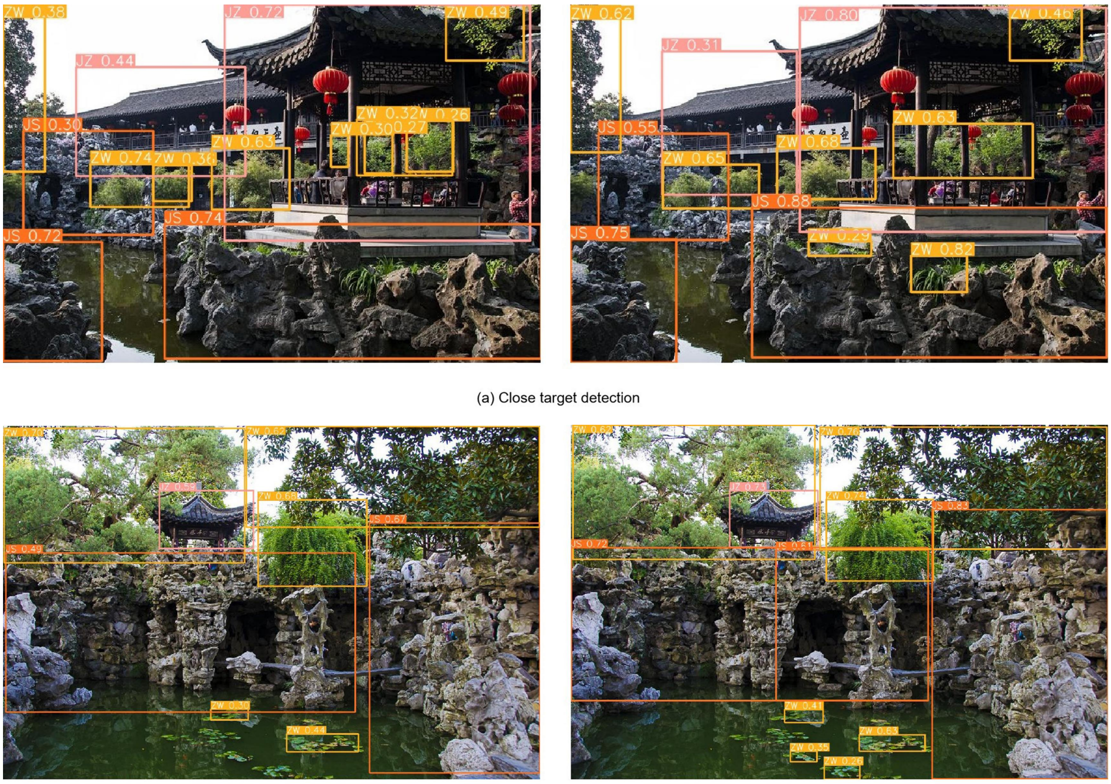
Overall, using BiFPN in the neck layer of detection models marks a significant improvement in computer vision technology. It boosts the effectiveness and scalability of combining features, and also increases the accuracy and stability of the model. This allows it to perform high-quality object detection across many difficult situations. Thus, BiFPN is an essential element in creating advanced object detection systems, enabling them to manage complex and diverse tasks more accurately and quickly.

(3) YOLOv8 Enhancement using DyHead

YOLOv8's integration of the DyHead, which utilizes advanced attention mechanisms in the head layer, marks a significant advancement in the representational capabilities of object detection systems. DyHead focuses on three critical types of attention: scaleaware, spatialaware, and taskaware. Each type of attention targets different aspects of the detection process, enhancing the model's ability to discern finer details and variations in object size, spatial relationships, and taskspecific requirements. This multifaceted attention mechanism allows YOLOv8 to manage information exchange more effectively across various feature levels and spatial positions, optimizing overall detection performance.

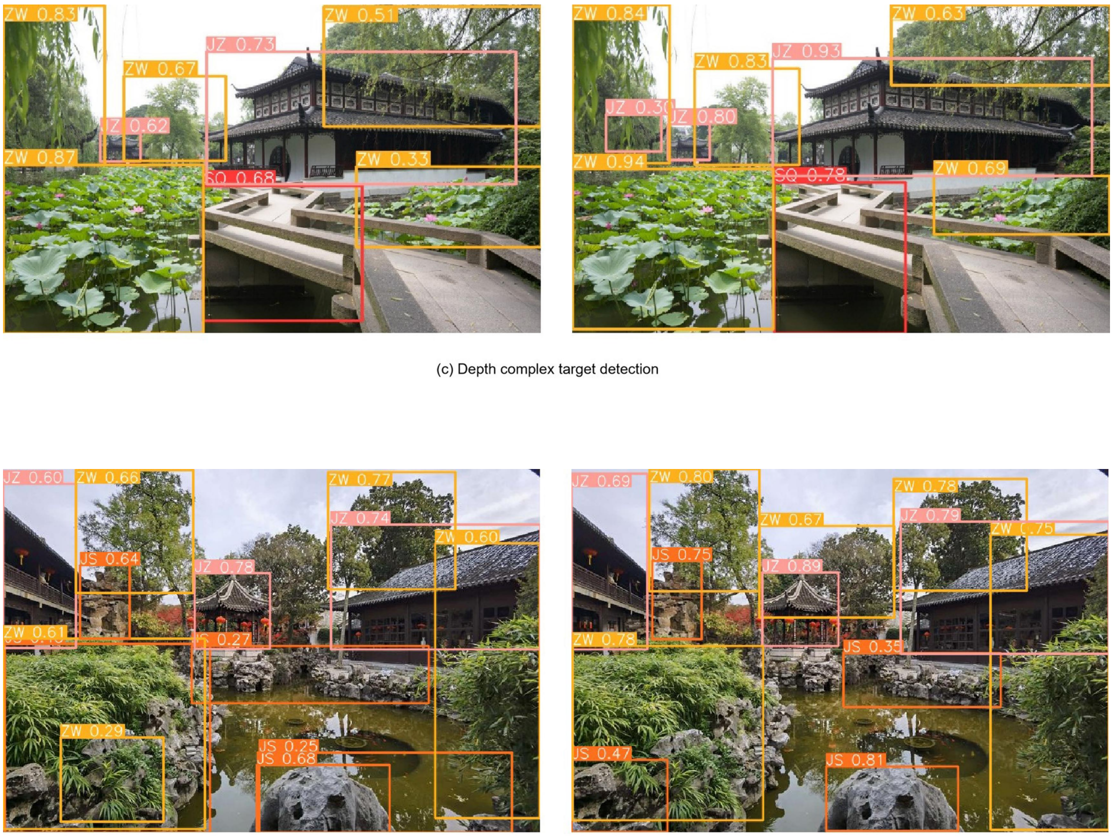
The scale attention in DyHead specifically addresses the challenge of detecting objects at different scales, which is crucial for applications ranging from pedestrian detection to vehicle identification in various environmental conditions. Spatial attention enhances the model's ability to focus on relevant areas within an image, improving precision by reducing background noise and distractions. Task attention, meanwhile, aligns the model′s processing power with specific detection tasks, ensuring that the computational resources are utilized most efficiently for the task at hand.

The incorporation of these attention mechanisms makes the DyHeadenhanced YOLOv8 a powerful tool in scenarios where traditional models might struggle with complexity and variability. The effectiveness of these enhancements is demonstrated through rigorous testing. Representative photographs from a test set are used to validate the improvements made by the YOLOv8modify version. The comparative results, as depicted in Figure 21, illustrate the clear distinctions in detection capabilities between the original YOLOv8n and the modified YOLOv8n-modify. The left and right images in these comparisons show how each model performs under identical conditions, highlighting the enhancements in detection precision and reliability introduced by the DyHead.



(a) Close target detection

(b) Mid-range target detection



(c) Depth complex target detection

(d) Complex background target detection

Figure 7-1: Comparison of YOLOv8n and YOLOv8n-modify in various scenes

Figure 7-1-a presents a detailed comparative analysis of detection capabilities focusing on closely positioned targets within a cluttered environment. In this particular test image, categories JZ and JS are placed in close proximity to each other amidst a chaotic scene of foreground and background obstructions. The left image in the comparison fails to detect ZW, which is concealed by the overlapping JS elements, thereby illustrating potential weaknesses in handling overlapping objects in dense scenarios. Conversely, the right image demonstrates superior detection capabilities by successfully identifying the concealed ZW, highlighting its enhanced sensitivity to obscured details. The confidence levels further quantify this difference: the left image shows a confidence of 0.72 in recognizing the primary building scene and only 0.32 for JS, suggesting lower reliability. Meanwhile, the right image boosts confidence to 0.80 for the building and 0.63 for JS, indicating a more robust detection mechanism.

Figure 7-1-b illustrates the detection performance with regard to midrange targets, encompassing elements labeled as JZ and JS in the midground and ZW in the distant background. The focal point is on JZ, centrally positioned in the image, where the detection confidence levels are markedly different: 0.59 for the left image and 0.73 for the right. This suggests that the right model better handles midrange details. Additionally, while the left image entirely misses ZW in the foreground pool—a critical miss in terms of situational awareness—the right image competently identifies this target, demonstrating its adeptness at handling varied depth cues and improving foreground detection in complex scenes.

In Figure 7-1-c, the comparison shifts to the detection results within a complex garden landscape, where the depth of field plays a crucial role, and objects like SQ extend the scene's depth. Here, ZW partially obscures JZ, creating a challenging detection scenario. The right image outshines the left by not only recognizing obscured and depthextended objects but also by achieving higher confidence levels across the board: it records a 0.1 higher confidence in detecting JS, a 0.2 increase for JZ, and a significant 0.36 boost for ZW. This illustrates the right model's enhanced capability to interpret and process intricate spatial relationships and partial obstructions more effectively than the left model.

Finally, Figure 7-1-d delves into scenarios with densely packed and complexly arranged targets, where the subjects, such as JZ, are presented in various orientations and depth perceptions ranging from close to long shots. The right image distinctly outperforms the left in recognizing the multifaceted nature of the scene. Specifically, it achieves a 0.11 higher confidence level in identifying JZ, a 0.56 higher level in recognizing JS, and a 0.17 increase for ZW. This superior performance underscores the right model's advanced capabilities in managing scenes with dense arrangements and varied target types, effectively handling depth cues and orientation differences to ensure accurate and reliable detections across a complex visual field.

7.1.2 Discussion of study 2

This section discusses the performance and implications of the Marigold depth estimation model as applied to traditional Jiangnan garden scenes, specifically the Zhuozheng Garden. The Marigold model's ability to accurately render complex natural and architectural elements provides a profound basis for exploring the integration of advanced depth estimation technologies in the preservation and presentation of cultural heritage sites.

Depth Continuity and Smoothness

Marigold's algorithm excels in creating depth maps that are not only visually appealing but also remarkably accurate. The model employs enhanced color gradation techniques that effectively delineate various spatial planes within the garden scene. This is particularly evident in the way Marigold handles the transition from foreground elements such as water bodies and vegetation to architectural features like pavilions in the background. The precision in these gradations allows for a depth perception that closely mimics human visual processing, making the digital representation both immersive and realistic.

Edge Preservation

The sharpness of edge definition is another significant strength of the Marigold model. By maintaining crisp boundaries around key architectural elements, the model ensures that these features stand out against more subtly textured backgrounds such as sky and water. This edge clarity is crucial in maintaining the structural integrity of heritage scenes, where the emphasis on specific architectural details can convey historical and artistic significance.

Handling of Complex Overlapping Structures

One of the standout features of the Marigold model is its ability to manage complex overlapping structures within the garden. Traditional Jiangnan gardens are known for their intricate layouts that often feature overlapping foliage and intertwined architectural elements. Marigold's depth maps effectively separate these overlapping layers, allowing each component to be distinctly appreciated. This separation enhances the viewer's understanding of the garden's layout and design principles, providing a clear insight into the spatial organization and aesthetic intent of the garden architects.

Response to Natural Lighting and Shadows

Marigold’s sophisticated handling of natural lighting and shadows adds another layer of depth and realism to the depth maps. The model’s ability to differentiate shadows from physical features allows it to accurately interpret and represent the dynamic lighting conditions typical in outdoor garden settings. This capability is particularly important in scenes where light plays a significant role in the overall aesthetic, such as the golden hour imagery common in many photographic captures of Jiangnan gardens.

Color and Texture Handling

Algorithms may interpret areas with similar color or texture as being at the same depth, which isn't always the case. Accurate depth maps should transcend these visual similarities to derive actual depth based on structure.

In Jiangnan gardens, similar textures and colors across different depth planes can confuse standard depth estimation techniques. The Marigold model addresses this by integrating texture and color differentiation features within its depth estimation process, enabling it to overcome the challenges posed by visual homogeneity in garden environments.

Multi-Scale Detail Preservation

The fidelity of Marigold in preserving multi-scale details is essential for a comprehensive representation of garden scenes. The model captures fine details such as individual leaves and branches, as well as larger structural elements like stone paths and water features. This multi-level detail preservation is vital for applications that require a detailed analysis of cultural sites, such as virtual tours, educational programs, and preservation planning. By accurately rendering both macro and micro aspects of the scene, Marigold aids in creating a thorough and engaging exploration of heritage sites.

Handling of Transparent and Reflective Surfaces

Depth mapping algorithms can misinterpret transparent or reflective surfaces, often either ignoring them or inaccurately representing their depth. This could be an issue in garden scenes where water bodies or windows are present.

Conclusion

The success of the Marigold model in accurately and beautifully rendering Jiangnan gardens has significant implications for the field of cultural heritage. By providing high-fidelity, three-dimensional visualizations of heritage sites, depth estimation models like Marigold can enhance public engagement and understanding of these cultural treasures. Additionally, the detailed depth maps generated by Marigold can be instrumental in conservation efforts, offering heritage conservators a powerful tool for monitoring and planning restoration works.

Furthermore, the application of such advanced technology in cultural heritage visualization opens up new possibilities for educational outreach. Virtual reality (VR) experiences powered by accurate depth maps can bring remote or fragile heritage sites to life for global audiences, expanding access and fostering a deeper appreciation for cultural diversity and history.

In conclusion, the Marigold depth estimation model represents a significant advancement in the application of machine learning and computer vision technologies to the field of cultural heritage. Its ability to combine aesthetic fidelity with technical accuracy makes it a valuable tool for both conservationists and cultural historians. The ongoing development and refinement of such technologies will undoubtedly continue to enhance our ability to preserve, understand, and celebrate the world’s cultural heritage in increasingly innovative and accessible ways.

7.2 Conclusions

SRO1: To construct an enhanced object detection algorithm tailored for the traditional gardens of the Jiangnan region.

The YOLOv8-modify integrates the Diverse Branch Block (DBB), Bidirectional Feature Pyramid Network（BiFPN）, and Dynamic Head modules （DyHead）to optimize model accuracy, feature fusion, and object detection representational capability, respectively. The enhancements elevated the model's accuracy by 8.7%, achieving a mean average precision (mAP) value of 57.1%.

SRO2: To analyze the depth of the complex spatial relationships in Jiangnan traditional gardens.
Marigold demonstrates its strength in providing a visually rich and detailed depth map that enhances the scene's natural beauty while maintaining the clarity and integrity of critical architectural elements.

The collaborative application of the YOLOv8 object detection system, and the

Marigold depth estimation model introduces a groundbreaking method for the analysis of space in Jiangnan traditional gardens. This integrative approach not only advances our understanding of these historic landscapes but also sets new standards for how we perceive and conserve complex cultural heritage sites. By harnessing the strengths of both advanced object detection and nuanced depth mapping, this methodology offers unprecedented insights into the spatial and contextual relationships within these gardens.

Innovations in Object Detection and Depth Analysis

The YOLOv8 model excels in identifying discrete elements within densely layered images typical of Jiangnan gardens. Its advanced attention mechanisms — scale-aware, spatial-aware, and task-aware — enable precise detection of diverse objects, ranging from subtle botanical details to prominent architectural features. This precision is vital for distinguishing individual components in a space where elements often overlap or blend subtly into one another.

Complementing YOLOv8, the Marigold model offers sophisticated depth perception capabilities that map the physical distances and relationships between these identified elements. Marigold's enhanced color gradation and edge clarity not only delineate objects in three-dimensional space but also reveal the nuanced interactions between them. This depth mapping is essential for visualizing the garden's layout as a cohesive whole, where each element is placed deliberately to create a harmonious and balanced environment.

Synergistic Integration for Comprehensive Analysis

The synergy between YOLOv8's object detection and Marigold's depth mapping facilitates a holistic analysis of space. This integration allows researchers and conservators to view Jiangnan gardens not just as collections of individual elements but as complex ecosystems where architectural and natural components are interdependent. The combined output of these models provides a multi-dimensional view that highlights both the physical and visual connections across the garden.

For instance, YOLOv8 can detect and categorize the garden's pavilions, rocks, trees, and pathways, while Marigold assigns accurate depth values to these elements, illustrating how they relate to each other within the garden's topography. This dual-layer analysis enriches our understanding of traditional Chinese garden design, revealing how elements are not only placed but also perceived in relation to one another.

Enhancing Heritage Conservation and Digital Humanities

This integrated approach revolutionizes heritage conservation practices by providing a tool that can analyze and document the spatial configurations of heritage sites with unprecedented accuracy and detail. For conservation efforts, this means interventions can be better planned and executed with respect to the original intent and design of the garden, preserving its historical authenticity while maintaining its aesthetic integrity.

In the realm of digital humanities, this methodology promotes the development of interactive and educational applications, such as virtual tours that offer a more immersive and informative experience. Users can explore a digital replica of a Jiangnan garden, understanding not only what each element is but also how it fits into the larger design scheme.

Future Prospects and Enhancements

The successful integration of YOLOv8 and Marigold encourages further technological innovations that could expand this approach to other complex heritage sites worldwide. Future enhancements could refine the accuracy of both object detection and depth mapping, perhaps incorporating machine learning algorithms that learn from vast datasets of cultural heritage spaces, thereby improving their predictive accuracy and applicability.

In conclusion, the combined use of YOLOv8 and Marigold provides a new paradigm for the analysis and conservation of Jiangnan traditional gardens. This innovative approach not only deepens our understanding of these cultural landscapes but also enhances our ability to protect and celebrate them, ensuring they are preserved and appreciated for generations to come.

Chapter 8: Limitations, future research and contributions

8.1 Limitations of the Study

8.1.1. Limited Detection of Small Targets

The YOLOv8n-modify model, designed with multilevel and multiscale feature extraction, has shown promising results in recognizing diverse elements within complex environments. However, it faces significant challenges in detecting smaller targets effectively. This limitation is particularly evident in scenarios involving intricate garden landscapes, where small but critical elements like minute rockeries and distant stone bridges might be overlooked. The inability to consistently identify these smaller targets could result from the model's feature representation scales not being sufficiently finegrained to capture the minute details essential for recognizing such elements. This could lead to gaps in data that are crucial for a comprehensive understanding and documentation of cultural heritage sites.

8.1.2 Robustness to Occlusion and Complex Scenes

The YOLOv8n-modify model's effectiveness is further tested under conditions of occlusion and within complex scenic backgrounds. For example, stone bridges or historical buildings that are partially obscured by lush greenery pose a significant challenge for the model. The natural occlusion caused by overlapping vegetation can prevent the model from accurately segmenting and identifying these structures. Similarly, water bodies within gardens present a unique challenge due to their reflective surfaces and varying degrees of transparency. These factors make it difficult for the model to differentiate between the water bodies and similar textural or reflective elements in the surroundings, such as glass or smooth stone surfaces. This issue is compounded by the dynamic nature of water, which can alter appearance based on lighting conditions and the presence of objects within or around it, leading to further confusion and misidentification.

8.1.3 Limitations of the Marigold algorithm

Depth map algorithms like Marigold, used to visualize and interpret spatial information, face several challenges that affect their effectiveness. One significant issue is the accuracy in depth estimation, particularly around complex features like overlapping objects and edges, which often results in artifacts or misrepresentations. Additionally, the use of color gradients to denote depth can be ambiguous without explicit scales, complicating the interpretation of spatial relations. These algorithms also struggle with transparent and reflective surfaces, frequently either omitting them or inaccurately gauging their depth. Furthermore, the computational demands for generating detailed and accurate depth maps are high, which can limit their use in real-time applications. Moreover, the quality of the depth map is heavily reliant on the resolution of the input image, with lower resolutions leading to coarser outputs that may not be suitable for precise tasks.

8.1.4 The Recommendations for Model Improvement

To address these challenges, several enhancements will be considered:

（1） Enhanced Feature Resolution: Improving the granularity of feature extraction, especially at smaller scales, could help the model detect smaller and more subtle elements within complex environments.

（2）Advanced Occlusion Handling: Implementing more sophisticated algorithms that can better handle occlusions, such as contextual awareness and predictive modeling that can infer the presence of partially obscured objects based on visible segments.

（3）Refined Training Strategies: Incorporating more diverse and challenging datasets that specifically include examples of occluded or complex scenes could train the model to better recognize and differentiate such conditions.

（4）Utilization of Supplemental Sensors: Integrating data from other sensory technologies, like LIDAR or structured light 3D scanning, could provide additional depth information that helps in distinguishing overlapping elements in a scene.

By tackling these issues, the YOLOv8n-modify model could be significantly enhanced, making it more reliable and effective in the context of cultural heritage preservation, particularly in the detailed and variable environments of Jiangnan traditional gardens.

8.2 Future Research

8.2.1Future Research on Cultural Heritage Protection

Given the vital role that object detection plays in protecting and preserving cultural heritage, future research will likely focus on improving the precision, adaptability, and applicability of these technologies in different conservation contexts.

(1)Advanced Documentation and Analysis

Research will delve into leveraging highresolution object detection to produce detailed documentation of cultural sites, providing comprehensive databases that capture intricate architectural details, sculptural features, and plant species. This rich data will support further analysis, informing conservation strategies that preserve both the physical and aesthetic integrity of heritage sites.

(1) Predictive Maintenance and Monitoring: Object detection will be coupled with predictive analytics, using historical data and current visual tags to forecast potential areas of concern. Machine learning algorithms could identify patterns of wear or damage and predict future degradation, allowing for timely maintenance that minimizes disruptions to the site's aesthetic and structural qualities.

(3) Restoration Simulation: Future research will explore integrating object detection

with digital modeling to simulate different restoration techniques. This virtual experimentation can help identify optimal materials and procedures for specific artifacts or structures, ensuring their historical authenticity and structural integrity are maintained without risking actual damage.

(4) CrossDisciplinary Approaches: Object detection will benefit from interdisciplinary collaborations, combining expertise in computer vision, art history, horticulture, and materials science. This will enable the creation of sophisticated models that understand both the technical and cultural aspects of heritage sites.

(5) Public Engagement and Accessibility: Researchers will investigate ways to enhance public engagement through virtual and augmented reality experiences powered by object detection. This will help create more interactive educational platforms that provide deeper insight into the historical and cultural significance of heritage sites.

(6) Disaster Preparedness and Management: Object detection could be expanded to include realtime data acquisition and simulation capabilities that assess potential risks from disasters like floods, earthquakes, and fires. This will lead to effective emergency response plans that protect vulnerable cultural assets.

These future research directions will enhance the conservation of cultural heritage sites, making preservation strategies more intelligent, efficient, and accessible while maintaining their rich historical legacy.

8.2.2. Advanced Areas of Algorithmic Research

（1） Advanced Semantic Analysis for Cultural Context Integration.

Future research could integrate deeper semantic analysis into the object detection model to distinguish between similar garden elements with different historical or cultural significances. By leveraging natural language processing (NLP) techniques alongside visual data, the model could analyze historical texts, garden blueprints, and other relevant documents to enrich the contextual understanding of detected objects. This approach would help in creating a more comprehensive database that not only catalogs physical features but also their cultural narratives.

（1）CrossDomain Adaptation for Broader Applicability

Exploring the adaptability of the Jiangnan garden model to other types of cultural heritage sites could be highly beneficial. Techniques such as transfer learning could be employed to adapt the developed model for other regions or types of heritage sites without starting from scratch. This could involve adjusting the model to recognize different architectural styles or garden designs found in other parts of China or globally, providing a scalable solution for cultural heritage preservation.

8.2.3 Digital Twin Research Directions

Building on the foundation of your current research on Jiangnan traditional gardens and incorporating morphological studies of classical Chinese gardens, further integration could explore several advanced avenues:

（1）Enhanced Spatial Analysis Using 3D Point Cloud Technology: Continue to develop and refine the 3D scanning of architectural elements within traditional gardens. By integrating these scans with other technologies like GIS, you can create a comprehensive spatial model. This model would not only depict the physical attributes but could also include temporal changes due to seasonality or human intervention, offering a dynamic view of garden evolution.

（2） RealTime Monitoring and Maintenance Systems: Using IoT sensors within the digital twin framework, develop systems that monitor the health of the garden architecture continuously. Sensors could measure environmental stresses like moisture, wind, and sun exposure, which affect the materials of traditional structures. This data can help predict and prevent damage, ensuring proactive maintenance and preservation.

（3） Simulation and Predictive Modeling: With the digital twin, simulate various conservation scenarios to predict the outcomes of different preservation techniques without physically testing them on the actual structures. This approach would not only save resources but also help in planning effective conservation strategies that minimize risk to the original structures.

（4）Integration of Cultural and Historical Data: Expand the scope of your database to include not just physical characteristics but also historical narratives and cultural significance linked to each element within the gardens. This enriched database would serve as a vital tool for educational programs, virtual tours, and detailed academic studies, enhancing the understanding of Jiangnan garden philosophy.

（5）Advanced Visualization Tools: Develop tools that utilize the data from 3D scans and GIS to create immersive, realistic visualizations of the gardens. These tools could be used for virtual reality experiences, providing a unique way for people to explore and learn about Jiangnan gardens from anywhere in the world.

（6）Interdisciplinary Collaboration: Engage with experts in fields like architecture, landscape architecture, cultural studies, and computer science to create a multidisciplinary approach to your research. This collaboration would enhance the methodological framework and provide comprehensive insights into the traditional and contemporary relevance of Jiangnan traditional gardens.

By pursuing these integrations and advancements in your research, you can significantly contribute to the preservation, understanding, and appreciation of traditional Chinese garden spaces, leveraging cuttingedge technology to bridge historical heritage with modern capabilities.

8.3 Contributions to advance the understanding of historic landscapes research

The integration of YOLOv8 for object detection and the Marigold depth estimation model represents a significant leap forward in the spatial analysis of Jiangnan traditional gardens, offering a methodological superiority over traditional approaches. This synthesis not only revolutionizes the precision and depth of spatial data available but also profoundly impacts the conservation, understanding, and dissemination of knowledge about these historic landscapes.

New Approach to space Analysis

Traditional spatial analysis techniques in heritage conservation often depend on physical surveys, two-dimensional images, and manual interpretation. These methods can be slow, require a lot of manpower, and are prone to mistakes. However, the innovative combination of YOLOv8 and Marigold employs advanced neural network algorithms to automate and refine the detection and spatial analysis of intricate garden scenes. This approach achieves precise identification and three-dimensional positioning of every element within the gardens, from tiny plants to large pavilions, providing a level of detail and accuracy that was previously unattainable.

The ability of these neural networks to quickly process large datasets also improves upon traditional methods, allowing for faster analyses and timely updates on the condition of the gardens. This is vital for continuous monitoring and management of these sites, helping to quickly address any changes or deterioration.

Enhanced Understanding of Spatial of Jiangnan traditional Gardens

The application of these advanced technologies allows for a deeper understanding of the spatial dynamics that characterize Jiangnan gardens. By accurately mapping the depth and relationships between various elements within the gardens, this method provides insights into the intentional design and aesthetic principles that underpin these spaces. This includes an appreciation of the 'borrowed scenery' techniques, the strategic placement of rocks and water bodies, and the intricate pathways that guide visitor movement through the gardens, all of which are designed to create specific views and experiences.

This level of detail extends the capabilities of spatial analysis beyond simple documentation, facilitating a more nuanced interpretation of how these gardens were designed to be experienced and interacted with. Such analyses not only contribute to academic research but also enhance the authenticity with which these gardens are restored and maintained.

Facilitation of Interdisciplinary Collaboration and Innovation

The integration of YOLOv8 and Marigold also encourages interdisciplinary collaboration between computer scientists, landscape architects, historians, and conservationists. This collaborative environment fosters innovation and allows for the application of these advanced tools across different fields, enhancing the scope and impact of research conducted on Jiangnan gardens.

Moreover, the data generated through this integrated approach can serve as a valuable educational resource, providing a basis for virtual tours, augmented reality experiences, and interactive learning platforms. These tools can bring these historic gardens to a global audience, increasing their accessibility and educational value.

In conclusion, the combined use of YOLOv8 and Marigold for spatial analysis in Jiangnan traditional gardens significantly surpasses traditional methodologies, offering a more detailed, accurate, and dynamic understanding of these cultural landscapes. This approach not only enhances the precision of conservation efforts but also deepens our understanding of the historical and cultural narratives embedded within these gardens, contributing profoundly to the field of cultural heritage research.

8.4 Contribution to Knowledge Science

The Knowledge Science at JAIST is designed to integrate various fields such as humanities, social sciences, cognitive science, information technology, natural sciences, and systems science. This interdisciplinary approach is centered on exploring how knowledge is created, accumulated, and utilized, aiming to generate ideas for future societal design and innovation.

The focus of the program on "knowledge creation" is particularly pertinent to addressing complex problems in what is increasingly referred to as a "knowledge society." This term reflects a societal shift where knowledge creation and innovation are paramount, driven by advancements in technology and a growing emphasis on intellectual property and innovation in the global economy.

The research on developing an advanced object detection algorithm for Jiangnan traditional gardens contributes to the field of Knowledge Science in several ways:

（1）Innovation in Knowledge Creation: By creating a new algorithm tailored for complex environments like Jiangnan gardens, and innovative use of Marigold algorithm，the research is pioneering new methods and technologies. This aligns with the Knowledge Science goal of generating innovative solutions to realworld problems.

（2）Interdisciplinary Research: The work intersects technology, cultural heritage, and environmental studies, exemplifying the interdisciplinary research encouraged in Knowledge Science. It integrates technical expertise with cultural and historical awareness, which is essential for holistic knowledge creation.

（3） Practical Application: The application of the research in cultural heritage preservation demonstrates how advanced scientific knowledge can have direct, practical implications. This is crucial for proving the relevance and impact of research in Knowledge Science, as it addresses tangible societal needs.

And this research contributes to both academic insights and practical applications within Knowledge Science:

Academic Contributions:

（1） Advancing Object Detection Methodologies: This study showcases a significant advancement in object detection technologies by adapting and enhancing YOLO v8 to effectively process complex garden environments with intricate structures and dense foliage. This adaptation not only tests the limits of YOLO v8 in novel applications but also expands the methodology for other researchers interested in similar complex environmental contexts.

（2） Dataset Development and Utilization: A specialized dataset of 4890 images that include various angles and lighting conditions of Jiangnan traditional gardens was constructed. This dataset is meticulously annotated and subjected to diverse data augmentation strategies to enhance model robustness and generalization capabilities.

This approach is crucial for the development of reliable object detection systems and provides a blueprint for future dataset constructions in other specialized fields.

（3） Interdisciplinary Approach: This work bridges the gap between technology and cultural heritage, illustrating how advanced computational models can be applied to the study and preservation of historical sites. This interdisciplinary approach not only enriches the field of Knowledge Science but also encourages the integration of diverse academic disciplines.

Practical Contributions:

（1）Interdisciplinary ProblemSolving Approach:

This research adopts an interdisciplinary problemsolving method, integrating knowledge from computer science, architectural design, and cultural heritage conservation. This comprehensive approach not only enhances the authors' expertise in various fields but also improves the efficiency of solving complex heritage conservation issues. By applying advanced computer vision technology, such as the optimized YOLOv8 algorithm, this study accurately identifies and classifies architectural and vegetation features in Jiangnan traditional gardens, providing scientific and technical support for the conservation of buildings and landscapes.

（2）ProjectOriented Approach to Learning New Domain Knowledge:

The projectoriented learning approach allows the researcher to quickly grasp new domain knowledge through practical operation. We must understand complex algorithms and programming skills and have a deep understanding of the history and culture of Jiangnan gardens to ensure that technical solutions are appropriately applied to cultural heritage conservation. This method accelerates the accumulation of knowledge, ultimately driving the practical and developmental aspects of cultural heritage protection through technological innovation.

Through these two practical contributions, the research not only advances heritage conservation at a technical level but also develops the ability to collaborate and innovate across disciplines, providing a solid foundation for future conservation projects. These contributions are significant extensions to both the practice and theory of heritage protection, offering valuable experiences and references for related fields.

List of publications and presentations

Papers published in journals
1.Gao, C., Zhang, Q., Tan, Z., Zhao, G., Gao, S., Kim, E., & Shen, T. (2024). Applying optimized YOLOv8 for heritage conservation: enhanced object detection in Jiangnan traditional traditional gardens. *Heritage Science*, 12(1), 31. (The main framework of the doctoral thesis) https://doi.org/10.1186/s40494024011441 ,Q1
2.Gao, C., Zhao, G., Gao, S. et al. Advancing architectural heritage: precision decoding of East Asian timber structures from Tang dynasty to traditional Japan. Heritage Science 12, 219 (2024). https://doi.org/10.1186/s40494-024-01332-z,Q1

International conference proceedings (including poster presentations)
1. Gao, C., Eunyoung, K., Zhao, G., & Gao, S. (2021, August). Analysis of the impact of the urban traffic noise on the vertical distribution of highrise residential buildings. In IOP Conference Series: Earth and Environmental Science (Vol. 831, No. 1, p. 012080). IOP Publishing. (Published)(The main framework of the minor research)
2. 14th International Symposium on Architectural Interchanges in Asia (ISAIA) 2024,Cultural Heritage for the Future: Strategies for the Digital Preservation and Restoration of Chinese Historical Buildings, 10pages, September 10-13, 2024, Kyoto, Japan.
3. 2024 9th International Conference on Civil Engineering and Materials Science (ICCEMS 2024),Using YOLOv8 for Building Damage Identification in Japan's Noto Region Following Earthquakes: A Deep LearningBased Approach,Accepted,10 pages, July 35, 2024, Singapore.

References

Badgujar, C. M., Poulose, A., & Gan, H. (2024). Agricultural Object Detection with You Look Only Once (YOLO) Algorithm: A Bibliometric and Systematic Literature Review (arXiv:2401.10379). arXiv. https://doi.org/10.48550/arXiv.2401.10379

Barlindhaug, G. (2022). Artificial Intelligence and the Preservation of Historic Documents. Proceedings from the Document Academy. https://doi.org/10.35492/docam/9/2/9

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection (arXiv:2004.10934). arXiv. https://doi.org/10.48550/arXiv.2004.10934

Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., & Li, H. (2024). End-to-end autonomous driving: Challenges and frontiers. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multiscale deep network. Advances in Neural Information Processing Systems, 27, 2366-2374.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 580-587. https://openaccess.thecvf.com/content_cvpr_2014/html/Girshick_Rich_Feature_Hierarchies_2014_CVPR_paper.html

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Jadhav, S., & Kurnthekar, B. M. (2022). Study of Restoration of the Historic Building. Journal of Recent Activities in Architectural Sciences. https://doi.org/10.46610/joraas.2022.v07i01.005

Karna, N. B. A., Putra, M. A. P., Rachmawati, S. M., Abisado, M., & Sampedro, G. A. (2023). Toward Accurate Fused Deposition Modeling 3D Printer Fault Detection Using Improved YOLOv8 With Hyperparameter Optimization. IEEE Access, 11, 74251–74262. https://doi.org/10.1109/ACCESS.2023.3293056

Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R. C., & Schindler, K. (2024). Repurposing diffusion-based image generators for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9492-9502).

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., & Bry, A. (2017). End-to-end learning of geometry and context for deep stereo regression. Proceedings of the IEEE International Conference on Computer Vision, 6675.

Lawrence, B., de Lemmus, E., & Cho, H. (2023). UAS-Based Real-Time Detection of Red-Cockaded Woodpecker Cavities in Heterogeneous Landscapes Using YOLO Object Detection Algorithms. Remote Sensing, 15(4), 883.

Li, Y., Wang, J., Huang, J., & Li, Y. (2022). Research on Deep Learning Automatic Vehicle Recognition Algorithm Based on RESYOLO Model. Sensors.

https://doi.org/10.3390/s22103783

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), Computer Vision – ECCV 2016 (pp. 21–37). Springer International Publishing. https://doi.org/10.1007/9783319464480_2

Liu, Y., Hou, M., Hou, M., Li, A., Dong, Y., Xie, L., & Ji, Y. (2020). Automatic Detection of TimberCracks in Wooden Architectural Heritage Using YOLOv3 Algorithm. ISPRS Archives.

Lo, K. H., Wang, Y. C. F., & Hua, K. L. (2017). Edge-preserving depth map upsampling by joint trilateral filter. IEEE transactions on cybernetics, 48(1), 371-384.

Ma, C., Du, K., & Hamdulla, A. (2023). Small target detection algorithm for flapping wing UAV based on improved YOLOv8. Second International Symposium on Computer Applications and Information Systems (ISCAIS 2023), 12721, 419–425.

Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. arXiv preprint arXiv:2003.08934.

Munin: Artificial Intelligence and the Preservation of Historic Documents. (n.d.). Retrieved May 30, 2024, from https://munin.uit.no/handle/10037/28342

Petracek, P., Kratky, V., Baca, T., Petrlik, M., & Saska, M. (2023). New Era in Cultural Heritage Preservation: Cooperative Aerial Autonomy. IEEE Robotics & Automation Magazine. https://doi.org/10.1109/MRA.2023.3244423

Qiu, Z., Zhao, Z., Chen, S., Zeng, J., Huang, Y., & Xiang, B. (2022). Application of an Improved YOLOv5 Algorithm in Real-Time Detection of Foreign Objects by Ground Penetrating Radar. Remote Sensing. https://doi.org/10.3390/rs14081895

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster RCNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems, 28.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 47(13), 742.

Shi, Y., Wang, N., & Guo, X. (2023). YOLOV: Making Still Image Object Detectors Great at Video Object Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 37(2), Article 2. https://doi.org/10.1609/aaai.v37i2.25320

Soeb, M. J. A., Jubayer, M. F., Tarin, T. A., Al Mamun, M. R., Ruhad, F. M., Parven, A., Mubarak, N. M., Karri, S. L., & Meftaul, I. M. (2023). Tea leaf disease detection and identification based on YOLOv7 (YOLOT). Scientific Reports, 13(1), 6078. https://doi.org/10.1038/s41598023332704

Sun, L., Wang, J., Xiong, R., Shi, Y., Zhu, Q., & Yin, B. (2021, July). Dual Regularization Based Depth Map Super-Resolution with Graph Laplacian Prior. In 2021 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

Urmashev, B., Buribayev, Z., Amirgaliyeva, Z., Ataniyazova, A., Zhassuzak, M., &

Turegali, A. (2021). Development of a weed detection system using machine learning and neural network algorithms. Eastern European Journal of Enterprise Technologies, 6(2), 114.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 1, I–I. https://doi.org/10.1109/CVPR.2001.990517

Wang, N., Liu, H., Li, Y., Zhou, W., & Ding, M. (2023). Segmentation and phenotype calculation of rapeseed pods based on YOLO v8 and mask convolution neural networks. Plants, 12(18), 3328.

Yiting, Li., Qingsong, Fan., Haisong, Huang., Zhenggong, Han., Qian, Gu. (2023). A Modified YOLOv8 Detection Network for UAV Aerial Image Recognition. Drones, doi: 10.3390/drones7050304

Yang, X., Zheng, L., Chen, Y., Feng, J., & Zheng, J. (2023). Recognition of Damage Types of Chinese Gray-Brick Ancient Buildings Based on Machine Learning—Taking the Macau World Heritage Buffer Zone as an Example. Atmosphere. https://doi.org/10.3390/atmos14020346

Zhang, C., & Viola, P. (2007). Multiple-Instance Pruning For Learning Efficient Cascade Detectors. Advances in Neural Information Processing Systems, 20. https://proceedings.neurips.cc/paper_files/paper/2007/hash/ffeabd223de0d4eacb9a3e6e53e5448dAbstract.html

Zhang, H., Liu, G., & Zhao, Q. (2021). Fusion of RGB and LiDAR data for accurate depth prediction in autonomous systems. Robotics and Autonomous Systems, 134, 103110.

Zhang, T., Hu, Y., Lei, T., & Hu, H. (2023). A GIS-based study on the spatial distribution and influencing factors of monastic gardens in Jiangxi Province, China. Frontiers in Environmental Science, 11, 1252231.

Zhu, Q., Yeh, M. C., Cheng, K. T., & Avidan, S. (2006). Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2, 1491–1498. https://doi.org/10.1109/CVPR.2006.119

Zhu, S., Brazil, G., & Liu, X. (2020). The edge of depth: Explicit constraints between segmentation and depth. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13116-13125).

Zhu, X., Lyu, S., Wang, X., & Zhao, Q. (2021). TPHYOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. 2778–2788. https://openaccess.thecvf.com/content/ICCV2021W/VisDrone/html/Zhu_TPHYOLOv5_Improved_YOLOv5_Based_on_Transformer_Prediction_Head_for_Object_ICCVW_2021_paper.html

Zheng, Y., & Wu, G. (2022). YOLOv4Lite-Based Urban Plantation Tree Detection and Positioning With High-Resolution Remote Sensing Imagery. Frontiers in Environmental Science. https://doi.org/10.3389/fenvs.2021.756227

Chinese references

Chen, C. Z. (2016). Gardens and the Cultural Landscape in China [M]. Nanjing: Jiangsu Phoenix Literature and Art Publishing.

Chen Fenfang. (2007). Analysis of literature on Chinese classical gardens.

Gong Xinjun. (2023). Digital expression of the aesthetic connotation of Jiangnan garden cultural heritage - taking Jichang Garden as an example. Beauty and Times (Urban Edition), 04, 118–120.

Gu, H. (2022). Architectural Narratology and the Space of Jiangnan Gardens [Ph.D. dissertation]. Suzhou University of Science and Technology.

Liu, H. Y. (2020). Quantitative Study on the Garden Elements and Environmental Characteristics of Zhuozheng Garden [Ph.D. dissertation]. Tsinghua University.

Peng, Yi Gang. (1996). Analysis of Chinese Classical Gardens. Beijing: China Architecture & Building.

Wang Jue. (2008). From Jiangnan gardens to modern landscapes.

Wang Ming. (2007). Constructing modern Suzhou garden architecture masterpieces. Chinese Urban Economy, 01, 79–81.

Wang Zhigang. (2021). Research on the literature and literary value of Jiangnan garden villas in the Qing Dynasty.

Wang, Z. (2016). Cultural Geography of Jiangnan Gardens [M]. Shanghai: Shanghai Scientific & Technical Publishers.

Zhang Qingping; Li Xia. (2018). Research on the protection of Jiangnan garden cultural heritage based on database construction. Architecture and Culture, 01, 65–67.

Zhang, J. J. (2004). Historical Research on Chinese Gardens [M]. Beijing: China Architecture & Building Press.

Zhou, W. Q. (2010). Classical Gardens of China: History and Artistic Conception [M]. Beijing: China Architecture & Building Press.